Assessing Causality Structures learned from Digital Text Media

Mariano Maisonnave
Depto. de Cs. e Ing. de la Computación, Instituto de Cs. e Ing. de la Computación (ICIC UNS–CONICET)
Bahía Blanca, Argentina
mariano.maisonnave@cs.uns.edu.ar

Fernando Delbianco
Depto. de Economía, Instituto de Matemática de Bahía Blanca (INMABB UNS-CONICET)
Bahía Blanca, Argentina
fernando.delbianco@uns.edu.ar

Fernando Tohmé
Depto. de Economía, Instituto de Matemática de Bahía Blanca (INMABB UNS-CONICET)
Bahía Blanca, Argentina
ftohme@criba.edu.ar

Ana G. Maguitman
Depto. de Cs. e Ing. de la Computación, Instituto de Cs. e Ing. de la Computación (ICIC UNS–CONICET)
Bahía Blanca, Argentina
agm@cs.uns.edu.ar

Evangelos E. Milios
Faculty of Computer Science, Dalhousie University
Halifax, Nova Scotia, Canada
eem@cs.dal.ca

ABSTRACT

In this paper we describe a framework to uncover potential causal relations between event mentions from streaming text of news media. This framework relies on a dataset of manually labeled events to train a recurrent neural network for event detection. It then creates a time series of event clusters, where clusters are based on BERT contextual word embedding representations of the identified events. Using these time series dataset, we assess four methods based on Granger causality for inferring causal relations. Granger causality is a statistical concept of causality that is based on forecasting. It states that a cause occurs before the effect, and the cause produces unique changes in the effect, so past values of the cause help predict future values of the effect. The four analyzed methods are the pairwise **Granger test**, **VAR(1)**, **BigVar** and **SiMoNe**. The framework is applied to the New York Times dataset, which covers news for a period of 246 months. This preliminary analysis delivers important insights into the nature of each method, identifies differences and commonalities, and points out some of their strengths and weaknesses.

CCS CONCEPTS

• **Mathematics of computing** → **Causal networks**; *Time series analysis*; • **Computing methodologies** → **Information extraction**.

KEYWORDS

Granger Causality, Event Detection, Time Series

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
DocEng '20, September 29-October 2, 2020, Virtual Event, CA, USA
© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8000-3/20/09…$15.00
https://doi.org/10.1145/3395027.3419594

ACM Reference Format:
Mariano Maisonnave, Fernando Delbianco, Fernando Tohmé, Ana G. Maguitman, and Evangelos E. Milios. 2020. Assessing Causality Structures learned from Digital Text Media. In *ACM Symposium on Document Engineering 2020 (DocEng '20), September 29-October 2, 2020, Virtual Event, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3395027.3419594

1 INTRODUCTION

Prediction and explanation are essential tasks in almost any scientific discipline and inferring causality relations is a major step towards achieving these tasks. In particular, the use of text data to predict events is a central theme within the field of machine learning. Text data collected through digital can be processed using information extraction techniques to identify mentions of events to define event variables. These and other relevant variables identified from external sources can be used to build prediction models. However, the generation of such models from large volumes of data often lacks an interpretation that reveals the causal impact among different variables [28]. While there are some contributions to the problem of causal modeling within the Computer Science discipline [1, 3, 21], most of the work developed in the area of machine learning has addressed the problem of "pure prediction", without emphasis on causal analysis. On the other hand, although the study of the concept of causality is a central and long-standing issue in the field of Econometrics [12], the relatively recent availability of large volumes of text data opens up new opportunities to conjecture on possible causality relations among events described in the news as well as to test them.

This paper analyzes four different methods that allow to uncover the causal relationship between news events obtained from the full New York Times archive (NYT). The analyzed methods infer causality between pair of variables in a time series by taking a statistical approach. The methods studied are the pairwise **Granger test** [13], **VAR(1)** [25], **BigVar** [20] and **SiMoNe** [4, 5].

