

Automated Construction and Analysis of Political Networks via open government and media sources.

MASTERS THESIS

DIEGO GARCIA-OLANO, Universitat Politècnica de Catalunya

Advisors: Marta Arias and Josep Lluís Larriba Pey

A joint research project with the DAMA and LARCA groups at
Universitat Politècnica de Catalunya (UPC) – Barcelona Tech
Department of Informatics (FIB)

Submitted in partial fulfillment of the requirements

for the degree of Master in Innovation and Research in Informatics

Specialization: Data Mining and Business Intelligence

June 30, 2015



DAMA-UPC. DATA MANAGEMENT
UNIVERSITAT POLITÈCNICA DE CATALUNYA

ACKNOWLEDGEMENTS

I would like to thank Marta Arias for being so gracious with her time and providing me with suggestions, editing, feedback and support through out this process. I would constantly make jokes about how she should be awarded for the number of Masters students she was advising concurrently in addition to myself, but I really wasn't joking. I would like to thank Josep Lluís "Larri" Larriba Pey for allowing me to work on this project, for inspiring me to work hard on it's initial phase for a Knight Foundation Media grant and for his support particularly with respect to my mother who was diagnosed with lung cancer near the beginning of this work, and who just yesterday, after months of treatment, was told she is in the clear! I would like to thank everyone from the DAMA and LARCA groups, but particularly Francisco Rodriguez Drumond with whom I brainstormed through out the process and discussed ideas. Additionally, I would like to thank Lluís Belanche for his excellent Machine Learning course, his in-class jokes, love of Prog Rock and for letting me help on his research involving Micro-bacterial source tracking. I would like to thank my dear friends Flora Lichtman, and Meghan Fergusson for helping make a video used for the grant application that involved a talking cactus wearing a Starburst candy wrapper as a bandana and his female companion, a Texan tumbleweed with boots on who sounds remarkably like my partner Cecilia. I would like to thank Chris Valdez for many reasons; for the initial design work of the Who You Elect tool, and for all that he has done for my mother during the past two years. I would like to thank Lou and Berta of Babelia, my favorite café/bookstore in Barcelona where a good deal of this work took place, for always smiling and being good sports when I tried to explain to them what I was doing. There are many others I would like to thank, particularly some of the other fantastic professors I've had the fortune of studying under at the UPC, but I will leave that for another time.

This thesis is dedicated to my mother Doctor Lilita Olano for everything, always.

ABSTRACT

In this work we present a tool that generates real world political networks from user provided lists of politicians and news sites. The tool downloads articles in which the politicians appear amongst the different news sites, processes them, enriches them with data obtained from various open sources and then generates various network visualizations, tools and maps that allow the user to explore and better understand those politicians and their surrounding environments. To demonstrate the capabilities of the tool for use in studying political and media landscapes, we construct a comprehensive list of current Texas politicians, select six news sites that convey a spectrum of representative political viewpoints publishing articles on the state, and examine the results produced by running the system with them as input. We propose a "Combined" co-occurrence distance metric to better assess the strength of the relationship between two actors in a graph and additionally provide automated summarization tools that utilize text-mining techniques to extract topics and issues characterizing the individual politicians. A similar topic modeling technique is also proposed as a novel way of labeling communities that exist within a politician's "extended" network. Finally we present media centric results of our case study that show who the different news sources publish articles about both from a geographic and individual perspective.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	1
ABSTRACT	1
1. INTRODUCTION	3
1.1. MOTIVATION	3
1.2. PROBLEM DESCRIPTION AND BACKGROUND	3
1.3. OVERVIEW OF SYSTEM: “WHO YOU ELECT”	4
1.4. DESCRIPTION OF CASE STUDY: TEXAS POLITICS	4
2. RELATED WORKS	5
3. AUTOMATED CONSTRUCTION OF NETWORKS	7
3.1 ADDING POLITICIANS FROM OPEN GOVERNMENT SOURCES	7
3.2 ADDING NEWS SOURCES	8
3.3 SET UP FOR DATA ACQUISITION BY TEMPLATE MODIFICATION	8
3.4 RUNNING AND STORING THE WEB SEARCH RESULTS FOR EACH ACTIVE ENTITY	9
3.5 PROCESSING ARTICLE RESULTS PER ACTIVE ENTITY	10
3.6 GENERATE INDIVIDUAL STAR AND EXTENDED GRAPHS FOR EACH INDIVIDUAL	12
3.7 ON PREPROCESSING THE CANDIDATE LIST OF POLITICIANS	12
4. OVERVIEW OF WHO YOU ELECT VISUALIZATION TOOLS	13
4.1 TABLE OF CONTENTS VIEW	13
4.2 MAPS OF TEXAS HOUSE, SENATE AND FEDERAL CONGRESSIONAL DISTRICTS	13
4.3 TEXAS COMMITTEES VIEW	15
4.4 INDIVIDUAL “STAR” NETWORK VIEW	15
4.4.1 CENTRAL ENTITY VIEW	16
4.4.2 SIDE BAR ENTITY ARTICLES TEXT VIEW	17
4.4.3 TOP ASSOCIATED DISTANCE METRICS	17
4.4.4 ARTICLE STATISTICS TEMPORAL VIEW	18
4.4.5 “COMBINED” METRIC & OTHER DISTANCE METRICS SUMMARY & COMPARISON VIEW	19
4.5 EXTENDED VIEW WITH COMMUNITY DETECTION	21
4.5.1 ENTITY-ENTITY GRAPH EXPANSION VIEW	22
4.5.2 COMMUNITY EXPANSION VIEW	25
4.5.3 ADDITIONAL CENTRALITY MEASURE TOOLS AND VISUALIZATION OPTIONS	25
4.5.4 COMMUNITY ANALYSIS VIEW	29
4.6 MEDIA ANALYSIS TOOLS	31
4.6.1 MEDIA ANALYSIS TABLE VIEWS	31
4.6.2 MEDIA ANALYSIS HEAT MAPS OF TEXAS HOUSE, SENATE AND FEDERAL DISTRICTS	33
5. ADDITIONAL ANALYSES	36
5.1 AUTOMATED SUMMARIZATION OF POLITICIANS	36
5.2 AUTOMATED SUMMARIZATION OF COMMUNITIES	38
5.3 MEDIA CENTRIC RESULTS OF CASE STUDY	39
6. CONCLUSIONS & FUTURE WORK	41
7. REFERENCES	43
8. APPENDICES	44

1. INTRODUCTION

1.1 Motivation

We live in an age of over-information, where we must constantly filter information, process it and make educated decisions be it at home, at work or when we vote. However in regards to voting, we are inundated by the media with news stories on a national level, which should in theory lead to a better informed populace, but more often than not, we only consume media outlets which reaffirm our political beliefs and thus entrench us in them, creating straw men of differing views and increasing polarization of voters. This situation is difficult, but at least manageable when a presidential election occurs, as we need only sort through the information related to the histories, views, and promises of a small field of candidates to make our decision.

The situation is however much worse on the local or state level where the opposite occurs. There are vastly more important decisions to be made, and we do not receive enough information regarding all candidates and election races in addition to the aforementioned issue of media bias, both self imposed through the choices we make as media consumers in addition to the biases of the media sources themselves. This overwhelming flood of information results in voters making uninformed or partially incomplete decisions at best or at worst sees them abstaining completely from the process. For instance, the last election for Governor of Texas in 2014 saw the Republican candidate Greg Abbott beat the Democrat one Wendy Davis by a 20% margin, 2.8 million votes versus 1.8 million votes [13]. However one must keep in mind that Texas has approximately 16.7 million eligible voters of whom 4.8 million voted thus giving the contest a dismal 27.5% of eligible voter participation. That number bumps up only to 32.5% if one uses registered voters, but that is still low. Another way of looking at it is to observe that the most powerful executive political position of Texas, a position that directly affects the daily lives and welfare of 27 million Texans was selected by 16% of its possible eligible voters.

