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], and 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

April 11, 2014

Contents

1 Cohesiveness through SVD approximation 2

2 Statistical significance of cohesiveness 5

3 Filtering of financial documents 7

4 Statistical significance (p-values) of Pearson correlation coefficients between indices 12

5 Steps from Toda-Yamamoto procedure used for Granger causality testing 13

6 Document-entity occurrences and cohesiveness - dependence on the choice of vocabulary and size of document corpus 14

1 Cohesiveness through SVD approximation

We start from the definition of the NCI index as Frobenious norm of similarity matrix $C = AA^T$:

$$\parallel C \parallel_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{m} \|C_{ij}\|^2} = \sqrt{tr(C^T C)} \tag{1}$$

,where tr denotes the trace of the matrix. Since, the matrix $C = AA^T$

$$\Longrightarrow tr(C^T C) = tr(AA^T AA^T)$$

By making a singular value decomposition $A = U \times S \times V^T$ we get:

$$tr(AA^T AA^T) = tr(USV^T VS^T U^T USV^T VS^T U^T) = tr(US^4 U^T) = tr(S^4).$$

which proves the equality. If we use only first k singular values then we get best low rank approximation of similarity matrix C by Eckart–Young theorem.

$$NCI = \sqrt{tr(S^4)} \approx \sqrt{\sum_{i=1}^{k} \sigma_i^4} \tag{2}$$

It is very important to note the run-time and memory improvement (see Figure 1) by using the singular approach on large matrices. In order to calculate the first k singular values, one can use the iterative Lanczos algorithm [4]. As the first singular values contain the most of the energy the approximation of NCI can be done with just a few values (see Figure 2).

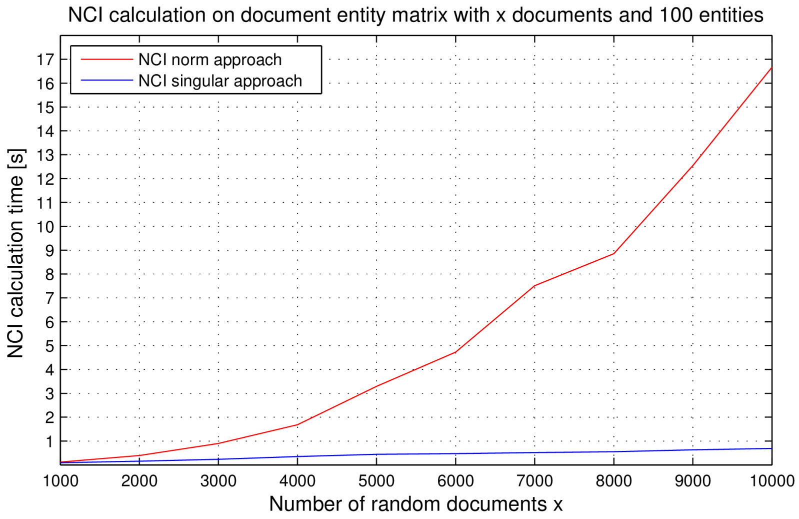

Figure 1: Comparison of run-time calculation of NCI index on matrix of 100 entities and different number of random documents. Run-time measurements were done in Matlab by using built-in optimized functions for matrix multiplication on Intel Core i7-3770 3.4 GHZ processor with 16 GB of RAM memory, finding matrix norm and sparse singular value decomposition with $k = 5$ values. The relative error of singular approach with $k = 5$ w.r.t norm approach was always below 1 percent of relative error. Memory consumption for case with 10k documents was around 1 GB for norm approach and around 100 MB for singular approach.

Table 1: **The computational time for calculating NCI index and occurrence volume**

	SVD (k=1)	SVD (k=5)	SVD (k=10)	Occurrence
Time[s] (financial)	7.2	23.6	40.2	0.06
Time[s] (all)	26.3	90.6	174.0	0.13

Total computational time for calculating daily NCI index (with first k singular) for 640 days from 24^{th} of October 2011 until 24^{th} of July 2013. There are total of 10 613 650 documents out of which 1 438 572 are financial. The measurements were done in Matlab on Intel Core i7-3770 3.4 GHZ processor with 16 GB of RAM memory.

