A. Details about Dataset

A.1. Data Source

We limit the scope of our study to papers related to Labour Economics. Specifically, we obtained 176 papers from the Journal of Labor Economics (JOLE) [4], covering 22 issues from Volume 37 Number 1 (January 2019) to Volume 41 Number 1 (January 2023). We separately downloaded one paper from Volume 34 Number 3 to be used as a test set for extraction only. All papers were accessed online through a paid subscription, and downloaded as HTML files. Subsequently, we extracted the paper contents using customized scripts and web scraping tools like `BeautifulSoup`. To create our `input_text` based off the papers, we applied a moving window of 5-sentences, moving with a step of 1-sentence per round, was used to generate multiple views of each section.

A.2. Annotation Guidelines

Our annotation guidelines mostly follow the Causal News Corpus (CNC) [5, 6], in that we focus on cause and effect arguments that comprise of consecutive words in the text, and that they must pass the five logical checks for causality. However, to suit our use-case of building a causal knowledge graph from Economics texts, we included many adaptations. Firstly, we allow cross-sentence annotations (up to relations falling within 5-sentences). Secondly, we also do not require arguments to contain events. Finally, our annotation philosophy prioritizes:

- Precision over recall in terms of causal meaning presented in the text (i.e. Causal relations phrased in ambiguous manner will corrupt our final causal graph and are therefore, not annotated.)
- Recall over precision in terms of how specific a causal relation is (i.e. If there is one state of the world where this causal relation could be true, we will annotate the causal relation. This means that causal relations presented in hypotheses are annotated.).

Finally, when in doubt, only causal relations that answered ”Yes” to the question: ”Will this causal relation and the sentences it originated from help inform an Economics academic?” are annotated. A consequence of this check means that the following causal relations were not annotated:

- Purpose types
- Explanations and justifications for a data or methodology choice
- If either argument does not contain any Economic concept or topic
- If given only the causal relations and sentences, the reader cannot understand why the relation is causal (i.e. External knowledge and information is needed).

A.3. Post-processing

Our annotations were done on the INCEPTION tool [17] and post-processed into CSV files. The main post-processing step is to convert the annotations into the 5-sentence `input_text` and the template-based `target_text` for S2S modelling.

[4] https://www.journals.uchicago.edu/toc/jole/current

B. Details about Extraction

An advantage of adopting a S2S framework, as opposed to a token classification framework, especially popular amongst other sequence labelling tasks like Named-Entity Recognition [18, 19], is the ability to extract any number of causal relations without having to modify the last few layers of the model.

C. Details about Clustering

SimCSE was trained to identify whether the relationship betweent two sentences suggests entailment, neutral, or contradition. SimCSE was evaluated against standard semantic textual similarity tasks, and achieved an average 81.6% Spearman's correlation, a 2.2% improvement compared to previous best results. Our embeddings had a feature dimension of 786 because the model is built on the BERT model, `bert-base-uncased` [20].

D. Experimental Setup

For extraction, for all three pre-trained models, `t5-base` [8], `bart-base` [9] and `pegasus-large` [10], had the same training parameters as follows:

- Epoch = 20
- Effective batch size = 16
- gradient_accumulation_steps 2

All other training parameters not mentioned here takes the defaults from Huggingface. In our ablation studies working on external datasets discussed later, we trained on Penn Discourse Tree Bank using 5 epochs and trained on FinCausal using 15 epochs.

E. Evaluation Metrics

- **ROUGE1 and ROUGE2:** ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics used to evaluate the quality of summaries by comparing them to reference summaries. ROUGE1 measures the overlap of unigram (individual words) between the generated summary and the reference summary, while ROUGE2 measures bigram (pair of consecutive words) overlap.
- **ROUGEL and ROUGELSUM:** These are extended versions of ROUGE metrics that consider the longest common subsequences (LCS) between the generated and reference summaries. ROUGEL calculates the LCS-based F1 score, while ROUGELSUM computes the sum of LCS scores over multiple reference summaries.
- **Adjusted Rand Index (ARI):** A measure used to evaluate the similarity or agreement between two data clusterings. It considers the pairs of data points and quantifies the agreement between the cluster assignments, taking into account the possibility of randomness. Scores range from -1 to 1.

