Macroeconomic forecasting with statistically validated knowledge graphs

Sonja Tilly[a,*], Giacomo Livan[a,b]

[a]*UCL, Computer Science Dep, 66 - 72 Gower St, Bloomsbury, WC1E 6EA London, UK*
[b]*Systemic Risk Centre, London School of Economics and Political Science, London, WC2A 2AE, UK*

ARTICLE INFO

ABSTRACT

This study leverages narrative from global newspapers to construct theme-based knowledge graphs about world events, demonstrating that features extracted from such graphs improve forecasts of industrial production in three large economies compared to a number of benchmarks. Our analysis relies on a filtering methodology that extracts "backbones" of statistically significant edges from large graph data sets. We find that changes in the eigenvector centrality of nodes in such backbones capture shifts in relative importance between different themes significantly better than graph similarity measures. We supplement our results with an interpretability analysis, showing that the theme categories "disease" and "economic" have the strongest predictive power during the time period that we consider. Our work serves as a blueprint for the construction of parsimonious – yet informative – theme-based knowledge graphs to monitor in real time the evolution of relevant phenomena in socio-economic systems.

1. Introduction

People are natural storytellers who rely on narrative to make decisions, particularly when faced with uncertainty. There is a large body of literature on narrative, with most works spanning disciplines such as psychology, cognitive sciences or political sciences (Brosch et al., 2013; Clore & Palmer, 2009; Bruner, 1990; King et al., 2017).

Keynes describes the state of mind that results in human actions as "animal spirits", which ultimately are reflected in economic indicators such as consumer confidence (Keynes, 2018). Shiller finds that viral narratives spread by newspapers play a causal role in economic activity (Shiller, 2017). Along the same lines, Conviction Narrative Theory (CNT) demonstrates that changes in narrative are a precursor to changes in economic growth (Tuckett et al., 2014). A recent study on CNT suggests that economic agents reassure themselves by building narratives supporting their expectations of the result of their actions (Nyman et al., 2018).

In recent years, the evolution of big data and natural language processing has enabled the quantification of news narrative and its potential to change the course of social systems. This has allowed to push econometric models such as vector autoregressions (Stock & Watson, 2001), or structural frameworks such as dynamic stochastic general equilibrium models (Christiano et al., 2005; Smets & Wouters, 2007) beyond conventional variables. In this respect, research increasingly explores features derived from narrative and their capacity of improving economic forecasts (Buono et al., 2018; Elshendy et al., 2018; Yang et al., 2020).

One possible way to capture and quantify narrative is via graphs that map the interactions between concepts or events pertaining to the object of study. A wide range of applications demonstrates that graphs are indeed very successful at capturing complex relationships (Emmert-Streib et al., 2018).

This study leverages narrative from global newspapers to construct theme-based knowledge graphs, demonstrating that features derived from such graphs improve forecasts of industrial production (IP) in three large economies. The findings are supported by an interpretability analysis, showing that disease and economy-related themes have the strongest predictive power.

2. Literature review

This section covers a selection of existing literature on graphs and their applications to the analysis of economic systems. A range of methods has been proposed to extract and interpret the information contained in large graphs, and this section addresses those techniques most pertinent to weighted undirected graphs, which are the object of this study.

Over recent years, there has been increasing interest in examining economic topics in terms of graphs. The study of economic graphs is an interdisciplinary field, spanning areas such as economics, the social sciences, computer science, statistics and business and management (Emmert-Streib et al., 2018). Existing studies cover a wide range of economic graphs, each one with their own meaning. For instance, a study on graph centrality and funding rates finds that interbank spreads are significantly affected by measures of centrality, with the effects of graph centrality increasing during the global financial crisis in 2008 (Temizsoy et al., 2016). Constantin *et al.* build a graph of European banks and estimate how negative shocks for one bank's returns depend on the impact of negative shocks on other banks' returns, effectively building an early warning model for bank dis-

---

Abbreviations. GDELT: Global Database of Events, Language and Tone; GKG: Global Knowledge Graph; CNT: Conviction Narrative Theory; Bi-LSTM: bi-directional long short term memory neural network; RNN: recurrent neural network; IP: industrial production; PLS: partial least squares

*Corresponding author

✉ sonja.tilly.19@ucl.ac.uk (S. Tilly); g.livan@ucl.ac.uk (G. Livan)

tress using network effects (Constantin et al., 2018). A paper by Carvalho *et al*. examines firm-level disturbance after the Japanese tsunami in 2011 and finds that the propagation of the shock over input-output linkages can account for a 1.2 percentage point decline in Japan's gross output in the year following the earthquake (Carvalho et al., 2016). Adamic *et al* use established graph analysis tools to describe the time-series dimensions of information and liquidity flows in the E-mini S&P stock index futures market, showing robust contemporaneous correlations between graph features and financial variables, with the former significantly leading (Granger-causing) intertrade duration, trading volume and other liquidity metrics (Adamic et al., 2017). Piccardi and Tajoli show that high centralization observed for complex products determines the hierarchy of the global trade graph as they form an important share of total trade. This implies an uneven distribution of trade links between countries, making them more vulnerable to shocks. The authors conclude that the present structure of the global trade graph is exposed to specific shocks, and quantify the impact of shock propagation from the central nodes (Piccardi & Tajoli, 2018). A study by Bonaccorsi *et al*. measures country centrality in multilayer graphs, demonstrates that these centralities are consistent with a North-South divide and positively correlated with economic variables such as GDP per capita (Bonaccorsi et al., 2019). Yang *et al*. extract entities from textual data and link them as knowledge graphs to macro variables. The authors show that incorporating the features extracted from knowledge graphs in a macroeconomic forecasting framework significantly improves predictions of macroeconomic variables such as inflation, net export growth or housing prices (Yang et al., 2020). Guo and Vargo use themes and location data from the Global Database of Events, Language and Tone (GDELT) to show that population, trade, cultural proximity, and geographic closeness drive international news attention (Guo & Vargo, 2020). Stern *et al*. examine intermedia agenda-setting by proposing a method to infer graphs of influence between different news sources on a given subject (Stern et al., 2020). The authors find that influence very much depends on the topic, with news outfits being agenda-setters (represented by central nodes) for some topics and followers (represented by peripheral nodes) for others. Campi *et al*. explore the determinants of specialisation in agricultural production, describing it as a time-sequence of bipartite graphs, linking countries to their produced agricultural products (Campi et al., 2020). The study concludes that agricultural production is a dense graph of well-defined and stable communities of countries and products that are characterised by environmental conditions as well as economic, socio-political and technological factors. A study by Bellomarini *et al*. analyses the impact of the COVID-19 outbreak on the graph of business relationships between Italian companies and identifies transactions that could lead to the takeover of strategic companies (Bellomarini et al., 2020). Gomez *et al*. apply graph analysis to understand business cycle synchronization in the European Union over the last 20 years, observing that co-movements and country inter-

actions have increased notably since the Eurozone crisis in 2011/12, while they remained relatively stable before (Matesanz Gomez et al., 2017).

