TECHNICAL UNIVERSITY OF CRETE

Extraction of Policy Networks using Web-based Features

by

Theodosis Moschopoulos

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Department of Electronics and Computer Engineering

February 23, 2012

Contents

List of Figures

List of Tables

Abstract

Policy making process in modern democratic systems is the outcome of the interrelations and interdependencies among political entities (i.e organizations, companies, groups or unions) from public or private sectors and from different levels of governance. Thus, the network perspective, namely policy network, is an efficient tool for political scientists to describe, analyze and explain various financial and social phenomena during the policy making process. A policy network can be described as a social graph with nodes the actors and edges the relations among them. The relations in a policy network serve as channels for communication, exchange of information, expertise, trust and other policy resources. Traditionally, policy networks are created manually after a series of arduous, time consuming steps including interviews and questionnaires. Furthermore, the manual creation of such networks is oftenly a high budget procedure which requires high-level of expertise. Another problem is that manually created policy networks suffer from subjective biases such as the respondent's will for participation, cultural and political issues even external factors such as the economical or political system.

In this work we propose a method for the automatic extraction of policy network. More specifically, given the actors of the network, our approach estimates the strength of relations among them. Our fundamental assumption is that the strength of such relations can be discovered automatically and in an unsupervised way through a variety of features that can be harvested from the web. Such features include webpage counts, outlinks and lexical information that is extracted from web queries, web documents or web snippets. In our work, we propose three types of metrics as well as their fusion i) page-count-based metrics that use the number of occurrences/co-occurrences of actors in web documents ii) text-based metrics that exploit the actors lexical context in web snippets iii) link-based metrics that use the outlinks cited in web documents where the actors exist. iv) the linear combination of the three types of metrics above. The proposed approach is automatic and does not require any external knowledge source, other than the specification of the word forms that correspond to the political actors. It is also language independent as it is not based on any knowledge about the language. Furthermore, the proposed approach reduces the biases emerged by the traditional methods (who depend on the answers of a small number of respondents) as it can integrate multiple points of view by exploiting the collective information of the web. Our approach is evaluated on two human-rated networks taken from the political science literature. The networks are located in Ireland and Greece and web queries are performed in English and Greek respectively. Furthermore, the extracted networks are visualized and qualitatively evaluated by political scientists. Based on the fact that relations in policy networks evolve through time, we apply our method to extract the networks for the years of a specific time period and visualize this evolution.

It is shown that our method can efficiently estimate the strength of relations that express cooperation (positive relations), while fail to estimate relations of antagonism (negative relations). Furthermore, our approach effectively identifies the most 'active' actors computing the degree of centrality which is a widely used measure in network analysis. Finally, the visualization of the policy networks as well as their evolution draw interesting results and conclusions from the perspective of political sciences.

Περίληψη

Η διαδικασία χάραξης πολιτικών σε σύγχρονα δημοκρατικά συστήματα είναι το αποτέλεσμα των σχέσεων και αλληλεξαρτήσεων μεταξύ των πολιτικών φορέων (π.χ. οργανισμοί, εταιρείες, ενώσεις ή ομάδες) από το δημόσιο ή τον ιδιωτικό τομέα και από τα διάφορα επίπεδα διακυβέρνησης. Έτσι το δίκτυο πολιτικής αποτελεί ένα αποτελεσματικό εργαλείο για τις πολιτικές επιστήμες για να περιγράψουν, αναλύσουν και να εξηγήσουν διάφορα οικονομικά και κοινωνικά φαινόμενα κατά τη διάρκεια της διαδικασίας χάραξης πολιτικών. Ένα δίκτυο πολιτικής (ή αλλιώς πολιτικό δίκτυο) μπορεί να περιγραφεί ως ένας κοινωνικός γράφος με κόμβους τους φορείς και άκρα τις σχέσεις μεταξύ των φορέων αυτών. Οι σχέσεις σε ένα πολιτικό δίκτυο χρησιμεύουν ως δίαυλοι επικοινωνίας, ανταλλαγής πληροφοριών, τεχνογνωσίας, εμπιστοσύνης και άλλων πηγών πολιτικής. Παραδοσιακά, τα πολιτικά δίκτυα χτίζονται με το χέρι μετά από μια σειρά επίπονων, χρονοβόρων διαδικασιών που συμπεριλαμβάνουν συνεντεύξεις και ερωτηματολόγια. Επιπλέον, η χειροκίνητη δημιουργία αυτών των δικτύων είναι μια κοστοβόρα διαδικασία η οποία απαιτεί υψηλό επίπεδο εμπειρογνωμοσύνης. Ένα άλλο πρόβλημα είναι ότι τα παραδοσιακά πολιτικά δίκτυα επηρεάζονται από προκαταλήψεις των ατόμων που συμμέτέχουν στη διαδικασία, όπως π.χ η θέληση του ατόμου για συμμετοχή στην έρευνα, πολιτιστικά και πολιτικά ζητήματα, αλλά και από εξωτερικούς παράγοντες, όπως το οικονομικό ή πολιτικό σύστημα.

Στην εργασία αυτή προτείνουμε μια μέθοδο για την αυτόματη εξαγωγή ενώς πολιτικού δικτύου. Πιο συγκεκριμένα, δεδομένων των φορέων του δικτύου η μέθοδος μας εκτιμά την ένταση των σχέσεων μεταξύ των φορέων αυτών. Η βασική μας υπόθεση είναι ότι η ένταση των σχέσεων μπορεί να εκτιμηθεί με αυτόματο τρόπο και χωρίς επίβλεψη μέσα από τα διαφορετικά χαρακτηριστικά που εξάγουμε από το διαδίκτυο. Τέτοια είναι ο αριθμός ιστοσελίδων, σύνδεσμοι και λεκτική πληροφορία που εξάγονται απο ερωτήματα, κείμενα του διαδικτύου και snippets. Στην εργασία αυτή προτείνουμε τρεις τύπους μετρικών καθώς και τον συνδιασμό τους i) page-count-based μετρικές που χρησιμοποιούν τον αριθμό εμφανίσεων/συνεμφανίσεων των φορέων σε κείμενα του διαδικτύου ii) text-based μετρικές που χρησιμοποιούν το λεκτικό περιεχόμενο σε snippets που βρίσκονται τα όνοματα των φορέων iii) link-based μετρικές οι οποίες χρησιμοποιούν τους συνδέσμους σε κείμενα που εμαφνίζονται οι φορείς iv) τον γραμμικό συνδιασμό των παραπάνω μετρικών. Η προτεινόμενη μέθοδος είναι αυτόματη και δεν απαιτεί καμία εξωτερική πηγή γνώσης εκτός από τις διαφορετικές μορφές ονομάτων των φορέων. Επίσης είναι αδιάφορη της γλώσσας καθώς δεν βασίζεται σε γνώση για τη γλώσσα. Επιπλέον, μειώνει τις προκαταλήψεις που εμφανίζονται στις παραδοσιακές μεθόδους εξαγωγής πολιτικών δικτύων (οι οποίες βασίζονται στις απαντήσεις ενώς μικρού αριθμού ερωτηθέντων) καθώς ενσωματώνει διαφορετικές οπτικές γωνίες κάνοντας χρήση της συλλογικής πληροφορίας του διαδικτύου. Η προτεινόμενη

μέθοδος αξιολογήθηκε σε δύο δίκτυα από την βιβλιογραφία των πολιτικών επιστημών. Τα δίκτυα βρίσκονται στην Ιρλανδία και την Ελλάδα και τα ερωτήματα που χρησιμοποιούμε γίνονται στα Αγγλικά και στα Ελληνικά αντίστοιχα. Επιλέον τα εξαγόμενα από τη μέθοδο μας δίκτυα απεικονίζονται και αξιολογούνται ποιοτικά από πολιτικούς επιστήμονες. Βασιζόμενοι στο γεγονός ότι οι σχέσεις στα πολιτικά δίκτυα εξελίσονται με το χρόνο, εφαρμόζουμε τη μέθοδο μας για να εξάγουμε δίκτυα για διαφορετικές χρονικές περιόδους και να απεικόνισουμε την εξέλιξη τους.

Αποδεικνύεται ότι η μέθοδος μας μπορεί να εκτιμήσει αποτελεσματικά σχέσεις συνεργασίας ενώ αποτυγχάνει να εκτιμήσει σχέσεις ανταγωνισμού. Επιπλέον, η μέθοδος μας μπορεί να προσδιορίσει τους πιο ενεργούς φορείς υπολογίζοντας το βαθμό κεντρότητας ένα ευρέως γνωστό μέτρο στην ανάλυση δικτύων. Τέλος, η απεικόνιση των δικτύων και της εξέλιξης τους οδηγούν σε ενδιαφέροντα αποτελέσματα και συμπεράσματα από τη μεριά των πολιτικών επιστημόνων.

Acknowledgements

First of all I would like to express my sincere gratitude to my supervisor, professor Alexandros Potamianos for his valuable guidance. I would also like to thank professors Vasilis Digalakis and Evripidis Petrakis for participating in the examining committee. Many thanks to Dr. Leeda Demetropoulou for her valuable help on this work from the side of the political scientist. I could not forget to thank my friend and colleague Elias for his help and useful advices and the rest of the group Nikos, Valia, Maria for spending 3 creative years of my life in the same office sharing our problems, our thoughts and our happiness.

Finally, I would like to thank my family and my friends who never stopped believing in me and supported me. This work is dedicated to all of them.

"Society does not consist of individuals but expresses the sum of interrelations, the relations within which these individuals stand."

Karl Marx.

Chapter 1

Introduction

Modern democratic governance has transformed from the hierarchical to more co-operative forms of public policy making[1]. During a policy making process, many partnerships are developed among organizations of different type (public-private) or governmental level (national-regional) in the policy arena[2]. The policy outcome then is the result of such political and economic interdependencies. A famous approach for the examination of the policy outcome is to express the set of these relations as a network. The policy network is used in political science to investigate social and financial phenomena, the creation of partnerships among actors[3] as well as explain and evaluate the different policy outcomes [1–3]. A policy network can be described as a social graph with nodes representing the actors and edges the linkages (or relations) among them. The relations in a policy network serve as channels for communication and the exchange of information, expertise, trust and other policy resources. In general, a policy network can be conceived as a special case of a social network, but we will see that there are significant differences between policy and social networks.

Policy network analysis is the procedure of analyzing, in a formal way, a policy making process using a network. The first step of the policy network analysis is the identification of the network under investigation. Traditionally it is a manual procedure performed by experts and requires refined techniques and extensive and time consuming collection of data through interviews and questionnaires. During the manual creation of policy networks, many subjective factors may be present, since this procedure relies strongly on the human subjects that participate in the interviews as the political scientists depend on a small number of respondents (sampled from the actors themselves). Such

[1]According to political science literature, policy making is the high-level development of official government policy.

[2]Policy arena is where the policy making process takes place i.e., a country or region.

[3]An actor in a policy network is a node assuming that the usual representation of a policy or a social network is a graph. In this work we use both terms interchangeably.

factors include personal opinions, the person's will for participation, even cultural issues. Overall, policy network identification requires a "large scale investment" that does not always "lead to breathtaking empirical and theoretical results". Furthermore, when lacking the resources for data collection and network analysis, political scientists often revert to qualitative analysis or construct the network topology using their intuition, significantly compromising the validity of their results.

It is mentioned above that policy networks can be considered as a special case of social networks. The fact is that networks in political sciences differ from their counterparts in social sciences in many aspects. In social networks, nodes usually represent persons and the edges the relations among them built on a ground of mutual understanding such as friendship or co-authorship. Actors in policy networks, on the other hand, can be organizations or even groups or unions of variable size and degree of formal organization. Furthermore, actors in policy networks might change name or even structure due to political[4] or economic factors. The relations among actors in policy networks usually signify the development of partnerships rather than a lax social relation. Relations in policy networks also depend on external factors such as the political environment [4], economic policies and funding at the local, national and supra-national level. Relations among policy actors can also be antagonistic rather than co-operative, or follow a more complex relationship of both co-operation and competition (sometimes refereed to by economists as coompetition). Often policy networks are studied at their infancy when the links among the actors are being formed and might not be directly observable through common action or direct communication. Another significant difference is that relations in policy networks evolute over time, e.g., some actors may intensify their relations of co-operation or create new relations with others. All these facts, imply that established features and algorithms for the extraction of social networks might not be directly applicable to policy network extraction. Moreover, the visualization of policy network evolution is a challenging task and of great importance to political scientists.

In this work we propose an algorithm of automatic extraction (or validation) of policy networks using information collected from the web. Specifically, the degree of relatedness (strength of link) between policy actors in a network is computed using three types of features on documents or snippets downloaded by a web search engine, namely: (i) the frequency of co-occurrence for each pair of actors (in web documents), (ii) the contextual similarity between snippets of web documents in which the actors appear, and (iii) the co-occurrence of hyperlinks presented in web documents that contain the actors. For each type of features and for their combinations, a variety of similarity metrics are used in order to estimate the link strength for each pair of actors. The proposed algorithm

[4]A recent example is the Kallikratis program of local government in Greece, according to which many political actors changed name and structure.

is not intended to substitute expert knowledge, but rather it should be viewed as a low-cost, semi-automated computational tool that can significantly contribute to policy network analysis. The proposed method aims to be efficient and minimizes subjective biases by exploiting the collective information from the web.

The research contributions of this work are summarized as follows:

- To our knowledge, this is the first comprehensive research effort towards the automatic extraction of policy networks.
- A variety of features extracted from the web data (hit counts, links and snippets) are proposed for estimating the relations among policy actors and their performance is examined on different types of relations. These features are motivated by recent research in the fields of information retrieval and natural language processing.
- The proposed method is unsupervised, semi-automatic and language independent. No previous knowledge resource is required except from the word-forms of actor names.
- Another important contribution is that the proposed features and metrics are evaluated against actual policy networks identified by expert political scientists. It is shown that the automatically extracted policy networks are capable of capturing the main relationships between policy actors are in broad agreement with networks built manually. The networks are also visualized and qualitatively evaluated by political scientists.
- The proposed method is applied on the extraction of policy networks over a selected time period. In this way we visualize the evolution of the policy networks and draw interesting conclusions.

The rest of the thesis is organized as follows, in Chapter 2 we refer to previous work from the fields of computational methods in politics, social network extraction, network visualization and semantic similarity computation. In Chapter 3, we formally define the metrics used and we present in full detail the proposed approach for policy network extraction. The experimental procedure is described in Chapter 4 and the results are presented and discussed in Chapter 5. Finally, we conclude our work giving new directions for further research in Chapter 6.

Chapter 2

Related Work

The approach of policy network extraction presented in the next chapters is based on previous work especially from the fields of: (i) traditional methods of network analysis, (ii) network visualization, (iii) computational methods in politics (iv) computational methods in social network extraction v) computational methods in semantic similarity. In this chapter we review the prominent methods or systems developed in the above research fields.

2.1 Traditional methods of network creation and analysis

Social networks were first introduced in 1930s [5] in the form of graphs to represent the interpersonal relationships among individuals. Since then social networks consist a powerful tool for many research fields such as social and political sciences and economics. In addition, the exploitation of graph theory gave birth to another research field, the social network analysis which is discussed next.

2.1.1 Network data gathering

Social network analysis is the field in social sciences who aims to examine and analyze social phenomena by applying measures and models from graph theory. The first step for social network analysis is the creation of the network. Traditionally social scientists manually build networks gathering data from interviews and questionnaires but other methods such as archives, diaries and electronic traces are used [6]. Network data are obtained with questions that ask a respondent to enumerate those individuals with whom he or she (or an organization for which he or she is an agent) has direct ties of a specific type. In cases of limited populations, respondents are asked to recognize their contacts from a list of individuals. Surveys and questionnaires have also been used by political scientists, to study interorganizational relations, through interviewing one or more individuals in an organization about communication, resource transfers, and joint activities with other organizations. Yet problems considering the selection of individuals arise. Most studies select a small number of individuals to report on the organization's relations to all other organizations, but the individuals' responses are biased on their specialty or activities. Unfortunately, the quality of network data obtained by surveys and questionnaires is far from perfect, and gathering such data often requires substantial research budgets [6]. On the other hand, exploiting archival sources of various kinds is cheaper. Information about relationships between banks or corporations can be obtained from records including the names of persons who are directors of major corporations e.g., organizations having one or more directors in common are assumed to be related. Archives of citations are used to identify communities of experts based on the notion that scientists whose work is cited by the same authors are assumed to be related. Yet, the problem of data quality and reliability remains an open issue [6].

2.1.2 Network creation and analysis

Having gathered the information required, social scientists manually create the social network using an adjacency matrix with each cell denoting the strength of the corresponding relation. To measure the strength of ties social scientists use strength indicators, measurable quantities that indicate the strength of a relation e.g., for the case of interpersonal relations these indicators are the amount of time, the emotional intensity, the intimacy (mutual confiding) and the reciprocal services considering the tie [7, 8]. The results of the indicators are then mapped to a values from a numerical scale. Once the network has been created the analysis takes places which consists of five levels [9]: i) *Actor level*, ii) *Dyadic level*, iii) *Triadic level*, iv) *Subset level*, v) *Network level*. All these levels of analysis are discussed next. The most widely used measure considering the actor's level is the centrality. According to [9] the most significant centrality measures are degree, closeness and betweenness centrality. The *degree centrality* of a node defines the total number of relations the node participates in. It is based on the idea that an actor with a large number of links is more active and might be of more interest than other actors. The *closeness centrality* is obtained by calculating the average geodesic distance of a node to all other nodes in the network. Closeness centrality thus characterizes whether the node of interest is easily reachable by any other node in the network. Easily reachable nodes are assumed to influence more than others the policy outcome. The *betweenness centrality* identifies whether an actor is either a broker or not by calculating the number of shortest paths that interconnect all other nodes and pass through the node of interest. Brokers are nodes of great importance as they often bridge different groups or communities in the network. Thus they serve as channels of communication between the different groups or communities. Considering the dyadic level, the most commonly used is the *structural equivalence* according to which actors of the same or similar social environment are clustered in the same block (or group) e.g., actors with the same friends. Other forms of equivalence such as *automorphic equivalence* or *maximal regular equivalence* exist in literature [9]. In the *triadic level* of analysis the network is assumed to be a set of triads (triangles of nodes) with the relations signed as positive or negative (e.g., a positive relation denotes friendship and a negative antagonism). Triads in signed networks follow the balance theorem [10], according to which in a balanced network every two positively related actors tend to be mutual friends or enemies with a third actor. In the *subset level* of analysis the most oftenly used measure is the number of *cliques* which is defined as the number of complete subgraphs in the network under investigation. Finally, at the *network level* of analysis the most popular measure is the density of the graph which is defined as the proportion of ties to the total number of possible ties.

2.2 Network visualization

The first network visualizations, namely sociograms, emerged in 1930s [5] and were handmade. In a sociogram the actors are represented by circles placed on horizontal lines connected by arrows that represented the relations among them. The central aspect in the first network visualizations was to be readable. A widely used approach towards the readability of sociograms required that arrow diagrams were drawn in which the most central actors were placed in the center and the researcher tried to reduce the number of cross-cutting connections to achieve the best possible clarity, a problem which is widely known as the *crossing minimization problem*. A lot of research has been made towards the solution of crossing minimization problem which is a problem critical to the readability of the graph. In [11], the nodes were grouped according to their centrality and then placed on concentric circles. In [12] a similar approach of network visualization was proposed placing the more central nodes farther outside the network. Other approaches developed are based on heuristics e.g., [13] and require NP-complete solutions. Heuristics such as *Barycenter* [14], *Median* [15] and *Random-Key* [16] have achieved satisfactory results. Although network visualizations (even in their early forms) gave new directions and possibilities to many research fields (psychology, social sciences, political sciences and economics), their creation was a tiresome and time-consuming procedure. The development of computers gave new possibilities in the network visualization. A great variety of visualization algorithms, techniques and tools have been proposed. The vast majority of modern visualization algorithms incorporate multidimensional scaling (MDS) [17], which efficiently visualize the network in 2 or 3 dimensions preserving the distances among the nodes. More specifically, a widely used class of algorithms for network visualization of general purpose are based on "spring embedder" layout. The most prominent spring-embedder algorithm has been proposed in [18] according to which each edge is represented as a spring model that can be compressed or stretched; the whole graph is a system consisting of the set of individual spring models. The desirable graph is the one with the minimum total spring model energy. In [18] it is shown that the proposed spring-embedder algorithm is a case of multidimensional scaling. Thus, to produce the graph, the algorithm iteratively minimizes a functional of the differences between the desirable distances of nodes and the actual ones. In [19], theoretical aspects for the problem of network visualization are discussed and directions for more effective visualizations are given. In [20], an algorithm for more effective visualization of policy networks is presented that incorporates centrality measures. More specifically, the basic layout of the produced graphs is based on the approach of [11] (all nodes are placed in concentric circles according to their centrality), then a three-phase layout algorithm is proposed that places the nodes of reciprocated[1] relations as a core graph and the nodes of non-reciprocated relations in the periphery of the concentric circles. At each phase of the algorithm an energy function (similar to [18]) is minimized. In addition, the output of the algorithm can be processed by the user for more exploratory analysis. Furthermore, tools such as PAJEK[2] or UCINET[3] have been developed that can compute statistical measures such as different centrality types, network densities and also visualize the networks.

The goal of all the above methods and techniques is the more readable and informative visualization of static networks i.e., the time dimension is not taken into account. Unfortunately, the structure of social or policy networks evolutes over time. The visualization of this evolution plays a key role for the explanation of policy outcomes and other phenomena. Thus dymanic network visualization is required. A great variety of methods and techniques have been proposed towards efficient dynamic network visualization. According to [21], the authors divide network movies into i) static flip books, where node position remains constant but edges change over time, and ii) dynamic movies, where nodes move as a function of changes in relations. Flip books are particularly useful in contexts where relations are sparse. For more connected networks, movies are often more appropriate. The basic idea for dynamic network visualization is to split the movie duration in overlapping slices and produce the desired animation via interpolating the changes in the node positions (coordinates) and the strength of ties among them. In

[1]Reciprocated relation is called the relation that is confirmed by both participating actors.
[2]http://vlado.fmf.uni-lj.si/pub/networks/pajek
[3]http://www.analytictech.com/ucinet

general all of the aesthetic for static graphs can be applied to dynamic graphs. In [21], the *SONIA* system is presented that implements many graph layout algorithms and can produce either flip book visualizations or dynamic movies allowing the user to tune a great number of parameters (slice duration, layout algorithm, node size e.t.c). In [22, 23], the problem of dynamic network visualization is discussed in detail and directions for more effective visualization of network evolution are given.

2.3 Computational methods in politics

The huge amount of political information available such as transcribe speeches web-based information and blogs has supplied political scientists with automated tools and methods for more efficient and quick political analysis, than that of traditional methods [24, 25].

Computational methods in political science have been developed the last few decades and the research has been focused mainly on the following directions:

- The *estimation of policy dimensions* of political parties, e.g., classify if a political party a right or left ideology.
- The *opinion mining*, e.g., to classify whether the piece of text under examination expresses support or opposition to a specific topic.
- The *selection of features* that express political conflict or opinion.
- The study of online *political activism*.

Many of state-of-the-art methods from natural language processing and information retrieval have been applied to solve the above challenging tasks. The methods of computational politics developed so far exploit textual data such as political manifestos, transcribed speeches and web-based data such as web documents and political blogs [26, 27].

One of the most prominent works in computational politics is in [28, 29], where the *Wordscores* system is proposed which extracts policy dimensions of political parties based on word frequencies from manifestos. Words taken from a set of training texts ('reference texts') are automatically scored according to their relative frequencies. In this way a dictionary with words their political dimension scores is created. The system scores each text from a test set ('virgin texts') using the average of the dimension scores of the words contained in the specific text. Then the computed numerical values of the virgin texts are translated in specific policy dimensions. In addition, the uncertainty of each score is measured as the variance between each word score and the text's total score. A advantage of *Wordscores* is that is fully automatic and does not require any previous knowledge for the language of the texts. However, its performance is highly dependent on the selection of appropriate training data. Another problem of *Wordscores* is that it cannot be easily used for time-series estimation of policy dimensions. Based on *Wordscores*, in [30] the *Wordfish* system is proposed which mines time-series policy dimensions of parties. Its difference from *Wordscores* is that the computation of word scores is less biased by the training data, as it is assumed that word frequencies are produced by a Poisson process.

Another active area relevant to political scientists is the opinion mining from political information. This area uses techniques and models from sentiment analysis and the data exploited are text, blogs or transcribed speeches. In [31], information about the relationships between discourse segments is used to enhance the performance of opinion classifiers from congressional transcripts. More specifically, their approach incorporates weighted links between speech segments of agreement in the classification function. However, they don't employ selection of lexical features for further classification improvement. In [32], a model for the extraction of political standpoints is proposed. The model scores opinion statements by incorporating subjective, topic and opinion features. Also the produced scores are used to extract sentences that best describe political opinions. More specifically the model's scoring function uses nouns, adjectives and verbs as features as well as their combination. It is shown that adjectives are the most important features for the mining of political standpoints at the level of sentences. In [33], a text classification algorithm is applied to legislative speeches to extract the words that indicate ideological positions. A basic SVM classifier is used to classify the Senators in liberal and non-liberal from congressional texts. The classifier was evaluated on different groups of features (nouns, adverbs, verbs e.t.c) and weighting schemes. It is shown that the tf-idf weighting scheme improves the classification accuracy for the examined feature groups.

Many methods in computational politics make use of lexical features (words or phrases) that are more descriptive for a specific political opinion. In [34], different techniques for selecting words that carry political conflict are discussed. The drawbacks of feature selection based on simple statistical measures (such as word frequency, tf-idf scoring, stop words removal) are presented and discussed. A model-based technique is proposed that uses Bayesian shrinkage as regularization process. The proposed approach is evaluated and applied on different tasks such as the examination of the word polarity through time. Computational methods in politics have used social media, such as the blogosphere and social network services (SNS). In [35], lexical features are combined with social information extracted from blogs to classify political sentiments during the 2008 U.S Presidential election. More specifically, at first step a baseline classifier was built from annotated blog posts and at a second step the entities included in each post were given sentiment scores. Furthermore, social information from the bloggers' mined network is incorporated. It is proved that the combination of sentiment scored features with social network features enhances the classification accuracy.

