Identifying Predictive Causal Factors from News Streams

Ananth Balashankar[1], Sunandan Chakraborty[2], Samuel Fraiberger[1,3], and Lakshminarayanan Subramanian[1]

[1]Courant Institute of Mathematical Sciences, New York University
[2]School of Informatics and Computing, Indiana University-Indianapolis
[3]World Bank, Washington DC
ananth@nyu.edu, sunchak@iu.edu, sfraiberger@worldbank.org, lakshmi@nyu.edu

Abstract

We propose a new framework to uncover the relationship between news events and real world phenomena. We present the *Predictive Causal Graph* (PCG) which allows to detect latent relationships between events mentioned in news streams. This graph is constructed by measuring how the occurrence of a word in the news influences the occurrence of another (set of) word(s) in the future. We show that PCG can be used to extract latent features from news streams, outperforming other graph-based methods in prediction error of 10 stock price time series for 12 months. We then extended PCG to be applicable for longer time windows by allowing time-varying factors, leading to stock price prediction error rates between 1.5% and 5% for about 4 years. We then manually validated PCG, finding that 67% of the causation semantic frame arguments present in the news corpus were directly connected in the PCG, the remaining being connected through a semantically relevant intermediate node.

1 Introduction

Contextual embedding models (Devlin et al., 2018) have managed to produce effective representations of words, achieving state-of-the-art performance on a range of NLP tasks. In this paper, we consider a specific task of predicting variations in stock prices based on word relationships extracted from news streams. Existing word embedding techniques are not suited to learn relationships between words appearing in different documents and contexts (Le and Mikolov, 2014). Existing work on stock price prediction using news have typically relied on extracting features from financial news (Falinouss, 2007; Hagenau et al., 2013), or sentiments expressed on Twitter (Mao et al., 2011; Rao and Srivastava, 2012; Bernardo et al., 2018), or by focusing on features present in a single document (Kalyani et al., 2016; Shynkevich et al., 2015). However, relationships between events affecting stock prices can be quite complex, and their mentions can be spread across multiple documents. For instance, market volatility is known to be triggered by recessions; this relationship may be reflected with a spike in the frequency of the word "recession" followed by a spike in the frequency of the word "volatility" a *few weeks later*. Existing methods are not well-equipped to deal with these cases.

This paper aims to uncover latent relationships between words describing events in news streams, allowing us to unveil *hidden* links between events spread across time, and integrate them into a news-based predictive model for stock prices. We propose the *Predictive Causal Graphs (PCG)*, a framework allowing us to detect latent relationships between words when such relationships are not directly observed. PCG differs from existing relationship extraction (Das et al., 2010) and representational frameworks (Mikolov et al., 2013) across two dimensions. First, PCG identifies unsupervised causal relationships based on consistent time series prediction instead of association, allowing us to uncover paths of *influence* between news items. Second, PCG finds inter-topic influence relationships outside the "context" or the confines of a single document. Construction of PCG naturally leads to *news-dependent* predictive models for numerous variables, like stock prices.

We construct PCG by identifying Granger causal pairs of words (Granger et al., 2000) and combining them to form a network of words using the Lasso Granger method (Arnold et al., 2007). A directed edge in the network therefore represents a potential influence between words. While predictive causality is not true causality (Maziarz, 2015), identification of predictive causal factors which prove to be relevant predictors over long periods of

time provides guidance for future causal inference studies. We achieve this consistency by proposing a framework for *Longitudinal Predictive Causal Factor* identification based on methods of honest estimation (Athey and Imbens, 2016). Here, we first estimate a universe of predictive causal factors on a relatively long time series and then identify time-varying predictive causal factors based on constrained estimation on multiple smaller time series. We also augment our model with an orthogonal spike correction ARIMA (Brockwell and Davis, 2002) model, allowing us to overcome the drawback of slow recovery in smaller time series.

We constructed PCG from news streams of around $700,000$ articles from Google News API and New York Times spread across over 6 years and evaluated it to extract features for stock price predictions. We obtained *two orders lower* prediction error compared to a similar semantic causal graph-based method (Kang et al., 2017). The longitudinal PCG provided insights into the variation in importance of the predictive causal factors over time, while consistently maintaining a low prediction error rate between 1.5-5% in predicting 10 stock prices. Using full text of more than 1.5 million articles of Times of India news archives for over 10 years, we performed a fine-grained qualitative analysis of PCG and validated that 67% of the semantic causation arguments found in the news text is connected by a direct edge in PCG while the rest were linked by a path of length 2. In summary, PCG provides a powerful framework for identifying predictive causal factors from news streams to accurately predict and interpret price fluctuations.