1.2 Problem description and background

There exists a vast literature on the use of network analysis to study important social and political phenomena [2,3,4,5,23]. This work was in fact inspired in part by a presentation of one such paper describing a novel use of natural language processing and network analysis techniques to describe the network of drug cartels in Mexico[1]. In it, the authors perform text mining, partially by hand and partially automated, on a single book about the subject “Los Señores del Narco” by Anabel Hernandez and then derive a visual network from the actors and links discovered therein as seen in Figure 1.1.

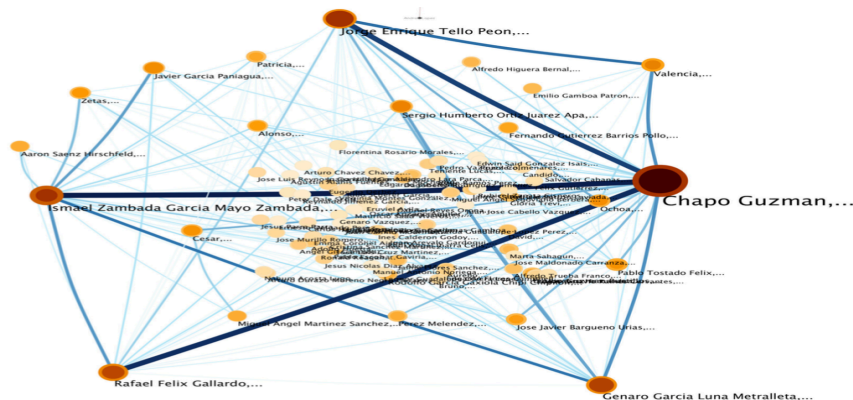


Fig. 1.1 Drug Cartels Network from “A literature-based approach to a narco-network”[1]

The larger nodes in the figure are the heads of cartels while the links connecting nodes imply a relationship between them; the thicker the link, the more important the relationship. In network analysis, nodes are sometimes referred to as vertices, links are sometimes referred to as edges, and

networks are referred to as graphs. They are used interchangeably and mean the exact same thing. The network derived from the book found 1037 nodes and 6405 links between them with Figure 1.1 being a filtered view highlighting the most important ones. The drug cartel network that was derived from the book was impressive for its visual synthesis of the content in the book. It in no way replaces the book by any stretch of the imagination, but it does provide a simple and concise high-level view of the actors and entities involved which is highly useful to someone new, but interested in the area.

This general idea of combining text mining and network science to both summarize content while allowing for the discovery of interesting relationships is a solid foundation on which to build upon that can handle the enormous, but under utilized, amount of information available in online news. Such processing of largely unstructured text combined with simple, flexible and powerful tools to explore and understand it in a targeted or broad way would be of great use for voters, journalists and researchers alike. The end goal of the work begun with this thesis is to provide a tool that can be both a hosted end solution for users to educate themselves and a generalized framework that may be built upon and deployed by interested parties for entirely new purposes of study. A solution that does not require a data center or anything else that may preclude adoption for financial or code proprietary reasons is also stressed.

1.3 Overview of System

In an effort to provide more insight into primarily local and state wide contests, but also including federal elections pertaining to a specific state, we decided to build a system “Who You Elect”¹ that could take one or many candidate names as input, along with a list of online news sources, and then retrieve all the articles pertaining to the candidates from them. Conversely the input could be a list of politicians or any group of people a researcher wishes to analyze. The system then using natural language processing, information retrieval and network analysis techniques automatically generates the network of all the politicians, organizations, businesses and locations associated with each candidate inputted. The system makes it easy to add new sources to pull content from and additionally provides two types of visualizations:

- An individual close up “star” view that allows a user to view the entities (politicians, businesses, etc.) most associated with a candidate along with the articles and textual context in which they co-occurred.
- An “extended” world view that is more of a global view of the network formed from articles pertaining to a candidate which in addition to showing links between himself and associated entities, also shows the links between those entities themselves and thus allows for the detection of communities and other traditional network analysis measures.

1.4 Description of Case Study: Texas Politics

For illustrative purposes, we decided to test our system on the political landscape of Texas in 2015. On a state level, the Texas Congress is composed of the Texas House of Representatives and Texas Senate. As of 2015, the state’s congress, or legislature by which it is also referred, is composed of 181 members, 31 senators and 150 representatives respectively. On the federal level within the United States Congress, Texas has 38 total members including 2 senators and 36 representatives. Additionally Texas has 27 other State Level Elected Officials including Governor, Lieutenant Governor, Attorney General, Comptroller of Public Accounts, Commissioner of General Land Office, Commissioner of Agriculture, 3 Railroad Commissioners, 9 Texas Supreme Court Justices, and 9 Court of Criminal Appeals Judges. Thus, in total we have 246 Texas politicians composed of 74 Democrats and 172 Republicans that we will be observing.

¹ Who You Elect tool using Texas specific case data at <http://www.whoyouelect.com/texas>

The organization of this document is as follows. Chapter 2 will cover related works and additional background. We will begin by describing the types of graphs the system constructs, along with rational for decisions made, and then look at other works that have focused on deriving networks from unstructured texts. Chapter 3 will describe the back end of the system, namely the methods by which we automatically gather, store and process the articles for the politicians we are studying. Chapter 4 focuses on presenting and explaining the frontend web tools developed and currently available at the WhoYouElect.com; particular attention is paid to demonstrating how interesting relationships may be discovered and verified by using the different tools available. Chapter 5 provides various media-centric statistics on the articles returned from the different new sources used in the case study and provides tools and methods for summarizing content found for the politicians studied via text analysis. Finally Chapter 6 contains a summary of conclusions and future work.

2. RELATED WORKS

The types of graphs we will be constructing are undirected heterogeneous networks with weighted edges in political contexts. Heterogeneous in this context just refers to the existence of different node types that will compose the graph. For instance, nodes in our system will be labeled as people, organizations, politicians, locations, bills, or miscellaneous. For an introduction and overview of the state of the art of heterogeneous networks and mining techniques see the following [6,7]. The graphs could be considered to be Social Networks involving political and nonpolitical actors or “noisy” Political Networks due to the inclusion of nodes and relationships not involving politicians, though in the end the distinction is largely semantic and unimportant.

Additionally, for the moment we are only considering one relationship type, i.e. at most one edge between each pair of nodes, based on various distance metrics and as such do not find ourselves in the multiplex context which is more adapt for studying complex networks. For an extensive examination of the field, uses, and visualization tools see [15,24]. Adapting the system to include more than one edge type is largely based on being able to create multiple edges between pairs of nodes, by perhaps leveraging linked datasets, and then characterize those relationship types effectively. The former task is largely an information retrieval one usually involving enriching datasets via a knowledge base such as Wikipedia, but usually suffers for queries lacking their own Wiki entries or incomplete data. Two recent works in this area seem promising; one exploring boosting specifically in the case of missing or incomplete linked data involves mining a knowledge base using additional textual context, i.e. “evidences”, for named entity disambiguation [20] and the other constructs a probabilistic model that captures the “popularity” of entities and the distribution of multi-type objects appearing in the textual context of an entity using meta-path constrained random walks over networks to learn weights associated with entity linking [21]. The later task of characterizing relationships between nodes is usually handled by deriving a topic model from the corpus of text available, all of the news articles gathered for a given politician in our case, and assigning the most probable “topic” to an edge based on learned Bayesian probability models [16,17] and the text contexts of the two nodes. A good overview of the idea and the use of Latent Dirichlet Allocation (LDA) or alternatively Latent Semantic Indexing (LSI) to produce topic models is found in [11]. A topic in this context simply refers to a collection or distribution of words that characterizes a predefined or latent concept. For instance, an example topic “Actor” could be defined by the set of words “Hollywood, movie, film, theatre, etc.” Chang’s paper [16] is particularly interesting as it both learns topics without supervision and assigns them to edges, by constructing “entity topics” and “pair topics” and then considering the two most probably “entity topics” assigned to a pair of nodes along with their most probably “pair topic” in order to final construct a description of their relationship, itself another set of disjoint words. Other approaches leverage the use of “phrases” [17] as opposed to single words to help improve performance, but the basis is the same. The topic model based approach has the advantages that it reduces the dimensionality of the search space since you are now working

