Quantifying the Impact of Cohesiveness in Financial News

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

[1] Ruđer Bošković Institute, Zagreb, Croatia
[2] Jožef Stefan Institute, Ljubljana, Slovenia
[3] Metropolitan College, Boston University, USA

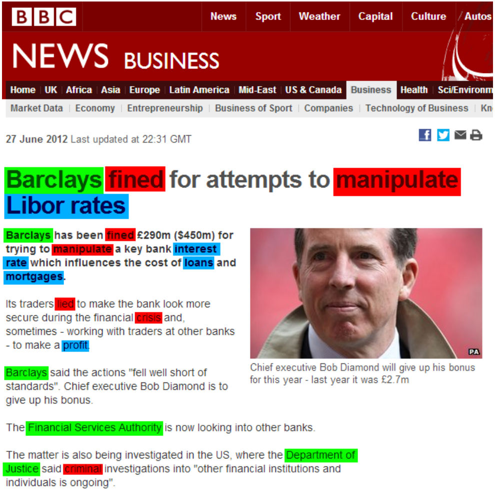
September 18, 2013

fcc
forecasting financial crises

Part I

Cohesiveness in a corpus of documents

Financial document - an example



27 June 2012 Last updated at 22:31 GMT

Barclays fined for attempts to manipulate Libor rates

Barclays has been fined £290m ($450m) for trying to manipulate a key bank interest rate which influences the cost of loans and mortgages.

Its traders lied to make the bank look more secure during the financial crisis and, sometimes - working with traders at other banks - to make a profit.

Barclays said the actions "fell well short of standards". Chief executive Bob Diamond is to give up his bonus.

Chief executive Bob Diamond will give up his bonus for this year - last year it was £2.7m

The Financial Services Authority is now looking into other banks.

The matter is also being investigated in the US, where the Department of Justice said criminal investigations into "other financial institutions and individuals is ongoing".

Figure: Entities in a news article: institutions (green), financial glossary terms (blue) and negative sentiment words (red).

Cohesiveness in a corpus of documents

Intuition: Large cohesion in a collection of financial news documents indicates a form of *herding effect* that either reflects on important event in the financial markets or can potentially elicit a response on financial market behavior.

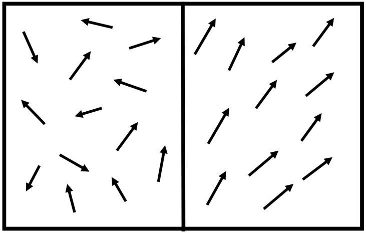
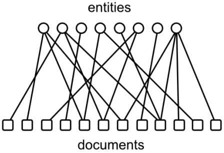
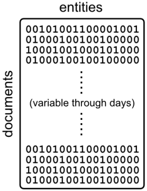


Figure: Financial documents on the Web represented as vectors of entities. “Normal” state on the left and “cohesive” state on the right.

Document-entity matrix

Document-entity matrix A - biadjacency matrix of a **bipartite-graph** between documents and entities





```
00101001100001001
01000100100100000
10001001000101000
01000100100100000
        ⋮
(variable through days)
        ⋮
00101001100001001
01000100100100000
10001001000101000
01000100100100000
```

A is a boolean matrix and it records whether each entity is present or not in the document. Its size is $m \times n$ where m is number of documents and n is number of entities:

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

Document-entity matrix

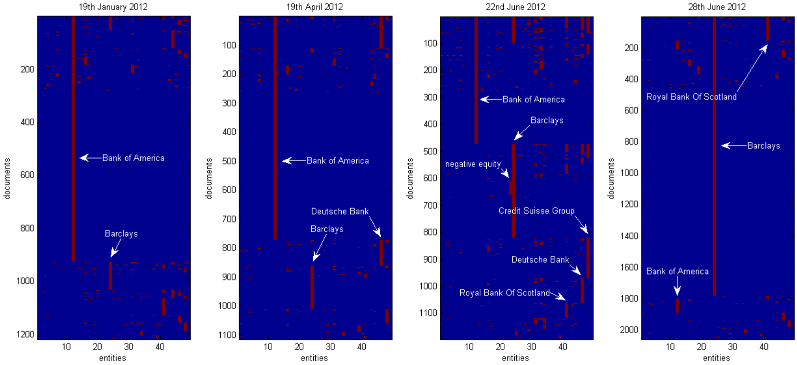
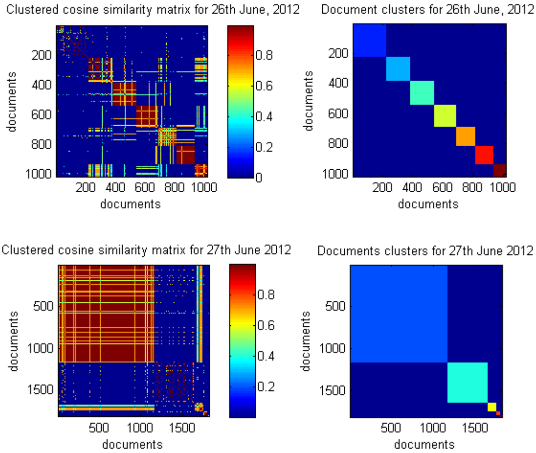


Figure: Document-entity matrix from four distinct days in 2012. Vocabulary consists of 13 banks listed on NYSE and 36 financial glossary terms.

Document-document similarity matrix



Clustered cosine similarity matrix for 26th June, 2012

Document clusters for 26th June, 2012

Clustered cosine similarity matrix for 27th June 2012

Documents clusters for 27th June 2012

Figure: Document-document similarity matrix for two distinct days in 2012. Vocabulary consists of 13 banks listed on NYSE and 36 financial glossary terms.

News cohesiveness index (NCI)

We define *News cohesiveness index* (NCI) in two equivalent ways...

Definition through Frobenius norm

Frobenius norm on the scalar similarity matrix $\| C \|_F = \| AA^T \|_F$ between all documents:

$$NCI = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{m} \|C_{ij}\|^2} \tag{2}$$

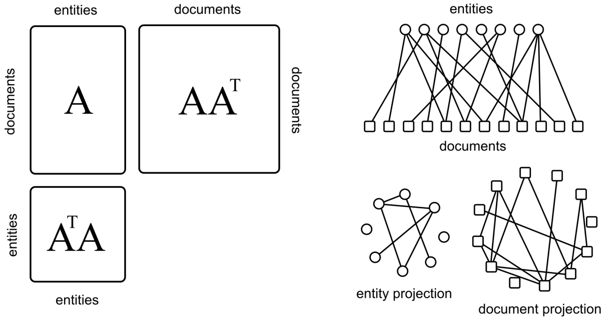
Definition through singular values

Function of singular values σ_i of matrix A:

$$NCI = \sqrt{\sum_{i=1}^{k} \sigma_i^4} \tag{3}$$

NCI - properties of scalar similarity

If we measure the similarity between document $\vec{x}_1$ and $\vec{x}_2$ as a scalar product $\langle \vec{x}_1, \vec{x}_2 \rangle$ then bipartite projection to documents AA^T and projection to entities $A^T A$ have the same NCI: $\| AA^T \|_F = \| A^T A \|_F$.



NCI measure captures the **intrinsic property of the document-entity bipartite graph** that is invariant to projection!

Part II

Results

Document collection pipeline

We use a document collection pipeline developed by our collaborators on Jozef Stefan intitute in Ljubljana as a part of FIRST[1] and FOC[2] projects.

Corpus used in this work: 80k financial documents from major news sources during 250 working days of 2012.

Type of entity	class_id	**Count**
Positive vocabulary terms	1	around 2000 entities
Negative vocabulary terms	2	around 2000 entities
Financial glossary terms	3	36 entities
Banks	4	56 entities
Funds	5	132 entities
Insurance	6	19 entities
Countries	7	36 entities

[1] http://project-first.eu/

[2] http://www.focproject.eu/

NCI and VIX

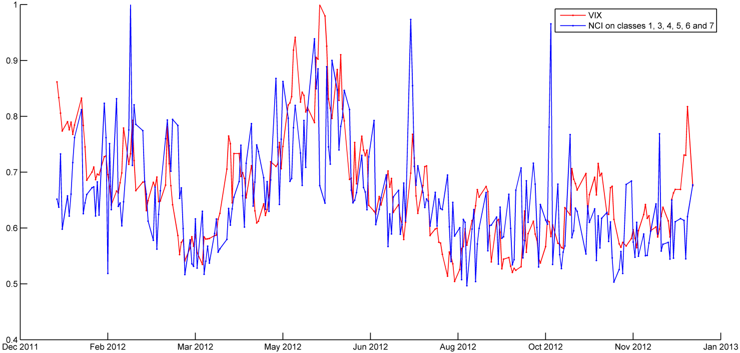


Figure: Normalized daily values of News Cohesiveness Index and implied volatility of S&P 500 companies (VIX) during 2012. High values of NCI indicate cohesiveness in financial news and we suspect that it can be used as a proxy for financial risk.