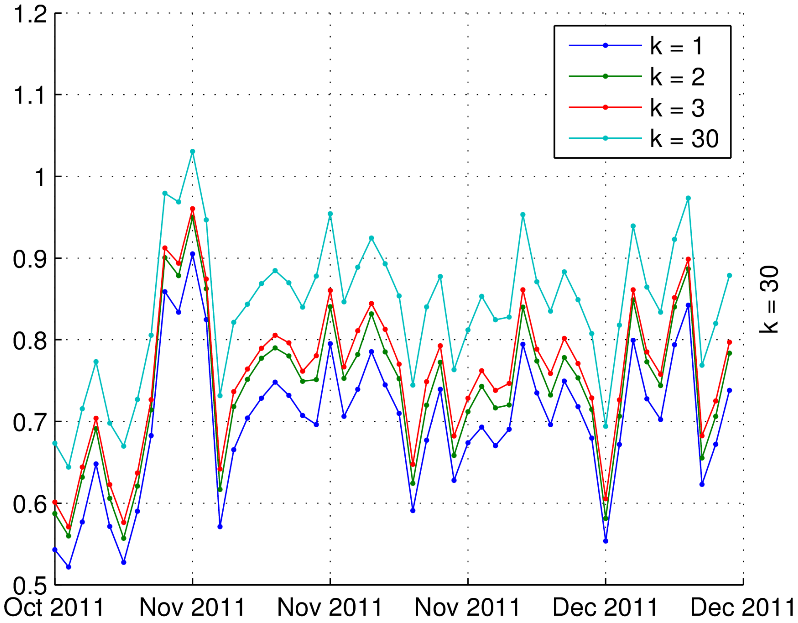

Figure 2: Comparison of accuracy of approximation of NCI index with k singular values.

Sometimes it is necessary to perform detailed analysis of which 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 cohesiveness for each group separately or between each pair of groups. Note, that even in this case we do not need to explicitly calculate similarity matrices but still use the singular values technique. Cohesiveness between any pair of the groups is calculated by exploiting the properties of Frobenius norm. If we have two semantic groups G_1 and G_2 and if the cohesiveness for each of them is $\parallel G_{11} \parallel_F$ and $\parallel G_{22} \parallel_F$ respectively, while for their combination is $\parallel G_{1212} \parallel_F$, then the cohesiveness between the two groups $\parallel G_{12} \parallel_F$ is:

$$\parallel G_{12} \parallel_F = \sqrt{\parallel G_{1212} \parallel_F^2 - \parallel G_{11} \parallel_F^2 - \parallel G_{22} \parallel_F^2}. \tag{3}$$

2 Statistical significance of cohesiveness

In this section we will quantify the statistical significance of NCI with respect to a cohesiveness null model of in our system. Let us first recall the definition of normalized NCI of document entity matrix A of size $m \times n$:

$$\frac{1}{m}NCI = \frac{1}{m} \parallel AA^T \parallel_F = \frac{1}{m} \parallel C \parallel_F = \frac{1}{m}\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m} \|C_{ij}\|^2},$$

Now, we will start with rather simple question. What is the probability that m documents have high NCI^* index just by "chance" in a system with n entities ? In the special case, when each of m independent random documents $\vec{x_i}$ contains exactly one entity from the vocabulary of size n which are equally likely the expected NCI^* index has the following upper bound:

$$\frac{1}{m}E[NCI^*] \leq \sqrt{\frac{1}{m} + \frac{1}{n}}.$$

Only for small values of m and n the expectation could be high and as the corpus of documents and entities is larger this component vanish.

