Multi-Level News Networks

Borut Sluban, Jasmina Smailović, Miha Grčar, Igor Mozetič

1 Introduction

1.1 Summary

This chapter describes how to construct time-varying, multi-layer networks linking entities from the online news. We demonstrate the approach on a collection of over 36 million news articles that were published around the world in the last four years. Our multifold approach identifies interesting events from thousands of daily news and models temporal interactions between the entities in the news. Informative news should answer the following questions: *'Who?'*, *'Where?'*, *'When?'*, *'What?'*, and possibly *'Why?'*. The temporal aspect of the network answers the *'When?'* question, whereas the entity co-occurrence layer answers the *'Who?'* or *'Where?'* questions. The summary layer answers the *'What?'* question, and the sentiment layer labels the links as 'good' or 'bad' news. We distinguish between the usual/common and unusual/exceptional patterns in the news. We compare the news network to empirical, real-world networks and show that geographical proximity highly influences the co-occurrence of countries in the news, and that countries with significant trade exchange tend to be jointly mentioned in a positive context. Finally, we propose an approach for identifying the most relevant events linking different entities, and show that top news are not as positive as general news. We demonstrate the evolution of the news network, the top news content and the associated sentiment in an interactive web portal.

1.2 Motivation

News informs people about current events around the world. It mostly covers topics like politics, business, sports, extreme natural or social disasters, but also reports on activities of various social groups or public personalities. News are spread and/or sold by various news agencies using different media. By monitoring news web sites from around the globe, we analyze the structure and the contents of news. While research in news analysis mainly addresses statistical properties and interlinking of news (Lloyd *et al.*, 2005; Flaounas *et al.*, 2011; Leban *et al.*, 2014), we focus on the following research questions: 1) How to extract the usual, 'everyday' patterns in the news on one hand, and the unusual, highly publicized events on the other hand? 2) What do the usual and

unusual news actually reflect? 3) Are there properties of the news that show significant difference between the usual and unusual news?

We apply a set of text mining, sentiment analysis and network analysis methods to answer the above questions. The theory of complex networks characterizes systems in the form of entities (nodes) connected by some interactions (links) (Albert and Barabási, 2002). Since news talk about numerous different entities (for example persons, companies, countries, etc.) and their mutual interactions, they can be interpreted as a complex network. The methods of complex network analysis strongly influenced and advanced research in social media, biology, and economics (Caldarelli, 2007; Jackson, 2010). In certain research areas the available data does not have an inherent network structure like transportation networks, computer networks, or social networks. Depending on the available data and the field of research, various types of networks can be constructed, however. A special case of networks extracted from data are co-occurrence networks in which nodes represent some entities (persons, companies, countries, etc.), and links represent an observation that these entities exist together in some data collection (for example database, news article, etc.) (Edmonds, 1997). For textual sources, it is important to extract unambiguous entities using effective entity resolution (Christen, 2012), and to extract the links between the entities that represent real relations, and are not created by chance. In our previous work, we developed a method to estimate the significance of co-occurrences, and a benchmark model against which their robustness is evaluated (Popović *et al.*, 2014).

This chapter builds on our preliminary research on extraction of entity co-occurrence networks from news (Sluban *et al.*, 2016*b*,*a*) and extends it by a comparative analysis of usual and unusual events in the news. We construct time-varying networks of entities appearing in worldwide news. We enrich the links between the entities by textual context and sentiment, thus creating different network layers. By comparing the layers with different network comparison methods we draw interesting conclusions which answer some aspects of the research questions.

The main difference between the usual and unusual patterns in the news is a baseline against which the properties of the news are contrasted. The essence of the usual news are the links between entities that co-occur significantly more often than expected by chance. This results in a network of connected entities that gradually varies through time. To see what shapes the usual everyday news, we compare this network and the network of associated sentiment to three empirical networks constructed from real-world data, as illustrated in Fig. 1.

On the other hand, we propose to identify unusual news as significant deviations of the news volume over several weeks, for any pair of entities. The network of unusual events between entities, enriched by most relevant news content and sentiment polarity, shown in Fig. 2, is analyzed separately and its properties are compared to those of the everyday-news network.

The proposed news modeling approaches show that geography and world trade influence the structure and the attitude of everyday news, and that news in general tend to be slightly positive, whereas top news are more neg-

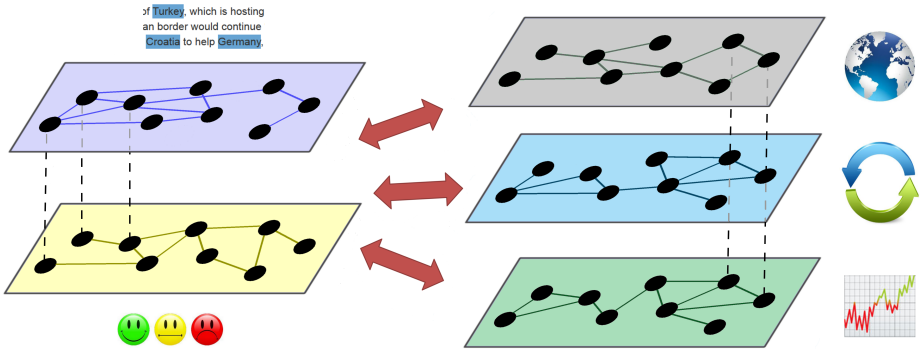

Figure 1: Analysis of the usual news. We compare two multiplex networks representing different types of relations between the same entities: significant co-occurrences and sentiment extracted from the news (at left), versus geographical proximity, high trade, and high correlation of financial indicators (at right, top to bottom).

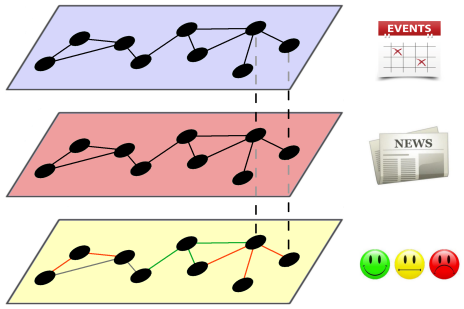

Figure 2: Analysis of the unusual news—major news events between pairs of entities, accompanied by their context in terms of news content and the associated sentiment polarity.

ative. Finally, we are also concerned with the presentation of such temporal multi-layer news networks. The network evolution over time, with drill-down inspection of details, is demonstrated in a public, interactive web portal at `http://newsstream.ijs.si/occurrence/major-news-events-map`. The portal facilitates access to over 38 million news, collected from 170 English news sites over a period of the last four and a half years.

The chapter is organized as follows. Related work is presented in the next subsection. We describe the methods for modeling and analyzing news in Section 2. Section 3 presents the results on everyday news and major news events, their differences, and the implemented interactive visualization of news networks. We conclude in Section 4 with ideas for future work.

1.3 Related work

This chapter builds on the work of several research fields: complex network analysis, text mining, and sentiment analysis. In this subsection we cover related research which is most relevant and provide references to broader overviews of the respective fields. The methods of complex network analysis are based on the methods developed in the fields of mathematics, computer science and statistical physics, and they have strongly influenced and advanced research in social media, biology, and economics (Caldarelli, 2007; Jackson, 2010). Particularly interesting are the co-occurrence networks which are extracted from data in the form of entities of interest (nodes) and their relations (links), inferred from their common context. The co-occurrence networks are used in diverse fields, such as linguistics (Edmonds, 1997), bioinformatics (Wilkinson and Huberman, 2004; Cohen *et al.*, 2005; Shalgi *et al.*, 2007), ecology (Freilich *et al.*, 2010), scientometry (Mane and Börner, 2004; Su and Lee, 2010), and socio-technological networks (Cattuto *et al.*, 2007; Zlatić *et al.*, 2009; Ghoshal *et al.*, 2009).

