Master DMKM Report

Mining parliamentary data and news articles to find patterns of collaboration between politicians and third party actors

Francisco Andrés RODRÍGUEZ DRUMOND

defended the July 7th, 2014

Supervision : Marta Arias, LARCA-UPC,
Josep Larriba, DAMA-UPC

Location : Universitat Politècnica de Catalunya,
Barcelona, Spain.

Abstract:

LOPEN IPSUM

Résumé :

LOPEN IPSUM

Contents

Hosting institution

This Master's thesis was done with the support and supervision of the DAMA and LARCA Research Groups at Universitat Politècnica de Catalunya (UPC).

DAta MAnagement group - DAMA

DAMA-UPC is part of the Computer Architecture Department (DAC). The main research topics of DAMA-UPC are oriented to performance, exploration and quality in data management, focusing particularly on large data volumes. Specifically, they have investigated the creation of new data structures, algorithms, methods and applications in the area of Data Management that make it easier to manipulate large amounts of data.

DAMA-UPC is a member of Tecnio since 2005. Tecnio is an initiative of ACC10, the Agency for Innovation and Internationalization of the Catalan Enterprise, belonging to Generalitat de Catalunya. It is supported by Generalitat de Catalunya as a Consolidated Research Group (SGR-1187) and by the Ministry of Education and Science of Spain. Since its creation it has worked with important technological partners such as IBM, Oracle technologies, the Ministry of Cience and Innovation, the Health Department of the Generalitat de Catalunya, among others; and is current participating in four European Research Projects.

Moreover, DAMA-UPC has been a pioneer in scientific collaboration, creating and organizing for three consecutive years the Workshop on Graph-based Technologies and Applications (Graph-TA), organizing the 17th International Database Engineering & Applications Symposium (IDEAS 2013) and the First University-Industry Meeting on Graph Databases (UIM-GDB). The work of DAMA has also give birth to Sparksee, a graph database, and Sparsity Technologies.

Laboratory of Relational Algorithmics, Complexity and Learnability - LARCA

LARCA is an international research group working on data mining, machine learning, data analysis, and mathematical linguistics. They typically approach problems from sound mathematical principles, using modelling tools and techniques from algorithmics, computational complexity, automata theory, logic, discrete mathematics, statistics, and dynamic systems.

LARCA has also participated in several European research projects and collaborated with several industrial partners including Gas Natural Fenosa, Xopie, 4dLife, Urbiotica and Vingenia. They also foster collaboration with other research groups in and outside of Spain, including ALBCOM: Algorithms, Computational Biology, Complexity and Formal Methods Research Group at CS-UPC; NLP: Natural Language Processing Research Group at CS-UPC; IAIA: Investigacion y Aplicaciones en Inteligencia Artificial group at U. Málaga; CCG: Computational Complexity Group at U. Zaragoza; MIDAS: Spanish Network on Data Mining and Learning; Machine Learning Group at the University of Waikato, New Zealand and the Real and functional Analysis Research Group of Universitat de Barcelona. [1]

[1] Text partially taken from `www.dama.upc.edu` and `https://recerca.upc.edu/larca/`

Acknowledgement

LOREM IPSUM

1 Introduction

Political decision-making is characterized by the indirect participation of third-parties to attempt to influence the contents and outcome of the discussions according to their interests. This process, known as *lobbying*, is regularly done by companies, Non-Governmental Organizations (NGOs) and in some cases citizens and foreign governments. For instance, when a law concerning human rights is being discussed in a parliament we can expect human rights NGOs to contact politicians to attempt to push their agenda.

Most of the times citizens are completely unaware of the lobbying activities of their representatives. This is due to the fact that it is hard to closely monitor the activities of politicians, which in many cases are opaque to people, to understand the intricacies behind their decisions and to grasp the implications they have on organizations and society. Knowing how decision makers are influenced is however of great interest for politologists, journalists and voters in general. By tracking these relations we would be able to better understand the way decisions are made by a government, have a better sense of who politicians are, who they serve and what their agenda is, and in some cases fight corruption.

Among the different powers of the classical liberal democracies, the Legislative Branch is particularly interesting in the context of lobbying and policy making analysis. This is due to the fact that parliaments are responsible for the writing and passing of laws and regulations, the ratification of international treaties and the oversight of other branches of government. This makes them particularly influential in the shaping of a society and its institutions. Because of this, there has been considerable work in the development of data mining tools for understanding the way legislative bodies work.

Social Network Analysis (SNA) has been widely used to understand the dynamics of complex social systems. This makes SNA a natural tool for modelling the dynamics of power in a parliament. There is a wealth of knowledge to be extracted from parliamentary data to model relationships between politicians, identify key players and sub-communities, predict the voting of bills and in general gain a better understanding of the legislative process. However, finding political relations between representatives and third party actors, who often do not appear in bills or in the transcripts of debates, remains a challenge.

News articles on the other hand contain a sizable amount of information about what happens in a given country and about its relevant actors. Because of this, there is plenty of work by the scientific community to develop methods to automatically extract useful knowledge from news articles. The development of applications to automatically generate Social Networks (SN) from corpora of news articles is particularly interesting in our context. This has been done extensively in the past in different areas including defense, counter-terrorism, news summarization and crime prevention. The working assumption is that by detecting patterns of co-occurrence of two entities in text we can establish that they are in some way related.

In this study we used that assumption and formulated the hypothesis that by analyzing and combining parliamentary data and news articles we can generate SN for the members of a Legislative Body and entities that are related to these politicians. Our purpose is consequently to propose a method to automatically detect links descriptive of the political closeness of politicians and relevant actors of a country, or in other words, patterns indicative of a possible lobbying activity. To do this, we automatically analyze news articles, the text of bills discussed in a parliament and the transcripts of the debates. In this document we present our proposal and show the results obtained by its application in the context of the Parliament of Catalonia.

The rest of this document is organized as follows: in chapter 2 we present a more detailed definition of our problem. This is followed by chapter 3, in which we present the state of the art in the automatic generation of SN from corpora of text and in the use of SNA in the context of political science. In chapter 4 we describe our proposal and present some relevant considerations about its implementation. Next, we present the obtained results in section 5 by means of case studies. Finally, we present in chapter 6 the conclusions of our work along with some suggestions for further work.

2 Problem definition

Social Networks (SN) are a powerful tool for understanding complex social systems, in which relationships between actors play a central role. Because of this, Social Network Analysis has become increasingly popular in the Political Science community for studying the dynamics of power. For instance, the authors of [6] studied a graph of co-sponsorship of bills to study interactions between congressmen in the US House of Representatives and found that by using network analysis tools it was possible to find highly influential politicians. Similarly, in [8] the authors developed a theory of influence diffusion across a legislative network of relations based on weak and strong links and found patterns useful for determining the success of a bill.

The growing popularity of SNA in Politicial Science has given birth to Policy Network Analysis, a discipline concerned in the creation and analysis of SN involving political actors.

2.1 Policy Networks: Social Network Analysis for Political Science

Policy Network Analysis (PNA) is the discipline in Political Science which focuses on the discovery and analysis of links between government and other members of a society, with a particular interest in the understanding of the policy making process and its outcomes. There are plenty of definitions of Policy Networks (PN) in the literature, which is due to the fact that there are many ways to characterize relationships between actors of a society and several ways to approach the analysis. However the Oxford Handbook of Public Policy proposed a definition which is widely accepted and encompasses many of the other proposals. They state that a PN is a "set of formal institutional and informal linkages between governmental and other actors structured around shared if endlessly negotiated beliefs and interests in public policy making and implementation"[9].

One of the main difficulties in PNA is that the generation of these PN is usually done in manual, cumbersome processes by experts. This involves the use of interviews, questionnaires and other instruments from the social sciences. During the actual generation of the network many subjective factors may come into play as the overall result depends on the people involved in the study. Consequently, PN generation involves a significant investment that does not always "lead to breathe taking empirical and theoretical results" [7]. The question is consequently, how can we use technology to improve the process of PN identification.

2.2 Towards the automatic generation of PNs: defining our objectives

The main objective of this study is to propose a method for the automatic generation of PN surrounding the Legislative Branch of government. We decided to focus on the Legistative Power for two reasons. First, as we have previously established, Parliaments write laws and regulations, and thus have an enormous power in the shaping of a society. This makes them the primary target of lobbyists.

Second, as we will later see in chapter 3, the challenges of automatic Social Network generation lie in i) understanding and characterizing the links found among the actors occurring in the SN and ii) discovering hidden relationships, which can only be detected by learning that two entities share a common topic. When outlining this study, we aimed to define the problem in a way that we could address these two difficulties and produce SN which are meaningful, easy to interpret and with as many relevant connections as possible.

As we will see later, bills can then be used as a cornerstone for detecting relationships between politicians and third-party actors.

The reader should know that for the rest of this document, we use the words bills and laws exchangeably to refer to the rules and regulations approved by a legislative body. We also use the terms entity, actors, political actors to refer to companies, non-governmental organizations, governments, advocacy groups, citizens and in general any person or association that is related to a bill, meaning that the contents of the bill affect their interests or policies.

2.3 What type of PN do we aim to generate?

Among the different types of PN that we could study, we are concerned with the generation of an *Entity-Entity graph* capturing possible relationships between two actors. More formally, we are interested in discovering pairs of actors that are i) related to a bill and ii) given the context of a bill, share similar roles or positions, meaning that they could possibly know each other, be in contact and cooperate or compete to push their agendas. Using bills allows us to examine the relationships between political actors accross a wide range of topics and discover fine-grained relationships which could otherwise be missed.

Additionally, in many cases entities have either a positive or negative position with respect to a law. In the case of politicians, their voting history for a particular bill is recorded and usually accessible through Parliament websites or open data. In the case of companies, NGOs and other types of actors, their position could be assessed either automatically, by employing sentiment analysis, or manually, by using expert knowledge. Using bills provides a framework for easily anotating the discovered links with this polarity, thus allowing more refined graph analysis techniques. For instance, in [19] we find a community detection algorithm which works with positive and negative links.

Determining the polarity of instances escapes the objective of this study. It is however relevant to mention this possibility as it is one of the advantages of using bill-centered PNs. For instance, we could argue two entities are positively related if they are positively or negatively related to a common bill; similarly, two entities could be negatively related if they have different polarities with respect to a common bill. These considerations are left for future work.