Now, we turn our attention to the statistical measurement of bias in our methodology. The process of constructing the financial entity vocabulary, gathering documents from the world wide web and filtering financial documents adds a statistical measurement bias to the cohesiveness signal, which we call the background cohesiveness signal. From all the news documents in the world wide web Ω we only consider a biased portion Ω_F of financial ones according to our financial filter. Therefore, the measurement of cohesiveness on the corpus of m documents within one day contain the cohesiveness of news of the particular day plus the superposition of the background cohesiveness. In order to measure the level of background cohesiveness we calculate the cohesiveness on a corpus of m independently sampled documents from a large period of time T. Having a large period of time (years), we sample m documents from a large sample of real documents that we have processed with our methodology pipeline. If the NCI over m documents within one day is significantly higher than the fluctuations of cohesiveness of m random bootstrapped documents from a large period of time T it suggests that the news on that particular day bring some extra correlations. Contrary, for those days when the NCI is not higher than the fluctuations of random bootstrapped documents, it means that the cohesiveness could be the results of our measurement bias. In Figure 3, we plot the NCI for particular day along with the estimated mean and standard deviation of background cohesiveness from bootstrapped documents from the end of 2011 until the end of 2013. We observe that the level of NCI is higher than the mean plus standard deviation of background cohesiveness in 80% of observed days. The mean number of documents per day in this period was ≈ 2000 and the total number of documents is approximately 1.4 million.

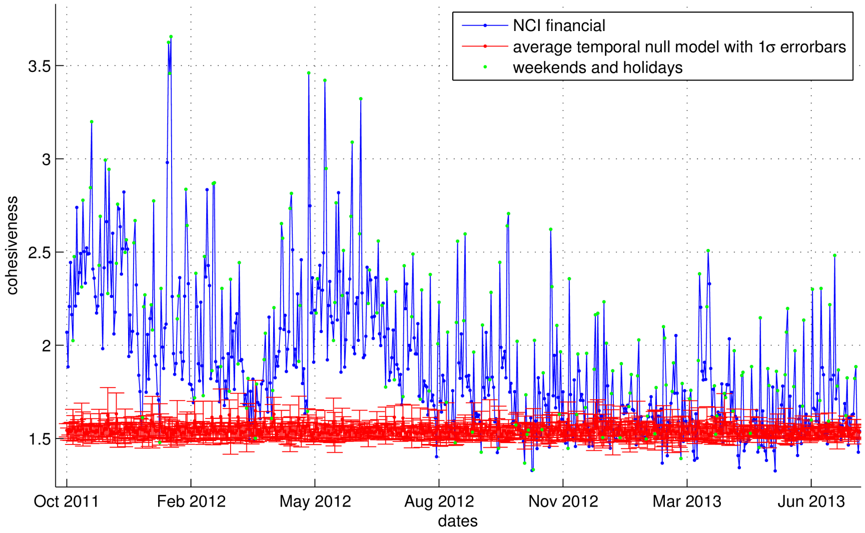

Figure 3: NCI index (blue) and expected NCI of background cohesiveness (red) model with 1 standard deviation.