with topic objects as opposed to articles, and the number of topics is intended to be much less than the number of documents, and additionally, it allows for the formation of relationships between entities even if they don't co-occur in the same document. The disadvantages are that labeling and verifying the resulting topics in itself is a largely manual process [16], the topics themselves can be very noisy and difficult to interpret, and precisely that it moves away from the article centric approach and in doing so creates groupings, similarly to community detection and clustering methods, which themselves impose a structure which may be warranted or not. Because the focus of this initial phase was to construct a working proof of concept and the inclusion of edge labeling via topic modeling would be a nice, but unnecessary step, it has been left for future work. We do some separate textual analysis based on topic modeling however to summarize politicians using their texts, but use it as a ground truth and complementary analysis tool. The use of topic modeling to find potentially interesting relations especially through the use of Congressional bills and their debates as separate topics could be particularly useful. The following recent work describes this idea in the context of Catalan political networks [26]. Similarly the role discovery technique proposed in [19] which is a sort of complement to community detection and uses the structural behavior of nodes to assign them "roles" to aid in the identification of nodes of interest was left for future work.

There exist a good deal of prior work that has focused on deriving information networks from unstructured news and other web texts [8,9,10,11]. Two introductions and overviews on the process of deriving networks from text may be found [22,25]. Of the prior works cited some rely on hand crafted networks "extracted through a time and effort consuming manual process based on interviews and questionnaires collected"[9], while others rely on organizations such as the European Media Monitor [8] providing them with access to an article lookup system that while impressive in the breadth of sources available, constrains them to only sources from that list. Additionally and specifically in reference to the European Media Monitor and another considered media aggregation service MIT's Mediacloud, in the cases where we noticed an overlap in available sources between their listings and the ones we are considering for our case study (the Houston Chronicle for instance), the search results from the original news source internal search engine always returned more results for specific entity queries than either the EMM or Mediacloud service which points to an additional quality assurance weakness¹. Other works avoid the actual aspect of retrieving content from news sites by using a point in time snapshot of curated news corpora released through the Linguistic Data Consortium or the New York Times[10, 11]. Similarly, [9] leverages paid-for search engine results, only grabbing the first twenty results for each query and then only utilizing the snippet of text present in Yahoo's search results page as opposed to all the content within the actual article itself. In the end we were unable to find any work that leveraged the publically available search engines present in most news websites. By utilizing this mechanism and allowing the flexibility to pull content from any news site which fulfills that requirement, our system allows site administrators to create a context in which to search and in doing so curate the content and thus satisfy the needs of different users.

Information retrieval aspects of the texts used in the prior works aside, the prior works all use NLP to extract entities and then leverage either similarity metrics based on some combination of entity co-occurrence, textual contexts of and shared between entities, and the correlation of entities and hyper links found in documents [8,9] or topical modeling [10,11] to infer relationships of interest. The textual context presented in [9] is limited in that it does not consider entire documents, rather only text snippets from search engine results, but is robust in its evaluation of metrics quantifying the use of co-occurrences metrics for labeling relationships as positive or negative. Based on the number of results, they produce four metrics of similarity: the Jaccard Coefficient, the Dice Coefficient, Mutual Information (MI) and the Google-based semantic relatedness [29] and evaluated them against a small hand-crafted Policy Network. Although interesting this approach does not scale as an evaluation approach. The works of [27,28] like [8] uses the EMM for content, but are novel in that as opposed to using similarity distance metrics or topic modeling, they also use natural language processing with

¹ see Appendix A0 for a sample comparison of a query comparing EMM vs HC result sets.

an initial seed of hand crafted syntactic templates to learn syntactic patterns which paraphrase certain predefined relations from articles and uses them to label relationships between entities. The work in [11] is of particular interest and leverages the CATHY framework (Construct A Topical HierarchY), which is a recursive clustering and ranking approach for topical hierarchy generation similar to the idea presented in [16] except with the hierarchical component which allows for relating of topics to their parents “more general” or children “more specific”. As mentioned before however, topic modelling approaches for relationship labeling and as a complementary view for better inclusion of bills in Congress are left for future work.

3. AUTOMATED CONSTRUCTION OF GRAPHS

In our case study, we constructed the graphs for 247 active Texan politicians. This political environment was chosen to illustrate the use of the system, but could just as easily have been a list of politicians for any city, state, or country. The system components which deal with the construction of graphs use only Python, some associated open libraries, MongoDB, and occasional Bash scripts, and as such is very light from a technical requirements view point. The general outline of the process for constructing graphs is shown in Figure 3.1

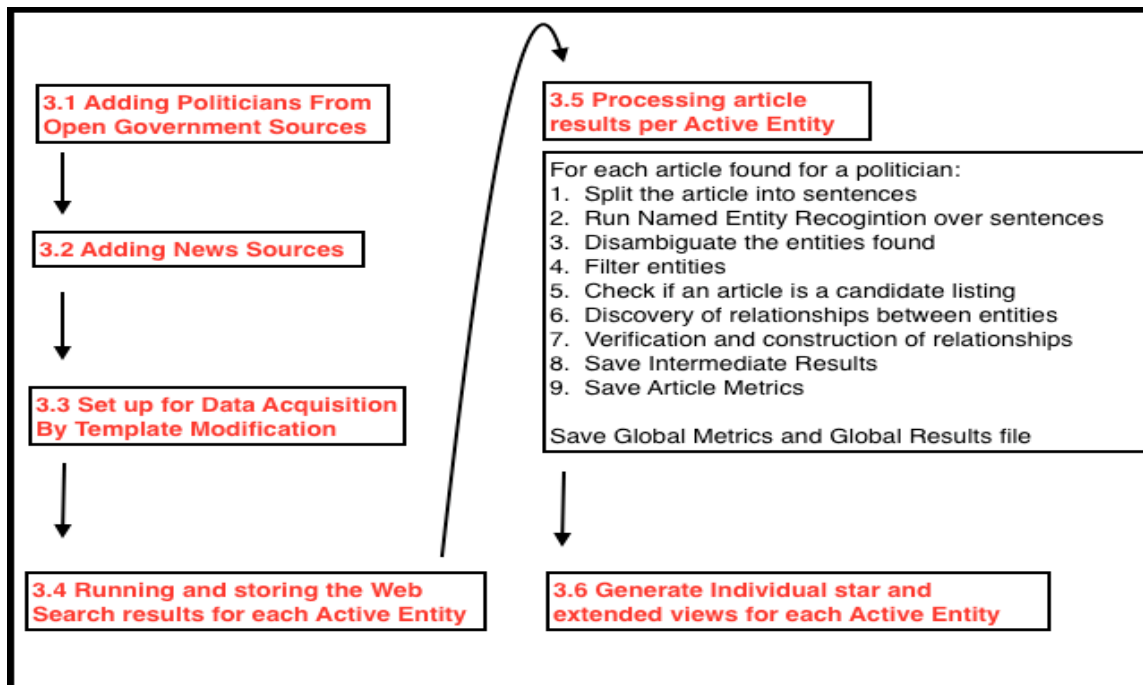


Fig 3.1 General Overview of Graph Construction Process

3.1 Adding Politicians From Open Government Sources

We first needed to generate the Texas specific list of congressmen. We leveraged the API for the Sunlight Foundation’s website Openstates.org to obtain a list of both active and inactive members of Texas congress returned as JSON¹ and saved them locally. Appendix A contains a sample record the OpenStates API returns for a Texas representative and also a sample record for a Federal Representative returned from the GovTrack.us API. The inactive members API returns a set of 112 prior State representatives and though they will not be analyzed and no articles specific to them will

¹ <http://openstates.org/api/v1/legislators/?state=tx&active=true>
<http://openstates.org/api/v1/legislators/?state=tx&active=false>

be retrieved, we still will include them as possible domain knowledge to leverage during disambiguation of entity types during the processing of articles stage. Throughout this work “entities” is an umbrella term encompassing any actors of interest including politicians, organizations, bills, etc.

