Economics Assistant for Robustness Checks (EconARC): Identifying Confounders from Causal Knowledge Graphs

Fiona Anting Tan, See-Kiong Ng

Institute of Data Science, National University of Singapore

Abstract

In Economics, authors conduct comprehensive robustness checks to prevent drawing misleading conclusions from their causal analyses by accounting for potential confounding factors. To assist in this process, we propose EconARC which offers automated identification of confounders from the literature. Our methodology involves extracting cause-and-effect arguments using a fine-tuned sequence-to-sequence model, clustering semantically similar arguments into topics, and utilizing the backdoor criterion on the causal graph to detect confounders. Our study is the first to employ text mining techniques for generating confounders in Economics, with implications for advancing Artificial Intelligence towards human-level capabilities like engaging in academic discourse.

Keywords

causal text mining, confounder detection, knowledge graphs, backdoor criterion

1. Introduction

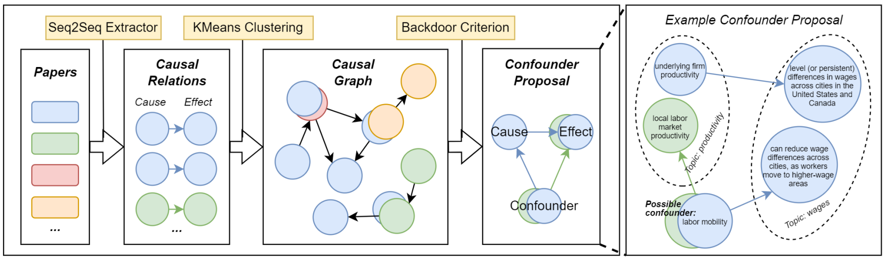
Seq2Seq Extractor
KMeans Clustering
Backdoor Criterion
Papers
...
Causal Relations
Cause
Effect
...
Causal Graph
Confounder Proposal
Cause
Effect
Confounder
Example Confounder Proposal
underlying firm productivity
level (or persistent) differences in wages across cities in the United States and Canada
local labor market productivity
Topic: productivity
can reduce wage differences across cities, as workers move to higher-wage areas
Possible confounder:
labor mobility
Topic: wages

Figure 1: Overview of EconARC

Causal inference relies on addressing confounders, which are variables affecting both the dependent and independent variables. In Economics, consideration of confounders is important, and often a critical part of referee reports. However, staying abreast of confounders in this field requires an extensive knowledge of the literature, which is a non-trivial task given the lengthy and vast number of Economics papers.

ISWC 2023 Posters and Demos: 22nd International Semantic Web Conference, November 6–10, 2023, Athens, Greece

tan.f@u.nus.edu (F. A. Tan); seekiong@nus.edu.sg (S. Ng)

CEUR Workshop Proceedings (CEUR-WS.org)

Existing literature on confounder detection from causal knowledge graphs prioritizes uncovering latent relationships using quantitative variables [1]. Our offers a novel approach by focusing solely on text, bridging the fields of causal text mining and causal identification. We propose the Economics Assistant for Robustness Checks (EconARC), overview shown in Figure 1, that automates robustness check proposal based on confounder identification. To our knowledge, we are the first work to use causal text mining techniques to generate confounders for Economics. We believe that EconARC will be a useful tool for Economics authors to review their paper prior to submission, and reviewers to obtain an unbiased, initial assessment of a paper. Our work also has implications for advancing Artificial Intelligence (AI) towards human-level understanding and inference tasks like engaging in academic discourse.

2. Our Approach

In this section, we outline our methodology, and provide additional details in the Appendix.[1]