To extract networks from textual data one needs to apply different text mining methods. First, detecting and disambiguating the entities of interest requires efficient entity resolution (Christen, 2012), and second, the categorization of more complex relation types requires semantic analysis of the context (Feldman and Sanger, 2006). Simple word co-occurrence networks have been extracted to model language structure (Ferrer i Cancho and Solé, 2001) or to measure the relatedness between languages (Liu and Cong, 2013). But textual data typically provides rich context to the entities it mentions, enabling the construction of various types of networks, like signaling networks in biological systems by extracting subject-predicate-object triplets (Miljković *et al.*, 2012), bisociative information networks for bridging concepts (Juršič *et al.*, 2012), or semantic networks for the purpose of text understanding and summarization (Sowa, 1991; Miller, 1995; Kok and Domingos, 2008; Shang *et al.*, 2011).

Sentiment analysis (Liu, 2015) can provide information on the emotional state of the entities, or the attitude towards these entities expressed by other entities (Smailović *et al.*, 2014; Ranco *et al.*, 2015; Mozetič *et al.*, 2016). Modeling the emotional dynamics in networks has been explored for interactions between users in social networks (Miller *et al.*, 2011; Zollo *et al.*, 2015), as well as for shared economic or political interests (Smailović *et al.*, 2015; Sluban *et al.*, 2015).

Global news represent a rich resource of textual data, which has been used to extract different types of entity co-occurrence networks. Modeling personal connections evident from the news yields a social network of historically (Özgür *et al.*, 2008) or politically (Traag *et al.*, 2015; Hicks *et al.*, 2015) influential individuals and communities. Tracking country mentioning in global news was used to identify geographic community structure of the world's news media (Leetaru, 2011).

There exist several Web platforms which collect and analyze news articles, such as Lydia (Lloyd *et al.*, 2005), NOAM (Flaounas *et al.*, 2011), Event Registry (Leban *et al.*, 2014) and European Media Monitor (EMM). The Lydia

system provides various analyses for entities identified in news articles. For selected entity it constructs a relational network of relevant entities and provides different visualizations of popularity, sentiment and geographical analysis (`http://www.textmap.com/`). The NOAM platform identifies topics, events, frequent phrases, and named entities in collected news articles. The system translates articles into English if they are written in some other language. Event Registry (`http://www.eventregistry.org/`) detects events and extracts main information about them from news articles written in various languages. The Web interface enables the search of events and exploration of corresponding articles, related events, and different visualizations. The EMM (`http://emm.newsbrief.eu/`) is a news aggregation and analysis systems developed by the European Joint Research Centre. It provides identification of different named entities, topic aggregation over several languages, and article exploration.

2 Methods

We describe a multi-stage approach to investigation of world news that combines text mining, network mining and sentiment analysis. The stages consist of news acquisition and entity recognition, network construction, event detection, content identification, and sentiment analysis.

2.1 News network layers

Modeling the news requires to monitor entities of interest in the news, detect their co-occurrences (links) over time, and identify the associated context in terms of content and sentiment polarity. We present a method for detection of significant co-occurrence links between entities and propose a time-aware method for detection of significant events about pairs of entities. These methods enable us to model the news as temporal networks connecting different entities. We propose also a method for identifying most relevant content of major news events and show how to assess the sentiment of a group of documents.

2.1.1 News acquisition and entity recognition

The news are collected by our data acquisition and processing pipeline implemented within the NewsStream platform (`http://newsstream.ijs.si`) (Kralj Novak *et al.*, 2015). The pipeline consists of several components for: (i) data acquisition, (ii) data cleaning, (iii) natural-language preprocessing, and (iv) semantic annotation. The pipeline is running continuously since October 2011, polling the Web and proprietary APIs for recent content, turning it into a stream of preprocessed text documents. News is acquired from 2,600 RSS feeds from 170 English language web sites, covering the majority of web news in English. On average, 25,000 news articles are collected per day. In the period from October 2011 to November 2015, more than 36 million unique documents were collected and processed.

News are about events related to individuals, social groups, countries, or companies, which we call entities. The process of identifying entities in textual documents requires three components: an ontology of entities and terms, gazetteers of the possible appearances of the entities in the text, and a semantic annotation procedure that finds and labels the entities. The ontology that we use for information extraction constitutes of three main categories: geographical entities, main protagonists (companies, politicians, etc.), and financial terms. Each entity in the ontology has associated a gazetteer, which is a set of rules that specify the lexicographic information about possible appearances of the entity in text. For example, 'The United States of America' can appear in text as 'USA', 'US', 'the United States', etc. The rules include capitalization, lemmatization, POS tag constraints, must-contain constraints (i.e., another gazetteer must be detected in the document or in the sentence) and followed-by constraints. Finally, a semantic annotation procedure recognizes the entities of interest. It traverses each document and searches for entities from the ontology. The gazetteers of the entities in the ontology provide information required for the disambiguation of different appearances of the observed entities, resulting in a mostly correct annotation of the entities.

2.1.2 Significant co-occurrences: Co-occurrence layer

Entities identified in a single piece of news (i.e., a document) can be connected with various types of relations. One of the simplest is their common appearance in the document, referred to as the co-occurrence of entities. Hence, for a selected set of entities $E = \{e_1, \ldots, e_l\}$ we construct a network layer of entity co-occurrences within a particular time frame – the *Co-occurrence layer*. We use the Significance algorithm proposed by Popović *et al.* (2014) to assess whether the co-occurrence of two entities is statistically significant.

Let the number of all documents with at least two entities be N. Let A and B be two entities that occur with at least one other entity in N_A and N_B documents, respectively. Let N_{AB} denote the number of the actual A and B co-occurrences. The expected number of co-occurrences is $\mathbb{E}(N_{AB}) = \frac{N_A N_B}{N}$. According to Popović *et al.* (2014), the standard deviation is

$$\sigma_{AB} = \sqrt{\frac{N_A N_B}{N}\left(\frac{N^2 - N(N_A + N_B) + N_A N_B}{N(N-1)}\right)} \tag{1}$$

and hence the standard significance score of the co-occurrence N_{AB} from the data is

$$Z_{AB} = \frac{N_{AB} - \mathbb{E}(N_{AB})}{\sigma_{AB}} \, . \tag{2}$$

For a selected threshold Z_0, one can distinguish significant ($Z_{AB} > Z_0$) from non-significant ($Z_{AB} < Z_0$) co-occurrence relations between the two entities.

2.1.3 Major event detection: Event layer

We use the daily volume of news documents as a proxy for identifying exceptional events in the news. Given a set of entities of interest $E = \{e_1, \ldots, e_l\}$, we identify all events related to all pairs of entities (e_i, e_j). We monitor the volume of news about these pairs and construct a network of exceptional events between the observed entities—the *event layer*.

A link in the co-occurrence layer denotes that the number of actual co-occurrences is significantly grater than expected by chance. The random co-occurrence baseline is estimated from the observed individual occurrences. Here we propose a different approach that compares the number of observed co-occurrences in a day to a longer time period.

We construct a time series of co-occurrence volumes $\mathbf{v}_{ij} = \{v_{ij}(t)\}_{t=0}^{T}$ for a pair (e_i, e_j). At a given time point $t = p$, we consider a window $W_h(p) = \{v_{ij}(p-h-1), \ldots, v_{ij}(p-1)\}$ of length h as a historical baseline, from which we calculate the expected volume at the time point p. We assume, for a pair of entities, that the volume of their co-occurrences in news is normally distributed around the average in a given time period. As the value of the average changes through time, we use the sliding window W_h to adapt to recent changes.

Given the co-occurrence volume time series $\mathbf{v}_{ij}$, and the size h of historical data to consider, we calculate the mean co-occurrence volume $\bar{v}_{ij}(p)$ in $W_h(p)$ and its standard deviation $\sigma_{ij}(p)$. Let $z_{ij}(p)$ denote the multiple of $\sigma_{ij}(p)$-deviations from the mean $\bar{v}_{ij}(p)$:

$$z_{ij}(p) = \frac{v_{ij}(p) - \bar{v}_{ij}(p)}{\sigma_{ij}(p)} \,. \tag{3}$$

The co-occurrence volume $v_{ij}(p)$ at the *peak* day p is unexpected and represents an exceptional event between the entities e_i and e_j, when $z_{ij}(p) > Z_0$, for a given Z_0.