Next, we needed the list of state wide elected officials who are not representatives (Governor, Lieutenant Governor, etc.) and made use of the Secretary of State of Texas website¹ and script² to generate a JSON file we then saved locally.

Finally, in order to get federal representatives and senators for the state of Texas, we leveraged the API at www.govtrack.us³ and saved the JSON file locally.

At this point we have four files referring to the active and inactive state representatives, state level elected officials, and federal level elected officials. We load them all and standardize formatting of fields and save them into our "entities" mongo database via a script⁴ that also creates a CSV file of only active entities⁵.

3.2 Adding News Sources:

Now that the entities have been added to our database, we select all the news sources to use for our case study. A subset of Texas newspapers with the highest circulation, the Dallas Morning News⁶, Houston Chronicle⁷, and the Austin American Statesman⁸ representing the spectrum of conservative, centrist and progressive areas within Texas were selected along with two sites, the Texas Observer⁹ and Texas Tribune¹⁰ which focus on Texas politics and issues. In addition, the New York Times¹¹ was selected to provide an outside context. This selection of sources was in the end entirely subjective, but made with the intent to present a reasonable mix of representative media consumed within the state of Texas.

3.3. Set up for Data Acquisition By Template Modification:

For each desired source we need to gather articles pertaining to our politicians list.

We’ve created two template web scrapers, one based on the Python package BeautifulSoup (BS) and another based on the Web.Selenium Python package which utilizes both Chrome and PhantomJS (PS) drivers. Each template version is a folder containing two files:

- 1. a file, based on BS or PS, which calls the internal search URL for a given news source, and then saves a JSON file of the article URLs, titles, and dates retrieved.
- 2. a file invoked after the first that then grabs the actual article content for each URL obtained from step one and saves it into a separate JSON file, one per article which contains the title, URL, article text, the news-source itself, time, and an identifier.

The reason two types of templates are warranted is based on the peculiarities of web scraping and limitations of each. If a website such as the Houston Chronicle practices good web standards and renders site content without the explicit use of JavaScript, since doing so breaks accessibility requirements for non JavaScript browsers including braille reader systems for seeing impaired users, we simply use the Beautiful soup version template to construct our scraper. Unfortunately not all websites render content as such so we take advantage of headless browser technologies via

¹ <http://www.sos.state.tx.us/elections/voter/elected.shtml>

² http://github.com/diegoolano/thesis/whoyouelect/js/htmltable_to_json.js

³ <https://www.govtrack.us/api/v2/role?current=true&state=TX>

⁴ http://github.com/diegoolano/thesis/generate_network/load_entities_from_file_into_db.py

⁵ http://github.com/diegoolano/thesis/generate_network/active-entities-list.csv

⁶ <http://dallasnews.com>

⁷ <http://www.chron.com>

⁸ <http://www.statesman.com>

⁹ <http://www.texasobserver.org>

¹⁰ <http://www.texastribune.org>

¹¹ <http://www.nytimes.com>

web.selenium to mimic the behavior of a user visiting a particular website to scrape the information we need. This approach allows for clicking around within a page to remove popup windows or follow AJAX pagination (i.e., "next" page results which are rendered in the same current page without going to a new URL and loading it), but has the down side of being slower than the BeautifulSoup approach, and is a little more work to program due to the nature of asynchronous page loads and the need for mechanisms which wait for certain DOM elements to load before proceeding. A user can always use the 2nd Selenium template if they want, but doing so may unnecessarily add more time to searching and storing of content so we provide both mechanisms.

Once a template version is selected, a user makes a copy of that folder, renames it to any name they wish to represent the new source, for instance "nytimes" for the New York Times, and then edits the two files within that folder so that they work with the page structure of the new source. See appendix B for more information on template editing and the structure of the file system where results are stored. This procedure is straightforward for a web developer with experience to setup and test however automating this step is planned for future work. Simplifying this procedure requires considerable effort via inferring structure of pages probabilistically and could lead to potential loss of information thus at this point its a manual procedure which has been simplified. Of the six sources we added, half were based on the beautifulsoup approach (Austin American Statesmen, Houston Chronicle, and Texas Tribune) and half on the selenium one (Dallas Morning News, NY Times, and Texas Observer).

3.4 Running and storing the Web Search results for each Active Entity

With the sources setup, we run the script "start-big-process-of-websearches.py" that goes through each active politician contained within the CSV file created before in Section 3.1. and calls the script "do_websearch_for.py" on them one by one.

This script takes a candidate names as input and then calls the web scrapers created in Section 3.3 for each news source concurrently and then once all the articles have been retrieved and stored locally by their source's webscraper mechanism, the script then calls "add_json_files_for.py" passing along the candidate name as an input. It is of interest to note that the script "do_websearch_for.py" can also be called with a candidate name and a single source in the case that we want to obtain and save the articles for a single source as opposed to all of the sources, which comes in useful when a given site is down for maintenance for instance.

The script "add_json_files_for.py" then takes the article JSON results from the prior step, runs some light post processing on each of them to detect the language of the article text and insert the candidate name as a key to index for quick future lookups before saving the result in the collection "db.texnews.english" or "db.texnews.spanish" within MongoDB based on the language detected respectively. Currently the Who You Elect system is tailored for articles in English, but also works for Spanish and any languages supported by MITIE¹, an open source Named Entity Recognition engine written in python that when given a document as input, identifies substrings that contain possible named entities and tags them as "Organization", "Location", "Person" or "Miscellaneous".

In addition to these tag types, in step 3.5 during processing we look for and include two other tag types, "Politician" and "Bill". We alter "Person" tags to be the more specific "Politician" tag type if we find that entity to be preceded or followed by a politician position title such as "Senator" Bob or Bob, the Senator from District 8, etc. More likely however, Person entities will be labeled as Politicians if the entity name is found in our entities database of active and inactive politicians that are labeled as Politicians when initially entered into the database in step 3.1. Although it could be read as such, we do not mean to imply that Politicians are not People. The Bill entity type refers to

¹

legislation and we use simple naïve heuristics looking for the phrases SB, HR or HB during the processing stage as these refer to Senate Bill, House Resolution and House Bill respectively. This is a rather simplistic solution that could be greatly improved by leveraging other open APIs provided by OpenStates and Govtrack.us, but it also proves to find legislation associated with a politician.

The decision to go with MITIE came from seeing how well the system identified entities over a hand selected group of articles related to Texas politics and then compared how it favored against other open source NER alternatives such as solutions from Stanford, Illinois, nltk, and pangar. In the end, in terms of speed and accuracy combined MITIE performed the best on our data. The possibility of training a NER system with “entity labeled” political documents could improve the overall performance of the system, and is considered for possible future work.