The development of social networking services and blogs made the study of online activism using computational methods possible, especially in countries under authoritarian regimes. In [36], online activism is studied by processing social information gathered from the Twitter network service during the Iran Elections of 2009. The automatically gathered tweets were given as input for three types of analysis: i) using the histogram of tweets, ii) using visualization of the generated networks and ii) using word frequencies. All three types of analysis led to many interesting conclusions for the behavior of online activism during Iranian elections. In [37], the Iranian blogosphere is analyzed and the different political poles are identified, using manually coded data, term frequencies and outlinks.

2.4 Computational methods of social network extraction

Political analysts apply network analysis techniques in manually mined policy networks. Regarding the fact that policy network extraction can be considered as a special type of social network extraction, we have to mention the work conducted in this active research area. For the web-based social network extraction methods, the most common feature used for relation identification is the frequency of co-occurrence of the related actors in web documents. However, other features such as lexical context, keyphrases, log files and e-mail information are also used. First studies in automatic web-based social network extraction were focused on networks of academics and researchers. The basic idea of these studies is that co-citation and co-authorship is an indication of relationship between scientists. In [38], the *Referral Web* system is developed whose goal is to connect a user to an expert of a specific topic exploiting the referral-chaining process. For an individual user a network of experts in a requested topic is constructed iteratively. In each iteration step the system constructs the path of links starting from the user's personal name to the name of an expert. The links are identified using the frequency of co-occurrence of names in web documents. In [39], the *Flink* system is developed, similar to the *Referral Web*, which creates the ego-network (personal social network) of a researcher based on the occurrences and co-occurrences of individual names. Furthermore, *Flink* associates researchers to certain areas of research using the frequency of the researcher name in pages of the topic of interest. In [40], the social network information is incorporated into the ontology representation leading to tripartite models of ontologies.

Apart from the relation identification, relation labeling is also important for the complete extraction of social networks. In [41], *Polyphonet* system is developed that uses web co-occurrences of names to extract the network of conference participants. *Polyphonet* labels the identified relations using simple classification approach and a predefined set of relationship types. To tackle the problem of name ambiguity the queries were expanded with characteristic words and keyphrases that were mined from the clusters produced from the clustering of the retrieved web documents. In [42], social networks are extracted in similar way to *Polyphonet* but the relationship types (for the relation labeling process) are enriched automatically after mining possible relation keywords (nouns and verbs) from the sentences in web documents where the two actors co-occur. The criterion of a relation keyword to be selected is the association score between the word and the relationship label and it is computed using the number of co-occurrences of the label-keyword pair. In [43], the relationship labels are automatically extracted using the collective context of clustered actor pairs. The relation identification process is still based on the web hit counts. The actor pairs are clustered according to their common context in web documents and the resulting clusters are used to extract the relation keywords. The possible words are scored according to a simple tf-idf scheme and those with the highest score are selected. In [44], a model is proposed, that learns the entity and relation topics from a social network and assigns descriptive words to them. The model was tested on large text corpora such as the Bible and Wikipedia.

Except from the Web documents as resource, social networks can be extracted from other types of data that carry social information such as email messages, threaded discussions and interaction activity in social networks. In [45], a system that process email messages to extract people names and mines the persons' personal pages or pages that the name exists. Contact information is then extracted from the pages using conditional random fields (CRFs). In addition, the system can extract keywords that describe each person's expertise, using the information gain between the person and the possible keywords. In [46], social network extraction is achieved through processing of threaded discussions between individuals. The proposed method identifies the relations between individuals and measures their strength. The relations are first mined using the number of references to personal names in postings. Then to measure the relation strength, the proposed method counts the number of common nouns and verbs in the postings. In [47], an unsupervised model is developed that estimates the relationship strength in online social networks, using the users' interaction activity (e.g., communication, tagging). More specifically, a relation is assumed to be a hidden effect of user profiles similarities, thus a link-based latent variable model is used to infer the possible relations among the users. In [48], the log files of shared workspaces are used to extract user-oriented and object-oriented social networks. In both cases of network extraction, the log records are translated in RDF triplets which are used for relation identification.

Other resources have been also used for social network extraction, such as news articles, literary fiction and blogs. In [49], social networks are extracted and updated over time using monolingual or multilingual news from articles. Live news of different languages are gathered in the form of text snippets using the *European Media Monitor* system. With the use of a multilingual named entity recognizer named entities are extracted which constitute the actors of the networks. Two actors are assumed to be related if they co-occur in a text snippet. In [50], the *SONEX* system is proposed that automatically extracts networks of entities from the blogosphere. According to *SONEX*, the entities that appear in the same sentence and in close proximity are assumed to be related. The entities are extracted using a standard named entity recognizer. *SONEX* also labels the extracted relations by clustering the entity pairs according to context. The relation labels are selected by the collective context of each produced cluster, an approach which is similar to [43]. In [51], social networks of persons from the literary fiction are extracted using quoted speech (dialogue interactions). Characters of literary fiction are identified using a standard named entity recognizer. Two characters are assumed to be related if they participate in a conversation or dialogue.

2.5 Computational methods in semantic similarity

The computation of semantic similarity between words or terms plays a key role in many tasks of natural language processing and computational linguistics. Consequently, a vast number of methods and metrics that measure semantic similarity has been proposed in literature. The proposed semantic similarity metrics can be classified into four main categories depending on the knowledge resources:

- Supervised *resource-based metrics*, that use human-built knowledge resources, such as handcrafted ontologies.
- Supervised *knowledge-rich text-mining metrics*, i.e., metrics that perform text mining but also rely on knowledge sources.
- Unsupervised *co-occurrence metrics*, i.e., metrics based on the assumption that the semantic similarity between the words or terms can be expressed as a function of their co-occurrence.

- Unsupervised *text-based metrics*, i.e., metrics that are fully text-based and rely on the assumption that the semantic similarity between words or terms can be expressed as a function of the shared context.

Many resource-based methods have been proposed that use human created resources e.g., Wordnet, for semantic similarity computation. Edge counting methods are based on the idea that the minimum number of edges two separate concepts can be used as a measure of their conceptual distance [52]. In [53], the semantic similarity between terms is computed as the combination of different conceptual properties of the taxonomic net. Furthermore, in [54], semantic similarity is computed as a function of depth and path length of the words in the Wordnet.

Considering the knowledge-rich text-mining metrics, in [55], page-count-based similarity scores and lexicosyntactic patterns are used to train an SVM classifier to classify synonymous and non-synonymous word-pairs and estimate their semantic similarity. The unsupervised co-occurrence metrics attempt to implement computational models for the notion of "word association," which is used in psycholinguistics, a procedure of lexical decision of human associative memory [56]. Specifically, in [57], a web-based unsupervised algorithm (PMI-IR) for recognizing synonyms is presented. PMI-IR uses point-wise mutual information (with different scoring functions) to measure the semantic similarity for pairs of words and capture the synonymy. It is proved that PMI-IR outperforms the LSA (for more details read [58]) for the case of synonymy identification. In [59], the *google similarity distance* is proposed according to which the semantic similarity between two words or terms is a function of the information distance for the specific word pair. The information distance can be expressed as a compression function. According to [59], the google distribution can play the role of the compression function. The universality of the google distribution makes the specific metric universal.

Many unsupervised similarity metrics use the vector space model (VSM) according to which the words or terms are represented as vectors in a high dimensional space [60]. Unsupervised approaches that are based on VSM are shown to work efficiently and their performance is very close to the supervised methods. In [56, 61], similarity scores for the Miller-Charles dataset are estimated in an unsupervised way, using page-count-based and context-based metrics. Their performance is evaluated on human ratings. It was proved that context-based metrics outperform the page-count-based metrics and the greater the corpus the more accuracy is succeeded.

Except from co-occurrence and context, links are also used as feature for the computation of semantic similarity. In [62], a novel link-based similarity measure for web-pages is proposed according to which the similarity between two pages can be computed using the common inlinks. In [63], traditional link-based similarity measures such as bibliographic coupling and co-citation were combined with content-based classifiers for improved web page classification.

Chapter 3

Extraction and Visualization of Policy Networks

In this chapter we describe in full detail our approach for the extraction of the policy networks. First, we formally define the metrics used for the estimation of the strength of relations (degree of relatedness) among the actors of the policy network. Second, we present the whole method which consists of the query formation, the data gathering, the computation of the strength of relations and finally the visualization of the networks. Our approach is automatic and requires no previous knowledge resource, other than the different word forms of the actors names whose extraction remains a manual procedure. Finally, we describe the extraction of the networks for the different years from a time period and how the extracted networks are used to visualize the network evolution.

3.1 Relatedness metrics

Next we describe and define the types of relatedness metrics for the computation of the strength of ties among political actors. The basic idea behind the proposed metrics is that the relationship strength between two actors of a policy network can be measured automatically using only web-based features. All the metrics defined in this section require a web search engine. The metrics are of three types: (i) page-count-based, (ii), text-based and (iii) link-based metrics. Each metric, explores different features, capturing different perspectives of web information. Page-count-based metrics use co-occurrence of the names (or acronyms) of the actors in web documents or snippets. Text-based metrics compute the lexical similarity between the context in which the political actors appear in web documents or snippets. The link-based metrics are based on the shared hyperlinks (links or outlinks) among the web documents that contain the actors of interest. Finally, we propose linear combination of the three metrics.

3.1.1 Page-count-based metrics

According to this type of metrics, the strength of relationship between two actors is estimated as an association ratio that is a function of the co-occurrence frequency of the actors in web documents. The assumption of these metrics is that *related actors tend to co-occur in web documents*. Co-occurrence implies that both actors deal with common policy issues or serve similar policy functions. We employ four page-count-based similarity metrics used in literature, namely: Jaccard coefficient, Dice coefficient, mutual information (as defined in [55]) and google-based semantic relatedness (see [64]). The four metrics are formally defined next. For all four page-count-based metrics, we use the notation in Table 3.1.

Notation	Description
$\{D\}$	set of all documents indexed by search engine
$\|D\|$	number of documents in $\{D\}$
a_i	a political actor a_i
$\{D_{a_i}\}$	set of documents indexed by a_i, $\{D_{a_i}\} \subseteq \{D\}$
$\|D_{a_i}\|$	number of documents in $\{D_{a_i}\}$
$\{D_{a_i,a_j}\}$	set of documents indexed by a_i and a_j, $\{D_{a_i,a_j}\} \subseteq \{D\}$
$\|D_{a_i,a_j}\|$	number of documents in $\{D_{a_i,a_j}\}$

Table 3.1: Definitions for indexed documents.

Jaccard coefficient: Generally this coefficient computes the similarity between sets. In our case, we consider the sets of web documents that are indexed by the actors of interest. Lets say that actors a_i and a_j exist (at least once) in D_{a_i} and D_{a_j} sets of web documents respectively. We can estimate how much related the actors a_i and a_j are by measuring the similarity between their corresponding document sets. Thus, the Jaccard coefficient S_J^P between actors a_i and a_j is defined as follows:

$$S_J^P(a_i, a_j) = \frac{|D_{a_i,a_j}|}{|D_{a_i}| + |D_{a_j}| - |D_{a_i,a_j}|}. \tag{3.1}$$

For identical actors the Jaccard coefficient assigns the maximum similarity score of 1. For unrelated actors a_i, a_j that never co-occur the Jaccard coefficient is 0.

Dice coefficient: This coefficient is closely related to the Jaccard coefficient and it is defined as:

$$S_D^P(a_i, a_j) = \frac{2|D_{a_i,a_j}|}{|D_{a_i}| + |D_{a_j}|}. \tag{3.2}$$

As before, (3.2) is equal to 1 and 0, for absolute similarity and dissimilarity, respectively.

Mutual information: Assuming that the number of occurrences of actors a_i and a_j ($|D_{a_i}|$, $|D_{a_j}|$) are random variables, then their point-wise mutual information [65], reflects the dependence between the occurrence of a_i and a_j as follows:

$$S_I^P(a_i, a_j) = \begin{cases} \log \frac{\frac{|D_{a_i,a_j}|}{|D|}}{\frac{|D_{a_i}|}{|D|}\frac{|D_{a_j}|}{|D|}} & \text{if } |D_{a_i,a_j}| > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

Eq. 3.3 compares the probability of observing actors a_i and a_j together (joint probability), with the probabilities of observing a_i and a_j independently. The greater the joint probability the greater the association and consequently the strength of relation between a_i and a_j. For identical actors the mutual information equals an unbounded positive value. If two actors never co-occur ($|D_{a_i,a_j}| = 0$), (3.3) is undefined and the similarity score is set to 0.

Google-based semantic relatedness: The "normalized google distance" is another page-count-based similarity metric that was proposed in [59, 66], defined as follows:

$$S_R^P(a_i, a_j) = \frac{\max\{\log|D_{a_i}|, \log|D_{a_j}|\} - \log|D_{a_i,a_j}|}{\log|D| - \min\{\log|D_{a_i}|, \log|D_{a_j}|\}}. \tag{3.4}$$

This metric is a dissimilarity measure, i.e., as the distance between two actors increases the metric takes smaller values. The scores assigned by (3.4) are unbounded, ranging from 0 to ∞. In [64], a variation of the normalized google distance was used, proposing a bounded similarity measure called "google-based semantic relatedness", defined as:

$$S_G^P(a_i, a_j) = e^{-2S_R^P(a_i,a_j)}, \tag{3.5}$$

$S_R^P(a_i, a_j)$ is computed according to (3.4). The google-based semantic relatedness is bounded in [0,1].

3.1.2 Text-based metric

In this section, we present a text-based metric that computes the strength of relation between two actors by examining the lexical context in web documents where the two actors are mentioned. Specifically, the bag-of-words model is applied for every actor. The lexical feature vectors are extracted from the left and right context of actors and the cosine similarity between the vectors is computed. The computed similarity score expresses the degree of relatedness between the two actors. The fundamental assumption behind this metric is that *related actors have similar syntactic, semantic and topical features.* For example, if two actors have similar political activities, then we expect these activities to be mentioned in the close lexical vicinity of the actors. The text-based metric applies a contextual window of predefined size, W, for extracting the lexical features for an actor a_i. In particular, a window containing the W words preceding and the W words following any instance of the actor a_i is used for feature extraction, i.e.,

$$[f_{W,L} \;\; \dots \;\; f_{2,L} \;\; f_{1,L}] \;\; a_i \;\; [f_{1,R} \;\; f_{2,R} \;\; \dots \;\; f_{W,R}],$$

where $f_{j,L}$ and $f_{j,R}$ represent the j^{th} feature (in this case word) that exist to the left and to the right context of a_i, respectively. Given a fixed value of W, a feature vector for a_i is built as $V_{a_i,W} = (v_{a_i,1}, v_{a_i,2}, ..., v_{a_i,N})$, where $v_{a_i,j}$ is a non-negative integer and W is the context window size. The feature vector has N elements, where N is the vocabulary size. The feature value $f_{a_i,j}$ corresponds to the occurrence of vocabulary word v_j within the left or right context window W of a_i. The value of $v_{a_i,j}$ can be a function of the frequency of occurrence of v_j in the context of a_i. More specifically the value of $v_{a_i,j}$ can be defined according to one of the weighting schemes defined in Table 3.2, namely, binary (B) or logarithm of word frequency (LTF). In Table 3.2, $c(f_{a_i,j})$ denotes the number of occurrences of the word (or feature) v_j in the left or right context of actor a_i, and $c(a_i)$ is the number of the occurrences of a_i. Note that the value of $v_{a_i,j}$ is set by considering all the occurrences of a_i in the corpus.

Scheme	Acronym	Equation ($c(f_{a_i,j}) > 0$)
Binary	B	1
Log of TF	LTF	$\frac{\log(c(f_{a_i,j}))}{\log(c(a_i))}$

Table 3.2: Weighting schemes for the text-based metric.

Once a weighting scheme is selected, the context-based metric S^T_W computes the similarity between two actors, a_i and a_j, as the cosine of their feature vectors, $V_{a_i,W}$ and $V_{a_j,W}$, as follows:

$$S^T_W(a_i, a_j) = \frac{\sum_{l=1}^{N} v_{a_i,l} v_{a_j,l}}{\sqrt{\sum_{l=1}^{N} (v_{a_i,l})^2} \sqrt{\sum_{l=1}^{N} (v_{a_j,l})^2}}, \tag{3.6}$$

where W is the context window length and N is the vocabulary size. The cosine similarity metric assigns 0 similarity score when a_i, a_j share no context (completely dissimilar actors), and 1 for identical actors (or actors sharing the same contexts).

3.1.3 Link-based metrics

In this section, we define two link-based relatedness metrics for computing the strength of reationship between two actors. The link-based metric exploits the hyperlinks that exist in the downloaded web documents. These hyperlinks are usually referred to as “outlinks” [63]. It is expected that hyperlinks will point to topically relevant web sites and documents [67]. Thus, the idea of using outlinks as features is that related actors will share common topics of interest, e.g., the pages of two related public organizations might refer to common hyperlinks. The links are being used in two different forms: (i) the full form where the whole path is specified (excluding the actual document specified in the link) and (ii) the base form where only the main website address is used. An example of the base and full form of an link is presented in Table 3.3.

Form	Example
Full	`www.ypes.gr/el/MediaCenter/Minister/`
Base	`www.ypes.gr`

TABLE 3.3: Forms of outlinks.

Usually the full form of a link points to a specific thematic area of the parent web site, while its base form corresponds to the top-level web site. In the example above, the full link points to the Press Secretary of the Ministry of Interior of Greece, while the corresponding base form points to the top-level site of the Ministry of Interior of Greece. The information that is provided by the full and base links is strongly related. However the full links are more topic-specific.

For each actor a_i, we consider the set of (full or base) links $\{O_{a_i}\}$ that appear in web documents where this political actor is mentioned. The similarity between two actors a_i and a_j is computed according to the overlap between the members of their link sets $\{O_{a_i}\}$ and $\{O_{a_j}\}$ respectively. For the computation of the link-based relatedness score, variations of (3.5) and (3.6) are employed as described next.

Google-based semantic relatedness using outlinks (S_G^L): We apply the metric of (3.5), using the set of links, instead of document sets. Specifically,

$$S_R^L(a_i,a_j) = \frac{\max\{\log|O_{a_i}|,\log|O_{a_j}|\} - \log|O_{a_i,a_j}|}{\log|O| - \min\{\log|O_{a_i}|,\log|O_{a_j}|\}}, \tag{3.7}$$

where $\{O_{a_i}\}$, $\{O_{a_j}\}$ and $\{O_{a_i,a_j}\}$ the set of links for actors a_i, a_j and jointly for both a_i and a_j, respectively, i.e., $\{O_{a_i,a_j}\}$ is the intersection of $\{O_{a_i}\}$ and $\{O_{a_j}\}$. We then normalize S_R^L into S_G^L using (3.5).

Cosine similarity using outlinks (S_T^L): Alternatively, for each actor a feature vector is built using the members of the set of links. The relationship strength between two actors is computed as the cosine of their feature vectors in the same fashion as (3.6) (here the window size parameter W is irrelevant and is not specified). The feature values can be set according to the weighting schemes defined in Table 3.2, i.e., using binary or logarithm of link frequency.

3.1.4 Linear fusion of relatedness metrics

Each of the aforementioned metrics use a different type of feature to estimate relatedness scores between two actors, i.e., actor co-occurrence for page-count metrics, lexical contextual similarity for text-based metrics and outlink similarity for link-based metrics. Here we propose a combination of these features using late integration, i.e., combine the relatedness scores from the three types of metrics. Only simple linear fusion is investigated here and the composite relatedness score S between actors a_i and a_j is defined as:

$$S(a_i, a_j) = \lambda_P S^P(a_i, a_j) + \lambda_T S^T(a_i, a_j) + \lambda_L S^L(a_i, a_j), \tag{3.8}$$

where S^P, S^T, S^L refers to the proposed page-count, text and link-based metrics, respectively, and λ_P, λ_T, λ_L are the corresponding weights. Two cases are investigated: equal weights (that sum up to 1) and inverse variance weighting (informative fusion). For informative fusion, the weights for each type of metric are set equal to the inverse variance, e.g., $\lambda_P = 1/\sigma_P^2$. The variance is computed across the relatedness scores for all actor pairs and a specific metric.

3.2 Extraction of Policy Networks

In the current section we present the whole process for the estimation of the degree of relatedness among the actors of a policy network as well as its visualization. Before the description of the proposed approach, we have to mention that throughout this work, the representation we use for the policy networks (extracted or human rated) is the adjacency matrix with each cell $r_{i,j}$ denoting the relatedness score $r_{i,j}$ for the corresponding pair of actors a_i and a_j. All the adjacency matrices are square N×N, where N denotes the number of actors, and symmetric meaning that $r_{i,j} = r_{j,i}$. In Figure 3.1, the form of adjacency matrices used in this work, is shown.

The basic idea of our approach is that the strength of the relations between the actors of a policy network can be estimated using the number of co-occurrences of actor names in the web, their lexical context or the number of outlinks. Our method takes as input

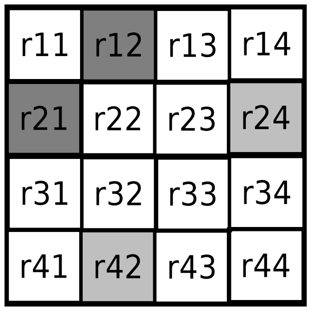


FIGURE 3.1: Policy networks are in the form of adjacency matrices (cells of the same color denote equal relatedness scores).

the actor names as they are given from political scientists and their word-forms, and extracts the policy networks as graphs. The process consists of four steps as shown in the flow diagram in Figure 3.2:

- *Step 0*: The derivation of the lexicalized forms that describe each actor is done manually.
- *Step 1*: Two types of queries are automatically created using the lexicalizations of actor names from *Step 0*, (i) individual queries (IND) and (ii) AND queries. Then the web data is downloaded as required for each of the relatedness metrics using a web search engine.
- *Step 2*: The relatedness scores are computed using the equations defined in Section 3.1.
- *Step 3*: The extracted networks are visualized as graphs.

Next, we provide a detailed description for each of these steps.

Step 0: Lexicalization of actors. A crucial step towards the successful extraction of the policy networks is the derivation of the lexicalized forms that describe each of the actors. These forms are usually multi-word terms consisting of more than three words and in many cases are also denoted by abbreviations. For example, the actor "Industrial Development Authority" is also denoted as "IDA". It is quite common for an actor to have alternative lexicalizations, e.g., the previous actor is also referred as "Industrial Development Agency". When only the official (long) names of actors are used very few relevant documents (hits) are returned by search engines. On the other hand, when only abbreviations are used, web queries return many results. However, most of the retrieved documents are irrelevant due to the inherent ambiguity in abbreviations (many unrelated entities often share the same abbreviation). In order to tackle both the data sparseness and term ambiguity problems, for each actor a number of lexicalized

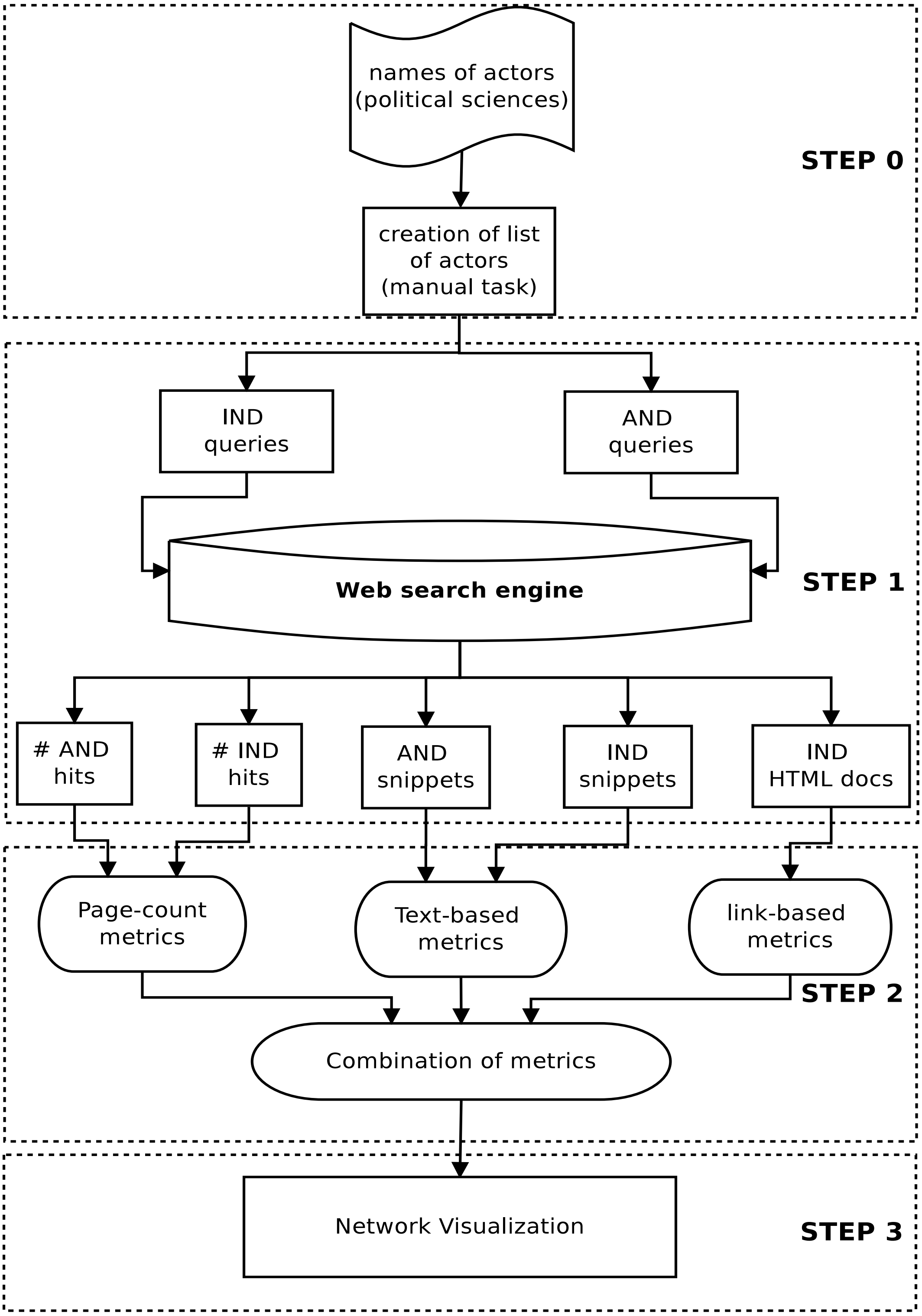


Figure 3.2: Diagram of the network extraction process.

forms that combine both multi-word terms and abbreviations is manually selected in collaboration with political scientists: The named entities more relevant to the actor were extracted manually from the top 20-30 hits and then the political scientist selected a set of them as candidates. Each lexicalization was then tested using web searches and looking at the relevance of the top 20-30 documents. The process was repeated 2-3 times to acquire the final set of lexicalization.