2 Related Work

Online news articles are a popular source for mining real-world events, including extraction of causal relationships. Radinsky and Horvitz (Radinsky and Horvitz, 2013) proposed a framework to find causal relationships between events to predict future events from News but caters to a small number of events. Causal relationships extracted from news using Granger causality have also been used for predicting variables, such as stock prices (Kang et al., 2017; Verma et al., 2017; Darrat et al., 2007). A similar causal relationship generation model has been proposed by Hashimoto et al. (2015) to extract causal relationships from natural language text. A similar approach can be observed in (Kozareva, 2012; Do et al., 2011), whereas CATENA system (Mirza and Tonelli, 2016) used a hybrid approach consisting of a rule-based component and a supervised classifier. PCG differs from these approaches as it explores latent inter-topic causal relationships in an unsupervised manner from the entire vocabulary of words and collocated N-grams.

Apart from using causality, there are many other methods explored to extract information from news and are used in time series based forecasting. Amodeo et al. (Amodeo et al., 2011) proposed a hybrid model consisting of time-series analysis, to predict future events using the New York Times corpus. FBLG (Cheng et al., 2014) focused on discovering temporal dependency from time series data and applied it to a Twitter dataset mentioning the Haiti earthquake. Similar work by Luo et al. (Luo et al., 2014) showed correlations between real-world events and time-series data for incident diagnosis in online services. Other similar works like, Trend Analysis Model (TAM) (Kawamae, 2011) and Temporal-LDA (TM-LDA) (Wang et al., 2012) model the temporal aspect of topics in social media streams like Twitter. Structured data extraction from news have also been used for stock price prediction using techniques of information retrieval in (Ding et al., 2014; Xie et al., 2013; Ding et al., 2015; Chang et al., 2016; Ding et al., 2016). Vaca et al. (Vaca et al., 2014) used a collective matrix factorization method to track emerging, fading and evolving topics in news streams. PCG is inspired by such time series models and leverages the Granger causality detection framework for the trend prediction task.

Deriving true causality from observational studies has been studied extensively. One of the most widely used algorithm is to control for variables which satisfy the backdoor criterion (Pearl, 2009). This however, requires a knowledge of the causal graph and the unconfoundedness assumption that there are no other unobserved confounding variables. While the unconfoundedness assumption is to some extent valid when we analyze all news streams (under the assumption that all significant events are reported), it is still hard to get away from the causal graph requirement. Propensity score based matching aims to control for most confounding variables by using an external method for estimating and controlling for the likelihood of outcomes (Olteanu et al., 2017).

More recently, (Wang and Blei, 2018) showed that with multiple causal factors, it is possible to leverage the correlation of those multiple causal factors and deconfound using a latent variable model. This setting is similar to the one we consider, and is guaranteed to be truly causal if there is no confounder which links a single cause and the outcome. This assumption is less strict than the unconfoundedness assumption and makes the case for using predictive causality in such scenarios. Another approach taken by (Athey and Imbens, 2016) estimates heterogeneous treatment effects by honest estimation where the model selection and factor weight estimation is done on two subpopulations of data by extending regression trees.

Our work is motivated by these works and applies methodologies for time series data extracted from news streams. PCG can offer the following benefits for using news for predictive analytics – (1) Detection of influence path, (2) Unsupervised feature extraction, (3) Hypothesis testing for experiment design.

3 Predictive Causal Graph

Predictive Causal Graph (PCG) addresses the discovery of *influence* between words that appear in news text. The identification of influence link between words is based on temporal co-variance, that can help answer questions of the form: "Does the appearance of word x influence the appearance of word y after δ days?". The influence of one word on another is determined based on pairwise causal relationships and is computed using the Granger causality test. Following the identification of Granger causal pairs of words, such pairs are combined together to form a network of words, where the directed edges depict potential influence between words. In the final network, an edge or a path between a word pair represents a flow of influence from the source word to the final word and this *influence* depicts an increase in the appearance of the final words when the source word was observed in news data.

Construction of PCG from the raw unstructured news data, finding pairwise causal links and eventually building the influence network involves numerous challenges. In the rest of the section, we discuss the design methodologies used to overcome these challenges and describe some properties of the PCG.

3.1 Selecting *Informative* Words:

Only a small percentage of the words appearing in news can be used for meaningful information extraction and analysis (Manning et al., 1999; Hovold, 2005). Specifically, we eliminated too frequent (at least once in more than 50% of the days) or too rare (appearing in less than 100 articles) (Manning et al., 2008). Many common English nouns, adjectives and verbs, whose contribution to semantics is minimal (Forman, 2003) were also removed from the vocabulary. However, named-entities were retained for their newsworthiness and a set of "trigger" words were retained that depict events (e.g. flood, election) using an existing "event trigger" detection algorithm (Ahn, 2006). The vocabulary set was enhanced by adding bigrams that are significantly collocated in the corpus, such as, 'fuel price' and 'prime minister' etc.