As a first step in the analysis, we used a dataset manually labeled by experts to train a recurrent neural network model (RNN) that was used to detect additional events on the NYT corpus. Then,

a time series of event clusters was created on BERT contextual word embedding representations of the event trigger names [7]. Finally, each of the four methods was applied on the time series of event clusters to find patterns of correlation that allow to uncover potential causal relations between the identified events.

2 RELATED WORK

There are several previous works that have addressed the problem of automatically or semi-automatically constructing causal networks or identifying causal relationships from large volumes of text [6, 11, 22–24]. Probabilistic graphical models [15] represent data and their dependency relationships, allowing to combine uncertainty and logical inference through the use of independence restrictions. Bayesian networks [21] are a kind of probabilistic graphical model where the graph encoding random variables and their conditional dependencies is directed and acyclic. The methods traditionally applied for the inference of graph structures are usually based on constraints [27], based on scores [10] or a combination of both [18].

In the field of Econometrics the technique known as *Granger Causality test* [12] is based on two principles: (1) a cause occurs before the effect, and (2) the cause produces unique changes in the effect, so past values of the cause help predict future values of the effect. In the case of structural models, different procedures have been developed to implement Granger's test, mostly based on extensions of the **VAR** model [14, 26]. A core idea in these approaches is the identification of Granger causal relations between variables with the possibility of using the "cause" to forecast values of the "effect" [9, 19]. These contributions focus only on quantitative data. Here we extend the application of this approach to relations among events, aiming to detect a network of causal links among them as in [2], although using the aforementioned methods.

3 A FRAMEWORK FOR THE ANALYSIS OF CAUSAL RELATIONS

3.1 Event Detection

The analyzed methods rely on a dataset of manually labeled events and a dataset of events automatically detected by an RNN model, to which we refer to as $\mathcal{E}_L$ and $\mathcal{E}_D$, respectively. We briefly outline here how these datasets were created (a detailed description can be found in [17]). The $\mathcal{E}_L$ dataset contains 2200 news extracts from the NYT archive associated with three episodes of real-world crises: the Mexican peso crisis of 1994, the Russian financial crisis of 1998, and the Asian financial crisis of 1997. Each of the news extracts was analyzed and labeled by four users that employed a consensus-based approach to identify event triggers (verbs or nouns that most clearly express the occurrence of the events). The resulting labeled dataset contains 94 event trigger names (represented by unique lemmas) and 2828 event mentions (total number of event occurrences).

The RNN model was trained using 2000 labeled news extracts while the remaining 200 labeled extracts were used to test the model. The labeling of the 2000 news extracts used for training was assisted by a simple active learning tool. It is worth mentioning that the RNN model developed for event detection combines state-of-the art features, such as BERT embeddings to define contextual word and contextual sentence embeddings, and achieves highly competitive results as reported in [17]. The code for the RNN model, as well as the dataset are made available to the research community for reproducibility and data reuse.[1]

The $\mathcal{E}_D$ dataset consists of the event mentions predicted by the RNN model. Only the 94 event trigger names identified in $\mathcal{E}_L$ were considered but the full NYT archive was used, resulting in a total of 21,449,746 event mentions covering a period of 246 months (from January 1987 to June 2007).

3.2 Event Clustering and Time Series Creation

We consider only the lowercase version of the event trigger names (the Spacy library was used for lemmatization). Since different event trigger names were associated with the same (or similar) type of event (e.g., *rise* and *increase*) we decided to group event trigger names that represented semantically similar events into a cluster represented by one event variable. To achieve this, we first generated BERT contextual word embeddings [7]. We used the sum of the last four hidden unit layers as the embedding for each word because, as shown in [7], this sum provides one of the best performance while maintaining a relative short embedding size (768 dimensions). Then we applied the DBSCAN clustering technique [8] on the contextual word embedding representations of the event trigger names. The minimum cluster size was set to two and cosine similarity was used to compare instances. Lastly, to avoid large clusters with mixed semantics, we set a low threshold value for two instances to be considered similar (epsilon = 0.15). As a result, we obtained 1222 clusters of event trigger names and considered each of these clusters as an event variable. To facilitate interpretation and visualization of the graphs, we limited the analysis only to those event trigger names that were mentioned more than 15 times in the dataset. This resulted in 94 different event trigger names and 44 clusters. Hence, the number of event variables considered in the reported analysis is 44, with clusters containing more than one event trigger names represented by one of those names.