Graph analysis is an effective means to represent and analyze complex systems. However, real-world graph data sets can be very large, both in terms of number of nodes and connections between them, making it difficult to identify those elements that are most crucial to the properties of such systems (often referred to as a graph's "backbone"). The literature proposes a range of approaches for extracting graph backbones. In this respect, Ghalmane *et al*. differentiate between "coarse-grained" and filter-based approaches to graph dimensionality reduction (Ghalmane et al., 2020). "Coarse-grained" methods are based on the concept of grouping graph nodes according to some criterion and keeping those with the properties of interest, while filter-based approaches define properties for nodes and edges and discard or preserve them based on the statistical significance of these attributes against a null hypothesis for the graph structure. Within the latter category, the filter proposed by Tumminnello *et al*. identifies edges that are statistically significant with respect to a null hypothesis of random interactions in a weighted graph, described by the hypergeometric distribution (Tumminello et al., 2011). The "disparity filter" proposed by Serrano *et al* works in a similar fashion, and relies on a null hypothesis of nodes distributing their activity uniformly across their neighbours in the graph (Serrano et al., 2009). Marcaccioli and Livan recently proposed a filter based on the Pólya urn model (which includes the disparity filter as a special case), which can be tuned to a given graph's specific topological properties, and demonstrate its effectiveness at preserving statistically significant links in real-world graphs such as the US airport network or the global input-output trade network (Marcaccioli & Livan, 2019).

2.1. Hypotheses formulation

In this study, we attempt to forecast industrial production (IP) leveraging data from the Global Database of Events, Language and Tone (GDELT). Specifically, we address two main research questions, i.e., (1) whether socio-economic dynamics can be captured by knowledge graphs based on themes from GDELT, and (2) whether features derived from such graphs are predictive of changes in economic activity. Accordingly, we formulate the two following hypotheses:

- H_1: Changes in the structural properties of knowledge graphs based on themes from GDELT are reflective of socio-economic changes.

- H_2: Features derived from GDELT knowledge graphs add value to forecasts of economic activity.

By tackling the above questions, we advance the literature on economic graphs in at least three ways. First, we demonstrate how themes from newspaper articles can be incorporated into macroeconomic forecasting models to improve predictions of economic activity. This use of news narrative is – to the best of our knowledge – new. Second, we contribute to the literature on economic graphs by proposing

a data-driven methodology to operationalise concepts such as information extraction, feature generation and macroeconomic forecasting. We complement our results with an interpretability study showing that disease and economy-related themes have the strongest predictive power. Third, we contribute to research on tracking change in social systems through news narrative by examining the evolution of graph statistics over time.

3. Data and methods

This section introduces GDELT as data source and outlines the filtering methodology that is used to isolate relevant signals. The GDELT Project is a research collaboration of Google Ideas, Google Cloud, Google and Google News, the Yahoo! Fellowship at Georgetown University, BBC Monitoring, the National Academies Keck Futures Program, Reed Elsevier's LexisNexis Group, JSTOR, DTIC and the Internet Archive. The project monitors world newspapers from a variety of perspectives, identifying and extracting items such as themes, emotions, locations, organisations and events. GDELT version two incorporates real-time translation from 65 languages and is updated every 15 minutes (*GDELT Project*, 2015). It is a public data set available on the Google Cloud Platform.

The GDELT Global Knowledge Graph (GKG) is a software suite that analyses global newspaper articles in real-time to extract entities such as persons, organizations, locations, dates, themes and emotions (Leetaru et al., 2014). The extraction of location data is done through a process called full text geocoding, developed by Leetaru (Leetaru, 2012). This process applies algorithms to parse through a news item and to identify textual mentions of locations using databases of places. Applying the same principle, themes are extracted from news articles using extensive lists of themes. GDELT contains c 13,000 themes from c 47,000 sources, from over a billion news articles scanned since 2015.

For each scanned news article, the themes field contains all themes the GDELT algorithm identifies, represented as a string of labels. GDELT themes are very nuanced and can be linked to distinct categories such as "economic", "disease" or "human rights" (see appendix 8.2 for a list of all theme categories). Theme categories describe events or conditions, except for four purely descriptive theme groups ("actor", "ethnicity", "language" and "animal"), which are removed. Table 1 illustrates the number of themes for the three countries we will consider in our analysis, both in the original data (no. of themes) and the number of themes after removing the descriptive ones (reduced no. of themes).

Country	no. of themes	reduced no. of themes
US	6880	3501
Germany	5313	2925
Japan	3330	1994

Table 1: Number of of GDELT themes

3.1. Predicted variables

This study models industrial production (IP) for the US, Germany and Japan. IP is a measure of economic activity, published on a monthly basis. It represents the output of industrial establishments, covers a broad range of sectors and tracks the monthly change in the volume of production output. The countries are selected for representing large, industrialised economies with diversified trade connections in three parts of the world – America, Europe and Asia.

3.2. Filtering methodology

We apply the filtering methodology introduced by Tilly *et al.* (Tilly et al., 2021) to extract observations from GDELT's GKG that are pertinent to economic growth. The methodology consists of three steps – first, a thematic keyword filter, second, a fine-grained filter using a neural network and third, data aggregation.

As a first step, a high level thematic filter based on the keyword "economic growth" is applied to GDELT themes to select relevant articles. On inspection of the filtered data by examining 100 randomly chosen original news articles, it becomes clear that this simple keyword filter retains too many news items that are not relevant to "economic growth".