2.1.4 Identification of top news: Summary layer

We attribute shallow semantics to the links in the network by a summary of the top news at peak days in the form of the most relevant titles. First we select all the news related to a particular link on a particular day. The titles of these news are merged into a single text document. One such merged document is created for each day in the past two months (excluding weekends). We apply the standard text preprocessing approach to compute the bag-of-words (BOW) vectors of these documents (Feldman and Sanger, 2006). In this process, we employ tokenization, stop word removal, stemming, and the *TF-IDF* weighting scheme (Salton, 1989). The TF-IDF scheme is the most common weighting scheme used in text mining. The TF (term frequency) weight, $TF_{d,k}$, denotes the number of times the word k occurs in the document d. The IDF (inverse document frequency) weight of the word k is computed as $IDF_k = log\frac{|T|}{m_k}$, where m_k is the number of documents in the collection T that contain the word k. The *TF-IDF* scheme, $TFIDF_{d,k} = TF_{d,k} \times IDF_k$, weights a word higher if it

occurs often in the same document (the *TF* component), and at the same time lower if it occurs in many documents from the corpus (the *IDF* component).

The *BOW* vector for the current day contains information about how important a certain word is with respect to the most relevant events on that day. Instead of showing the top-ranked words, we propagate the weights to the news titles and thus rank the titles by their relevance. The weight-propagation formula is simple: we compute the average of the word-weights in a title c. The weight of the title, w_c, is thus computed as:

$$w_c = \frac{1}{|c|} \sum_{k \in c} TFIDF_{d^*,k} \ , \tag{4}$$

where k enumerates the words in the title c, and d^* represents the merged documents for the day in question. Note that this technique tends to penalize long titles. In our case, this is a desirable property because we would like to find short and to-the-point titles that best describe the most important event(s). The most distinguished titles at peak days represent the *summary layer* of the constructed network.

2.1.5 Lexicon-based sentiment analysis: Sentiment layer

The sentiment polarity of a document is computed from the number of predefined sentiment terms (positive and negative) in the document. The sentiment terms are from the Harvard-IV-4 sentiment dictionary (Tetlock *et al.*, 2008). For a document d, the sentiment polarity s_d is calculated as:

$$s_d = \frac{pos_d - neg_d}{pos_d + neg_d} \, , \tag{5}$$

where *pos* and *neg* are the numbers of positive and negative dictionary terms found in the document d, respectively. The sentiment polarities of a set of documents can then be aggregated over time. The aggregate sentiment of a pair of entities (e_i, e_j) in a certain time period T is computed from the news documents $\{d, (e_i, e_j) \in d\}$ at days $t \in T$:

$$s_{ij}(T) = \frac{1}{n} \sum_{t \in T} \sum_{d \in t} s_d \ , \tag{6}$$

where n is the total number of documents selected in the time period T. Based on the analysis of the sentiment distribution, we determine the thresholds n_0 and p_0 for the creation of positive and negative sentiment links.

2.2 Empirical network layers

We observe the same set of entities E as in the 'News network' layer, but the information regarding their mutual interactions is not acquired from the news. In particular, we explore three data sources to construct the empirical network

layers: the geographical proximity of the entities, their interaction in terms of mutual trade, and correlations between their financial indicators.

Geo layer. The simplest among the 'Empirical network' layers is the geographical proximity. Each entity has a predominant geographical location, place of residence, address or area. A link between two entities, A and B, is established if a selected proximity measure is above a given threshold. Examples of proximity measures include geographical distance $d(A, B)$, inverse distance $\frac{1}{d(A,B)}$, or inverse squared distance $\frac{1}{d(A,B)^2}$.

Trade layer models the interaction between entities as the amount of mutual trade. The amount of trade from e_i to e_j is denoted by $r(e_i, e_j)$. In total, e_i is engaged in $r(e_i)$ worth of trade with all other entities, where $r(e_i) = \sum_{e_j \in E \setminus \{e_i\}} r(e_i, e_j)$. The cumulative, non-directed amount of trade between e_i and e_j is $r(e_i, e_j) + r(e_j, e_i)$. For e_i, its relative share of trade with e_j is $R_{ij} = \frac{r(e_i,e_j)+r(e_j,e_i)}{r(e_i)}$. A trade link between two entities e_i and e_j is established if any of their relative trade shares, R_{ij} or R_{ji}, is above a given threshold R_0.

Financial layer. Certain entities have an associated time-varying financial indicator (e.g., price, trade volume), which is represented as a time series f. A basic approach to measure similar trends in the movement of financial indicators is the Pearson correlation (Pearson, 1895) between time series f_i and f_j of entities e_i and e_j, over a period of T time points:

$$\rho_{ij} = \frac{\sum_{t=1}^{T} (f_{i,t} - \bar{f}_i)(f_{j,t} - \bar{f}_j)}{\sqrt{\sum_{t=1}^{T} (f_{i,t} - \bar{f}_i)^2 \sum_{t=1}^{T} (f_{j,t} - \bar{f}_j)^2}} \ , \tag{7}$$

where $\bar{f}_i$ and $\bar{f}_j$ are the average values of the respective series. The *Financial layer* is constructed using a threshold value ρ_0, which determines whether the indicator time-series of two entities are sufficiently correlated ($\rho_{ij} > \rho_0$) to form a link between them.

2.3 Network comparison measures

A comparison of network layers $\{L_1, \ldots, L_m\}$ can be done by measuring the link overlap between the layers. Let $l(L_a)$ and $l(L_b)$ be the sets of links in layers L_a and L_b, where a link is defined as a pair of nodes it connects, e.g., (e_i, e_j). Then

$$o(L_a, L_b) = \frac{|l(L_a) \cap l(L_b)|}{|l(L_b)|} \tag{8}$$

is the size of their link overlap relative to layer L_b.

Considering for each layer not only the links, but also their weights (strength of the relation), a comparison of the top strongest links in each layer can be adapted. Let $sl(L_a)$ and $sl(L_b)$ be sorted lists of links from layers L_a and L_b, ordered by descending weights, and let $sl_k(L)$ denote the first k elements of $sl(L)$. Then *precision-at-k* ($prec_k$) (Raghavan *et al.*, 1989) is defined as:

$$prec_k(L_a, L_b) = \frac{|sl_k(L_a) \cap sl_k(L_b)|}{k} \ . \tag{9}$$

If for all pairs of layers L_a and L_b, $a, b \in \{1, \ldots, m\}$, the same k is selected, then a meta-network can be constructed with nodes representing layers L_a, $a \in \{1, \ldots, m\}$ and links representing the relation between the layers, where $prec_k(L_a, L_b)$ values are weights of the links, indicating the strength of the relationship.

Other comparisons of the network layers, induced on the 'strongest' links for a particular relation type, are based on the most important nodes in each layer. In one approach, we measure the importance of nodes in terms of their *centrality*, as denoted by the *eigenvector centrality measure* (Bonacic, 1972). Let A be the adjacency matrix of nodes $e_1, \ldots, e_n$ in the network. The components of the eigenvector of the largest eigenvalue λ solving the equation $A\mathbf{x} = \lambda\mathbf{x}$ hold the centrality values of the corresponding nodes. Nodes connected to better-connected nodes get higher centrality values. This measure can be used to compare the most central nodes between pairs of layers. Another approach to identify the most important nodes of a network is the *k-core decomposition* (Seidman, 1983). This is an iterative process, pruning all the nodes with degree smaller than k. The remaining part of the network which holds only nodes with degree greater or equal to k is called the k-core. The core with the largest k is called the main core of the network. Comparing the main cores of different network layers will be used to assess the similarity between the layers.