Step 1: Retrieval of web data. Once the set of actor names is created, we search the web in order to retrieve data, for this purpose we use a web search engine. In this work, the search engine used is the Yahoo Search API[1]. To retrieve the web data required two different query types are given as input to the search engine: (i) individual (IND), e.g., "a_i", and (ii) conjunctive (AND), e.g., "a_i AND a_j". The IND type concerns individual actors, while the AND type requires the co-existence of the two actors in the returned data. We consider three different types of information returned by the search engine: (i) page counts (hits), (ii) URLs of web documents, and (iii) their corresponding snippets. In order to acquire the outlinks of the web documents, we employ a further downloading step using the returned URLs. The outlinks are extracted using HTML::SimpleLinkExtor[2]. Examples of IND and AND queries are presented in Table 3.4.

Query type	Query
IND	"Industrial Development Authority" OR "Industrial Development Agency" OR "IDA"
	"Limerick City Council" OR "Limerick City Co"
AND	("Industrial Development Authority" OR "Industrial Development Agency" OR "IDA") AND ("Limerick City Council" OR "Limerick City Co")

Table 3.4: Examples of IND and AND queries used in our approach.

Step 2: Computation of relatedness. Relatedness scores are computed according to the metrics defined in Chapter 3: (i) page-count-based (S^P), (ii) text-based (S^T), and (iii) link-based (S^L). For the S^P metrics, we use the page counts that are returned by IND and AND queries. The S^T metric is applied on snippets that are retrieved using either IND or AND queries. The S^L metric operates on the outlinks of documents that are downloaded using IND queries. For each relation under examination, we compute the corresponding relatedness score using one of the three types of metrics. The computed relatedness scores have different value ranges and are all normalized using simple min-max normalization (see Eq 4.1). The final scores are then stored in matrices (see Figure 3.1). The scores are linearly combined according to the Eq. 3.8. Matrices are also created for the linear combinations of metrics. In Algorithm 1 the pseudocode of the process for the retrieval of web data, the computation of relatedness scores and their

[1] http://search.cpan.org/~timb/Yahoo-Search-1.11.3/lib/Yahoo/Search.pm

[2] http://search.cpan.org/~bdfoy/HTML-SimpleLinkExtor-1.23/lib/SimpleLinkExtor.pm

linear combination is presented. Algorithm 1 takes as input the set actors A and the set of lexicalizations for each actor a_i. Specifically, in lines 1 to 6 (see Algorithm 1) for each actor individually the IND query is created (function *BuildINDQuery*), with the IND query as input the hits for the actor are returned by the search engine (*GetPageCounts*), the M top-ranked URLs of the web documents that the actor exists (*GetDocumentsURLs*) and the corresponding snippets (*GetDocumentSnippets*). In lines 7 to 10, for each retrieved URL using IND queries we download the corresponding document (*DownloadDocument*) and we extract the outlinks (*ExtractOutlinks*). In lines 11 to 17, for each actor pair the corresponding AND query is built (*BuildANDQuery*) and the number of hits for the specific actor pair is retrieved (*GetPageCounts*) as well as the M top-ranked snippets (*GetDocumentSnippets*). In lines 18 to 26, for each actor pair the page-count-based (*PageCountSimilarity*), the text-based for IND and AND snippets (*TextSimilarity*) and the link-based (*LinkSimilarity*) relatedness scores are computed as well as their combination (*CombineSimilarities*).

Step 3: Visualization of extracted networks. In this step, the extracted networks are displayed as graphs. The nodes correspond to the political actors and the edges to the relations among them. The graphs are created by giving the produced matrices with the relatedness scores from Step 2, to the NEATO program [3], a free tool for undirected graphs that implements the general purpose Kamada-Kawai algorithm [18]. According to Kamada-Kawai algorithm, each edge is represented as a spring model that can be compressed or stretched. Assuming we have a graph G consisting of a set of vertices V. The energy or a spring is the squared difference of its natural and actual lengths

$$Dist(x_u, x_v|\lambda_{uv}) = \frac{c}{\lambda_{uv}^2}(d(u,v) - \lambda_{uv})^2, \tag{3.9}$$

where λ_{uv} and $d(u,v)$ the desired and actual distance respectively of vertices u and v. The whole graph is a dynamical system consisting of the set of individual spring models whose energy is computed by Eq 3.9. The desirable graph is the one with the minimum total spring model energy. Thus, to produce the graph, the algorithm iteratively minimizes a functional of the differences between the desirable distances of nodes and the actual ones

$$U_{KK} = \sum_{u,v \in V} Dist(x_u, x_v|d_G(u,v)), \tag{3.10}$$

where $d_G(u,v)$ is the length of the shortest path from vertices u to v in graph G.

NEATO is a convenient visualization tool that can be easily programmed to represent the relation strength. In this work the relation strength is indicated by the line thickness

[3] http://www.graphviz.org

Algorithm 1 Algorithm for the computation of relatedness scores between actors.

Require: A // Set of actors
Require: $R(a_i)$ // Lexicalized form of actor a_i
Require: M // Number of web documents/snippets
// Step 1: Retrieve web data
for each actor $a_i \in A$ **do**
  $Q_{I,i} \leftarrow$ BuildINDQuery($R(a_i)$) // IND queries
  $C_{I,i} \leftarrow$ GetPageCounts($Q_{I,i}$)
  $U_{I,i} \leftarrow$ GetDocumentURLs($Q_{I,i}$, M)
  $N_{I,i} \leftarrow$ GetDocumentSnippets($Q_{I,i}$, M)
end for
for each URL $u_{i,k} \in U_{I,i}$, $k = 1, \ldots, M$ **do**
  $D_{I,i} \leftarrow$ DownloadDocument($u_{i,k}$)
  $O_{I,i} \leftarrow$ ExtractOutlinks($D_{I,i,k}$)
end for
for each actor $a_i \in A$ **do**
  for each actor $a_j \in A$ **do**
    $Q_{A,i,j} \leftarrow$ BuildANDQuery($R(a_i)$, $R(a_j)$) // AND queries
    $C_{A,i,j} \leftarrow$ GetPageCounts($Q_{A,i,j}$)
    $N_{A,i,j} \leftarrow$ GetDocumentSnippets($Q_{A,i,j}$, M)
  end for
end for
// Step 2: Compute similarities
for each actor $a_i \in A$ **do**
  for each actor $a_j \in A$ **do**
    $S^P \leftarrow$ PageCountSimilarity($C_{I,i}$, $C_{I,j}$, $C_{A,i,j}$)
    $S^T \leftarrow$ TextSimilarity($N_{I,i}$, $N_{I,j}$), or
    $S^T \leftarrow$ TextSimilarity($N_{A,i,j}$)
    $S^L \leftarrow$ LinkSimilarity($O_{I,i}$, $O_{I,j}$)
    $S \leftarrow$ CombineSimilarities(S^P, S^T, S^L)
  end for
end for
return S^P, S^T, S^L, S

of the edges in the produced graphs. Strong relations correspond to greater line thickness while weak relations to smaller. We have five different levels of line thickness. Each level corresponds to value range in [1,3] and is defined between two thresholds. We define four thresholds in (1,3). The values of the four thresholds are computed so as ratings are equally distributed in the five levels. In Figure 3.3, the five levels of line thickness are presented with their corresponding thresholds.

3.3 Capturing the evolution of networks through time

So far our method extracts the policy network without taking account the influence of time parameter. In reality, policy networks evolve over time and the visualization

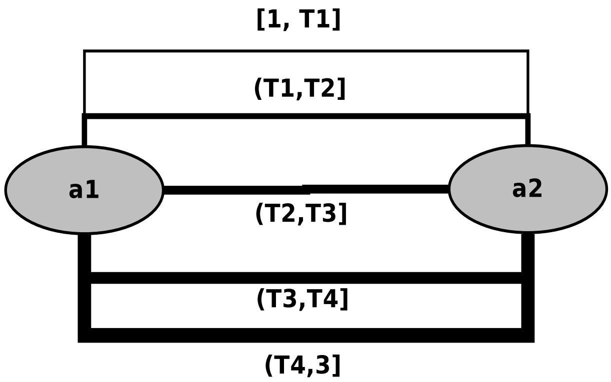


FIGURE 3.3: Visual notation of the five levels of line thickness and the 4 corresponding thresholds. From weak to strong relations.

or animation of their evolution can help political scientists to gain insights into useful conclusion and theories [68, 69]. In this work we visualize the evolution of the policy networks based on our approach for policy network extraction from the previous section. The goal of the produced animation is to show how the relation strengths and degree of centralities (which is a measure of activeness and importance, see Eq. 4.5) evolute over time. Finally, a video is created to better visualize the evolution of the policy networks.

More specifically, to complete this task, a time period for examination is selected. For each one of the years in the selected time period all three types of web data (page counts, snippets and links) are retrieved. To obtain the necessary web information, we add the year of interest at the end of each AND and IND query e.g., "a_i AND a_j +1995", creating in this way a set of AND and IND queries for each year. The process of the data retrieval remains the same with that in Section 3.2. Following the procedure described in Section 3.2, the relatedness scores for each year are estimated and a network is extracted. Thus for each year we have a set of computed relatedness scores and a set of degree of centralities.

To eliminate any noise in the computed relatedness scores and degree of centralities and achieve a smooth changes, we apply a moving average according to which the relatedness score of a relation between two actors on a specific year t is the average of $\frac{w}{2}-1$ previous and $\frac{w}{2}-1$ next years. The same moving average is used to smooth the degree centralities. The smoothed relatedness score $r^s_{i,j}(t)$ for a relation between two actors a_i and a_j and a year t is computed by taking the central moving average in a window of w years

$$r^s_{i,j}(t) = \frac{\sum_{t-(\frac{w-1}{2})}^{t+(\frac{w-1}{2})} r^e_{i,j}(t)}{w}, \tag{3.11}$$

where $r^e_{i,j}(t)$ is the estimated (unsmoothed) relatedness score in [1,3] for the relation between a_i and a_j and year t. In addition, the smoothed degree of centrality $c^s_i(t)$ for an actor a_i is computed

$$c_i^s(t) = \frac{\sum_{t-(\frac{w-1}{2})}^{t+(\frac{w-1}{2})} c_i^e(t)}{w}, \tag{3.12}$$

where $c_i^e(t)$ denotes the computed (unsmoothed) degree centrality for actor a_i and year t. For each actor a_i we have a vector of degree centralities of T elements (where T is the number of years in the selected time period), each degree centrality is then mapped in [1,4] using the Eq. 4.1. In the visualization of the evolution we aim at visualize the change in activity of the nodes (actors) over the years of the selected time period. Thus the normalized value of the degree centrality of the specific actor and year is inserted as height of the corresponding node so as the increase (or decrease) of actor activity be represented by an increase (or decrease) of the node's height in the graph.

Finally, the different time snapshots of the network are rendered and a video that shows its evolution is produced. For each year of the selected time period a graph is created then the whole set of graphs are rendered and the movie is produced. During the process of rendering, for each year t we reproduce 50 frames. In the produced animation when passing from year t to $t+1$ we apply the fade-in/fade-out technique on the 50 frames of the graph for year t. After the fade-in/fade-out a visual effect follows which aims at showing the nodes that will change in activity (increase or decrease the centrality) in year $t+1$. (i.e., blinking the nodes that change their degree of centrality) and then the animated network comes to a steady state (for year $t+1$). The above scenario is represented schematically in Figure 3.5. Finally, in Figure 3.4, the flow diagram of the whole procedure for the creation of the animation of the network evolution is presented.

3.4 Summary

In this chapter, we present our approach towards the estimation of the relation strengths and the visualization of policy networks. We first formally define the relatedness metrics used to estimate the strength of relation in a policy network. The metrics are all web-based and are of three types: i) page-count-based ii) text-based and iii) link-based. Page-count-based metrics use the number of occurrences/co-occurrences of the actors of interest in the web, the text-based metric exploits the lexical context that the actors share in web snippets and the link-based metrics depend on the common outlinks extracted from the web documents that contain the actors. We also propose the linear combination of the metrics with two different weighting schemes, equal weights and inverse variances. Then we present the basic steps of our approach; the lexicalization of actor names, the query formation. the data acquisition, the relatedness computation

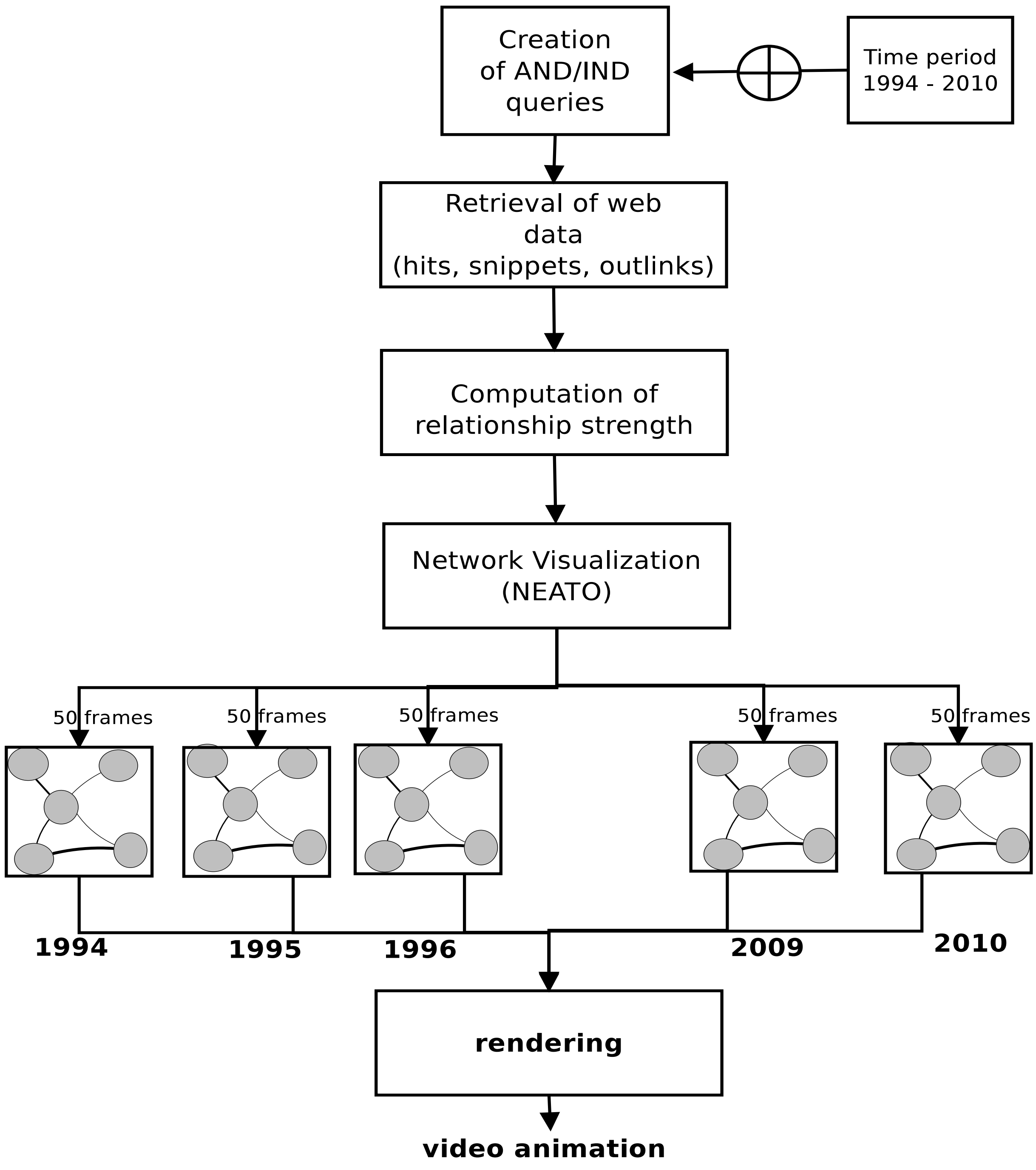


FIGURE 3.4: Diagram of the creation of animation of the network evolution.

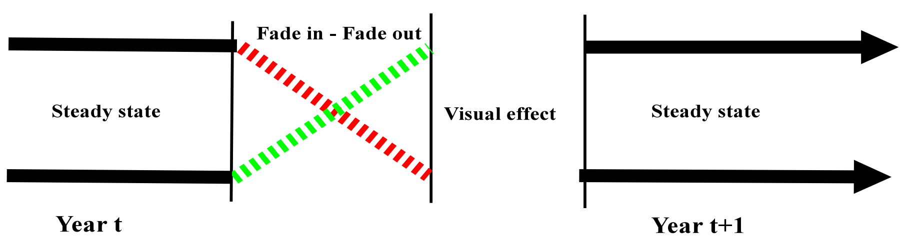


FIGURE 3.5: Animation of change from year t to year $t + 1$.

and finally the network visualization. We describe each one in full detail. We also describe how our approach is applied for the extraction of networks for different years in a time period of interest and how we animate the evolution of the extracted networks.

Chapter 4

Methodology

In the previous chapter, our approach for the computation of relation strength and the visualization of policy networks is presented. In this chapter, we first describe the policy networks under examination and we then proceed to the experimental setup. Finally, we define the metrics we use to evaluate the performance of our method.

4.1 Policy Network Corpus

Two policy networks from the political science literature were used to evaluate our approach. Both networks examine the patterns of adaptation and institutional policy learning during the third Community Support Framework (3rd CSF, 2000-2006), in two EU country members, namely, Ireland and Greece. Both case studies took place during the time period 2001-2003. The networks were extracted through a time and effort consuming manual process based on interviews and questionnaires collected during the Fifth Framework Project ADAPT (EU Enlargement and Multi-level Governance in European Regional and Environmental Policies). The same (translated) questionnaires were used for the analysis of the transformation of regional development policy-making procedures and institutional setting in Ireland and Greece.

The first network is based on the research conducted by Rees et al. [70], and includes the main governmental and non-governmental political actors involved in regional policy-making in Ireland and specifically in the Mid-West Region. The network consists of 37 public and private actors representing institutions at the local, regional and national levels. Relations among institutions are undirected; thus the network is represented by a 37×37 symmetric matrix (it has the form of Figure 3.1). Each matrix element denotes the strength of the relation between the corresponding actors. Not all possible relations

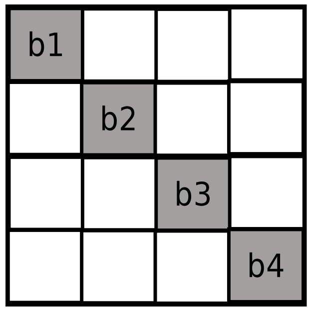


FIGURE 4.1: Ireland case study: The adjacency matrix given from political scientists and the blocks of positive and negative relations.

were investigated by the political scientists[1]. Each examined relation is rated with a score of '1','2' or '3' corresponding to a weak, medium or strong relationship. According to [70], actors in the network matrix were clustered in four blocks of structural equivalence (the diagonal blocks in Figure 4.1), so that relations that indicate friendship or cooperation be located in the same block, while relations that express antagonism be located in different blocks. We followed the same separation in our work. The relations between the actors of the same block are denoted as "positive", while the relations between actors from different blocks are denoted as "negative". The relationship strength ranges from '1' to '3' (weak to strong) for both positive and negative relations. In our work, we present separately results for positive and negative relations. In Figure 4.1, the blocks of positive and negative relations are shown. Positive relations are located in the diagonal blocks colored in gray while negative in the non-diagonal white blocks.

The second network is based on the study by Getimis and Demetropoulou [71] that focuses on the South Aegean region in Greece. The objectives of this research are very similar to those of the Irish case[2]. The Aegean network consists of 21 political private and public actors from the local, regional and national levels. As in the Irish case, relations are assumed symmetric and the network is represented by a 21×21 symmetric matrix. Each element denotes the strength of relation between the corresponding actors using the same "1" to "3" (weak to strong) scale. Unlike [70], there is no such separation between the actors. It has to be mentioned that although the questionnaires were identical in both case studies, the histograms of the ratings given from political scientists are different. In the Ireland case study, the relations have been rated uniformly while in the Aegean case the majority of relations were rated as weak ('1') or medium ('2'). In Figure 4.2, the histograms of ratings for the two policy networks Aegean and Ireland are presented.

[1] All possible relations between N actors is $\frac{N(N-1)}{2}$ but only a subset is investigated. This is the common practice in the political science literature. Only those actor pairs that are judged by the experts to be related are examined formally.

[2] Questionnaires in both case studies are the same but translated in English and Greek.

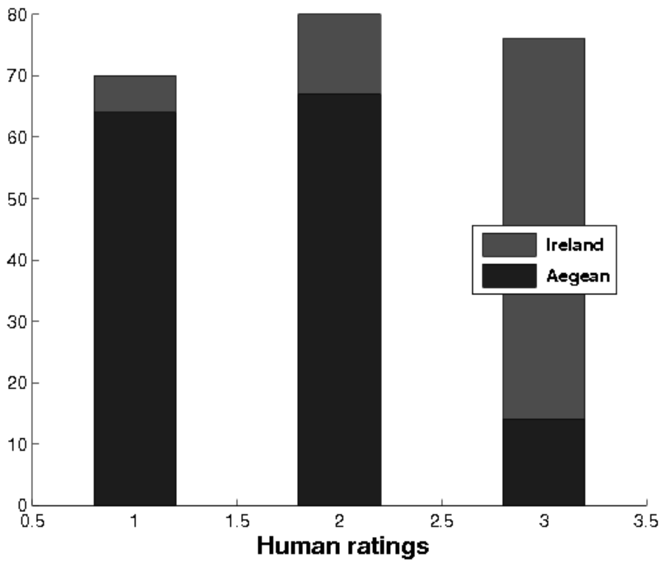


FIGURE 4.2: Histograms of ratings for Ireland and Aegean.

4.2 Experimental Setup

For the Ireland case study, the network contains 37 political actors and 226 rated actor pairs. In this work, we have focused on a subset of the network containing only 24 actors for which all relatedness metrics can be effectively computed, i.e., each actor generates an adequate number of web hits. We exclude the 13 actors for which the search engine returns less than 500 URLs of HTML documents. In the corresponding 24×24 submatrix there are 85 rated relations corresponding to 19 positive relations (denoted henceforth as “pos”) and 66 negative relations (denoted as “neg”).

Similarly for the Aegean case study, there are 21 political actors and 145 rated relations in the policy network. Using the same criterion as above, 3 actors were excluded; for the remaining 18 actors [3] we examine the 109 rated relations (all “pos”). The same policy network extraction algorithm is applied to both networks.

As discussed in Chapter 3, actors might appear with different names or abbreviations (e.g., acronyms) in web documents. The most common actor names were given by political scientists and then manually refined using web queries. For each of the common actor names, an individual web query is posed and the returned documents are inspected for alternative wordforms of the actor. This is an iterative procedure where each alternative name is tested and the relevance of the returned documents is evaluated. The goal is to select a list of names, abbreviations and acronyms for each actor that is not overly ambiguous. At the end, each actor name is represented as a regular expression with the list

[3] The names and acronyms of the selected actors for both policy networks are included in the Appendix

of alternative names connected via OR conjunctions. Despite our best efforts to select a list of unambiguous actor names, there are issues related to organizations in different countries that share the same name or acronym. This is especially true for the case of Ireland where confusion arises with similar names in the US or UK. In order to reduce ambiguity we include the pragmatic constraint "Ireland" via an AND conjunction in the actor's regular expression.

For the case of Aegean, the initial actor names are translated to Greek. The different names of actors are manually extracted according to the procedure described above. Even though the use of Greek terms tackles the problem of ambiguity, it strengthens data sparsity. Considering this fact, we also take account of the different cases of actors in the individual queries. Finally, for both policy networks, the AND queries are created by connecting the IND queries of the related actors using AND conjunction.

For the computation of the hit-based metrics, we use the returned hit counts from AND and IND type queries using the regular expressions for each actor, as presented above. Similarly for text-based metrics, we use the snippets returned by these AND and IND type queries. Specifically, we requested from the search engine to retrieve the 500 top-ranked URLs for each IND and AND query A snippet (characteristic portions of the document as selected by the search engine containing the actor name) was downloaded for each URL (from IND or AND queries). In our experiments, we report results using the top 100, 200 or 500 top-ranked snippets for each AND or IND query. A window $W = 10$, i.e., ten words to the left and ten words to the right of the actor, is used. Stop words[4] are excluded from the list of contextual features. For the computation of the link-based metrics the base outlinks extracted from the downloaded HTML documents from IND queries are used. Table 4.1 summarizes the statistics for different domains according to the downloaded URLs for both case studies. Looking at Table 4.1, we observe that for the Ireland case study the downloaded snippets and IND queries only the 32% of the downloaded documents and snippets come from URLs in *.ie* domain, 42% from *.com* and a 15% from other domains. In case of AND queries, the percentage of *.ie* URLs has increased and the percentage of *.com* and URLs at other domains has decreased. For Aegean case study, we observe that the majority of downloaded documents and snippets come from *.gr* and *blog* domains, while for AND queries the percentage of blogs has increased.

