*Natural Language Engineering* **1** (1): 000–000. Printed in the United Kingdom



# *Network Analysis of Narrative Content in Large Corpora*

SAATVIGA SUDHAHAR, GIANLUCA DE FAZIO,
ROBERTO FRANZOSI, NELLO CRISTIANINI

*Intelligent Systems Laboratory, University of Bristol, UK*
*email: saatviga.sudhahar@bristol.ac.uk, nello.cristianini@bristol.ac.uk*
*Department of Sociology, Emory University, Atlanta, USA*
*email: rfranzo@emory.edu, gdefazi@emory.edu*

( *Received 15 March 2013*)

**Abstract**

We present a methodology for the extraction of narrative information from a large corpus. The key idea is to transform the corpus into a network, formed by linking the key actors and objects of the narration, and then to analyse this network to extract information about their relations. By representing information into a single network it is possible to infer relations between these entities, including when they have never been mentioned together. We discuss various types of information that can be extracted by our method, various ways to validate the information extracted, and two different application scenarios. Our methodology is very scalable, and addresses specific research needs in social sciences.

**Keywords:** network analysis, story grammar, semantic triplets, spectral graph partitioning

## 1 Introduction

The analysis of text, most notably news content, is a fundamental research task, for example in the social sciences, but also in the humanities and the political sciences. Often this tasks is performed by hand manually (in a process known as "coding" in that literature) before any quantitative analysis can be performed.

One important set of tasks involves the identification of basic narrative information in a corpus. That is identifying the key actors and objects and their relations. We will refer to actors and objects generally as entities. This can be approximated by identifying the "SVO triplets" (subject-verb-object) that appear in a text. For example in the sentence "A dog bit a man" we would extract the triplet "Dog-Bite-Man". There are various applications in the detection of such semantic triplets, and we will focus mostly on the study of the networks that result from linking together all entities of a given narration (the resulting structure is sometimes called a semantic graph).

For example in Quantitative Narrative Analysis (QNA)(Franzosi 1987)(Earl*et al.*

2004) the fundamental idea is to find the actors and their relations by extracting all SVO triplets. While this is only a subset of the narrative information contained in a text, the set of all SVO triplets does contain information about the key entities and actions described in that text. In the QNA literature the SVO structure of a text is also called a "story grammar".

The present article describes a scalable methodology to extract narrative networks from large corpora, discusses various issues relative to the validation of the resulting information, and shows two different applications of this methodology, to the analysis of crime stories and of political stories. Our methodology is focused on the extraction of high quality triplets, and contains a series of filtering steps to ensure that only highly reliable information is identified. This high precision comes at the cost of smaller recall, as we will see, but does create networks that capture valuable information from a corpus.

The contribution of this study is not in the improvement of tools for the processing of language (e.g. parsers) but in the developement of a new methodology for the extraction of knowledge from a large corpus. The information we extract (e.g. the political relations among actors) is not found in any individual document, but inferred from information distributed across the corpus, by effect of analysing a large network assembled by using all the documents. We test our approach on a corpus of 200,000 articles about the 2012 US elections, as well as on small corpora relative to the past 6 election cycles, always extracting statistically significant relations that result from the collective analysis of all the documents. We also present a study of crime stories from the New York Times corpus.

In section 2 we discuss related work. In section 3 we present the key conceptual framework behind our methodology, and we will describe the software pipeline that we have used to implement it. We have used existing tools whenever possible, for the various stages of the pipeline.

In section 4 we will describe some of the network properties that we can extract from the data. These include the centrality of entities, their tendency to be subjects rather than objects, the division of entities in different camps, and more.

In section 5 we will discuss the thorny issue of validating the methodology. This is difficult as there is very limited data that we can access, but we propose a multi strategy approach to validation: validating the entire pipeline (by computing the p-value of certain network properties that we measure); hand validation of a small subset of triplets; and study of the existing literature that has validated various sub components of the pipeline.

In section 6 we present an experimental study of the 2012 US elections and the past 6 elections cycles, focusing only on verbs of two types signalling positive or negative attitude from an actor to another actor or object and showing that the resulting network does capture the actual political relations among entities with a very high degree of significance (hence addressing some of the validation problems discussed in section 5).

In section 7 we discuss the validation of entity spectrums, obtained from the elections data that produces a ranking of entities from the left to the right of the political spectrum.

In section 8 we present a study of crime stories from the New York Times, distinguishing between crimes against property and those against person. Again we show that valuable information can be extracted by turning a corpus into a network.

Section 9 discusses the limitations of this approach, its relations with pre-existing methods, and draws the conclusions from this study.

## 2 Related Work

Our approach builds on an idea presented in (Rusu *et al.* 2007) for purposes of triplet extraction. They discuss various ways of extracting triplets using different parsers like Stanford Parser, OpenNLP, Link Parser and Minipar. In this approach, Subject-Verb-Object (SVO) triplets are extracted from text by a parser, and then used to generate a semantic graph that captures narrative structures and relations contained in a text. These semantic graphs have been then used for document visualisation and construction of document summaries using SVM classifiers (Rusu *et al.* 2008). (Trampus and Mladenic 2011) describe extraction of event templates by identifying small subgraphs of these semantic graphs. Another work by (Dali 2009) describes a question answering system where the answer generated is described by a semantic graph, by its automatically generated summary and by a list of facts which stand for SVO triplets. Hence the kind of information that can be extracted from semantic-graph representations of large amount of text data is remarkable. We made use of this idea and developed it in many ways for the extraction of narrative information from a large corpus.

Our methodology extracts narrative networks from large corpora and studies its properties as mentioned above. A similar approach has also recently been used in the humanities for the analysis of novels, called "Distant Reading" (Moretti 2011). In that domain, novels are turned into networks, whose nodes are the actors of the narration, and whose links are the verbs. Topological properties of the network are used to identify protagonists, antagonists and so on. A recent study has compared the actor networks resulting from three mythological epics (Padraig and Ralph 2012). The bottleneck for these studies has always been the extraction of the triplets, a work that is labour intensive and therefore limits these studies to small samples. A fully automated crime data mining framework was developed and network centrality measures was used by (Chen 2004) to analyse Crime data and detect key members in criminal groups. But this study was limited to only 36 criminal reports from the Pheonix police department in Arizona.

## 3 Network Inference

The main idea of our methodology is to identify the entities (actors and objects) and actions that form the narration contained in a text or a corpus. In the election part of the study we are interested in extracting attitudes (positive or negative) of actors (e.g. persons, organisations, etc) towards other actors or other objects which

may include ideas, issues, events, etc. This is based on classical approaches in the social sciences.

Heider (1946) says that "we shall understand by attitude the positive or negative relationship of a person $P$ to another person $O$ or to an impersonal entity $X$ which may be a situation, an event, an idea, or a thing". In the literature of QNA (Franzosi 1987) the only actors and objects that are kept are those that fit in a SVO structure, and the most general structures that can play the role of S or of O in a sentence are noun phrases. For example, in the sentence: "The customer enjoyed the product" we have two noun phrases (i.e. "The customer" and "the product") and a verb (enjoy). In other words, we extract the most general set of actors and objects that are compatible with the existing definition in QNA literature.

In the social sciences there is an interest in identifying social actors which are animate entities (e.g. "the mob", " the pope" or "a woman") as opposed to inanimate ones (e.g. "the earthquake", "recession" ). In this study we will not make any effort to distinguish between animate and inanimate entities, although this remains an important research question which we will discuss in the conclusions. It is often possible to use information about their syntactical role (e.g. subject vs object) to identify actors from other entities.

For the elections study we will focus on statements in the text where a certain actor expresses positive or negative attitude towards another actor or an object, in the form of a SVO triplet. Noun phrases can appear on both sides of this triplet, and we will use them as candidate actors/objects (note that this set also includes named entities). We can easily distinguish between actors and pure objects by separating those noun phrases that have been seen "to act" (by being the subject in a triplet) from those that have not been seen to act (by being only or mostly seen as the object of the triplet). Similarly, we consider as actions the verbs found in the text, and we focus in this study only on transitive verbs, although it would be technically possible to operate also on intransitive ones.

While this is a design choice, it is one that we have observed to cover many entities in our validation, achieving 62% precision and 57% recall. The entire approach is based on the idea that we extract explicitly stated information, ignoring metaphores and indirect allusions, relying on the fact that we analyse vast amounts of data and focus only on relations that are supported by a large number of articles.

We will use a parser to extract SVO triplets. The process of parsing has greatly been improved in recent years but it is still a difficult task to automate and it can result in erroneous SVO triplets. In order to increase the precision of our system we will only accept triplets that have been seen a certain amount of times in the corpus. As this step lowers the recall of the system, we precede it with two steps aimed at reducing the number of different noun phrases that can be found: anaphora and co-reference resolution (these steps will be explained in detail below). We will also introduce a weighting scheme to identify those actors and actions that are relevant to a given analysis (e.g. the most specific to the corpus at hand, or the most frequent).

We also assign verbs to a small set of categories, by using lists of verbs, so that there are only a small number of verb types in the triplets. These can be seen as

expressing a relation among the actors (in the example "The customer enjoyed the product", there is a relation of 'approval' from the customer to the product).

The resulting set of triplets reduced in size by each of the above steps can then be assembled into a network. The topology of the network will represent some simplified information that is contained in the corpus, perhaps in distributed and implicit manner. From the topological structure among nodes it will be possible to infer relations among actors that appear in the same corpus, but perhaps never in the same document. The analysis of networks can, in this sense, replace other forms of inference.

### *3.1 Key Definitions*

We will call entities all noun phrases/proper nouns that have been seen as subject or object in a SVO triplet. Entities include both actors and objects in the context of social science. We will call actions the verbs that have been seen within a SVO triplet. We call two equivalent triplets that refer to the same entities in different ways, or use different verbs to express the same action. A parser is a software that identifies syntactic structures in natural language.