3 Filtering of financial documents

Taxonomy class	Count	Class id
Object		53.
| Company	103	11.
| | Fund	114	2.
| | Industry	533	5.
| | Bank	85	6.
| | Insurance	37	9.
| | Other	152	10.
| Geographical region	32	14.
| | City	37	13.
| | | Capital city	243	7.
| | Country	239	17.
| | | PIIGS Country	5	29.
| | | BRICS Country	5	34.
| | Continent	7	27.
| Organization	29	20.
| Financial Instrument	17	12.
| | Over the counter	1572	24.
| | Stock	977	36.
| | Currency	15	52.
| | Digital Currency	5	26.
| | | Digital Currency Market	4	49.
| | Index		15.
| | | Stock Index	18	51.
| Eurocrisis		16.
| | Protagonist	28	4.
| | Finance		8.
| | | Financial Markets		45.
| | | | Financial markets term	9	21.
| | | | Neg Financial Sentiment	2	31.
| | | | Financial Crisis Term	5	40.
| | | | Pos Financial Sentiment	4	42.
| | | Financial Institution		46.
| | | | EU Mechanism	11	23.
| | | | EU Financial Institution	8	25.
| | | | Rating Agency	4	32.
| | | | Other Financial Institution	3	33.
| | | | Int Financial Institution		37.
| | | | US Financial Institution	3	38.
| | | | Central Bank	3	39.
| | | Public Finance		47.
| | | | Monetary Policy Term	8	30.
| | | | Fiscal Policy		48.
| | | | | Workforce term	9	18.
| | | | | Public Spending		44.
| | | | | | Loan Risk Term	5	19.
| | | | | | Budget Term	17	22.
| | | | | | Loan Term	6	28.
| | | | | | Financial Failure Term	6	35.
| | | | | | Public Debt Term	4	41.
| | | | | | Loan Insurance Term	3	43.

Table 2: Structure of the taxonomy of entities/terms with number of entities per class and class id that we use in this work. Underlined categories define our semantic partitions.

Table 2 shows the taxonomy of entities/terms with the corresponding class id number and the number of entities in each group. This taxonomy was manually defined within the scope of EU FP7 project FIRST (`http://project-first.eu/`) with the entities related mainly to the Euro crisis. The rest of the taxonomy was added within the scope of EU FP7 project FOC (`http://www.focproject.eu/`) and it expands the financial vocabulary in order to cover terminology related to the U.S. economy and its financial crisis.

In order to filter the financial documents we have made a gold standard of approximately 3500 randomly selected documents in the period of interest (from 24th of October 2011 until 24th of July 2013) manually labeled as financial (650 documents), non-financial (1514 documents) and neutral. Out of 2164 financial and non-financial documents we use 50% from each group for learning the model and 50% from each group for testing. We built a machine learning model that classifies documents as financial or non-financial depending on the number of entities from each taxonomy class they contain. Document is classified as financial if *any* of the following rules is satisfied:

1. `(49.>=2 AND 18.<=0 AND 17.<=3 AND 10.>=1 AND 3.<=3)`
2. `(49.>=1 AND 16.>=1 AND 11.>=1 AND 44.>=1 AND 49.>=4)`
3. `(51.>=1 AND 18.=0 AND 14.<=4 AND 11.>=1 AND 45.=0 AND 53.>=6)`
4. `(49.>=1 AND 18.=0 AND 51.>=2)`
5. `(51.>=1 AND 13.=0 AND 47.=0 AND 11.>=1 AND 16.>=1)`

The rules were obtained by using the Ripple-Down Rule learner [1] within the machine learning framework Weka [2] (`http://www.cs.waikato.ac.nz/ml/weka/`). Each rule is a logical conjunction over entity class numbers and their occurrence frequency within a document. For example, in the last line of our financial filter we have the following entity classes: 51. (stock index), 13. (city), 47. (public finance), 11. (company), 16. (eurocrisis). This model achieves a recall of over 50%, with precision of well over 80% on a test set. Figure 4 demonstrates the effect of filtering on entity and document histograms.