In the experiments that follow, we evaluate the performance of our metrics by keeping all the relations ('1','2','3') denoted as '3-levels' or by keeping only the weak and strong relations ('1','3') denoted as 'high-low'. Table 4.2 summarizes the number of relations in the evaluation process for both policy networks.

[4]For a stop word list for the Greek language see [72].

Dataset	Query type	Domains (%)					
		ie/gr	*org*	*gov*	*blog*	*com*	others
IRELAND	IND	**32**	7.4	2.5	1.1	**42**	**15**
	AND	**43**	5	4.4	2.3	**34**	**10**
AEGEAN	IND	**72**	3	0.6	**11**	9	4.2
	AND	**58**	3.4	0.4	**26**	7.7	2.9

Table 4.1: Statistics at domains for both case studies.

Experiment	Number of Relations	
	IRELAND (pos/neg)	AEGEAN
3-levels	19/66	109
high-low	14/40	62

Table 4.2: Relations examined on both policy networks for '3-levels' and 'high-low' experiment.

To examine the evolution of both policy networks over time, we selected the 1994-2010 time period which can be divided into three subperiods:

1. **1994-1999**: the years before the exploitation of the 3rd CSF funding for both countries Ireland and Greece.
2. **2000-2006**: the years of the 3rd CSF funding. Also this period includes the time political research took place (2001-2003).
3. **2007-2010**: the time period after the 3rd CSF funding.

For each date from the selected time period we gather the required data for both policy networks using the conjunctive (AND) and individual (IND) queries for the selected actors and relations and adding the specific year at the end of each AND or IND query i.e., a_i AND a_j +1995, a_i +1995. For both case studies, the hits are used for the computation of the page-count relatedness scores, the top-ranked 100, 200, 500 IND and AND snippets are used for the computation text-based relatedness scores and the extracted outlinks from the web documents (from IND queries) for the link-based scores.

4.3 Flight traffic and Co-citation networks

Apart from the Ireland and Aegean networks which are manually created by political scientists, we tested our method on two other toy networks: i) a flight traffic network of airports ii) a co-citation network of researchers. The creation of the ground-truth networks as well as the procedure of web data gathering is presented in the next two sections.

4.3.1 Flight traffic network

From a publicly available list of the busiest US airports, we selected 20 airports. The ground-truth network of the 20 airports was manually created using data from the US Bureau of Transportation and Statistics [5]. The weights on the edges of the network are the number of flights between the participating airports, during the year 2010. From the 190 pairs of airports (relations), the pairs where the number of flights is 0 were excluded. The remaining airport pairs are 173. For the computation of the hit-based metrics we used the hit-counts from AND and IND type queries. For the text-based metrics we requested from the search engine to retrieve the 1000 top-ranked URLs for each AND and IND query and we downloaded the snippets of the corresponding URLs. We report experiments for 100, 200 and 500 (top-ranked) snippets for AND, IND queries. For the link-based metrics, we downloaded the HTML documents using the retrieved URLs from IND queries to extract the outlinks. We evaluated the metrics conducting two sets of experiments: i) using the whole set of the 173 relations (in tables is denoted as 'All'). ii) using only the pairs of pairs of airports that share less than 200 flights and more than 20000 flights (in tables is denoted as 'Weak-Strong'). We did this to investigate the performance of our method for the case of the whole range of relations and for the case of the weak and strong relations only. Table 4.3 summarizes the relations in each case.

Flight traffic network	
Relation set	Num. relations
All	173
Weak-Strong	39 (27 weak, 12 strong)

Table 4.3: Number of relations for the flight traffic network.

4.3.2 Co-citation network

As a second toy network, we manually created a co-citation network of researchers. More specifically, we selected 10 researchers. The full names of the researchers were lexicalized following the procedure described in the paper. The individual query for each actor was created by connecting his wordforms (including his full name) using OR conjunctions. The AND queries were created by connecting the corresponding individual queries with AND conjunctions.

The ground-truth network was manually created by weighting the edges with the number of the papers where the participating researchers are co-cited. To do this, we used the database of our references. The pairs of researchers that are not co-cited at all

[5] www.transtats.bts.gov

were excluded. More specifically, from the 45 pairs of researchers (relations) 34 were considered. For the computation of the hit-based metrics we used the hit-counts from AND and IND type queries. For the text-based metrics we requested from the search engine to retrieve the 1000 top-ranked URLs for each AND and IND query and we downloaded the snippets of the corresponding URLs. We report experiments for 100, 200 and 500 (top-ranked) snippets for AND, IND queries. For the link-based metrics, we downloaded the HTML documents using the retrieved URLs from IND queries to extract the outlinks.

4.4 Evaluation Metrics

Let $H = (h_1, h_2, ..., h_M)$ and $K = (k_1, k_2, ..., k_M)$ be the vectors of human rated and automatically computed relatedness scores, respectively, where M is the total number of relations. Scores k_i may be computed by any of the relatedness metrics presented in Chapter 3 or their fusion. In order to match the range of human ratings all relatedness scores are linearly scaled as follows:

$$e_i = \frac{2(k_i - k_{min})}{k_{max} - k_{min}} + 1, \tag{4.1}$$

where k_{min}, k_{max} is the min and max scores (for a specific metric), respectively, and e_i is the normalized relatedness score that takes continuous values in [1,3].

To measure the correlation between the human ratings and normalized relatedness scores we use the Pearson Correlation Coefficient defined as:

$$r_{H,E} = \frac{\sum_{i=1}^{M}(h_i - \bar{H})(e_i - \bar{E})}{\sqrt{\sum_{i=1}^{M}(h_i - \bar{H})^2 \sum_{i=1}^{M}(e_i - \bar{E})^2}}, \tag{4.2}$$

where $\bar{H}$ and $\bar{E}$ denote the sample mean of H and E respectively, and $E = (e_1, e_2, ..., e_M)$ is the vector of values produced by (4.1). In addition to correlation, the Mean Square Error (MSE) is also used to measure the distance between human ratings and normalized relatedness. The MSE is averaged over all investigated relations, as follows:

$$MSE = \frac{1}{M}\sum_{i=1}^{M}(h_i - e_i)^2. \tag{4.3}$$

Note that the MSE ranges between 0 and 4. Smaller MSE values denote better agreement between the manual and automatically computed ratings.

Considering the high-low experiment, we define the precision/recall measures used. The precision is defined as the fraction of the correctly classified strong relations and the recall is the fraction of the relations classified as strong. Precision (P) and recall (R) are formally defined:

$$P = \frac{|C|}{|S|}, \quad R = \frac{|C|}{|F|}, \tag{4.4}$$

Where C is the set of the correctly classified strong relations, F denotes the set of the relations rated as strong and S is the set of relations classified as strong. In the experimental results in Chapter 5 we compute the F-measure using the known definition $F - meas = \frac{2PR}{P+R}$.

A widely used measure in social network analysis is the degree of centrality that indicates the importance of an actor in a network [73]. In policy networks graphs, vertices represent actors and edges represent the relations between actors. The degree of centrality for each actor a_i is defined:

$$DC_{a_i} = \frac{1}{(A-1)} \sum_j w_{i,j}, \tag{4.5}$$

where A is the number of actors (vertices), and $w_{i,j}$ is the weight (rating) of the relation (edge) between actors a_i, a_j. The degree of centrality is computed for both the original and extracted networks. The two degree of centrality vectors (extracted vs. original) are compared in terms of correlation and MSE using (4.2) and (4.3), respectively.

Chapter 5

Experimental Results

In this chapter, we present evaluation results for page-count, text-based and link-based relatedness metrics, as well as, their fusion on the Ireland and Aegean corpora. This chapter is organized in two parts. In the first part (Sections 5.1 and 5.2), we present the baseline scores and discuss the results of the evaluation of our approach using the evaluation metrics from Chapter 3 for both relationship strength and extracted degree of centrality. We also measure the correlation and MSE (on both relationship strength and degree of centrality) between the extracted and original networks for a selected time period. The results of the proposed method for the flight traffic and co-citation network are also presented and discussed. In the second part (Section 5.3), the differences between the manually created and extracted networks are visualized via network graphs and discussed. Finally, indicative screenshots of the extracted networks for years from the selected time period (1994-2010) are presented and discussed.

5.1 Baseline

We consider two methods for the definition of the baseline results.

- **Baseline 1:** We created 10000 random vectors of '1','2','3' and we computed the correlation and MSE between each random vector and the human rated and we took the mean of the correlation and MSE scores. The random vectors of '1','2','3' were created in two ways: i) assuming that ratings have equal apriori and ii) using their actual apriori. The baseline results are shown in the next Table 5.1 for both Ireland and Aegean and '3-levels' (all pairs included) and 'high-low' (only '1', '3' included) experiments.

- **Baseline 2:** We rated all the relations with '1','2' or '3' and measured the MSE with the human rated in each case. In this case the correlation is undefined and cannot be measured according to Eq 4.2. The baseline results are shown in Table 5.2 for both Ireland and Aegean networks.

Baseline 1					
Experiment	Apriori	IRELAND (pos/neg)		AEGEAN	
		Correlation	MSE	Corr.	MSE
3-levels	equal	**0.0013/0.0015**	1.45/1.30	-0.0006	1.25
	actual	0.0005/0.0015	**1.11/1.20**	-0.0009	**0.89**
high-low	equal	-0.0004/0.0007	1.99/1.99	0	2.00
	actual	**0.0023/0.0007**	**1.35/1.97**	-0.0002	**1.24**

Table 5.1: Correlation and MSE scores for Ireland and Aegean network, Baseline 1.

Baseline 2 (only MSE)			
Experiment	Rating	IRELAND (pos/neg)	AEGEAN
3-levels	1	2.58/1.48	0.87
	2	**0.74/0.61**	**0.57**
	3	0.89/1.73	2.27
high-low	1	3.14/**1.80**	**0.77**
	3	**0.86**/2.20	3.23

Table 5.2: MSE scores for both Ireland and Aegean networks, Baseline 2.

5.2 Evaluation in terms of correlation and MSE

5.2.1 Page-count-based metrics

5.2.1.1 Results on relationship strength

The performance of the four page-count-based metrics (Jaccard S_J^P, Dice S_D^P, mutual information S_I^P, google-based relatedness S_G^P) is shown in Table 5.3 in terms of correlation and average MSE for the Ireland and Aegean policy networks. Results are shown separately for '3-levels' (all pairs included) and 'high-low' (only pairs with scores 1 or 3 included). For the case of Ireland where also negative (antagonistic) relations exist in the network, results are shown separately for positive and negative relations.

For Ireland and the positively related actor pairs, the google (S_G^P) and mutual information (S_I^P) metrics outperform the Jaccard (S_J^P) and Dice (S_D^P) metrics both in terms of correlation and (especially) MSE. The highest correlation of 0.61 is achieved by S_I^P (the corresponding score for the 'high-low' experiment is 0.66). For the negatively related actor pairs, the results are relatively poor, all correlations are below 0.30. The S_J^P and

Page-count-based metrics					
Relationship strength					
Experiment	Metric	IRELAND (pos/neg)		AEGEAN	
		Correlation	MSE	Corr.	MSE
3-levels	Jaccard (S_J^P)	0.29/0.28	1.77/1.14	0.35	0.53
	Dice (S_D^P)	0.29/**0.29**	1.75/1.12	**0.37**	**0.51**
	Mutual Info (S_I^P)	**0.61**/0.09	**0.42**/0.77	0.24	1.14
	Google (S_G^P)	0.49/0.17	0.69/**0.70**	0.35	0.91
high-low	Jaccard (S_J^P)	0.30/0.34	2.20/1.38	0.41	0.55
	Dice (S_D^P)	0.30/**0.35**	2.18/1.36	0.43	**0.53**
	Mutual Info (S_I^P)	**0.66**/0.10	**0.54**/1.19	0.44	0.81
	Google (S_G^P)	0.56/0.19	0.92/**1.08**	**0.52**	0.61

TABLE 5.3: Correlation and MSE on relationship strength for the page-count-based metrics.

S_D^P metrics achieve somewhat higher correlations here, although their MSE is higher than S_G^P and S_I^P. As expected, higher correlation scores are achieved for the 'high-low' experiment rather than the '3-levels' experiment, however, the MSE is usually higher for the 'high-low' experiment. Overall, good correlation is achieved for positive relations using page-count metrics (especially for S_I^P, S_G^P), however, page-count metrics perform poorly for negative relations, in fact the MSE scores in that case are worse than the baseline (see Table 5.2).

For the case of Aegean, S_J^P, S_D^P and S_G^P achieve similar performance in terms of correlation and for the '3-levels' experiment, while in terms of MSE S_J^P and S_D^P outperform the S_G^P and S_I^P. For the 'high-low' experiment better correlation scores are achieved (compared to the '3-levels' experiment) and the MSE is lower (S_I^P, S_G^P) or stays at about the same levels (S_J^P, S_D^P). Overall, correlation results are lower than those achieved for the positive relations in the Ireland network and reach the maximum value of 0.52 for the 'high-low' experiment using the S_G^P metric. In terms of MSE similar conclusions can be reached; for the Aegean case study and the '3-levels' experiment, the S_D^P metric achieves the minimum MSE at about 0.51 (which is slightly better than the baseline in Table 5.2) compared to 0.42 for the S_I^P in the Ireland case study.

Another experiment considering the performance of page-count-based metrics is the investigation of the affect of the domain confinement. In this experiment we confined the queries for both case studies to different domains, *ie/gr*, *org*, *gov*. The evaluation results on relationship strength are presented in Table 5.4. We present the evaluation results for mutual information (S_I^P) and google (S_G^P) metrics for the case of Ireland and Aegean respectively.

From Table 5.4 we observe that the selected page-count-based metrics achieve better performance when a restriction in domain is applied. More specifically, for the case of

Page-count-based metrics					
Relationship strength					
Experiment	Domain	IRELAND (pos/neg)		AEGEAN	
		Correlation	MSE	Corr.	MSE
3-levels	any	0.61/0.09	0.42/0.77	0.35	0.91
	ie/gr	**0.66**/0.18	**0.42**/0.77	**0.48**	**0.46**
	org	0.42/**0.20**	0.92/**0.76**	0.41	0.94
	gov	0.64/0.08	0.53/0.85	0.27	0.96
high-low	any	0.66/0.10	**0.54**/1.19	0.43	**0.53**
	ie/gr	**0.68**/**0.29**	0.55/1.11	**0.56**	0.65
	org	0.47/0.23	1.14/1.14	0.41	1.63
	gov	0.65/0.14	0.67/**1.07**	0.35	1.85

TABLE 5.4: Performance of page-count-based metrics after the restriction at different domains.

Ireland when the queries are restricted in *ie* domain the correlation increases from 0.61 (no domain restriction) to 0.66 for the positive relations and the '3-levels' experiment. For the negative relations the correlation score increases from 0.09 to 0.18 but the performance still remains poor. For both positive and negative relations the MSE remains the same if the restriction in *ie* domain is applied.

For the Aegean case study the domain restriction in *gr* increases the correlation score from 0.35 (no domain restriction) to 0.48 for the '3-levels' experiment which is a significant increase of 0.13 . The same observation holds for the 'high-low' experiment. Restriction in *gr* enhances the performance of the metrics in terms of MSE where the scores are better than the baseline (see Table 5.2) Overall, *org* and *gov* do not seem to improve the performance of the page-count-based metrics, we believe that this happens due to fact that the actors of the networks do not occur oftenly in web documents of *org* and *gov* domains (see Table 4.1). In general, the domain restriction especially at *ie* and *gr* domains enhances the performance of the page-count metrics. Overall, it is shown that in both case studies Ireland and Aegean the domain restriction tackles somewhat the problem of ambiguity and enhances the performance of the metrics.

5.2.1.2 Results on degree of centrality

The performance of the page-count-based metrics on the degree of centrality is presented in Table 5.5 in terms of correlation and MSE. For the case of Ireland, mutual information (S_I^P) as well as google metric (S_G^P) achieves high correlation score (0.98) for positive relations for both '3-levels' and 'high-low' experiments. For negative relations, all four metrics have similar performance in terms of correlation and for both '3-levels' and 'high-low' experiments. In terms of MSE and for the positively related actor pairs, S_I^P

Page-count-based metrics					
Degree of centrality					
Experiment	Metric	IRELAND (pos/neg)		AEGEAN	
		Correlation	MSE$\times 10^{-2}$	Corr.	MSE$\times 10^{-2}$
3-levels	Jaccard (S_J^P)	0.94/0.98	6/5.6	0.85	7.4
	Dice (S_D^P)	0.94/**0.98**	5.9/5.4	0.85	**6.2**
	Mutual Info (S_I^P)	**0.98**/0.97	**0.6**/1.8	0.88	38
	Google (S_G^P)	0.98/0.98	1.5/**1.5**	**0.90**	28
high-low	Jaccard (S_J^P)	0.93/0.96	5.5/2.8	0.90	1.4
	Dice (S_D^P)	0.93/**0.96**	5.5/2.8	0.90	**1.2**
	Mutual Info (S_I^P)	**0.98**/0.95	**0.7**/1.5	0.89	6
	Google (S_G^P)	0.97/0.95	1.7/**1.3**	**0.92**	3.3

Table 5.5: Correlation and MSE on degree of centrality for page-count-based metrics.

performs better among all other metrics achieving the lowest MSE scores 0.6×10^{-2} and 0.7×10^{-2} for the '3-levels' and 'high-low' experiment respectively. Considering the negative relations, S_G^P achieves the lowest MSE scores 1.5×10^{-2} and 1.3×10^{-2} for the '3-levels' and 'high-low' experiment respectively.

For the case of Aegean and in terms of correlation, the S_G^P metric achieves the highest score among all other metrics 0.90 and 0.92 for the '3-levels' and 'high-low' experiment respectively. In terms of MSE, the S_D^P metric performs better than any other metric achieving 1.2×10^{-2} and 6.2×10^{-2} for the '3-levels' and 'high-low' experiment respectively.

5.2.2 Text-based metrics

In this section, we present the performance of the text-based metric using snippets downloaded from the web using conjunctive queries containing both actors (AND) or individual queries for each actor (IND). Various context window sizes (W) were evaluated experimentally and best results were achieved around window size $W = 10$, i.e., ten words to the left and ten words to the right of the actor of interest are the contextual features used. Results on relationship strength are reported for both case studies in Tables 5.6 and 5.7 for AND and IND queries respectively. For the degree of centrality the results are shown in Tables 5.8 and 5.9 for AND and IND queries respectively. The results presented are for $W = 10$, for different number of downloaded snippets (100, 200 or 500) and for the binary (B) and logarithm of term frequency (LTF) weighting schemes.

Text-based metric					
Relationship strength					
AND queries					
Experiment	Number of snippets	Correlation		MSE	
		Weighting schemes			
		B	LTF	B	LTF
IRELAND (pos/neg)					
3-levels	100	**0.29**/0.29	0.26/**0.31**	**0.94**/**0.65**	0.95/0.67
	200	**0.30**/0.28	0.26/**0.29**	**0.95**/**0.66**	0.96/0.69
	500	**0.29**/0.29	0.26/**0.30**	0.99/**0.68**	**0.97**/0.69
high-low	100	**0.36**/0.40	0.31/**0.41**	**1.09**/0.89	1.10/**0.87**
	200	**0.39**/0.39	0.33/**0.39**	**1.10**/0.91	1.10/**0.93**
	500	**0.42**/0.39	0.36/**0.40**	1.12/0.91	**1.09**/**0.89**
AEGEAN					
3-levels	100	**0.20**	0.19	**0.53**	0.57
	200	0.18	**0.19**	**0.54**	0.59
	500	0.17	**0.19**	**0.52**	0.58
high-low	100	**0.19**	0.16	1.05	**1.00**
	200	0.16	**0.17**	1.16	**1.14**
	500	0.14	**0.18**	**1.10**	1.12

Table 5.6: Correlation and MSE on relationship strength for the text-based metric as a function of the number of snippets for AND queries ($W = 10$).

Text-based metric					
Relationship strength					
IND queries					
Experiment	Number of snippets	Correlation		MSE	
		Weighting schemes			
		B	LTF	B	LTF
IRELAND (pos/neg)					
3-levels	100	0.06/**0.33**	**0.10**/0.30	1.06/**0.55**	**0.81**/0.59
	200	**0.33**/0.32	0.29/**0.34**	1.51/**0.57**	**1.39**/0.58
	500	0.09/0.34	**0.13**/**0.35**	1.17/0.70	**0.98**/**0.64**
high-low	100	-0.08/**0.43**	**-0.04**/0.40	1.34/**0.81**	**1.03**/0.86
	200	**0.29**/0.41	0.20/**0.43**	1.86/**0.85**	**1.72**/0.86
	500	**0.04**/0.44	0.04/**0.45**	1.47/0.99	**1.60**/**0.93**
AEGEAN					
3-levels	100	0.37	**0.40**	**0.41**	0.43
	200	0.36	**0.38**	0.46	**0.41**
	500	0.32	**0.36**	0.46	**0.42**
high-low	100	0.49	**0.56**	0.70	**0.52**
	200	0.44	**0.49**	0.73	**0.63**
	500	0.41	**0.48**	0.81	**0.70**

Table 5.7: Correlation and MSE on relationship strength for the text-based metric as a function of the number of snippets for IND queries ($W = 10$).

5.2.2.1 Results on relationship strength

For the Ireland case study and positive relations, text-based metrics perform relatively poorly especially for the '3-levels' experiment. Comparing the results in Tables 5.6 and 5.7, AND queries outperform IND queries consistently on positive relations, especially in terms of correlation (except for the case of the 200 top-ranked snippets where the opposite is true). Considering the evaluation results of AND snippets in Ireland case study (Table 5.6), the B scheme outperforms somewhat the LTF for both positive relations while the opposite holds for the negative but the differences in performance are small. This observation holds in terms of both correlation and MSE and for both '3-levels' and 'high-low' experiment. Furthermore, correlation on positive relations increases as more AND snippets are considered, especially for the 'high-low' experiment, although the improvement going from 100 to 500 snippets is modest (from 0.36 to 0.42 at best). More specifically, the highest correlation of 0.42 is achieved for the 'high-low' experiment when using AND queries, the binary weighting scheme and the 500 top-ranked snippets. For negative relations, similar but somewhat higher correlation scores are achieved, up to 0.45. Here the best results are achieved when using individual (IND) queries. Also there is little or no difference in the performance of B and LTF weighting schemes. Note that although the correlation scores for negative relations are low they are higher than those achieved using page-count metrics (see Table 5.3) or link-based metrics (see Table 5.10). Similar conclusions considering negative relations can be drawn for the MSE scores. Overall, for the case of Ireland only the top-ranked IND snippets achieve better MSE results than the baseline and this is observed only for the negative relations. For the positive relations the text-based metric achieves worse MSE scores than the baseline in all cases.

For the Aegean case study, slightly higher correlation scores are achieved, up to 0.56 for the 'high-low' experiment. The best results, in terms of correlation and MSE, are obtained for the individual (IND) queries, in fact, performance for conjunctive queries (AND) here is very poor (correlation below 0.2 and MSE greater 0.50 are achieved throughout). Also the performance of both weighting schemes does not improve when a larger number of snippets is used, best correlation and MSE results are obtained (mostly) for 100 or 200 snippets (a sign of data sparseness). The LTF scheme outperforms the B scheme especially for the 'high-low' experiment, although the differences are small. Overall, moderate correlation scores are achieved using text-based metrics for the South Aegean case study, at the same level or better than those achieved using page-count metrics (see Table 5.3). In general, the text-based metric is shown to perform better on positive relations for the Aegean case study than for Ireland. In fact for Aegean the text-based metric on IND snippets (and especially for the LTF scheme) achieves better MSE scores than the baseline.

5.2.2.2 Results on degree of centrality

Text-based metric					
Degree of centrality					
AND queries					
Experiment	Number of snippets	Correlation		$MSE \times 10^{-2}$	
		Weighting schemes			
		B	LTF	B	LTF
IRELAND (pos/neg)					
3-levels	100	**0.89/0.98**	0.88/0.98	2.7/**0.66**	**2.7**/0.89
	200	**0.89/0.98**	0.89/0.98	2.7/**0.66**	**2.6**/0.93
	500	**0.88/0.98**	0.88/0.98	2.7/**0.76**	**2.6**/0.87
high-low	100	**0.90/0.97**	0.89/0.97	2.5/0.69	**2.4/0.65**
	200	**0.91/0.97**	0.90/0.97	2.5/**0.74**	**2.4**/0.88
	500	**0.91/0.97**	0.90/0.97	2.6/0.72	**2.5/0.67**
AEGEAN					
3-levels	100	**0.79**	0.78	**5.6**	6.1
	200	**0.79**	0.79	**5.7**	7.4
	500	**0.80**	0.80	**4.6**	7.0
high-low	100	**0.75**	0.70	9.7	**8.4**
	200	**0.75**	0.72	**1.1**	1.1
	500	**0.76**	0.73	**1.0**	1.0

Table 5.8: Correlation and MSE on degree of centrality for text-based metrics as a function of the number of snippets for AND queries ($W = 10$).