Hence, a further, more precise filter is required as a second step. We use the same bidirectional long short term memory (Bi-LSTM) architecture as proposed in the original filtering methodology. This algorithm is chosen as it exhibits the best performance in terms of precision, recall and F_1 score among a range of algorithms explored. For this step, the raw GDELT data is preprocessed. Each string of theme labels is split into lower case tokens. The tokens for every item are then label-encoded so that the themes are given numbers between zero and $N-1$. For out-of-vocabulary words, an "unknown" token is assigned. The length for each token sequence is standardized to address the variable length of these sequences by setting a maximum length of 5,000 tokens and padding. Then, 1,000 news items are manually classified into relevant and non-relevant to "economic growth" looking up the original text using the url contained in the DocumentIdentifier field (encoded as one or zero, respectively). The encoded themes represent the predictor, and the classification into relevant/non-relevant represent the predicted data, respectively.

Model performance is assessed using k-fold cross-validation as it provides a robust estimate of the performance of a model on unseen data. The training data set is divided into ten subsets. Models are trained on all subsets except one which is held out, and performance is evaluated on the held out validation data set. The process is repeated until all subsets have been used as the held out validation set.

As performance metrics, precision (the number of true positives divided by the number of true positives and false positives), recall (the number of true positives divided by the number of true positives and the number of false negatives) and the F_1 score are used:

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} .$$

Table 2 shows the performance of the Bi-LSTM classifier.

Classifier	Precision	Recall	F1
Bi-LSTM	0.8853	0.9375	0.9101

Table 2: Bi-LSTM performance

A long short term memory (LSTM) is a type of recurrent neural network (RNN) structure that can span longer distances without the loss of short term capacity (Hochreiter & Schmidhuber, 1997). The algorithm enforces constant error flow during backpropagation through internal states of special units and addresses the vanishing gradient problem. This issue originates from the repeated use of the recurrent weight in RNNs, and prevents these models from learning long term dependencies.

Like a RNN, a LSTM passes on information as it propagates forward. However, the operations within the LSTM's cells differ compared to those of a RNN. A LSTM incorporates a cell state that can be considered the "memory" as well as a set of gates. These gates control when information enters the memory, when it is output, and when it is forgotten. Through this structure, the LSTM is able to learn what information to keep or to forget during training, and keep relevant information in memory for an extended period. While a unidirectional LSTM preserves information from one direction as it only runs forward, a Bi-LSTM runs the inputs forwards and backwards simultaneously. Combining two hidden states allows a Bi-LSTM to retain information from both directions at any time (Schuster & Paliwal, 1997). Research shows that the Bi-LSTM architecture outperforms unidirectional algorithms in tasks where context matters (Graves & Schmidhuber, 2005).

Using the filtered data from step two, for each calendar month from March 2015 to December 2020, theme-based undirected weighted graphs are constructed, applying country filters for the US, Germany and Japan, respectively. In these graphs, the themes represent nodes, the co-occurrences of themes represent edges and the count of the co-occurrences represent weights. For comparison purposes, monthly graphs based on unfiltered GDELT data are also generated.

3.3. Statistical validation of graph elements

This section outlines the statistical validation procedure we apply to extract the backbone of each monthly graph.

The monthly theme-based graphs are connected and contain c 4,000 nodes and c 1 million edges. Given their large scale and the potentially large amount of noise contained in them, these graphs require a filtering mechanism in order to extract statistically relevant sets of edges, i.e., the so-called "backbone".

To this end, we applied the disparity filter (see Section 2), the most widely adopted backbone extraction method. The method starts by normalizing the weights of each edge (i, j) with respect to the total strength of node i (i.e., the total weight on edges starting from node i), so that they sum up to one. The method's null hypothesis states that the normalised weights belonging to edges of a node i are random and distributed according to the uniform distribution over $[0, 1]$. For each edge, the probability α_{ij} of the edge occurring under the null hypothesis, i.e., a p-value, is calculated. For edges whose p-value is smaller than a set threshold α, the null hypothesis is rejected and they are retained in the backbone. In this study, the threshold α is set to 0.05. Since multiple tests are conducted, the resulting p-values are adapted according to the Benjamini-Hochberg (BH) procedure to account for multiple hypothesis testing (Benjamini & Yekutieli, 2005). It should also be noted that, due to the fact that the normalisation of edge weights is done with respect to one of the two nodes they are connected to, each edge needs to be tested twice. The links retained in the backbone are those for which the null hypothesis can be rejected at least once.

Applying the disparity filter to the theme-based graphs reduces the number of nodes and edges in each monthly graph by c 50% and c 90%, respectively, while still returning connected graphs.

3.4. Graph features

For each monthly graph, the eigenvector centrality is calculated for all nodes. This measure is appropriate as the graphs are connected and reflects the broader influence of a node on the filtered graph (Constantin et al., 2018). Then, a $T \times K$ matrix is constructed, where T represents the number of observations (i.e. calendar months) and K stands for the number of nodes represented by themes.

The portrait divergence for each graph compared to the previous month's graph is calculated. This measure is a method for comparing graphs based on the graph's portrait, which encodes the distribution of the shortest-path lengths in a graph. The portrait divergence is a comprehensive measure of how the topological features of two graphs differ (Tantardini et al., 2019).

3.5. Explanatory variables

When forecasting IP, the monthly percentage changes in the baltic dry index and the crude oil price, respectively, are incorporated to control for macroeconomic effects. The baltic dry index is considered a leading indicator for economic growth, representing global trade volume (Bildirici et al., 2015). Van Eyden *et al.* find that there is a significant link between changes in oil price and economic activity in OECD countries (Van Eyden et al., 2019).

3.6. Data preprocessing

In this section, we summarize the data preparation methods.

The values for IP are used as predicted variable for each of the three aforementioned economies, with index values reflecting the monthly percentage change. The augmented

Dickey Fuller unit root test is applied to 20 years of monthly data for IP as well as the explanatory variables and stationarity is not rejected at 5% significance for any of them.

Any missing values in the $T \times K$ eigenvector centralities matrix are filled by zero, which is the lowest possible centrality score. The augmented Dickey Fuller unit root test is applied to eigenvector centralities, portrait divergences and stationary is not rejected at 5% for any of the variables. The eigenvector centralities are standardized by removing the mean and scaling to unit variance.

4. Analysis

In this section, we set out the analysis conducted to establish if the graph features derived from GDELT data have predictive power.

