SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
INFORMATION TECHNOLOGY
COMPUTER SCIENCE
COMPLEX NETWORKS

Received
17 February 2014

Accepted
2 May 2014

Published
22 May 2014

Correspondence and requests for materials should be addressed to T.Š. (tomislav.smuc@irb.hr)

Cohesiveness in Financial News and its Relation to Market Volatility

Matija Piškorec[1], Nino Antulov-Fantulin[1], Petra Kralj Novak[2], Igor Mozetič[2], Miha Grčar[2], Irena Vodenska[3] & Tomislav Šmuc[1]

[1]Laboratory for Information Systems, Division of Electronics, Ruđer Bošković Institute, Croatia, [2]Department of Knowledge Technologies, Jožef Stefan Institute, Slovenia, [3]Department of Administrative Sciences, Metropolitan College, Boston University, USA.

Motivated by recent financial crises, significant research efforts have been put into studying contagion effects and herding behaviour in financial markets. Much less has been said regarding the influence of financial news on financial markets. We propose a novel measure of collective behaviour based on financial news on the Web, the News Cohesiveness Index (NCI), and we demonstrate that the index can be used as a financial market volatility indicator. We evaluate the NCI using financial documents from large Web news sources on a daily basis from October 2011 to July 2013 and analyse the interplay between financial markets and finance-related news. We hypothesise that strong cohesion in financial news reflects movements in the financial markets. Our results indicate that cohesiveness in financial news is highly correlated with and driven by volatility in financial markets.

The exponential growth of online media, expansion of communication and mobility-tracking capabilities have spawned research regarding the utility of the big data available from these sources. Big-data analytics aims to provide tools for better understanding large techno-social systems[1,2], improve predictions of different socio-economic outcomes and optimise processes. For example, Gonzales et al.[3] use 100,000 trajectories of mobile phone users to explain human mobility patterns. Ginsberg et al.[4] use Google search queries to help detect outbreaks of influenza epidemics in areas with a large population of web-search users. Whereas the aforementioned work estimates the current state of disease spread, other works focus on the predictive value of online information. For example, Goel et al.[5] demonstrate that Google search query volumes significantly improve predictions for the revenue of featured movies, video game sales and rank of songs. Similar to the above studies, our work explores the relationship between large corpora of online news and financial markets.

In this context, previous studies have analysed the relationship of search query volumes of specific terms with movements in financial markets of related items[6]. Bordino et al.[7] demonstrate that daily trading volumes of stocks traded on the NASDAQ 100 are correlated with the daily volumes of Yahoo queries related to the same stocks and that query volumes can anticipate peaks of trading by one or more days. Dimpfl et al.[8] report that Internet search queries for the term "dow" obtained from Google Trends can help predict the Dow Jones Industrial Average (DJIA) realised volatility. Vlastakis et al.[9] study information demand and supply using Google Trends at the company and market level for 30 of the largest stocks traded on the NYSE and NASDAQ 100. Chauvet et al.[10] devise an index of investor distress in the housing market, the housing distress index (HDI), which is also based on Google search query data. Preis et al.[11] demonstrate how Google Trends data can be used to design a market strategy or define a future orientation index[12].

In principle, different effects between information sources and financial markets are expected when considering news, blogs or even Wikipedia articles[13]. Andersen et al.[14] characterise the response of US, German and British stock, bond and foreign exchange markets to real-time US macroeconomic news. Zhang and Sikena exploit[15] blog and news data to build a sentiment model using large-scale natural language processing. They study how a company's media frequency, sentiment polarity and subjectivity anticipate or reflect stock trading volumes and financial returns. Chen et al.[16] investigate the role of social media in financial markets, focussing on single-ticker articles published on Seeking Alpha, which is a popular social-media platform among investors. Mao et al.[17] compare a range of different online sources of information (Twitter feeds, news headlines and volumes of Google search queries) using sentiment-tracking methods and compare their values for financial prediction of market indices, such as the DJIA, trading volumes, implied market volatility (VIX) and gold prices. Casarin and

Squazzoni[18] compute the Bad News Index as the weighted average of negative sentiment words in the headlines of three distinct news sources.

Recent crisis motivated a number of studies that have focussed on co-movements in financial markets as phenomena that are characteristic of financial crises and that reflect systemic risk in financial systems[19–24]. Harmon et al.[22] demonstrate that the last economic crisis and earlier large single-day panics were preceded by extended periods of high levels of market mimicry, which is direct evidence of uncertainty and nervousness and of the comparatively weak influence of external news. Kennet et al.[23] define an index cohesive force (ICF), which represents the balance between stock correlations and partial correlations after subtracting the index contribution, and demonstrate that financial markets transitioned to a risk-prone state at the end of 2001 that was characterised by high values of ICF.

The idea of cohesiveness as a measure of news importance is simple: if many sources report the same events, then the high number of reports should reflect the event's importance and correlate with the main trends in financial markets. However, to capture the trends of systemic importance, one must be able to track different topics over the majority of relevant online news sources. In other words, one needs (i) access to the relevant news sources and (ii) a comprehensive vocabulary of terms that are relevant to the domain of interest. We satisfy the second prerequisite for a systemic approach through the use of a large vocabulary of financial terms that correspond to companies, financial institutions, financial instruments and financial glossary terms. To satisfy the first prerequisite, in our analysis, we rely on financial news documents that are extracted by a novel text-stream processing pipeline, NewStream (http://newstream.ijs.si/), from a large number of Web sources. These texts are then filtered and transformed into a form that is convenient for computing our cohesiveness measure.