The performance of the text-based metric on the degree of centrality is presented in Tables 5.8 and 5.9 for AND and IND queries respectively. For the Ireland case study and positive relations, IND queries outperform AND queries in terms of correlation and for the '3-levels' experiment. However, for the 'high-low' experiment the opposite is true. Comparing the Tables 5.8 and 5.9 for Ireland case study and positive relations, we observe that the AND queries outperform IND queries consistently in terms of MSE, while in terms of correlation the performance is similar. Furthermore, for the negative relations the performance of AND and IND queries is similar in terms of correlation, while in terms of MSE the AND queries outperform IND queries consistently. Considering only the case of AND queries (see Table 5.8), the performance of B and LTF schemes is similar in terms of correlation and for both '3-levels' and 'high-low' experiments. Nevertheless, in terms of MSE, B outperforms LTF on positive relations, while the opposite is observed for the negative relations. For both weighting schemes and both types of relations (positive/negative), the performance remains the same for the different number of snippets in terms of correlation, while in terms of MSE there is a

Text-based metric					
Degree of centrality					
IND queries					
Experiment	Number of snippets	Correlation		MSE$\times 10^{-2}$	
		Weighting schemes			
		B	LTF	B	LTF
IRELAND (pos/neg)					
3-levels	100	**0.90**/**0.98**	0.89/0.98	3.0/**0.38**	**2.0**/0.42
	200	**0.94**/**0.98**	0.91/0.98	4.8/**0.4**	**4.7**/0.59
	500	**0.90**/**0.98**	0.90/0.98	3.4/1.6	**2.6**/**1.1**
high-low	100	**0.86**/**0.97**	0.86/0.97	2.8/**0.40**	**2.1**/0.51
	200	**0.92**/**0.97**	0.88/0.97	4.3/**0.46**	**4.1**/0.60
	500	**0.87**/**0.97**	0.86/0.97	**3.2**/1.2	3.4/**0.95**
AEGEAN					
3-levels	100	**0.90**	0.90	**1.8**	3.9
	200	0.87	**0.88**	3.9	**2.3**
	500	0.86	**0.88**	4.0	**2.5**
high-low	100	0.90	**0.91**	4.7	**2.5**
	200	0.88	**0.89**	5.2	**3.7**
	500	0.86	**0.88**	7.0	**5.5**

Table 5.9: Correlation and MSE on degree of centrality for text-based metrics as a function of the number of snippets for IND queries ($W = 10$).

slight perfomance degradation is observed. Considering now the case of IND queries (see Table 5.9), the performance of both weighting schemes is similar in terms of correlation. In terms of MSE, similar conclusions with those for AND queries (in Table 5.8) can be drawn. Finally, the performance of IND queries decreases as larger number of snippets is considered.

For the Aegean case study, IND queries clearly outperform AND queries in terms of correlation and both '3-levels' and 'high-low' experiments. The same holds in terms of MSE and the '3-levels' experiment, while for the 'high-low' experiment IND queries are shown to perform better. Considering only the AND queries, B scheme outperforms the LTF in terms of both correlation and MSE. Larger number of AND snippets has positive effect on the performance of the metric in terms of correlation and MSE and both '3-levels' and 'high-low' experiments. Considering now the case of IND queries, LTF scheme performs consistently better than B in terms of correlation and MSE and both '3-levels' and 'high-low' experiments. Contrary to AND queries, a larger number of IND snippets degrades the performance of the metric.

5.2.3 Link-based metrics

5.2.3.1 Results on relationship strength

Link-based metrics						
Ratings						
Experiment	Correlation			MSE		
	S_G^L	S_T^L		S_G^L	S_T^L	
		B	LTF		B	LTF
IRELAND (pos/neg)						
3-levels	**0.62**/0.01	0.34/**0.21**	0.36/0.18	**0.36**/0.83	0.79/0.71	0.79/**0.69**
high-low	**0.85**/0.001	0.61/**0.25**	0.59/0.22	**0.25**/1.24	0.79/1.04	0.84/**1.02**
AEGEAN						
3-levels	0.27	**0.36**	0.25	0.84	**0.41**	0.57
high-low	0.23	**0.46**	0.27	1.34	**0.80**	1.20

TABLE 5.10: Correlation and MSE on relationship strength for the link-based metrics.

The performance of link-based metrics using outlinks at *base* form is shown in Table 5.10 in terms of correlation and MSE for Ireland and Aegean. The following metrics are evaluated: google-based semantic relatedness using outlinks S_G^L and cosine similarity using outlinks S_T^L (with binary B and log term frequency LTF weighting schemes). For the case of Ireland and for positive relations, very good correlation performance is achieved especially for the 'high-low' experiment at 0.85. Cosine similarity achieves good performance for the 'high-low' experiment at 0.62 (less so for the '3-levels' experiment). There is no major performance difference between the B and LTF weighing schemes. For the negative relations, very poor results are achieved, throughout, with the binary cosine similarity metric achieving the best performance at 0.25. Overall, the outlinks perform the best out of all evaluated metrics for positive relations in the Ireland network, but fail to identify negative relations.

For the South Aegean network, results are not as impressive. Good performance is achieved only for the S_T^L metric (using B weighting), up to 0.46 for the 'high-low' experiment, while the google outlink metric performs poorly (unlike Ireland). Note that in terms of average MSE performance cosine similarity using outlinks S_T^L performs the best out of all metrics (page-count and text-based). Overall, outlinks produce good results for both case studies; for Ireland S_G^L performs the best, while for Aegean S_T^L provides the best results.

Another experiment considering the performance of link-based metrics is the investigation of the affect of the domain confinement. In this experiment we confine the queries (AND and IND) to different domains for both case studies, *ie/gr*, *org*, *gov*. The evaluation results on relationship strength are presented in Table 5.11. For both case studies

Link-based metrics					
Relationship strength					
Experiment	Domain	IRELAND (pos/neg)		AEGEAN	
		Correlation	MSE	Corr.	MSE
3-levels	any	**0.62**/0.21	**0.36**/0.71	0.36	0.41
	ie/gr	-0.48/0.22	1.19/**0.70**	0.33	0.43
	org	0.20/**0.24**	0.86/0.88	0.23	0.51
	gov	0.41/0.16	0.68/0.67	**0.41**	**0.41**
high-low	any	**0.85/0.25**	**0.25/1.04**	**0.46**	0.80
	ie/gr	-0.54/0.25	1.53/1.04	0.43	0.82
	org	0.50/0.28	0.79/1.19	0.26	0.84
	gov	0.72/0.19	0.50/1.10	0.45	**0.72**

Table 5.11: Performance of link-based metrics after the restriction at different domains. The evaluation results are for relationship strength.

we present the evaluation results for the link-based metrics that achieve the best performance in terms of correlation. More specifically, for Ireland we selected the S_G^L and the binary weighted S_T^L for positive and negative relations respectively. For the Aegean we selected the binary weighted cosine similarity S_T^L. The results are shown in Table 5.11.

For the Ireland case study and positive relations we observe that the domain restriction does not enhance the performance of the S_G^L metric. More specifically, the correlation score without the domain restriction remains the highest (0.62). It is interesting that restricting the outlinks at *ie* domain the estimated relationship strengths are negatively correlated with the human ratings. Furthermore, restriction at *org* and *gov* domains leads to low correlation scores 0.20 and 0.41 respectively (‘3-levels’ experiment). A somehow good correlation score 0.72 is achieved (in ‘high-low’ experiment) using the *gov* outlinks but is still lower than the baseline correlation score 0.85. The same conclusions are drawn in terms of MSE. For Ireland case study and negative relations improvement is not significant for both ‘3-levels’ and ‘high-low’ experiment.

For the case of South Aegean, the performance of the binary weighted S_T^L metric improves in terms of correlation from 0.36 to 0.41 for the ‘3-levels’ experiment using only the *gov* outlinks. We believe that this happens because that web pages that contain the actors that are related refer to common governmental web sites. In terms of MSE the selected metric achieves 0.41 and 0.72 for the ‘3-levels’ and the ‘high-low’ experiment respectively, which is lower than the corresponding baseline (no domain restriction).

5.2.3.2 Results on degree of centrality

Considering the evaluation results for the degree of centrality (see Table 5.12), for the Ireland case study and positive relations, the S_G^L has the best performance among all

Link-based metrics						
Degree of centrality						
Experiment	Correlation			MSE$\times 10^{-2}$		
	S_G^L	S_T^L		S_G^L	S_T^L	
		B	LTF		B	LTF
IRELAND (pos/neg)						
3-levels	**0.97**/0.97	0.94/**0.97**	0.94/0.97	**0.50**/1.9	1.4/1.2	1.5/**0.91**
high-low	**0.98**/0.93	0.97/**0.95**	0.96/0.94	**0.48**/1.3	1.5/0.88	1.6/**0.78**
AEGEAN						
3-levels	0.89	**0.89**	0.84	20	**2.1**	7.0
high-low	0.90	**0.92**	0.89	14	**6.8**	11

Table 5.12: Correlation and MSE on degree centrality for the link-based metrics.

other metrics in terms of correlation and MSE. This holds for both '3-levels' and 'high-low' experiments. For negative relations, on the other hand, the S_T^L with B weighting scheme achieves the highest correlation score among all other schemes for both '3-levels' and 'high-low' experiments. In terms of MSE and for negative relations, S_T^L with B scheme outperforms the S_G^L for the '3-leves' experiment, while for the 'high-low' experiment S_T^L is better using the LTF scheme. Finally, for the Aegean case study the S_T^L metric with B scheme outperforms S_G^L in terms of correlation and MSE for both '3-levels' and 'high-low' experiments.

5.2.4 Combination of metrics

5.2.4.1 Results on relationship strength

Next we investigate the performance for the linear combination of the three types of metrics, namely, page-count, text and link-based metrics. For each case study, we have selected the metric that performs best in terms of correlation. Specifically for the Ireland case study and for positive relations, we have selected mutual information S_I^P as the best performer among the page-count-based metrics, binary weighting using the 200 top-ranked snippets (AND queries) as the best text-based metric and the S_G^L as the best link-based metric. For negative relations, we have selected the Dice S_D^P page-count metric, LTF weighting using the 500 top-ranked snippets (IND queries) from the text-based metrics and the S_T^L with B scheme as link-based metric. Similarly for the Aegean case study, we have selected the google page-count metric S_G^P, the LTF weighted text-based metric using 100 snippets (IND queries), and the binary weighted cosine similarity S_T^L link-based metric, respectively. The results are presented in Table 5.13 for the two networks, using equal weights. First the performance of the individual metrics is shown (first three rows), then their two-way combinations are shown with equal weights or inverse of variances (next three lines) and finally the three way linear combination results are shown.

Fusion of metrics (equal weights)							
Relationship strength							
Experiment	Weights			IRELAND (pos/neg)		AEGEAN	
	λ_P	λ_T	λ_L	Correlation	MSE	Corr.	MSE
3-levels	1	0	0	0.61/0.29	0.42/1.12	0.35	0.91
	0	1	0	0.30/0.35	0.95/0.64	0.40	0.43
	0	0	1	0.62/0.21	0.36/0.71	0.36	0.41
	0	0.5	0.5	0.51/0.35	0.57/0.58	0.42	**0.38**
	0.5	0	0.5	**0.74**/0.27	**0.26**/0.84	0.39	0.52
	0.5	0.5	0	0.63/**0.42**	0.51/**0.57**	**0.45**	0.41
	0.33	0.33	0.33	0.68/0.37	0.36/0.68	0.44	0.42
high-low	1	0	0	0.66/0.35	0.54/1.36	0.52	0.61
	0	1	0	0.39/0.45	1.10/0.93	0.56	**0.52**
	0	0	1	0.85/0.25	**0.25**/1.04	0.46	0.80
	0	0.5	0.5	0.71/0.39	0.60/0.86	0.57	0.59
	0.5	0	0.5	**0.86**/0.31	0.26/1.13	0.53	0.70
	0.5	0.5	0	0.68/**0.46**	0.66/**0.83**	**0.62**	0.63
	0.33	0.33	0.33	0.81/0.40	0.43/0.95	0.60	0.62

TABLE 5.13: Correlation and MSE on relationship strength for individual metrics and their linear fusion.

For the Ireland case study and for positive relations, the two-way combination of the page-count and link-based metrics achieves the highest correlation on positive relations both for the '3-levels' and 'high-low' experiments at 0.74 and 0.86, respectively. The three-way combination with equal linear weights performs somewhat worse, which is expected due to the poor baseline performance of text-based metrics. For negative relations and for the '3-levels' experiment, the two-way combination of page-count-based and text-based metrics achieves the highest correlation at 0.42, followed closely by the three-way combination at 0.37. The results are very similar also for 'high-low' experiment with correlation up to 0.46. Overall, simple linear fusion outperforms the individual metrics, achieves very good performance for positive relations and acceptable performance for negative relations.

For the South Aegean case study and for equal weights, the two-way combination of the page-count and text-based metrics achieves the best performance in terms of correlation, while the three-way combination is a close second. All metric combinations achieve a consistent performance improvement in terms of correlation over the baseline, however, this is not always the case in terms of average MSE. Overall, the performance of the combined metrics is good and achieves correlation of up to 0.62 for the 'high-low' experiment. Unlike Ireland where the link-based metrics perform the best for positive relations, here the text-based metric is the best performer and combinations that contain it achieve the highest correlation scores.

We have also investigated linear metric fusion using inverse variance weighting. The results are presented in Tables 5.14 and 5.15 for Ireland and Aegean respectively. But for both case studies the results are similar to those of equal weights in Table 5.13. For the three-way combination the correlation results are 0.80/0.40 and 0.59 for the inverse variance weights for the positive/negative Ireland and Aegean case studies, respectively ('high-low' experiment). This is within 0.01 of the equal weighting scores.

Fusion of metrics (inverse variances)					
Relationship strength					
Experiment	Weights			IRELAND (pos/neg)	
	λ_P	λ_T	λ_L	Correlation	MSE
3-levels	1	0	0	0.61/0.29	0.42/1.12
	0	1	0	0.30/0.35	0.95/0.64
	0	0	1	0.62/0.21	0.36/0.71
	0/0	2.49/3.83	2.43/2.56	0.52/0.34	0.54/**0.59**
	3.43/4.26	0/0	2.43/2.56	**0.74**/0.28	**0.26**/0.89
	3.43/4.26	2.49/3.83	0/0	0.66/**0.41**	0.44/0.68
	3.43/4.26	2.49/3.83	2.43/2.56	0.71/0.36	0.53/0.74
high-low	1	0	0	0.66/0.35	0.54/1.36
	0	1	0	0.39/0.45	1.10/0.93
	0	0	1	0.85/0.25	**0.25**/1.04
	0/0	2.81/2.69	2.33/2.57	0.69/0.39	0.66/**0.86**
	2.71/3.25	0/0	2.33/2.57	**0.85**/0.32	0.27/1.17
	2.71/3.25	2.81/2,69	0/0	0.68/**0.45**	0.68/0.92
	2.71/3.25	2.81/2.69	2.33/2.57	0.80/0.40	0.46/1.00

TABLE 5.14: Correlation and MSE on relationship strength for individual metrics and their linear fusion, Ireland case study. Weights are given using the inverse variances of metrics.

Assigning equal weights to the metrics is a simple and naive approach to linearly combine the metrics. Thus using a brute force approach, we found the combination of weights that achieves the highest correlation and lowest MSE score for both case studies Ireland (positive and negative relations) and Aegean. More specifically, from all possible two-way and three-way combinations of weights that sum up to 1, we selected the best in terms of correlation and MSE for each case. The results are presented in Tables 5.16 and 5.17 for Ireland and Aegean respectively. From Tables 5.16 and 5.17 it is shown that selecting the weights following the brute force approach enhances the performance but not significantly. Only for the case of Ireland and positive relations the correlation goes from 0.85 to 0.88 and 0.26 to 0.21 the corresponding MSE score for the two-way combination ($\lambda_P = 0.3$, $\lambda_L = 0.7$) of page-count and link-based metrics.

Fusion of metrics (inverse variances)					
Relationship strength					
Experiment	Weights			AEGEAN	
	λ_P	λ_T	λ_L	Correlation	MSE
3-levels	1	0	0	0.35	0.91
	0	1	0	0.40	0.43
	0	0	1	0.36	0.41
	0	6.94	6.59	0.42	**0.38**
	9.20	0	6.59	0.39	0.57
	9.20		0	**0.45**	0.46
	9.20	6.94	6.59	0.44	0.45
high-low	1	0	0	0.52	0.61
	0	1	0	0.56	**0.52**
	0	0	1	0.46	0.80
	0	3.34	4.22	0.57	0.61
	4.56	0	4.22	0.53	0.69
	4.56	3.34	0	**0.61**	0.62
	4.56	3.34	4.22	0.59	0.65

Table 5.15: Correlation and MSE on relationship strength for individual metrics and their linear fusion, Aegean case study. Weights are given using the inverse variances of metrics.

IRELAND						
Experiment	Relation type	Weights			Correlation	MSE
		λ_P	λ_T	λ_L		
3-levels	pos	0.5	0	0.5	0.74	0.26
	neg	0.5	0.5	0	0.42	0.57
high-low	pos	0.3	0	0.7	**0.88**	**0.21**
	neg	0.3	0.7	0	0.47	0.78

Table 5.16: Ireland case study, combinations with highest correlation and lowest MSE scores.

AEGEAN					
Experiment	Weights			Correlation	MSE
	λ_P	λ_T	λ_L		
3-levels	0.3	0.5	0.2	0.45	0.37
high-low	0.3	0.5	0.2	0.61	0.58

Table 5.17: Aegean case study, combinations with highest correlation and lowest MSE scores.

5.2.4.2 Results on degree of centrality

Fusion of metrics (equal weights)							
Degree of centrality							
Experiment	Weights			IRELAND (pos/neg)		AEGEAN	
	λ_P	λ_T	λ_L	Correlation	MSE$\times 10^{-2}$	Corr.	MSE$\times 10^{-2}$
3-levels	1	0	0	0.97/0.97	0.6/5.4	0.90	28
	0	1	0	0.89/0.98	2.7/1.1	0.90	3.9
	0	0	1	0.96/0.97	0.5/1.3	0.89	2.1
	0	0.5	0.5	0.94/0.98	1.3/**0.6**	0.90	**1.8**
	0.5	0	0.5	**0.98**/0.97	**0.2**/2.8	0.91	8.8
	0.5	0.5	0	0.97/**0.98**	1.1/0.9	**0.92**	4.1
	0.33	0.33	0.33	0.98/0.98	0.6/1.7	0.91	4.2
high-low	1	0	0	0.97/0.95	0.7/2.7	0.92	3.3
	0	1	0	0.91/**0.97**	2.6/1.0	0.91	**2.5**
	0	0	1	0.98/0.95	0.5/0.9	0.92	6.9
	0	0.5	0.5	0.96/0.97	1.4/0.5	0.93	4.1
	0.5	0	0.5	**0.99**/0.95	**0.3**/1.5	0.92	5.2
	0.5	0.5	0	0.97/0.97	1.3/**0.5**	0.93	4.9
	0.33	0.33	0.33	0.98/0.96	0.9/0.9	**0.93**	4.7

Table 5.18: Correlation and MSE on degree of centrality for individual metrics and their linear fusion.

The degree of centrality results are shown in Table 5.18 for the two networks, using individual metrics and their linear fusion with equal weights (the evaluation results for the inverse variances are not shown are they are very similar). In general, the agreement between the original and extracted networks is very good both in terms of correlation and MSE. More specificaly, for the Ireland case study, correlation of up to 0.99 is achieved. The lowest correlation score of 0.89 holds for the text-based metric and for positive relations; all other individual and combined metrics score over 0.94. For negative relations, agreement between the original and extracted network centralities is excellent (over 0.97) for all metrics and their combinations. The results are also very good for the Aegean case study, achieving correlations between 0.90 and 0.93. There are no significant differences in performance between metrics or their combinations. Overall, all metrics (with the possible exception of the text-based metric for positive relations in Ireland) perform equally well for degree of centrality computation and provide very good to excellent correlation results.

5.2.4.3 Classification of weak and strong relations

Considering the 'high-low' experiment (ratings '1' and '3'), we tested our approach on the classification of weak and strongly rated relations. We consider two cases of the experiment: i) using the average of the extracted ratings as decision threshold and ii) training the threshold using the apriori of the two classes. In Table 5.19, the apriori of the two classes '1' and '3' are presented. In Table 5.20 the precision/recall and F-measure scores are presented for both cases (i) and (ii) of the experiment and both case studies.

Dataset	apriori of '3'	apriori of '1'
IRELAND (pos/neg)	0.78/0.45	0.22/0.55
AEGEAN	0.20	0.80

Table 5.19: Apriori probabilities of the strong ('3') and weak ('1') relations.

Dataset	Threshold	Precision	Recall	F-measure
IRELAND(pos/neg)	Avg	1/0.70	0.81/0.66	0.89/0.68
	Apriori	**1/0.70**	**1/0.66**	**1/0.68**
AEGEAN	Avg	0.44	1	0.61
	Apriori	**0.66**	**0.66**	**0.66**

Table 5.20: Precision/Recall and F-measure scores for the classification of weak vs. strong relations for the 'high-low' experiment.

From the two tables above it is shown that our method can efficiently classify the positive relations (in case of Ireland). For negative relations the precision and recall are not high but are over the baseline (according to the apriori of the classes). In the case of Aegean, precision and recall scores are not so high but again are over the baseline.

5.2.5 Flight traffic and co-citation networks

In this section, we present the correlation and MSE scores considering the flight traffic and co-citation networks. The results are presented for the three types of metrics: i) page-count-based (Tables 5.21 and 5.22), ii) text-based (Tables 5.23 and 5.24), iii) link-based (Tables 5.25 and 5.26), as well as for their linear combination (using equal weights) in Table 5.27.

5.2.5.1 Results on page-count-based metrics

Considering the page-count-based metrics for the case of the flight traffic network (see Table 5.21), we observe that the performance of the metrics is low throughout. More specifically, mutual information (S_I^P) performs slightly better in terms of correlation than any other metric, while Jaccard (S_J^P) and Dice (S_D^P) achieve significantly lower MSE. The same observations hold for the whole set of relations and for the 'weak-strong' relations set. Yet the performance of the metrics in the 'weak-strong' relation set is better than considering the whole set of 173 relations in both terms of correlation

Flight traffic network			
Relation set	Metric	Correlation	MSE
All	S_J^P	0.20	**0.32**
	S_D^P	0.20	0.34
	S_I^P	**0.22**	0.81
	S_G^P	0.21	0.81
Weak-Strong	S_J^P	0.30	**0.48**
	S_D^P	0.30	0.49
	S_I^P	0.32	0.63
	S_G^P	**0.34**	0.63

Table 5.21: Correlation and MSE for the flight traffic network and the page-count-based metrics.

Co-citation network		
Metric	Corr.	MSE
S_J^P	0.49	0.27
S_D^P	**0.49**	**0.26**
S_I^P	0.23	0.76
S_G^P	0.42	0.44

Table 5.22: Correlation and MSE for the researchers co-citation network and the page-count-based metrics.

and MSE. For the case of the co-citation network (see Table 5.22), Jaccard (S_J^P) and Dice (S_D^P) achieve better correlation scores than mutual information (S_I^P) or google-based semantic relatedness (S_G^P). The same observation also holds in terms of MSE. In general, it is observed that different metrics work efficiently in each case of network.

5.2.5.2 Results on text-based metric

Flight traffic network									
Text-based metric									
Relation set	Num. snippets	Correlation				MSE			
		B		LTF		B		LTF	
		AND	IND	AND	IND	AND	IND	AND	IND
All	100	-0.02	-0.10	**0.03**	-0.06	**0.56**	0.66	0.93	0.57
	200	**0.00**	-0.11	-0.01	-0.13	**0.60**	0.69	0.96	0.57
	500	**-0.01**	-0.03	-0.05	-0.09	**0.61**	0.82	1.08	0.63
Weak-Strong	100	**0.28**	0.04	0.27	0.03	**0.85**	1.34	1.10	1.25
	200	**0.25**	0.01	0.22	-0.04	**0.85**	0.86	1.01	1.00
	500	**0.19**	0.10	0.09	0.01	0.99	**0.76**	1.26	1.01

Table 5.23: Correlation and MSE for the text-based metric for the flight traffic network using different query types and number of snippets ($W = 10$).

Considering the text-based metric and the flight traffic network (see Table 5.23), the metric on AND and IND snippets achieves zero or negative correlation scores considering

Co-citation network								
Text-based metric								
Num. snippets	Correlation				MSE			
	B		LTF		B		LTF	
	AND	IND	AND	IND	AND	IND	AND	IND
100	0.46	0.32	**0.53**	0.32	0.51	0.26	0.51	**0.26**
200	**0.61**	0.23	0.57	0.31	0.34	0.28	0.47	**0.27**
500	**0.64**	0.31	0.64	0.36	0.34	0.42	0.39	**0.33**

Table 5.24: Correlation and MSE for the text-based metric for the researchers co-citation network using different query types and number of snippets ($W = 10$).

the whole set of relations ('All'). On the other hand, for the case of 'Weak-Strong' relations the metric performs better than the set of all relations but the correlation remains low. For the 'Weak-Strong' relations the performance of the metric is better on AND snippets than IND. Also it is observed that for the case of AND snippets, the performance of the metric deteriorates as more snippets are considered. Furthermore, considering AND snippets the B weighting scheme outperforms LTF in both correlation and MSE. For the case of the co-citation network (see Table 5.24), the performance of the metric on AND snippets is significantly better (0.64 for the 500 top-ranked snippets) than IND in terms of correlation. On the other hand, the metric applied on IND snippets performs better than AND in terms of MSE. It is also observed that correlation of AND snippets increases (and the MSE decreases) as more snippets are considered (for B weighting scheme going from 0.46 to 0.64).

5.2.5.3 Results on link-based metrics

Flight traffic network						
Link-based metrics						
Relation set	Correlation			MSE		
	S_G^L	S_T^L		S_G^L	S_T^L	
		B	LTF		B	LTF
All	-0.20	-0.21	**-0.18**	0.68	0.64	**0.61**
Weak-Strong	-0.25	-0.29	**-0.19**	1.15	**0.99**	1.01

Table 5.25: Correlation and MSE for the link-based metrics and the flight traffic network.