2.1. Dataset

Our study experiments on 177 papers from 23 issues of the Journal of Labor Economics (JOLE)[2]. *(1) Annotating Causal Relations:* One of our authors, who is an Econometrics graduate, annotated 5 papers (2,223 sentences) for training and 1 paper (560 sentences) for testing with causal relations. We restricted our annotations to causal relations that appear across 5-sentences in the same section of the paper, and for cause and effect arguments to be consecutive spans. We mainly adapted the annotation guidelines from the Causal News Corpus (CNC) [2, 3] with the key difference being that our annotated causal relations must be helpful to an Economics academic. A consequence of this rule means that we differ from CNC in areas like our arguments need not contain events, and we do not annotate: (1) Purpose relations, (2) justifications for a data or methodology choice, etc. In total, 522 and 76 causal relations were annotated in the training and test set respectively. *(2) Annotating Argument Topics:* For two papers from the training set, the same annotator assigned open-ended topic labels to each argument. 356 arguments were annotated to 119 topics. Topic labels were as general as *"education"*, to more specific labels like *"greater upward mobility"*, *"areas where fathers tend to be richer within the bottom half of households and whose sons did better accordingly"*, etc.

Sequences without annotations will be referred to as our Out-of-Sample (OOS) set. The OOS set comprises of 83,676 sentences from 171 papers.

2.2. Extraction of Causal Relations

We fine-tune SOTA sequence-to-sequence (S2S) pre-trained language models (PTMs), like `t5-base` [4], `bart-base` [5] and `pegasus-large` [6]. Given the `input_text`, the model learned to generate the `target_text`. These texts are described below:

1. `input_text`: An input sequence that is 5-sentences long with a "`summarize: `" prefix.

[1] Our repository is available at https://github.com/tanfiona/EconARC.
[2] https://www.journals.uchicago.edu/toc/jole/current

2. `target_text`: If no causal relations were annotated within the `input_text`, return "No key causal relations". Else, return a line-separated list of causal relations in the format of "Key causal relations:`\n1. Cause: <FIRST_CAUSE_SPAN>\tEffect:<FIRST_EFFECT_SPAN>\n`..."

2.3. Knowledge Graph Creation

To prevent a sparse graph, we grouped arguments with similar meaning into a topic. Similar to [7, 8], we approached this task by (1) generating word embeddings and (2) clustering the embeddings. We concatenated the annotated causal relations from the training set and the inferred causal relations from the OOS set together when performing clustering. For (1), to encode arguments into embeddings, we experimented with PTMs like the supervised pre-trained language model by SimCSE [9] and the encoder portion of our fine-tuned T5 extraction model. For (2), we condensed our embeddings into 400 components[3] using Principal Components Analysis (PCA) and used Mini-Batch K-Means [10] to perform our clustering. We explored various levels of K (5000 to 15000, jumping by gaps of 2500). We removed relations where the cause and effect have the same topic to avoid nodes with self-loops. Our knowledge graph (KG) $G = (V, E)$ is a collection of nodes $V = \{(v_1, v_2, ..., v_n)\}$ and directed edges $E = \{(v_1, v_2), (v_2, v_3), ...\}$. A directed edge (v_x, v_y) represents the presence of causality between the two nodes, where v_x is the cause argument and v_y is the effect argument. The edges are also weighted by support s, indicating the count of relations expressing causality from v_x to v_y in the dataset.

2.4. Confounder Detection

Given an ordered pair of variables (X, Y) in a directed acyclic graph G, a set of variables Z satisfies the **backdoor criterion** relative to (X, Y) if no node in Z is a descendant of X, and Z blocks every path between X and Y that contains an arrow into X [1]. Backdoor paths may make X and Y dependent despite lacking causal influences from X. To estimate the causal relationship of X on Y, we need to condition on a set of nodes Z such that Z (1) blocks all spurious paths between X and Y, (2) leaves directed paths between X and Y unchanged, and (3) creates no new spurious paths. In other words, the causal effect of X on Y is given by the formula: $P(Y = v_y|do(X = v_x)) = \sum_{v_c} P(Y = v_y|X = v_x, Z = v_z)P(Z = v_z)$. This formula describes the distribution of Y given an intervention ($do(X = v_x)$) that sets X to the value v_x, thereby removing X's dependence on Z.

Given a Cause and Effect argument, we automatically identify potential confounders using adapted backdoor criterion scripts from DoWhy [11], a Python package for Causal Inference. A depth-first search algorithm to explore paths between the Cause and Effect pair and determines the variables that need to be conditioned on to block all paths between them. Since our whole graph is too large, we had to restrict our search space to improve run times: For each node in the graph G, we obtained a subgraph (sG) by considering nodes falling within the radius of 2 units around it. For each node in sG that is not the center node, we designate it as the target node, while the center node was fixed as the source. This setup enables our search for backdoor variables between every pair of nodes in G within a feasible run time. However, our methodogy fails to identify backdoor variables that lie outside of each subgraph.