After completing the clustering and filtering stages, we built two monthly time series, from $\mathcal{E}_L$ and $\mathcal{E}_D$ to which we refer to as ϵ_L and ϵ_D, respectively. Since the number of event variables is 44 and the NYT dataset spans 246 months, the dimension of both time series is 44×246. For each time series, the ith event variable is represented as $EV_i = \{EV_{i,1}, EV_{i,2}, ..., EV_{i,246}\}$, where $EV_{i,j}$ is the number of mentions of the ith event variable in the jth month. To compute $EV_{i,j}$ in ϵ_L we searched for each of the event trigger names associated with the ith event variable occurring in the $\mathcal{E}_L$ dataset during the jth month. The ϵ_D time series was computed from $\mathcal{E}_D$ in the same way.[2]

3.3 Inference Methodologies

In this paper we assume the definition of *causality* introduced in Econometrics by the Nobel Prize winner Clive Granger [12]. Given two time series corresponding to the values of two variables, X and Y, X is said to Granger-cause Y if (1) the values of X precede those of Y, and (2) the values of X allow to forecast the values of Y.

In this sense, *causality* is associated with the capability of *forecasting* values of Y based on past values of X. In our case, where

[1] The code is available at http://cs.uns.edu.ar/~mmaisonnave/resources/ED_code/ and the dataset is available at http://cs.uns.edu.ar/~mmaisonnave/resources/ED_data/.
[2] The code is available at https://cs.uns.edu.ar/~mmaisonnave/resources/TSE_code/ and the time series at https://cs.uns.edu.ar/~mmaisonnave/resources/TSE_data/.

the variables represent events, the intuition is that when an event causes another (e.g., *crisis* causes *deficit*), past values of the time series associated with one event variable will have unique information that would help to forecast the other. In this paper we extend this to the case of several variables [16]. This amounts, according to [13], to running statistical F tests on the lagged coefficients of an X variable, for the same number of lags in Y.

The main potential issue with this approach on many variables, not ordered a priori, is that evaluating Granger-causal relations between pairs of variables may yield too many causal relations that might be, with high probability, spurious. The alternative is to test a system of equations, in the style of **VAR** models (Vector Auto Regression) [25], where X is now a vector of variables, A the matrix of coefficients of the system of equations, $\mathbf{b}$ is a vector of constants and ε a vector of perturbations, distributed normally around 0:

$$X_t = X_{t-1}\mathbf{A} + \mathbf{b} + \varepsilon_t, \quad t \in \mathbb{N}.$$

This is not without disadvantages, ensuing from the dimensionality problem of aggregating all the possible relations and the corresponding lags (the values of t) in each equation of the system.

The next sensible step is thus to keep the system but penalize the addition of new variables in each equation. Penalization, or regularization, is introduced to generate more stable, and more interpretable models, avoiding overfitting and producing models less sensitive to the noise in our data. Penalization in this context means that when the positive impact in forecasting introduced by a variable is small, the increase could be due to statistical noise rather than a true causal relation. Therefore, we introduce penalization to consider the most robust links only. We follow here two different, although essentially similar, penalization techniques. On one hand, **BigVar**[3] [20], and on the other **SiMoNe**[4] [4, 5]. Both postulate a **VAR(1)** model (i.e. a one-period lag) and deem *causal* those variables that are statistically significant.