Structure of semantic components

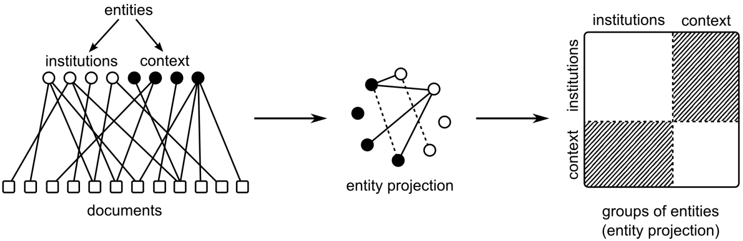


Figure: Dividing entities into semantic components - in this case into entities corresponding to financial institutions and entities corresponding to financial context (financial glossary terms).

Structure of semantic components

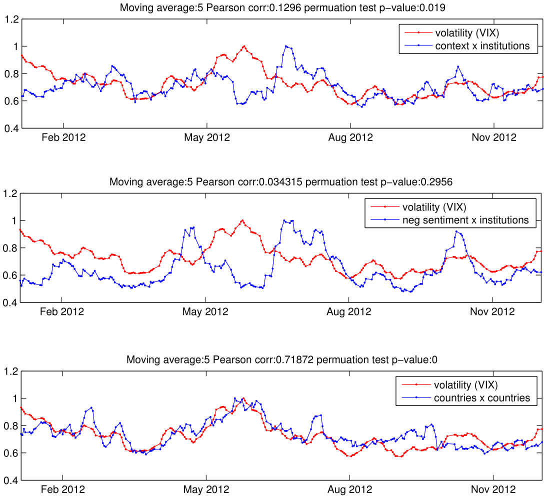


Figure: Correlations between individual semantic components and VIX.

Null models for NCI

We quantify statistics of *NCI* on random documents $\{\vec{d_1}, .., \vec{d_m}\}$ in order to estimate its *significance*.

Uniform null-model (NCI_u)

- Each document $\vec{d_i} = (1, 0, ..., 1)$ is a random binary vector, where degree is preserved and all entites are equally likely
- Documents $\{\vec{d_1}, .., \vec{d_m}\}$ are mutually independent
- Can be considered as a measure of noise in the system

Temporal null-model (NCI_c)

- Generates the statistics of NCI index calculated on m bootstrapped documents from independent days in a year.

Null models for NCI

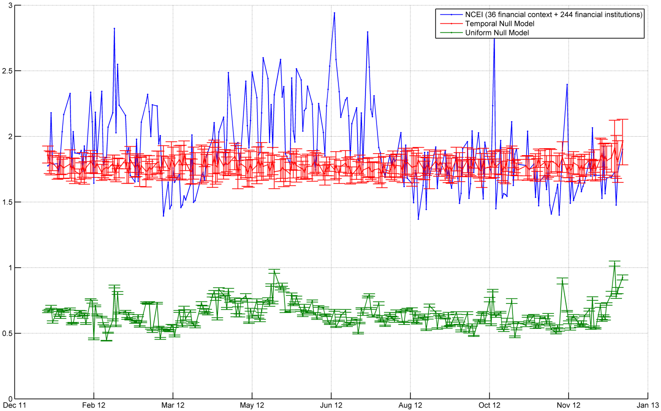


Figure: Null models for NCI.

Summary and future directions

Summary:

1. NCI measures average mutual similarity of texts in the corpus. In comparison with simple entity occurrences/co-occurrences it aims to be a *systemic measure*.
2. NCI captures the intrinsic property of the bipartite graph that is invariant to projection.
3. Singular computation of NCI enables fast real-time computation on large datasets and potentialy has a physical interpretation.

Thank you for you attention!
Questions?