[3] With 400 components, only 5.005^{-05} of variance is dropped.

(A) Extraction (Seq2Seq Model)

PTM	ROUGE1	ROUGE2	ROUGEL	ROUGELsum
T5	**79.90**	**77.65**	**79.25**	**79.65**
Pegasus	66.27	63.07	65.51	65.73
BART	76.86	73.76	75.97	76.48

(B) Clustering (MiniBatch K-Means Model)

PTM	K	ARI	FMI	NMI
SimCSE	7500	**23.03**	**32.95**	**82.47**
SimCSE	10000	18.32	27.51	80.54
T5	7500	15.81	21.53	77.64
T5	10000	12.19	21.50	80.11

Table 1
Performance metrics for (A) cause-effect extraction in a S2S framework and (B) argument clustering using Mini-Batch K-Means. Scores are reported in percentages (%). Top score per column is in bold. Explanations for evaluation metrics are available in the Appendix.

3. Results & Conclusion

Panel A of Table 1 report scores for extraction. Across all metrics, our best model was the S2S model that fine-tuned T5, scoring 79.90% for ROUGE1 and 76.25% for ROUGEL[4]. Hence, we used this best model on our OOS set to obtain predicted causal relations. Panel B of Table 1 report scores for clustering. Across all metrics, our best model uses SimCSE embeddings and performs K-Means clustering for 7500 topics. This best model scored 23.03% for ARI and 82.47% for NMI. Using the SimCSE embeddings consistently supercedes using T5's, suggesting benefits in clustering arguments that were converted to embeddings that convey semantic similarity. For our best model, our KG comprises of 7498 unique nodes, 37557 edges, and an edge support ranging from 1 to 10 and averaging at 1.207.[5] Finally, we apply the backdoor criterion detection algorithm to identify confounders. For our dataset of 176 papers, we identified 152 papers and 676 confounders for authors to consider reviewing. These confounders lie 1 to 4 steps away from either the cause or effect argument of the main relation. We also detected 161 papers and 1408 confounders that the authors themselves describe within their paper, which reveal that confounders and robustness checks are definitely a key concern and covered by most authors.

In conclusion, EconARC successfully uses causal text mining techniques to automatically identifying confounders. EconARC will be a useful tool for economics authors or for referees to critically evaluate the validity of their causal identification strategy. This tool will also help mitigate reviewers' bias to some extent because the feedback provided is automated. We hope to expand the coverage of our work to more journals and to more branches of Economics, and to evaluate our system with Economic academics.

References

[1] J. Pearl, M. Glymour, N. P. Jewell, Chapter 3: The effects of interventions, in: Causal inference in statistics: A primer, John Wiley & Sons, 2016.

[2] F. A. Tan, A. Hürriyetoğlu, T. Caselli, N. Oostdijk, T. Nomoto, H. Hettiarachchi, I. Ameer, O. Uca, F. F. Liza, T. Hu, The causal news corpus: Annotating causal relations in event sentences from news, in: Proceedings of the Thirteenth Language Resources and Evalua-

[4] We use ROUGE evaluation metrics because the task was framed as an open-ended generation task. W

[5] Due to limited space, we provide experimental details, explanation of our evaluation metrics, ablation studies and provide qualitative examples of confounders in the Appendix.

tion Conference, European Language Resources Association, Marseille, France, 2022, pp. 2298–2310. URL: https://aclanthology.org/2022.lrec-1.246.

[3] F. A. Tan, H. Hettiarachchi, A. Hürriyetoğlu, N. Oostdijk, T. Caselli, T. Nomoto, O. Uca, F. F. Liza, S.-K. Ng, RECESS: Resource for extracting cause, effect, and signal spans, in: Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Bali, Indonesia, 2023.