4 RESULTS

We run four processes on our dataset. In the first place, we run pairwise **Granger tests** with 4 lags. The results, shown on Figure 1 and 2 for ϵ_L and ϵ_D, respectively, and in Table 1, yield a dense network of relations. Then, we estimate a **VAR(1)** model, which provides a sparser graph, although with still a large number of edges. We can reasonably hint that some of those relations are still spurious. So, finally, we run two estimations of the **VAR(1)** model, with the penalties imposed by **BigVar** and **SiMoNe**, respectively. In the former case we use a basic penalization on the coefficients of the matrix A, assuming a forecast horizon of one period. In the case of **SiMoNe** we choose the best model (i.e. the graph) according to an information criterion (BIC), penalizing the structure of connections among nodes and not the nodes themselves. This technique requires imposing a network structure by an a priori version P of matrix A.

With both **BigVar** and **SiMoNe** we obtain sparser but not identical graphs. This is because the selection of non-zero coefficients is not independent of the method of penalization applied.

We also looked into the question of how similar the inferred causal structures are. Consider two causal structures represented by the graphs $G_1 = (N_1^E, E_1^C)$ and $G_2 = (N_2^E, E_2^C)$ where N_i^E and E_i^C represent the set of nodes (event clusters) and edges (causal relations), respectively, in graph G_i for $i = 1, 2$, with $N_1^E = N_2^E$ (i.e., the causal relations are defined on the same set of event clusters). We use Jaccard similarity, computed as $\text{Jaccard}(E_1^C, E_2^C) = |E_1^C \cap E_2^C| / |E_1^C \cup E_2^C|$, to measure the similarity between G_1 and G_2. Table 2 presents the Jaccard similarities computed between each pair of structures obtained by the analyzed methods.

Granger		VAR(1)		BigVar		SiMoNe	
ϵ_L	ϵ_D	ϵ_L	ϵ_D	ϵ_L	ϵ_D	ϵ_L	ϵ_D
273	538	135	188	40	21	17	167

Table 1: Number of relations (edges) in causality structures inferred from ϵ_L and ϵ_D.

	Granger		VAR(1)		BigVar		SiMoNe	
	ϵ_L	ϵ_D	ϵ_L	ϵ_D	ϵ_L	ϵ_D	ϵ_L	ϵ_D
Granger	1	1	0.172	0.093	0.068	0.016	0.018	0.055
VAR(1)			1	1	0.048	0.005	0	0.134
BigVar					1	1	0.096	0.022
SiMoNe							1	1

Table 2: Jaccard similarity between sets of relations (edges) in causality structures inferred from ϵ_L and ϵ_D.

5 DISCUSSION AND FUTURE WORK

The analysis we carried out should be understood as a first step towards the goal of inferring causal relations among events detected in news text streams. The results found at this point are crucially dependent on tools chosen for carrying out the task. It is rather evident that the relations inferred are not robust and vary according to the estimation technique applied.

The **VAR** procedure seems appropriate to the task of detecting Granger causality in a set of clusters of events. This is also the case of the further step of penalizing some elements of the model to mitigate the dimensionality problem of estimating models with many variables and many lags, even with only a subset of the possible clusters of events. But **BigVar** and **SiMoNe**, the tools used to implement the penalization, differ in the nature of the elements that become penalized. The different results, indicated by the low Jaccard similarity indexes among the outcomes of the different techniques point towards the existence of different causal structures underlying the same set class of events. The penalization used by each method uncovers a distinct kind of Granger causal relation. All together, they provide a rich picture of the relationship among events, providing ways to forecast future ones in terms of past and current data in the news.

The next step in the research agenda reported here involves the development of criteria to select the most appropriate penalization strategy according to the data analyzed. A possibility to be explored is to penalize *both* the coefficients in the matrix and the network structure. The goal is to obtain sparse causal graphs, yielding intuitively acceptable cause-effect relations.

ACKNOWLEDGEMENTS

This research work was supported in part by CONICET (Argentina), a LARA Google Research grant, the Emerging Leaders in the Americas Program (ELAP-Canada) and Universidad Nacional del Sur (PGI-UNS 24/N051 and 24/E145).