4.1. Granger causality analysis

The Granger causality between the eigenvector centralities and the predicted variable is assessed to establish if there are statistical relationships between those variables (Granger, 1969). Granger causality is testing for precedence rather than true causality, hence it may be found even in the absence of an actual causal connection (Leamer, 1985).

For this test, the null hypothesis stipulates that lagged graph features are not Granger-causing IP at a significance level of 5%, while the alternate hypothesis states that lagged graph features are Granger-causing IP at the same significance level.

The Granger causality for lags up to a maximum of three months is examined. The *p*-values are adjusted using the Benjamini-Hochberg (BH) procedure to control for multiple hypothesis testing (Benjamini & Yekutieli, 2005).

4.2. Forecasting

As further step in our analysis, we predict IP for the US, Germany and Japan adopting a factor-augmented autoregressive framework as proposed by Girardi *et al.* (Girardi et al., 2016).

As a first step, we predict IP using an autoregressive model including the predicted variable and the explanatory macroeconomic variables. This framework allows modeling a $T \times K$ multivariate time series Y, where T denotes the number of observations and K the number of variables. The framework is defined as

$$Y_t = v + A_1 Y_{t-1} + \cdots + A_p Y_{t-p} + u_t \qquad (1)$$

where A_i is a $K \times K$ coefficient matrix, v is a constant and u_t is white noise.

As it is not possible to incorporate the large number of eigenvector centralities into the autoregressive framework described in Eq. (1), we extract factors from this broad set of features for inclusion into the model. Therefore, as a second step, we apply Partial Least Squares (PLS) to reduce dimensionality and to extract relevant information from the eigenvector centralities. This technique is suitable for data sets whose number of features is considerably larger than

the number of observations, and collinearity of features exists (Cubadda & Guardabascio, 2012). PLS incorporates information from predicted variable and predictors when computing scores and loadings, which are chosen to maximise the covariance between predicted variable and predictors (De Jong, 1993).

PLS is implemented on the residuals derived from the autoregressive model including the predicted variable and the explanatory macroeconomic variables only. The residuals include the part of the predicted variable that is not explained and therefore, applying PLS to the eigenvector centralities based on GDELT themes provides additional information to the predictors. The orthogonal relationship between the predicted variable and the residuals maintains the orthogonality between the factors extracted by PLS and the autoregressive components. For each country variable, cross-validation analysis shows that the residual sum of squares is increasing in a model with more than five factors, indicating that five PLS factors are appropriate (Tobias, 1995). The first five PLS components account for around 70% of the variation in the respective predicted variables.
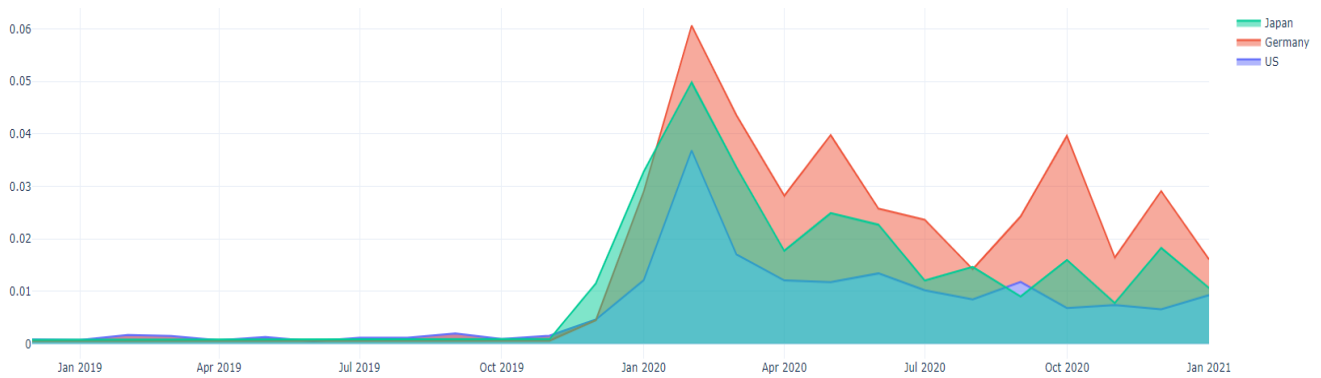
In a third step, for each country, the predicted variable, the explanatory variables and the five PLS components derived from the eigenvector centralities are employed as input into the autoregressive framework described in Eq. (1) to form a factor augmented autoregressive model (Colladon et al., 2019).

For each model, the the Akaike (AIC) and the Bayesian (BIC) information critera are applied to identify the optimal lag length. Both metrics follow the concept that the inclusion of a further term should improve the model although the model should also be penalised for adding to the number of parameters to be computed. Once the improvement in goodness-of-fit exceeds the penalty term, the statistic related to the information criterion decreases. Therefore, we select the lag that minimises the information criterion (Brooks & Tsolacos, 2010).

We construct three benchmarks for performance comparison – first, an autoregressive framework including the predicted variable and explanatory macroeconomic variables, second, an autoregressive framework incorporating the predicted variable, explanatory macroeconomic variables and the respective country's portrait divergence scores, and third, an autoregressive framework incorporating the predicted variable, explanatory macroeconomic variables and five PLS components derived from eigenvector centralities based on unfiltered GDELT data.

Performance is assessed using walk-forward cross-validation and the root mean squared error (RMSE). The data set is split into three folds. In the kth split, this cross-validation technique returns the first k folds as training set and the $(k + 1)^{\text{th}}$ fold as test set, appropriate for time series data.

The modified Diebold Mariano test proposed by Harvey, Leybourne and Newbold (Harvey et al., 1997) is used to gauge whether model forecasts are significantly different from each other.

Figure 1: Influence of COVID-19 related themes within monthly graphs.

5. Research findings

This section presents the findings from the analysis set out in the previous section.

5.1. Evolution of graph features

Theme-based knowledge graphs can be employed to identify change in specific aspects of social systems, illustrated here by the the example of the COVID-19 pandemic. Fig 1 shows the evolution of the median eigenvector centrality of COVID-19 symptom related themes. These themes were selected manually for their relevance, such as "pneumonia", "fever" or "cough", within monthly graphs for the US, Germany and Japan (see section 8.1 for a complete list of themes used). The monthly median eigenvector centralities reflect the timing and impact of events in specific geographies on the example of the COVID-19 outbreak in 2020. Prior to end of 2019, the median eigenvector centralities associated with COVID-19 symptoms were consistently low for all three countries, then picking up notably as the virus spread across the globe, first in Japan, then in Germany and at last in the US.