[4] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[5] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7871–7880. URL: https://aclanthology.org/2020.acl-main.703. doi:10.18653/v1/2020.acl-main.703.

[6] J. Zhang, Y. Zhao, M. Saleh, P. J. Liu, PEGASUS: pre-training with extracted gap-sentences for abstractive summarization, in: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 11328–11339. URL: http://proceedings.mlr.press/v119/zhang20ae.html.

[7] S. Sia, A. Dalmia, S. J. Mielke, Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too!, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 1728–1736. URL: https://aclanthology.org/2020.emnlp-main.135. doi:10.18653/v1/2020.emnlp-main.135.

[8] Z. Zhang, M. Fang, L. Chen, M. R. Namazi Rad, Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 3886–3893. URL: https://aclanthology.org/2022.naacl-main.285. doi:10.18653/v1/2022.naacl-main.285.

[9] T. Gao, X. Yao, D. Chen, SimCSE: Simple contrastive learning of sentence embeddings, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 6894–6910. URL: https://aclanthology.org/2021.emnlp-main.552. doi:10.18653/v1/2021.emnlp-main.552.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[11] A. Sharma, E. Kiciman, Dowhy: An end-to-end library for causal inference, CoRR abs/2011.04216 (2020). URL: https://arxiv.org/abs/2011.04216. arXiv:2011.04216.

[12] J.-C. Klie, M. Bugert, B. Boullosa, R. E. de Castilho, I. Gurevych, The inception platform:

Machine-assisted and knowledge-oriented interactive annotation, in: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, Association for Computational Linguistics, 2018, pp. 5–9. URL: http://tubiblio.ulb.tu-darmstadt.de/106270/, event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

[13] Y. Yang, A. Katiyar, Simple and effective few-shot named entity recognition with structured nearest neighbor learning, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6365–6375. URL: https://aclanthology.org/2020.emnlp-main.516. doi:10.18653/v1/2020.emnlp-main.516.

[14] C. Li, Y. Liu, Improving named entity recognition in tweets via detecting non-standard words, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Beijing, China, 2015, pp. 929–938. URL: https://aclanthology.org/P15-1090. doi:10.3115/v1/P15-1090.

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[16] B. Webber, R. Prasad, A. Lee, A. Joshi, The penn discourse treebank 3.0 annotation manual, Philadelphia, University of Pennsylvania (2019).

[17] D. Mariko, H. Abi-Akl, E. Labidurie, S. Durfort, H. De Mazancourt, M. El-Haj, The financial document causality detection shared task (FinCausal 2020), in: Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation, COLING, Barcelona, Spain (Online), 2020, pp. 23–32. URL: https://aclanthology.org/2020.fnp-1.3.

[18] D. Mariko, H. A. Akl, E. Labidurie, S. Durfort, H. de Mazancourt, M. El-Haj, The financial document causality detection shared task (FinCausal 2021), in: Proceedings of the 3rd Financial Narrative Processing Workshop, Association for Computational Linguistics, Lancaster, United Kingdom, 2021, pp. 58–60. URL: https://aclanthology.org/2021.fnp-1.10.

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) [6], 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) [2**?**], 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 [12] 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.

[6] 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 [13, 14], 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` [15].

D. Experimental Setup

For extraction, for all three pre-trained models, `t5-base` [4], `bart-base` [5] and `pegasus-large` [6], 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) [16] and FinCausal 2020+2021 (FC) [17, 18] 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.

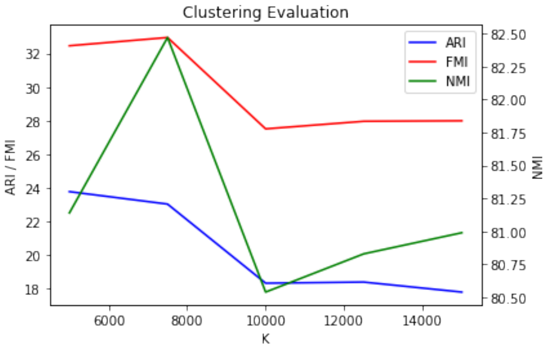

Figure 2: Clustering using different K values

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.

From Figure 2, the ideal number of clusters is 7500 for 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.