3.2 Time-series Representation of Words:

Consider a corpus D of news articles indexed by time t, such that D_t is the collection of news articles published at time t. Each article $d \in D$ is a collection of words W_d, where i^{th} word $w_{d,i} \in W_d$ is drawn from a vocabulary V of size N. The set of articles published at time t can be expressed in terms of the words appearing in the articles as $\{\alpha_1^t, \alpha_2^t, ..., \alpha_N^t\}$, where α_i^t is the sum of frequency of the word $w_i \in V$ across all articles published at time t. α_i^t corresponding to $w_i \in V$ is defined as, $\alpha_i^t = \frac{\mu_i^t}{\sum_{t=1}^{T} \mu_i^t}$ where $\mu_i^t = \sum_{d=1}^{|D_t|} TF(w_{d,i})$. α_i^t is normalized by using the frequency distribution of w_i in the entire time period. $\mathcal{T}(w_i)$ represents the time series of the word w_i, where i varies from 1 to N, the vocabulary size.

3.3 Measuring Influence between Words

Given two time-series X and Y, the Granger causality test checks whether the X is more effective in predicting Y, than using just Y and if this holds then the test concludes X "Granger-causes" Y (Granger et al., 2000). However, if both X and Y are driven by a common third process with different lags, one might still fail to reject the alternative hypothesis of Granger causality. Hence, in PCG, we explore the possibility of causal links between all word pairs and detect triangulated relations to eliminate the risk of ignoring confounding variables, otherwise not considered in the Granger causality test.

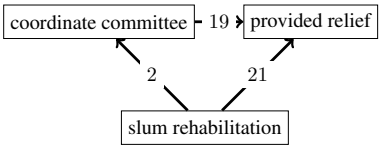


Figure 1: PCG highlighting the underlying cause

However, constructing PCG using an exhaustive set of word pairs does not scale, as even after using a reduced set of words and including the collocated phrases, the vocabulary size is around $39,000$. One solution to this problem is considering the Lasso Granger method (Arnold et al., 2007) that applies regression to the neighborhood selection problem for any word, given the fact that the best regressor for that variable with the least squared error will have non-zero coefficients only for the lagged variables in the neighborhood. The Lasso algorithm for linear regression is an incremental algorithm that embodies a method of variable selection (Tibshirani, 1994).

If we define V to be the input vocabulary from the news dataset, N is the vocabulary size, x is the list of all lagged variables (each word is multivariate with a maximum lag of 30 days per word) of the vocabulary, w is the weight vector denoting the influence of each variable, y is the predicted time series variable and λ is a sparsity constraint hyperparameter to be fine-tuned, then minimizing the regression loss below leads to weights that characterize the influential links between words in x that predicts y,

$$\mathbf{w} = argmin \frac{1}{N} \Sigma_{(\mathbf{x},y)\in V} |\mathbf{w}.\mathbf{x} - y|^2 + \lambda ||\mathbf{w}|| \quad (1)$$

To set λ, we use the method based on consistent estimation used in (Meinshausen and Bühlmann, 2006). We select the variables that have non-zero co-efficients and choose the best lag for a given variable based on the maximum absolute value of a word's co-efficient. We then, draw an edge from all these words to the predicted word with the annotations of the optimal time lag (in days) and incrementally construct the graph as illustrated in Figure 1.

3.4 Topic Influence Compression

To arrive at a sparse graphical representation of PCG, we compress the graph based on topics (50 topics in our case). Topics are learned from the original news corpus using unsupervised Latent Dirichlet Allocation (LDA)(Blei et al., 2003). Influence is generalized to topic level by calculating the weight of inter-topic influence relationships as a total number of edges between vertices of two topics. If we define θ_u and θ_v to be two topics in our topic model and $|\theta_u|$ represents the size of topic θ_u, i.e. the number of words in the topic whose topic-probability is greater than a threshold (0.001), then the strength of influence between topics θ_u and θ_v is defined as,

$$\Phi(\theta_u, \theta_v) = \frac{\text{\# Edges between words in } \theta_u \text{ and } \theta_v}{(|\theta_u| \times |\theta_v|)} \quad (2)$$

$\Phi(\theta_u, \theta_v)$ is termed as *strong* if its value is in the 99^{th} percentile of Φ for all topics. Any edge in the original PCG is removed if there are no strong topic edges between the corresponding word nodes. This filtered topic graph has only edges between topics which have high influence strength. This combination of inter-document temporal and intra-document distributional similarity is critical to obtaining temporally and semantically consistent predictive causal factors.

4 Prediction Models using PCG

In this section, we present three approaches for building prediction models using PCG namely (1) direct estimation using PCG (2) longitudinal prediction which incorporates short term temporal variations and (3) spike augumented prediction which estimates spikes over a longer time window.