2.4 Stop Online Piracy Act (SOPA): a motivating example

As an illustrative example, consider the *Stop Online Piracy Act* (SOPA). SOPA was a bill introduced in 2011 in the United States to combat online copyright infringement and online trafficking in counterfeit goods. If a website was found to infringe the law, SOPA allowed court orders to require Internet service providers to block access, prevent search engines from listing them and forbid advertising networks or other payment facilities from conducting business with the website. SOPA also made the unauthorized streaming of copyrighted content a criminal offense, imposing a penalty of up to five years in prison.

Due to its controversy, SOPA was widely discussed internationally and many organizations raised their voices in favor and against the bill[2]. In broad terms, organizations which rely on copyright strongly supported the bill. This includes, for instance:

- **Pharmaceutical companies and associations** like Pharmaceutical Research and Manufacturers of America (PhRMA), Pfizer, Alliance for Safe Online Pharmacies (ASOP).
- **TV channels**: ABC, NBC, ESPN, Disney, among others.
- **The music industry**, including the Recording Industry Association of America, the American Federation of Musicians, Sony Music Entertainment, Universal Music, etc.

Similarly, organizations that advocate for freedom and liberties, or whose business would be negatively affected by increased regulations voiced their opposition against the bill. Among these, we highlight:

[2]A list of organization in favour and against can be found in `http://en.wikipedia.org/wiki/List_of_organizations_with_official_stances_on_the_SOPA_and_PIPA`

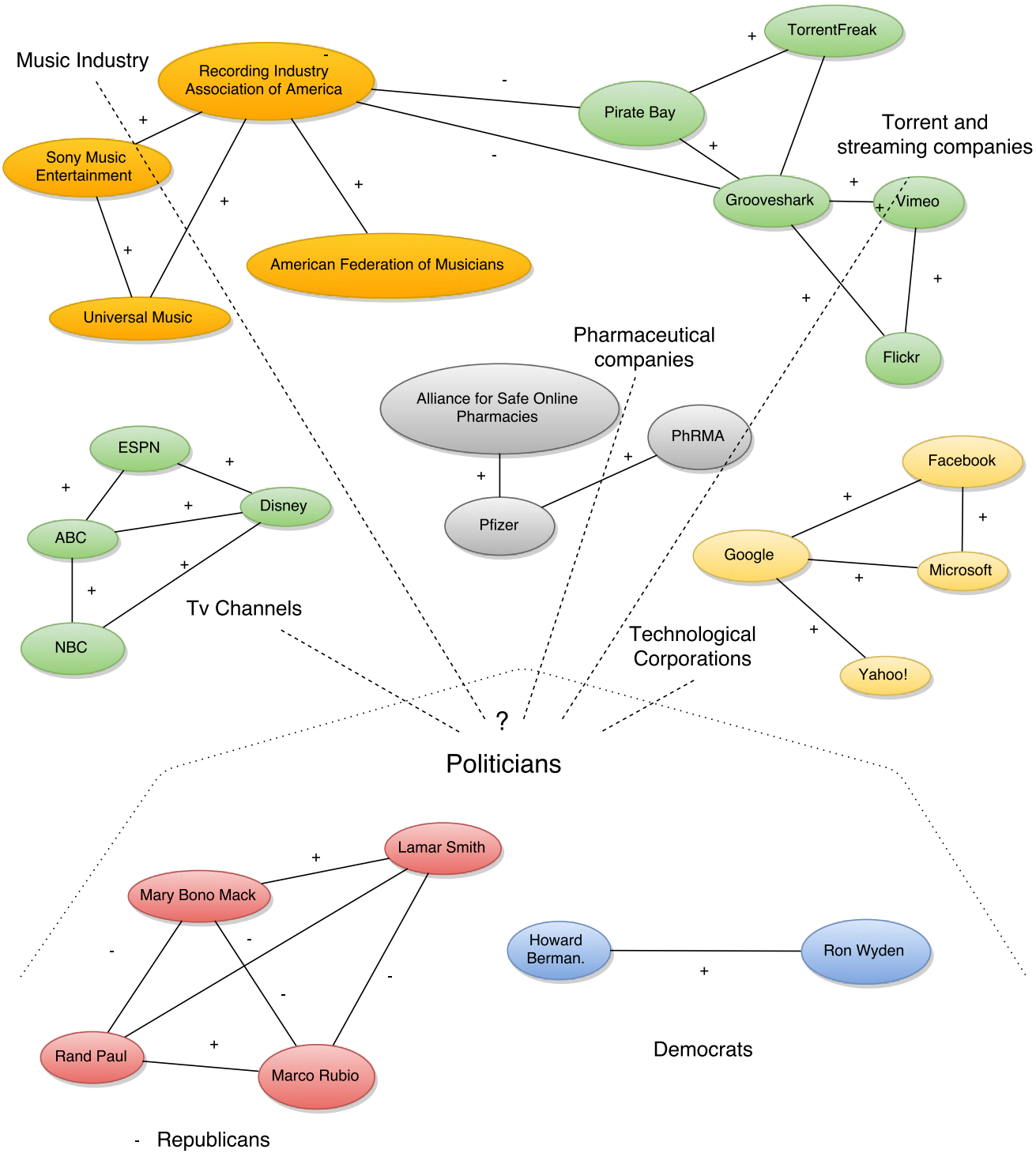

Figure 1: Entity-entity graph for SOPA.

- **Torrent and streaming companies**: such as Pirate Bay, TorrentFreak, Vimeo, Grooveshark, Flickr, etc
- **Technological corporations**, such as Facebook, Yahoo!, Microsoft, Google.

Finally, we can also know the position of congressmen by looking at their voting history or their positions. In the case of SOPA, Republicans Lamar Smith and Mary Bono Mack, and Democrat Howard Berman were among the sponsors of the bill, while Democrat Ron Wyden and Republicans Marco Rubio,

Rand Paul were against.

Based on this information, which we have manually gathered, we could generate an *Entity-Entity graph* as the one shown in figure 1. This graph was generated in a manual way taking into account the relationship of the entity with respect to the bill as defined by the categories listed above.

Note that we deliberately chose not to draw edges between the politicians and the rest of the political actors. This is due to the fact that we do not have the knowledge to suggest that a politician and a company might be related. We decided to draw only links between organizations of the same type, with the exception of the relationship between the Recording Industry Association of America and Grooveshark, TorrentFreak and PirateBay. Since the latter three are used to illegally download music we can infer that they are related. Our objective in this study is to produce a system that can automatically detect these and other obscure relationships automatically by using news articles.

What type of analysis can we do with the *Entity-Entity graph*? We can use the traditional SNA tools to individually analyze bills or alternatively look at the overlapping of these graphs accross different bills. This graph is however particularly useful for detecting possible lobbying relationships. For instance if a politician (or group of politicians) has a high number of links with the political actors of a specific community (imagine for instance that Lamar Smith had links with most of the Music Industry group) then we could possibly infer that there is a lobbying relationship.

2.5 Is it really possible to detect lobbying in an automated way? An important caveat

It is important to clarify that despite being motivated by the detection of patterns that can be suggestive of a lobbying process, in practice we aim to find relationships of political similarity or dissimilarity between two entities. The relations found by our method do not necessarily imply that the two actors are in direct liaison; to do so we would need to closely monitor all the activities of politicians and organizations to verify with whom they are in contact. This is naturally unreasonable due to privacy considerations.

We aim to detect links between entities that indicate that they are both related to a particular political decision – in our case bills –, from which we could then establish political affinity or aversion. Naturally, if two entities have highly similar political views in a broad set of issues then one can believe that they could be collaborating. This could be verified by investigative journalists and by considering other sources of information like donations information, speeches, etcetera. Regardless of the verification of a lobbying process, our method is intended to be an unbiased, low-cost, semiautomated tool to aid the process of Policy Network generation and analysis.

3 Related works

There is an increasing interest in the development of applications to generate Social Networks (SN) relating entities occurring in corpora of text. This is due to the fact that there is a growing wealth of knowledge contained in text documents which is difficult to exploit due to their unstructured nature. By generating graphs that reflect the information contained in these documents, we produce a structured view which can be analyzed using Social Network Analysis (SNA) tools.

In this section we present the state of the art in the generation of SN from corpora of text and some related applications that use SNA for political analysis. First in section 3.1 we present the entity co-occurrence approach, a simple technique that has been widely used with good results. We then show in 3.2 relevant work by the scientific community to characterize the link between two entities in a way that can be used by SNA. We also describe in 3.3 methods that can be used to relate entities that despite not being mentioned in the same document might be linked by means of the broader context provided by the whole corpus of documents. Next, we present in section 3.4 studies that use SNA for political analysis and that are related to our project. Finally, we succinctly present in 3.5 some considerations about how the state of the art was taken into account when making our proposal.

3.1 Entity co-occurrence: a first step towards the creation of SN

One of the most widely used approaches to generate SN from text consists in relating entities based on their co-occurrence in a certain context. The underlying assumption of this approach is that if two entities are consistently mentioned together then they are probably related. There are several variants depending on the granularity of the context; one can look for co-occurrence within a certain sentence, paragraph, document or cluster of documents. The choice depends on the amount of data available – finer granularity requires more documents to produce more relations – and on the application.

The authors of [5] propose a method to automatically generate a social network of narco-traffickers in Mexico based on the co-occurrence of names in books about the topic. They do Entity Recognition (ER) to produce a list of entities which is then manually curated and used to determine links between drug dealers. The weight of the relationship is the count of repetitions of the co-occurrences of two entities within a certain distance. They use different network analysis tools to show how the obtained graph closely resembles the different cartels and their chain of command.

Similarly, the Joint Research Centre of the European Commission has done extensive work in extracting entities and inter-entities relations from newspapers written in different countries of the European Union. In [15] we find a summary of their work, which is explained in depth in [13] and [16]. Essentially, they look for co-occurrence of entities within previously built clusters of articles that represent a story. They take into account entity coreference and use different heuristics to improve the entity recognition and disambiguation processes. They also produced a formula to measure the strength of a link based not only on the number of co-occurrences but also the frequency of the entities in the clusters and the corpus. By doing this, they aim to weight down relationships in which one of the entities is frequently mentioned, so that only relevant relationships are chosen.