Considering the link-based metrics and the flight traffic network (see Table 5.25), all metrics achieve constantly negative correlation scores. This observation holds for both 'Weak-Strong' relations and for the whole set of 173 relations ('All'). We believe that this is due to the fact that the web documents that contain the actors (airports) share many outlinks, thus the number of common outlinks is not a good feature in this case. Considering the co-citation network (see Table 5.26), the S_G^L metric achieves better

Co-citation network					
Link-based metrics					
Correlation			MSE		
S_G^L	S_T^L		S_G^L	S_T^L	
	B	LTF		B	LTF
0.32	0.12	0.13	0.96	**0.32**	0.61

TABLE 5.26: Correlation and MSE for the link-based metrics for the co-citation network.

correlation than any other metric, but the correlation scores of all metrics are low. In terms of MSE the S_T^L metric with B weighting scheme achieves the lowest MSE score.

5.2.5.4 Combination of metrics

Similarly to the case of Ireland and Aegean, we examined the performance of the linear combination of the three types of metrics for the flight traffic and co-citation networks. For the case of the flight traffic network, we selected the google-based semantic relatedness (S_G^P) from the page-count metrics, the binary weighting scheme on the 100 top-ranked AND snippets for the text-based metrics and the binary weighted S_T^L from the link-based metrics. The results in terms of correlation and MSE are presented in Table 5.27. For the case of flight traffic network, the presented results are only for the 'weak-strong' relation set.

Fusion of metrics (equal weights)						
Relationship strength						
Weights			Flight traffic		Co-citation	
λ_P	λ_T	λ_L	Correlation	MSE	Corr.	MSE
1	0	0	0.34	**0.63**	0.49	0.26
0	1	0	0.28	0.85	0.64	0.34
0	0	1	-0.29	0.99	0.32	0.96
0	0.5	0.5	-0.07	0.77	0.63	0.50
0.5	0	0.5	0.09	0.90	0.49	0.26
0.5	0.5	0	**0.37**	0.81	**0.64**	**0.15**
0.33	0.33	0.33	0.18	0.93	0.64	0.21

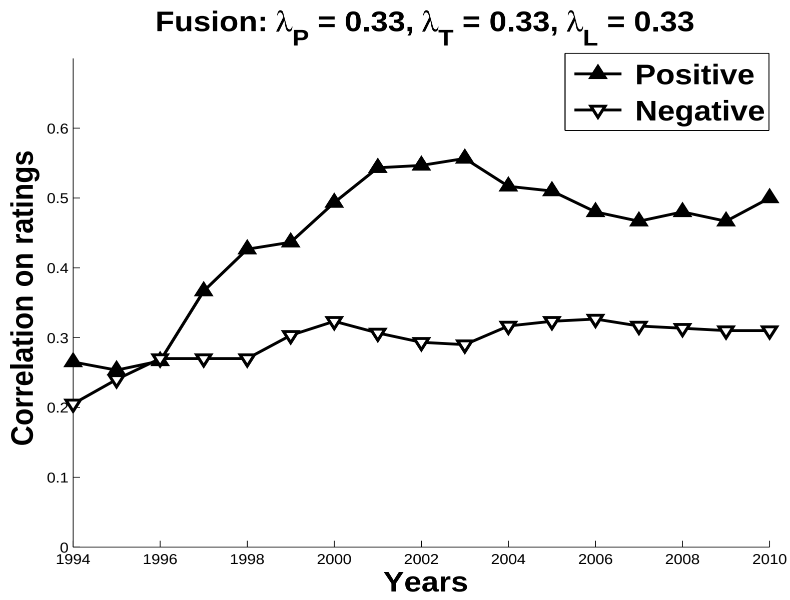
TABLE 5.27: Correlation and MSE on relationship strength for individual metrics and their linear fusion, flight traffic and co-citation networks.

From Table 5.27, it is observed that for the case co-citation network the combination of page-count and text-based metrics reduces the MSE of the individual metrics significantly (from 0.34 tp 0.15), while the correlation score remains the same. On the other hand, for the case of the flight traffic network none of the examined combination schemes enhances the performance significantly.

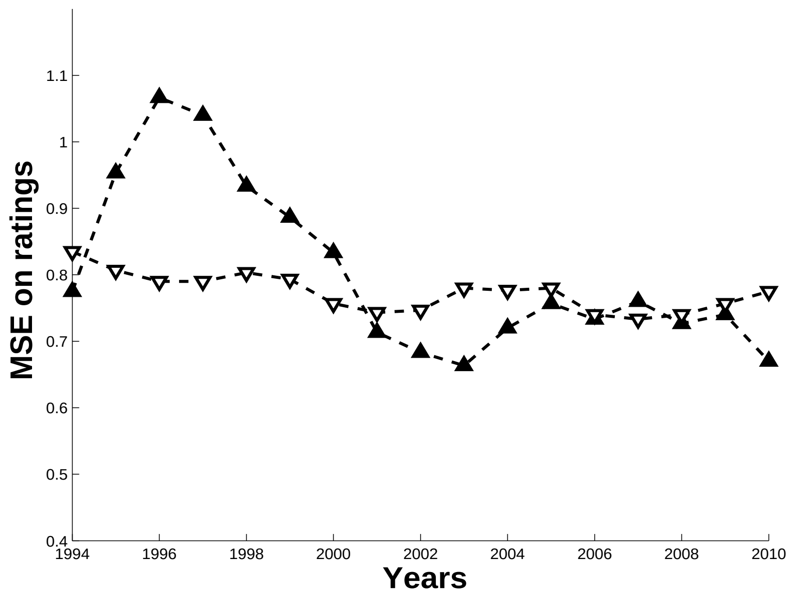
5.2.6 Evaluation over a time period

In this section we evaluate the performance of the linear fusion metrics using equal weights in terms of correlation and MSE for the selected time period 1994-2010. The main goal of this experiment is to investigate how the performance of our method is affected by the insertion of time parameter. For each year in 1994-2010 period we measure the correlation and MSE score between the corresponding extracted network and the human-rated. Thus, we can also examine during which years the extracted networks best match the human-rated. The evaluation metrics were applied on computed relationship strengths and degree of centrality. For each case study, we selected the metrics that can better capture the change in the performance of the proposed method. For Ireland case study and positive relations, we have selected the mutual information (S_I^P) from the page-count-based metrics, the LTF weighting scheme on the 200 AND snippets for the text-based metric and the binary weighted cosine similarity S_T^L from the link-based metrics. For the negative relations, we have selected the dice (S_D^P) from the page-count metrics, the LTF weighting scheme on the top-ranked 200 IND snippets for the text-based metric and the binary weighted cosine similarity S_T^L from the link-based metrics. For the South Aegean case study we have selected the google (S_G^P) from the page-count metrics, the LTF weighting scheme on the top-ranked 200 AND snippets for the text-based metric and the binary weighted cosine similarity S_T^L from the link-based metrics. The correlation and MSE on ratings are presented in Figures 5.1 and 5.2 for Ireland and South Aegean networks respectively. The presented evaluation results are for the '3-levels' experiment as for the 'high-low' experiment the same conclusions hold. To avoid the presence of noise, the correlation and MSE scores are smoothed using a moving average window of 3 years including the current year i.e., we take the average of the previous, current and next date.

For the Ireland case study and positive relations, for the time period 1994-1998 a low correlation score is observed (and high MSE score) as the number of web documents containing the actors is small. The low performance of the metrics during this time period is reasonable as many of the actors do not exist (or co-exist) in web documents. For the time period 2001-2006 the correlation gets its maximum values (and the MSE its lowest values) and for the 2006-2010 the correlation and MSE scores remain stable. This observation clearly shows that the extracted policy network is in best agreement with the original when the political research took place (2001-2003). Furthermore, after the end of the CSF (year 2006) the actors retained the strength of the relations among them. For the case of negative relations the correlation score consistently increases (and the MSE consistently decreases) showing that the relations among the actors simply evolute without changing significantly. For the Aegean case study, a slight increase in



(a)



(b)

FIGURE 5.1: Evaluation results on relationship strength for **Ireland** case study, linear fusion of metrics : Correlation (a) and MSE (b).

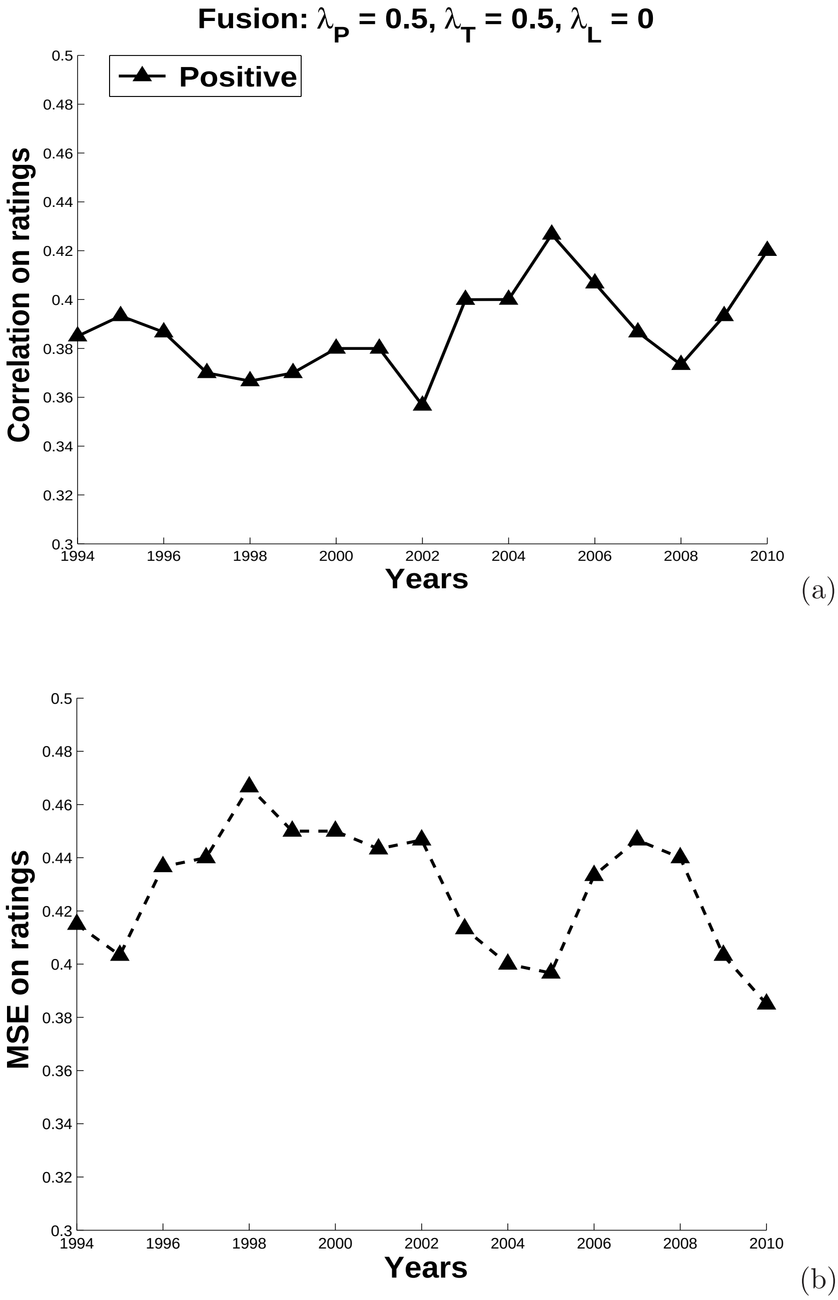
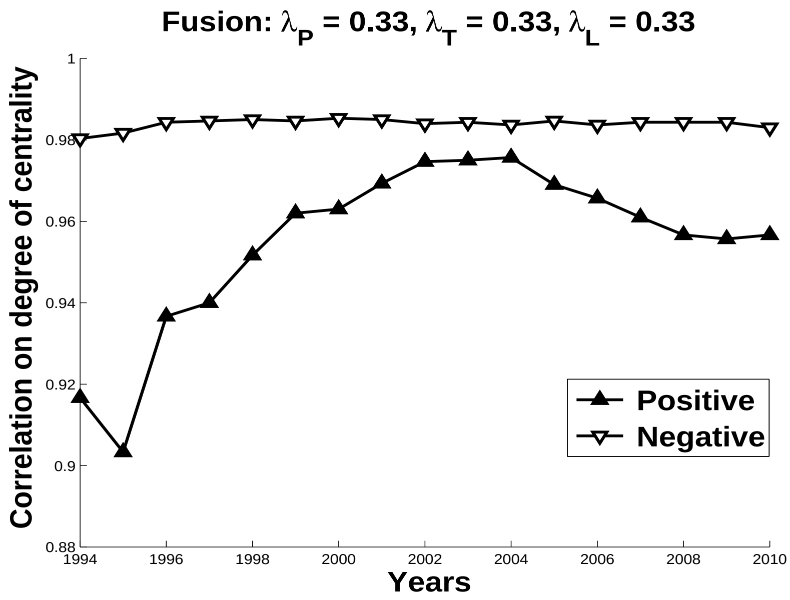
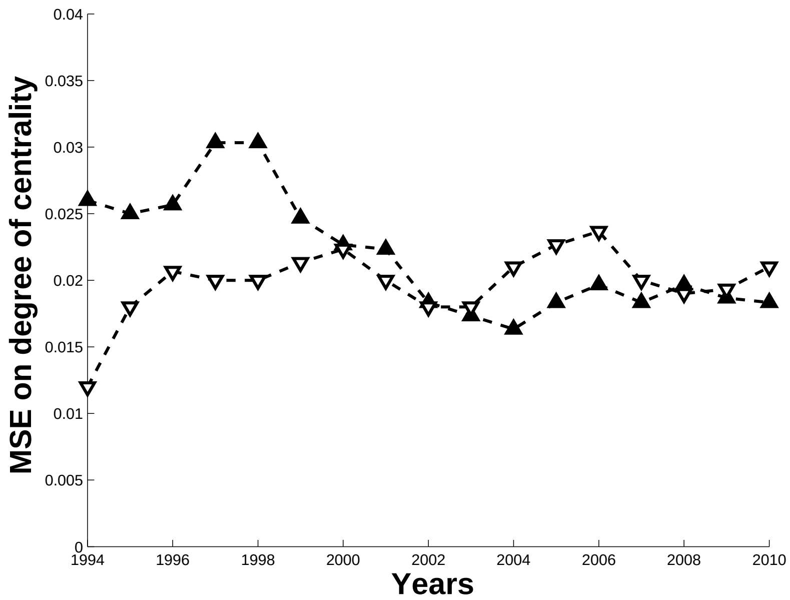


FIGURE 5.2: Evaluation results on relationship strength for **South Aegean** case study, linear fusion of metrics : Correlation (a) and MSE (b).

the correlation is observed during the 2003-2006 time period. This observation shows that the extracted network best matches the human rated in the time period that is close to the period of research. After the end of the CSF the actors in the Aegean policy network did not retained their relation strengths as the case of Ireland. Finally, the conclusions above are in consistency with the evaluation results on the degree of centrality that are presented in Figures 5.3 and 5.4 for Ireland and Aegean respectively.

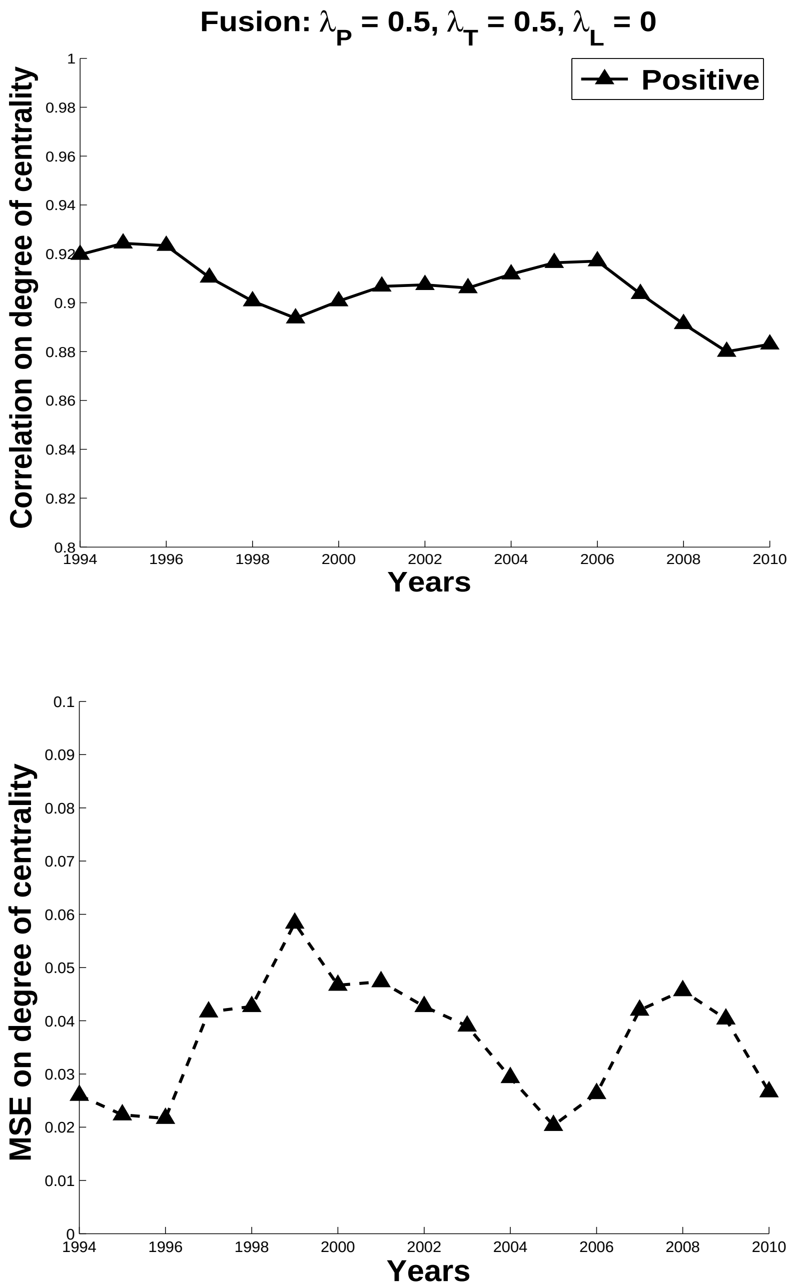


(a)



(b)

FIGURE 5.3: Evaluation results on degree of centrality for **Ireland** case study, linear fusion of metrics : Correlation (a) and MSE (b)



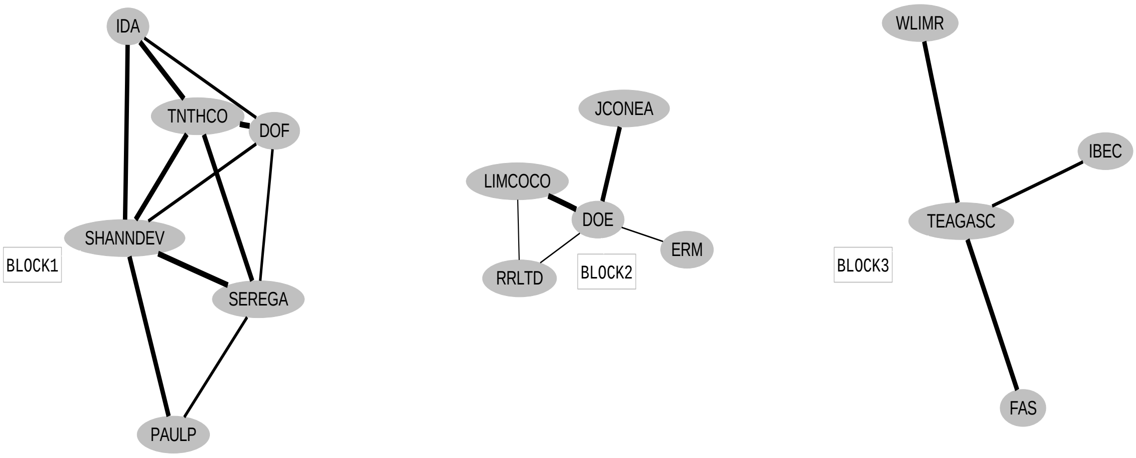
(a)

(b)

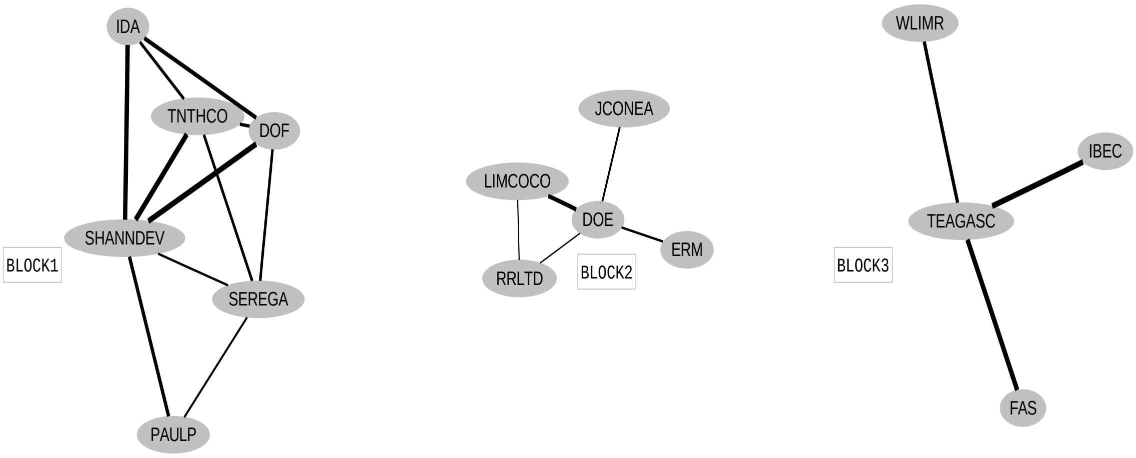
Figure 5.4: Evaluation results on degree of centrality for **Aegean** case study, linear fusion of metrics : Correlation (a) and MSE (b)

5.3 Network Visualization

In this section, the manually annotated and automatically extracted networks for both case studies are displayed as graphs. In Figures 5.5, 5.6 and 5.7, the graphs of the original and extracted policy networks are shown for Ireland (positive and negative relations) and the Aegean, respectively. The nodes on the graphs correspond to political actors in the relations under examination while the edges show the relations among them. The nodes are labeled using the acronyms of the actors supplied by political scientists We use the relatedness scores from the three-way linear combination of all metrics using equal weights (see '3-levels' experiment in Table 5.13). The relatedness scores take values in [1,3] according to (4.1). For the visualization of both original and extracted networks, we followed the procedure in Chapter 3.



(a)



(b)

FIGURE 5.5: Ireland case study network graphs for **positive relations**: (a) original and (b) extracted network.