5.2. Granger causality test results

The eigenvector centralities from theme-based knowledge graphs and the IP index values for three countries are tested for Granger causality, with a maximum lag of three months. Table 3 exhibits the number of BH-adjusted p-values that exhibit significance at 5% for each country's IP index.

Country \ Data set	GC (filt)	Reverse GC	GC (unfilt)
US	673	583	94
Germany	783	742	67
Japan	552	556	99

Table 3: IP: Number of significant BH-adjusted p-values. GC refers to Granger causality; filt (unfilt) refers to filtered (unfiltered) GDELT data

The Granger causality analysis results in a considerably larger number of statistically significant BH-adjusted p-values for those features generated from filtered GDELT data compared to those from unfiltered GDELT data, suggesting that the filtering methodology applied in section 3.2 isolates relevant signals.

Strong reverse Granger causality exists between IP and the eigenvector centralities generated from monthly theme-based knowledge graphs. This is unsurprising given that press reporting is bi-directional. Events – extracted by GDELT as themes – covered by global newspapers impact macroeconomic variables such as IP while the state of the economy – as measured in macroeconomic variables – is commented on in the press.

5.3. Forecast error analysis

For the US, Germany and Japan, the respective eigenvector centrality matrices are condensed into five factors applying PLS. They are then incorporated into the forecasting framework described in section 4.2 to predict IP. All models have a lag of one month, determined by evaluating AIC and BIC.

The columns of Table 4 show the performance metric (RMSE) from the factor augmented models compared to three benchmarks, which are based on autoregressive models – the first contains the predicted variable and explanatory macroeconomic variables only (referred to as BM1), the second includes the predicted variable, explanatory macroeconomic variables and portrait divergence scores (referred to as BM2) and the third incorporates the predicted variable, explanatory macroeconomic variables and five components from eigenvector centralities derived from unfiltered GDELT data (BM3). The numbers in the cells represent the RMSE in percentage terms for each model and its benchmarks. Blue (red) cells denote cases in which the models outperform (underperform) the respective benchmarks. In the column "Sign.", numbers in parentheses correspond to the number of significant coefficients associated with GDELT factors in the model in Eq. (1), with the asterisks denoting the level of their statistical significance.

IP for	Model	BM1	BM2	BM3	Sign.
US	1.6139	1.6341	1.6376	1.6144	***(1), **(1)
Germany	2.6942	2.7079	2.7082	2.7159	*(1)
Japan	2.4978	2.5300	2.5321	2.5049	*(1)

Table 4: Results of the model in Eq. 1 applied to IP. Numbers represent the RMSE (%). Blue (red) cells denote cases in which the model outperforms (underperforms) the benchmark. Numbers in parentheses correspond to the number of significant coefficients associated with GDELT factors in the model in Eq. (1) ($***$ denotes at least one GDELT sentiment factor with p-value < 0.01, $**$ < 0.05, $*$ < 0.1).

Rows one to three in Table 4 show that the models containing the filtered GDELT factors outperform all three benchmarks for the US, Germany and Japan, respectively. All factor augmented models contain at least one statistically significant GDELT factor at 0.1, 0.05 and 0.01, each.

In 2020, IP for all three countries experienced high levels of volatility as the lockdowns imposed by governments around the world severely curbed economic activity. In the cross-validation, the last validation set incorporates the period of the COVID-19 outbreak in 2020. Predictions on this last validation set exhibit a much larger error metric across countries than those predictions on the validation sets that exclude the outbreak. However, performance dynamics during the COVID-19 outbreak remain the same in that the factor enhanced models outperform their benchmarks.

Table 5 contains the p-values from the modified Diebold Mariano test. This test is used to to gauge if factor enhanced model forecasts are significantly different from the benchmark predictions as set out in section 5.3.

IP for	Data set	Model - BM1	Model - BM2	Model - BM3
US		0.0000	0.0000	0.0000
Germany		0.0103	0.0093	0.0001
Japan		0.0000	0.0000	0.0887

Table 5: p-values from modified Diebold Mariano test

According to the modified Diebold Mariano test, all model forecasts for IP are statistically different to BM1, BM2 and BM3 predictions at 1 % or 10 % significance, respectively.

Results suggest that features derived from narrative-based knowledge graphs improve IP forecasts for three large economies. In particular, the factor augmented autoregressive models deliver consistently better performance compared to those models that contain a single graph similarity measure (BM2) or factors based on unfiltered GDELT data (BM3). This indicates that eigenvector centralities are sufficiently nuanced and thus well suited to capturing the changing dynamics in theme-based graphs while this is not achieved by the portrait divergence, which is a single score measuring the change in topological properties.

5.4. Drivers of GDELT factors

In this section, we examine the loadings corresponding to each component to gain an understanding of the relationship between themes and the PLS components extracted from GDELT data. Loadings correspond to the strength of relationship between the original eigenvector centralities for each graph node (represented by a GDELT theme) and the PLS components, quantifying the importance of the underlying themes in each of them.
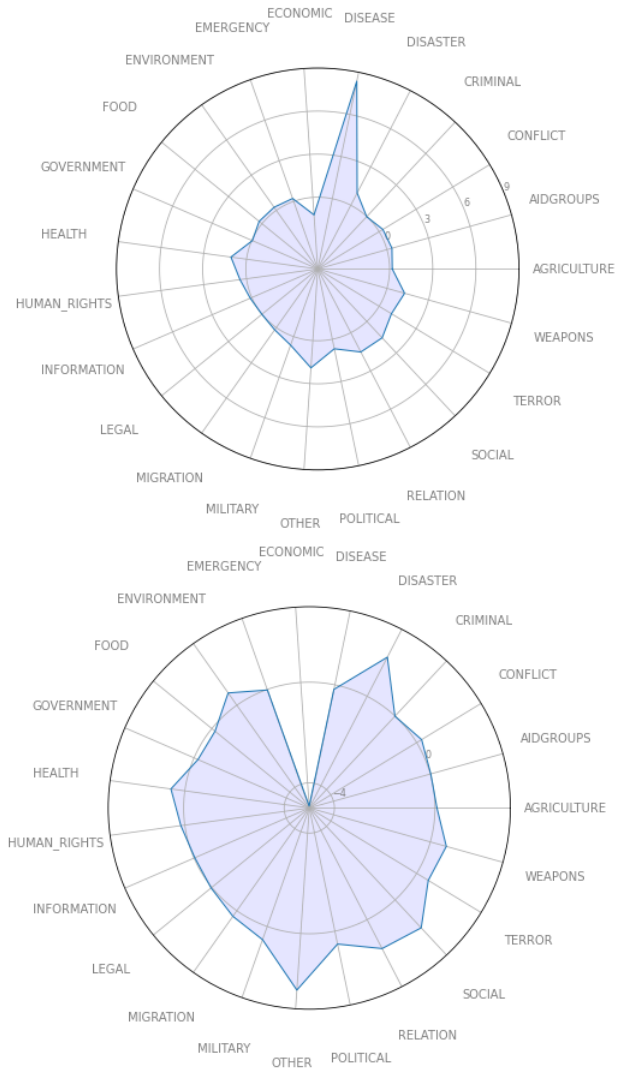