Another interesting aspect of their proposal is the use of Wikipedia for validating the obtained graphs. Because we are in presence of a knowledge discovery task for which we do not have a ground truth set, it is difficult to evaluate if the detected links between two entities are meaningful. The definition of a meaningful link is itself not an easy task. The authors of the Joint Research Centre look for the Wikipedia pages of the detected entities and verify if there are links between pages that correspond to the links detected by the system. They define a "strong" relationship as one in which there is a reciprocal presence of a link. This allows the creation of a ground truth set which can be used to evaluate the system with the standard precision and recall metrics.

On a different note, the authors of [3] present a technique to measure the semantic similarity of two words or phrases by using the Google search engine. They propose a metric based on information distance and Kolmogorov complexity that uses the count of search hits returned by looking up two words individually and together. The authors show how their approach is useful for distinguishing between colors, numbers, names of paintings and names of books, among others.

The main advantage of this approach is that it is able to measure the similarity of two entities based on the whole corpus of documents in the World Wide Web. The drawback is that the number of search hits returned by Google is a gross and often highly inaccurate estimate of the real count. Particularly, it is usually the case that a search with more terms returns a higher count of hits than a search with a subset of these terms. The reason for this is that when adding more terms the search is more fine-grained, allowing for a more refined estimate of documents. More specifically, when having more search terms it is necessary to go deeper through the posting lists which leads to more accurate and larger result estimates. The data centers or the indices used when answering the query also affect the number of expected hits returned. This makes approaches that depend too much on the exact count returned by the search engine unreliable.

There are two shortcomings in taking a co-occurrence approach. First, we are often interested in characterizing the link between two entities in terms of strength and meaning to produce richer graphs. It is true that the co-occurrence approach allows a human user to manually inspect the documents in which two entities co-occur. We are however particularly interested in mechanisms that can infer and represent the semantic nature of a link in a way exploitable by SNA tools with as little human participation as possible. Second, we are also interested in methods that do not rely on direct co-occurrence within a same document (or a pre-computed cluster of documents), but that can also discover meaningful relationships across a corpus.

3.2 What links two entities? Enriching the SN with the strength and semantics of the relations

As previously said, the co-occurrence approach relies on the assumption that if two entities are mentioned in the same context then they are probably related. Co-occurrence may indeed be suggestive that two entities are related, but if there is a relationship we need to characterize that relationship before we can perform SNA. To illustrate this, in the context of our application finding a link between a politician and a third party might indicate that they are closely aligned politically, that they are in opposition, that they participated together in a meeting, that they mentioned each other, among others. Having more information about the found relationships is consequently of great importance to be able to do better analysis.

The efforts of the scientific community to characterize relationships between two entities have been mostly concentrated on Natural Language Processing methods to analyze the context in which the entities co-occur.The authors of [18] propose a method which uses dependency trees to learn patterns that relate two entities co-occurring in a sentence according to a pre-defined type of relationship. They work with two examples of relationships: “support” and “meeting”. By working with a small, manually obtained number of seed instances of the relationship – tuples of entities –, they look for sentence co-occurrence in a group of news articles and extract patterns by using their SyntNet GSL algorithm over the dependencies found by a dependency parser.

Similarly, in [14] the authors propose a method to automatically detect quotation relationships in news articles. They aim to do this by also finding linguistic patterns that are often used when expressing citations: quotation markers and reporting verbs. The authors of [21] illustrate the use of kernel methods for relation extraction. They first produce a shallow parse representation of the texts which is then used by kernels designed specifically to work on parse trees, which have been defined in [4]. These kernels are able to implicitly enumerate all possible subtrees of two parse trees, find which are the most common subtrees,

weight them and compute a similarity measure based on these. By using a pre-obtained ground truth set, they are able to train classifiers to determine, given two entities and a tree describing the sentence they co-occur in, if the entities have relationships of the type person-affiliation and organization-location.

Determining the strength of a relationship is also of interest for SNA applications. To the best of our knowledge, there are no explicit efforts within the scientific community to assess the strength of relations inferred from corpora of text in a systematic way. Most of the applications are mostly concerned with link detection, which is often done by computing and theresholding similarity measures between entities. These similarity measures can be seen as measures of strength; defining a formal method to determine the strength would require however to have previously annotated SNs or a model for specifying what a strong/weak relation is. This is hard, particularly when we take into account the difficulty of manually producing measures of strength that are unbiased and inexpensive. While the question of determining the strength of relationships between two entities has been addressed for online Social Networks in which interactions may be indicative of the strength[20], this remains a challenge in the area of SN generation.

3.3 Going beyond the co-occurrence approach: finding links between entities across documents

There is also a number of techniques to address the need to find links between entities without depending on their direct co-occurrence within a same document or pre-computed cluster of documents. A widely used approach is automatically inferring topics present in a corpus of text and to verify co-occurrence in articles related to these topics. The authors of [12] propose a method that uses Latent Dirichlet Allocation (LDA) to produce a topic model of the documents. In LDA a document is regarded as a finite mixture of topics and represented as a vector in which each component constitutes the probability that the document belongs to a given topic. This model is then used to calculate an entity-entity measure of affinity that is used to find links between entities. The reported results are motivating; the authors show how the use of topics allows to discover more links between entities and to characterize them by means of the topics. They also report however that LDA has one significant disadvantage: the obtained topics may be hard to interpret and may not be sufficiently semantically cohesive.

An alternative to the use of LDA is Latent Semantic Indexing (LSI). By producing a vectorized representation of entities (in which for instance we store information about their co-occurrence in documents), we can use Singular Value Decomposition (SVD) to find a lower dimensional space and estimate the similarity of entities based on latent concepts. This is useful for noise-reduction and for finding semantic relationships between entities. The drawback is that the found relationship may be hard to interpret. The authors of [1] provide an example of the use of LSI for the generation of graphs of terrorists networks.

3.4 SNA and political analysis. Has anyone done this before?

In [2] we find a recent survey on the automatic extraction of Policy Networks. They mention several approaches for the generation of SN which we have already mentioned in this chapter (or that are at least closely related to the cited studies) and some other applications using NLP for political science analysis. For instance, in [10] the authors propose a method to automatically extract and characterize relationships between politicians and locations (relations, for instance, of the type (Barack Obama, President, United States); they do so by looking for co-occurrence of persons and locations in web documents, extracting keywords from their context, clustering similar pairs of (Person, Location) and identifying relevant labels.

In the survey great attention is paid to the work in [11], as it is the only one, to the best of our knowledge, which specifically addresses the task of automatic Policy Network extraction. The authors of this proposal work with two PNs previously created by experts in a manual, time-consuming process. They evaluate the use of three type of metrics for SN generation that can be produced by using a Web Search Engine:

- ***Co-occurrence metrics***, which measure the degree to which two political actors co-occurr in web pages by looking them up individually and in conjunction in a search engine. Based on the number of results, they produce four metrics of similarity: the *Jaccard Coefficient*, the *Dice Coefficient*, *Mutual Information* and the *Google-based semantic relatedness* [3].
- ***Text-based metrics***, which use a vectorial representation of political actors in which components are the frequency of occurrence of words in a certain context of the snippets returned by a search engine. They use cosine similarity to produce a similarity measure from these vectors.
- ***Link-based metrics***, which exploits the hyperlinks of the web pages returned by the search engines to measure the degree of association between actors. The assumption is that if two actors are mentioned in webpages that have links to the same webpages, then they are probably similar. To measure this, they use a version of the *Google-based semantic relatedness*.

The manually created SNs contain positive and negative edges, which correspond to relationships of political affinity and aversion, and are annotated with measures of strength. This is particularly useful for understanding how the different methods for graph generations perform.

In general, they obtained better results for when detecting affinity relationships than negative relationships. They also found that using link-based metrics and co-occurrence metrics are the best alternatives for positive relationships while text-based metrics are the best option for negative relationships. When comparing the four proposed measures for co-occurrence based similarity they found that *Mutual Information* is the best alternative for positive links, whereas the *Dice Coefficient* is the best option for negative links.

The main difference with our work is that while they work with a predefined list of political actors from the policy networks they use for validation, we are also interested in the discovery of relevant entities and in finding ways to characterize their links. Also, the authors of this proposal did not present any alternatives for automatically inferring the sign of the relationships detected by their system. We can use parliamentary data and news articles to address these two needs.

3.5 Framing our work with respect to the state of the art

As we have seen, generating social networks from corpora of text is a topic widely addressed by the scientific community. There are two important considerations with which researchers are concerned: i) discovering the largest number of relationships possible while ensuring the discovered relationships are meaningful and ii) characterizing the relationships between entities in a way exploitable by SNA tools.

The task of discovering patterns of political closeness between politicians and third parties is conditioned by the fact that these relationships are usually hidden and unknown by journalists and the public. Entity co-occurrence in a given context thus entails a risk of not revealing all the relevant relationships. Similarly, characterizing the relationships by means of NLP techniques would also require knowledge by the writer of the document of the existence of a relationship between two entities. Finally, if we model topics from the whole set of documents we could come up with some that are not related to the Parliament.

To address this, we use bills as the cornerstone that allows us to link politicians and third parties. By using bills, we can model interpretable and semantically cohesive topics that allow us to address the two considerations mentioned earlier in this section. We can specifically i) detect meaningful links between entities across documents and ii) characterize the found links by using the bill. Furthermore, analyzing the relationships between political actors in the specific context of a bill allows us to discover fine-grained relationships that would otherwise be missed if we analyze the whole set of documents at the same time. In section 4 we present our proposal and how the state of the art was used for specific tasks.

4 Generating bill based Policy Networks

In this study we propose to use bills approved by a Parliament as the element linking political actors. To generate a PN, we take as input news articles and Parliamentary data and automatically model topics based on the contents of the bills, extract relevant politicians that participated in the drafting of the bill, find articles related to each of these topics, perform entity recognition and normalization and compute similarity measures between relevant entities to produce *Entity-entity graphs*. Figure 2 shows the pipeline of the whole process, from the obtention of the data to the production of the desired graphs. In this chapter we describe how we carry out each of these tasks.