3.5 Processing article results per Active Entity

The script “get_articles_for_person_then_find_and_save_relations.py” is called following the completion of “add_json_files.py” and is the heart of the processing step of the articles retrieved for a given entity. The script is passed in the name of the politician or candidate to process, and first queries the mongodb to gather all of the articles found for the person. It pre-filters out “sports” articles by way of looking for “/sports/” in the URL to avoid noise and uninteresting results, and then processes the remaining articles one by one in the following manner:

1. **Find the date** retrieved for an article and assert its validity and if necessary, change it to follow the format YYYY-MM-DD. In the rare case where it is not valid, make a note in debug logs and assign a dummy date of 2000-01-01.
2. **Split the article into sentences**, and verify that the article text is non-empty otherwise skip it. This step may seem frivolous, but in actuality is of critical importance because we are dealing with uncertain input and one of the key challenges to using real data is that at times search engines return dead links.
3. **Run Named Entity Recognition**¹ over the sentences to return the named entities and tags per sentence and verify that the candidate we are searching for appears. This step identifies the cases where a pay wall exists for a given article and the returned article text contains only the first few lines of the actual article that does not include the name of the candidate himself. By virtue of the article having been returned by candidate query in the first place, we can assume the person’s name does appear in the full article, especially if we leverage the meta-data for each politician that we gathered from the open government APIs in the first step. We however do not take this liberty and skip articles entirely if the candidates name does not appear.
4. **Disambiguate the entities** found. As the prior step found entities and tags by sentence, we must now consolidate the entities list and do a sort of heuristic co-reference resolution to infer that two entities found in actuality refer to the same entity. For instance, if we find a “Barack Obama” tagged “politician” entity in one sentence and in the next sentence we find a “Mr. Obama” entity, we remove the second entity and use the first in its place. This step is of particular importance and uses some heuristics to discover interesting data common to political texts in English such as political parties, positions and location relationships. For instance, if the entity “D-Houston” is found in a sentence, the system knows to look forward or backward within the sentence for a politician since in American political texts in that construct implies “Democrat from Houston”. Its important to note that this stage uses heuristics that are largely language specific and as such, would need to be adapted or just discarded for use with other languages. By the end of this step, we have a dictionary of all the unique “disambiguated” entities discovered in the article.

¹ The NER and Disambiguate Entities functionality resides predominantly in dama_ner.py

5. **Filter entities**¹ further. This step is optional and is essentially a look up table that could be used in future work by a user to add uninteresting entities discovered that don't need to be added to the overall graph for analysis. For instance, in our case the entities for "Texas Legislature" and "Congress" and other entities that provide little distinguishing capacity or that were erroneously recognized are further excluded.
6. **Check if an article is a candidate listing** or something similar in which case skip it. A simple ratio of unique entities divided by the number of sentences in the article was used, and if that ratio was 10 or greater, the article was skipped. This metric was developed during testing of the system when it was noticed that the occasional article would take very long compared to the normal use case, and on inspection, it would be an article that contained an unusually high number of disambiguated entities with respect to the number of sentences in the article. These articles tended to be something along the lines of a listing of awards given to high school students of a district where a House Representative was present at the ceremony. The system would discover a great deal of unique entities, one per student, and then spend a great deal of time constructing the relationship between them even though the information was mostly just noise that is of little interest to the analysis step. Even in the case where the listing was a list of candidates for offices or winners of statewide races, it would create a great deal of spurious relationships where they didn't exist. A future improvement could be to simply handle this latter case differently than others, by for instance only considering "same sentence" or "near" relations in these contexts.
7. **Discovery of relationships**² is the next step of processing. At this point, the system goes through each instance found of the politician being processed and for each of them, gathers all of the entities that occur on the same sentence as it, within 3 sentences of it, or farther than three sentences away. At the end of this step we have a multidimensional result array with the number of rows equal to the number of times the main politician was found in the article, and with five columns, one for "same sentence" entities, "near prior" entities, "near next" entities, "far prior" entities, and "far next" entities respectively. Each cell of the matrix is itself an array of entities. For instance if two instances of the main politician are found in the article, the second row and third column will refer to all the entities found within three sentences after the second occurrence.
8. **Verification and construction of relationships.** At this point we take the matrix from the prior step and then first construct the relationships for each instance of the main politician found with the other entities. See Appendix C for the JSON structure of a single relationship. All the relationships that have been created with respect to the main politician will be used later to construct the "star" individual view graph. Next, if the initial script "get_articles_for_person_then_find_and_save_relations.py" was called with the option "include_larger", the system additionally constructs all the relationships between all the other non-main politician entities in order to construct the "extended" graph visualization.
9. **Save Intermediate Results.** At the end of the prior step we have a results object (see appendix D) for the article that we need to save to our global results set for the politician. This process goes through the relationships formed in this article and for each sees if it exists in the global result set already and if so adds its information as a new "instance" for that existing edge. If it doesn't exist in the global result set, a new edge gets added. Additionally, there exist actually two functions to handle this procedure. It was discovered during development that while most articles take well under a second to process and save, there were some which took a large amount of time; "large" meaning anything over 5 seconds. It was observed that these articles were ones whose product

of candidate instances found x # of entities x # of sentences

¹ The optional filter entities, verification/construction of relationships and saving to intermediate object is done mostly in `verify_and_save_relations.py`.

² Discovery of relations functionality is predominantly in `entity_funcs.py`

was found to be usually greater than a certain threshold (8000 is the threshold we ended up using). So in addition to the original function that worked well in most cases, we developed one specific to these larger instances and optimized it to work well for them by pre-computing the hashes for lookups. This pre-computation was such that the function behaved slightly worse on the normal articles from before so at the beginning we choose which function to used based on the product above.

10. **Save Article Metrics.** Once an article has been processed and saved, or alternatively skipped, metrics for the article are saved internally.

Once all articles have been processed, we save the resulting article metrics to our global metrics file (see APPENDIX E) and then save our global result set and some other variables into a python “pickle” file, essentially an internal tar file for python that we compress to save space and place in “data/pickles/PoliticianName.pickle”. This later step was done largely as a time saving mechanism during development of the system but also allows for a developer to access the processed data immediately without having to reprocess articles.

3.6. Generate individual star and extended views for each individual

Once the articles have been processed and saved from the prior step, the script “generate_single_network.py” is then called with the politician’s name, and an input stating whether or not to generate the larger “extended” graph. This script constructs the individual star and extended views graph files that are then used by the frontend interactive visualizations that are built on D3.js. The script also updates the two configuration files, “config.json” and “configdesc.json” in whoyoelect/js/, that are used by the frontend system, and has an option which tells the system to look for and save images and short descriptions for each entity in the small graph by leveraging the Wikipedia API¹. This option is to enrich the individual “star” view with photos of politicians, organizations, etc. along with the first few sentences of their Wiki article. The option was turned off for all, but one of the active politicians for time consideration purposes though future work could automate this and additionally leverage the API to help identify and verify entity types. It is of interest to note that one may leverage the photo URLs returned by Wikipedia to disambiguate entities even further. For instance when searching for the photo corresponding to “U.S.” or “United States” on Wikipedia, the system maps both queries to the same page and returns the same image.

3.7 On preprocessing the candidate list of politicians

As an aside, during the development and testing of the process detailed in section 3.4, it was discovered that some candidates were receiving less article results than expected because their official name as obtained from the OpenStates.org API was not in actuality the name they usually went by and were referred to by in the media. Thus for some of the entities in question, we removed middle names, accent marks and in a few cases “Jr.” This decision was done by seeing how many results were returned for querying with three versus two names for people with more than two word names, querying with or without Jr. and querying with or without accent marks. Almost every time the second name, without Jr. and without accent marks returned more results. The exceptions in this case involved "Eddie Lucio Jr." since his son "Eddie Lucio III" is also a representative, and the cases of Jimmie Don Aycock, and Sheila Jackson Lee who go by their full names. Undoubtedly this will produce more noise, but the processing stage should filter it out. This process was done manually, but could easily be automated in a future iteration. The updates to the names were also applied to the affected entities in the database.