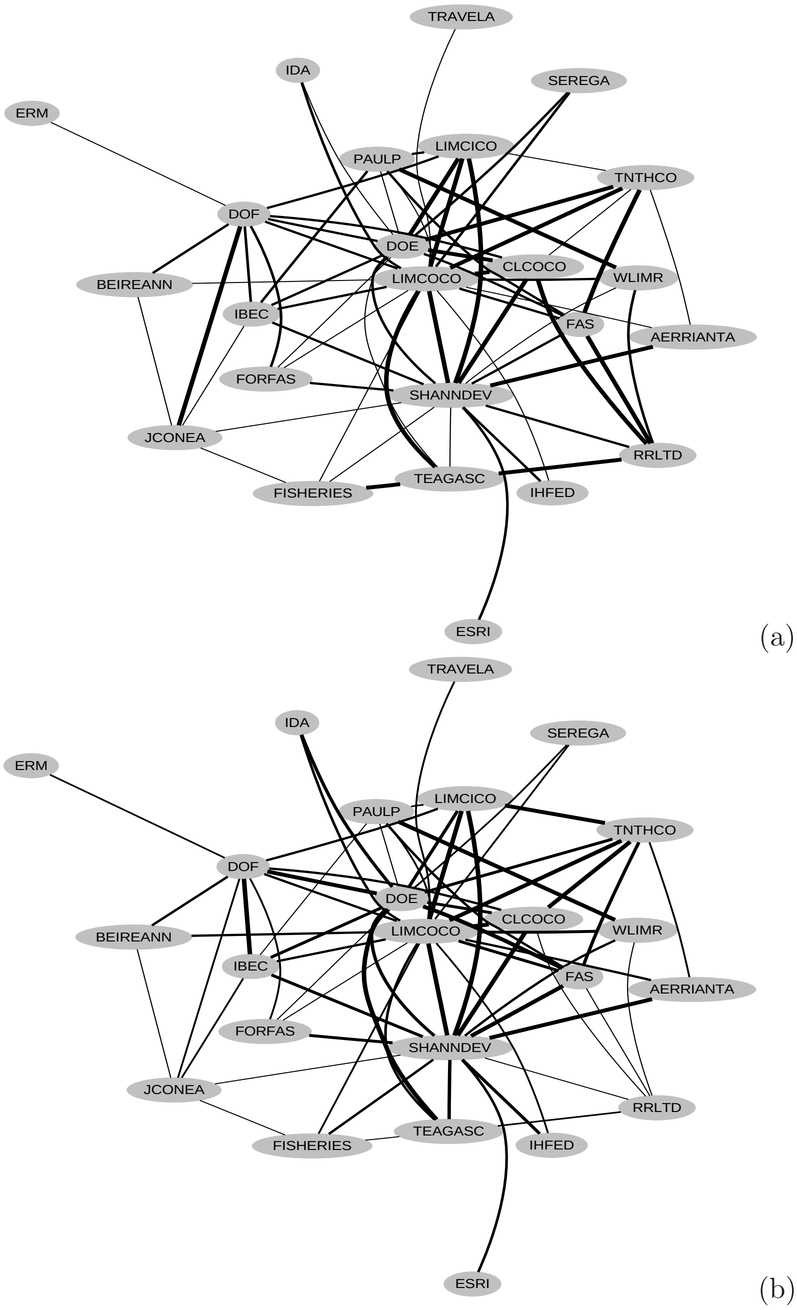
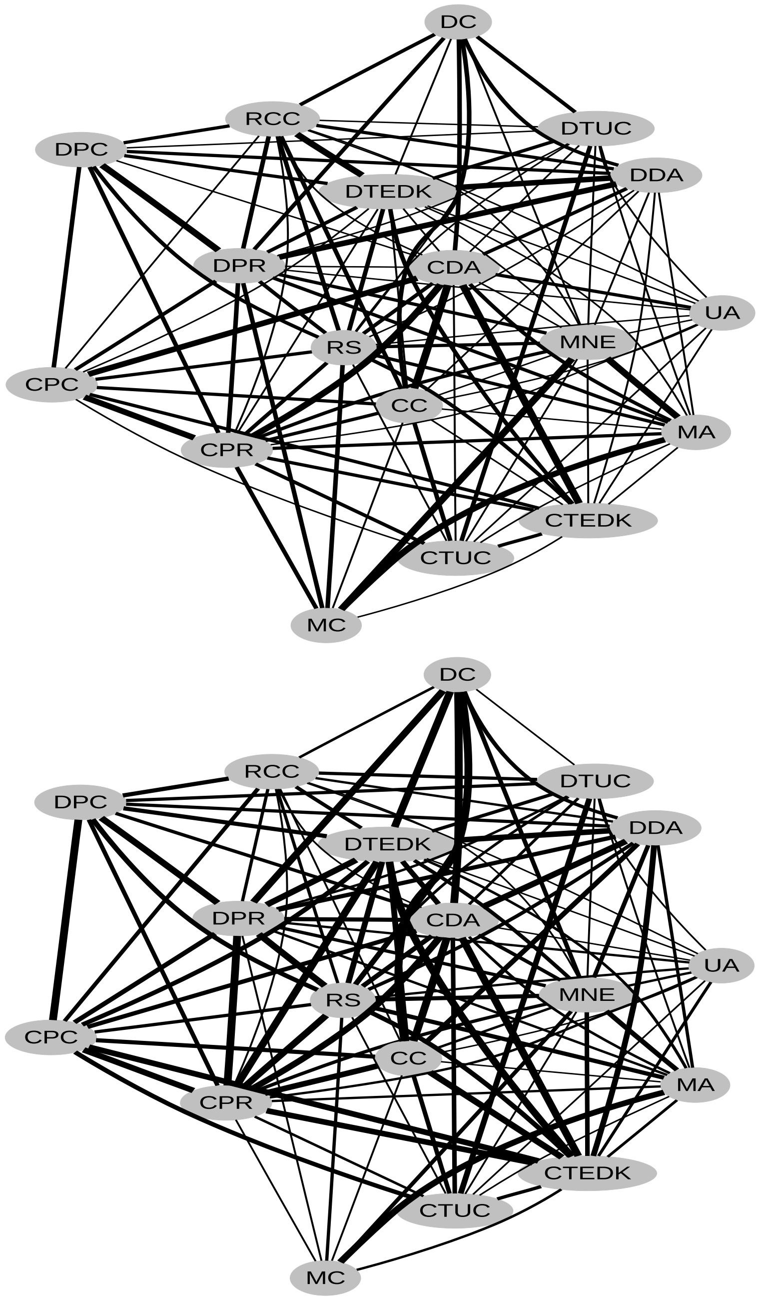


FIGURE 5.6: Ireland case study network graphs for **negative relations**: (a) original and (b) extracted network.

The graphs for the positive relations of the Ireland network are shown in Figures 5.5a, 5.5b for the original and extracted networks respectively. Note that three sub-graphs each corresponding to one of the diagonal blocks (positive relations) of the relatedness matrix are shown. Overall, there is good agreement for the strength of relations between actors, as expected, from the high correlation scores. More specifically, considering positive relations, the SHANNDEV actor which is a central node (see Figure 5.5b graph BLOCK1) is strongly connected to IDA and TNTHCO and less strongly connected to PAULP (compared to the original). In addition, DOF is not so strongly connected to SEREGA in both extracted and original graphs. Only SEREGA appears somewhat less connected to the rest of extracted network (compared to the original). Actor RRLTD less connected to DOE and LIMCOCO (see Figure 5.5b graph BLOCK2) in both extracted and original networks. DOE is strongly connected to LIMCOCO but less connected to JCONEA (contrary to the original network). TEAGASC which is the central node in graph of BLOCK3 is strongly connected to WLIMR and IBEC in both original and extracted networks.

The negative relations that appear in the off-diagonal blocks are lumped together in a single network shown in Figures 5.6a, 5.6b for original and extracted networks respectively. Despite the very low correlation scores achieved for negative relations, the original and extracted graphs looks reasonable similar, e.g., the actors in the {DOE, LIMCOCO, CLCOCO} clique are strongly interconnected in both graphs. Furthermore, peripheral actors (e.g., TRAVELA, ERM, ESRI) are less connected to other more central actors in both extracted and original graphs. The central actors (e.g., SHANNDEV, LIMCOCO, CLCOCO) are strongly interconnected in both extracted and original policy networks.

A qualitative analysis of the Aegean graphs in Fig. 5.7 reveals very similar connectivity patterns for most of the actors in the original and extracted network. For example, the actors {CDA, DTEDK, DPR, RS, CC} have high connectivity and are central in both graphs, while the actors {UA, CTUC, MC, RCC} have weaker relations and are peripheral (again in both). However, there are some actors that have increased their relationship strength and connectivity, and have become more central in the extracted network, e.g., {DC, CTEDK}. Overall, the qualitative analysis of the extracted graphs shows good agreement with those from political scientists. The ultimate judge of the quality and usefulness of the extracted networks are of course the political scientists, their views on the extracted networks are discussed in Chapter 6.



(a)

(b)

FIGURE 5.7: South Aegean case study network graphs: (a) original and (b) extracted network.

5.4 Visualization of network evolution

In this section we visualize the network evolution for the Ireland and Aegean case studies. We present indicative snapshots of the extracted networks for 2001, 2003, 2007, and 2010 years for both case studies. The extracted networks presented are for the fusion of metrics using equal weights and for the '3-levels' experiment. More specifically for Ireland case study and positive relations, we selected the two-way combination of the page-count and link-based metrics ($\lambda_P = 0.5$, $\lambda_T = 0$, $\lambda_L = 0.5$) as it produces good correlation and MSE results (see Table 5.13). For the negative relations, we selected the two-way combination of the page-count and text-based metrics ($\lambda_P = 0.5$, $\lambda_T = 0.5$, $\lambda_L = 0$). For the Aegean case study we selected the two-way combination with equal weights on the page-count and text-based metrics as in this case good performance is achieved (see Table 5.13). In Figures 5.8, 5.9 and 5.10 the networks for Ireland are presented for positive and negative relations respectively. In Figure 5.11 and 5.12, the networks for Aegean are presented.

For Ireland case study and positive relations (see Figure 5.8), we observe that the degree centrality for the majority of actors of national level changes for all selected years. Considering the actors of national level, JCONEA (BLOCK2) increases its activity in 2007 (one year after the end of the 3rd CSF). Regional actors such as TEAGASC, FAS (BLOCK3) and IDA (BLOCK1) and local actors such as WLIMR and TNTHCO have also increased their activity on the network. Furthermore, it is observed that regional and local actors such as SHANNDEV which is a central actor in BLOCK1 and SEREGA reduce their activity. For Ireland case study and negative relations (see Figures 5.9 and 5.10), we observe that the degree centrality for the majority of actors increases. It is interesting that peripheral actors of local level such as TRAVELA, TNTHCO, WLIMR and RRLTD have increased their activity significantly during and after the 3rd CSF. The same observation holds for the regional actors of regional level BEIREANN, FISHERIES, TEAGASC.

For Aegean case study (see Figures 5.11 and 5.12), we observe that all actors increase their activity for the years 2001 and 2003. National actors MNE, MA and MC are more active during the years of the 3rd CSF (years 2001 and 2003), while their activity decreases after the 3rd CSF (years 2007 and 2010). This reduction in the activity of national actors might be the result of the reduction of fundings after the CSF. Peripheral actors of local level such as DC, DPC, CTUC, CTEDK, DDA, DTUC have increased their activity significantly after the year 2001 and stayed active during the 2003, 2007 and 2010 years. Overall, it is observed that local actors central or peripheral have increased their connectivity over the years.

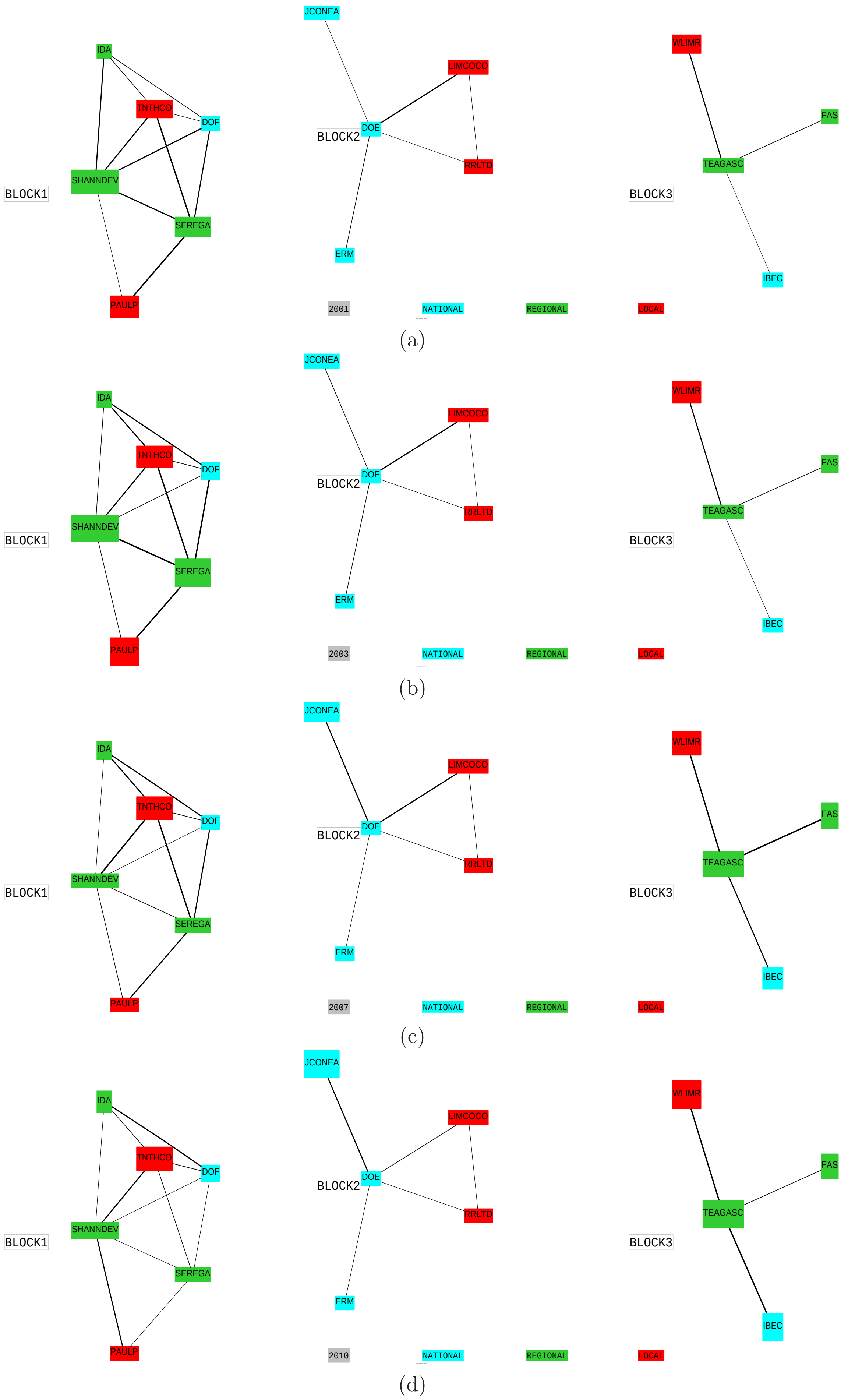


FIGURE 5.8: Snapshots for Ireland network (positive relations): (a) 2001, (b) 2003, (c) 2007, (d) 2010.

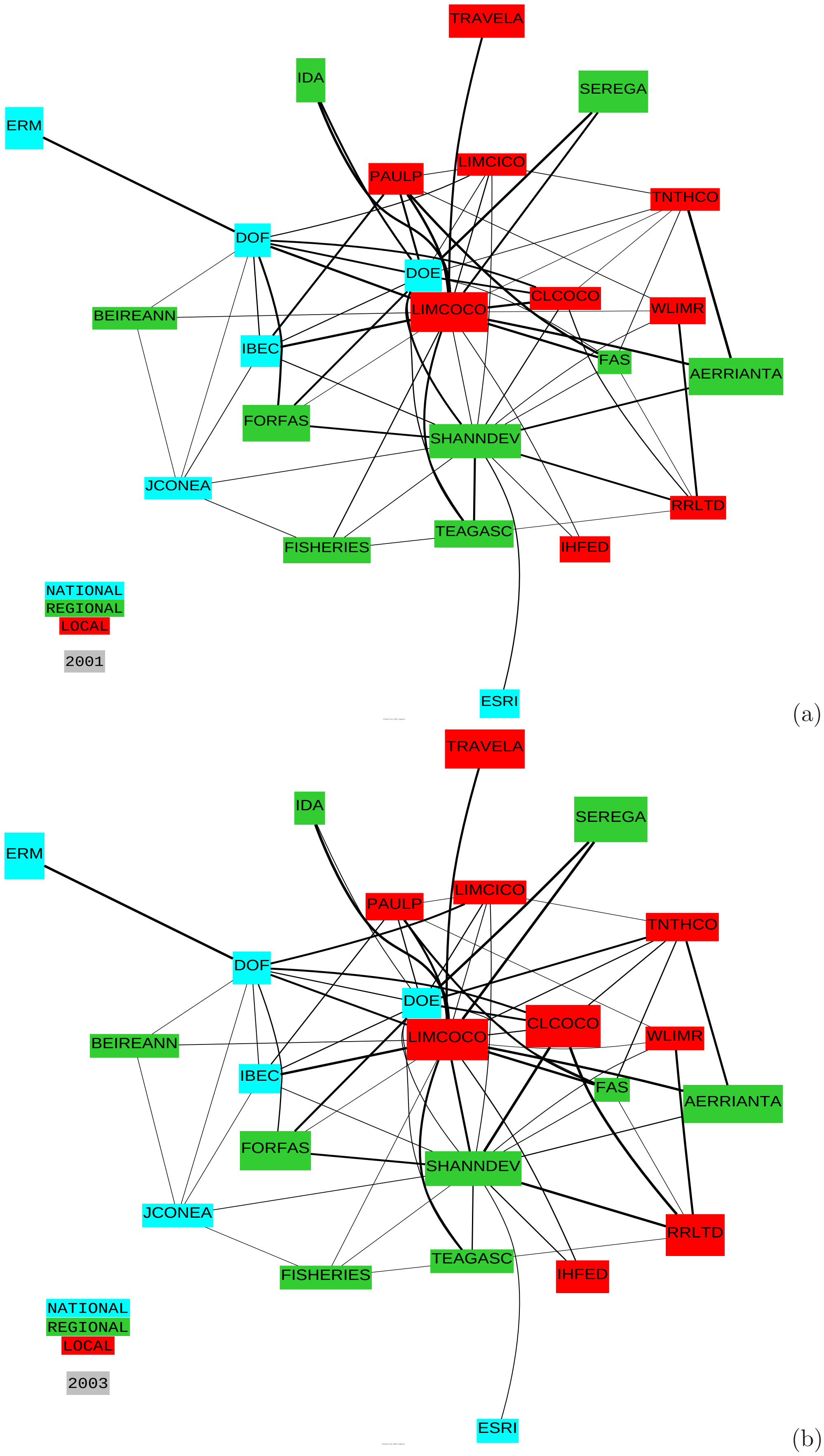


FIGURE 5.9: Snapshots for Ireland network (negative relations): (a) 2001, (b) 2003.

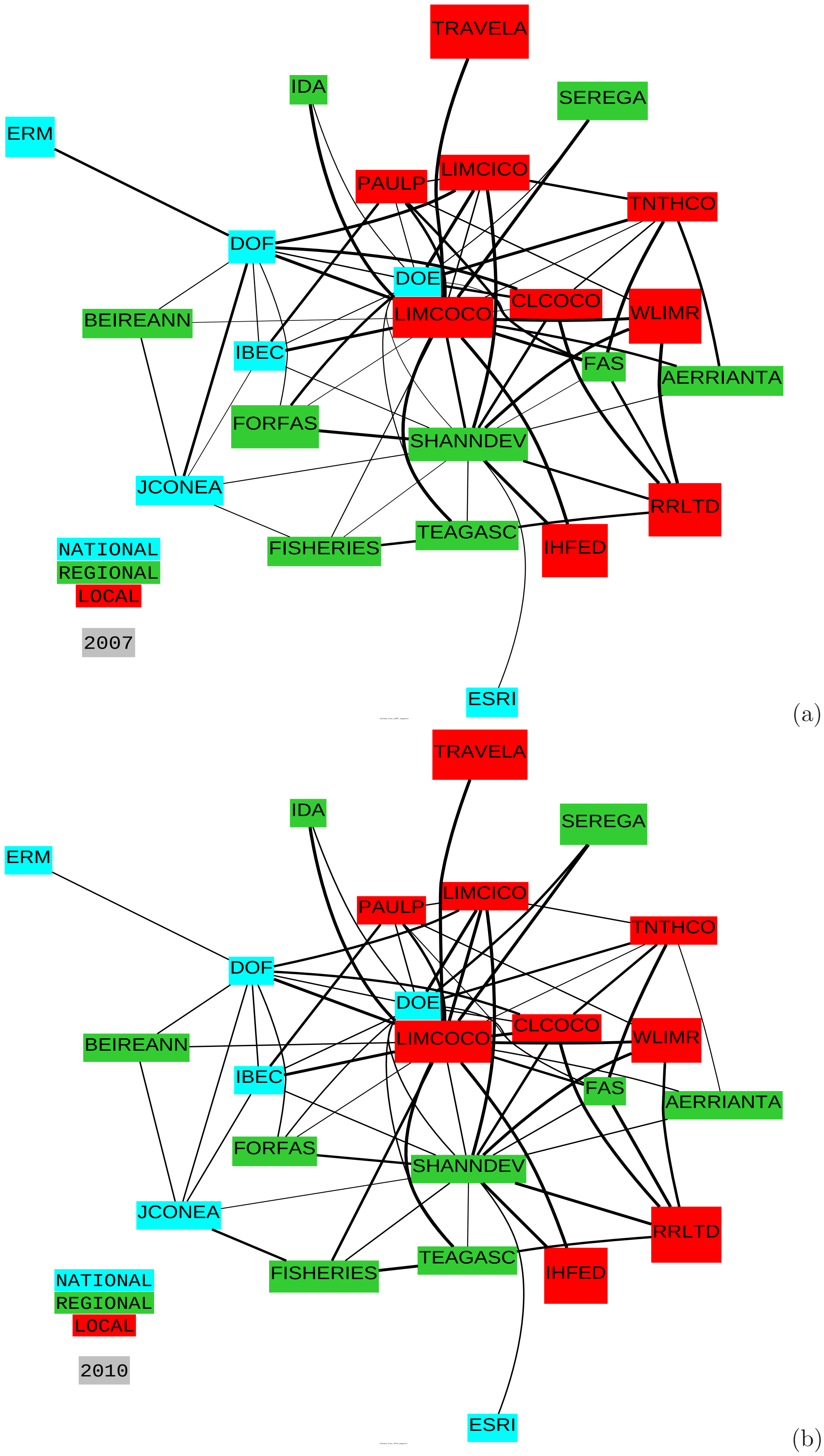


FIGURE 5.10: Snapshots for Ireland network (negative relations): (a) 2007, (b) 2010.

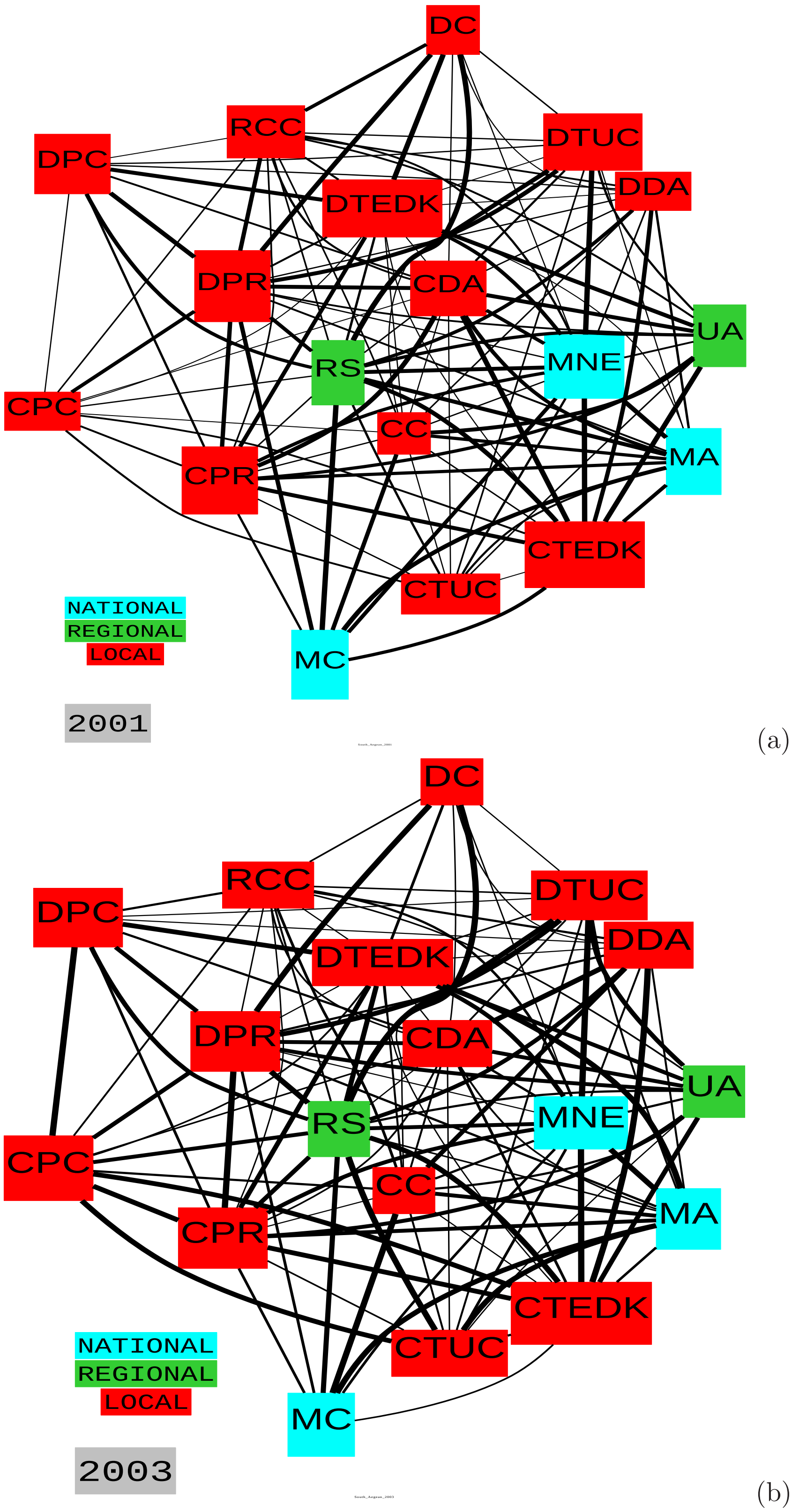
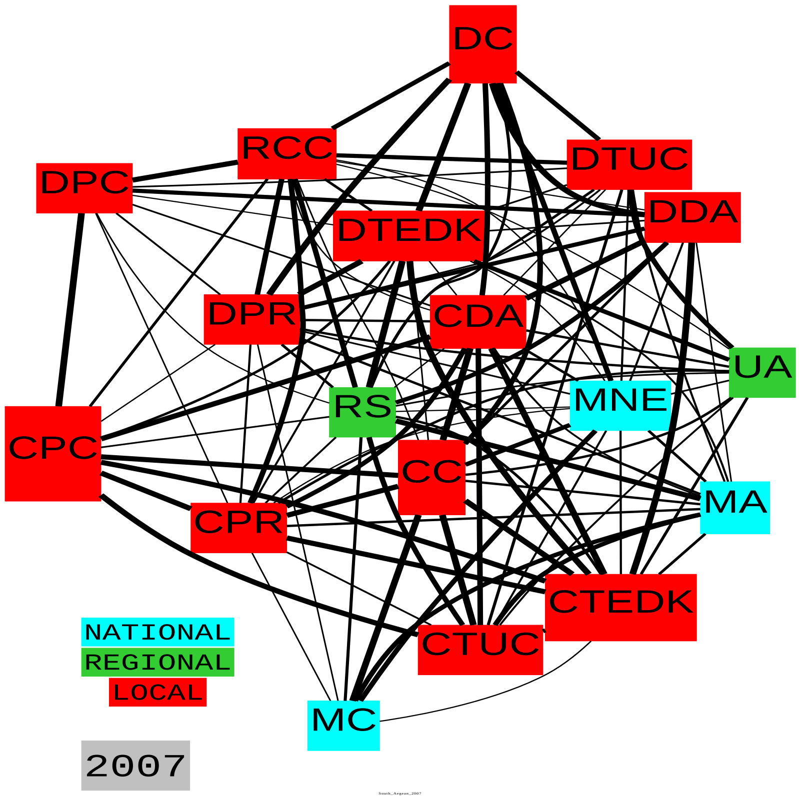
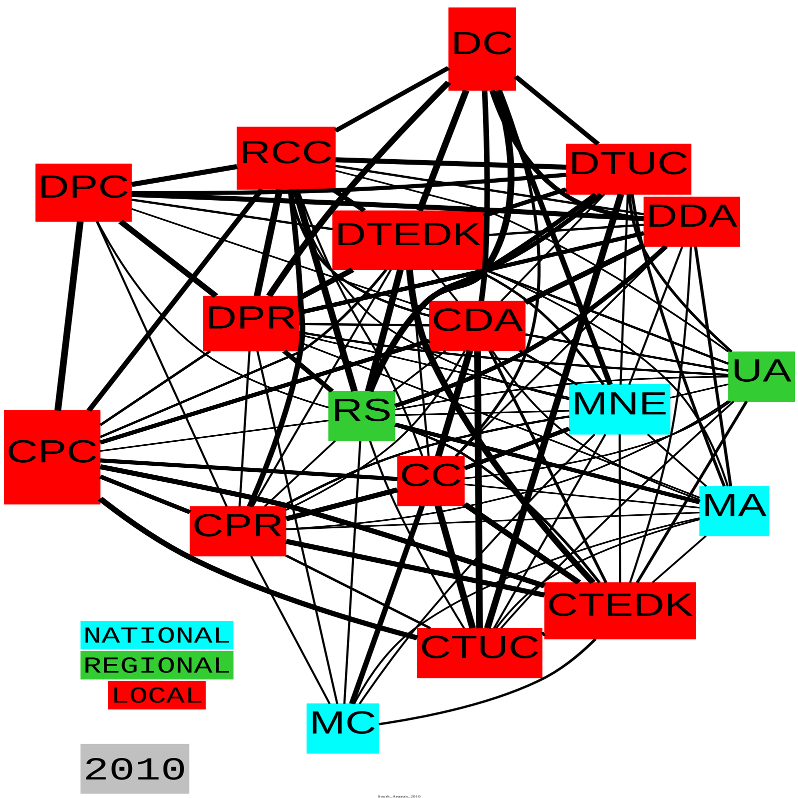


FIGURE 5.11: Snapshots for Aegean network: (a) 2001, (b) 2003.