By using a parser, we can extract a list of all SVO triplets in a corpus. One of the main problems is to recognise equivalent triplets. While a full solution to this kind of problem is very difficult, we can introduce some pre-processing steps aimed at alleviating it.

First we perform co-reference resolution of named entities, which is the process of determining whether two expressions in natural language refer to the same entity in the world (Soon *et al.* 2001). Then we perform anaphora resolution of pronouns. Anaphora is a cohesion which points back to some previous item. The "pointing back" (reference) is called an anaphor and the entity to which it refers is its antecedent. The process of determining the antecedent of an anaphor is called anaphora resolution (Mitkov 1999). The example below will help clarify these steps.

Consider the sentence:
"Romney praised Paul Ryan. He recalled the excitement of the country in electing Obama four years ago. Ryan criticized Obama for rejecting a deficit reduction plan."

After coreference and anaphora resolution, the sentence is rewritten as follows:
"Romney praised Ryan. Romney recalled the excitement of the country in electing Obama four years ago. Ryan criticized Obama for rejecting a deficit reduction plan."

After parsing, we can identify the following SVO triplets:
"Romney praise Ryan"
"Romney recall excitement"
"Ryan criticize Obama"

Since the verb "recall" is not a positive/negative attitude we would create a small directed network out of these triplets like shown in Figure 1 where the edge be-

Fig. 1: Network


tween "Romney" and "Ryan" denote praise (positive attitude) and the edge between "Ryan" and "Obama" denote criticize (negative attitude). The real network would also have a positive/negative weight with the sign on the edges based on the number of positive and negative triplets extracted (explained in section 3.2). Note incidentally that while Romney and Obama do not appear in the same triplet, this set contains implicit information about their relation which we may want to access.

### *3.2 Reliable and Relevant Triplets*

Each of the steps described so far may introduce errors in the process of extracting narrative information. We will discuss the difficult issue of validation of our results, and of the methodology, in Section 4. However there is an obvious step to reduce the amount of errors in our output: if sufficient input data is available, we can filter away all uncertain or irrelevant results, to keep only those of more interest for our task.

This introduces the need to quantify the reliability and the relevance of a triplet, or perhaps an actor, an object or an action. We will do this by introducing a weighting scheme.

We will define both the weight of an entity or action. These quantities, that can be changed for different applications and tasks, will allow us to rank and select the most relevant or reliable information to include in our network.

**Relevance (of entities or actions).** Relevance of entities or actions to a given topic can be gauged by comparing their relative frequency in the corpus at hand with that of a background corpus. For example if we want to emphasize sport-related verbs we could compare the relative frequencies of all verbs in a corpus of sports articles with those in a background corpus, selecting those verbs that are most specific of sport. One possible choice of weight is shown in Equation 1:

$$w_a = \frac{f_a(T_1)}{f_a(T_2)} \tag{1}$$

where $w_a$ refers to the weight of the entity/action; $f_a(T_1)$ and $f_a(T_2)$ refer to the frequency of the entity/action in a given corpus $T_1$ and a background corpus $T_2$ .

**Reliability (of Triplets).** We can select reliable (and relevant) triplets by various means. One is to use their frequency in the corpus: triplets seen in more than $k$ independent documents could be considered acceptable. Another method is to choose those triplets that include key-entities and key-actions. We combine these and consider triplets containing key entities/actions, which have been seen in more

than $k$ independent documents as reliable. The decision on k is explained later. The highest ranking candidates according to Equation 1 are considered as key entities and key actions. For example a reliable SVO triplet could be defined in these ways:

S (Key entity) V (key action) O (entity)
S (entity) V (key action) O (Key entity)
S (Key entity) V (key action) O (Key entity)

**Strength of Relations.** In the elections study we will map verbs to positive/negative attitude between entities. Once we have identified a set of reliable triplets by the methods above, we can use them to assess the strength of a relation between two actors. For example, we could have two lists of verbs, one signalling actions compatible with positive attitude and the other signalling actions compatible with negative attitude. Then we could just count every triplet as a vote in favour of positive or negative attitude, and calculate a weight for each of the two possible relations.

As we define the extent to which one actor $a$ supports/opposes an object or another actor $b$, we need to combine the number of positive and negative statements observed in the data going from node $a$ to node $b$. There are various ways to do this, and they correspond to slightly different interpretations of the meaning of that score. A possible approach to quantifying the weight of a relation between entity $a$ and entity $b$ is to consider also a confidence interval around our estimate of the value of that relation. This will relate to the estimation of the parameter of a Bernoulli distribution, so that we can then calculate the confidence interval around this estimate by using standard methods.

The math for this was worked out in 1927 by Edwin B. Wilson. According to it the Wilson score confidence interval(Wilson 1927) for a Bernoulli parameter is given by,

$$\mathrm{w} = \left( \hat{\mathrm{p}} + \frac{\mathrm{z}^2{}_{\alpha/2}}{2\mathrm{n}} \pm \mathrm{z}_{\alpha/2} \sqrt{\frac{\hat{\mathrm{p}}(1-\hat{\mathrm{p}})}{\mathrm{n}} + \frac{\mathrm{z}^2{}_{\alpha/2}}{4\mathrm{n}^2}} \right) / (\frac{\mathrm{z}^2{}_{\alpha/2}}{\mathrm{n}} + 1) \tag{2}$$

Here $\hat{\mathrm{p}}$ is the fraction of positive observations, $z_{\alpha/2}$ is the (1-$\alpha$/2) quantile of the standard normal distribution, and n is the total number of observations. For a confidence level of 95% the value for $z_{\alpha/2}$ is 1.96. This could be approximated to 2 and a simplified version of the Wilson score interval could be obtained by considering the number of positive (P) and negative (N) triplets found between any two entities $a$ and $b$. Equation 3 shows the simplified version.

$$\mathrm{w} = \frac{\mathrm{P}+2}{\mathrm{P}+\mathrm{N}+4} \pm \frac{2\sqrt{\frac{\mathrm{P.N}}{\mathrm{P+N}}+1}}{\mathrm{P}+\mathrm{N}+4} \tag{3}$$

As we can see that this range consists the mean $m$ that is, $\frac{P+2}{P+N+4}$ and the actual interval $i$ that is, $\frac{2\sqrt{\frac{P.N}{P+N}+1}}{P+N+4}$ on either side of the mean. When a positive/negative relation is supported by many independently generated triplets ($k$), $i$ becomes smaller and the resulting network would contain the most reliable information. Hence we introduce a threshold to the percentage of $i$ and accept relations only if $i$ lies below

this threshold. Details on how we select this threshold is explained in the Validation section. In this way the value for $k$ remains very high implicitly.

The final score of our links should be associated as a function of the proportion of positive triplets, which is possible in this case since: (P-N)/(P+N) = 2 P/(P+N)-1; then to observe that P/(P+N) is the rate of positive mentions, and then to treat the estimation of this quantity like the estimation of parameters of a Bernoulli distribution. This score is computed using the lower bound of the Wilson interval. Since the correction is quation 3 for 95% confidence the final weight on the links become,

$$S = 2\left(\frac{P+2}{P+N+4} - \frac{2\sqrt{\frac{P.N}{P+N}+1}}{P+N+4}\right) - 1 \tag{4}$$

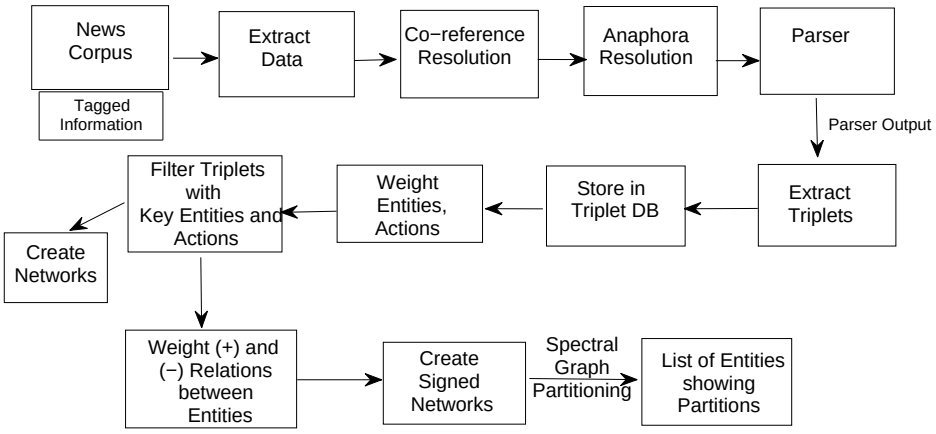
The above methods can be used to select either a set of SVO triplets, or a set of binary relations that we consider as sufficiently supported by the corpus, calculate weights and use them to assemble a network.

### *3.3 Software Pipeline*

We have described in the subsections 3.1 and 3.2 all the conceptual steps that we do in order to turn a corpus into a network of actors, objects and actions. We describe here the software pipeline that we have used in our experiments. The two guiding principles were for us to re-use existing tools where possible, and to make a system that can scale to large corpora. Figure 2 shows the system pipeline. Each component of the pipeline is explained in detail.

- News Corpus - The system uses articles contained in an available news corpus to perform the task.