¹ http://en.wikipedia.org/w/api.php?action=query&titles=ENTITY_NAME&prop=pageimages&format=json&pithumbsize=200

4. Overview of Who You Elect Visualization Tools

4.1 Table of Contents View

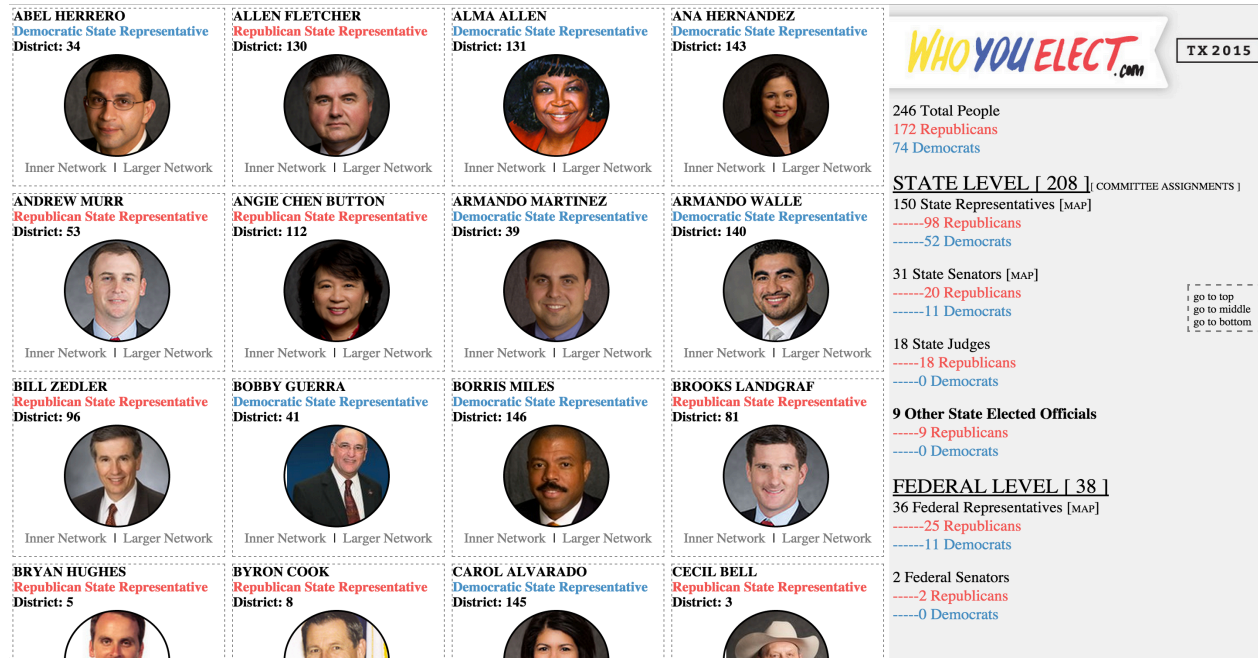


Fig. 4.1 Table of Contents View Filtered by State Representatives

After gathering and processing results for 246 politicians, it became readily apparent that a user would need a way to quickly browse and filter the type of politicians they were looking to study. The right hand side in gray shows statistics for all politicians and allows the user to click on any header to display only that type of politician in the left hand side. For instance Fig 4.1, shows the result when the link for “150 State Representatives” is selected. The right hand side also contains links to 3 interactive maps that show district information for the Texas House, Senate and US Congress, and a link to Committee Assignments for Texas Congress.

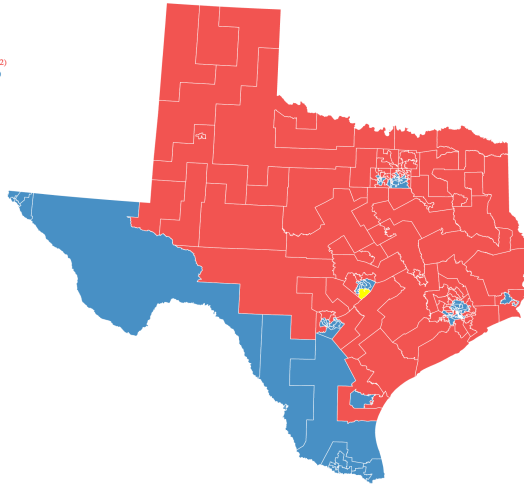
The left hand side contains the full name, party affiliation, position and the district represented if any for each politician studied. Each politician’s area also contains a link to their “Inner” or “star” network, and their “Larger” or “extended” network.

4.2 Maps Of Texas House, Senate And Federal Congressional Districts

Figures 4.2, 4.3 and 4.4 display the district boundaries for the Texas House, Senate and Federal House of Representatives. They are color coded by the political party of the Congressman representing the district; red for Republican districts and blue for Democrat ones. The three maps allow users to scroll over districts on the map or by hovering along the list of politicians along the left hand side to display the Representative or Senator for that district along with their neighboring Congressman. By overlaying the maps together we can establish a ground truth of which politicians we would expect to see associated solely by spatial proximity and a more thorough use of this idea is planned for future work. The maps are based on the district available from the Open States API and open geographic data from naturalearthdata.com. The Federal Map is based on a D3 example from Mike Bostock¹ along with the district data from GovTrack.us.

¹ <http://bl.ocks.org/mbostock/8814734>

Abel Herrera (District: 34)
 Allen Fletcher (District: 130)
 Alma Allen (District: 131)
 Ana Hernandez (District: 143)
 Andrew Marr (District: 53)
 Angie Chen Button (District: 112)
 Armando Martinez (District: 39)
 Armando Walle (District: 140)
 Bill Zedler (District: 90)
 Bobby Guerra (District: 41)
 Borris Miles (District: 146)
 Brooks Landgraf (District: 41)
 Bryan Hughes (District: 5)
 Byron Cook (District: 8)
 Carol Alvarado (District: 145)
 Cecil Bell (District: 3)
 Celia Israel (District: 50)
 Cesar Blanco (District: 76)
 Charles Anderson (District: 56)
 Charlie Geren (District: 99)
 Chris Paddie (District: 9)
 Chris Turner (District: 101)
 Cindy Burkett (District: 113)
 Craig Goldman (District: 97)
 Dada Pichan (District: 21)
 Dan Flynn (District: 2)
 Dan Huberty (District: 127)
 David Simpson (District: 7)
 Dawson Drake (District: 46)
 DeWayne Burns (District: 58)
 Debbie Riddle (District: 150)
 Dennis Bonnen (District: 25)
 Dennis Paul (District: 129)
 Diego Bernal (District: 123)
 Donna Howard (District: 48)
 Doug Miller (District: 73)
 Drew Darby (District: 72)
 Drew Springer (District: 68)
 Dustin Burrows (District: 43)
 Dwayne Bohae (District: 138)
 Ed Thompson (District: 29)
 Eddie Lacio II (District: 38)
 Eddie Rodriguez (District: 21)



WHO YOU ELECT .COM TX 2015

TEXAS STATE HOUSE DISTRICTS:
 98 Republican
 52 Democrats
 150 TOTAL

District 51

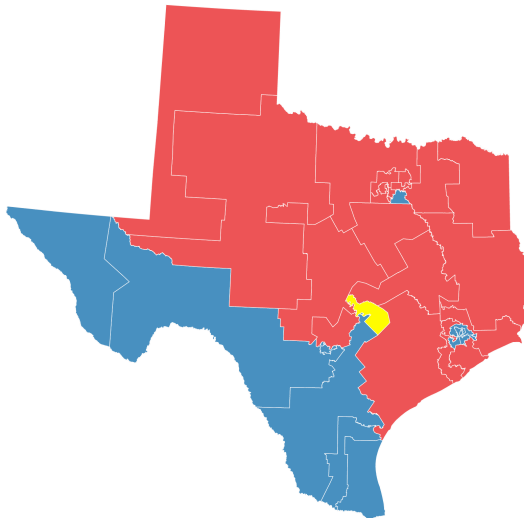
Democratic Representative Eddie Rodriguez

with neighbors

District: 17 Republican John Cyrrier	District: 45 Republican Jason Isaac
District: 46 Democratic Dawanna Duke	District: 47 Republican Paul Workman
District: 48 Democratic Donna Howard	District: 49 Democratic Elliott Naish
District: 50 Democratic Celia Israel	