Our News Cohesiveness Index (NCI) captures the average mutual similarity between the documents and entities in the financial corpus. If we represent documents as sets of entities, then there are two alternative views regarding similarity: (i) two documents are more similar than some other two documents if they share more entities, and (ii) two entities are more similar than some other two entities if they co-occur in more documents. We construct the NCI such that the overall similarity in a corpus of documents is equal regardless of the view that we choose to adopt.

There is already strong evidence that links the co-movement of financial instruments to the volatility and uncertainty in financial markets[23], thereby also reflecting the degree of systemic risk. Systemic risk is the risk that is associated with the whole financial system as opposed to any individual entity or component. It can be defined as any set of circumstances that pose a threat to the stability of the financial system and have the potential to initiate a financial crisis[27]. We hypothesise that the cohesiveness of financial news partially reflects this systemic risk.

We analyse the NCI in the context of different financial indices, in terms of their volatility and trading volumes, and Google search query volumes. We demonstrate that the NCI is highly correlated with the volatility of the main US and EU stock market indices, in particular their historical volatility and VIX (the implied volatility of the S&P500).

Results

News cohesiveness index. To measure the herding effects in financial news, we introduce the News Cohesiveness Index, which is an indicator that quantifies the cohesion in a collection of financial documents. A starting point for calculating the NCI is a *document-entity matrix* that quantifies occurrences of entities in each individual document collected over a certain period of time. We use the concept of an entity (instead of e.g., a term) to represent different lexical appearances of some concept in texts. In our case, we use a vocabulary of entities that includes financial glossary terms, financial institutions, companies and financial instruments. The full taxonomy of entities is available in Section 3 of the Supplementary Information. We start with the definition of an occurrence, which determines whether some entity is present in some document, regardless of how many times it occurs in the document. This makes the document-entity matrix A a binary matrix:

$$A_{i,j} = \begin{cases} 1 & \text{if entity } e_j \text{ is in document } d_i \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

A is an $m \times n$ matrix, where m is the number of documents published in the selected time period and n is the total number of entities that we monitor. The document-entity matrix A also corresponds to a biadjacency matrix of a bipartite graph between documents and entities. An edge between document d_i and entity e_j exists if the entity e_j appears in the document d_i.

The overall similarity in the collection of documents should be equal regardless of whether we choose to view it as the similarity either between the *documents* or between the *entities*. To achieve this goal, we define the similarity as the *scalar product* of either document pairs $\langle d_i, d_j \rangle$ or entity pairs $\langle e_i, e_j \rangle$, where the scalar product between vectors $a = [a_1, a_2, \ldots, a_n]$ and $b = [b_1, b_2, \ldots, b_n]$ is defined as $\langle a,b \rangle = \sum_{i=1}^{n} a_i b_i$. Now, we define the NCI as the Frobenius norm of the scalar similarity matrix between all pairs of documents $C^d_{ij} = \langle d_i, d_j \rangle$ or pairs of entities $C^e_{ij} = \langle e_i, e_j \rangle$:

$$NCI = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m} \left\| C^d_{ij} \right\|^2} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} \left\| C^e_{ij} \right\|^2}. \quad (2)$$

The Frobenius norms of both the document-document similarity matrix $C^d = AA^T$ and the entity-entity similarity matrix $C^e = A^T A$ are equal. Therefore, cohesion is conserved whether we measure it as the *document* or *entity* similarity:

$$\left\| C^d \right\|_F = \left\| AA^T \right\|_F = \left\| A^T A \right\|_F = \left\| C^e \right\|_F. \quad (3)$$

In the network representation, these two similarity matrices correspond to two projections of a bipartite graph of the original document-entity matrix, as illustrated in Figure 1. Moreover, one can exploit properties of the Frobenius norm of the scalar similarity matrix and express cohesiveness as a function of the singular values of the document-entity matrix A (a proof of this claim is presented in Section 1 of the Supplementary Information):

$$NCI = \sqrt{\sum_{i=1}^{k} \sigma_i^4}, \quad (4)$$

where σ_i are the k largest singular values of matrix A in a singular value decomposition:

$$A = U \times S \times V^T.$$

The matrices U and V are unitary matrices of the left and right singular vectors of matrix A, and S is a diagonal matrix with singular values σ_i of A. Note that the NCI index is a characteristic property of the corresponding document-entity matrix because it is calculated from its singular values σ_i.

Calculating the NCI through a singular-values approximation can be beneficial for large document-entity matrices because this approach is much more efficient in terms of computational time and memory consumption compared with the explicit calculation of the similarity matrix. We can incrementally calculate only the first k values until we reach the desired accuracy of the NCI (see Section 1 of the Supplementary Information). In practice, only a small number

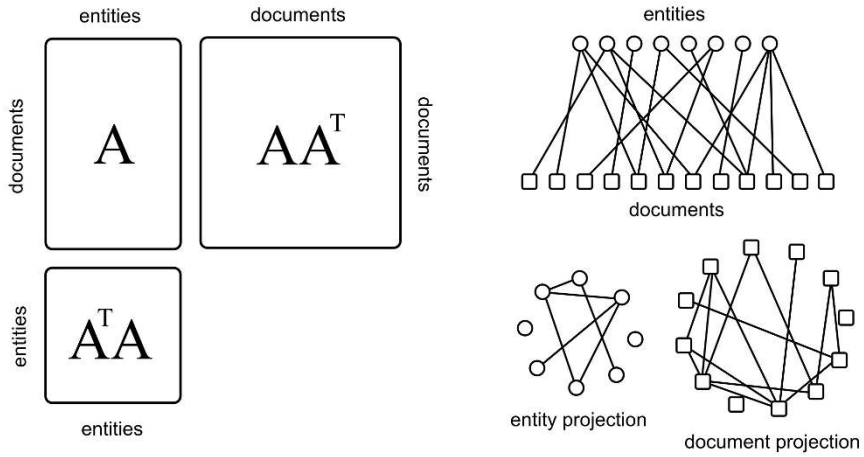

Figure 1 | Matrix and network representations of the document-entity matrix. Matrix representations of the document-document and entity-entity similarity matrices (left) and the corresponding network representations of the entity and document projections (right). The Frobenius norms of the two similarity matrices correspond to the sum of the squares of the connection weights in the two projections. The norms are equal, which indicates that cohesiveness is conserved in both projections.