Fig. 2: System Pipeline



- Extract Data - We could first extract the content from articles that are specific to a domain which is of interest to the analysis. e.g. Crime, Elections, Sports etc.
- Co-reference Resolution - The text in every individual article is processed for named entity co-reference resolution. The Orthomatcher module in ANNIE Information extraction system in GATE (General Architecture for Text Engineering)(Cunningham 2002) distribution is used to perform this task.
- Anaphora Resolution - Once the co-references have been resolved the Pronominal resolution module in ANNIE is used to perform anaphora resolution. The system solves pronouns in all forms that are identified by GATE.
- Minipar Parser - We use the parser Minipar (Lin 1998) to parse the above processed text. The parser tags each word of the sentence with its grammatical relation to it. Minipar has its own limitations since it cannot parse sentences more than 1024 characters long. On the other hand, we found that this length exceeds the size of a typical sentence in the news which is made of approximately 500 characters.
- Extract Triplets - From the Minipar parser output we extract words tagged with s (subject), i (verb) and obj (object of the verb) relations. A SVO triplet is formed out of these words if the s, i, obj relations are found in the sentence in this chronological order.
- Store in Triplet DB - All extracted triplets are stored in the Triplets database along with the article information from which they were extracted. This includes, article date, title, content and article feed URL. We also store the Minipar parser output for each article.
- Weight Entities and Actions - Entities (subject/object of triplet) and actions are weighted according to Equation 1 and this weight is used to rank and select the highest ranking candidates as key entities and key actions.
- Filter Triplets with Key Entities and Actions - We then filter the triplets that have key entities as subjects/objects and key actions as verbs.
- Create Networks - Directed networks are created with the triplets where the nodes are entities and the edges are actions linking them. To create networks we use Cytoscape (Shannon*et al.* 2007) which is a general platform for complex network analysis and visualization. We also used JUNG[1] for automatically generating networks and analysing network properties.
- Weight Positive and Negative Relations between Entities - Positive and negative relations indicate friendship or hostility between actors like mentioned before. In order to identify the strength of these relations we introduce a weighting method which is shown in Equation 4. This would result in entities linked by a positive/negative link with weights.
- Create Signed Networks - We create signed networks where nodes are entities and edges are the positive/negative links with weights.
- Spectral Graph Partitioning - Signed networks are partitioned using spectral graph partitioning methods to assess the degree to which actors/objects

[1] JUNG: http://jung.sourceforge.net/

belong to or in favor of one of two parties, in the assumption that the networks are naturally organised into two main communities. This is explained in section 4.4. We used the JAMA[2] matrix package for java to perform this task.

- List of entities showing partitions - Once the network is partitioned we obtain a list of entities which shows the association of them to one of the two communities in the network.

We do not discuss here the problem of validation of the software, because we leave it for Section 5.

## 4 Network Analysis

There are many advantages in representing the information extracted from a corpus in the form of a network (or semantic graph). One of them is that several types of relations among entities can easily be calculated, without requiring any explicit form of logical or other inference. Another advantage is that the overall shape of the network can reveal much about the properties of the corpus, and allow comparisons with other corpora. For example the role played by an actor (say a hero or a villain) within the narration might be reflected by its topological position within the network (Padraig and Ralph 2012).

### *4.1 Finding Central Actors/Objects*

The most obvious application of network analysis to the extraction of corpus-level narrative information is to identify the most central actors/objects to the narration. There are several well known measures of node centrality in a network, and each of them can be used to capture some different aspects of narrative centrality.

Betweenness centrality measures how important a node is by counting the number of shortest paths of which it is a part (Mihalcea and Radev 2011). In-Degree and Out-Degree measure the count of the nodes number of inward and outward ties to other nodes. Link analysis algorithms like HITS (Hyperlink-Induced Topic Search) (Kleinberg 1998) produce two network measures called authority and hub. The authority score indicates the value of the node itself and hubs estimates the value of the links outgoing from the node. PageRank (Brin and Page 1998) is a way of deciding on the importance of a node within a graph. When one node links to another one, it is casting a vote for that other node. The higher the number of votes that are cast for a node, the higher the importance of the node.

### *4.2 The Subject/Object Bias of an Entity*

Another source of information about entities is how often they appear as subjects or objects in the narration. This can give information about their role in the news

[2] JAMA: http://math.nist.gov/javanumerics/jama/

narrative: that is, its tendency to be portrayed as an active or passive element in the story. We make use of the subjects and objects in the collected triplets to do this.

We can compute the subject/object bias of an entity $S_a$ by finding the distance between the absolute frequencies of entities as subjects and objects like in Equation 5.

$$S_a = \frac{f^T_{subj}(a) - f^T_{obj}(a)}{f^T_{subj}(a) + f^T_{obj}(a)} \tag{5}$$

$f^T_{subj}(a)$ and $f^T_{obj}(a)$ refer to the frequency of entity $a$ as a subject and object in a given corpus T. This quantity $S_a$ is in the interval [-1,+1] where a positive score indicates subjectivity and a negative score indicates objectivity.

### *4.3 Lists of Verbs*

One way to identify higher level relations among entities is to classify the verbs into categories. For example we could have verbs expressing friendship (or being compatible with a relation of friendship) or hostility. The sighting of a single triplet containing one such verb would not allows us to conclude that such a relationship exists, but the sighting of several independent such triplets (possibly in different documents) would start increasing the evidence towards that.

An important problem is therefore to create lists of verbs that are organised by type. We have experimented with verbs that denote political support and political opposition, and with verbs that denote crimes, but this can be extended to virtually any domain. Currently our lists are generated by using pre-existing resources such as ontologies, or by hand. For example we used VerbNet (Kipper*et al.* 2006) to obtain English verbs and annotated them with tags: crime against person, crime against property, political support and political opposition. Crime-related verbs were obtained from Wikipedia lists[3]. Verbs denoting political support/opposition were obtained by manually going through the actions in triplets that were extracted from the New York Times elections data (Sandhaus 2008). Synonyms of these verbs were also added to the corresponding lists using the online thesaurus dictionary.

### *4.4 Spectral Analysis of Networks*

We may be interested in assessing the degree to which actors or objects are in favor of one of two parties, in the assumption that the network is naturally organised into two main communities. We would expect that the actors in the same community will have positive attitudes towards each other, while actors in different communities will have negative attitudes towards each other. In the case of objects, certain issues or concepts could be favored by one of two parties. We are interested in partitioning the graph into two classes such that nodes in the same class are linked by positive edges and nodes in different classes are linked by negative edges. The

[3] Crime related verbs: http://en.wikipedia.org/wiki/Offence_against_the_person

division of a network into two parts can be a computationally expensive step, but it can be relaxed to a simple algebraic task by introducing the approximation that the adjacency matrix is symmetric and positive definite, an assumption that can be readily satisfied: given a network with its adjacency matrix $A$, we make it symmetric by adding it to its transpose resulting in matrix $M{=}A{+}A^T$.

In matrix $M$ where $M_{ij} \in$ {-1,+1} we want to assign each node to one of the two classes -1,+1 as mentioned. This leads to the following optimisation problem.

$$argmax_{y \in \{-1,+1\}^m} \sum_{ij} M_{ij} y_i y_j \tag{6}$$

We relax this problem (which is NP hard) by allowing the membership function of each node to assume values in $\mathbb{R}$ ($y_i \in \mathbb{R}$) while keeping the norm of y fixed to avoid trivial solutions. The problem now reduces to the following optimisation problem.

$$max_y \frac{y^T M y}{y^T y} \tag{7}$$

This is equivalent to the eigenproblem My = $\lambda$y, by Rayleigh quotient since M is symmetric and positive definite by construction, and therefore is efficiently solvable.

The real value assigned to each node in the eigenvector can be interpreted as the degree to which it belongs to one of the two classes. Each eigenvector corresponds to a possible bi-partitioning of the graph, with the quality of the partition being represented by the corresponding eigenvalue. Therefore it is natural to make use of the first eigenvector, possibly looking at the second one when the eigenvalues are very similar.

Results of spectral graph partitioning methods on real networks will be presented in the experiments section.

## 5 Validation of the Pipeline

Estimating the performance of this methodology is a difficult and important task. There are no accessible corpora that have been annotated in terms of SVO triplets that we can use in order to measure precision and recall of our method, and there are no other networks of actors/objects that have been generated by hand, based on a corpus. This is not an unusual situation, as most new tasks do not come with a gold-standard benchmarking dataset attached. However there are various things we can measure, in order to increase our confidence in the method, and obtain a rigorous statistical estimate of performance.

In Validation 1 we estimate the probability P(T) of a given triplet T extracted once by our tool and was not in the source text. When the probability of this event is known, we can estimate the probability of detecting the same spurious triplet multiple times, in the assumption that the extraction process is independent (ie: it is applied on independently written text). While this is obviously a simplifying assumption, we feel that it is a reasonable one, and it allows us to obtain a ballpark estimate for the probability of a spurious triplet being seen $k$ times in $k$ independent observations of text by trivial calculation of probability of joint independent events.

Secondly we can easily measure precision of our tool by examining by hand the number of errors in its results. This can be readily done. What cannot be readily done is to estimate the number of missing results, as this would require actually having the set of all true triplets, which we do not have. We will call this Validation 2.

Thirdly, we can apply rigorous statistical testing to properties of the network that we know must be true. If the resulting network has the expected properties, then we know that the entire process for its production must have been extracting valid information, even without estimating the performance of each individual step. In one of the following Sections, we will report experiments on the 2012 US elections and the past 6 election cycles, and each time we measure if the two main parties "Democrats" and "Republicans" are correctly separated in the network of political support. These entities are chosen because they are present in every election and always on distinct camps. This removes subjectively choosing the test statistics and increases rigour. We apply statistical hypothesis testing to that experiment, obtaining a p-value that very strongly rejects the null hypothesis. We call this Validation 3.

In other words, by following a multi-strategy approach, we can increase our confidence that the system is extracting valid and valuable information. In particular, the statistical significance study on the elections network and the precision estimates performed on the triplets we have extracted point in the same direction: that our system extracts very precise information that represents the true relations among the actors in the corpus. Estimating the recall would be harder, but since we work in the setting of very large datasets, we choose to focus on obtaining high precision rather than high recall.