3 Results and Discussion

The proposed methods are used to 1) extract the usual, 'everyday' patterns in the news on the one hand, and the unusual, highly publicized events on the other hand, 2) analyze what do usual and unusual news actually reflect, and 3) discover whether any properties of the news show significant differences between the usual and unusual news. The results are presented in three parts. First, we analyze the everyday news, second, we focus on major news and the differences towards everyday news, and finally, we show some visualizations of the network snapshots.

3.1 Analysis of everyday news

3.1.1 Network construction

We monitored 50 countries in the news as entities of interest and constructed a network of their co-occurrences using the significance algorithm presented in Section 2.1.2. From the news, we also created a sentiment layer of the country co-occurrences showing the sentiment of the joint context in which both entities co-occur. The network layers are sampled in monthly snapshots over a time period of two years.

The construction of the empirical network, which should reflect the real-world context of the news, was done using data from external sources. For the *Geo layer* we used the *is-a-neighbour-of* relation to link the selected countries. Links representing common terrestrial boarders were extended also with a few

links between countries that are considered adjacent in the local geographical context, such as Australia and New Zealand, South Korea and Japan, or Italy and Malta.

Trading relations between the countries were obtained from the UNCTAD website (`http://unctadstat.unctad.org`), the United Nations statistics data center, providing yearly aggregations of trade data. Our *Trade layer* was constructed from trade links that present relatively important trade relations (greater than 10%, i.e., $R_0 = 0.1$) for at least one of the connected countries.

We consider 50 countries that issue sovereign bonds, and which are insured by Credit Default Swaps (CDS), i.e., an insurance for the case when the bond issuer defaults and is unable to repay the debt. To construct the *Financial layer* we used the time series of their CDS prices, which are often considered a good proxy for the risk of default of the country issuing bonds (Pan and Singleton, 2008; Aizenman *et al.*, 2013). We create links between countries whose correlation between their CDS time series is above 0.9 ($\rho_0 = 0.9$). In order to ensure enough data for reliable correlation results, we use a three-months time window for each snapshot, and assigned it to the last month (e.g., the Nov-Dec-Jan window for the 'Jan' snapshot).

3.1.2 Co-occurrence vs. empirical layers

The results are presented for a multiplex network of 50 country nodes, for the time period of two years, 2012 and 2013. The co-occurrence network L_{CO} varies with time, and we used a one-month time window. The *Geo layer* L_{Geo} and the *Trade layer* L_{Tr} are static. The yearly aggregated trade data from 2012 were used for 2013 as well. The *Financial layer* L_{CDS} varies — we used three-months time window.

First, we present the analysis of overlapping links between the network layers L_{CO}, L_{Geo}, L_{Tr}, and L_{CDS}. We are interested in the number of links from the 'empirical network' that appear in the news as country co-occurrences over time. The relative overlaps $o(L_{CO}, L)$ for $L \in \{L_{Geo}, L_{Tr}, L_{CDS}\}$ are presented in Fig. 3. We see that most of the Geo layer links coincide with the country co-occurrences in the news, whereas on average less than half of the links between the countries in the Trade and Financial layers also appear in the co-occurrence layer.

Next, we investigate how is the sentiment associated with the country co-occurrences related to the empirical network. Using the sentiment analysis approach presented in Section 2.1.5 we find that there is a strong bias towards positive sentiment in the news. We set thresholds n_0 and p_0 to two standard deviations apart from the average sentiment polarity in the documents, thus selecting only links that reflect the most negative and most positive sentiment in the context of two countries. The negative sentiment layer turns out to be predominantly small, even for a slightly less restrictive threshold n_0 (at 90% st. dev. from the average) and therefore has mostly low overlap with the empirical layers. On the other hand, the comparison of the positive sentiment layer with the empirical layers results in a larger number of common links, as shown in

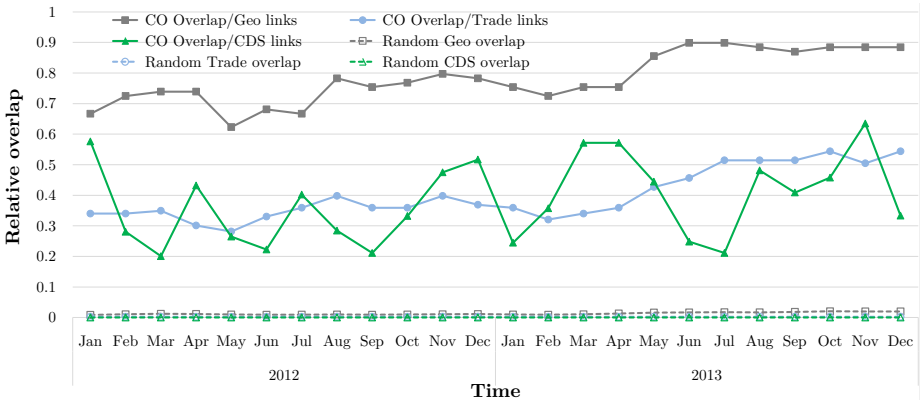

Figure 3: Relative size of the empirical layer links present in the co-occurrence layer.

Fig. 4. Positive sentiment between the countries has the largest overlap with the trade relations, followed by geographical proximity and to the smallest extent by the correlation between the CDS time series.

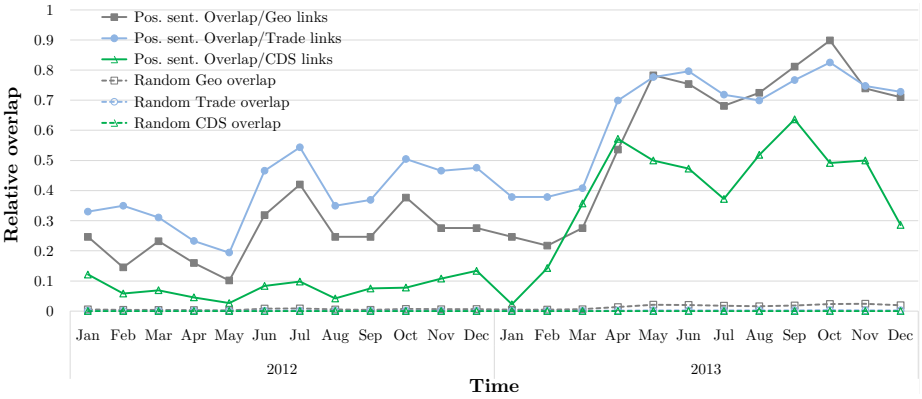

Figure 4: Relative size of the empirical layers' links present in the positive sentiment layer.

Comparison of the most important nodes in each layer shows similar results. The comparison of the main k-cores results in the largest overlap between the co-occurrence and Geo layer cores, and positive sentiment and the Trade layer cores, see Figures 5 and 6. The co-occurrence layer cores overlap with the Geo layer cores in central European countries, and with the Trade layer cores in western European countries. The overlaps between the co-occurrence and CDS layers show common presence of some eastern European countries in 2012,

but no regular presence in 2013. Several countries regularly appear in the core overlap between the positive sentiment and the Trade layers (CN, DE, US, UK, JP, BR, FR, and AU). Germany is also almost all the time (23 months) in the core overlap of the positive sentiment and the Geo layers.

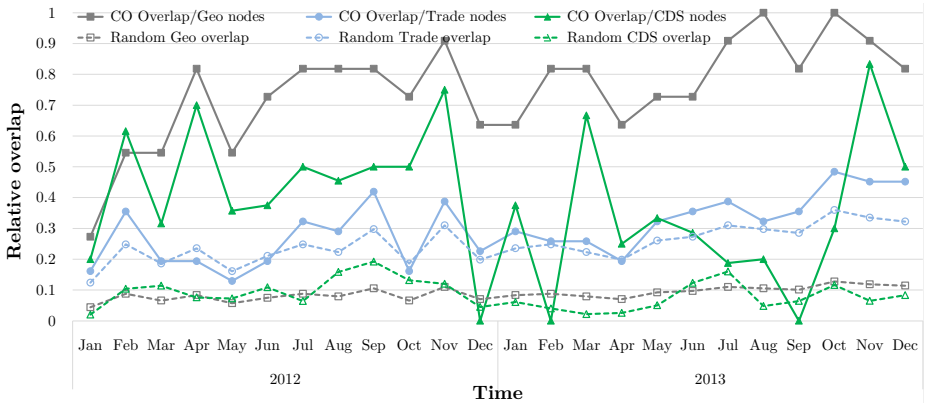

Figure 5: Relative size of the empirical layers' links present in the positive sentiment layer.