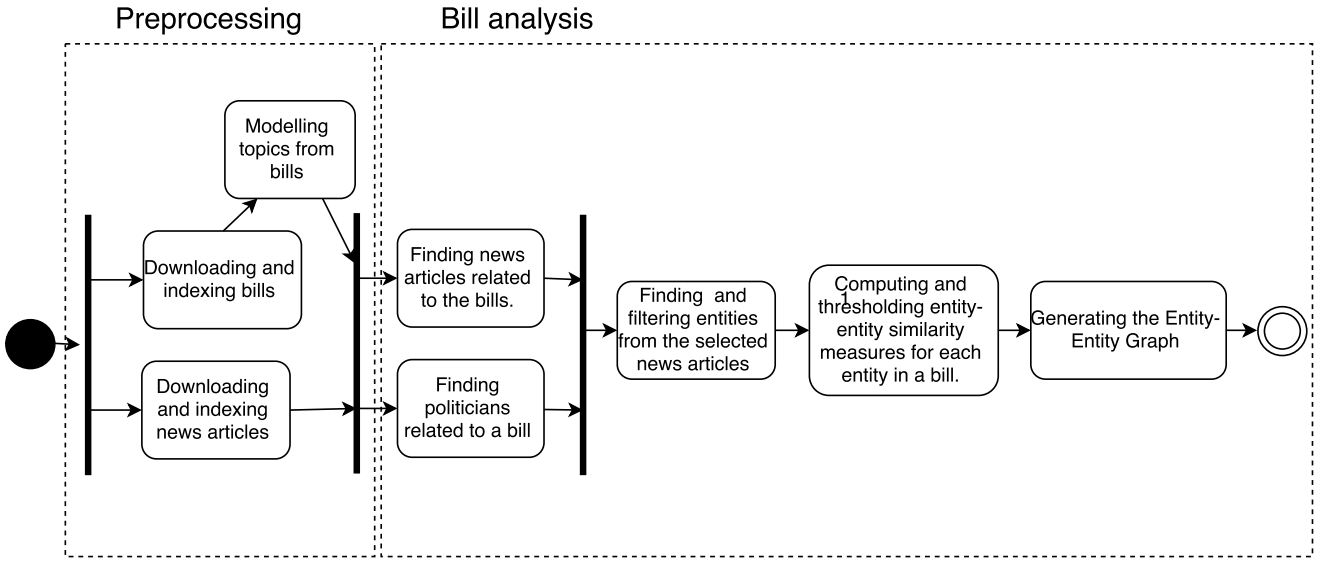

Figure 2: The pipeline of our bill analysis.

4.1 Preprocessing: getting the data ready for analysis

The first step consists of obtaining and preprocessing the news articles and bills so that they can be analyzed efficently. To do this, we developed scrappers to automatically visit the websites of the main newspapers of Catalonia and download all the news articles of the last ten years. This requires a different approach for each newspaper, particularly to determine how to get to every news article from their home page and the Xpath necessary to get the contents of the news articles without noise. We also made a scrapper to download all the bills that have been approved by the Catalan Parliament.

In Catalonia many newspapers publish their articles in Catalan or Spanish, or in some cases both. For this study, we chose to work only with news articles written in Spanish. The reason is that there are better Entity Recognition tools for Spanish. To automatically classify the articles by language, we use the number of stopwords in Spanish and Catalan of each article. After doing this process we obtained a total of 758276 articles in Spanish.

Once we have downloaded the news articles, we index them using ElasticSearch[3], a distributed search engine based on Apache Lucene. ElasticSearch uses inverted indices to fastly query documents. Based on the types of queries that we do, we make two indices treating news article and individual paragraphs as documents.

[3]https://www.elastic.co/

4.1.1 Modelling topics from bills

We model each bill as a topic by extracting a list of weighted keywords. This allows us to find news articles that are related to each bill. To find keywords, we consider n-grams of size *1,2,3* and compute the TF-IDF score of these n-grams with respect to the set of all bills approved by the Parliament. TF-IDF is a measure of relevance of a keyword in a specific document which takes into account its frequency inside the document and accross the whole corpus of documents. Formally, it is defined as:

$$tf - idf(keyword, document, corpus) = tf(keyword, document) * idf(keyword, corpus)$$

Where tf is usually the number of times that the keyword occurs in the document and idf is the *inverse document frequency*, a term which aims to penalize words that are used in many documents, and is defined as:

$$idf(keyword, corpus) = log(\frac{|corpus|}{|\{d \in corpus | keyword \in d\}|})$$

For computing the TF-IDF of a keyword, we consider each bill as a document.

To enhance the quality of the keyword extraction we could also consider the transcripts of the debates of each bill. By doing this we could discover words and phrases that describe the contents of the bills in an informal, non-legislative way which is possibly more similar to the jargon used by newspapers. This is however a minor improvement and is left for future work.

After extracting keywords, we rank them and keep the top 1000 keywords. We keep the TF-IDF score of the keywords as a weight.

4.2 Analyzing a bill

To analyze a bill, we find relevant politicians and related news articles, detect relevant entities and compute entity-entity similarity measures to generate the desired Social Networks. In this section we describe this process in detail.

4.2.1 From bills to news articles: modelling topics from bills

In order to find entities that are related to the bills, we first find news articles that are related to the bills. To do this, we use the extracted keywords to query the corpus of news articles. To determine the relevance of an article, we consider the vector space model representation of each bill and news article, and compute the cosine similarity between the vectors.

To improve the phase of keyword extraction, we use Rocchio's algorithm to discover new keywords from the corpus of news articles. Rocchio's algorithm is widely used in Information Retrieval systems to expand queries defined by users by adding news terms found in the top relevant documents retrieved by an initial search. To determine what a relevant document is there are two alternatives: (i) asking the user for feedback – an option which is not possible in our application – and ii) assume that the top documents retrieved by the query are relevant.

After defining a set of relevant documents $|R|$ and a set of non-relevant documents, we update the query vector q according to this formula:

$$q_{new} = \alpha * q + \beta * \frac{1}{|R|} * \sum_{d \in R} d + \gamma * \frac{1}{|NR|} * \sum_{d \in NR} d$$

Where d is a document in the vector space model, α, β, γ are manually fixed parameters usually satisfying $\alpha > \beta > \gamma = 0$.

Once a new query vector is computed, we execute a new query and rank the news articles according to the new query. After doing this, we need to determine the number of articles to use based on the score. To do this, we plot the score of the articles with respect to their ranking and find the best trade-off point according to the following method.

4.2.2 Finding the best trade-off point between relevance and size

When plotting a relevance as a function of the ranking of an article, we observe that all bills follow a pattern similar to the one in figure 3. That is, the relevance score is very high for the first articles and then drops quickly. We want to use as many relevant articles as possible, avoiding at the same the inclusion of articles that are not related to the bill and that could introduce noise.

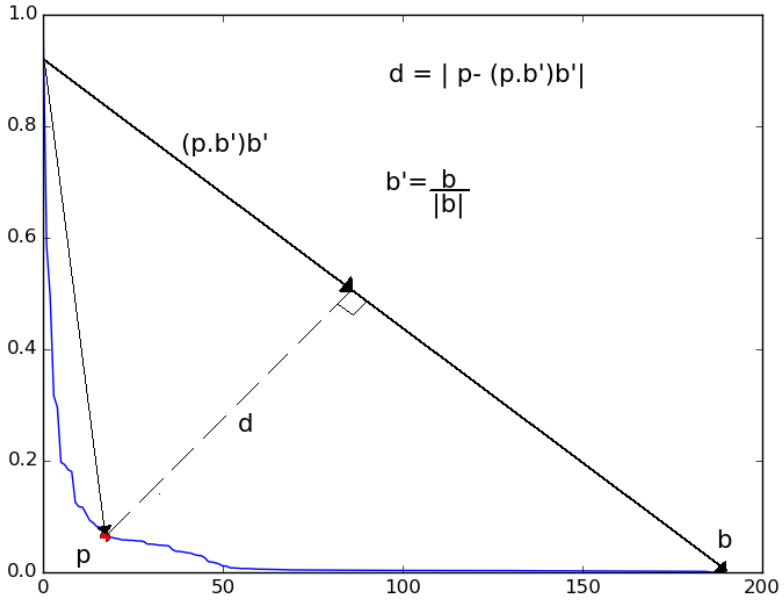

Figure 3: Determining the number of articles to use for a given bill.

To solve this, we observe that the plot ressembles an elbow and pick a value a point near the region where the two arms of the elbow meet. One way to do this is by selecting the point p that maximizes d as shown in figure 3. In the figure:

- b is a vector pointing from the point of highest relevance to the point of lowest relevance.
- b' is an unitary vector in the direction of b
- p is a vector pointing from the point of highest relevance to each point in the curve.
- $(p.b')b'$ is the projection of p onto b.
- $|p - (p.b')b'|$ is the distance of every point of the curve to the line going from the point of highest relevance to the point of lowest relevance.

The threshold is chosen by determining the point that maximizes d according to the formula:

$$threshold = \underset{p}{\text{argmax}} |p - (p.b')b'|$$

4.2.3 From news articles to political actors: entity recognition and preprocessing

After finding articles related to bills, they are analyzed to find entities that are in turn related to the bills. For this purpose we use MITIE[4], a state of art tool Named Entity Recognition tool created by the MIT. Given a document, MITIE identifies substrings that contain possible named entities and tags them as *Organization, Location, Person* or *Miscellaneous*.

Before carrying on the detected entities need to be pre-processed before they can be used. The names found may i) not be correctly delimited (resulting in truncated names or in names that contain excess text) ii) be ambiguous (an entity may have more than one name and a name may refer to more than one entity) and iii) may be noise and not refer to a real entity.

Name disambiguation is in itself a complex and interesting research topic which goes beyond the scope of this study. We have however implemented some heuristics we briefly describe below:

1. **Entity Normalization**: entities are brought to a canonical form to address spelling variations. This involves i) punctuation sign removal; ii) double, leading and trailing whitespaces removal; iii) leading and trailing stop words removal; iv) string camelization (all characters are put in lowercase except for the first letter of every word, which is in uppercase).
2. **Mapping organization initials to the whole name**: when an organization name is detected we aim to detect if there are any other names composed by its initials. Specifically, we aim to exploit a widely found pattern: organizations often have the full name followed by the initials inside parenthesis.
3. **Mapping partial names with full names**: when person names are detected we classify them into *full names* (containing more than 1 word) and *short names* (containing one word, which is possibly the last or first name of the full name). We then link short names with the nearest, previous full name such that the short name is contained inside the full name.
4. **Expanding names based on the news corpus**: to address the issue of truncated names (eg 'Word Life Fund' may be truncated and processed as 'World' and 'Life Fund') we: i) look up every name and its surrounding context in the corpus ii) extract sentences in the top articles and iii) find the longest substring matching these sentences.