**Validation 1.** We have used a corpus covering the Civil Rights movement in the Northern Ireland. For that corpus (which contains little or no repetitions) a previous analysis had been done (De Fazio 2012) and therefore 72 manually extracted triplets were available. We applied our methodology, without filtering 'reliable' triplets, due to the limitations of the data. Our method extracted 66 triplets out of which 41 were correct while 31 were missed. This gives us 62% precision and 57% recall in the very unfavourable case when we cannot use any filtering for reliable triplets. This means that there is a probability of 38% of a triplet being incorrect, if it has been seen just once. If we use this figure as the error rate for triplets seen once, we can use it in a model for the probability of error in triplets seen more than $k$ times, which would be $0.38^k$. This is true under the assumption that the triplets seen more than $k$ times are independently generated. By only selecting triplets that are seen at least 3 times we achieve 5% error rate. We implicitly use even higher values for $k$ when sufficient data is available which was explained in section 3.

**Validation 2.** We have analysed by hand 75 triplets coming from the 2012 US Election campaign, and checked how many were actually present in the articles that were indicated by our pipeline as supporting them. 72 out of 75 were actually present in the article achieving 96% precision. This gives us a clear estimate of precision after our filtering step, but no estimation of recall, which we expect to be low.

**Validation 3.** In the following Sections we will describe two experiments, one of which identifies the key entities in US elections data for the past years by applying spectral analysis to the resulting network of entities, the experiment produces a ranking of all entities from the left to the right of the political spectrum. We observe by hand that in each case the two candidates are maximally separated (an event that would be very improbable by chance). We have therefore run a non-parametric statistical test (Siegel 1957) based on directly sampling the distribution rather than introducing assumptions like in a student's t-test. The details in designing this statistical test and the p-value computations are reported in Section 7. Here we also show the effect on p-values for different thresholds to the percentage of interval in Equation 4 which was discussed earlier and prove that we remove noise and keep signal by applying our filtering step.

**Remarks.** Finally we can corroborate our findings by identifying in the literature the performance rates of the main modules that we have deployed. This would allow us to have confidence in our pipeline. The precision and recall results are on average 96% and 93% for the ANNIE orthomatcher (co-reference resolution module) and 66% and 46% for the ANNIE pronominal anaphora resolution module (Bontcheva *et al.* 2002). An evaluation with the Susanne corpus shows that MINIPAR is able to achieve about 89% precision and 79% recall (Lin 1998).

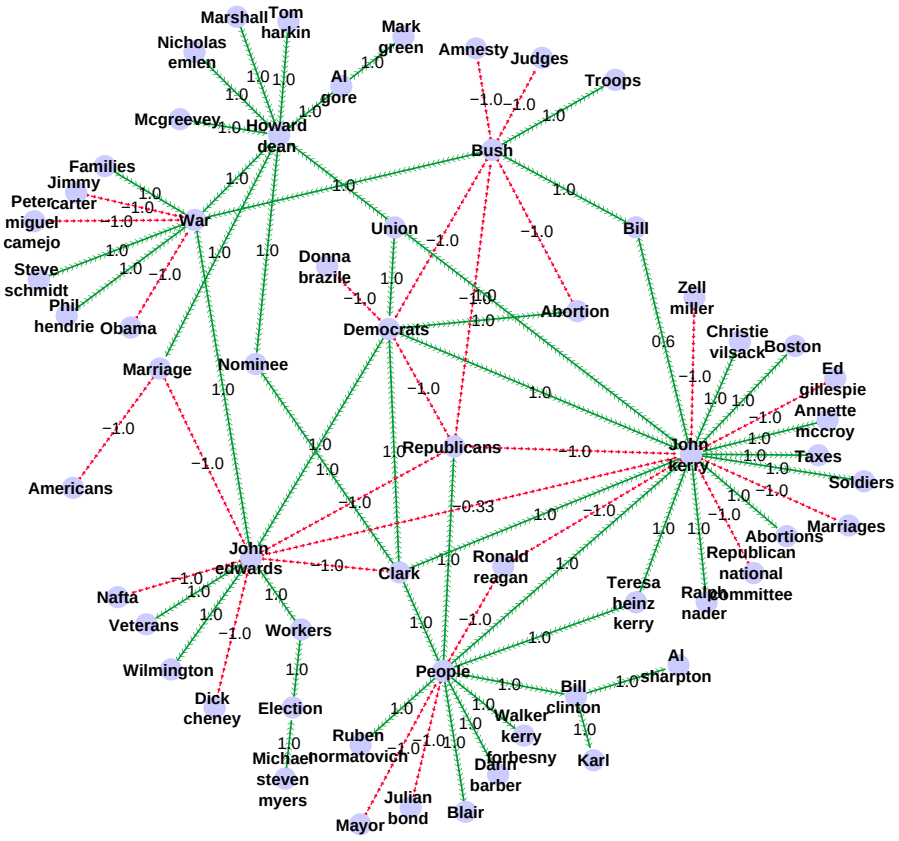
## 6 Experiment 1: Analysis of US Elections

We present here the results on experiments done with the past six (1988-2008) US Presidential election data from the New York Times corpus (Sandhaus 2008) and also with 200,000 articles on the 2012 US elections data obtained from our News Outlets Analysis and Monitoring System (NOAM) (Flaounas *et al.* 2011). Experiments were done separately on data from January to August (during primaries) and from August to September (after the conventions). For this experiment we define key entities as those that were most mentioned in this domain instead of comparing the relevance of actors with a background corpus. This was because entities in elections are also key entities in many other domains (e.g. Obama). Hence we used their absolute frequencies and selected the top 100 most frequent entities as key entities in this domain. Then we filter triplets that contain the key entities and actions that denote positive/negative attitudes using our verb lists.

Prior to this if there is a negation preceding a verb in a triplet like "Romney not support cuts", the not support is replaced with the verb "oppose". Again if there is "not oppose" in a triplet it is replaced with the verb "support". Using our weighting method in Equation 2 we assigned positive and negative weights to the links between key actors denoting the strength of friendship/hostile relations between them. From this we were able to create endorsement networks where nodes represent actors/objects and edges represent positive/negative attitudes between them.

Figure 3 and 4 show the endorsement networks obtained from year 2004 U.S Presidential election data from January to August and August to November. We observed that in each year there were many hubs representing candidates campaign-

Fig. 3: Network with Positive and Negative edges between Entities (U.S. Presidential election Data: January to August 2004)

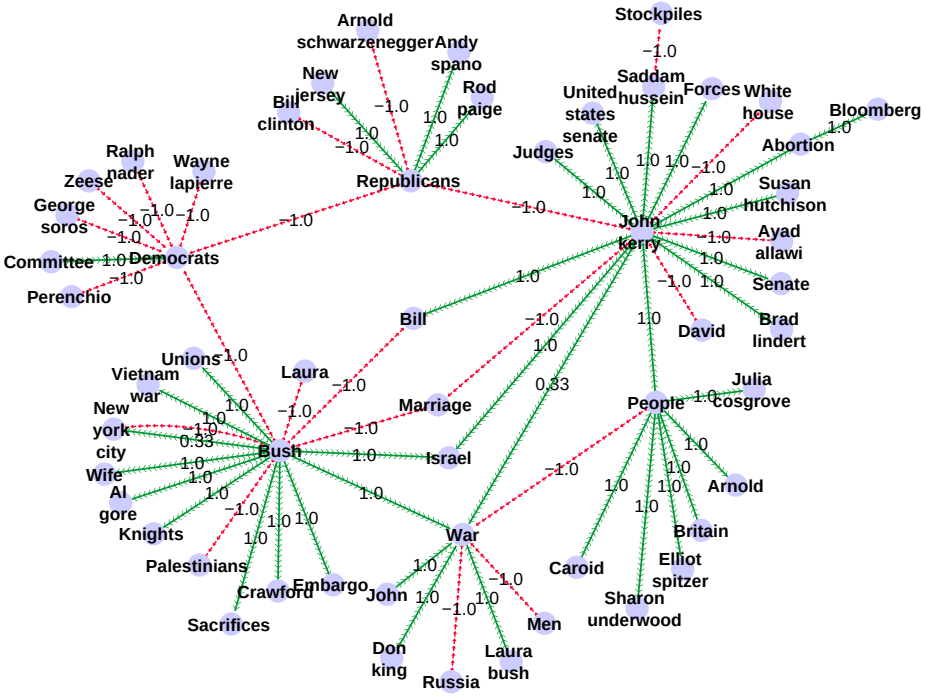



ing in different states in the network for the period of January - August while there were only two main hubs during August-November showing the two main opposing candidates from the Republicans and Democrats.

### *6.1 Spectral Graph Partitioning*

We applied our spectral graph partitioning technique to the networks obtained in the previous election cycles after the conventions. The output was two lists of actors ordered according to the 1st and 2nd eigenvectors. With regard to party associations we observed that the first eigenvector ordering of the vertices during the period of August-November gave more accurate results than the second eigenvector for all the years except year 2004. Table 1 shows a smaller version of the lists obtained for year 2000, 2004 and 2008 after removing the actors/objects in the middle of the list. The full lists obtained from 1988 to 2008 during August to November are shown in Table 6 and 7 in the appendix. Here we could see the two main opposing candidates in the top and bottom sides of the list representing the "Democrats"

Fig. 4: Network with Positive and Negative edges between Entities (U.S. Presidential election Data: August to November 2004))




and "Republicans". It is also interesting to see topics like "Abortion" and "War" take sides, with "Abortion" being more associated with the Democrats and "War" with the Republicans.

