CAUSALITY LINK

Automatic Detection of Corpus Incoherence Through Causal Knowledge Graph

Olav Laudy, Chief Data Scientist; Pierre Haren, CEO; Eric Jensen, CTO

In this paper, we describe a method to detect corpus incoherence on the points of view of multiple related indicators.

We have described in other papers available on www.causalitylink.com how, through our universal data model of finance concepts and our natural language processing system, we are able to extract four types of data structures from a growing corpus of 84 million documents that are important for this paper: indicators, trends, events and causal links.

The indicators and trends describe the permanent variations of what can be called the "signal", that is the myriad of ever-changing data that are used to describe the financial world: the GDP of countries; the revenues of companies, by country or product; the demand, production and price of commodities; and a large number of other variables. These data can correspond to past measures, or forecasts from different authors, so they all have a date at which they were published, and a date at which they became true, or will become known.

The causal links represent the "model" that people have built over time of causal relationships between these indicators and expressed in the documents we analyzed. Causal links can represent accounting rules, such as "the growth of the sales of Ford explain its increasing profit", manufacturing constraints, such as "the increase in the price of steel has increased the costs of goods sold by Ford", and many other relationships that humans have discovered between indicators.

The automatically generated knowledge graph linking our indicators (in the tens of millions as of today) through causal links (about 6 million today) is a large knowledge graph that can be used for multiple purposes, to explain and potentially predict the movements of the different indicators.

One such usage is the subject of this paper: the detection of some types of transient incoherence in the graph.

The signal

Our NLP system detects statements about the evolution of indicators, or trends. Authors will either comment on the past ("yesterday, the stock price of Tesla increased by 5% to $735 per share"), or forecast the indicator in the future ("demand for new vehicles in the US will be strongly reduced in Q2 2020"). The trends can be either up (when the indicator goes up), down, flat, or unknown. We define Positive Trend Percentage ("PTP") for an indicator to be the percentage of positive trends over the total of positive and negative trends measured for a certain time horizon. PTPs vary between 0 and 100. At 50, we have a split consensus. Closer to 100 indicates a bullish consensus, and below 50 a bearish consensus for that indicator. Leveraging the past and future trends with the PTP, two key metrics available are "Past PTP" and "Future PTP".

One of the things we witnessed while developing this signal was the excellent correlation between known indicators (such as stock prices) and the Past PTP about the same stock price, which corresponds to statements made about the positive or negative evolution of that stock price in the past. We attribute this to an interesting feature of the "wisdom of the crowds". In our case: the higher the percentage of trend on a given day, the higher the number of mentions of that change during the day.

Figure 1 shows this interesting correlation between the evolution of the daily stock price of Ford (in red) and our Past PTP (in blue) extracted from statements in our corpus for the same indicator.

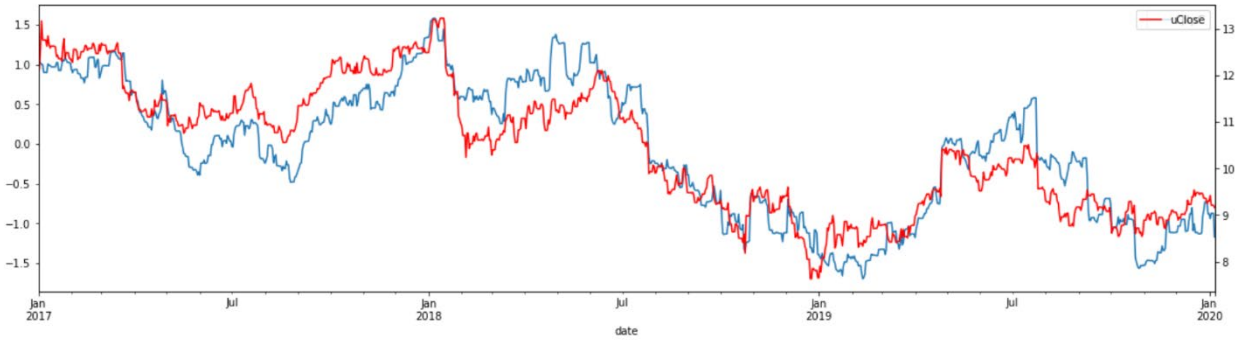

Figure 1: Comparison between the stock price of Ford and the Past PTP of the stock-price

Another promising characteristic of our PTP is that it is available as a continuous measure over time, even if the underlying indicator is only published quarterly, as is the case for almost all company-related fundamental indicators. This is because inside a large set of documents analyzed every day (we integrate 40,000 documents in 18 languages every day), there is a constant chatter about these indicators which measures the current understanding and expectations by the market of the past and future evolution of these indicators.

The one-hop model

Our causal model enables us to extract, for every "target" indicator, the "driver" indicators which influence this target. We call them the "drivers" of that "target".