of singular values is required to calculate the NCI up to the desired precision.

Because the number of documents changes each day, whereas the number of entities stays constant, all NCI indices in our analyses are normalised by dividing them by the number documents in the corpus, *m*. We have statistically confirmed that the NCI is significantly above the level of fluctuations of the cohesiveness random null model (see Section 2 of the Supplementary Information).

Semantic partitions of NCI. It is sometimes interesting to perform a detailed analysis of which groups of entities or documents contribute the most to the overall cohesiveness. For this purpose, we can divide entities or documents into groups using any appropriate semantic criteria and calculate the cohesiveness for each group separately or between pairs of groups. Semantic partitions in the entity projection are created via grouping of entities in mutually disjoint groups, which are defined by their taxonomic labels (hence, this type of partition is referred to as a semantic interpretation). Conversely, semantic partitions in the document projection can be created by grouping documents by their publication date. Figure 2 illustrates the concept of partitioning in the context of different projections.

We can calculate the cohesiveness separately for each semantic group or a combination of semantic groups. Even in this case, we do not need to explicitly calculate similarity matrices (see Section 1 of the Supplementary Information). Following the taxonomy of entities described in Section 3 of the Supplementary Information, we define four semantic groups: companies, regions, financial instruments and Euro crisis terms. We use the notation [company], [region], [instrument] and [eurocrisis] when referring to the cohesiveness of each semantic group and notation in the form [eurocrisis]x[region] when referring to the cohesiveness between two semantic groups. We refer to the cohesiveness calculated within or between any of the groups as *semantic components*. Figure 3 shows the most frequent entities in each of the semantic partitions as determined based on the news corpus collected over the analysed period. The most frequent entities are the ones that define the geographic regions that correspond to the world's leading financial markets: United States, China, Europe, United Kingdom, London, Japan and Germany. We thus concentrate

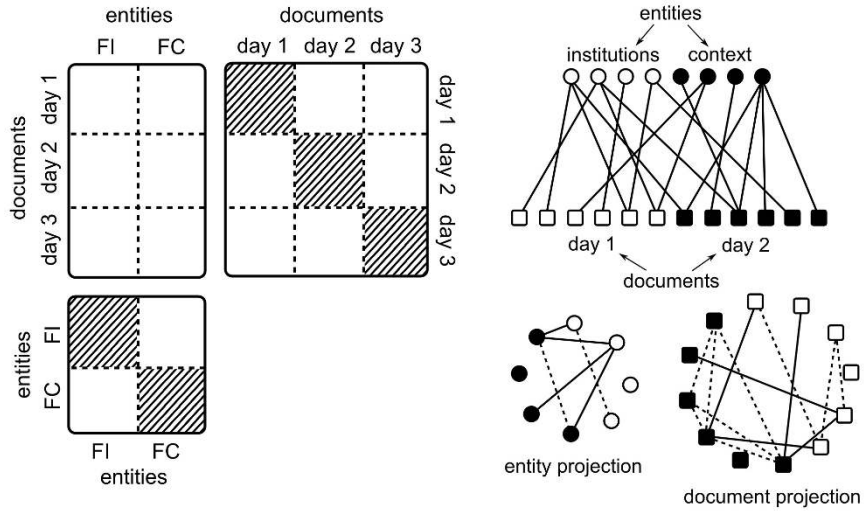

Figure 2 | Semantic partitioning. Semantic partitioning for two entity semantic groups - "Financial Institutions" and "Financial Context" - and three document semantic groups - "day 1", "day 2" and "day 3". The Frobenius norms of the shaded regions quantify the cohesiveness within each semantic group, whereas the Frobenius norms of all other regions quantify the cohesiveness between pairs of semantic groups.