### *6.2 Plotting eigenvectors and subject/object bias of Entities*

To exploit the information coming from the eigenvectors we tried plotting the first and second eigenvectors in a two-dimensional scatter plot to see the actual positions of entities (actors/objects) in the eigenvector space. Since there are many campaigns during the primaries we dont expect to see a clear separation of entities showing their association to a party like what we got after the conventions. But still it is interesting to visualise the entities in the eigenvector space. Figure 5 shows the plot obtained for year 2004 from January to August during primaries. Figure 6 is a zoomed-in version of the lower right corner of Figure 4. The distances between entities in terms of eigenvectors explain the relationship between them. The more the distance is, the more likely that they were opposing each other at some point. Figure 12 in the appendix illustrates the plot obtained for year 2008.

Table 1: Lists of entities (actors/objects) showing party association identified in the U.S.Presidential Election data according to 1st/2nd Eigenvector cuts from 2000 - 2012

| 2000 | 2004 | 2008 | 2012 |
|---|---|---|---|
| Al gore | Democrats | Obama | Obama |
| Democrats | John kerry | Democrats | Clinton |
| Abortion | Bill | People | Democrats |
| Unions | People | Christ | Voters |
| Marriage | Palestinians | Senate | Majority |
| Government | Laura | Camp | Crowds |
| John robert | Marriage | Reasoning | Overhaul |
| National endowments | Committee | Bill | Marriage |
| Georgie yin | Russia | Drilling | Abortion |
| Protecting the earth | Men | Range | Taxes |
| ... | ... | ... | ... |
| ... | ... | ... | ... |
| ... | ... | ... | ... |
| Mcclellan | Saddam hussein | Bridge | Family Research Council |
| Blacks | United states senate | Project | Cuts |
| Dingell | Forces | Bombings | Conservatives |
| Amnesty | Ralph nader | Republicans | United States |
| Clarence thomas | George soros | Surge | Israel |
| Ralph nader | Perenchio | Mccain | Mccain |
| Pharmaceutical | Israel | Sarah palin | Governor |
| Vietnam war | Al gore | John maccain | Ryan |
| People | Unions | | Republicans |
| Son | Knights | | Romney |
| Republicans | Crawford | | |
| Bush | Embargo | | |
| | Wife | | |
| | Vietnam war | | |
| | Republicans | | |
| | War | | |
| | Bush | | |

We also plotted the subject/object bias of entities against their eigenvector space. We assigned a subject/object bias score for each entity in the eigenvector space according to Equation 5. In this way a positive score indicates subject bias and negative score indicates object bias. Figure 7 shows the scatter plot obtained for year 2004 where entities are plotted against their 2nd largest eigenvector and sub-

Fig. 5: Eigenvector 1 vs Eigenvector 2 of entities in 2004 (January - August)

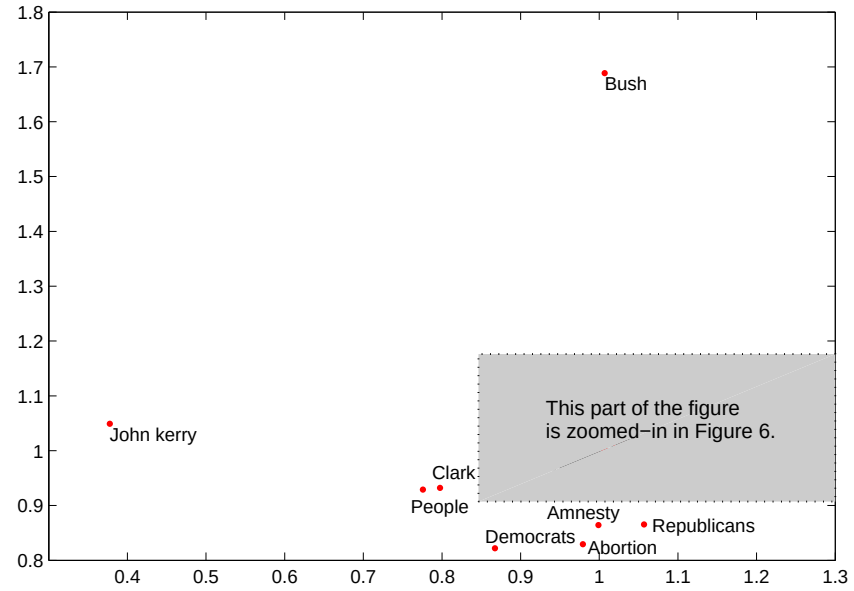



Fig. 6: Zoomed in version of the lower right hand corner of Figure 3

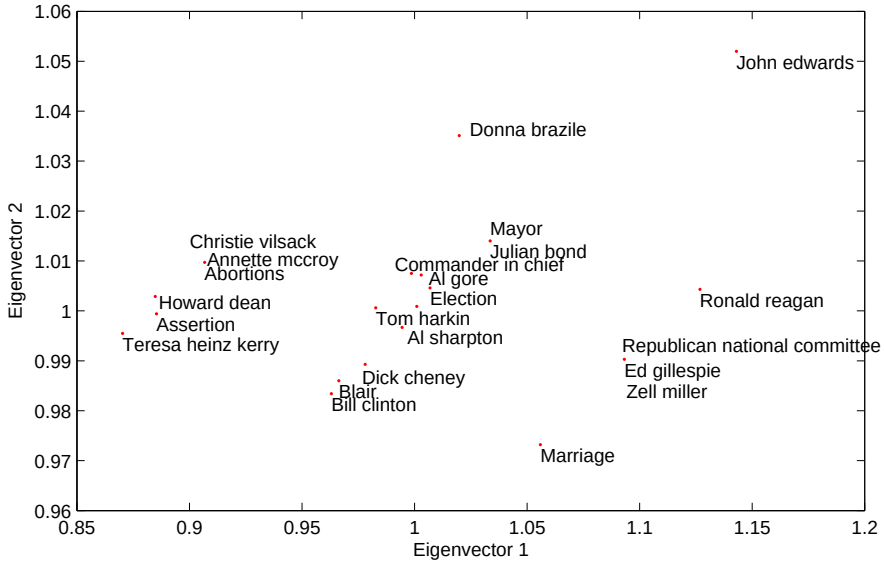
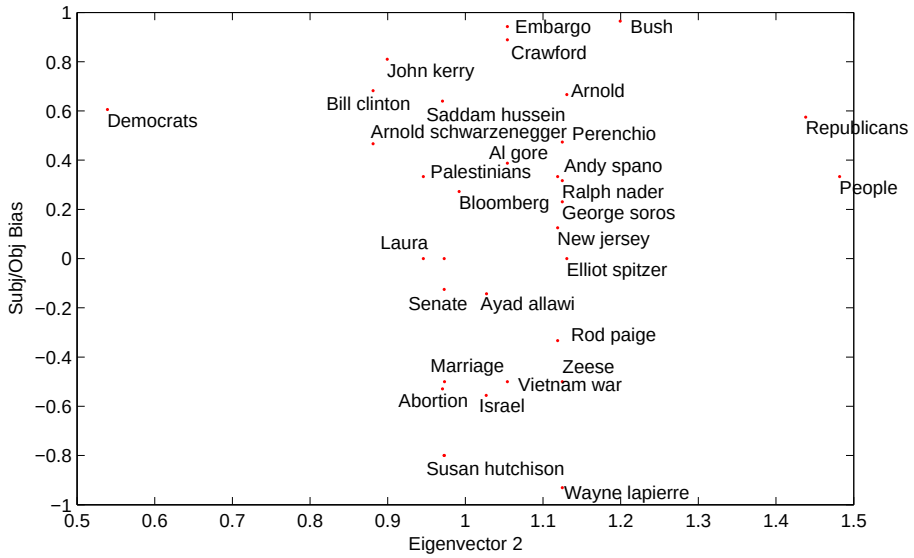


Fig. 7: Eigenvector 2 (vs) Subject/Object Bias of entities in 2004 (August- November)




ject/object bias scores. We plot the 2nd eigenvector for 2004 since it gave much cleaner ordering of entities compared to the 1st eigenvector.

What we observed here is that topics like "Vietnam War", "Marriage" and "Abortion" are most often mentioned as objects while named entities are often subjects. Figure 13 in the Appendix shows the scatter plot obtained for year 2008.

## 7 Validation of the Entity Spectrum

In designing a statistical test, some design choices must be made arbitrarily and upfront. The most notable ones are of course the choice of null hypothesis and the choice of test statistic. Our test statistic was intended to measure the extent to which our analysis captures the division into two camps of the US political actors/objects. We were interested in making our choices as objective and as general as possible, so we did not want to arbitrarily pick and choose specific actors by hand, or assign them to a political part, for each election cycle. This would also create issues with words such as "President", or "Senate", which might change political leaning in different cycles. Instead we settled on the obvious choice: the two political parties ("Democrats" and "Republicans") are mentioned in each election cycle, and we decided to use the "distance" between them as a test statistic. This initial design choice allows us to design a rigorous statistical test in a way that contains no subjective choices on our behalf.

Permutation testing (also known as randomisation test, or exact test) (Good

2005)(William 1990) is a central part of non-parametric statistics. It directly obtains the distribution of the test statistic under the null hypothesis by calculating all possible values of the test statistic under rearrangements of the treatments (labels) on the observed data points. This removes the need to know the analytical form of this distribution, as done in parametric testing, and hence to apply rigorous statistical testing to situations where an analytical form of the distribution is not available. The availability of high computing power is making non-parametric testing standard in many modern applications. An early example of permutation testing is Fisher's exact test, more recent examples include bootstrapping and jackknifing.

The basic idea of all randomisation tests is to use the null-hypothesis that all treatments (labels) are interchangeable, by measuring the value of the test statistic under (ideally) all possible permutations. In practice a large sample of random permutations is used. The one-sided pvalue of the test is calculated as the proportion of sampled permutations where the test statistic is greater than or equal to that in the original dataset.

We obtain the eigenvalue distance $d$ between the "Republicans" and the "Democrats" in the entity list. We compare it with distance $d_1$ obtained by taking the distance between the same actors from 100 randomised networks. Here we use two random network models, Erdös-Rényi and Random re-wiring to generate the random networks.