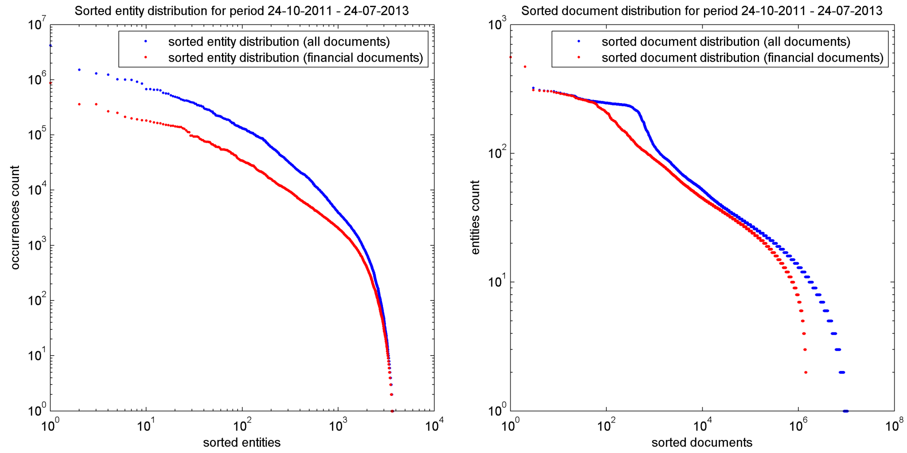

Figure 4: Document and entity histograms in original corpus (NewStream) and filtered (strictly financial corpus) from the October 2011 until July 2013.

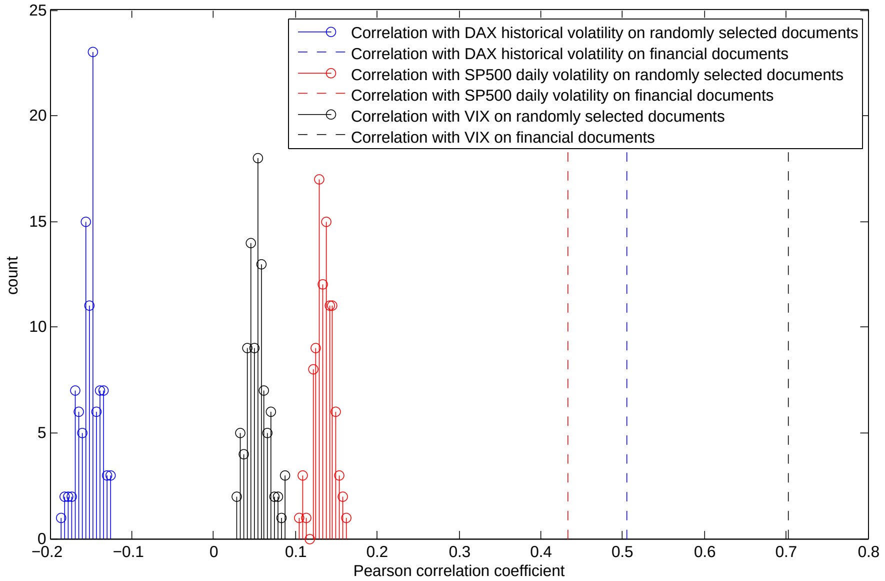

Figure 5: Correlations of three financial indices (DAX historical volatility, S&P 500 daily volatility and VIX) with NCI calculated on documents randomly selected from the original corpus and on filtered financial documents from the October 2011 until July 2013. Number of random documents for each day is equal to the number of filtered financial documents for that day.

Figure 5 shows correlations of three financial indices (DAX historical volatility, S&P 500 daily volatility and VIX) with NCI calculated on documents randomly selected from the original corpus and on filtered financial documents from the October 2011 until July 2013. It demonstrates that selection of financial documents is indeed crucial for obtaining high correlation with financial indices, as compared to equal number of randomly selected documents from the original corpus.

In the Figure 6, we show the most relevant domains and the corresponding number of downloaded news in our NewStream pipeline.