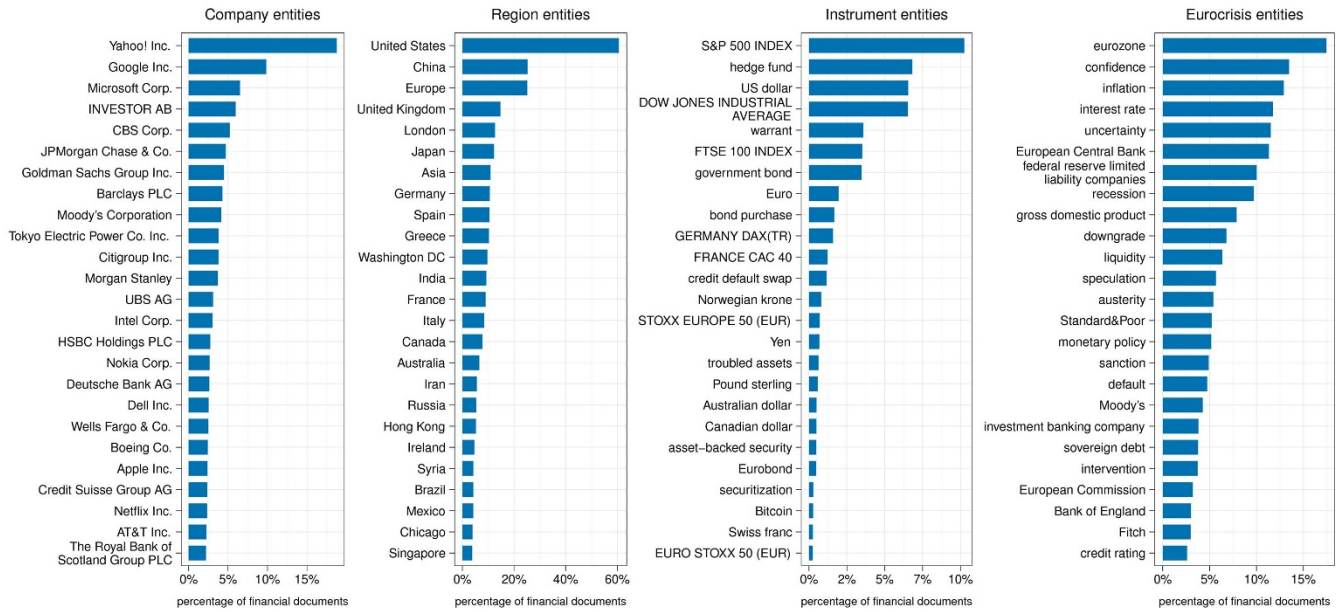

Figure 3 | Occurrences of the 25 most frequent entities in each of the semantic partitions. The most frequent entities are the ones that define the geographic regions that correspond to the world's leading financial markets: United States, China, Europe, United Kingdom, London, Japan, and Germany. We thus concentrate our further analysis on the financial indicators that correspond to the aforementioned markets. Considering the frequency of the term United States, it is no surprise that the majority of other frequent entities, from companies to instruments, are also tied to the US financial market and related terminology.

our further analysis on the financial indicators that correspond to the aforementioned markets.

NCI in relation to financial markets and query volumes. To assess the NCI's utility as a financial market indicator, we use correlation analysis and Granger causality tests against the set of different financial market indicators. The analysis should also provide deeper insight into the interplay between news and trends in financial markets. We adopt the terminology from[9] and treat our news-based indicators (NCI variants and entity occurrence) as indicators of the information supply in online media, whereas volumes of Google search queries are treated as indicators of information demand.

We group the indicators as follows:

- **Information supply indicators:** cohesiveness index based on all the news from NewStream (NCI), cohesiveness index based only on filtered financial news from NewStream (NCI-financial), total entity occurrences based on the aggregate from all news documents and total entity occurrences based on strictly financial documents from NewStream.
- **Information demand indicators:** these are volumes of Google search queries (GSQ) for 4 finance/economy-related categories from Google Finance (Google Domestic Trends – Finance and Investment, Bankruptcy, Financial Planning and Business).
- **Financial market indicators:** these include daily realised volatilities, historical volatilities and trading volumes of major stock market indices (S&P 500, DAX, FTSE, Nikkei 225 and Hang Seng) and the implied volatility of the S&P500 (VIX).

The details of the preparation of individual indicators are given in the Methods section.

We start the analysis with a simple comparison of the NCI calculated using all news and the NCI calculated on filtered financial news. Figure 4 shows the dynamics of NCI and NCI-financial in comparison with VIX (the implied volatility of S&P 500, which is the so-called "fear factor"[25]) and demonstrates that the selection of financial documents is crucial for achieving a high correlation ($R = 0.703$) between the two indices. Selecting financial documents also improves the correlation with other financial indices as shown in Figure 5. For more details regarding the selection of financial documents and how it affects correlations with several other indices, see Section 3 of the Supplementary Information.

Figure 5 shows the Pearson correlation coefficients between different information indicators and financial market indicators. The corresponding p-values are calculated using a permutation test and are available in Section 5 of the Supplementary Information. All correlations reported in this article have p-value $< 10^{-4}$ unless explicitly stated.

In Figure 5, we show that the correlations between (i) financial indices and total entity occurrences and (ii) financial indices and the NCI calculated using all documents are very low around $R < 0.15$. On the other hand, the NCI-financial exhibits much higher correlation with financial indices, with $R > 0.7$ for the implied volatility of the S&P 500 measured by the VIX index. The NCI-financial correlations with financial market volatility indices are much stronger compared to the GSQ categories correlations with volatility measures with $R < 0.3$. In contrast with the NCI-financial, the GSQ categories exhibit stronger correlations with stock market volumes ($0.3 < R < 0.4$).

A more in-depth picture of news cohesiveness is obtained when observing the individual semantic components of NCI-financial and their correlation patterns with financial and Google search query indicators. The semantic components based on the [region] and [eurocrisis] taxonomy categories all have correlation patterns similar to those of NCI-financial (with $R > 0.7$ for [eurocrisis] and $R > 0.5$ for [region]; see Figure 5). This result indicates that these components are most important for the behaviour of NCI-financial. Conversely, semantic components based on [company] and [instrument] exhibit quite different and, in many cases, opposite correlation patterns (with correlations that are close to 0 or even negative). It is interesting to note that both the NCI-financial and GSQ indicators have strong negative correlations with the Nikkei 225 volatility and trading volume (as much as -0.4 for NCI-financial and -0.5 for GSQ-unemployment).