In the Erdös-Rényi (Erdös and Rényi 1960) model, all edges are removed from the network first. Each pair of the nodes is connected with an edge at random where the edge is chosen uniformly from the set of removed edges. Here the degrees of nodes are not preserved. In Random rewiring we randomly reshuffle links, while keeping the in and out degree of each node constant. A convenient numerical algorithm performing such randomization consists of first randomly selecting a pair of directed edges A→B and C→D. The two edges are then rewired in such a way that A becomes connected to D, while C connects to B (Sergei and Kim 2002). We do this rewiring $m.10$ times for creating each random network where $m$ is the number of edges existing in the graph.

We check for the number of times $r$ that $d_1 \geq d$ in the 100 $r$ random networks to calculate the pvalue. We expect

$$pvalue = \frac{r}{100} \tag{8}$$

Table 2 shows the resulting pvalues obtained for experiments performed on 2012 and the previous election cycles. It shows that the pvalue is very low for the same signal appearing by chance.

We also checked the effect on pvalues when different thresholds are introduced to the percentage of interval $i$ in our weighting Equation 4 which is $\frac{2\sqrt{\frac{P.N}{P+N}+1}}{P+N+4}$. The selection of the optimum threshold for $i$ would be based on the following. It should produce a pvalue less than 0.01, contain atleast 100 nodes in the network and the relations in the network should be supported by many number of positive/negative triplets (we report the average number of positive ($k_p$) and negative ($k_n$) triplets

Table 2: pvalues for distance $d_1 \geq d$ over 100 random networks according to two different random graph models

| Year | pvalue(Random-rewiring) | pvalue(Erdös-Rényi) |
|---|---|---|
| 2012 | 0 | 0 |
| 2008 | 0 | 0 |
| 2004 | 0.01 | 0 |
| 2000 | 0.05 | 0.01 |
| 1996 | 0.06 | 0 |
| 1992 | 0.05 | 0 |
| 1988 | 0 | 0 |

Table 3: pvalues for distance $d_1 \geq d$ over 100 random networks according to two different random graph models

| Threshold($i$) | No of nodes($n$) | Avg No of Pos Triplets($k_p$) | Avg No of Neg Triplets($k_n$) | pvalue (Random-rewiring) | pvalue (Erdös-Rényi) |
|---|---|---|---|---|---|
| <9% | 70 | 44 | 213 | 0 | 0 |
| <10% | 80 | 42 | 193 | 0 | 0 |
| <12% | 131 | 37 | 155 | 0.03 | 0.01 |
| <13% | 150 | 35 | 146 | 0 | 0 |
| <15% | 188 | 31 | 127 | 0 | 0 |
| <17% | 269 | 28 | 112 | 0 | 0 |
| <20% | 298 | 27 | 105 | 0 | 0 |
| <23% | 421 | 23 | 91 | 0 | 0 |
| <29% | 656 | 21 | 77 | 0 | 0 |
| <35% | 1195 | 17 | 63 | 0.36 | 0.09 |

obtained per relation in the network). Table 3 shows the results obtained for different thresholds of $i$. According to our selection criteria the optimum threshold for $i$ is 13%. But its interesting to see that upto 29% for $i$ the networks which are larger still produce perfect entity spectrums.

## 8 Experiment 2: Analysis of Crime Stories

In this experiment we applied our pipeline to the analysis of nearly 100,000 crime-related stories that appeared in the New York Times corpus between 1987 and 2007 (Sandhaus 2008). Our pipeline identified the key entities and actions, by weighting the entities in crime stories against their frequency in a background corpus Top News (280K articles) according to Equation 1. We selected the top 300 ranking candidates as "key entities and key actions". This threshold was based on the number of triplets extracted which contained the key entities and actions for network analysis. The higher the threshold is the higher the number of extracted triplets. We decided on this threshold value since we wanted to have a compact network with the most important information only.

Here we present results from experiments performed on crime data in year 2002. Figure 8 shows the top 20 key subjects, objects and actions in Crime in 2002 ranked according to their weights. When examined carefully we see that the application exposes a critical crime story that occurred during that year. Sexual abuse scandal in Boston archdiocese was a major chapter in the crime news in early 2002. Actors like "Diocese", "Detectives", "Archdiocese", "Cardinals", "Bishops" and actions such as "Molest", "Plead" and "Abuse" reveal that.

In order to create networks in Crime we filtered only the triplets that contained the "key entities" and "key actions". The networks created had subjects and objects as nodes and the verbs linking them as edges. Every relation in the network had a direction from the subject to object. Figure 9 illustrates a sub network for year 2002 which highlights the interactions particularly between the subject "Priest" and other objects in the whole network. By analysing the properties of these kind of networks we can identify the most central entities in a given corpus.

### *8.1 Measures of Importance*

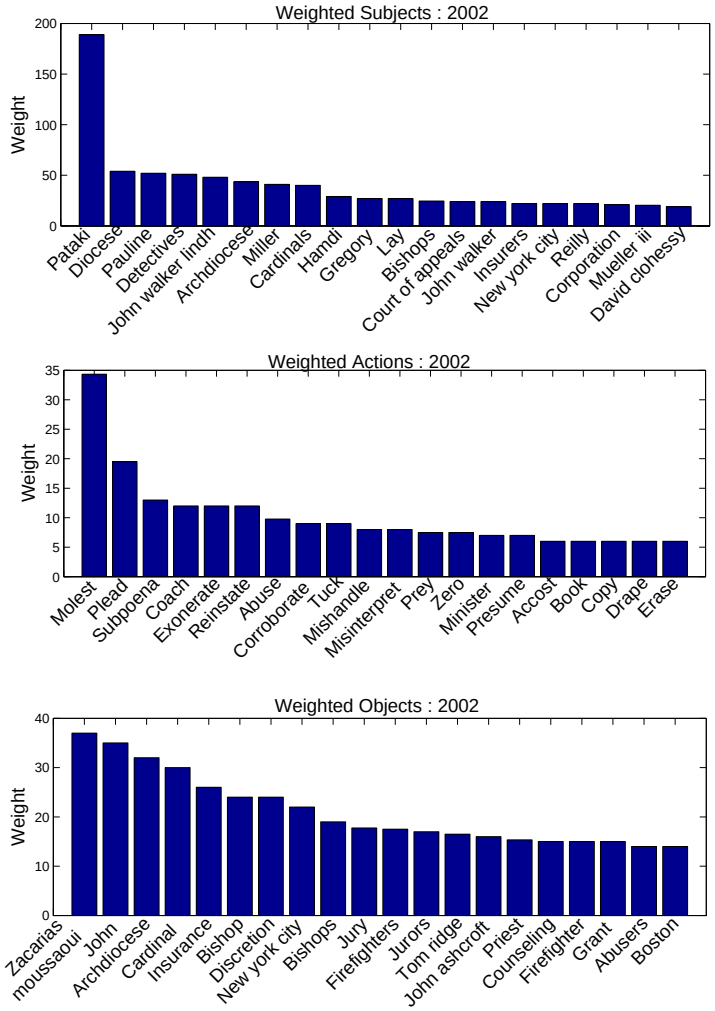
In order to identify the central entities in crime we ranked all entities according to various network centrality measures like Betweenness Centrality, In-Degree, Out-Degree, HITS(Hyperlink-Induced Topic Search, also known as hubs and authorities) and PageRank.

Table 4 shows the top 10 ranked entities for each network measure computed in Crime data for 2002. It shows actors like "Priest", "Archdiocese", "Prosecutors", "Zacari" and objects such as "Law" and "Cases" have been most central in the data, reflecting the leading crime story of that year in the US.

### *8.2 Measures of Importance over Time*

To detect changes of roles of entities and actions in crime over the 20 years we performed an analysis for each key entity by looking at how their centrality measures vary over time. We discovered that network measures like Out-Degree and Hub picked up the most central and interesting entities out of the data. Hence we used them and the frequency count of each entity to perform the analysis. Figure 10

Fig. 8: The top ranked key subjects, objects and actions in Crime 2002



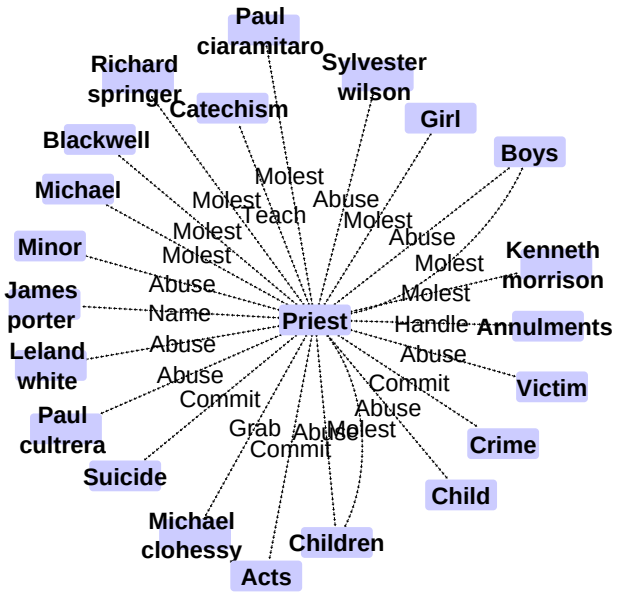

shows the time series graphs for Archdiocese plotted against its Frequency, Out-Degree and Hub values and actions Molest, Plead and Abuse plotted against their frequencies in 20 years. It clearly demonstrates that there has been a peak in all these measures during 2002 when the news stated a lot about the involvement of the "Priest" and "Archdiocese" in the Boston sexual scandal.

### *8.3 Verb Types*

We considered the roles different entities play in crime by classifying verbs into two different types also known as action spheres, such as "Crime against Person" and "Crime against Property". Here are some examples of verbs in these categories.