[3] https://CRAN.R-project.org/package=BigVAR
[4] https://CRAN.R-project.org/package=simone

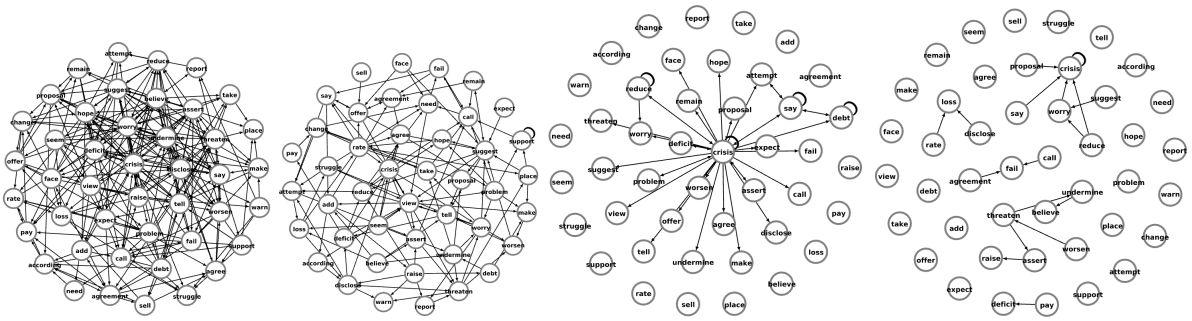

Figure 1: Causality structures inferred from ϵ_L. From left to right: Granger, VAR(1), BigVar and SiMoNe.

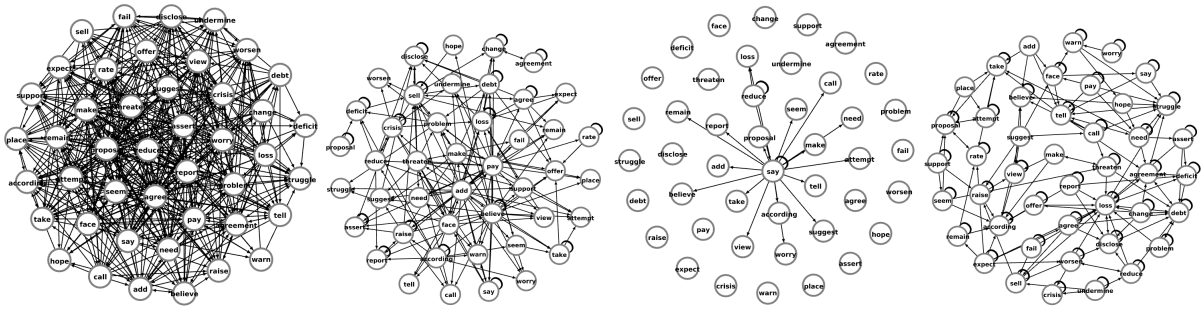

Figure 2: Causality structures inferred from ϵ_D. From left to right: Granger, VAR(1), BigVar and SiMoNe.

REFERENCES

[1] Jayadev Acharya, Arnab Bhattacharyya, Constantinos Daskalakis, and Saravanan Kandasamy. 2018. Learning and testing causal models with interventions. In *Advances in Neural Information Processing Systems*. MIT press, Montréal, Canada, 9447–9460.

[2] Ananth Balashankar, Sunandan Chakraborty, Samuel Fraiberger, and Lakshminarayanan Subramanian. 2019. Identifying Predictive Causal Factors from News Streams. In *Proceedings of EMNLP-IJCNLP 2019*. Association for Computational Linguistics, Hong Kong, China, 2338–2348. https://doi.org/10.18653/v1/D19-1238

[3] Elias Bareinboim and Judea Pearl. 2015. *Causal inference from big data: Theoretical foundations and the data-fusion problem.* Technical Report. DTIC Document.

[4] Camille Charbonnier, Julien Chiquet, and Christophe Ambroise. 2010. Weighted-LASSO for structured network inference from time course data. *Statistical applications in genetics and molecular biology* 9, 1 (2010), 15.