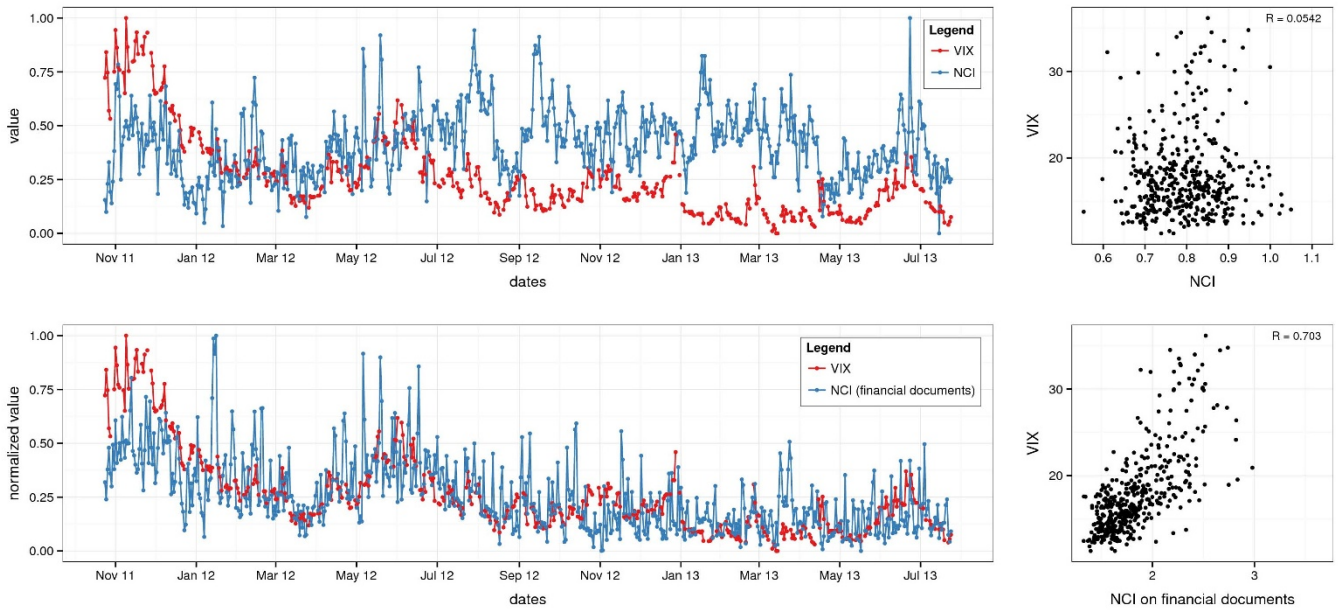

Figure 4 | Comparison of the NCI and VIX time series. NCI, which is calculated using all news (top panel); NCI-financial, which is calculated using strictly financial news (bottom panel); and their correlation with VIX (right panels) are shown. The time series for NCI covers 640 days, from 24th October 2011 to 24th July 2013. The time series for VIX covers 439 working days in the same period. The NCI-financial, obtained by financial document filtering, exhibits much stronger correlation with the VIX compared to the NCI.

We have performed a more detailed analysis of the correlations with several financial indices when using different variants of entity occurrences and NCI-financial that are calculated on subsets of the vocabulary and the document space. For more details, see Section 6 of the Supplementary Information.

In addition to the correlation analysis, we also perform Granger causality tests. The Granger causality test (G-causality test) is frequently used to determine whether a time series $Y(t)$ is useful for forecasting another time series $X(t)$. The idea of the G-causality test is to evaluate whether $X(t)$ can be better predicted using the histories of both $X(t)$ and $Y(t)$ rather than using only the history of $X(t)$ (i.e., $Y(t)$ Granger-causes $X(t)$). The test is performed by regressing $X(t)$ on its own time-lagged values and on those that include $Y(t)$. An F-test is used to determine whether the null hypothesis that $X(t)$ is not Granger-caused by $Y(t)$ can be rejected.

In Figure 6, we show the results of pairwise G-causality tests between information supply and demand indicators and financial indicators. The cells of the table give both the directionality ($X \rightarrow Y$,

Figure 5 | Pearson correlation matrix between all indices. The indices used include the NCI computer using all documents, NCI-financial (calculated using selected financial documents) and its semantic components, entity occurrences, the implied volatility of the S&P 500 (VIX), the realised historical and the daily volatilities of the main stock market indicators (S&P 500, NASDAQ 100, FTSE, DAX, Nikkei and Hang Seng) and Google search query indicators (Business and Industrial, Bankruptcy, Financial Planning, Finance and Investing and Unemployment). The corresponding p-values for all correlations are given in Section 4 of the Supplementary Information.