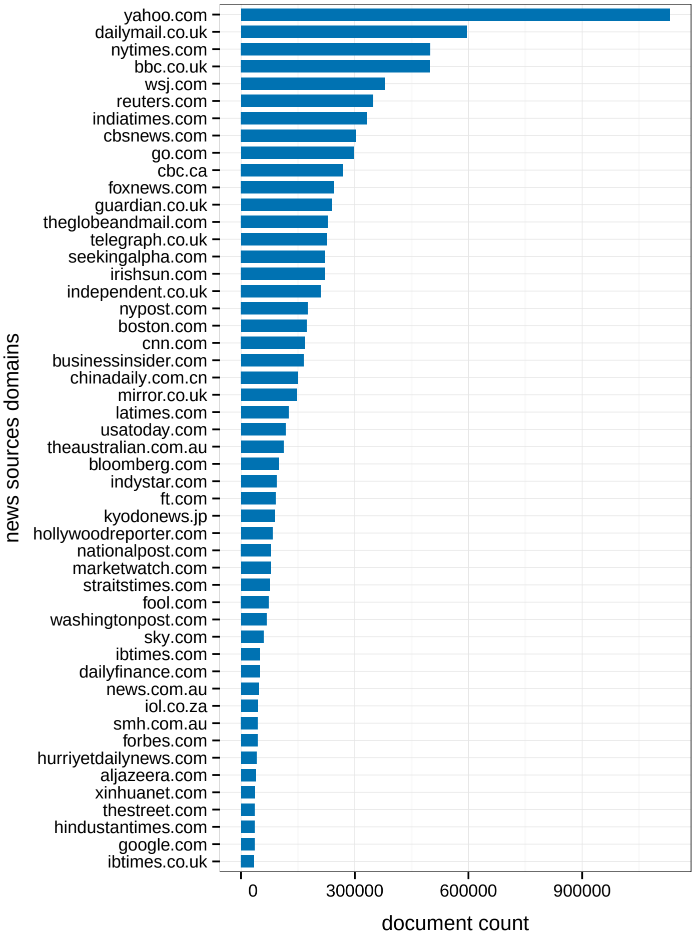

Figure 6: Domains from which most documents were downloaded by the NewStream pipeline from the October 2011 until July 2013.

4 Statistical significance (p-values) of Pearson correlation coefficients between indices

We done a series of independent permutations (10^4) on our indices and estimate the Pearson correlation on permuted data. Then, we estimate the p-value as the proportion of correlations on permuted data that are larger than the correlation on original data.

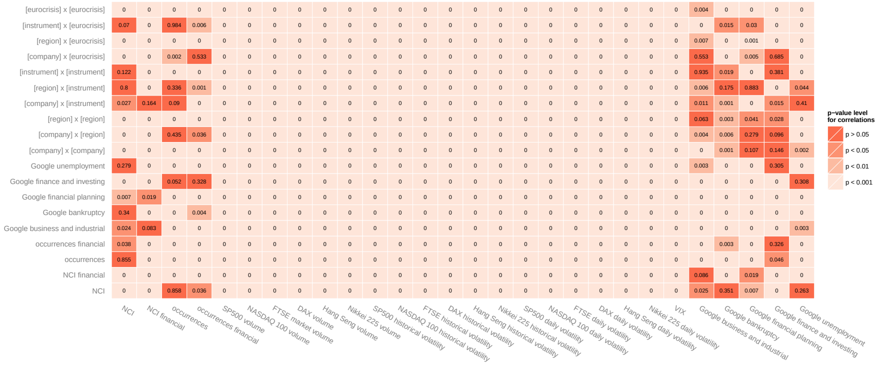

Figure 7: **p-values for Pearson correlations between indices.** p-values are calculated with permutation testing.

5 Steps from Toda-Yamamoto procedure used for Granger causality testing

We based our Granger causality testing on the Toda-Yamamoto procedure [3]:

1. We first test each of the time-series to determine their order of integration using the ADF test (Augmented Dickey Fuller) and KPSS test (Kwiatkowski–Phillips–Schmidt–Shin) to have a cross-check. We determined that for all of them maximum order of integration is 1.

2. We set up vector autoregressive (VAR) models using time series as obtained (no differencing - i.e. in the levels of the data).

3. We then determined the maximum lag length for the variables in the VAR model - p, using `VARselectfunction` and choose the p based on the average minimum error lag as determined by 4 information criteria (AIC, SC, HQ, FPE). Maximum lag order determined in this fashion for the whole dataset wasused for setting of all VAR models $p = 7$.