By doing this we find a list of entities for which we have a list of aliases and a list of tag counts. After doing this, we choose for every entity the tag with the highest frequency as the type of the entity.

4.2.4 Extracting the authors and influencers of a bill from Parliamentary data.

While news articles are an excellent source to determine third party actors which are affected or interested in a bill, there is also a wealth of information to be obtained from open data released by Parliaments. By using the transcripts of their meetings, we can determine who are the authors and the politicians that participated in the drafting of a bill. These politicians might be of great importance and still not show up often in the news articles.

In the case of the Catalan Parliament, all the transcripts and procedings of the meetings of the different comissions are available online. The Parliament website also allows to look for the speeches of individual politicians or to look for all the interventions concerning a bill. By interventions we mean speeches, authorships or amendment writing. Given a bill, we label a politician as relevant if they have at least one intervention about that bill.

[4]https://github.com/mit-nlp/MITIE

The extraction of relevant politicians could be perfomed automatically by scraping the website of the Parliament or by querying their database. However, due to time constraints this functionality is left for future work. We manually look for each of the bills we analyzed and store the list of relevant politicians in the database.

4.2.5 Selecting relevant actors: filtering noisy entities

After obtaining a list of entities for each bill, we need to select the entities that are most likely to be related to the bill. Our method for entity selection yields thousands of entities for each of the topics. This makes filtering them necessary for two reasons.

First, noise is bound to be present in the list of entities found by our system. This includes entity names that do no correspond to real entities, entities that occur in articles related to the bills but that are not really related to them and entities from articles that are completely unrelated to the bills. As we will see later in section 4.2.6 to compute entity-entity similarity measures it is better to work with a list of entities as less noisy as possible.

Second, to produce an entity-entity graph we need to compute similarity measures between each pair entities; the cost of each comparison depends on the method chosen and the number of comparisons to make grows quadratically with the number of entities.

To select relevant entities, we first filter out entities that occur less than a given threshold in the whole corpus of news articles. This is useful for removing phrases that are not real entities but that were incorrectly classified as such by the ER tool. Then we compute a gross and inexpensive similarity measure between entities and use it to perform clustering. We do this to find a group of entities that are strongly related among themselves and with the bill. The similarity measure is computed as follows:

1. First we generate a binary matrix in which rows correspond to entities and columns correspond to news articles related to the bill that the entities occur in. Each cell contains a one iff the entity occurs in the document; otherwise it contains a zero.

2. After generating an *Entity-Document* matrix, we perform Latent Semantic Index (LSI) to reduce the number of dimensions. LSI uses Singular Value Decomposition to identify relationships between terms and concepts in a corpus of text. The idea is that terms (in our case entities) that occur in the same context (news articles) tend to have a similar meaning. LSI takes as an input the *Entity-Document* matrix B and transforms it into three matrices:

   - An $m \times r$ term-concept matrix T, where m is the number of terms and r is the number of concepts.
   - An $r \times r$ singular values matrix S.
   - An $n \times r$ concept-document matrix, where n is the number of documents.

   Such that B,T,S and D satisfy the condition:

   $$B \approx TSD^T$$

   .

   By doing LSI we can use the term-concept matrix T to identify that two entities are similar even if they do not co-occur in any documents. The reasoning is that there might be more than one topic (or concept) in the set of retrieved news articles. By automatically identifying possible concepts, we can establish the similarity between two entities based on their relations to the latent concepts as opposed to the stricter requirement of direct document co-occurence.

3. We compute an entity-entity similarity measure by taking the cosine similarity of the entity vectors in the concept space, i.e. the cosine similarity of the rows of the T matrix.

Computing the similarity measure on the concept space is relatively cheap computationally as it involves computing the dot product between two vectors with few components. We can further reduce the computational cost of selecting relevant actors by using the fact that any entity which is related to the topic must be vaguely related to the politicians that participated in the drafting and discussion of the bill. That is, rather than computing the N^2 comparisons involving thousands of entities, we can first compute the similarity of each of the N entities with a list of seed politicians and discard entities which do not have a similarity higher than a certain threshold with at least one of the politicians. We use an $\epsilon = 0.1$ in our system. This is a safe assumption as the seed politicians must be strongly related to the concepts to which the bill is related and are also rows of the B matrix.

After computing the similarity measure between all pairs of entities, we perform Hierarchical Aglomerative Clustering (HAC) to identify the set of relevant actors. We define the dissimilarity between two entities as one minus their cosine similarity. More specifically, we perform Unweighted Pair Group Method with Arithmetic Mean (UPGMA), which computes a dendogram that reflects the structure in the dissimilarity matrix. UPGMA uses a bottom-up approach, starting from clusters containing each individuals and progressively merging two clusters at a time until every individual is in the same cluster. At each iteration, the nearest two clusters are combined into a higher-level cluster. UPGMA uses the mean dissimilarity to merge clusters. That is, the dissimilarity between any two clusters A and B is taken to be the average of all dissimilarity between pairs of objects "x" in A and "y" in B.

When performing HAC, it is necessary to set a threshold or number of clusters. To select this threshold, we use the silhoutte, a widely usely measure of the quality of a cluster based on the dissimilarities between objects inside a cluster and their dissimilarity with the nearest clusters. Formally, the silhoutte of an observation (in our case an entity) is defined in [17] as:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

where $a(i)$ is the average dissimilarity of the entity i with the rest of the entities in its cluster and $b(i)$ is the lowest average dissimilarity of i with any cluster of which it is not a member.

It can be shown that $-1 < s(i) < 1$. Particularly, in order to have a silhoutte close to 1, we need $a(i) > b(i)$, which means that entity i is significantly closer to the members of its cluster than to the members of the nearest cluster. A high value of the silhoutte thus indicates a better cluster quality. We can compute the silhoutte of a cluster by taking the average of all its entities.

By taking into account the seed politicians we can choose the best cluster that groups entities related to the bill. We do this by choosing the cluster that i) contains all the seed politicians, and ii) has the highest average silhoutte. Because the silhoutte can have several local minima and be unstable we also request the cluster containing the seed politicians to have a minimum size (we use a minimum size of 100).

After performing this step we are able to reduce the number of entities to analyze from thousands to hundreds. This makes the fine-grained similarity measures we use in section 4.2.6 faster to compute and more accurate.

4.2.6 From a list of relevant actors to a Policy Networks: computing and thresholding similarity measures

Once we have produced a list of relevant political actors, we need to compute and threshold a similarity measure to produce a Social Network. Thresholding this similarity measure is an optional step; depending

on the type of analysis we wish to perform we could consider all the possible relationships between all the pairs of entities and keep the similarity measure as a weight. We present a method for thresholding this measure however, as it leads to a richer visualization and the analysis and applications that we present in this thesis better with sparse, unweighted graphs.

As we previously said in section 2.4, we are interested in generating *Entity-Entity graph* capturing relationships between entities that given the context of a bill, share similar roles or positions. The challenge of this procedure is thus being able to determine which are the relevant relationships between the entities without i) producing very dense graphs (high number of edges) and ii) missing important relationships.

As we saw in chapter 3, the state of the art in SN generation uses co-occurence and topic modelling to compute the similarity between two entities. In this study we use a text-based similarity measure similar to the one used by the authors of [11]. Its computation consists of the following steps:

1. For every entity, look for every article that mentions it which is related to the bill. We consider that an article is related to the bill if it is selected after applying the elbow criterion specified in section 4.2.2.

2. For every article the entity is mentioned in, look for every paragraph that mentions the whole name of the entity or any of its aliases.

3. For every paragraph that mentions the entity, generate $1, 2, 3$ grams with the contents of the paragraph. Then use these N-grams to generate a vector space representation of the entity capturing the contexts it is talked about. We use TF-IDF to generate these vectors, using sublinear tf scaling in which the TF component is calculated as follows:

   $$tf = 1 + log(frequency)$$

   This version of TF-IDF is widely used when the weight of a keyword should not be linearly dependent on its number of occurrences. For instance, a keyword occurring 30 times should not be 30 times more important than a keyword occurring only once. This is useful in our context as we are comparing N-grams of different sizes and using normal TF-IDF could give much more weight to N-grams of size one over N-grams of size two or three, as n-grams of size one are more likely to have a higher frequency. The N-grams of size three could however be more useful to describe the entities. Also note that this version gives a higher weight to the IDF component of the formula, as $1 + log(x) > x$ for $x > 1$. This means that we also give a higher weight to rare terms in comparison with the typical TF-IDF formula.

4. The similarity between two entities is finally computed by taking the cosine similarity of the vectors representing each entity.

This similarity measure has several of the advantages of the different approaches we studied in the literature. Specifically, it takes into account:

- Entity co-occurence in a context: if two entities are mentioned together, they will be represented in their corresponding vectors as N-grams. Furthermore by using TF-IDF the co-occurence is weighted according to its frequency – if two entities co-occur often their corresponding components will be high – and the frequency of each entity accross all the documents – if an entity is often referred to, then the IDF will penalize its weight –.

- The semantic role of the entity with respect to the topic, as captured by the words that are used to talk about that entity and the bill.

Note that by using only documents that specifically talk about a bill, we are able to produce fine-grained similarity measures. If we tried to determine the similarity based on the whole set of documents, the relationships between politicians and actors that are rarely talked about could be missed. The N-grams that describe each entity are computed and weighted taking into account only news articles that talk about the bill.

The last step before computing the PN consists of thresholding the similarity measure to discard relationships between entities that have a low weight. Rather than using a global threshold for all the relationships, we fix a threshold for each entity taking into account its similarities. For each entity e we:

1. Rank the rest of the entities with respect to their similarity to e.
2. Plot the similarity measure of every entity with respect to e as a function of the ranking.
3. Find an elbow according to the method defined in 4.2.2. This elbow captures the most relevant entities for e, i.e., it detects the point after which all entities have a low similarity. This decision was made empirically after studying the similarity-ranking plots of several entities for different bills; most of the plots follow the shape of figure 3, meaning that the elbow is an useful technique for filtering relevant entities.