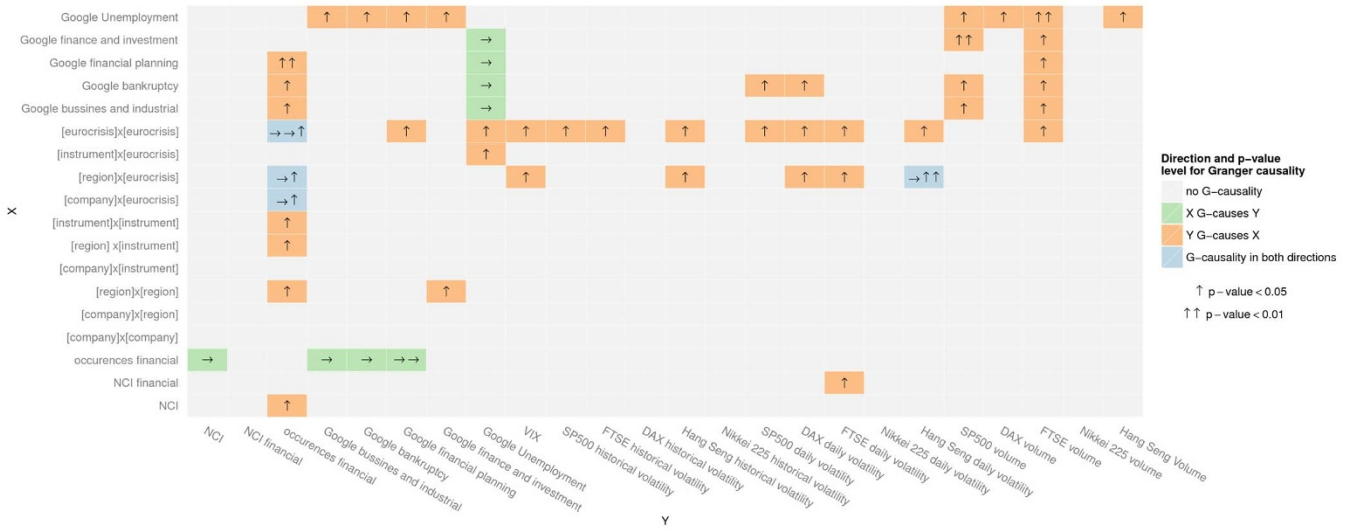

Figure 6 | Granger causality tests. Results of Granger causality tests for the mutual impacts between information and financial indicators. Colours indicate the direction of G-causality ($X \rightarrow Y$ or $Y \rightarrow X$) and bidirectional G-causality ($X \leftrightarrow Y$) at two levels of significance (F-test p-value < 0.01 and p-value < 0.05).

$Y \rightarrow X$ or bidirectional, $X \leftrightarrow Y$) and significance at two levels of the F-test (p-values ≤ 0.01 and ≤ 0.05). From Figure 6, we observe that the Granger causality is almost exclusively directed from the financial indicators to the information indicators, with a single bidirectional exception between the [region]x[eurocrisis] semantic component of the NCI-financial and the Hang Seng daily realised volatility.

Our financial news indicator NCI-financial seems to be G-caused solely by the FTSE daily volatility. However, two of the semantic components, [eurocrisis]x[eurocrisis] and [region]x[eurocrisis], are strongly G-caused by the implied volatility and the historical and daily volatilities of most of the major stock market indices. However, the GSQ categories seem to be mostly G-causality-driven by trading volumes, almost exclusively of the US and UK financial markets (S&P 500 and FTSE).

GSQ indicators seem to be divided into two groups in terms of their Gcausality: (i) those that are G-caused mainly by trading volumes (Business and Industrial, Bankruptcy, Financial Planning and Finance and Investment) and total entity occurrences in the news and (ii) those that are strongly G-caused by all other GSQ categories (Unemployment). The total entity occurrence in the news seems to be the strongest G-causality driver of the GSQ volumes, whereas two of the semantic components of the NCI-financial are G-caused by the GSQ categories of Finance and Investment and Financial Planning.

Discussion

In this work, we introduce a new indicator, based on a concept of cohesiveness in a large collection of news and blogs documents obtained from major Web news sources. In contrast with indicators introduced by other authors, which are often based on sentiment modelling[15,18], the NCI measures the cohesiveness in the news by calculating the average similarity in the financial news.

The analysis of Granger causality tests over a set of financial and information-related indicators suggests that NCI-financial is related to the volatility of the market. In our analysis, the most important semantic components of the NCI-financial are mainly G-caused by the implied volatility (VIX) and historical and daily volatilities. This result implies effects from both short- and long-term risks in the financial market. The only exception (bidirectional causality between [region]x[eurocrisis] and the Hang Seng daily volatility) might be explained as a time-zone effect. This does not seem to be the case for GSQ indicators, which are mainly driven by trading volumes, with the exception of GSQ Unemployment, which seems to be driven primarily by the search volumes of other GSQ categories. Similar to the findings of some previous studies[18,26], in which aggregate sentiment or financial headline occurrence were used as measures of the state of the financial market, NCI-financial seems to be primarily caused by trends in the financial market rather than the opposite. We find that similar results hold for the GSQ categories that quantify the information demand.

The G-causality patterns suggest the presence of circular interplay between information supply and information-demand indicators. For example, total entity occurrence G-causes three of the GSQ categories (Business and Industry, Bankruptcy and Financial Planning), whereas Financial Planning and Unemployment G-cause the semantic components [instrument]x[eurocrisis] and [eurocrisis]x[eurocrisis], which suggests feedback mechanisms between the news and search behaviours.

However, one has to bear in mind that the results of G-causality tests reflect the average of lagged correlations between indicators over the specific time period (in our case, from 24th October 2011 until 24th July 2013). It is also possible that the direction of causality between information and financial indicators changes in time, but such a change was difficult to detect in our data because of the limited length of the time series.

The correlation results confirm the main hypothesis that the cohesiveness of the financial news is a signal that is strongly correlated with the volatilities of the major financial markets. In particular, the NCI-financial correlation with VIX is very important because of VIX's role as a proxy for uncertainty in global market conditions. In situations in which this uncertainty is high, liquidity shocks triggered by some important events can lead to chains of defaults of individual financial institutions and a systemic crisis. The connection between extreme values of implied volatility in times of market turmoil and news regarding important economic and political events has been previously reported[28,30].