It is therefore possible to pick a target (such as the stock price of a particular company, or the price of some commodity), retrieve from the graph the drivers of that target, as well as the strength and direction of the relations between each driver and the target, and build a one-hop causal model linking these drivers with the target.

A linear model that computes the expected evolution of the chosen target through the expected evolution of its drivers (their Future PTPs) can be generated by weighing the importance of the evolution of each driver by the strength of the driver-to-target causal link, or elasticity. The resulting synthetic number or "background PTP" represents over time the coherent view of the evolution of the target if our authors are right in their explanations of the past movement of the target and if these forces are still acting today.

We can then compare this background PTP with the future for any indicator. This provides an answer to the question: are the views of our authors consistent between one target indicator and its network of drivers?

Our current results indicate that in general, our automatically generated causal models for companies align well with the evolution of the most important indicators of these companies, such as their revenue, their profits, and their stock price, as shown in Figure 2.

Because the background PTP is a linear combination of several other future PTPs, it is reasonably stable, and seems to correspond to a rational point of view, versus the target itself, which can be subject to all sorts of biases.

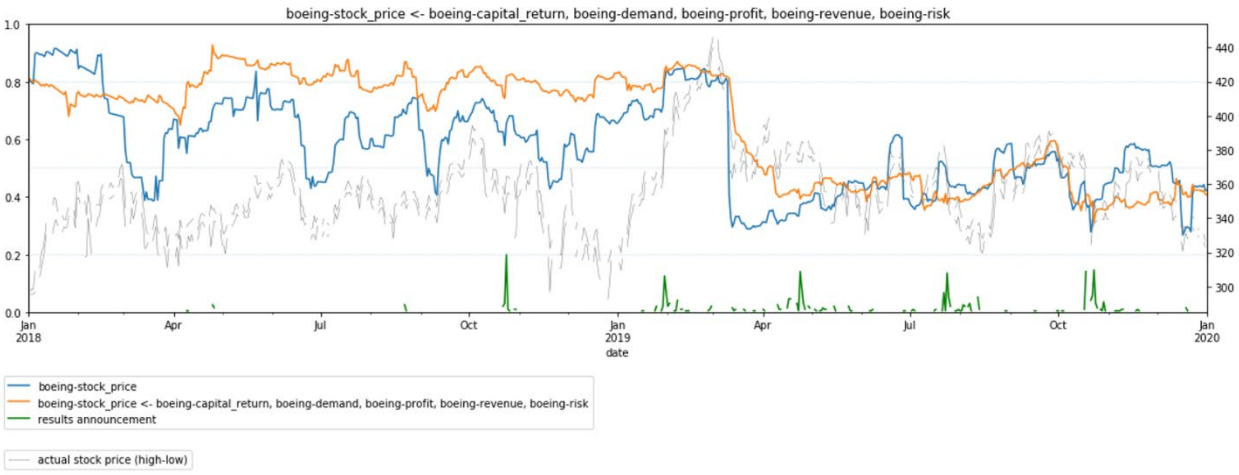

Figure 2: For Boeing, comparison of stock price, stock-price PTP and stock-price background PTP

In Figure 2, we compare Boeing's actual stock price high and low (in grey) with the PTP for the stock-price (in blue) and the background PTP (in orange), which leverages the PTPs of the following five most important Boeing stock price drivers according to the knowledge graph: revenue, demand, profit, capital return and risk.

Two-hop model

Financial analysts who are tracking the performance of companies are keeping an eye on the direct drivers of the stock price and these drivers mostly represent standard accounting knowledge as demonstrated in the drivers of Boeing example of Figure 2.

It is therefore interesting to leverage further the causal graph and expand the graph to the drivers of the drivers and utilize a more specific "two-hop model".

We present in Figure 3 a simplified "two-hop model" for Boeing. This model is extracted from the causal graph by ranking the most important drivers in decreasing order.

From this model, we can generate a “two-hop background PTP” as another linear combination of the drivers-of-drivers like the “one-hop background PTP” but leveraging information about external indicators two hops away rather the internal ones only one-hop away from the target.