Figure 2: Top: Significant PLS components explained by theme categories (Factor 1, US). Bottom: Significant PLS components explained by theme categories (Factor 4, US)

GDELT themes represent specific events or conditions such as ''inflation'', "economic growth" or "disease" and are mapped to 22 distinct theme categories using the GDELT naming convention. For example, themes containing the word "disease" are mapped to the disease category; themes including the word "weapon" are mapped to the weapons category, etc (see section 8.2 for detailed list).

For each component, the loadings are summed according

to the 22 distinct theme categories set out above. Mapping GDELT themes to theme categories delivers insights into the relationship of these categories with each PLS component.

In Fig. 2 we show radar charts of the theme categories associated with the loadings of the statistically significant PLS components used to forecast IP in the US. Table 4 illustrates that the corresponding model outperforms the three benchmarks and exhibits substantial statistical significance.

This example demonstrates that the factors we use to model IP can be linked to distinct theme categories. Thus, change in such categories over time helps explain movements in IP, which is a monthly proxy of economic activity.

Of the 22 distinct theme categories, "disease" and "economic" exhibit the strongest relationships for factor 1 and factor 4, respectively and therefore are their most important drivers with the strongest predictive power. This also applies to the statistically significant PLS components for the models predicting German and Japanese IP, which are driven by "economic" and "disease" categories, respectively. These charts can be made available on request.

6. Discussion

Our study presents a novel method of incorporating themes from global newspaper narrative into macroeconomic forecasts. We apply an effective filtering methodology for extracting relevant signals from a large volume of data. The filtered data is used to build theme-based weighted undirected knowledge graphs for each calendar month and for three large economies, respectively. For comparison purposes, monthly graphs based on unfiltered GDELT data are generated. For each monthly graph, we calculate the eigenvector centralities and portrait divergences, creating a monthly frequency time series of features. On the example of the COVID-19 outbreak end of 2019, we show that the change in graph centralities is indicative of genuine real-world change. This finding supports hypothesis H_1.

The monthly eigenvector centralities for each of the three economies exhibit robust Granger causality, indicating a statistical relationship between graph features and the IP values in corresponding countries. For each country, we reduce the eigenvector centralities into five components applying PLS and incorporate them into a factor augmented autoregressive framework. Results show that features derived from theme-based graphs significantly improve IP forecasts for the US, Germany and Japan. In particular, the factor augmented models consistently outperform the benchmark models that contain a single graph similarity measure only (portrait divergence) as well as the benchmark models that incorporate five PLS components derived from unfiltered GDELT data. These results support hypothesis H_2. Grouping over one thousand GDELT themes into 22 distinct theme categories helps interpret the relationship between those categories and each PLS factor, with themes related to "disease" and "economic" being their most important drivers.

Our work makes contributions to three main streams of research. First, our study expands existing research on using big data and machine learning in macroeconomics (Giannone et al., 2008; McCracken & Ng, 2016; Baker et al., 2016; Coulombe et al., 2020). We leverage machine learning, graph analytics and natural language processing to represent the state of the economy in terms of themes derived from news narrative, and demonstrate how these themes can be incorporated into macroeconomic forecasting models to improve predictions of economic activity. This use of news narrative is – to the best of our knowledge – new. Second, we contribute to the literature on economic graphs, which have a wide range of applications and often span across several disciplines (Emmert-Streib et al., 2018). Our work adds to existing research on information extraction, feature generation and macroeconomic forecasting using textual data at scale, presenting a data-driven approach to operationalising such concepts and to integrating them into macroeconomic forecasting frameworks. We supplement our findings with an interpretability analysis illustrating that disease and economy-related themes have the strongest predictive power. Third, we expand the research on tracking change in social systems through narrative. As nodes in a graph, themes convey detailed information about the state of a system, while the evolution of their properties over time is indicative of changes. These narrative-based graphs can be modified to track specific phenomena such as "inflation", "migration" or "human rights violations" in real-time and, therefore, have a broad range of applications.

6.1. Limitations and ideas for further research

Some potential limitations of our study should be acknowledged. First, the forecasting framework focuses on linear relationships between predictors and predicted variables. Exploring non-linear modelling techniques may generate further understanding of the interaction between these variables and could be an extension to this project. Second, our work represents a proof of concept and does not focus on performance optimisation. In addition, our study is limited to predicting one macroeconomic variable for three large economies. Its pertinence to real-world applications could be enhanced by expanding to a broader range of variables and economies. Third, GDELT has a relatively short track record, going back to February 2015. The limited amount of observations is likely to affect results, in particular when modelling monthly frequency data. Fourth, this study explores a limited number of graph features and graph analysis techniques. Our work could be expanded by considering alternative centrality and similarity measures for inclusion into the forecasting framework.

7. Conclusions

This study proposes a new method of incorporating themes derived from global newspaper narrative into macroeconomic forecasts, and demonstrates that factors derived from these themes significantly improve predictions of economic activity for three large economies. The interpretation of the factors extracted from eigenvector centralities shows that themes associated with the "disease" and "economic" categories have

the strongest predictive power and therefore help explain changes in economic activity. The theme-based knowledge graphs used in this study represent a nuanced picture of the state of the economy at a given point in time. They can be modified to monitor specific aspects of social systems in real-time and thus have a wide range of applications.