Training Data	ROUGE1	ROUGE2	ROUGEL	ROUGELsum
JOLE	**79.90**	**77.65**	**79.25**	**79.65**
PDTB	65.29	61.94	64.40	64.85
FC	75.63	74.18	75.48	75.45
PDTB+FC	76.18	74.64	75.85	75.96
PDTB+FC+JOLE	78.43	76.30	77.68	78.08

Table 2
Performance metrics for cause-effect extraction in a Seq2Seq framework for JOLE test set, when trained on various training sets. Scores are reported in percentages (%). Top score per column is in bold.

- **Fowlkes-Mallows Index (FMI):** An evaluation metric for clustering algorithms. It measures the similarity between two clusterings based on the proportion of pairs of data points that are assigned to the same cluster in both clusterings. Scores range from 0 to 1.
- **Normalized Mutual Information (NMI):** A measure of the mutual information between two clusterings, which quantifies the amount of shared information. It normalizes the mutual information by considering the cluster sizes and the overall entropy of the data. Scores range from 0 to 1.

F. Ablation Studies: Extraction

We investigate the effectiveness of training on external datasets annotated with causal relations to develop a predictive model for causal relations in JOLE. We utilize the Penn Discourse Tree Bank V3.0 (PDTB) [21] and FinCausal 2020+2021 (FC) [22, 23] datasets, converting their annotations into 17873 and 3653 S2S examples, respectively.

Evaluation metrics from Table 1 showcase our results on the JOLE test set. We observe that the best performing model is one that was trained on JOLE alone, achieving a ROUGE1 score of 79.90%. Training on FC yields scores close to the best model, with a ROUGE1 score of 75.63%. However, since FC annotates only causal relations at the sentence-level, it is not reasonable to expect such a model to detect causal relations occurring across sentences that exists in JOLE.

Despite the large number of examples available from PDTB, they do not translate into satisfactory performance on JOLE. Differences in annotation guidelines could explain this discrepancy: PDTB includes Purpose types, while our study does not; and PDTB only annotates causal relations when arguments form clauses, we do not have such a restriction.

Additionally, we explore stepwise training on PDTB, FC,and JOLE (PDTB+FC+JOLE). Surprisingly, despite increased optimization steps and exposure to training examples, the PDTB+FC+JOLE model (ROUGE1 score of 78.43%) does not surpass the performance a model that was trained solely on JOLE. This finding suggests the importance of a specialized dataset tailored to our study, given that our annotation rules aim to cater specifically to assisting Economics academics.

K	ARI	FMI	NMI
5000	**23.77**	32.46	81.14
7500	23.03	**32.95**	**82.47**
10000	18.32	27.51	80.54
12500	18.38	27.96	80.83
15000	17.79	27.99	80.99

Table 3
Performance metrics for argument clustering using MiniBatch K-Means for JOLE. Top score per column is in bold.

G. Ablation Studies: Clustering

Table 3 provides clustering evaluation metric values corresponding to Figure 2, where we experimented with various levels of K to cluster the SimCSE embeddings.

H. Examples of Counfounders Proposed

In this Section, we provide examples of one paper and some of confounders that we detected based on other papers in the literature to be reviewed.

H.1. Local Labor Markets in Canada and the United States

This paper appears in the JOLE Vol 37, No S2. We detected 5 potential confounders based on external sources, and 2 potential confounders mentioned within the paper itself. We report 3 cases of the confounders and include our interpretation of it. In the future, we hope to automatically generate these interpretations too. Apart from proposing confounders, the papers containing these confounders are highly relevant readings that the authors should consider reading up on too.

1.

- **Cause to Effect:** While the magnitude of the effect is slightly smaller in logarithms in Canada, <effect>the effect on the unemployment rate in percentage points is more similar</effect>, since <cause>Canada has on average a higher unemployment rate</cause>.
- **Confounder:** during the Great Recession
- **Confounder to Cause:** Second, we present new evidence on <effect>unemployment dynamics in Canada</effect> <cause> during the Great Recession</cause> and compare Canada to the United States.
  - Paper source: "Long Time Out: Unemployment and Joblessness in Canada and the United States" in JOLE Vol 37, No S2.
- **Confounder to Effect:** Some of the adjusted/unadjusted gap is cyclical; <cause>during the Great Recession</cause> <effect>the gap between the unadjusted and adjusted US unemployment rates rose to 0.4–0.5 percentage points</effect>.