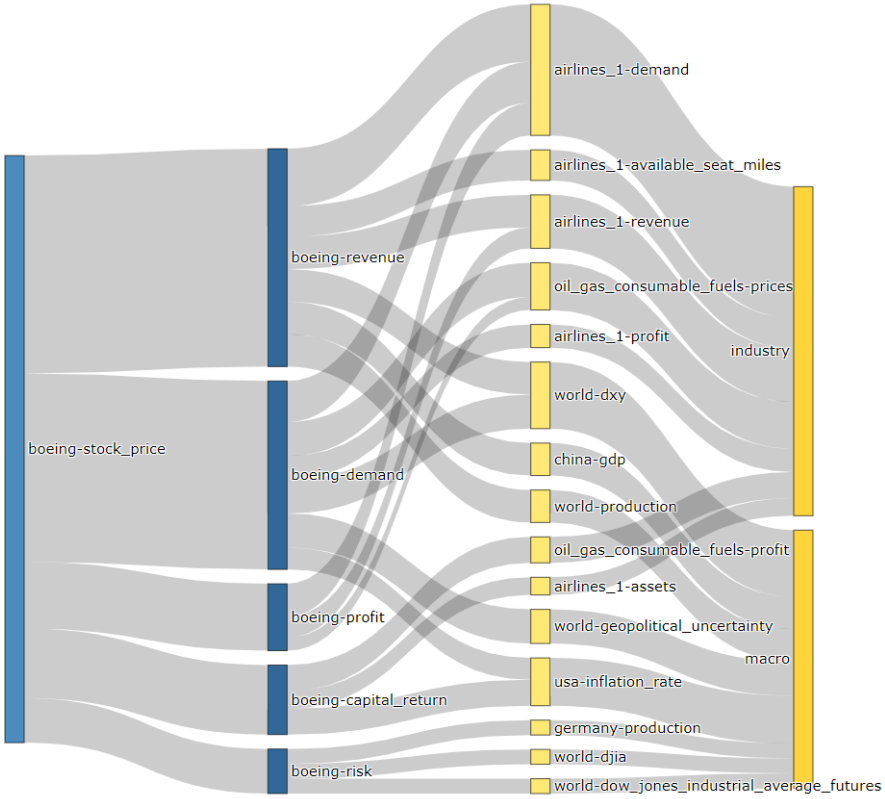

Figure 3: The two-hop model of indicators influencing Boeing stock price

It is now possible to generate a comparison between the one-hop background PTP, and the two-hop background PTP, which is leveraging the PTPs of indicators that are not specific to Boeing. At this point, we have not included the indicators linked to the partner companies, or the competition, nor have we taken into account the events that we detect but do not yet integrate in the computation of the background PTP.

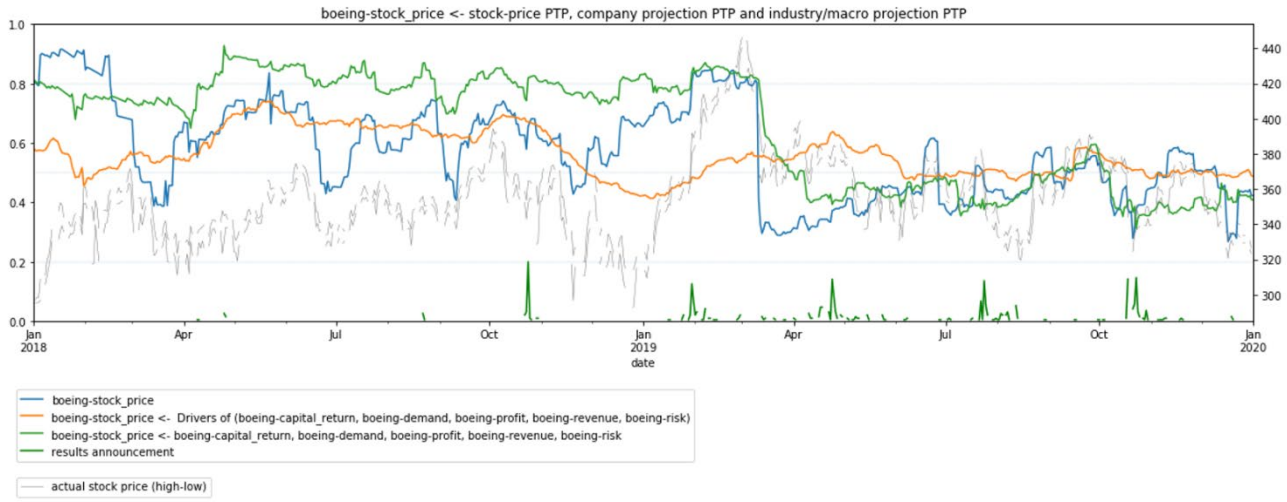

Figure 4: Boeing stock price vs the stock-price PTP, the one-hop and the two-hops background PTPs

The resulting two-hop background PTP curve (in orange in Figure 4) is smoother than the one-hop background PTP curve as it incorporates more indicators, and it corresponds to a model of the background forces acting on Boeing, as opposed to the internal forces acting on Boeing (the green curve) and the sentiment on the Boeing stock price (the blue curve).

Detection of transient incoherence and two-hop models

When comparing these two-hop background PTP and direct PTP curves, we can sometimes observe a vertical gap between the two signals, which most of the time is closed after a few months by a movement of the target.