4. We check whether the VAR models are correct wrtserial correlation in the residuals using Portmanteau test (`serial.test` function).

5. We use same lag order for VAR modelswith addition of 1 lags of each of the variables in each of the equations.

6. We test for Granger non-causality using all pairs of time series (X and Y). We test the hypothesis that the coefficients of the first p lagged values of X are zero in the Yequation, using Wald test (`wald.test` function). Then we do the same for the coefficients of the lagged values of Y in the equation. Rejection of the null hypotheses (coefficients are zero) implies a rejection of Granger non-causality, i.e. a rejection supports the presence of Granger causality.

Note: We skipped tests for cointegration of time series as they are required only as a cross-check for validity of results, and are not needed for performing the GC tests.

6 Document-entity occurrences and cohesiveness - dependence on the choice of vocabulary and size of document corpus

In this section we provide a more detailed picture of cohesiveness with respect to the choice of vocabulary and document corpus. We base our analysis on the correlations with implied (VIX) and realized volatility of S&P 500 index. The analysis bellow shows Pearson correlations over the whole analysed period, however different aggregations over document-entity matrices (defined below) and NCI are based on entity occurrences and document corpora on daily basis. First, we define different aggregations of the document-entity occurrence matrix A $(m \times n)$:

1. Total entity occurrence over all documents: $\sum_j \sum_i A_{ij}$,
2. Normalized total entity occurrence: $\frac{1}{m} \sum_j \sum_i A_{ij}$,
3. Normalized cohesiveness diagonal in entity projection: $\frac{1}{m}\sqrt{\sum_i (A^T A)^2_{ii}}$, where the term $(A^T A)_{ii}$ corresponds to the number of documents in which entity i occurs,
4. Normalized cohesiveness diagonal in document projection: $\frac{1}{m}\sqrt{\sum_j (AA^T)^2_{jj}}$, where the term $(AA^T)_{jj}$ corresponds to the number of entities that occurred in document j,
5. NCI: Normalized cohesiveness: $\frac{1}{m}\sqrt{\sum_i \sum_j (AA^T)^2_{ij}} = \frac{1}{m}\sqrt{\sum_i \sum_j (A^T A)^2_{ij}}$.

Aggregations 2, 3 and 4 can be understood as approximations of cohesiveness. Figure 8 illustrates behaviour of these aggregations for different fractions of entities from the vocabulary (plots: A and B) and different fraction of documents from the corpus (plots C and D). In plots A and B, we start from having only the most frequent entities and add those less frequent until we end with the complete vocabulary. We aggregate over all documents. In plots C and D, we start from having only the documents with highest number of entities and gradually add those that have smaller number of entities from the vocabulary. We aggregate taking into account whole vocabulary of entities.

If we focus first on plots A and B, we can observe that NCI and its close approximation 3 (a diagonal part of cohesiveness in entity $A^T A$ projection) give very stable and high correlations over a very broad choice of entities from the vocabulary. Note, also that the approximation 4 (a diagonal part of cohesiveness in document AA^T projection) gives very poor results. Contrary, the NCI measures has the same stable performance both in entity $(A^T A)$ and document (AA^T) projection. On the other hand total entity occurrence exhibits very low correlation with volatility and has its maximum for very low number of most frequent entities. Its normalized version seems to be a very good approximation of cohesiveness, but its behaviour is again dependent on the choice of entities.

One has to bear in mind that frequent entities are determined on a daily basis - i.e. the most frequent entities in principle change from day to day, which means that we cannot observe this kind of behaviour with small vocabulary. In other words, although only small number of entities is responsible for overall cohesiveness, we need larger vocabulary in order to capture concept drift in the news.

If we now observe plots C and D in Figure 8, we see that relatively large fraction of documents is needed to obtain rather stable and high correlations of NCI with volatility. In these plots aggregation 1 and 2 attain much lower correlations than in plots above, which is the consequence of aggregation over all entities from the vocabulary.