4.1 Direct Prediction from PCG

One straightforward way of using PCG for prediction modeling is to use the Lasso regression equation used for identifying the predictive causal factors directly. We first adopt this approach by restricting the construction of PCG to the nodes of concern, which significantly speeds up the computation. This inherently ignores any predictive causal factor which only has an indirect link to the outcome node, as theorized by the Granger Causality framework. In this case, we split the data into a contiguous training data, and evaluate on the remaining testing data. If y represents the target stock time series variable and x represents a multivariate vector of all lagged feature variables, w represents the coefficient weight vector indexed by the feature variable $z \in x$ and time lag m, p, q in days, a represent a bias constant and ϵ_t repre-

sents an i.i.d noise variable, then we predict future values of y as follows.

$$y_t = a + \sum_{i=1}^{m} w_{y,i} y_{t-i} + \sum_{z \in x} \sum_{j=p}^{q} w_{z,j} z_{t-j} + \epsilon_t \quad (3)$$

4.2 Longitudinal Prediction via Honest Estimation

In scenarios where heterogenous causal effects are to be estimated, it is important to adjust by partitioning the time series into subpopulations which vary in the magnitude of the causal effects (Athey and Imbens, 2016). In a news stream, this amounts to constructing the word influence networks given a context specified by a time window. This naive extension however can be quite computationally expensive and can limit the speed of inference considerably. However, if the set of potential causal factors are identified over a larger time series, learning their time varying weights over a shorter training period can significantly decrease the computation required.

Hence, we do a two staged honest estimation approach similar to (Athey and Imbens, 2016). In the first stage, multiple sets (instead of trees as in (Athey and Imbens, 2016)) of predictive causal factors $F(Tr_m)$, that provide overall reduction in root mean squared error ($RMSE$), over training data Tr_m, are gathered for model selection through repeated random initialization of regression weights. For any $f \in F(Tr_m)$, we define f to be a set of predictive factors which when trained over data Tr_m to predict future time series values of the target y, achieves $RMSE(Tr_m, f) < \delta$, for a hyperparameter $\delta > 0$.

$$F(Tr_m) = \bigcup\nolimits_f \{RMSE(Tr_m, f) < \delta\} \quad (4)$$

From these sets of predictive causal factors $F(Tr_m)$, we choose the set of features $f_{LPC}(Tr_m, Tr_e)$, which gives the least expected root mean squared error on time windows w of length W uniformly sampled from the unseen training data used for estimation Tr_e through cross validation. This model selection procedure depends on the size of the smaller time windows W used to fine-tune the factors. The size of this time window is a hyperparameter to trade off long-term stability and short-term responsiveness of the predictive causal factors. We chose the time window based of 30 days for our stock price prediction due to prior knowledge that many financial indicators have a monthly cycle and hence responsiveness within that cycle is desired.

$$E(Tr_e, f) = \sqrt{\mathop{\mathbb{E}}_{w \in Tr_e, |w|=W} MSE(w, f)} \quad (5)$$

$$f_{LPC}(Tr_m, Tr_e) = min_{f \in F(Tr_m)} \Big(E(Tr_e, f) \Big) \quad (6)$$

We then evaluate the model on an unseen time series Te, where the learnt predictive causal factors and their weights are used for inference.

4.3 Spike Prediction

One drawback of using a specific time window for estimating the weights of the predictive causal factors is the lack of representative data in the window used for training. This could mean that predicting abrupt drops or spikes in the data would be hard. To overcome this limitation, we train a composite model to predict the residual error from honest estimation by training on differences in consecutive values of the time series. Let ($\Delta y = y_t - y_{t-1}, \Delta f = f_t - f_{t-1}$) denote time series of the differences of the consecutive values of the labels and the feature variables and let $[\Delta y, \Delta f]$ denote the concatenated input variables of the model. We use a multivariate ARIMA model M of order (p, d, q) (Brockwell and Davis, 2002) where p controls the number of time lags included in the model, q denotes the number of times differences are taken between consecutive values and r denotes the time window of the moving average taken to incorporate sudden spikes in values. Let the actual values of the time series of label y be y^*, the predictions of the honest estimation model be $\hat{y}$, a training sample of significantly longer time window Tr_s with $|Tr_s| >> W$, then the composite model is trained to predict the residuals, $res = y^* - \hat{y} = E(Tr_s, f)$.

$$M = ARIMA(p, d, q) \quad (7)$$

$$M.fit(E(Tr_s, f), [\Delta y, \Delta f]) \quad (8)$$

$$\hat{res} = M.predict(Tr_e) \quad (9)$$

Augmenting this predicted residual ($\hat{res}$) back to $\hat{y}_{Tr_e}$ gives us the spike-corrected estimate $\hat{y_s}$.

$$\hat{y_s} = \hat{y}_{Tr_e} + \hat{res} \quad (10)$$

5 Results

In this section, we present the results from direction prediction models from PCG, followed by improvement in stock price prediction due to longitudinal and spike prediction from news streams and

compare it to a manually tuned semantic causal graph method. We analyze the time varying factors to explain the gains achieved via honest estimation.

5.1 Data and Metrics

The news dataset[1] we used for stock price prediction contains news crawled from 2010 to 2013 using Google News APIs and New York Times data from 1989 to 2007. We construct PCG from the time series representation of its 12,804 unigrams and 25,909 bigrams over the entire news corpora of more than 23 years, as well as the 10 stock prices[2] from 2010 to 2012 for training and 2013 as test data for prediction. The prediction is done with varying step sizes (1,3,5), which indicates the time lag between the news data and the day of the predicted stock price in days. The results shown in Table 1 is the root mean squared error (RMSE) in predicted stock value calculated on a 30 day window averaged by moving it by 10 days over the period and directly comparable to our baseline (Kang et al., 2017). To evaluate the time-varying factors over a larger time window, we present average monthly cross validation RMSE % sampled over a 4 year window of 2010-13 in Table 3. Please note that the results in Table 3 are not comparable with (Kang et al., 2017) as we report a cross validation error over a longer time window.