Crime against Person: Murder, Kill, Torture, Rape, Assault
Crime against Property: Steal, Extort, Rob, Embezzle, Confiscate

Fig. 9: Interactions between the entity 'Priest' and other entities in the network

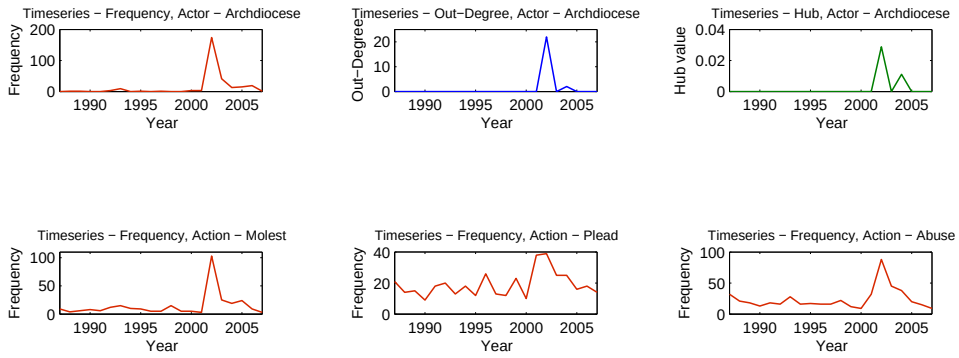



For each type we filtered triplets containing actions related to the type and visualised them in a network. We then ranked the subjects and objects found in the filtered triplets according to their frequencies to find the highly ranked entities in these two types of crimes.

Table 4: Top 10 ranked entities according to Network Centrality measures for Crime data in 2002

| Betweeness Centrality | In-Degree | Out-Degree | Hub | Authority | PageRank |
|---|---|---|---|---|---|
| Law | Cases | Priest | Law | Cases | Cases |
| Archdiocese | Case | Judge | Archdiocese | Case | Court |
| Complaint | Letter | Law | Priests | Letter | Lawsuit |
| Suit | Allegations | Prosecutors | Suit | Questions | Anyone |
| Jurors | Boys | Jury | Abuse | Allegations | Nothing |
| Prosecutors | Child | Lawyers | Firm | Acts | Law |
| Diocese | Questions | Priests | Bishop | Law | Properties |
| Priests | Accusations | Archdiocese | Scandal | Suit | Play |
| Lawyers | Children | Church | Complaint | Nothing | Sorts |
| City | Law | Department | Diocese | Boys | Dying |

Fig. 10: Time series graphs for actors "Archdiocese", "Priest" and actions "molest", "plead" and "abuse"




The top 10 ranked (based on frequency) subjects and objects involved in crime in 2002 against person and against property are shown in Table 5. We found that "Men" are most commonly responsible for crimes against person, while "Women" and "Children" are most often victims of those crimes.

It is also encouraging to see that all key objects of crime against person are indeed persons, and similarly most key objects of crimes against property are indeed non persons. The subjects are nearly all persons, with the exceptions of a few organisations. All this provides an extra reliability check.

Table 5: Top10 ranked subjects and objects in crime against person and against property in 2002

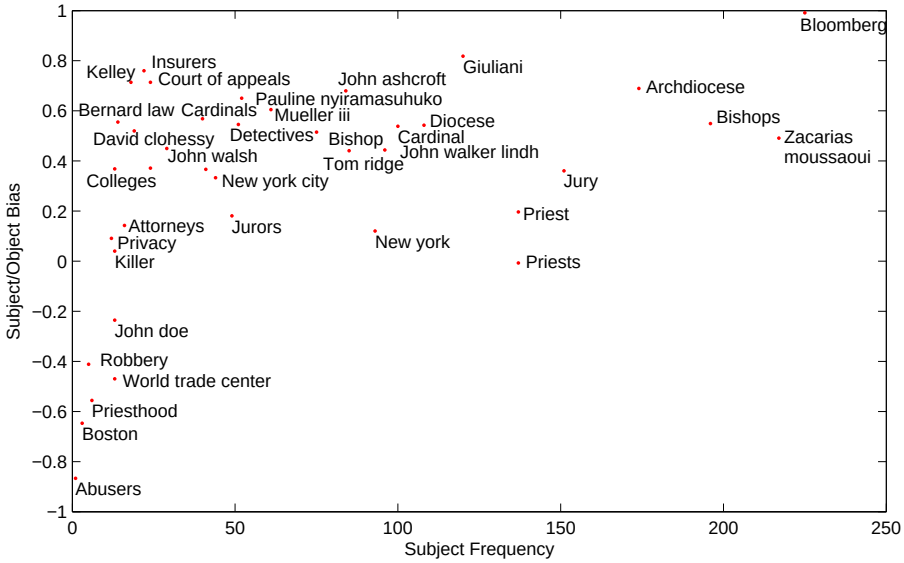
| **Crime against Person** | | **Crime against Property** | |
|---|---|---|---|
| Subject | Objects | Subjects | Objects |
| Priest | People | Man | Money |
| Man | Boy | Police | Bank |
| Troops | Child | Soldiers | Records |
| Reyes | Girl | Winona Ryder | Millions |
| Geoghan | Man | Priest | Weapons |
| Shanley | Woman | People | Wallet |
| Forces | Jogger | Jason Bogle | Trade Secret |
| Police | Victim | Investigators | Steven Seagal |
| United States | Minors | Employee | Most |
| Others | Me | Agents | Man |

### *8.4 Subject/Object Bias of Entities in Crime*

For crime stories we again compute the subject/object bias using Equation 5. Figure 11 illustrates the subject/object bias of entities in crime for year 2002 against their subject frequencies in a 2-dimensional scatter plot.

We find that "Archdiocese", "Bishops" are very subjective with a very high frequency and "World trade center", "Priesthood" and "Abusers" were very objective in that year. Generally we see all the named entities on the subjective side.

Fig. 11: Scatter Plot showing the subject, object bias in data for year 2002. For ease of visualisation we removed NY Governor Pataki from set, as it had a very high subject bias




## 9 Conclusions

The task of extracting narrative information from a corpus has applications in many domains. This information includes the identification of the key entites in a narration, the key actions that are narrated, and the overall relational structure among them. It can be applied to the analysis of the political relations among political actors, as we saw in our study of the US Elections, or in the extraction of information from historical text, or from literary text, among other things.

We have presented a method to automate the creation of large networks of entities, testing their validity, and analysing properties of the underlying text. We can for example identify the most central entities, those who tend to be subjects or objects, and the relations among them.

We have also presented a method to map actions to action-types, by making use of verb lists. This greatly simplifies the networks by only allowing for few types of edges, as was the case for the network of political support or the network of crime in our experiments. Future work will focus on distinguishing actors from objects.

The contribution of this study is in the developement of a new methodology for the extraction of knowledge from a large corpus and not in the improvement of tools for the processing of language. Among various sanity checks we have performed, we have seen that our method always correctly separates the two candidates and the two parties in US election data, and correctly identifies people as objects of crimes against person (as opposed to crimes against property, for example). Among potentially interesting findings for social investigation, we have seen that this network identified men as frequent perpetrators, and women and children as victims, of violent crime, a finding that might have relevance for social sciences.

More generally, we believe that this method can automate the labour intensive "coding" part of the task of Quantitative Narrative Analysis and of Distant Reading among other tasks, and therefore have relevance in the social sciences and the humanities.

## References


Bastian M., Heymann S. and Jacomy M. (2009) *Gephi an open source software for exploring and manipulating networks.* In International AAAI Conference on Weblogs and Social Media, San Jose, California

Bontcheva K., Dimitrov M., Maynard D., Tablan V. and Cunningham H. *Shallow Methods for Named Entity Coreference Resolution.* In Workshop TALN 2002, Nancy, France.

Brin S. and Page L. *The anatomy of a large-scale hypertextual (web) search engine.* In Seventh International World Wide Web Conference.

Chen H., Chung W., Xu J., Wang G., Qin Y. and Chau M. *Crime data mining: a general framework and some examples* IEEE Computer, 37 (4) (2004), pp. 5056

Cristianini N. (2011) Automatic discovery of patterns in media content. In *22nd Annual Symposium on Combinatorial Pattern Matching*, Lecture Notes in Computer Science, 6661, Springer 2011, Palermo, Italy

Cunningham H. (2002) *GATE, a General Architecture for Text Engineering* In Computer and the Humanties, SpringerLink 36:223–254

Dali L. and Fortuna B. (2008) *Triplet extraction from sentences using svm.* In Conference on Data Mining and Data Warehouses (SiKDD), Las Vegas, USA

Dali L., Rusu D., Fortuna B., Mladenic D. and Grobelnik M. (2009) *Question Ansewering Based on Semantic Graphs.* In 18th International World Wide Web Conference, Madrid, Spain

De Fazio G. (2012) *Political radicalization in the making: The civil rights movement in northern ireland,1968-1972.* PhD thesis, Department of Sociology, Emory University, Atlanta, Georgia

Earl J., Martin A., McCarthy J. and Soule S. (2004) *The use of newspaper data in the study of collective action.* In Annual Review of Sociology 30:65-80

Erdös P. and Rényi A. (1960). *On the evolution of random graphs.* Mathematical Institute of the Hungarian Academy of Sciences 5: 1761.