Fig. 4.2 Interactive Map of Texas House Districts with District 51 selected

Bob Hall (District: 2)
 Brandon Creighton (District: 4)
 Brian Birdwell (District: 22)
 Carlos Ossati (District: 19)
 Charles Perry (District: 28)
 Charles Schwertner (District: 5)
 Craig Estes (District: 30)
 Don Huffines (District: 16)
 Donna Campbell (District: 25)
 Eddie Lacio Jr (District: 27)
 Jane Nelson (District: 12)
 Joan Huffman (District: 17)
 John Whitmore (District: 15)
 Jose Menendez (District: 26)
 Jose Rodriguez (District: 29)
 Juan Hinojosa (District: 20)
 Judith Zaffirni (District: 21)
 Kel Seliger (District: 31)
 Kelly Hancock (District: 9)
 Kevin Elife (District: 1)
 Kirk Watson (District: 14)
 Konni Burton (District: 10)
 Larry Taylor (District: 11)
 Lois Kolkhorst (District: 18)
 Paul Bonomaccari (District: 7)
 Robert Nichols (District: 3)
 Rodney Ellis (District: 13)
 Royce West (District: 23)
 Sylvia Garcia (District: 6)
 Troy Fraser (District: 24)
 Van Taylor (District: 8)



WHO YOU ELECT .COM TX 2015

TEXAS STATE SENATE DISTRICTS:
 20 Republican
 11 Democrats
 31 TOTAL

District 14

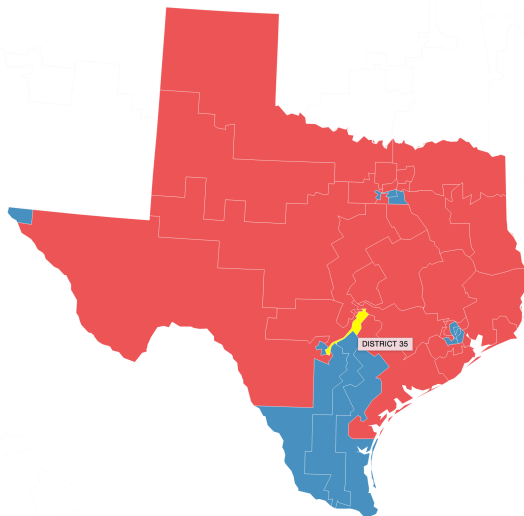
Democratic Representative Kirk Watson

with neighbors

District: 5 Republican Charles Schwertner	District: 18 Republican Lois Kolkhorst
District: 21 Democratic Judith Zaffirni	District: 24 Republican Troy Fraser
District: 25 Republican Donna Campbell	

Fig. 4.3 Interactive Map of Texas Senate Districts with District 14 selected

Al Green (District: 9)
 Beto Rousee (District: 16)
 Bill Flores (District: 17)
 Blake Farenthold (District: 27)
 Brian Babin (District: 36)
 Eddie Johnson (District: 30)
 Filemon Vela (District: 34)
 Gene Green (District: 29)
 Henry Cuellar (District: 28)
 John Krasakofsky (District: 5)
 Joaquin Castro (District: 20)
 Joe Barton (District: 6)
 John Carter (District: 31)
 John Culshaw (District: 7)
 John Ratcliffe (District: 4)
 Kay Granger (District: 12)
 Kenny Marchant (District: 24)
 Kevin Brady (District: 6)
 Lamar Smith (District: 21)
 Lloyd Doggett (District: 35)
 Louie Gohmert (District: 1)
 Mac Thornberry (District: 13)
 Marc Veasey (District: 33)
 Michael Burgess (District: 26)
 Michael McCaul (District: 10)
 Mike Conaway (District: 11)
 Pete Olson (District: 22)
 Pete Sessions (District: 32)
 Randy Neugebauer (District: 19)
 Randy Weber (District: 14)
 Roger Williams (District: 25)
 Ruben Hinojosa (District: 15)
 Sam Johnson (District: 3)
 Sheila Jackson Lee (District: 18)
 Ted Poe (District: 3)
 Will Hurd (District: 23)



WHO YOU ELECT .COM TX 2015

TEXAS FEDERAL CONGRESSIONAL DISTRICTS:
 25 Republican
 11 Democrats
 36 TOTAL

District 35

Democratic Representative Lloyd Doggett

with neighbors

District: 28 Democratic Henry Cuellar	District: 27 Republican Blake Farenthold
District: 15 Democratic Ruben Hinojosa	District: 23 Republican Will Hurd
District: 20 Democratic Joaquin Castro	District: 25 Republican Roger Williams
District: 21 Republican Lamar Smith	District: 10 Republican Michael McCaul

Fig. 4.4 Interactive Map of Texas Federal Congressional Districts with District 35 selected

On a side note the effect of gerrymandering, particularly at the Federal level, is quite apparent when the maps are view simultaneously; notice the lack of blue in the southwest region. Gerrymandering is “the practice of redrawing legislative boundaries so that the resultant political landscape features built-in electoral advantages for a specific constituency.”[32] Districts are generally redrawn every 5 or 10 ten years depending on the district in question, and the political party in control of that legislature largely gets to decide them thus making the process a remarkably important political affair. For background and interesting work on visualizing and resolving gerrymandering, quantifying its efficacy, and even a theory of optimal partisan gerrymandering see [30,31,32,33,34] .

4.3 Texas Committees View

The Committee View, seen in Figure 4.5, is based on two APIs available via OpenStates.org, one that gathers all the committee data available for a state¹ and another² that takes a committee id from the first and returns a list of members who are a part of it.

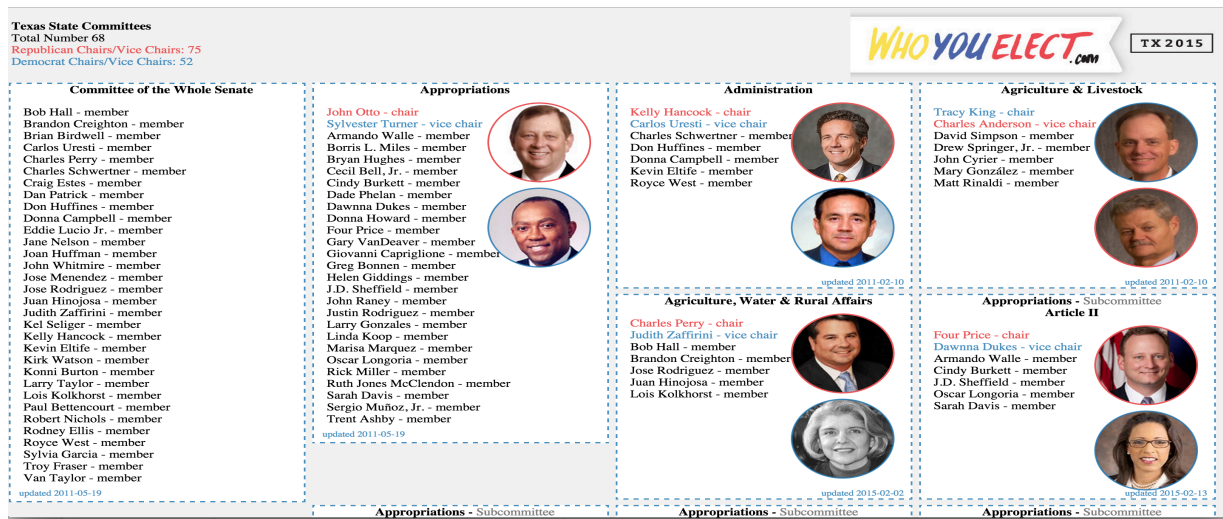


Fig. 4.5 Committee Assignments for the Texas Congress

A script was written to automate the downloading of all committee assignments³. Senate members are appointed to committees by the Lieutenant Governor, and the Speaker of the House’s duties include the appointment of chairships and committee membership. The current Speaker of the House is Republican Joe Straus, the Speaker Pro Tempore who fills in in case of absence of the Speaker is Republican Dennis Bonnen, and the current Lieutenant Governor is Republican Dan Patrick. These committee assignments, in the same way spatial proximity of representatives does, provide us with certain associations we expect to find. As an aside, although it may seem that some of the contents are out of date since the “updated date” appears old for certain committees in the visualization, it has been verified that these are the current assignments for the 84th Legislature⁴ .