5.2 Prediction Performance of PCG

To evaluate the causal links generated by PCG, we use it to extract features for predicting stock prices using the exact data and prediction setting used in Kang et al. (2017) as our baseline.

Baseline: Kang et al. (2017) extract relevant variables based on a semantic parser - SEMAFOR (Das et al., 2010) by filtering causation related frames from news corpora, topics and sentiments from tweets. To overcome the problem of low recall, they adopt a topic-based knowledge base expansion. This expanded dataset is then used to train a neural reasoning model which generates sequence of cause-effect statements using an attention model where the words are represented using word2vec vectors. (Kang et al., 2017)'s CGRAPH based forecasting model - C_{best} model uses the top 10 such generated cause features, given the stock name as the effect and apply a vector auto-regressive model on the combined time series of text and historical stock values.

[1]https://github.com/dykang/cgraph

[2]https://finance.yahoo.com

Table 1: 30 day windowed average of stock price prediction error using PCG

Step size	C_{best}	PCG_{uni}	PCG_{bi}	PCG_{both}
1	1.96	0.022	0.023	0.020
3	3.78	0.022	0.023	0.022
5	5.25	0.022	0.023	0.021

Comparison with the baseline: Compared to the baseline, note that our features and topics were chosen purely based on distributional semantics of the word time series. Once the features are extracted from PCG, we use the past values of stock prices and time series corresponding to the incoming word edges of PCG to predict the future values of the stock prices using the multivariate regression equation used to determine Granger Causality. As compared to their best error, PCG from unigrams, bigrams or both obtain two orders lower error and significantly outperforms C_{best}. The mean absolute error (MAE) for the same set of evaluations is within 0.003 of the RMSE, which indicates that the variance of the errors is also low. We attribute this gain to the flexibility of PCG's Lasso Granger method to produce sparse graphs as compared to CGRAPH's Vector Auto Regressive model which used a fixed number (10) of incoming edges per node already pre-filtered by a semantic parser. This imposes an artificial bound on sparsity thereby losing valuable latent information. We overcome this in PCG using a suitable penalty term (λ) in the Lasso method.

Key PCG factors for 2013: The causal links in PCG are more generic (Table 2) than the ones described in CGRAPH, supporting the hypothesis that latent word relationships do exist that go beyond the scope of a single news article. The nodes of CGRAPH are tuples extracted from a semantic parser (SEMAFOR (Das et al., 2010)) based on evidence of causality in a sentence. PCG poses no such restriction and derives topical (unfriended, FB) and inter-topical (healthcare, AMZN), sparse, latent and semantic relationships.

Inspecting the links and paths of PCG gives us qualitative insights into the context in which the word-word relationships were established. Since PCG is also capable of representing other stock time series as potential influencers in the network, we can use this to model the propagation of shocks in the market as shown in Figure 2. However,

Table 2: Stock price predictive factors for 2013 in PCG

Stock symbol	Prediction indicators
AAPL	workplace, shutter, music
AMZN	healthcare, HBO, cloud
FB	unfriended, troll, politician
GOOG	advertisers, artificial intelligence, shake-up
HPQ	China, inventions, Pittsburg
IBM	64 GB, redesign, outage
MSFT	interactive, security, Broadcom
ORCL	corporate, investing, multimedia
TSLA	prices, testers, controversy
YHOO	entertainment, leadership, investment

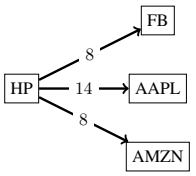


Figure 2: Inter-stock influencer links where one stock's movement indicates future movement of other stocks (time lag annotated edges)

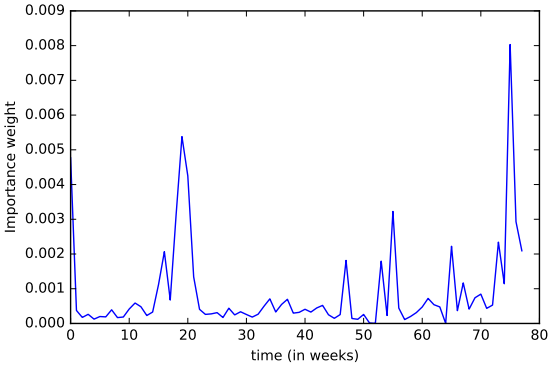
these links were not used for prediction performance to maintain parity with our baseline.