Because of the growing complexity and interconnectivity of the global financial system and global economy, it is less likely that we will arrive at a single measure of systemic risk; it is more plausible that we will understand systemic financial risk as a collection of measures[30]. Based on this reasoning and the strong correlation between the NCI-financial and the VIX, we hypothesise that the NCI-financial can be used as a news-borne measure that reflects the degree of systemic risk.

Methods

Data. Access to structured information regarding the financial market with its various instruments and indicators is available for several decades, but the systematic quantification of unstructured information hidden in news from diverse Web sources is of relatively recent origin.

We base our analyses on a newly developed text processing pipeline, New-Stream, which was designed and implemented within the scope of the EU FP7 projects FIRST (http://project-first.eu/) and FOC (http://www.focproject.eu/). NewStream continuously downloads articles from more than 200 worldwide news sources, such as yahoo.com, reuters.com, nytimes.com and bbc.co.uk. It extracts the content, stores complete texts of articles and extracts finance-related entities. It is a domain-independent data acquisition pipeline but is biased towards finance by the selection of news sources and the taxonomy of entities that are relevant to finance.

For the purpose of filtering, efficient storing and analytics, we created an expert-based financial taxonomy and vocabulary of entities and terms that contains the names of relevant financial institutions and companies and finance-and economics-specific terms. The NewStream pipeline has collected approximately 10,000 to 30,000 documents per day since October 2011. In our analyses, we use over 1,400,000 finance-related texts from 24th October 2011 until 24th July 2013. The full structure of the taxonomy, and the list of the domains from which most documents were downloaded are presented in Section 3 of the Supplementary Information.

Filtering of financial documents. Despite the pipeline's bias towards financial news sites, many articles are only indirectly related to finance, such as politics or sports articles. To obtain a clean collection of financial texts, we developed a rule-based model that uses taxonomic categories as features to describe documents. The model was trained on a gold standard of 3500 randomly selected documents that were manually labelled as financial (650 documents), non-financial (1514 documents) or neutral. This model has a recall of over 50% and a precision of well over 80%. It selects approximately several thousand financial documents per day. The rule-based model for filtering financial documents is explained in Section 3 of the Supplementary Information.

Financial indicators. We analyse the NCI in comparison with the financial market indicators of worldwide markets and Google search query volumes. For that purpose, we downloaded the following stock market indices from the Yahoo Finance web service: (http://finance.yahoo.com/): the high, low, open, and close prices and volume of the S&P 500, DAX, FTSE, Nikkei 225 and Hang Seng indices. We also used the implied volatility of the S&P 500 (VIX). The implied volatility is calculated for the next 30 days by the Chicago Board Options Exchange (CBOE, http://www.cboe.com/) using the current prices of indices options. Historical (realised) volatilities are calculated from the past prices of the indices themselves. We use the daily prices of individual indices to calculate a proxy for the daily realised volatility.

$$\text{daily volatility} = \frac{High_t - Low_t}{0.5(Close_t + Close_{t-1})}. \tag{5}$$

Historical (realised) volatilities are calculated as the standard deviations of the daily log returns in the appropriate time window:

$$\text{Historical volatility} = \sqrt{\frac{1}{n}\sum_{t}^{window}\left(\log\left(\frac{p_t}{p_{t-1}}\right)\right)^2}, \tag{6}$$

where p_t are the daily prices and n is the time window. In our analyses, we used a window of 21 working days.

Google search query volumes. Almost all previous studies used search query volumes of specific terms. Instead, we used Google search query volumes of predefined term categories from the Google Finance web site. We chose five categories from Google Domestic Trends that are related to the financial market: Business and Industrial, Bankruptcy, Financial Planning, Finance and Investing and Unemployment. We downloaded YOY (year-over-year) change values for these categories from the Google Finance web service (https://www.google.com/finance).

Granger causality testing. We used functions from the R packages *tseries*, *lmtest*, *vars* and *urca* to calculate indices, construct joint time series dataset, determine correlations and study the Granger causality relations. We followed the methodology of Toda and Yamamoto[29] for Granger causality testing of non-stationary series. Details of the procedure are given in Section 5 of the Supplementary Information.