This analysis also emphasizes specific properties of news corpora: i) that most of the cohesiveness signal is based on a small fraction of top most frequent entities for the particular day and ii) that one needs rather large fraction of documents to get stable (high) correlation with volatility. This analysis supports further the hypothesis that the cohesiveness is a more robust measure of news importance than entity volume, and is thus more appropriate measure for systemic risk reflected in financial news.

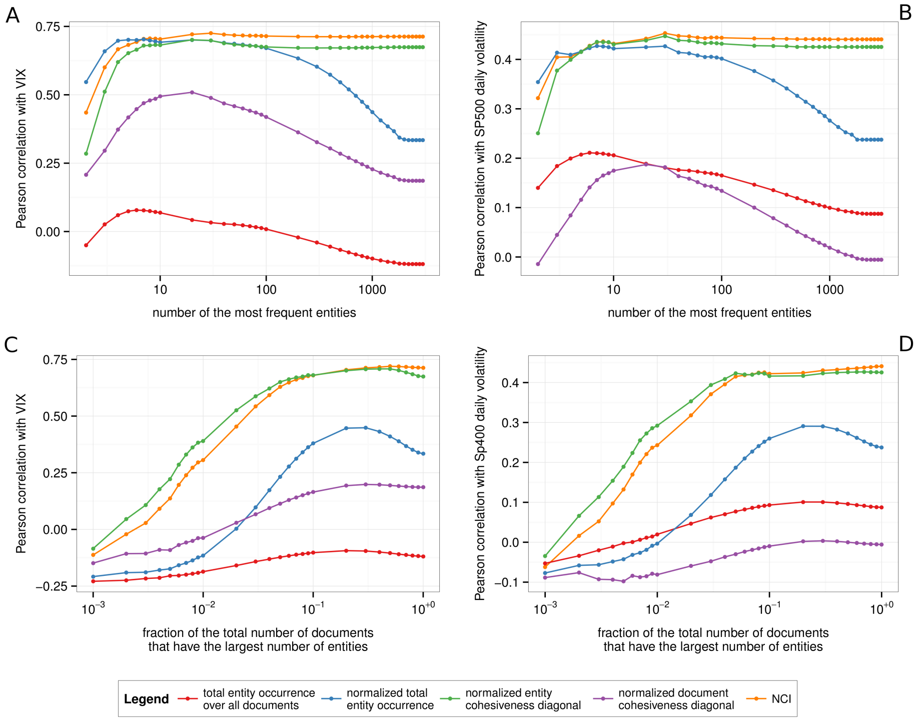

Figure 8: Pearson correlations of total entity occurrence over all documents (1), normalized total entity occurrence (2), normalized diagonal of cohesiveness in entity projection (3), normalized diagonal of cohesiveness in document projection (4) and NCI measure (5) with VIX and daily realized volatility. Plots A and B: Influence of the choice of entities, using the whole document corpus (per day). Rather high correlation of NCI and its close approximation (3) with implied and realized daily volatility is obtained for very low number of most frequent entities for a particular day. Total entity occurrence (1) and its normalized variant (2) are not stable with respect to the choice of vocabulary. Plots C and D: Influence of the choice of the documents, while using whole vocabulary. We observe that relatively large fraction of documents is needed in order to get high correlation of NCI and its approximation (3) with volatilities. Total entity occurrence (1) and its normalized variant (2) do not achieve are not stable with respect to the choice of vocabulary.

References

[1] Brian R. Gaines and Paul Compton. Induction of ripple-down rules applied to modeling large databases. *Journal of Intelligent Information Systems*, 5:211–228, 1995.

[2] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11, 2009.

[3] Hiro Y. Toda and Taku Yamamoto. Statistical inference in vector autoregression with possibly integrated processes. *Journal of Econometrics*, 66:225–250, 1995.

[4] Kesheng Wu and Horst Simon. Thick-restart lanczos method for symmetric eigenvalue problems, 1998.