5.3 Time Varying Causal Analysis

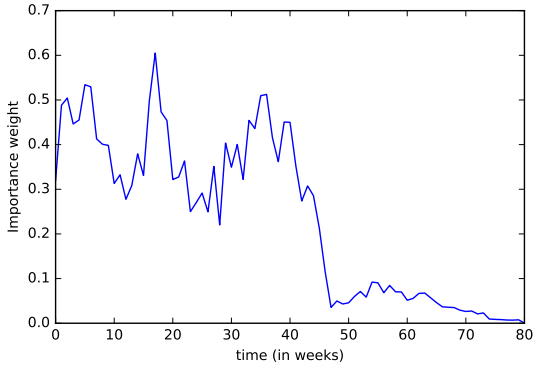
We quantitatively evaluate the time varying variant of PCG by using it to extract features for stock price prediction for a longer time window.

We present average root mean squared errors in Table 3 for different values of time windows of size W (50,100). For model selection, we used 50% of the time series and then used multiple time series of length W, disjoint from the ones used for time-varying factor selection and took average of the test error on the next 30 data points using the weights learnt. We repeat this using K-fold cross validation (K=10) for choosing the model selection data and present the average errors. The variation in importance weights of predictive causal factors for stock prices ("podcast", AAPL) and ("unfollowers", GOOG) is shown in Figure 3 which illustrates several peaks (for weeks) when the factor in the news was particularly predictive of the company's stock price and not significant during other weeks.

The time series and error graph shown for multiple stocks shows that the RMSE errors range between 1.5% 5% for all the test time series as shown in Figure 4. However, sudden spikes tend to display higher error rates due to the lack of training data which contain spikes. This issue is mitigated when the time window for training is increased. Increasing the window more than 100 did not improve the RMSE and came at the cost of training time. But, incorporating the spike residual PCG model, which predicts the leftover price value from the first model, provides significant improvements over the model without spike correction as seen in the last column on Table 3. Thus, we are able to achieve significant gains with an unsupervised approach without any domain specific feature engineering by estimating using an ARIMA model (p, d, q) = (30, 1, 10).



(a) "podcast" for predicting AAPL stock



(b) "unfollowers" for predicting GOOG stock

Figure 3: Temporal variation in importance weight of predictive causal factors

6 Interpretation of Predictive Causal Factors

In order to qualitatively validate that the latent inter-topic edges learnt from the news stream is also humanly interpretable, we constructed PCG from the online archives of Times of India (TOI) [3], the most circulated Indian English newspaper. We used this dataset as, unlike the previous dataset

[3] https://timesofindia.indiatimes.com/archive.cms

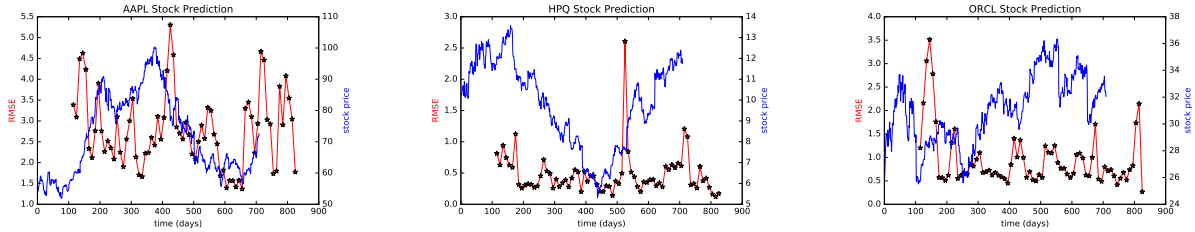


Figure 4: RMSE of stocks for longitudinal causal factor prediction without spike correction. Spikes in RMSE can be seen along with spikes in stock prices like HPQ.

Table 3: Variation in stock price prediction error (RMSE) % with window size and spike correction

Stock	W=50	W=100	W=100 + spike
AABA	2.87	2.07	1.61
AAPL	2.95	2.84	2.28
AMZN	3.03	2.99	2.41
GOOG	2.67	2.36	1.91
HPQ	6.77	3.34	2.44
IBM	2.19	2.07	1.65
MSFT	3.03	9.45	4.80
ORCL	2.94	2.21	1.65
TSLA	5.56	5.52	4.32

which provided just the time series of words, we also have the raw text of the articles, which allowed us to perform manual causal signature verification. This dataset contains all the articles published in their online edition between January 1, 2006 and December 31, 2015 containing 1,538,932 articles.

6.1 Inter-topic edges of PCG

The influence network we constructed from the TOI dataset has 18,541 edges and 7,190 uni-grams and bi-gram vertices. We were interested in the inter-topic non-associative relationships that PCG is expected to capture. We observe that a few topics (5) influence or are influenced by a large number of topics. Some of these highly influential topics are composed of words describing "Agriculture", "Politics", "Crime", etc. The ability of PCG to learn these edges between topical word clusters purely based on temporal predictive causal prediction further validates its use for design of extensive causal inference experiments.