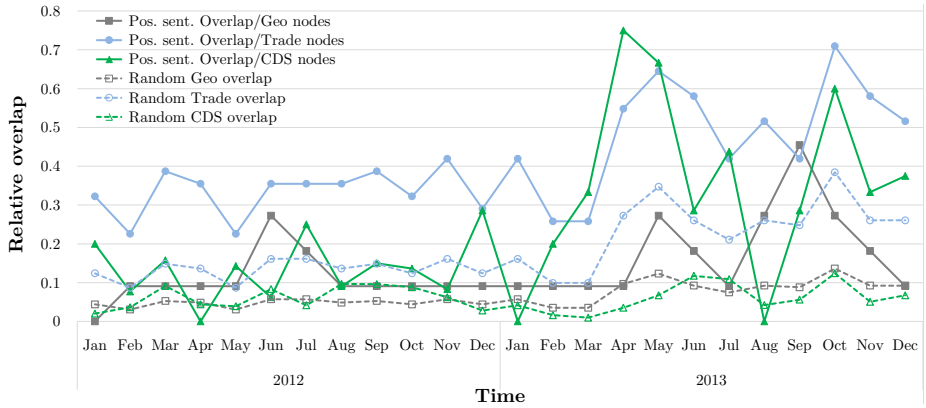

Figure 6: Relative size of the empirical layers' links present in the positive sentiment layer.

Most central nodes of the co-occurrence layer coincide with the Geo layer in central European countries (AT, CZ, HU, SK, SI), with the CDS layer in few eastern European countries, and with the Trade layer only Finland appears often among the top ten most central nodes. For the positive sentiment layer, the common most central nodes are Germany and Russia for the Geo layer, and some of the largest economies (CN, DE, FR, JP, RU, US) for the Trade layer.

Finally, we use the *precision-at-k* method to measure the link overlap of the strongest relations in each layer, in terms of the highest excess over random co-occurrence, most positive sentiment, highest mutual trade volume, and highest correlation between financial indicators. Limited by the number of links in the Geo layer, k was set to 69. Fig. 7 illustrates the relations between the layers weighted by the *precision-at-69* values.

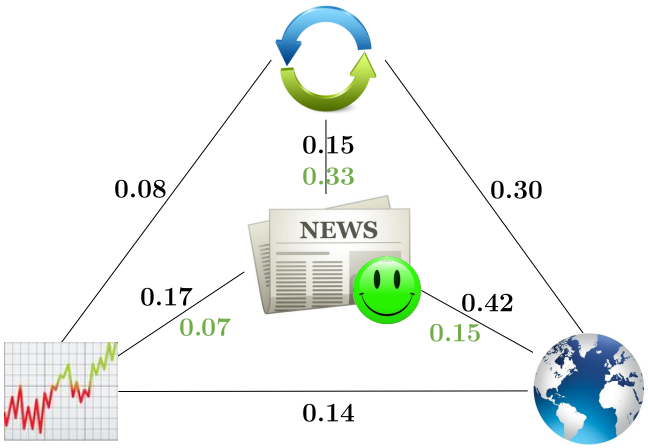

Figure 7: A meta network between the news and empirical network layers.

We can observe similar results as in Figures 3 and 4. The News layer has the highest overlap of strongest links with the Geo layer. One can infer that neighbouring countries tend to appear together in the news. On the other hand, the positive sentiment layer has the highest overlap with the Trade layer, suggesting that countries with high mutual trade tend to appear together in a positive context in the news. We can also observe a relatively strong relation of trade between neighbouring countries, whereas relations between other layers are weaker.

3.2 Major event news

In this section we focus on the detection of major events in the news. We describe the construction of a temporal network of different countries as entities of interest, over a period of the last four years. The network reveals the semantics of the relations between the countries in terms of the extracted contents and sentiment.

3.2.1 Significant events

For all country pairs in the news on NewsStream we selected the news with at least 3 occurrences of both entities and at least one entity in the title. We detect significant events by comparing the daily news volume to the volume of

the previous two months (44 weekdays, weekends excluded due to much lower volume). We assume a normal distribution of the entity co-occurrence volume around the average number of co-occurrences over the past two months. We set $Z_0 = 3$ to identify most outstanding news production increases about a pair of entities. Hence, if on a particular day the news volume exceeds the average volume of the past two months by more than three standard deviations, this day is identified as a significant event day—peak day for the observed pair of entities.

In Fig. 8 we show the volume of news containing the entities 'China' and 'United Stated' in the period between November 2011 and October 2015. The significant increases in the news volume are peaks above the gray line, indicating $\bar{v}_{\text{CN-US}} + 3 \cdot \sigma_{\text{CN-US}}$.

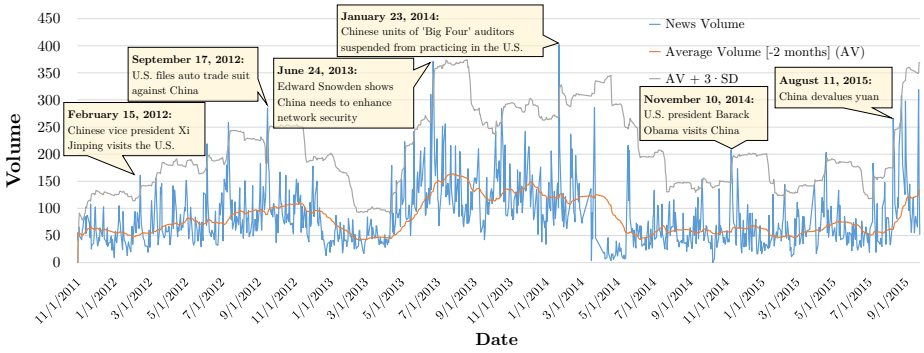

Figure 8: Volume of news articles about 'China' and 'United States'. Significant events are at days peaking more than three standard deviations above the average volume in the previous two months. Labels at certain peak days describe an event concerning China and the United States on that day, as manually identified in the news.

We added labels to some of the volume peaks in Fig. 8 to illustrate what kind of significant news events related to China and the United States happened on the particular dates. These facts were obtained by manually inspecting news and official press releases of both countries. In the period between January 2012 and October 2015, 17,997 significant events between 217 countries were detected. We analyze these events in terms of the automatically detected relevant content and the associated sentiment in the following subsections.

3.2.2 Contents and sentiment

We identify the most relevant and distinguishing topics for each significant event day, as described in Section 2.1.4. However, evaluating the obtained results in terms of their relevance proved to be quite challenging, as there is no exact ground truth for the type of events that we are detecting. One publicly available

resource of historical major news events that we could find is provided by the *Europe Media Monitor* (EMM).

Hence, we compare our top news results to the major news timeline of EMM, `http://emm.newsexplorer.eu/NewsExplorer/timelineedition/en/timeline.html`. The overlap with all the EMM major news is 45%, and 60% with major news topics mentioning at least two countries in the topic title. These differences are mostly due to the following reasons. As major news events of EMM are not limited to country relations (links), they include also news events mentioning only one country or none at all. Our approach, on the other hand, detects a wider range of more specialized events, which relate pairs of countries (entities). Topics persisting for several days with low evolution are avoided by our approach as we are looking for significant new events. A country's involvement in a certain topic may be overlooked in the evaluation process due to unresolved indirect mentioning, like 'Merkel' or 'VW' instead of Germany.