8. Appendix

8.1. COVID-19 related themes

The themes listed below represent COVID-19 related symptoms. Applying these themes, we calculate the median eigenvector centralities for each calendar month and illustrate the evolution of the eigenvector centralities over time on the example of the COVID-19 outbreak.

- `WB_2165_HEALTH_EMERGENCIES`
- `WB_1406_DISEASES`
- `TAX_DISEASE_CORONAVIRUS`
- `TAX_DISEASE_EPIDEMIC`
- `TAX_DISEASE_OUTBREAK`
- `TAX_DISEASE_INFECTION`
- `TAX_DISEASE_PNEUMONIA`
- `TAX_DISEASE_FEVER`
- `TAX_DISEASE_INFECTIOUS`
- `TAX_DISEASE_FLU`
- `TAX_DISEASE_COUGH`

8.2. GDELT theme categories

All GDELT themes are assigned to one of 26 theme categories. For example, the "Weapons" category includes 81 themes such as "`TAX_WEAPONS_GUNS`", "`TAX_WEAPONS_BOMB`" and "`TAX_WEAPONS_SUICIDE_BOMB`". We removed themes belonging to "Actor", "Language", "Animal" and "Ethnicity" for our analysis as they are purely descriptive.

- Economic
- Disease
- Actor
- Language
- Ethnicity
- Animal
- Disaster
- Social
- Relation
- Political
- Health
- Weapons
- Military
- Terror
- Environment
- Food
- Government
- Aid groups
- Information
- Conflict
- Emergency
- Human rights
- Migration
- Legal
- Criminal
- Other (various events or conditions, c 6% of themes)

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Adamic, L., Brunetti, C., Harris, J. H., & Kirilenko, A. (2017). Trading networks. *The Econometrics Journal*, *20*(3), S126–S149. doi: https://doi.org/10.1111/ectj.12090

Baker, S., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The quarterly journal of economics*, *131*(4), 1593–1636. doi: https://doi.org/10.1093/qje/qjw024

Bellomarini, L., Benedetti, M., Gentili, A., Laurendi, R., Magnanimi, D., Muci, A., & Sallinger, E. (2020). Covid-19 and company knowledge graphs: assessing golden powers and economic impact of selective lockdown via ai reasoning. *arXiv preprint arXiv:2004.10119*.

Benjamini, Y., & Yekutieli, D. (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, *100*(469), 71–81. doi: https://doi.org/10.1198/016214504000001907

Bildirici, M. E., Kayıkçı, F., & Onat, I. Ş. (2015). Baltic dry index as a major economic policy indicator: the relationship with economic growth. *Procedia-Social and Behavioral Sciences*, *210*, 416–424. doi: https://doi.org/10.1016/j.sbspro.2015.11.389

Bonaccorsi, G., Riccaboni, M., Fagiolo, G., & Santoni, G. (2019). Country centrality in the international multiplex network. *Applied Network Science*, *4*(1), 1–42. doi: https://doi.org/10.1007/s41109-019-0207-3

Brooks, C., & Tsolacos, S. (2010). *Real estate modelling and forecasting*. doi: https://doi.org/10.1017/CBO9780511814235

Brosch, T., Scherer, K. R., Grandjean, D. M., & Sander, D. (2013). The impact of emotion on perception, attention, memory, and decision-making. *Swiss medical weekly*, *143*, w13786. doi: https://doi.org/10.4414/smw.2013.13786

Bruner, J. S. (1990). *Acts of meaning* (Vol. 3). Harvard University Press.

Buono, D., Kapetanios, G., Marcellino, M., Mazzi, G. L., & Papailias, F. (2018). Evaluation of nowcasting/flash estimation based on a big set of indicators..

Campi, M., Dueñas, M., & Fagiolo, G. (2020). How do countries specialize in agricultural production? a complex network analysis of the global agricultural product space. *Environmental Research Letters*, *15*(12), 124006. doi: https://doi.org/10.1088/1748-9326/abc2f6

Carvalho, V. M., Nirei, M., Saito, Y., & Tahbaz-Salehi, A. (2016). Supply chain disruptions: Evidence from the great east japan earthquake. *Columbia Business School Research Paper*(17-5). doi: http://dx.doi.org/10.2139/ssrn.2883800

Christiano, L. J., Eichenbaum, M., & Evans, C. L. (2005). Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of political Economy*, *113*(1), 1–45. doi: https://doi.org/10.1086/426038

Clore, G. L., & Palmer, J. (2009). Affective guidance of intelligent agents: How emotion controls cognition. *Cognitive systems research*, *10*(1), 21–30. doi: https://doi.org/10.1016/j.cogsys.2008.03.002

Colladon, A. F., Guardabascio, B., & Innarella, R. (2019). Using social network and semantic analysis to analyze online travel forums and forecast tourism demand. *Decision Support Systems*, *123*, 113075. doi: https://doi.org/10.1016/j.dss.2019.113075

Constantin, A., Peltonen, T. A., & Sarlin, P. (2018). Network linkages to predict bank distress. *Journal of Financial Stability*, *35*, 226–241. doi: https://doi.org/10.1016/j.jfs.2016.10.011

Coulombe, P. G., Leroux, M., Stevanovic, D., & Surprenant, S. (2020). How is machine learning useful for macroeconomic forecasting? *arXiv preprint arXiv:2008.12477*.

Cubadda, G., & Guardabascio, B. (2012). A medium-n approach to macroeconomic forecasting. *Economic Modelling*, *29*(4), 1099–1105. doi: https://doi.org/10.1016/j.econmod.2012.03.027

De Jong, S. (1993). Simpls: an alternative approach to partial least squares regression. *Chemometrics and intelligent laboratory systems*, *18*(3), 251–263. doi: https://doi.org/10.1016/0169-7439(93)85002-X

Elshendy, M., Colladon, A. F., Battistoni, E., & Gloor, P. A. (2018). Using four different online media sources to forecast the crude oil price. *Journal of Information Science*, *44*(3), 408–421. doi: https://doi.org/10.1177/0165551517698298

Emmert-Streib, F., Tripathi, S., Yli-Harja, O., & Dehmer, M. (2018). Understanding the world economy in terms of networks: A survey of data-based network science approaches on economic networks. *Frontiers in Applied Mathematics and Statistics*, *4*, 37. doi: https://doi.org/10.3389/fams.2018.00037

Gdelt project. (2015). Retrieved from https://www.gdeltproject.org/ Accessed15May2020

Ghalmane, Z., Cherifi, C., Cherifi, H., & El Hassouni, M. (2020). Extracting backbones in weighted modular complex networks. *Scientific Reports*, *10*(1), 1–18. doi: https://doi.org/10.1038/s41598-020-71876-0

Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, *55*(4), 665–676. doi: https://doi.org/10.1016/j.jmoneco.2008.05.010

Girardi, A., Guardabascio, B., & Ventura, M. (2016). Factor-augmented bridge models (fabm) and soft indicators to forecast italian industrial production. *Journal of Forecasting*, *35*(6), 542–552. doi: https://doi.org/10.1002/for.2393

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, 424–438. doi: https://doi.org/10.2307/1912791

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 ieee international joint conference on neural networks, 2005.* (Vol. 4, pp. 2047–

2052).