6.2 Causal evidence in PCG

To validate the causal links in PCG, we extracted 56 causation semantic frame (Baker et al., 1998) arguments which depict direct causal relationships in the news corpus. We narrowed down the search to words surrounding verbs which depict the notion of causality like "caused", "effect", "due to" and manually verified that these word pairs were indeed causal. We then searched the shortest path in PCG between these word pairs. For example, one of the news article mentioned that "Gujarat government has set aside a suggestion for *price hike* in electricity due to the Mundra Ultra Mega *Power Project*." and these corresponding causation arguments were captured by a direct link in PCG as shown in Table 4. 67% of the word pairs which were manually identified to be causal in the news text through causal indicator words such as "caused", were linked in PCG through direct edges, while the rest were linked through an intermediate relevant node. As seen in Table 4, the bi-gram involving the words and the intermediate words in the path provide the relevant context under which the causality is established. The time lags in the path show that the influence between events are at different time lags. We also qualitatively verified that two unrelated words are either not connected or have a path length greater than 2, which makes the relationship weak. The ability of PCG to validate such humanly understood causal pairs with temporal predictive causality can be used for better hypothesis testing.

Table 4: Comparison with manually identified influence from news articles

Pairs in news	Relevant paths in PCG
price, project	price-hike –(19)– power-project
land, budget	allot-land –(22)– railway-budget
price, land	price-hike –(12)– land
strike, law	terror-strike –(25)– law ministry
land, bill	land-reform –(25)– bill-pass
election, strike	election –(21)– Kerala government –(10)– strike
election, strike	election –(18)– Mumbai University –(14)– strike
election, strike	election –(20)– Shiromani Akali –(13)– strike

7 Conclusion

We presented PCG, a framework for building predictive causal graphs which capture hidden rela-

tionships between words in text streams. PCG overcomes the limitations of contextual representation approaches and provides the framework to capture inter-document word and topical relationships spread across time to solve complex mining tasks from text streams. We demonstrated the power of these graphs in providing insights to answer causal hypotheses and extracting features to provide consistent, interpretable and accurate stock price prediction from news streams through honest estimation on unseen time series data.

References

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*. Association for Computational Linguistics, Sydney, Australia, 1–8. https://www.aclweb.org/anthology/W06-0901

Giuseppe Amodeo, Roi Blanco, and Ulf Brefeld. 2011. Hybrid Models for Future Event Prediction *(CIKM '11)*. 1981–1984.

Andrew Arnold, Yan Liu, and Naoki Abe. 2007. Temporal Causal Modeling with Graphical Granger Methods. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '07)*. ACM, New York, NY, USA, 66–75. https://doi.org/10.1145/1281192.1281203

Susan Athey and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects: Table 1. *Proceedings of the National Academy of Sciences* 113, 7353–7360. https://doi.org/10.1073/pnas.1510489113

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1 (ACL '98/COLING '98)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 86–90. https://doi.org/10.3115/980845.980860

Ivo Bernardo, Roberto Henriques, and Victor Lobo. 2018. Social Market: Stock Market and Twitter Correlation. In *Intelligent Decision Technologies 2017*, Ireneusz Czarnowski, Robert J. Howlett, and Lakhmi C. Jain (Eds.). Springer International Publishing, Cham, 341–356.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (March 2003), 993–1022.

Peter J. Brockwell and Richard A. Davis. 2002. *Introduction to Time Series and Forecasting* (2nd ed.). Springer.

Ching-Yun Chang, Yue Zhang, Zhiyang Teng, Zahn Bozanic, and Bin Ke. 2016. Measuring the Information Content of Financial News. In *COLING*. ACL, 3216–3225.

Dehua Cheng, Mohammad Taha Bahadori, and Yan Liu. 2014. FBLG: A Simple and Effective Approach for Temporal Dependence Discovery from Time Series Data *(KDD '14)*. 382–391. https://doi.org/10.1145/2623330.2623709

Ali F. Darrat, Maosen Zhong, and Louis T.W. Cheng. 2007. Intraday volume and volatility relations with and without public news. *Journal of Banking and Finance* 31, 9 (2007), 2711 – 2729. https://doi.org/10.1016/j.jbankfin.2006.11.019

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A. Smith. 2010. Probabilistic Frame-Semantic Parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Los Angeles, California, 948–956. https://www.aclweb.org/anthology/N10-1138

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2014. Using Structured Events to Predict Stock Price Movement: An Empirical Investigation.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep Learning for Event-Driven Stock Prediction. In *IJCAI*. AAAI Press, 2327–2333.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2016. Knowledge-Driven Event Embedding for Stock Prediction. In *COLING*. ACL, 2133–2142.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally Supervised Event Causality Identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 294–303. http://dl.acm.org/citation.cfm?id=2145432.2145466

Pegah Falinouss. 2007. Stock Trend Prediction using News Events. *Masters thesis* (2007).

George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research* 3, Mar (2003), 1289–1305.

Clive WJ Granger, Bwo-Nung Huangb, and Chin-Wei Yang. 2000. A bivariate causality between stock prices and exchange rates: evidence from recent Asianflu. *The Quarterly Review of Economics and Finance* 40, 3 (2000), 337–354.