Some significant event days in August and September 2015, for four country pairs, are presented in Tables 1 and 2. Each event is characterized by the top news headlines detected by our approach and the associated sentiment calculated from the texts of the top news.

Table 1: Content and sentiment of the most relevant news on significant event days between China and the United States, in August 2015.

Link	Day	News title	Sentiment
CN - US (-0.189)	Aug 11 2015	China devalues yuan	-0.022
		China Devalues Renminbi	0.139
		China devalues yuan by 2%	0.056
	Aug 12 2015	China devalues its currency again	-0.333
		China Currency Falls for 2nd Day After Surprise Devaluation	-0.333
		China currency falls again for 2nd day after surprise devaluation	-0.228
	Aug 24 2015	Alarm bells ring as China sinks, dollar tumbles	-0.290
		Great fall of China sinks world markets	-0.356
		Great fall of China sinks world stocks, dollar tumbles	-0.300
	Aug 25 2015	Global markets rebound after China cuts rates	-0.193
		Global markets rebound after China cuts interest rates	-0.234
		Dow jumps 300 points as China cuts interest rates	-0.171

The first example describes the events related to significant increases of news volume about China and the United States in August 2015. As China devalued its currency, the *Renminbi* (or *Chinese yuan* as better known internationally),

Table 2: Content and sentiment of the most relevant news on significant event days between France and Russia, France and Egypt, and between Germany and the United States, in August and September 2015.

Link	Day	News title	Sentiment
FR - RU (0.265)	Aug 6 2015	France to pay Russia under $1.31 billion over warships	0.286
		France to pay Russia under 1.2 billion euros over warships	0.256
		France says several nations interested in Mistral warships	0.254
FR - EG (-0.072)	Sep 23 2015	France sells 2 disputed warships to Egypt	-0.091
		France sells warships to Egypt after Russia deal scrapped	-0.020
		France to sell warships to Egypt after Russia deal scrapped	-0.103
DE - US (-0.015)	Sep 21 2015	VW rocked by US emissions scandal as stock slides 17 percent	0.039
		VW Rocked by U.S. Emissions Scandal as Stock Slides 17%	-0.036
		VW shares plunge on emissions scandal US widens probe	-0.026
	Sep 24 2015	Will Volkswagen scandal tarnish Made in Germany image?	0.007
		After year of stonewalling Volkswagen stunned U.S. regulators with confession	-0.042
		Insight - After year of stonewalling Volkswagen stunned U.S. regulators with ...	-0.030

the U.S. media discussed the possible impacts on its economy. Similar increased media coverage of the two countries can be observed when the effects of China's actions become evident and when China cuts its interest rates. Notice the changes in the sentiment of these news, ranging from neutral (initially) and negative (perceived effects) to less negative (relief after shock). The second two illustrative examples are about the events concerning French-built warships, which were not delivered to Russia, but were later sold to Egypt. These events are also accompanied by different sentiment polarity. The fourth news example highlights the 'emissions scandal' of a German automobile producer VW, which broke out in the United States in September 2015.

We construct the *major event network* from significant news events between country pairs. We use the top three most relevant news articles at those peak days and their associated sentiment to construct the summary and sentiment layers.

3.2.3 Comparison to everyday news

We compare the 'everyday' news and the 'major event' news in terms of network structure and the sentiment of the news. For the comparison between the everyday and major event news networks, the links of the event network are merged for each month and filtered to the selected 50 countries. Details of the structural comparison are presented in Table 3. Structural analysis shows that the major event network is on average less densely connected, having less than 40% links as the everyday network. On average only a quarter of the event links are also in the everyday network, which is due to the greater sensitivity (daily resolution) of the event detection approach, which can identify individual peak days that are left undetected by the averaging over several days as done by Popović *et al.* (2014). The major event network has also a lower clustering coefficient, which suggests that major events in the news tend to involve less countries than everyday news. The most interesting structural difference between the two networks is regarding their assortativity (Newman, 2002). In the everyday network countries with similar presence in the news tend to co-occur, whereas in the major event network unequally represented countries tend to co-occur, which supports the unusual nature of the detected events.

Table 3: Comparison of the structural properties of the Everyday news and the Major event news networks. Shown are the average values with standard deviations of network density, clustering coefficient, and network assortativity, over a period of two years.

Network	Density	Clustering	Assortativity
Everyday news	0.23 ± 0.07	0.65 ± 0.06	0.53 ± 0.07
Major event news	0.12 ± 0.03	0.27 ± 0.12	-0.22 ± 0.10

We examine the differences in the sentiment distribution of everyday news and the major event news. We include also the 'title' news, the base for event detection mentioning at least one of the relevant countries in their title, and the 'peak' news, i.e., all the news on significant event days. Fig. 9 shows the sentiment distributions.

All four sentiment distributions are approximately normal, and very similar. There is an evident positive sentiment bias in everyday news, while the peak news are slightly negative. Title news also show a minor positive bias, whereas major event news are on average less positive but contain proportionally more extremely positive and extremely negative news articles. These statistics are summarized in Table 4.

We test the null hypothesis that a pair of news populations has equal mean sentiment. We apply the Welch's t-test (Welch, 1947) which is robust for skewed distributions, and large sample sizes (Fagerland, 2012). The results are in Table 5. With t values > 10, the degrees of freedom $\gg 100$, and the p-value ≈ 0, the null hypothesis can be rejected for all pairs of news populations. We conclude, with high confidence, that the four populations of news have significantly different sentiment means, though some of these differences are very small.

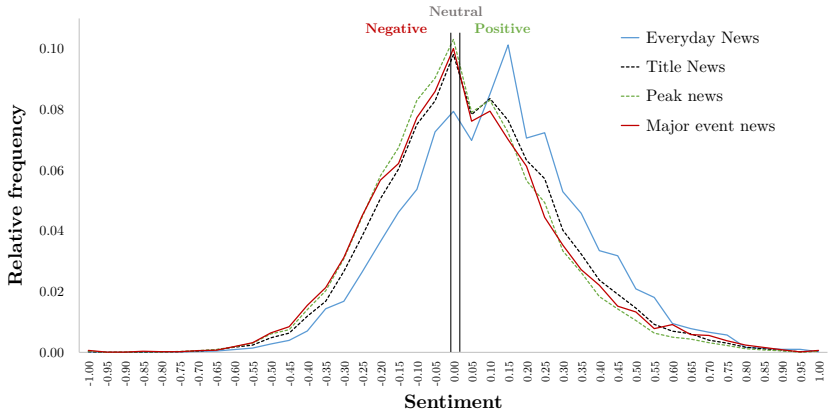

Figure 9: Comparison of sentiment distributions of everyday news articles, title news, peak day articles, and major event news, i.e., most relevant articles at peak days.

Table 4: The sentiment distributions for different sets of news relating pairs of countries. There is the number of documents, $\bar{s}$ is the sentiment mean, SD standard deviation, and SEM standard error of the mean.

News corpus	Documents	$\bar{s}$	SD	SEM
Everyday news	7,391,204	0.092	0.243	0.0001
Title news	1,590,388	0.029	0.239	0.0002
Peak news	279,432	-0.002	0.232	0.0004
Major event news	48,864	0.011	0.249	0.0011

Table 5: The results of t-tests for comparison of sentiment means. DF is the estimated degrees of freedom.

News corpora		t	DF
Everyday news	vs. Title news	300.56	2,355,313
Everyday news	vs. Peak news	209.88	303,211
Everyday news	vs. Major event news	71.64	49,480
Title news	vs. Peak news	64.78	391,099
Title news	vs. Major event news	15.83	51,655
Major event news	vs. Peak news	10.60	64,477

We introduce a neutral zone around the sentiment mean $\bar{s}$, to distinguish 'bad' from 'good' news. As $\bar{s}$ is the sample mean, then the population mean is in the interval $\bar{s} \pm 9SEM$ with very high confidence. We classify the sentiment of the top news into three discrete classes: *negative* if $-1 \leq s < 0$, *neutral* if $0 \leq s \leq +0.02$, and *positive* if $+0.02 < s \leq +1$. The neutral zone is used to distinguish between the negative and positive sentiment of major event news in

the network visualization.