We can observe this phenomenon with General Electric in Figure 5.

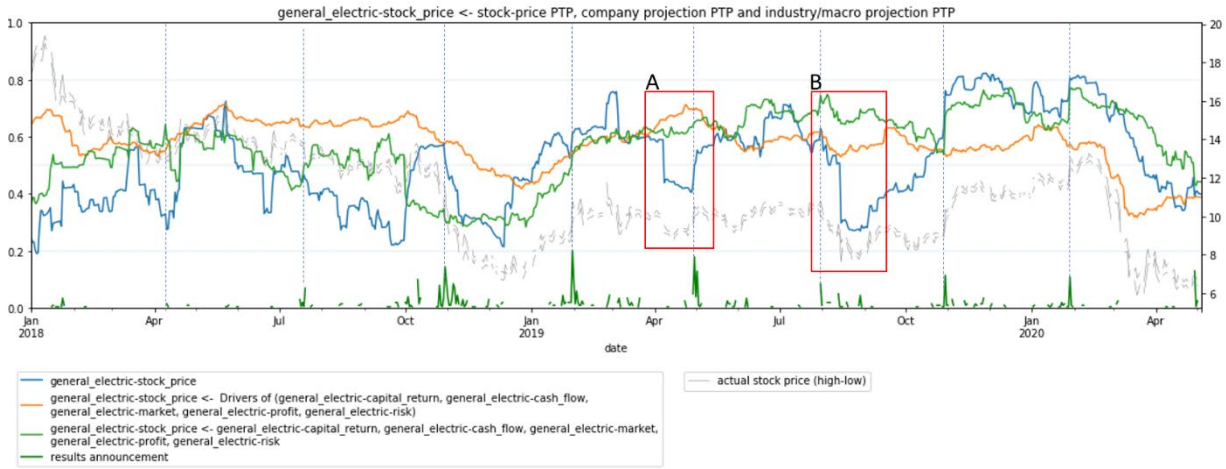

Figure 5: GE stock price vs the stock-price PTP, the one-hop and the two-hop background PTP

For example, we see that in May 2019 (red block A), the reports on GE's stock price took a very negative turn, while the reporting on the company's KPIs as well as the industry and macro effects indicated no change. In fact, the external forces showed to be a good tailwind. The PTPs, as good explanatory AI requires, are linked to the sentences of the articles in the corpus. Investigating the links, we find that the drop in stock price was due to a downgrade by a prominent Goldman Sachs analyst. The stock recovered a month later with an above expected Q2 results announcement. The incoherence demonstrated here shows how the two projected PTP signals can be used to shine a new perspective on a sudden stock drop.

The stock price took a hit again in September 2019 (red block B), due this time to the announcement of a lawsuit. Again, while the sentiment on the stock-price tanked, the actual background PTP stayed relatively stable reflecting an unchanged business environment. It was not until the consequences of the coronavirus started to degrade simultaneously all the PTPs for the network of drivers that our background PTP plunged.

The detection of such transient incoherence can be readily automated using our causal graph data available over the past 6 years. We are in the process of refining this process and measuring the performance of a simulated portfolio which would trade upon this new signal. We will publish our results when they become available.

How close are corpus incoherence to market incoherence?

Any specific instance of the Causality Link Research Assistant is only as good as the corpus it ingests. Our base corpus relies on the Naviga content aggregator corpus, as well as a smaller number of government and public web sites publications. With over 84 million texts over the past 5 years, this content provides an excellent base level of general information about public companies worldwide and the countries in which they operate.

We are actively pursuing discussions with other content providers to include their content in our offering.

We are also able to ingest proprietary content for specific customers and compare or aggregate the results achieved on that specific content with our base results for these customers only.

Consistent with the theory of the "wisdom of crowds", we find that in general, the more content, the higher the quality of our results compared to the market. Our PTPs for the companies that are most talked about have a higher quality of correlation with the actual values of the underlying indicators as we showed in Figure 1.

This gives a strong indication that the aggregation of a very large corpus of texts coming from various sources tends to produce results converging towards an accurate representation of the market view. In a sense, it is a case where the corpus representation matches more and more accurately the market representation, and therefore the results on corpus incoherence exhibit a high probability to represent market incoherence. Further statistical results demonstrating this concept will be made available in future publications.

Conclusion

In this paper, we have introduced a novel way to leverage the combination of an automatically generated signal (the PTP) with an automatically generated causal model of the relations between the PTPs that produces a novel set of alerts relying on transient corpus incoherence detection.

The same methodology applied to different corpus leads to different alerts, thus providing a strong incentive for potential users to contribute their private corpus to their specific version of our Research Assistant in order to leverage the knowledge contained in private documents to achieve the best results.

We believe this is another demonstration of the power of causal graphs applied to the finance sector.