Michael Hagenau, Michael Liebmann, and Dirk Neumann. 2013. Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems* 55, 3 (2013), 685 – 697. https://doi.org/10.1016/j.dss.2013.02.006

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, and Jong-Hoon Oh. 2015. Generating Event Causality Hypotheses Through Semantic Relations. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15)*. AAAI Press, 2396–2403. http://dl.acm.org/citation.cfm?id=2886521.2886654

Johan Hovold. 2005. Naive Bayes Spam Filtering Using Word-Position-Based Attributes.. In *CEAS*. 41–48.

Joshi Kalyani, H. N. Bharathi, and Rao Jyothi. 2016. Stock trend prediction using news sentiment analysis. *CoRR* abs/1607.01958 (2016). arXiv:1607.01958 http://arxiv.org/abs/1607.01958

Dongyeop Kang, Varun Gangal, Ang Lu, Zheng Chen, and Eduard Hovy. 2017. Detecting and Explaining Causes From Text For a Time Series Event. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2758–2767. https://www.aclweb.org/anthology/D17-1292

Noriaki Kawamae. 2011. Trend analysis model: trend consists of temporal words, topics, and timestamps *(WSDM '11)*. 317–326.

Zornitsa Kozareva. 2012. Cause-effect Relation Learning. In *Workshop Proceedings of TextGraphs-7 on Graph-based Methods for Natural Language Processing (TextGraphs-7 '12)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 39–43. http://dl.acm.org/citation.cfm?id=2392954.2392961

Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. *CoRR* abs/1405.4053 (2014). arXiv:1405.4053 http://arxiv.org/abs/1405.4053

Chen Luo, Jian-Guang Lou, Qingwei Lin, Qiang Fu, Rui Ding, Dongmei Zhang, and Zhe Wang. 2014. Correlating Events with Time Series for Incident Diagnosis *(KDD '14)*. 1583–1592. https://doi.org/10.1145/2623330.2623374

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. 2008. *Scoring, term weighting, and the vector space model*. Cambridge University Press, 100123. https://doi.org/10.1017/CBO9780511809071.007

Christopher D Manning, Hinrich Schütze, et al. 1999. *Foundations of statistical natural language processing*. Vol. 999. MIT Press.

Huina Mao, Scott Counts, and Johan Bollen. 2011. Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data. *Arxiv preprint* (2011).

Mariusz Maziarz. 2015. A review of the Granger-causality fallacy. *The Journal of Philosophical Economics* 8, 2 (2015), 6. https://EconPapers.repec.org/RePEc:bus:jphile:v:8:y:2015:i:2:n:6

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13)*. Curran Associates Inc., USA, 3111–3119. http://dl.acm.org/citation.cfm?id=2999792.2999959

Paramita Mirza and Sara Tonelli. 2016. CATENA: CAusal and TEmporal relation extraction from NAtural language texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, 64–75. http://www.aclweb.org/anthology/C16-1007

Alexandra Olteanu, Onur Varol, and Emre Kiciman. 2017. Distilling the Outcomes of Personal Experiences: A Propensity-scored Analysis of Social Media. In *Proceedings of The 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing* (computer-supported cooperative work and social computing ed.). Association for Computing Machinery, Inc.

Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics Surveys* (2009).

Kira Radinsky and Eric Horvitz. 2013. Mining the web to predict future events *(WSDM '13)*. ACM, 255–264.

Tushar Rao and Saket Srivastava. 2012. Analyzing Stock Market Movements Using Twitter Sentiment Analysis. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012) (ASONAM '12)*. IEEE Computer Society, Washington, DC, USA, 119–123. https://doi.org/10.1109/ASONAM.2012.30

Y. Shynkevich, T. M. McGinnity, S. Coleman, and A. Belatreche. 2015. Stock price prediction based on stock-specific and sub-industry-specific news articles. In *2015 International Joint Conference on Neural Networks (IJCNN)*. 1–8. https://doi.org/10.1109/IJCNN.2015.7280517

Robert Tibshirani. 1994. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B* 58 (1994), 267–288.

Carmen K Vaca, Amin Mantrach, Alejandro Jaimes, and Marco Saerens. 2014. A time-based collective factorization for topic discovery and monitoring in news *(WWW '14)*. 527–538.

Ishan Verma, Lipika Dey, and Hardik Meisheri. 2017. Detecting, Quantifying and Accessing Impact of News Events on Indian Stock Indices. In *Proceedings of the International Conference on Web Intelligence (WI '17)*. ACM, New York, NY, USA, 550–557. https://doi.org/10.1145/3106426.3106482

Yu Wang, Eugene Agichtein, and Michele Benzi. 2012. Tm-lda: efficient online modeling of latent topic transitions in social media *(KDD '12)*. ACM, 123–131.

Yixin Wang and David M. Blei. 2018. The Blessings of Multiple Causes. *CoRR* abs/1805.06826. http://dblp.uni-trier.de/db/journals/corr/corr1805.html#abs-1805-06826

Boyi Xie, Rebecca J. Passonneau, Leon Wu, and Germán Creamer. 2013. Semantic Frames to Predict Stock Price Movement. In *ACL (1)*. The Association for Computer Linguistics, 873–883.