After computing a set of relevant entities for each entity, we decide that two entities $e1$ and $e2$ are related iff they are included in the set of relevant entities of the other entity. That is, $e1$ must be contained in the list of relevant entities of $e2$ and viceversa. This requirement implies that two entities must be *strongly* related to be added to our final graph. Note that an interesting variant consists of considering instead directed graphs in which we draw an edge from a node $e1$ to a node $e2$ if $e2$ is a relevant entity for $e1$ or if $e1$ is a relevant entity for $e2$. Studying this alternative is left for future work.

We do this procedure for every entity related to each bill to generate our desired Policy Networks. According to the tag of the entities (PERSON, ORGANIZATION, LOCATION) we build different types of graphs, showing the relations between the entities of one category (PERSON-PERSON, for instance) or between entities of different categories (PERSON-ORGANIZATION for example). We then analyze these graphs using SNA tools. In chapter 5 we show the obtained results for three bills considered as case studies.

5 Results

The evaluation of our system is conditioned by the fact that there is no ground truth available to assess its quality. To evaluate our system, we decided to present three bills as case studies and to show the insight and short-comings of our method in the generation of Policy Networks. For each of these bills, we present a brief description so that the reader understands their context and show some of the most relevant PNs that we generated.

5.1 BCN-World

BCN-World is a tourist and entertainment project which was announced in 2012. The project included several hotels, casinos and gambling houses, theme and water parks, golf courses, a beach club, theaters, convention centers, shopping malls, restaurants, among others. BCN-World was planned to be built in Tarragona, a province of Catalonia which is located to the south of Barcelona.

The project was largely promoted by Veremonte, a british based investment group, and Convergencia i Uniò (CiU), a right-wing party currently governing in Catalonia. Veremonte coordinated the efforts of several other investment groups, including La Caixa - a Catalan bank -; Melco and Caesars Entertainment - two Asian and American based leisure and gambling corporations -; Hard Rock; Value Retail - a luxury outlet shopping company- and Melia Hotels.

Because of the enormous size of the project, its implementation required the modification of the law regulating every aspect of a Touristic Complex. The new bill, written in 2014, took into account all the negotiations between the investors and the Catalan Parliament. For instance, given the significant involvement of gambling companies in the project, lowering the gambling industry taxes from 55% to 10% was a necessary pre-condition for the execution of the project.

The negotiations between the investors and the Catalan government were conditioned by the fact that the ruling party, CiU, does not have a majority and has required the support of a left wing party called Esquerra Republicana de Catalunya (ERC). ERC was opposed from the very start to the bill. Because of this, CiU had to look for the support of other political parties, and found an ally in the Partit dels Socialistes del Catalunya (PSC, the Socialist Party of Catalunya). Despite initially being against the project, the PSC changed its position because it rules in the cities located near BCN-World and its mayors were interested in the execution of the project as a means to create employment in their districts. After intense negotiations including the mayors and senior politicians of the PSC, CiU got the necessary votes to approve the new law regulating Touristic Complexes.

5.1.1 Organization-Organization PN

Figure 5 shows the PN relating the main organizations concerned by the BCN-World bill. Due to space constraints, we show only the giant component of this network. The nodes of the graph are coloured according to the communities detected by performing modularity clustering. The size of the nodes and the font of the names are proportional to their Pagerank centrality. Pagerank is a metric widely used in Social Network Analysis to assess the importance of a node. The Pagerank algorithm assigns to each node a value that is computed based on the number and quality of its links to other nodes. The idea is that a node is as important as the number of links it receives from other important nodes.

There are several insights that can be derived from the Organization PN. To start with, three main communities are detected. The first community, in dark red and in the upper left part of the graph, is mostly composed of political parties and other NGO which are linked to these. ERC, Ciu, CUP, ICV-EUIA, PP, PPC and Ciutadans are political parties, while ACENCAS (Associació Catalana d'Addiccions Socials) is the Catalan Asociation of Social Addictions, Camara Catalana (Cambra Catalana) is the Catalan Chamber of Commerce, Parlament is the Parliament, Govern refers to the Government and Tripartito

and Sociedad Centre Medics Selva Maresme are noisy entities which are not related to the bill. All of the political parties are related among themselves and with the Government, the Parliament and the Chamber of Commerce. ACENCAS, an organization which was opposed to the bill, is shown to be linked with CUP, another left wing party which was very strongly against BCN-World.

The second community, shown in dark pink and in the lower right section, is mostly constituted by the companies interested in investing in BCN-World. Note that none of these companies are related to the political organizations except for Veremonte. This is due to the fact that Veremonte, as we previously established, was the intermediary between the investors and the government. Its bridging role is clearly shown in this PN.

The last community, in light pink and the lower left part, is composed mostly of institutions from Tarragona, the province where BCN-World was set to be built. This includes the local government (Diputación de Tarragona, in catalan Diputació de Tarragona), the local chamber of commerce (Camara de Comercio de Tarragona, in catalan Cambra de Comerç de Tarragona), the Confederation of Tarragonian Business (CEPTA), the Port of Tarragona and Universitat Rovira i Virgili (an university based in Tarragona). There are other organizations that are not related to Tarragona, including the Spanish Association of Accountants, La Caixa (a Catalan bank) and Pimec which stands for Small and Median Businesses. Note however the important role of Diputacion de Tarragona as determined by its pagerank value. This confirms its role as a broker between the Catalan Government and Veremonte.

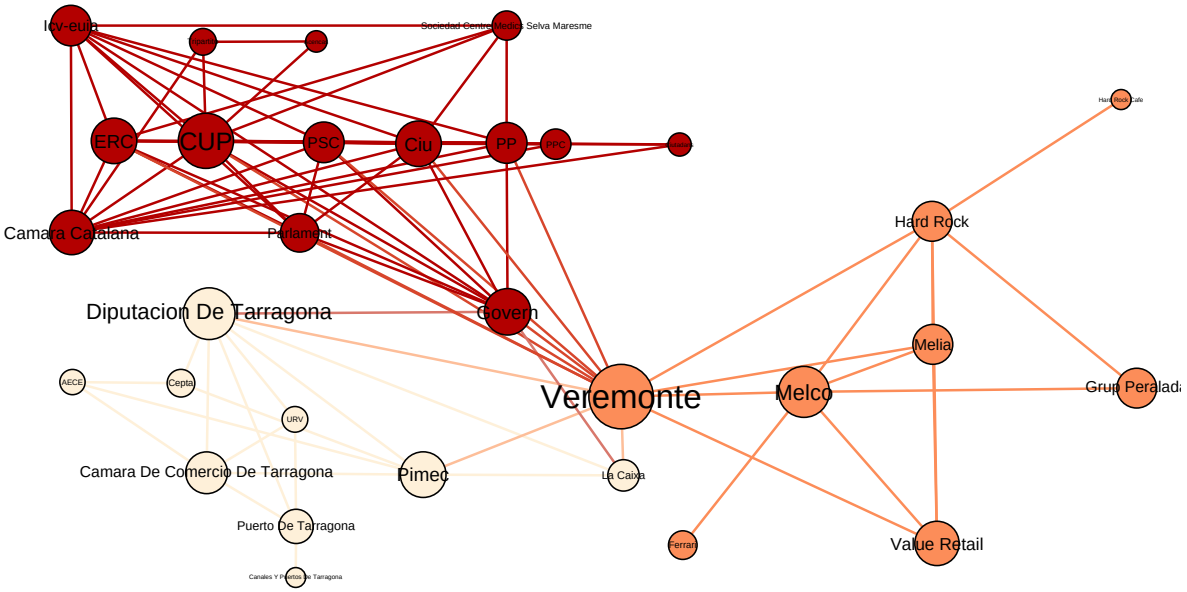

Figure 4: Organizations PN of the BCN-World Bill.

5.1.2 Person-Organization PN

Figure 5 shows the PN relating the relationships between the most important persons and organizations related to BCN-World. This graph was constructed by considering all possible relationships between persons and organizations, selecting the most influential persons and organizations as measured by their pagerank and filtering out every node that is not relevant or directly connected to a relevant node. Once

more, we show only the giant component due to space constraints.

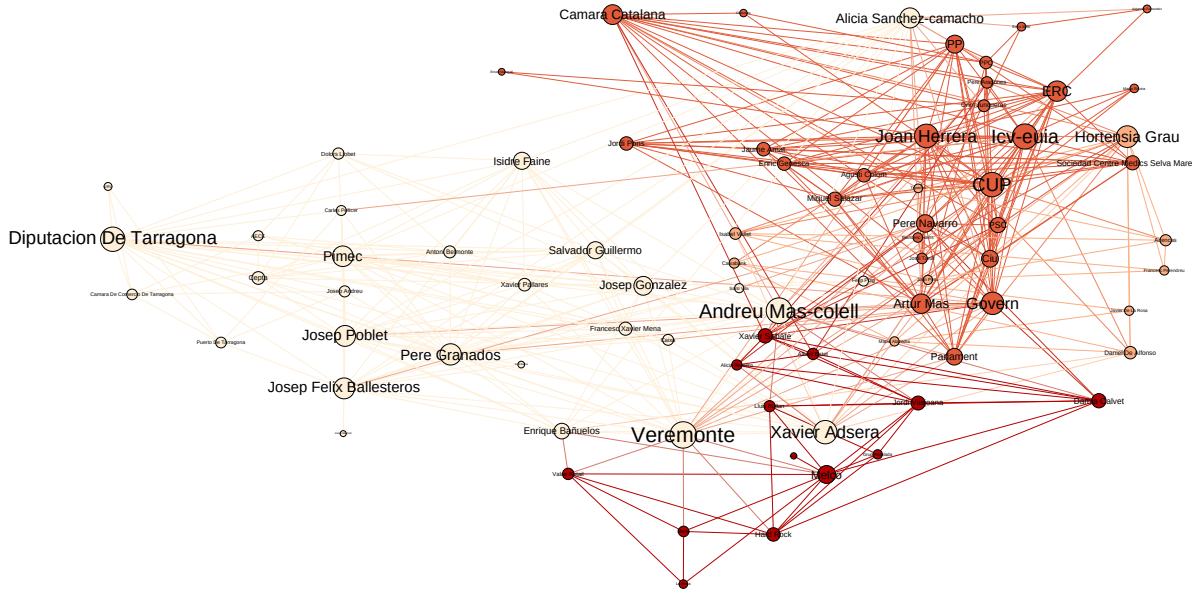

Figure 5: Person-Organization PN of the BCN-World Bill.