Flaounas I., Ali O., Turchi M., Snowsill T., Nicart F., Tijl D.B. and Cristianini N. (2011) *Noam: News outlets analysis and monitoring system.* In ACM SIGMOD International Conference on Management of Data, Athens, Greece

Franzosi R. (1987) *The press as a source of socio-historical data: Issues in the methodology of data collection from newspapers.* In Historical Methods 20:5–16

Franzosi R. (1998) *Narrative as data. linguistic and statistical tools for the quantitative study of historical events.* In New methods in Historical Sociology/Social History Special issue of International Review of Social History 43:81–104

Franzosi R. (2010) *Quantitative narrative analysis.* Sage Publications Inc, Quantitative Applications in the Social Sciences 162:200

Good P. (2005) *Permutation, Parametric, and Bootstrap Tests of Hypotheses* 3rd edition. New York: Springer Series in Statistics.

Kipper K., Korhonen A., Ryant N. and Palmer M. (2006) *Extensive classifications of english verbs.* In 12th EURALEX International Congress, Turin, Italy

Kleinberg J.(1998) *Authoritative sources in a hyperlinked environment.* In 9th ACM-SIAM Symposium on Discrete Algorithms, San Francisco, California

Kunegis J., Schmidt S., Lommatzsch A., Lerner J., De Luca E. and Albayrak S. (2010) Spectral analysis of signed graphs for clustering, prediction and visualization. In SIAM International Conference on Data Mining, Columbus, Ohio, USA

Leskovec J., Huttenlocher D. and Kleinberg J. (2010) *Signed networks in social media.* In 28th CHI(ACM Conference on Human Factors in Computing Systems), Atlanta, Georgia, USA

Lin D. (1998) *Dependency-based evaluation of minipar.* In Workshop on the Evaluation of Parsing Systems, Granada, Spain

Mihalcea R., Radev D. (2011) *Graph-Based Natural Language Processing and Information Retrieval.* In Cambridge University Press

Mitkov R. (1999) *Anaphora resolution: The state of the art.* Tech. rep., School of Languages and European Studies, University of Wolverhampton, UK

Moretti F.(2011) *Network theory, plot analysis.* In New Left Review. 68:80–102

Padraig M. C. and Ralph K. *Universal properties of mythological networks.* In Europhysics Letters 99:28002

Rusu D., Dali L., Fortuna B., Grobelnik M. and Mladenic D. (2007) *Triplet extraction from sentences.* In10th International Multiconference Information Society - IS 2007, Ljubljana, Slovenia

Rusu D., Fortuna B., Grobelnik M. and Mladenic D. (2008) *Semantic graphs derived from triplets with application in document summarization.* In Conference on Data Mining and Data Warehouses (SiKDD), Las Vegas, USA

Sandhaus E. (2008) The new york times annotated corpus. New York Times

Shannon P., Markiel A., Ozier O., Baliga N., Wang J., Ramage D., Amin N., Schwikowski B. and Ideker T. (2003) *Cytoscape: A software environment for integrated models of biomolecular interaction networks.* In Genome Research 13:2498–2504

Sergei M. and Kim S. *Specificity and Stability in Topology of Protein Networks.* In Science 296(5569):910–913

Seigel S. (1957) *Nonparametric Statistics* In The American Statistician 11(3):13–19

Soon W., Ng H. and Lim D. (2001) *A machine learning approach to coreference resolution of noun phrases.* In Computational Linguistics 27:521–544

Trampus M. and Mladenic D. (2011) *Learning event patterns from text.* In Informatica 35

William J.W. (1990) *Construction of Permutation Tests* In Journal of the American Statistical Association. 85:693–698

Wilson, E.B. (1927). *Probable inference, the law of succession, and statistical inference.* In Journal of the American Statistical Association, 22:209212.

Yang B., Cheung W. and Liu J. (2007) *Community mining from signed social networks.* In IEEE Transactions on Knowledge and Data Engineering 19:10.

Table 6: Lists of entities(actors/objects) showing party association identified in the U.S.Presidential Election data according to 1st/2nd Eigenvector cuts from 1988 - 1996

| 1988 | 1992 | 1996 |
|---|---|---|
| Bush | Bill clinton | Bob dole |
| Policies | Democrats | Reagan |
| Republicans | Ross perot | People |
| Reagan | Persian gulf war | Bush |
| White house | Brady bill | Mayor |
| Secretary | Barbara bush | Charles vaughn |
| Bob dole | Victor morrone | Dave winkler |
| Slade gorton | Students | Darlene stermer |
| State department | Robert abrams | Bill knapp |
| Republicans administration | Russell feingold | Lucy smith |
| School | Military | Amendment |
| Attacks | Perry | John sakelaris |
| Tax | Laurie pawlowski | Reuven frank |
| Senate | People | Sandra eash |
| Judith lichtman | Media | Mario rizzo |
| Civil rights | Civil rights act | Michelle carr |
| Electoral college | Dean alger | Bryant |
| Lloyd bentsen | Paula zahn | Smith |
| Dan quayle | Clinton presidency | Republicans |
| Spencer tracy | Burt monroe | Jack kemp |
| Lowell | Jim maser | Liberal president clinton |
| Bill | Bob packwood | Roger clinton |
| Abortion | Abortion | Reliance |
| Democrats | Hillary clinton | Presidential debate commission |
| Dukakis | Diane english | Lamm |
| | War | Charlotte morrisom |
| | Americans | Derrick rhamad |
| | America | Steve forbes |
| | Dan quayle | Scott reed |
| | Republicans | Blawenburg |
| | Buchanan | Philbrook |
| | Newt gingrich | Wilkinson |
| | Fred mosley | Westbrook |
| | Jorge mas | Daniel kovalik |
| | Edward habecker | Betsy |
| | Homosexuality | Cuomo |
| | Vietnam war | Beth vogl |
| | Jack colhoun | Mckinley |
| | Michel | Ross perot |
| | White house | Clinton administration |
| | Bush | Media |
| | | Democrats |
| | | Bill clinton |

Table 7: Lists of entities(actors/objects) showing party association identified in the U.S.Presidential Election data according to 1st/2nd Eigenvector cuts from 2000 - 2012

| 2000 | 2004 | 2008 | 2012 |
|---|---|---|---|
| Algore | Democrats | Obama | Obama |
| Democrats | John kerry | Democrat | Clinton |
| Abortion | Bill | People | Democrats |
| Unions | People | Christ | Voters |
| Marriage | Palestinians | Senate | Majority |
| Government | Laura | Camp | Crowds |
| John robert | Marriage | Reasoning | Overhaul |
| National endowments | Committee | Bill | Marriage |
| Georgie yin | Russia | Drilling | Abortion |
| Protecting the earth | Men | Range | Taxes |
| Bill | Abortion | Barack | Vice President |
| Military | Saddam hussein | Bridge | People |
| Fidel castro | United states senate | Project | White House |
| Rendell | Forces | Bombings | Campaign |
| Ann mcfall | Judges | Republicans | Investments |
| Ross perot | Susan hutchison | Surge | Family Research Council |
| Lieberman | Brad lindert | Mccain | Cuts |
| Vicki simon | Senate | Sarah palin | Conservatives |
| Bipartisanship | Bill clinton | John mccain | United States |
| Ann hazlet | Arnold schwarzenegger | | Israel |
| Robert alphin | Arnold | | Mccain |
| Ellen burt | Elliot spitzer | | Governor |
| Carmen obando | Sharon underwood | | Ryan |
| Countries | Caroid | | Republicans |
| Colorado | Britain | | Romney |
| Dick cheney | Julia cosgrove | | |
| Mcclellan | Bllomberg | | |
| Blacks | Stockpiles | | |
| Dingell | New jersey | | |
| Amnesty | Andy spano | | |
| Clarence thomas | Rod paige | | |
| Ralph nader | Ayad allawi | | |
| Pharmaceutical | David | | |
| Vietnam war | White house | | |
| People | New york city | | |
| Son | John | | |
| Republicans | Laura bush | | |
| Bush | Don king | | |
| | Wayne lapierre | | |
| | Zeese | | |
| | Ralph nader | | |
| | George soros | | |
| | Perenchio | | |
| | Israel | | |
| | Al gore | | |
| | Unions | | |
| | Knights | | |
| | Crawford | | |
| | Embargo | | |
| | Wife | | |
| | Vietnam war | | |
| | Republicans | | |
| | War | | |
| | Bush | | |

Fig. 12: Eigenvector 1 vs Eigenvector 2 of entities in 2008 (January- August)

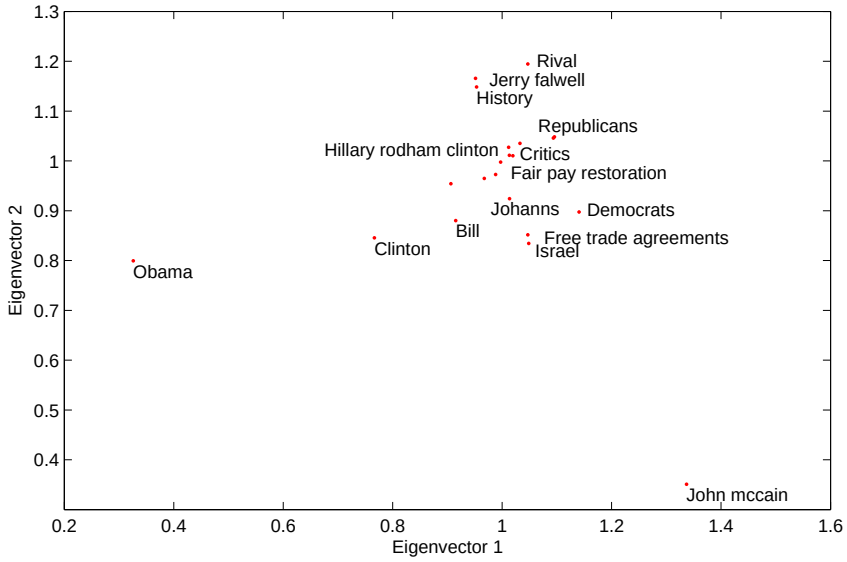



Fig. 13: Eigenvector 1 vs Subject/Object Bias of entities in 2008 (August- November)