Guo, L., & Vargo, C. J. (2020). Predictors of international news flow: Exploring a networked global media system. *Journal of Broadcasting & Electronic Media*, *64*(3), 418–437. doi: https://doi.org/10.1080/08838151.2020.1796391

Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of forecasting*, *13*(2), 281–291. doi: https://doi.org/10.1016/S0169-2070(96)00719-4

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780. doi: https://doi.org/10.1162/neco.1997.9.8.1735

Keynes, J. M. (2018). *The general theory of employment, interest, and money*. Springer.

King, G., Schneer, B., & White, A. (2017). How the news media activate public expression and influence national agendas. *Science*, *358*(6364), 776–780. doi: https://doi.org/10.1126/science.aao1100

Leamer, E. E. (1985). Self-interpretation. *Economics and Philosophy*, *1*(2), 295–302. doi: doi.org/10.1017/S0266267100002546

Leetaru, K. H. (2012). Fulltext geocoding versus spatial metadata for large text archives: Towards a geographically enriched wikipedia. *D-lib Magazine*, *18*(9), 5. doi: https://doi.org/10.1045/september2012-leetaru

Leetaru, K. H., Perkins, T., & Rewerts, C. (2014). Cultural computing at literature scale: encoding the cultural knowledge of tens of billions of words of academic literature. *D-lib Magazine*, *20*(9), 8. doi: https://doi.org/10.1045/september2014-leetaru

Marcaccioli, R., & Livan, G. (2019). A pólya urn approach to information filtering in complex networks. *Nature communications*, *10*(1), 1–10. doi: https://doi.org/10.1038/s41467-019-08667-3

Matesanz Gomez, D., Ferrari, H. J., Torgler, B., & Ortega, G. J. (2017). Synchronization and diversity in business cycles: a network analysis of the european union. *Applied Economics*, *49*(10), 972–986. doi: https://doi.org/10.1080/00036846.2016.1210765

McCracken, M. W., & Ng, S. (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, *34*(4), 574–589. doi: https://doi.org/10.1080/07350015.2015.1086655

Nyman, R., Kapadia, S., Tuckett, D., Gregory, D., Ormerod, P., & Smith, R. (2018). *News and narratives in financial systems: exploiting big data for systemic risk assessment.* Retrieved from https://www.bankofengland.co.uk/working-paper/2018/news-and-narratives-in-financial-systems/Accessed30October2019

Piccardi, C., & Tajoli, L. (2018). Complexity, centralization, and fragility in economic networks. *PloS one*, *13*(11), e0208265. doi: https://doi.org/10.1371/journal.pone.0208265

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, *45*(11), 2673–2681. doi: https://doi.org/10.1109/78.650093

Serrano, M. Á., Boguná, M., & Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the national academy of sciences*, *106*(16), 6483–6488. doi: https://doi.org/10.1073/pnas.0808904106

Shiller, R. J. (2017). Narrative economics. *American Economic Review*, *107*(4), 967–1004. doi: https://doi.org/10.1257/aer.107.4.967

Smets, F., & Wouters, R. (2007). Shocks and frictions in us business cycles: A bayesian dsge approach. *American economic review*, *97*(3), 586–606. doi: https://doi.org/10.1257/aer.97.3.586

Stern, S., Livan, G., & Smith, R. E. (2020). A network perspective on intermedia agenda-setting. *arXiv preprint arXiv:2002.05971*. doi: https://doi.org/10.1007/s41109-020-00272-4

Stock, J. H., & Watson, M. W. (2001). Vector autoregressions. *Journal of Economic perspectives*, *15*(4), 101–115. doi: https://doi.org/10.1257/jep.15.4.101

Tantardini, M., Ieva, F., Tajoli, L., & Piccardi, C. (2019, 11). Comparing methods for comparing networks. *Scientific Reports*, *9*. doi: https://doi.org/10.1038/s41598-019-53708-y

Temizsoy, A., Iori, G., & Montes-Rojas, G. (2016). Network centrality and funding rates in the e-mid interbank market (16/08). *London, UK: Department of Economics, City, University of London. This is the published version of the paper. This version of the publication may differ from the*

final published version. doi: https://doi.org/10.1016/j.jfs.2016.11.003

Tilly, S., Ebner, M., & Livan, G. (2021). Macroeconomic forecasting through news, emotions and narrative. *Expert Systems and Applications.* doi: https://doi.org/10.1016/j.eswa.2021.114760

Tobias, R. D. (1995). An introduction to partial least squares regression. In *Proceedings of the twentieth annual sas users group international conference* (Vol. 20).

Tuckett, D., Ormerod, P., Smith, R., & Nyman, R. (2014). Bringing social-psychological variables into economic modelling: Uncertainty, animal spirits and the recovery from the great recession. *Economic Growth eJournal.* doi: https://doi.org/10.2139/ssrn.2408155

Tumminello, M., Micciche, S., Lillo, F., Piilo, J., & Mantegna, R. N. (2011). Statistically validated networks in bipartite complex systems. *PloS one*, *6*(3), e17994. doi: https://doi.org/10.1371/journal.pone.0017994

Van Eyden, R., Difeto, M., Gupta, R., & Wohar, M. E. (2019). Oil price volatility and economic growth: Evidence from advanced economies using more than a century's data. *Applied Energy*, *233*, 612–621. doi: https://doi.org/10.1016/j.apenergy.2018.10.049

Yang, Y., Pang, Y., & Huang, G. (2020). The knowledge graph for macroeconomic analysis with alternative big data. *arXiv preprint arXiv:2010.05172.*