  - Paper source: "Unemployment, Marginal Attachment, and Labor Force Participation in Canada and the United States" in JOLE Vol 37, No S2.
- **Interpretation:** The higher unemployment rate in Canada causes the effect on the unemployment rate in percentage points to be more similar to the United States. However, depending on the time horizon of the study, the occurrence of the Great Recession could be a confounding variable. The Great Recession has shown to correlate with the unemployment dynamics in Canada. The Great Recession is also shown to correlate with the gap between the unemployment rates in United States. Therefore, if the main paper of interest includes data that covers the Great Recession period, some robustness checks to control for this confounding variable is needed to ensure that the observed similarities in the effect on the unemployment rate in percentage points are skewed by the confounding effects of the downturn from the Recession.

2.

- **Cause to Effect:** Column 7 provides evidence that <cause>immigrant population growth in response to local labor demand shocks</cause> may explain <effect>the relatively higher population elasticity in Canada</effect>.
- **Confounder:** duration of stay in the country/ the Great Gatsby Curve in the United States
- **Confounder to Cause:** ... immigrants who stayed in the United States for 11–15 years have lower initial relative earnings than those who left the country prior to 2010. [1] <cause>Conditioning on staying in the United States</cause> increases <effect>measures of assimilation for all education groups</effect>. [2] In this single cross section, <effect>immigrant employment rates</effect> increase <cause>with duration in the United States for all education levels</cause>.
  - Paper source: "Immigrant Earnings Assimilation in the United States: A Panel Analysis" in JOLE Vol 39, No 1.
- **Confounder to Effect:** Chetty et al. (2014) point out that <cause><effect>the Great Gatsby Curve</effect> is present within the United States</cause>, and figure 9 suggests that it <effect>also exists within Canada as well as across the joint landscape of the two countries</effect>.
  - Note that the predicted Effect span is disjointed
  - Paper source: "Intergenerational Mobility Between and Within Canada and the United States" in JOLE Vol 37, No S2.
- **Interpretation:** While the two confounder arguments ("Conditioning on staying in the United States" and "the Great Gatsby Curve is present within the United States") may seem unrelated at first, it is important to consider the potential confounding effects of an immigrant's prior exposure to the United States. An immigrant's prior exposure to the United States could be a confounding variable that the authors should take into account. Immigrants who have been exposed to the United States may have experienced greater assimilation, better employment opportunities, or different levels of income inequality

compared to those who have not been exposed. These factors could influence their decision to move to Canada for career opportunities or affect their susceptibility to income inequality, which in turn could affect the observed population elasticity. For example, immigrants who have experienced better assimilation or employment in the United States may be more likely to consider moving to Canada for career reasons. On the other hand, immigrants who have been exposed to higher income inequality in the United States (as indicated by the Great Gatsby Curve) may be more or less inclined to move to Canada depending on their preferences and aspirations. These factors related to prior exposure to the United States could confound the relationship between immigrant population growth and population elasticity in Canada. Considering these confounding variables and controlling for them in the analysis would help ensure that the observed relationship between immigrant population growth and population elasticity in Canada is not affected by the effects of immigration exposure to the United States.

3.

- **Cause to Effect:** Using an augmented version of the Rosen (1979)–Roback (1982) model, Albouy (2016) and Albouy, Leibovici, and Warman (2013) argue that <effect>level (or persistent) differences in wages across cities in the United States and Canada</effect> are largely driven by <cause>underlying firm productivity</cause>.
- **Confounder:** labor mobility
- **Confounder to Cause:** Table 8 provides a summation of the calculations concerning the effect of <cause>labor mobility</cause> on </effect>LLM productivity<effect>.
  - Paper source: ""Good" Firms, Worker Flows, and Local Productivity" in JOLE Vol 37, No 3.
- **Confounder to Effect:** In this case, <cause>labor mobility</cause> <effect>can reduce wage differences across cities, as workers move to higher-wage areas</effect> (Ganong and Shoag 2018).
  - Paper source: Main paper
- **Interpretation:** Labor mobility acts as a confounder because it influences both the cause (underlying firm productivity) and the effect (level differences in wages). Labor mobility can directly impact productivity by reallocating workers from low-productivity regions or industries to high-productivity regions or industries. Labor mobility can facilitate the diffusion of knowledge and technology across regions. Therefore, labour mobility can affect firm productivity. This confounder-to-cause relation was not addressed by the main paper.