This graph is bigger in number of nodes but is similar to the Organization-Organization PN. To start with, it features three communities which are related to the communities previously found: a community formed by investors (in the lower right corner), a community of politicians and their political parties (in the upper right corner) and a community of organizations and people related to Tarragona (in the left side).
Note that the most relevant actors from the Organization-Organization PN retain their importance: Veremonte, Diputació de Tarragona, CUP and the Government are highly central. Cambra Catalana (The Catalan Chamber of Commerce) is now slightly more important, as it has many connections with politicians.

When it comes to analyzing the role of persons, there are many interesting insights to be gained. First, the most relevant people are Andreu Mas-Colell, the Advisor of Economy and Knowledge of the Catalan Government, and Xavier Adsera, a senior advisor to Veremonte and the president of the BCN-World group. Other relevant entities include Josep Felix Ballesteros, Pere Granados and Josep Poblet, three PSC and CiU mayors of the main towns in Tarragona; Joan Herrera and Hortensia Grau the spokesmen of ICV-EUIA, a party against the project; Artur Mas, the president of the Catalan Government; Alicia Sánchez-Camacho, spokeswoman of PP, the ruling party in Spain and Isidre Faine and Enrique Bañuelos, presidents of La Caixa and Veremonte.

Something which is particularly interesting is how there intermediary role of Veremonte is reinforced in this PN, showing links with most of the nodes in the other two communities and being also connected to them through Xavier Adsera. By inspecting the news articles, we realize that Adsera was the main negotiator for Veremonte and participated in several key meetings, for instance with Ballesteros, Granados and Poblet, to convince the PSC to vote in favor of the law.

An Organization-Person-Person-Organization graph pattern like this might be particularly interesting for detecting lobbying relationships. We find for instance that Veremonte is directly conected to the Catalan Government; these are two organizations that are in turn related to Adsera and Artur Mas who are brokers in charge of connecting their respective organizations.

Besides from these interesting insights, this graph also shows some shortcomings of our method. First, there is a giant community made of politicians which does not seem to exhibit any kind of internal sub-structure. Almost all of the politicians of the Parliament are connected with every other politician, instead of, for instance, showing only connections among politicians of the same party.

A second shortcoming of this graph is that it fails to detect some relationships which are very important. For instance none of the previously mentioned mayors are related to their political party, something which is particularly important to establish their role as negotiators. This might be because they are mostly referred in the news as mayors of important cities of Tarragona (note how they are strongly linked to most of the institutions of the Tarragona community) instead of being related to their political parties.

5.2 Law of Popular Non-referendary Consults

The Law of Popular Non-referendary Consults (hereinafter LPNC) is a law which was approved in September 2014 to allow the Catalan Government to call for non binding electoral consults to catalan citizens. This law was particularly controversial as it was approved with the main objective of making a consult regarding Catalonia'secession from Spain. Shortly after its approval, Artur Mas, president of the Catalan Government, called for such a consultation to take place on November 9th. The Spanish Government immediately declared its unconstitutionality, which the Supreme Court of Spain availed.

The LPNC is particularly interesting because although its a law regulating any type of consultation on public affairs, its controversy lies on its use for the November 9th election. The affected parties are consequently mainly politicians and organizations which are in favor or against Catalan independence.

5.3 Organization PN

Figure 5 shows the Organization PN for the LPNC. As previously, colors identify communities and font and node size identify importance as measured by the Pagerank of the node. We show only the giant component of the graph.

The first relevant observation is that the entities shown in this graph are mostly political parties and NGOs related to the independence of Catalonia. Modularity Clustering yields three communities, shown in the upper left, upper right and bottom sections of the graph. The two communities in the the upper right section are made on the one hand by parties of Spain, Catalonia and other regions of Spain which have independentist claims and, on the other hand, institutions that were involved in determining its constitutionality: the Spanish Senate and House of Representatives (*Senado* and *Congreso*), the Catalan Parliament(*Parlament*) and the Constitutional Court of Spain (*Tribunal Constitucional*). The single community in the bottom of the graph mostly shows pro-independence groups: *Barcelona Decideix*, *Reagrupament*, *Solidaritat* and *Moviment Arenyenc*.

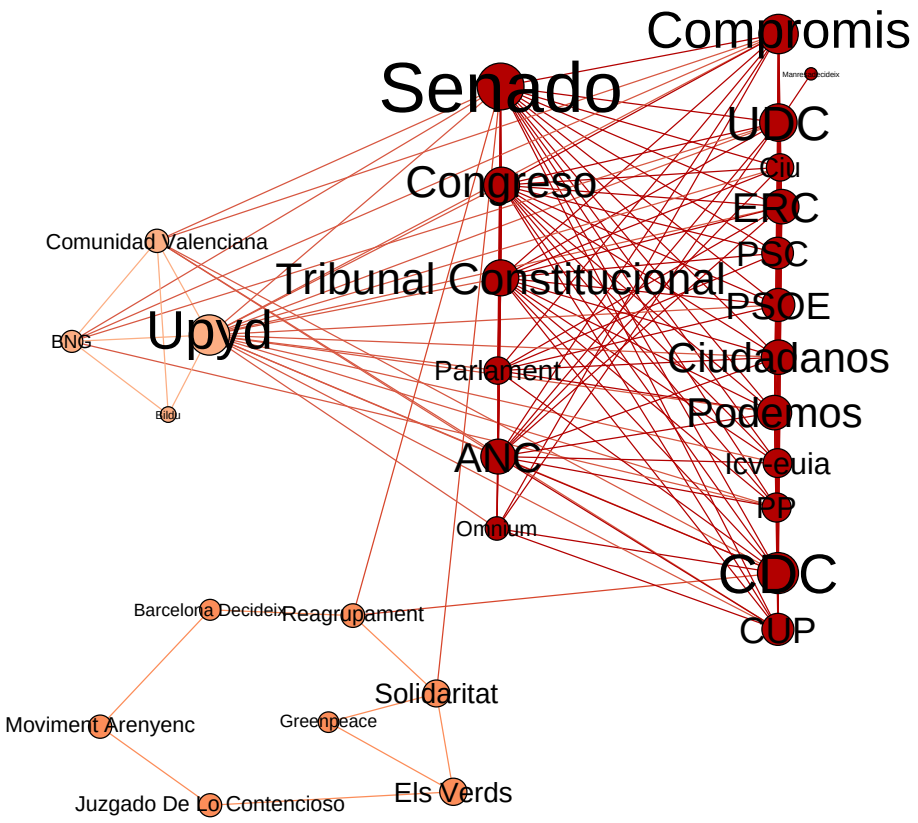

Figure 6: Organization PN for the LPNC bill

Broadly speaking, these communities correctly identify the two types of organizations that are involved with the law, with some exceptions: *ANC* (Assemblea Nacional Catalana, in English National Catalan Assembly) and *Omnium* are NGOs which rather than being connected with the rest of the NGOs are connected with other political parties. This might seem counter-intuitive at first, but if may be because they work more with political parties than with the rest of NGOs. Similarly *Juzgado de lo Contencioso* (Litigation Court) would intuitevely be related with the rest of the political institutions. Finally, note how *Greenpeace* and *Els Verds*, despite not being related, are present in the graph. These might be due to the fact they are too Non-Governmental Organizations, devoted though to environmentalist causes.

When it comes to assessing the importance of organizations, we find that almost all the political parties, with the exceptions of parties from other regions of Spain, are equally relevant as measured by Pagerank. Also note the importance of the Spanish Senate and Congress, the Catalan Parliament and the Constitutional Court. Among the pro-Independence NGOs, *ANC* is the most preeminent.

5.4 Person PN

Figure 7 shows the Person PN for the LPNC. When analyzing the community structure of the PN, we see that there are seven communities, identified with different colours. In the case of this PN its hard to make sense of the clusters of people identified by modularity clustering. Politicians of different parties and of different regions or instutions, are spread around the graph.

On the other hand, looking at individual, widely known politicians, it is possible to make sense of the relations we see. For instance, Mariano Rajoy, the prime minister of Spain, is connected to senior members of PP, his political party, and of other relevant parties. Similarly, senior pro-independence politicians like Artur Mas, Oriol Junqueras, Josep Antoni Duran tend to be connected together.

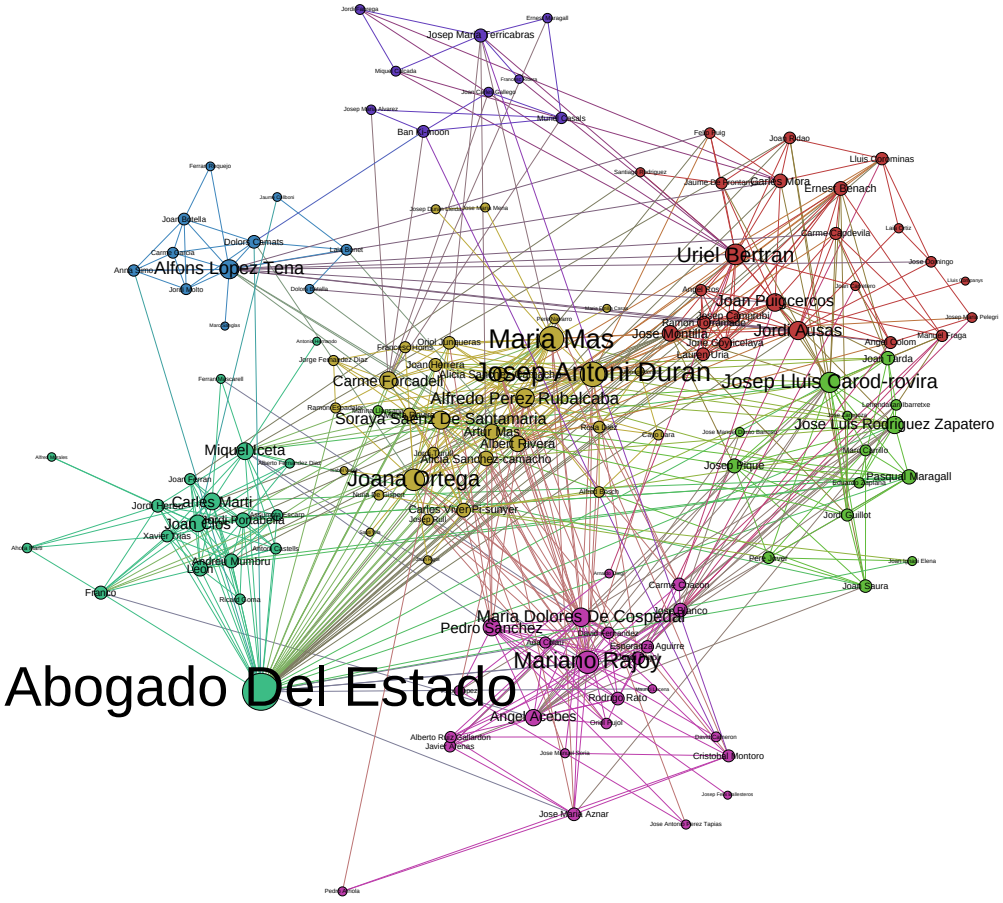

Figure 7: Person PN for the LPNC bill

6 Conclusions

In this work we present a method for automatically generating bill-centered Policy Networks relating politicians and other relevant third parties. The generated Policy Networks can be used for doing political analysis, i.e., detecting influential actors, understanding the flow of information surrounding a bill and identifying possible lobbying relationships. By using parliamentary data and news articles, we automatically detect relevant entities and discover their relationships by studying their semantic similarity and sentence co-occurence.