3.2.4 Network visualization

A network visualization offers an insight to better understand and analyze complex systems by enabling the user to easily inspect and comprehend relations between individual units and their properties (Rossi and Magnani, 2015). In addition to a single layer network visualization (Batagelj and Mrvar, 2004; Bastian *et al.*, 2009), also multi-layer network visualization is becoming increasingly popular for highlighting various aspect of complex systems (Secrier *et al.*, 2012; Krzywinski *et al.*, 2012; De Domenico *et al.*, 2014; Piškorec *et al.*, 2015).

We implemented a spatio-temporal visualization of the country co-occurrence network, constructed from the detected major event news, their most relevant content, and the associated sentiment. The visualization is implemented within the NewsStream portal, to facilitate the inspection of various aspects of the network: time dimension, news content, news sentiment, and geography. The network is embedded into the world map, and the interface includes functionalities to explore different aspects of the network. Figs. 10 and 11 show two instances of the network in time and space. The visualization is an extension of the NewsStream portal, and is publicly accessible at `http://newsstream.ijs.si/occurrence/major-news-events-map`.

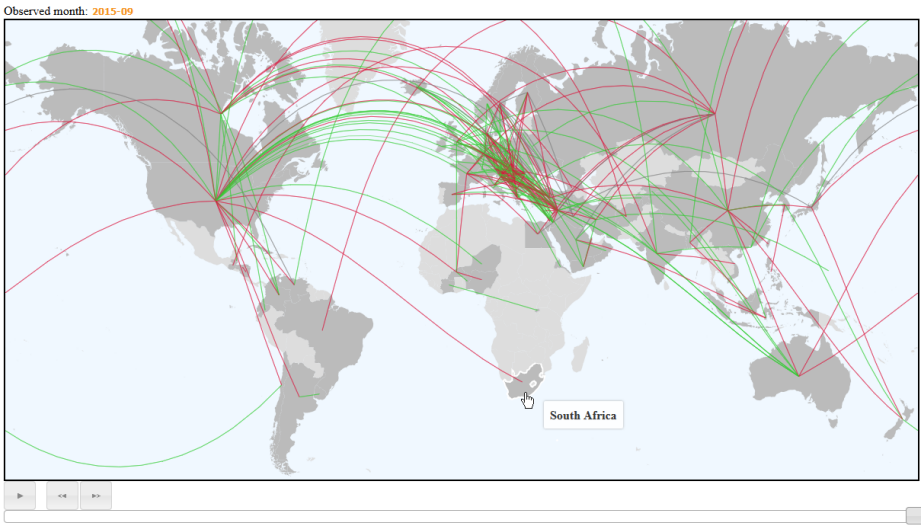

Figure 10: Temporal country co-occurrence network of major news events during September 2015.

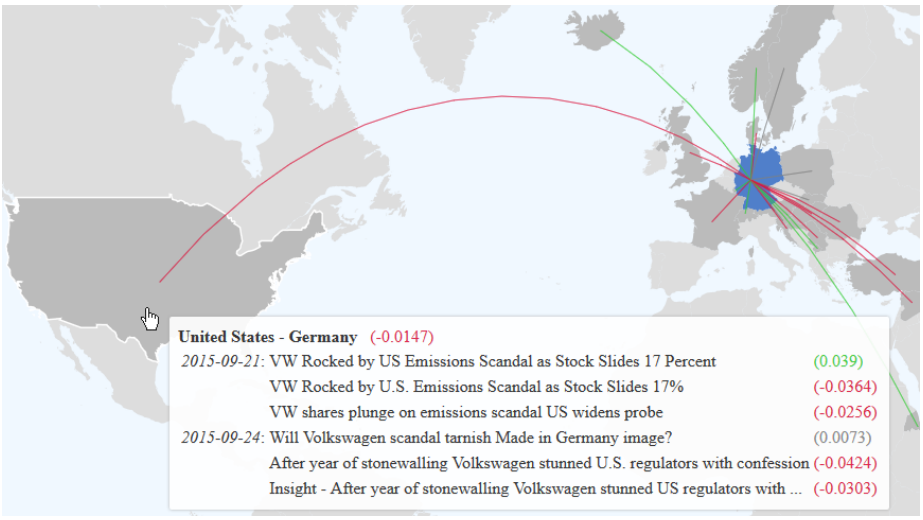

Figure 11: The most significant news about Germany and the United Stated in September 2015.

4 Conclusions

We describe several methods for modeling and analyzing news by means of text mining, network mining, and sentiment analysis. The resulting temporal multi-layer news network reveals the dynamic relations between various entities appearing in the news. It enables to capture usual and unusual news events about different entities, summarize the relations between them in terms of most relevant articles, and assign the sentiment to the corresponding context. From the corpus of over 36 millions news articles published in the last four years we constructed a time-varying network between countries which were mentioned in the news. We show that countries which are geographically close tend to co-occur in everyday news, and that countries having good trade relations tend to be mentioned in a positive context. Exploring unusual patterns in news, on the other hand, we find that major news events are more negative than everyday news, which have an evident positive sentiment bias. Furthermore, in major event news there are more co-occurrences between unequally mentioned countries than in everyday news. Finally, we implemented an interactive network visualization that supports the spatio-temporal exploration of the constructed networks.

We plan to broaden the range of semantic relations extracted from text in order to construct a public knowledge network from news. Another direction of future research is to study the role of news in the policy making process. As news is shaping the opinions in policy debates, we plan to extend the news network with (in)direct ownership structure of the media companies, and analyze how this influences the reported news.

Acknowledgment

This work was supported in part by the European Commission FP7 project MULTIPLEX (no. 317532), and by the Slovenian ARRS programme Knowledge Technologies (no. P2-103).

References

Aizenman, J., Hutchison, M., and Jinjarak, Y. (2013). What is the risk of European sovereign debt defaults? Fiscal space, CDS spreads and market pricing of risk. *Journal of International Money and Finance*, **34**(C), 37–59.

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, **74**(1), 47.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *Proc. Intl. AAAI Conference on Weblogs and Social Media*.

Batagelj, V. and Mrvar, A. (2004). Pajek – analysis and visualization of large networks. In *Graph Drawing Software*, pp. 77–103. Springer.

Bonacic, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, **2**, 113–120.

Caldarelli, G. (2007). *Scale-Free Networks: Complex webs in nature and technology*. Oxford University Press.

Cattuto, C., Schmitz, C., Baldassarri, A., Servedio, V.D.P., Loreto, V., Hotho, A., Grahl, M., and Stumme, G. (2007). Network properties of folksonomies. *AI Communications*, **20**(4), 245–262.

Christen, P. (2012). *Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Data-Centric Systems and Applications. Springer.

Cohen, A.M., Hersh, W.R., Dubay, C., and Spackman, K. (2005). Using co-occurrence network structure to extract synonymous gene and protein names from medline abstracts. *BMC bioinformatics*, **6**(1), 103.

De Domenico, M., Porter, M.A., and Arenas, A. (2014). Muxviz: a tool for multilayer analysis and visualization of networks. *Journal of Complex Networks*, **3**, 159–176.

Edmonds, P. (1997). Choosing the word most typical in context using a lexical co-occurrence network. In *Proc. 35th Annual meeting of ACL*, pp. 507–509.

Fagerland, M.W. (2012). t-tests, non-parametric tests, and large studies – a paradox of statistical practice? *BMC Medical Research Methodology*, **12**(78), 1–7.

Feldman, R. and Sanger, J. (2006). *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, New York, USA.

Ferrer i Cancho, R. and Solé, R.V. (2001). The small world of human language. *Proc. Royal Society of London*, **268**(1482), 2261–2265.