[5] Julien Chiquet, Alexander Smith, Gilles Grasseau, Catherine Matias, and Christophe Ambroise. 2009. Simone: Statistical inference for modular networks. *Bioinformatics* 25, 3 (2009), 417–418.

[6] Rahim Dehkharghani, Hanefi Mercan, Arsalan Javeed, and Yucel Saygin. 2014. Sentimental causal rule discovery from Twitter. *Expert Systems with Applications* 41, 10 (2014), 4950–4958.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[8] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*. AAAI Press, Portland, OR, USA, 226–231.

[9] Hao Fang. 2018. *Multivariate density forecast evaluation and nonparametric Granger causality testing*. Universiteit van Amsterdam, Amsterdam, Netherlands.

[10] Nir Friedman and Daphne Koller. 2003. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine learning* 50, 1-2 (2003), 95–125.

[11] Roxana Girju and Dan Moldovan. 2002. Text Mining for Causal Relations. In *In Proceedings of the FLAIRS Conference*. AAAI Press, Pensacola, FL, USA, 360–364.

[12] Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* 37, 3 (1969), 424–438.

[13] Clive WJ Granger. 1988. Some recent development in a concept of causality. *Journal of econometrics* 39, 1-2 (1988), 199–211.

[14] Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Hoyer Patrik. 2010. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research* 11 (2010), 1709—-1731.

[15] Daphne Koller and Nir Friedman. 2009. *Probabilistic graphical models: principles and techniques*. MIT press, Cambridge, MA, USA.

[16] Kumar Mainali, Sharon Bewick, Briana Vecchio-Pagan, David Karig, and William F Fagan. 2019. Detecting interaction networks in the human microbiome with conditional Granger causality. *PLoS computational biology* 15, 5 (2019), e1007037.

[17] Mariano Maisonnave, Fernando Delbianco, Fernando Tohmé, Ana Maguitman, and Evangelos Milios. 2020. Improving Event Detection using Contextual Word and Sentence Embeddings. arXiv preprint arXiv:2007.01379.

[18] Nicolai Meinshausen and Peter Bühlmann. 2006. High-dimensional graphs and variable selection with the lasso. *The annals of statistics* 34, 3 (2006), 1436–1462.

[19] Peter Molenaar. 2019. Granger causality testing with intensive longitudinal data. *Prevention Science* 20(3) (2019), 442–451.

[20] William Nicholson, David Matteson, and Jacob Bien. 2017. Bigvar: Tools for modeling sparse high-dimensional multivariate time series. arXiv preprint arXiv:1702.07094.

[21] Judea Pearl. 2009. *Causality*. Cambridge university press, Cambridge, England.

[22] Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning Causality for News Events Prediction. In *Proceedings of the 21st International Conference on World Wide Web* (Lyon, France) *(WWW '12)*. ACM, New York, NY, USA, 909–918. https://doi.org/10.1145/2187836.2187958

[23] Olivia Sanchez-Graillet and Massimo Poesio. 2004. Acquiring Bayesian Networks from Text. In *LREC*. ELRA, Lisbon, Portugal.

[24] Craig Silverstein, Sergey Brin, Rajeev Motwani, and Jeff Ullman. 2000. Scalable techniques for mining causal structures. *Data Mining and Knowledge Discovery* 4, 2-3 (2000), 163–192.

[25] Christopher A Sims. 1980. Macroeconomics and reality. *Econometrica: journal of the Econometric Society* 48, 1 (1980), 1–48.

[26] Peter Spirtes and Kun Zhang. 2016. Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics* 3, 3 (2016). https://doi.org/10.1186/s40535-016-0018-x

[27] Harald Steck et al. 2001. *Constraint-based structural learning in Bayesian networks using finite data sets*. Ph.D. Dissertation. Technischen Universität München.

[28] Hal R Varian. 2014. Big data: New tricks for econometrics. *The Journal of Economic Perspectives* 28, 2 (2014), 3–27.