1. Vespignani, A. Predicting the Behavior of Techno-Social Systems. *Science* **325**, 425–428 (2009).
2. Mitchell, T. M. Mining Our Reality. *Science* **326**, 1644–1645 (2009).
3. Gonzalez, M. C., Hidalgo, C. A. & Barabasi, A.-L. Understanding individual human mobility patterns. *Nature* **453**, 779–782 (2008).
4. Ginsberg, J. *et al.* Detecting influenza epidemics using search engine query data. *Nature* **457**, 1012–1014 (2009).
5. Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M. & Watts, D. J. Predicting consumer behavior with web search. *PNAS* **107**, 17486–17490 (2010).
6. Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A. & Jaimes, A. Correlating financial time series with micro-blogging activity. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* (Seattle, Washington, February 2012), WSDM' **12**, 513–522 (ACM, 2012).
7. Bordino, I. *et al.* Web search queries can predict stock market volumes. *PLoS ONE* **7**, e40014 (2012).
8. Dimpfl, T. & Jank, S. Can internet search queries help to predict stock market volatility? Paper presented at *Finance Meeting EUROFIDAI-AFFI, Paris* (2012). Available at http://ssrn.com/abstract=1941680. Accessed December 21, 2013.
9. Vlastakis, N. & Markellos, R. N. Information demand and stock market volatility. *J. Bank. Financ.* **36**, 1808–1821 (2012).
10. Chauvet, M., Gabriel, S. A. & Lutz, C. Fear and loathing in the housing market: Evidence from search query data (2013). Available at http://ssrn.com/abstract=2148769. Accessed January 14, 2014.
11. Preis, T., Moat, H. & Stanley, E. Quantifying trading behavior in financial markets using Google trends. *Sci. Rep.* **3**, 1684; DOI:10.1038/srep01684 (2013).
12. Preis, T., Moat, H. S., Stanley, H. E. & Bishop, S. R. Quantifying the advantage of looking forward. *Sci. Rep.* **2**, 350; DOI:10.1038/srep00350 (2012).
13. Moat, H. S. *et al.* Quantifying Wikipedia usage patterns before stock market moves. *Sci. Rep.* **3**, 1801; DOI:10.1038/srep01801 (2013).
14. Andersen, T. G., Bollerslev, T., Diebold, F. & Vega, C. Real-time price discovery in stock, bond and foreign exchange markets. *J. Int. Econ.* **73**, 251–277 (2007).
15. Zhang, W. & Skiena, S. Trading strategies to exploit news sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (Washington, D.C., May 23–26, 2010)*, 375–378 (The AAAI Press, 2010).
16. Chen, H., De, P., Hu, Y. J. & Hwang, B.-H. Wisdom of crowds: The value of stock opinions transmitted through social media. *Rev. Financ. Stud.* (2013). Available at http://ssrn.com/abstract=1807265. Accessed January 14, 2014.
17. Mao, H., Counts, S. & Bollen, J. Predicting financial markets: Comparing survey, news, twitter and search engine data. *arXiv:1112.1051* (2011).
18. Casarin, R. & Squazzioni, F. Being on the field when the game is still under way: The financial press and stock markets in times of crisis. *PLoS ONE* **8**, e67721 (2013).
19. Battiston, S., Puliga, M., Kaushik, R., Tasca, P. & Caldarelli, G. DebtRank: Too Central to Fail? Financial Networks, the FED and Systemic Risk. *Sci. Rep.* **2**, 541; DOI:10.1038/srep00541 (2012).
20. Huang, X., Vodenska, I., Havlin, S. & Stanley, H. E. Cascading Failures in Bi-partite Graphs: Model for Systemic Risk Propagation. *Sci. Rep.* **3**, 1219; DOI:10.1038/srep01219 (2013).
21. Quax, R., Kandhai, D. & Sloot, P. M. A. Information dissipation as an early-warning signal for the Lehman Brothers collapse in financial time series. *Sci. Rep.* **3**, 1898; DOI:10.1038/srep01898 (2013).
22. Harmon, D. *et al.* Predicting economic market crises using measures of collective panic. (2011). Available at http://ssrn.com/fabstract=1829224. Accessed January 14, 2014.
23. Kenett, D. Y. *et al.* Index cohesive force analysis reveals that the US market became prone to systemic collapses since 2002. *PLoS ONE* **6**, e19378 (2011).
24. Zheng, Z., Podobnik, B., Feng, L. & Baowen, L. Change is cross-correlations as an indicator for systemic risk. *Sci. Rep.* **2**, 888; DOI:10.1038/srep00888 (2012).
25. Vodenska, I. & Chambers, W. J. Understanding the relationship between VIX and the S&P 500 index volatility. Paper presented at *26th Australasian Finance and Banking Conference 2013* (Sydney, Australia, December 17–19 2013). Available at http://ssrn.com/abstract=2311964. Accessed January 18, 2014.
26. Da, Z., Engelbergz, J. & Gaox, P. The sum of all fears - investor sentiment and asset prices (2011). Available at http://rady.ucsd.edu/faculty/directory/engelberg/pub/portfolios/FEARS.pdf. Accessed January 14, 2014.
27. Kaufman, G. Banking and currency crisis and systemic risk: A taxonomy and review. *Finan. Markets, Inst. Instruments* **9**, 69 (2000).
28. Neely, C. Using implied volatility to measure uncertainty about interest rates. *Fed. Reserve Bank St. Louis Rev.* **87**, 407–425 (2005).
29. Toda, H. Y. & Yamamoto, T. Statistical inference in vector autoregression with possibly integrated processes. *J. Econometrics* **66**, 225–250 (1995).
30. Lo, A. Hedge funds, systemic risk, and the financial crisis of 2007–2008: Written testimony to the U.S. House of Representatives Committee on Oversight and Government Reform. (2008). Available at http://ssrn.com/abstract=1301217. Accessed January 13, 2014.

Acknowledgments

This work was supported in part by the European commission as part of the FP7 projects FOC (Forecasting Financial Crises, Measurements, Models and Predictions, grant no. 255987, and FOC INCO, grant no. 297149), the EU-FET project MULTIPLEX (Foundational Research on MULTIlevel comPLEX networks and systems, grant no. 317532) and by the Croatian Ministry of Science, Education and Sport project "Machine Learning Algorithms and Applications". We would like to thank the following people for helpful discussions: Stefano Battiston, Vinko Zlatić, Guido Caldarelli, Michelangelo Puliga, Tomislav Lipić and Matej Mihelčić.

Author contributions

All authors contributed to the writing and editing of the manuscript. M.P., N.A.F. and T.S. performed the modelling and analyses. P.K.N., I.M. and M.G. were involved in gathering and processing of the data. I.V. and T.S. were involved in interpreting the results.

Additional information

Dataset availability: All data and codes that we used in our analysis are freely available from http://lis.irb.hr/foc/data/data.html.

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Piškorec, M. *et al.* Cohesiveness in Financial News and its Relation to Market Volatility. *Sci. Rep.* **4**, 5038; DOI:10.1038/srep05038 (2014).