Summarize results aqui.

our method is intended to be an unbiased, low-cost, semiautomated toal to aid the process of Policy Network generation and analysis.

Faltan articulos.

6.1 Contributions

Our contributions are the following:

1. The use of bills as a cornerstone relating political actors, which allows to i) understand better the discovered relations and ii) and to find fine-grained relationships which would otherwise be missed.
2. An unsupervised method for automatically detecting relevant entities of a given topic from a corpus of documents – in our case news articles – given a set of seed entities.
3. A framework for labeling the discovered relationships according to the polarity of the political actors and generating signed graphs, which allows the use of Signed Social Network Analysis tools.

6.2 Future work

The work we have presented is a first step with many improvements which are left for future work. These improvements consists of i) refining the method and its evaluation with expert insight, ii) finding ways to enhance the PN generation phase and iii) producing a list of use cases as a base to tailor the method to the needs of political analysts and journalists.

6.2.1 Towards a more rigorous evaluation and problem definition

The first step for improving our method consists in validating our results by a group of experts. In this report we show how our results are reasonable and indicate the potential of bill-centered PN generation. However it is import to acknowledge that the analysis we made might be deceiving as it is on the one hand prone to confirmation bias, meaning that when analyzing a PN we might look for information that confirms our own sets of beliefs, and on the other hand can lead to missing links, which at this stage we are not possible to detect.

Consequently, we need a more rigorous and systematic approach for evaluation, meaning producing a set of gold standard PNs and list of relevant entities so that we can validate the results in a quantitative way. Assessing the quality of our proposal in a more formal way is key to any of the other possible improvements. Furthermore, manually generating a gold standard can shed light on ways to improve each of the steps we have presented in our method.

6.2.2 Improving the PN generation phase

Based on the obtained results, improving the way the entity similarity measures are computed and thresholded is arguably the most important improvement for future work. We have seen many cases in which

the discovered relations are meaningful and lead to important discoveries in terms of communities, influence and paths of communication between relevant actors. In some others, there are both missing and irrelevant relationships; this can be due either to the quality of the similarity measure or it's threshold. It is important to assess which of these two problems needs to be addressed. For instance, given a set of gold standard PNs we could look at the statistical correlation of the similarity scores between entities and the values of the gold standard relations, as in the work presented in [11]. If a high correlation is found, then it is better to focus on the improvement of the similarity measure thresholding, which can be done by trying different methods and using the one that yields the best results with respect to the gold standard.

To improve the similarity measures there are many things which can be done. We decided to use a text-based similarity measure comparing entities based on the words they appear in because it captures i) co-occurrence and ii) semantic relatedness. However, it might be that entity co-occurrence should have a stronger weight for detecting relevant relationships. We have seen, for instance, how politicians tend to be related to many other politicians without necessarily being related by the bill; it might well be that because of their condition of politician and parliamentaries, they share a great number of keywords. On the other hand, using LSI for computing similarity measures between entities might also be useful; this requires however defining how many concepts should be used as a parameter.

6.2.3 Generative models: a major improvement and food for thought

On another hand, there are many efforts of the scientific community in finding ways to generate graphs and predict missing links by using information about the topology of the desired graphs. Generative models allow to reason in a statistical manner by thinking of the probability of a graph given a set of parameters θ that describe the underlying structure of a family of graphs. For instance, the *Stochastic Block Model* assumes that there are a number of communities (blocks). This allows to statistically infer the set of parameters describing that community structure (($P(\theta|G)$). Similarly, given a set of parameters, it allows to generate the most likely graph (($P(G|\theta)$).

A major improvement of this work consists in studying i) what types of generative models best describe the topology of policy networks ii) how to enrich these models with the computed similarity measures and iii) evaluating the use of these methods for generating PNs. Doing this could allow us to compute PNs not only based on the media coverage of a law, but also on the generalization of what a bill-centered PN looks like. Also, it allows us to avoid the problem of thresholding similarity measures.

6.2.4 How to use PNs for political analysis: use-case driven PN generation

There are plenty of Social Network Analysis tools which can be used to analyze our PN. In this case we have illustrated the use of modularity clustering and Pagerank to detect groups of actors and influential actors, but there are other ways to analyze these graphs. One thing to think about is whether it is useful to threshold the similarity measures. For example, we could look for the shortests paths connecting two organizations that necessarily traverse PEOPLE nodes, in which case the weight of an edge is important. Similarly, it would be interesting to explore if it is worth it to work with directed graphs.

Furthermore, it could interesting to analyze the time evolution of these graphs to understand how the dynamics of a law evolve and answer questions such as how does the influence of an actor change over time, what new actors appear and how do their dynamics change, among others. For instance, the PN of a specific bill could be very different before it is approved and after it is approved, e.g. could be the case that a bill has a negative impact on a society, triggering new organizations to be involved with it.

Another aspect of SNA which could be explored is the use of multiplexes, which is becoming increasingly popular in the network analysis community. Multiplexes are networks composed of different layers in which we find the same nodes but different types of links; it is interesting to study from this perspective

the correlation between layers generated for different bills or to employ multiplex SNA tools.

References

[1] Roger B Bradford. Application of latent semantic indexing in generating graphs of terrorist networks. In *Intelligence and Security Informatics*, pages 674–675. Springer, 2006.

[2] Mrs Vaishali Chaudhari and J Ratanaraj Kumar. A survey paper on automatic extraction of policy network using web link and documents. In *International Journal of Engineering Research and Technology*, volume 3. ESRSA Publications, 2014.

[3] Rudi L Cilibrasi and Paul MB Vitanyi. The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on*, 19(3):370–383, 2007.

[4] Michael Collins and Nigel Duffy. Convolution kernels for natural language. In *Advances in neural information processing systems*, pages 625–632, 2001.

[5] Jesús Espinal-Enríquez, J Mario Siqueiros-García, Rodrigo García-Herrera, and Sergio Antonio Alcalá-Corona. A literature-based approach to a narco-network. In *Social Informatics*, pages 97–101. Springer, 2014.

[6] James H Fowler. Connecting the congress: A study of cosponsorship networks. *Political Analysis*, 14(4):456–487, 2006.

[7] Patrick Kenis, Volker Schneider, et al. Policy networks and policy analysis: scrutinizing a new analytical toolbox. *Policy networks: Empirical evidence and theoretical considerations*, pages 25–59, 1991.

[8] Justin H Kirkland. The relational determinants of legislative outcomes: Strong and weak ties between legislators. *The Journal of Politics*, 73(03):887–898, 2011.

[9] Michael Moran, Martin Rein, and Robert E Goodin. *The Oxford handbook of public policy*. Oxford Handbooks Online, 2008.

[10] Junichiro Mori, Takumi Tsujishita, Yutaka Matsuo, and Mitsuru Ishizuka. Extracting relations in social networks from the web using similarity between collective contexts. In *The Semantic Web-ISWC 2006*, pages 487–500. Springer, 2006.

[11] Theodosis Moschopoulos, Elias Iosif, Leeda Demetropoulou, Alexandros Potamianos, and Shrikanth Shri Narayanan. Toward the automatic extraction of policy networks using web links and documents. *Knowledge and Data Engineering, IEEE Transactions on*, 25(10):2404–2417, 2013.

[12] David Newman, Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. Analyzing entities and topics in news articles using statistical topic models. In *Intelligence and Security Informatics*, pages 93–104. Springer, 2006.

[13] Bruno Pouliquen, Ralf Steinberger, and Jenya Belyaeva. Multilingual multi-document continuously-updated social networks. In *Proc. Workshop: Multi-source, Multilingual Information Extraction and Summarization*, pages 25–32, 2007.

[14] Bruno Pouliquen, Ralf Steinberger, and Clive Best. Automatic detection of quotations in multilingual news. In *Proceedings of Recent Advances in Natural Language Processing*, pages 487–492, 2007.

[15] Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, and Tamara Oellinger. *arXiv preprint cs/0609066*, 2006.

[16] Bruno Pouliquen, Hristo Tanev, and Martin Atkinson. Extracting and learning social networks out of multilingual news. In *Proceedings of the Social Networks and Application tools workshop (Skalica, Slovakia, Septembe*. Citeseer, 2008.

[17] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[18] Hristo Tanev. Unsupervised learning of social networks from a multiple-source news corpus. *MuLTISOuRcE, MuLTILINguAL INfORMATION ExTRAc-TION ANd SuMMARIzATION*, page 33, 2007.

[19] VA Traag and Jeroen Bruggeman. Community detection in networks with positive and negative links. *Physical Review E*, 80(3):036115, 2009.

[20] Rongjing Xiang, Jennifer Neville, and Monica Rogati. Modeling relationship strength in online social networks. In *Proceedings of the 19th international conference on World wide web*, pages 981–990. ACM, 2010.

[21] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. Kernel methods for relation extraction. *The Journal of Machine Learning Research*, 3:1083–1106, 2003.

ANNEXES