(a)



(b)

FIGURE 5.12: Snapshots for Aegean network: (a) 2007, (b) 2010.

Chapter 6

Conclusions and Future work

In this chapter, we attempt to interpret the results from Chapter 5 both from an engineering and a political science standpoint. Important parameters that affect the quality of the automatically extracted network (in addition to the relatedness metrics used) include data sparseness, lexical ambiguity for actors, language, type of network relations, and network evolution over time. When comparing manually annotated policy networks with automatically extracted ones, human biases also come into play, e.g., cultural biases of the interviewers and interviewees, (non-linear) scaling of the relatedness metrics by the political scientists. In the following sections, we discuss these factors and we summarize the conclusions of this work. Finally we give research directions for future work.

6.1 Conclusions

One of the major problems in both policy networks Ireland and especially Aegean is data sparsity. In both policy networks the actor names are named entities and with different word-forms such as full names (consisting of more than three words) and acronyms. Inserting only the full names in our queries, the search engine returns small number of hits, while inserting acronyms makes the queries suffer from ambiguity. Thus there is no obvious solution to more effective query formulation than using both forms of actors' names. Furthermore the problem of ambiguity remains for the Ireland case study as the same abbreviations are used in other countries such as US and UK. On the other hand, in the case of Aegean, the use of Greek language somewhat tackles ambiguity. A third crucial factor who influences the performance of the metrics and our method in general is the fact that that Greek and Irish respondents might conceived differently the notion of 'weak' and 'strong' relations i.e., in Ireland case study the relations were rated uniformly, while in the Aegean the majority of relations were rated with '1' or '2' (weak or medium) and only a low number of relations were rated with '3'. This difference shows that Greek and Irish respondents conceived differently the notion of 'weak' and 'strong' relations. Another factor that plays significant role is the fact that the original networks have been created during the 2001-2003 period and the ratings given express the relationship strengths for the specific time period.

Keeping the above factors in mind, we discuss the performance of our method in each case. Considering the Ireland case study, the page-count-based metrics perform well on positive relations. It is shown that positively related actors co-operate much often as they co-occur in many web documents and page-count-based metrics can effectively estimate such relations. However the performance of page-count metrics in Aegean case study is modest. The difference in the performance of the page-count metrics in the two case studies depends mainly on data sparsity which is more intense on Aegean. Link-based metrics are highly correlated with positive relations for Ireland. Positively related actors have common 'friends'. These friends are represented by the outlinks that are referred to the web pages that contain the positively related actors. On the other hand this is not true for Aegean case study, as the web pages indexed by the actors do not refer to common outlinks which is an issue of data sparseness. The performance of page-count and especially link-based metrics is low on negative relations. This happens because negatively related actors do not co-occur oftenly in web documents or the pages containing them do not refer to common links. Considering the text-based metrics, they work somewhat better than the page-count and link-based only for negative relations. Generally, text-based metrics work efficiently for relations of semantic similarity (or dissimilarity) but they do not perform well on more lax relations such co-operation.

In general, linear combinations produce better results especially when we combine the two best metrics. In addition, the metrics as well as their combinations perform better (in terms of correlation) for the 'high-low' experiment than the '3-levels' experiment showing that the proposed metrics efficiently discriminate the strong relations from the weak. However, they fail to discriminate relations of medium strength. Furthermore, the geopolitical domain restriction (*ie* for Ireland, *gr* for Aegean) enhances the performance of page-count-based metrics in both case studies as it solves somewhat the ambiguity problem. In general, the metrics achieve excellent results considering the identification of the more active actors in both case studies. The correlation and MSE results on degree centrality of nodes (actors) are very good. Another interesting conclusion is that performance of the method depends on the year parameter. The relations evolute over time and even the manually extracted networks are biased on the time period the research took place. More specifically, for the Ireland case study the ratings on positive relations are biased on the period of the political research (or other factors such as funding). On the other hand, negative relations simply evolute. Considering the case of the South Aegean, the ratings are shown to be biased on the period 2004-2006 which is the time period of the 3rd CSF. With the use of network visualization we can qualitatively evaluate our method. The visualization of the extracted networks are very close to the human rated. Finally, by presenting indicative snapshots of the networks we gave the possibility to visualize the change of actors' activity over time which is crucial for the explanation of the policy making process.

6.2 Future Work

Even though the conclusions of this work are interesting from both points of data mining and political science, many problems remain open. At first the automatic extraction of all possible word-forms of the actors can be applied to reduce data sparsity and ambiguity in the downloaded data. The extracted word-forms can be used for a more efficient query expansion. A drawback of the proposed method is that no feature selection has been applied on lexical or link-based features. We believe that applying a classical feature selection method to find the most discriminative features would enhance the performance the metrics and our approach in general. Referring to features, we believe that it is of great challenge to identify words that express the existence of a positive or a negative relation. In general the metrics are shown to perform well on positive relations contrary to negative relations. Thus, a method or metric that efficiently estimates negative relations is an open problem. As the simple linear fusion of the metrics is proved to work better than the individual metrics, it is worth investigating other more sophisticated approaches of fusing different sources of information.

Appendix A

Appendix

In appendix the tables with the actors' names (as given from political scientists) their acronyms and the manually retrieved lexicalizations are inserted.

Num	Acronym	Lexicalizations
1	SHANNDEV	Shannon Development, Shannon Free Airport Development Company
2	SEREGA	South-East Regional Assembly, Southern & Eastern Regional Assembly, S&E Regional Assembly
3	PAULP	Paul Partnership, Paul Partnership Ltd
4	IDA	Industrial Development Authority, Industrial Development Agency, IDA Ireland
5	DOF	Department of Finance, An Roinn Airgeadais
6	TNTHCO	North Tipperary County Council, North Tipperary County Co, North Tipperary Co Co
7	ERM	Environmental Resource Management, Environmental Resources Management
8	RRLTD	Rural Resources, Rural Resources Ltd
9	LIMCOCO	Limerick County Council, Limerick County Co, Limerick Co Co
10	JCONEA	Joint Committee on European Affairs, Joint Committee European Affairs

Table A.1: The actor names and their lexicalizations for the **Ireland** network (1-10).

Num	Acronym	Lexicalizations
11	DOE	Department of the Environment, Department of Environment
12	WLIMR	West Limerick Resources
13	IBEC	Irish Business and Employers Confederation, IBEC
14	TEAGASC	
15	FAS	FÁS, Foras Áiseanna Saothair, Foras Aiseanna Saothair, Irish National Training and Employment Authority
16	FISHERIES	Shannon Regional Fisheries Board, Regional Fisheries Board
17	AERRIANTA	AER Rianta International, Dublin Airport Authority, Aer Rianta
18	FORFAS	FORFÁS, Forfás, FORFAS
19	TRAVELA	Irish Travel Agents Association, Travel Agents Association
20	BEIREANN	Bus Éireann, Bus Eireann
21	IHFED	Irish Hotel Federation, Irish Hotels Federation
22	LIMCICO	Limerick City Council, Limerick City Co
23	CLCOCOC	Clare County Council, Clare Co Co
24	ESRI	Economic and Social Research Institute, ESRI

Table A.2: The actor names and their lexicalizations for the **Ireland** network (11-24).

Num	Acronym	Lexicalizations
1	MNE	Υπουργείο Οικονομίας, Υπουργείο Οικονομικών, Υπουργείο Οικονομίας · Οικονομικών
2	MA	Διαχειριστική Αρχή Κοινοτικού Πλαισίου Στήριξης, Διαχειριστική Αρχή ΚΠΣ
3	MC	Επιτροπή Παρακολούθησης Κοινοτικού Πλαισίου Στήριξης, Επιτροπή Παρακολούθησης ΚΠΣ
4	MOU	Μονάδα Οργάνωσης · Διαχείρησης Κοινοτικού Πλαισίου Στήριξης, Μονάδα Οργάνωσης Διαχείρισης ΚΠΣ, ΜΟΔ
5	RS	Περιφέρεια Νοτίου Αιγαίου, Περιφερειάρχης Νοτίου Αιγαίου
6	UA	Πανεπιστήμιο Αιγαίου
7	CPC	Νομαρχιακό Συμβούλιο Κυκλάδων, Νομαρχιακού Συμβουλίου Κυκλάδων
8	CPR	Νομαρχία Κυκλάδων, Νομαρχιακή Αυτοδιοίκηση Κυκλάδων
9	CTUC	Εργατικό Κέντρο Κυκλάδων, Εργατικό Κέντρο Νομού Κυκλάδων
10	CTEDK	ΤΕΔΚ Νομού Κυκλάδων, ΤΕΔΚ Κυκλάδων, Ένωση Δήμων και Κοινοτήτων Νομού Κυκλάδων
11	CC	Επιμελητήριο Νομού Κυκλάδων, Επιμελητήριο Κυκλάδων
12	CDA	Αναπτυξιακή Εταιρεία Κυκλάδων, Αναπτυξιακή Εταιρεία Νομού Κυκλάδων, Αναπτυξιακή Κυκλάδων
13	RCC	Δημοτικό Συμβούλιο Ρόδου, Δημοτικό Συμβούλιο Ροδίων
14	DPC	Νομαρχία Δωδεκανήσου, Νομαρχιακή Αυτοδιοίκηση Δωδεκανήσου, Νομαρχία Δωδεκανήσων, Νομαρχιακή Αυτοδιοίκηση Δωδεκανήσων
15	DC	Επιμελητήριο Δωδεκανήσου, Επιμελητήριο Δωδεκανήσων
16	DTEDK	ΤΕΔΚ Νομού Δωδεκανήσου, ΤΕΔΚ Δωδεκανήσου, Ένωση Δήμων και Κοινοτήτων Νομού Δωδεκανήσου, ΤΕΔΚ Νομού Δωδεκανήσων
17	DDA	Αναπτυξιακή Εταιρεία Δωδεκανήσου, Αναπτυξιακή Εταιρεία Νομού Δωδεκανήσου, Αναπτυξιακή Εταιρεία Δωδεκανήσων, Αναπτυξιακή Εταιρεία Νομού Δωδεκανήσων
18	DTUC	Εργατικό Κέντρο Ρόδου, Εργατικό Κέντρο Ροδίων

Table A.3: The actor names and their lexicalizations for the **Aegean** network.

Bibliography

[1] H. Moschitz and M. Stolze, “Organic Farming Policy Networks in Europe: Context, Actors and Variation,” *Development of organic farming policy in Europe*, vol. 34, no. 3, pp. 258–264, 2009.

[2] M. Brockhaus, R. Vignola, and F. Kalame, “Decision Making and Policy Networks in Forestry-dependent Development Sectors: Mainstreaming Adaptation to Climate Change into Policy,” in *TroFCCA Policy Research Workshop*, 2007.

[3] M. Schneider, J. Scholz, M. Lubell, D. Mindruta, and M. Edwardsen, “Building Consensual Institutions: Networks and the National Estuary Program,” *American Journal of Political Science*, vol. 47, no. 1, pp. 143–158, 2003.

[4] C. S. Thomas, *First World Interest Groups: A Comparative Perspective*. Westport: Greenwood Press, 1993.

[5] J. L. Moreno, *Who Shall Survive?* Washington, DC: Nervous and Mental Disease Publishing Company, 1934.

[6] P. V. Marsden, “Network Data and Measurement,” *Annual Review of Sociology*, vol. 16, pp. 435–463, 1990.

[7] P. V. Marsden and K. E. Campbell, “Measuring Tie Strength,” *Social Forces*, vol. 63, no. 2, pp. 481–502, 1984.

[8] M. S. Granovetter, “The Strength of Weak Ties,” *American Journal of Sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.

[9] S. Wasserman and K. Faust, *Social Network Analysis*. Cambridge: Cambridge University Press, 1994.

[10] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[11] M. L. Northway, “A Method for Depicting Social Relationships Obtained by Sociometric Testing,” *Sociometry*, vol. 3, no. 2, pp. 144–150, 1940.

[12] A. S. Klovdhal, “A Note of Images of Networks,” *Social Networks*, vol. 3, no. 3, pp. 197–214, 1981.

[13] P. Eades and K. Sugiyama, “How to Draw a Directed Graph,” *Journal of Information Processing*, vol. 13, no. 4, pp. 424–437, 1991.

[14] K. Sugiyama, S. Tagawa, and M. Toda, “Methods for Visual Understanding of Hierarchical System Structures,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. 11, no. 2, pp. 109–125, 1981.

[15] P. Eades and N. C. Wormald, “Edge Crossings in Drawings of Bipartite Graphs,” *Algorithmica*, vol. 11, no. 4, pp. 379–403, 1994.

[16] H. Nagamochi, “An Improved Approximation to the One-Sided Bilayer Drawing,” in *Graph Drawing*, vol. 2912 of *Lecture Notes in Computer Science*, pp. 406–418, 2004.

[17] J. B. Kruskal, “Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis,” *Psychometrika*, vol. 29, pp. 1–27, 1964.

[18] T. Kamada and S. Kawai, “An Algorithm for Drawing General Undirected Graphs,” *Information Processing Letters*, vol. 31, no. 1, pp. 7–15, 1989.

[19] U. Brandes, P. Kenis, J. Raab, V. Schneider, and D. Wagner, “Explorations Into the Visualization of Policy Networks,” *Journal of Theoretical Politics*, vol. 11, no. 1, pp. 75–106, 1999.

[20] U. Brandes, P. Kenis, and D. Wagner, “Communicating Centrality in Policy Network Drawings,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 9, no. 2, pp. 241–253, 2003.

[21] J. Moody, D. McFarland, and S. Bender-deMoll, “Dynamic Network Visualization,” *American Journal of Sociology*, vol. 110, no. 4, pp. 1206–1241, 2005.

[22] J. Branke, “Dynamic Graph Drawing,” in *Drawing Graphs* (M. Kaufmann and D. Wagner, eds.), vol. 2025 of *Lecture Notes in Computer Science*, pp. 228–246, 2001.

[23] R. Shannon and A. J. Quigley, “Considerations in Dynamic Graph Drawing: A Survey,” 2007. Last accessed: 11/5/2011.

[24] I. Budge, D. Robertson, and D. Hearl, *Ideology, Strategy and Party Change: Spatial Analyses of Post-War Election Programmes in 19 Democracies*. Cambridge University Press, 1987.

[25] I. Budge, A. H.-D. Klingemann, J. Bara, E. Tannenbaum, R. Fording, D. Hearl, H. M. Kim, M. McDonald, and S. Mendes, *Mapping Policy Preferences: Parties, Estimates for Parties, Electors and Governments, 1945–1998.* Oxford University Press, 2001.

[26] L. Zhu, "Computational Political Science Literature Survey." (last accessed on 28/2/2011).

[27] B. Monroe and P. Schrodt, "Introduction to the Special Issue: The Statistical Analysis of Political Text," *Political Analysis*, vol. 16, no. 4, pp. 351–355, 2008.

[28] M. Laver, K. Benoit, and J. Garry, "Extracting Policy Positions from Political Texts Using Words as Data," *American Political Science Review*, vol. 97, no. 2, pp. 311–331, 2003.

[29] K. Benoit and M. Laver, "Estimating Irish Party Policy Positions Using Computer Wordscoring: The 2002 Elections - A Research Note," *Irish Political Studies*, vol. 18, no. 1, pp. 97–107, 2003.

[30] J. B. Slapin and S.-O. Proksch, "A Scaling Model for Estimating Time-Series Party Positions from Texts," *American Journal of Political Science*, vol. 52, no. 3, pp. 705–722, 2008.

[31] M. Thomas, B. Pang, and L. Lee, "Get Out the Vote: Determining Support or Opposition from Congressional Floor-debate Transcripts," in *Proc. of the 2006 Conf. on Empirical Methods in Natural Language Processing*, pp. 327–335, 2006.

[32] B. Chen, L. Zhu, D. Kifer, and D. Lee, "What Is an Opinion About? Exploring Political Standpoints Using Opinion Scoring Model," in *Proc. of the 24th AAAI Conf. on Artificial Intelligence*, pp. 1007–1012, 2010.

[33] D. Diermeier, J.-F. Godbout, B. Yu, and S. Kaufmann, "Language and Ideology in Congress," *Midwest Political Science Association*, 2007.

[34] B. Monroe, M. Colaresi, and K. Quinn, "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict," *Political Analysis*, vol. 16, no. 4, pp. 372–403, 2008.

[35] W. Gryc and K. Moilanen, "Leveraging Textual Sentiment Analysis with Social Network Modelling: Sentiment Analysis of Political Blogs in the 2008 U.S Presidential Election," in *Proc. of the "From Text to Political Positions" Workshop*, 2010.

[36] D. Gaffney, "#iranElection: Quantifying Online Activism," in *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*, 2010.

[37] J. Kelly and B. Etling, "Mapping Iran's Online Public: Politics and Culture in the Persian Blogosphere," *Berkman Center Research Publication*, no. 2008-1, 2008.

[38] H. Kautz, B. Selman, and M. Shah, "The Hidden Web," *AI Magazine*, vol. 18, no. 2, pp. 27–36, 1997.

[39] P. Mika, "Flink: Semantic Web Technology for the Extraction and Analysis of Social Networks," *Journal of Web Semantics*, vol. 3, no. 2, pp. 211–223, 2005.

[40] P. Mika, "Ontlogies Are Us: A Unified Model of Social Networks and Semantics," *Journal of Web Semantics*, vol. 5, no. 1, pp. 5–15, 2007.

[41] Y. Matsuo, J. Mori, and M. Hamasaki, "POLYPHONET: An Advanced Social Network Extraction System from the Web," in *Proc. of the 15th International World Wide Web Conf.*, pp. 397–406, 2006.

[42] Y. Jin, Y. Matsuo, and M. Ishizuka, "Extracting Social Networks among Various Entities on the Web," in *Proc. of the European Conf. on The Semantic Web: Research and Applications*, pp. 251–266, 2007.

[43] J. Mori, T. Tsujishita, Y. Matsuo, and M. Ishizuka, "Extracting Relations in Social Networks from the Web Using Similarity Between Collective Contexts," in *Proc. of the 5th International Semantic Web Conference*, pp. 487–500, 2006.

[44] J. Chang, J. Boyd-Graber, and D. M. Blei, "Connections between the Lines: Augmenting Social Networks with Text," in *Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 169–178, 2009.

[45] A. Culotta, R. Bekkerman, and A. Mccallum, "Extracting Social Networks and Contact Information from Email and the Web," in *Proc. of the 1th Conference on Email and Anti-Spam*, 2004.

[46] A. Gruzd and C. Haythornthwaite, "Automated Discovery and Analysis of Social Networks from Threaded Discussions," in *Proc. of the International Network of Social Network Analysis*, 2008.

[47] R. Xiang, J. Neville, and M. Rogati, "Modeling Relationship Strength in Online Social Networks," in *Proc. of the 19th International World Wide Web Conference*, pp. 981–990, 2010.

[48] P. Nasirifard, V. Peristeras, C. Hayes, and S. Decker, "Extracting and Utilizing Social Networks from Log Files of Shared Workspaces," in *Proc. of the 10th IFIP Working Conference on Virtual Enterprises*, pp. 643–650, 2009.

[49] B. Pouliquen, R. Steinberg, and J. Belyaeva, “Multilingual Multi-document Continuously-updated Social Networks,” in *Proc. of the International Conference in Recent Advances in Natural Language Processing*, pp. 25–32, 2007.

[50] F. Mesquita, Y. Merhav, and D. Barbosa, “Extracting Information Networks from the Blogosphere: State-of-the-Art and Challenges,” in *Proc. of the 4th International Conf. on Weblogs and Social Media, Data Challenge Workshop*, 2010.

[51] D. K. Elson, N. Dames, and K. R. McKeown, “Extracting Social Networks from Literary Fiction,” in *Proc. of the 48th Annual Meeting of the Association of Computational Linguistics*, pp. 138–147, 2010.

[52] R. Rada, H. Mili, E. Bicknell, and M. Blettner, “Development and Application of a Metric on Semantic Nets,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. 19, no. 1, pp. 17–30, 1989.

[53] Y. Li, Z. Bandar, and D. Mclean, “An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 4, pp. 871–882, 2003.

[54] Xiao-Ying Liu, Yi-Ming Zhou, and Ruo-Shi Zheng, “Measuring Semantic Similarity in Wordnet,” in *Proc. of the sixth International Conference on Machine Learning and Cybernetics*, pp. 3431–3435, 2007.

[55] D. Bollegala, Y. Matsuo, and M. Ishizuka, “Measuring Semantic Similarity between Words Using Web Search Engines,” in *Proc. of the 16th International World Wide Web Conference*, pp. 757–766, 2007.

[56] E. Iosif and A. Potamianos, “Unsupervised Semantic Similarity Computation between Terms Using Web Documents,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 11, pp. 1637–1647, 2010.

[57] P. D. Turney, “Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL,” in *Proc. of ECML-2001*, pp. 491–502, 2001.

[58] T. K. Landauer and S. T. Dumais, “A solution to Plato’s problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge,” *Psychological Review*, vol. 104, no. 2, pp. 211–240, 1997.

[59] P. Vitányi, “Universal Similarity,” in *Proc. of Information Theory Workshop Coding and Complexity*, pp. 238–243, 2005.

[60] P. D. Turney and P. Pantel, “From Frequency to Meaning: Vector Space Models of Semantics,” *Journal of Artificial Intelligence Research*, vol. 37, pp. 141–188, 2010.

[61] E. Iosif and A. Potamianos, "Unsupervised Semantic Similarity Computation using Web Search Engines," in *Proc. of the IEEE/WIC/ACM Int'l Conf. on Web Intelligence*, pp. 381–387, 2007.

[62] Z. Lin, M. R. Lyu, and I. King, "PageSim: A Novel Link-based Measure of Web Page Similarity," in *Proceedings of the 15th International Conference on World Wide Web*, pp. 1019–1020, 2006.

[63] P. Calado, M. Cristo, E. Moura, N. Ziviani, B. Ribeiro-Neto, and M. A. Gonçalves, "Combining Link-Based and Content-Based Methods for Web Document Classification," in *Proc. of the 12th International Conf. on Information and Knowledge Management*, pp. 394–401, 2003.

[64] J. Gracia, R. Trillo, M. Espinoza, and E. Mena, "Querying the Web: A Multiontology Disambiguation Method," in *Proc. of the 6th International Conference on Web Engineering*, pp. 241–248, 2006.

[65] K. W. Church and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics*, vol. 16, no. 1, pp. 22–29, 1990.

[66] R. Cilibrasi and P. Vitányi, "The Google Similarity Distance," *IEEE Trans. Knowledge and Data Eng.*, vol. 19, no. 3, pp. 370–383, 2007.

[67] S. Chakrabarti, *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan-Kauffman, 2002.

[68] D. Rempe and S. Fink, "Visualizing Networks in Time Using Dynamic Network to Test Theories of European Integration."

[69] S. L. T. McGregor, "Modeling the Evolution of a Policy Network Using Network Analysis," *Family and Consumer Sciences Research Journal*, vol. 32, no. 4, pp. 382–407, 2004.

[70] N. Rees, B. Quinn, and B. Connaughton, "Ireland's Pragmatic Adaptation to Regionalization: the Mid-West Region," *Regional and Federal Studies*, vol. 14, no. 3, pp. 379–404, 2004.

[71] P. Getimis and L. Demetropoulou, "Europeanization towards New Forms of Regional Governance in Greece," *Regional Studies Association International Conference*, 2003.

[72] F. Lazarinis, "Engineering and Utilizing a Stopword List in Greek Web Retrieval," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 11, pp. 1645–1652, 2007.

[73] M. E. J. Newman, "Analysis of Weighted Networks," *Physical Review E*, vol. 70, no. 5, p. 056131, 2004.