Flaounas, I., Ali, O., Turchi, M., Snowsill, T., Nicart, F., De Bie, T., and Cristianini, N. (2011). NOAM: News outlets analysis and monitoring system. In *Proc. ACM SIGMOD Intl. Conf. on Management of Data*, pp. 1275–1278.

Freilich, S., Kreimer, A., Meilijson, I., Gophna, U., Sharan, R., and Ruppin, E. (2010). The large-scale organization of the bacterial network of ecological co-occurrence interactions. *Nucleic Acids Res.*, **38**(12), 3857–3868.

Ghoshal, G., Zlatić, V., Caldarelli, G., and Newman, M.E.J. (2009). Random hypergraphs and their applications. *Physical Review E*, **79**(6), 066118.

Hicks, J., Traag, V.A., and Reinanda, R. (2015). Turning digitised newspapers into networks of political elites. *Asian Journal of Social Science*, **43**(5), 567–587.

Jackson, M.O. (2010). *Social and economic networks*. Princeton University Press.

Juršič, M., Sluban, B., Cestnik, B., Grčar, M., and Lavrač, N. (2012). Bridging concept identification for constructing information networks from text documents. In *Bisociative Knowledge Discovery*, pp. 66–90. Springer.

Kok, S. and Domingos, P.M. (2008). Extracting semantic networks from text via relational clustering. In *Machine Learning and Knowledge Discovery in Databases*, pp. 624–639. Springer.

Kralj Novak, P., Grčar, M., Sluban, B., and Mozetič, I. (2015). Analysis of financial news with NewsStream. Technical Report IJS-DP-11965, Jožef Stefan Institute, Ljubljana, arXiv: 1508.00027.

Krzywinski, M., Birol, I., Jones, S.J.M., and Marra, M.A. (2012). Hive plots–rational approach to visualizing networks. *Briefings in Bioinformatics*, **13**(5), 627–644.

Leban, G., Fortuna, B., Brank, J., and Grobelnik, M. (2014). Event registry: Learning about world events from news. In *Proc. 23rd Intl. Conf. on World Wide Web*, pp. 107–110.

Leetaru, K.H. (2011). Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, **16**(9).

Liu, B. (2015). *Sentiment Analysis - Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Liu, H. and Cong, J. (2013). Language clustering with word co-occurrence networks based on parallel texts. *Chinese Sci Bull*, **58**(10), 1139–1144.

Lloyd, L., Kechagias, D., and Skiena, S. (2005). Lydia: A system for large-scale news analysis. In *String Processing and Information Retrieval*, pp. 161–166. Springer.

Mane, K.K. and Börner, K. (2004). Mapping topics and topic bursts in PNAS. *Proc. National Academy of Sciences*, **101**(Suppl 1), 5287–5290.

Miljković, D., Stare, T., Mozetič, I., Podpečan, V., Petek, M., Witek, K., Dermastia, M., Lavrač, N., and Gruden, K. (2012). Signalling network construction for modelling plant defence response. *PLoS ONE*, **7**(12), e0051822.

Miller, G.A. (1995). WordNet: A Lexical Database for English. *Commun. ACM*, **38**(11), 39–41.

Miller, M., Sathi, C., Wiesenthal, D., Leskovec, J., and Potts, C. (2011). Sentiment flow through hyperlink networks. In *Proc. 5th Intl. Conf. on Weblogs and Social Media*. The AAAI Press.

Mozetič, I., Grčar, M., and Smailović, J. (2016). Multilingual Twitter sentiment classification: The role of human annotators. *PLoS ONE*, **11**(5), e0155036.

Newman, M.E.J. (2002). Assortative mixing in networks. *Phys. Rev. Lett.*, **89**, 208701.

Özgür, A., Cetin, B., and Bingol, H. (2008). Co-occurrence network of Reuters news. *Intl. Journal of Modern Physics C*, **19**(05), 689–702.

Pan, J. and Singleton, K.J. (2008). Default and recovery implicit in the term structure of sovereign CDS spreads. *The Journal of Finance*, **63**(5), 2345–2384.

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proc. Royal Society of London*, **58**, 240–242.

Piškorec, M., Sluban, B., and Šmuc, T. (2015). MultiNets: Web-Based Multilayer Network Visualization. In *Proc. European Conf. on Machine Learning and Knowledge Discovery in Databases*, pp. 298–302. Springer.

Popović, M., Štefančić, H., Sluban, B., Kralj Novak, P., Grčar, M., Mozetič, I., and Zlatić, V. (2014). Extraction of temporal networks from term co-occurrences in online textual sources. *PLoS ONE*, **9**(12), e99515.

Raghavan, V., Bollmann, P., and Jung, G.S. (1989). A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems*, **7**(3), 205–229.

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., and Mozetič, I. (2015). The effects of Twitter sentiment on stock price returns. *PLoS ONE*, **10**(9), e0138441.

Rossi, L. and Magnani, M. (2015). Towards effective visual analytics on multiplex and multilayer networks. *Chaos, Solitons & Fractals*, **72**(0), 68–76.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Boston, USA.

Secrier, M., Pavlopoulos, G.A., Aerts, J., and Schneider, R. (2012). Arena3D: visualizing time-driven phenotypic differences in biological systems. *BMC Bioinform.*, **13**, 45.

Seidman, S.B. (1983). Network structure and minimum degree. *Social Networks*, **5**(3), 269–287.

Shalgi, R., Lieber, D., Oren, M., and Pilpel, Y. (2007). Global and local architecture of the mammalian microrna–transcription factor regulatory network. *PLoS computational biology*, **3**(7), e131.

Shang, Y., Li, Y., Lin, H., and Yang, Z. (2011). Enhancing biomedical text summarization using semantic relation extraction. *PLoS ONE*, **6**(8), 1–10.

Sluban, B., Grčar, M., and Mozetič, I. (2016*a*). Temporal multi-layer network construction from major news events. In *Proc. 7th Workshop on Complex Networks CompleNet*, pp. 29–41. Springer.

Sluban, B., Smailović, J., Battiston, S., and Mozetič, I. (2015). Sentiment leaning of influential communities in social networks. *Computational Social Networks*, **2**(9), 1–21.

Sluban, B., Smailović, J., and Mozetič, I. (2016*b*). Understanding financial news with multi-layer network analysis. In *Proc. European Conf. on Complex Systems 2014*, pp. 193–207. Springer.

Smailović, J., Grčar, M., Lavrač, N., and Žnidaršič, M. (2014). Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences*, **285**, 181–203.

Smailović, J., Kranjc, J., Grčar, M., Žnidaršič, M., and Mozetič, I. (2015). Monitoring the Twitter sentiment during the Bulgarian elections. In *Proc. IEEE Intl. Conf. on Data Science and Advanced Analytics*, pp. 1–10. IEEE.

Sowa, J.F. (1991). *Principles of Semantic Networks: Explorations in the Representation of Knowledge*. Representation and Reasoning. Morgan Kaufmann.

Su, H.-N. and Lee, P.-C. (2010). Mapping knowledge structure by keyword co-occurrence: a first look at journal papers in technology foresight. *Scientometrics*, **85**(1), 65–79.

Tetlock, P.C., Saar-Tsechansky, M., and Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, **63**(3), 1437–1467.

Traag, V.A., Reinanda, R., and van Klinken, G. (2015). Elite co-occurrence in the media. *Asian Journal of Social Science*, **43**(5), 588–612.

Welch, B.L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika*, **34**(1–2), 28—35.

Wilkinson, D.M. and Huberman, B.A. (2004). A method for finding communities of related genes. *Proc. National Academy of Sciences*, **101**(Suppl 1), 5241–5248.

Zlatić, V., Ghoshal, G., and Caldarelli, G. (2009). Hypergraph topological quantities for tagged social networks. *Physical Review E*, **80**(3), 036118.

Zollo, F., Kralj Novak, P., Del Vicario, M., Bessi, A., Mozetič, I., Scala, A., Caldarelli, G., and Quattrociocchi, W. (2015). Emotional dynamics in the age of misinformation. *PLoS ONE*, **10**(9), e0138740.