4.4 Individual “Star” Network View

When “Inner Network” is selected from the “Table of Contents” view, the politician’s processed graph is visually displayed with the entities (ie, people, politicians, organizations, locations, etc.) which have most co-occurred with him being placed closer to his central position. By “most co-occurred” we refer to overall counts independent if the co-occurrences were in the same sentence, near sentences or

¹ openstates.org/api/v1/committees/?state=tx
² <http://openstates.org/api/v1/committees/TXC000010/>
³ [js/committees/committee-ids.sh](http://openstates.org/js/committees/committee-ids.sh)
⁴ http://www.house.state.tx.us/_media/pdf/committee.pdf

farther. Figure 4.6 shows the graph produced for Democratic State Representative of District 51, Eddie Rodriguez, who we will be using as a running example in this thesis.

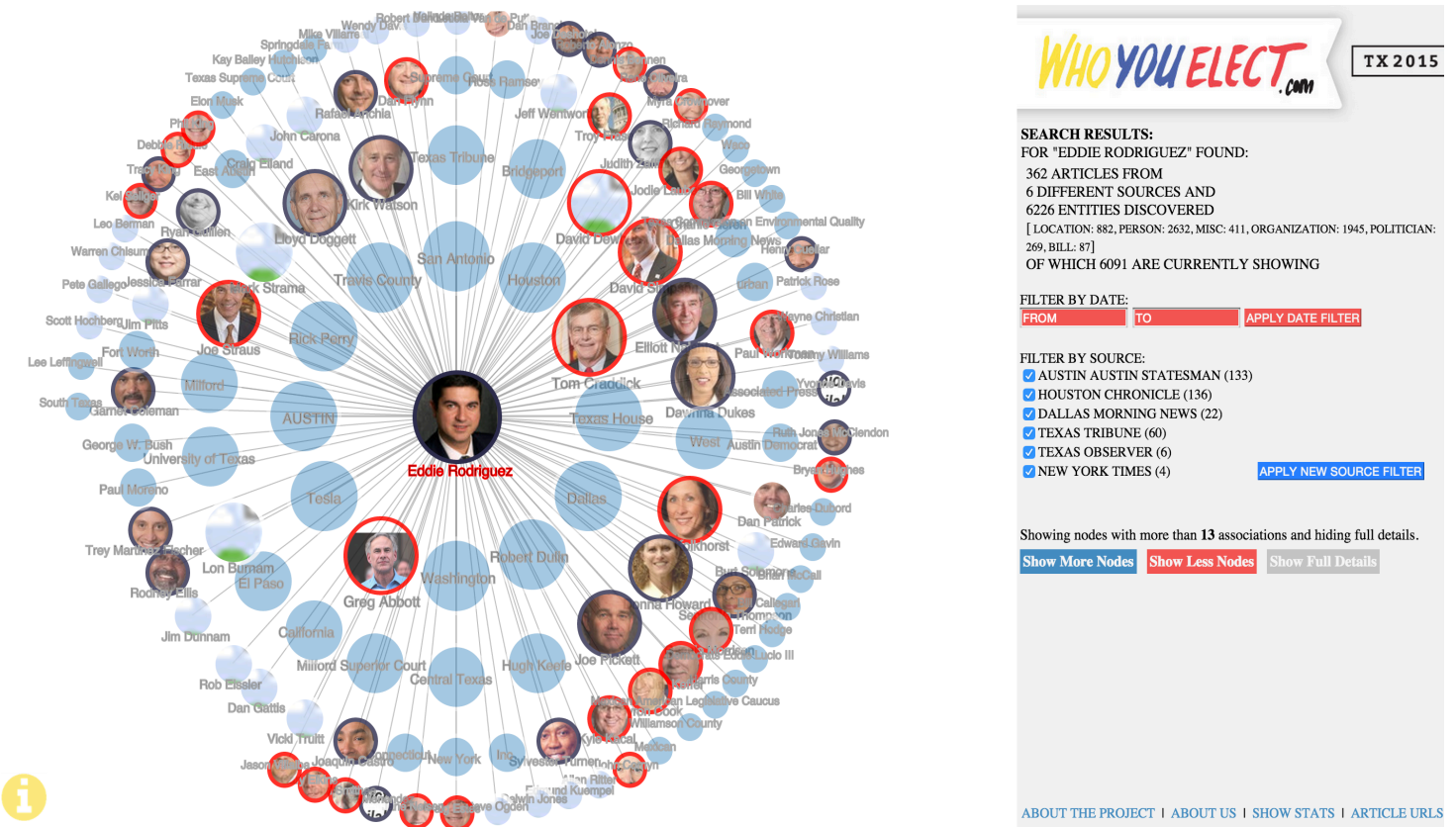
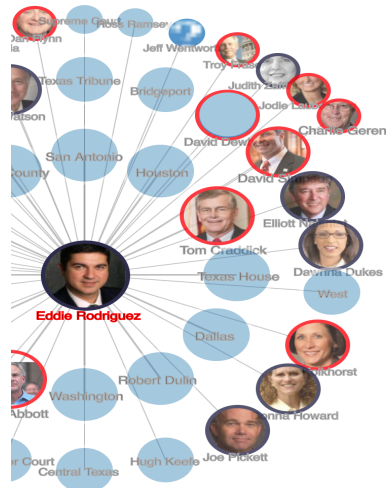


Fig. 4.6 Individual Star Network Landing Page for Representative Eddie Rodriguez

In the right hand side area we see that 362 articles were obtained from the six sources with the bulk coming from the Houston Chronicle (136), Austin American Statesman (133) and Texas Tribune (60). Additionally we see that 6226 entities were discovered along with the exact break down of counts for each entity types shown. We are presented with the option to filter by date range, filter by which news sources to include, show more, less or all (“Show Full”) nodes on the screen. The “ARTICLE URLS” link at the bottom right displays a pop up window of a sortable table of all 362 articles including their URL, date, number of sentences, number of unique entities, and number of relations created from it. This area is particularly interesting for someone interested in studying the media aspect of which sources reported on an politician and when. The functionality for “SHOW STATS” is of particular interest and is explained in detail in Sections 4.4.2 through 4.4.4. When hovering over a circle on the left hand side, the user sees how many co-occurrences that entity shares with Eddie Rodriguez. Additionally, politicians have their circles shaded with the color of their party affiliation, red for Republican and blue for Democrats. The bottom left yellow icon displays a general user manual.

4.4.1 CENTRAL ENTITY VIEW

The results of clicking on the center node Eddie Rodriguez, the politician under study, are seen in Figure 4.6.1. In it we can see the entities’ information, name, party, position, and district number, and additionally we see a map of the district represented. Immediately beneath the map is a link that displays the same pop up window of statistical findings as the “SHOW STATS” link explained in 4.4.3. In addition, we see a sortable table of the findings visualized in the graph with the exception that all entities are shown. Entity types are color coordinated and may be filtered by clicking on the appropriate header.



EDDIE RODRIGUEZ
Democratic State Representative
District: 51

Eddie Rodriguez is a Democratic member of the Texas House of Representatives, serving since 2003. Before serving in the legislature, Rodriguez was an aide to state representative Glen Maxey.

[hide map](#)

[CLICK HERE to show stats summary popup](#)
Topic Frequency And Extended Network Analysis Coming Soon!

Most Associated with:
[POLITICIAN | PERSON | ORG | MISC | BILL | LOC | ALL]

Name	Occurrences	Type
AUSTIN	241	location
Rick Perry	96	person
Travis County	91	misc
San Antonio	84	location
Houston	83	organization
Tom Craddick	77	Republican, Representative
Texas House	66	location
Dallas	60	location
Robert Dulin	57	person
Washington	47	location

Fig. 4.6.1 Individual Star Network Center View for Representative Eddie Rodriguez

4.4.2 SIDE BAR ENTITY ARTICLES TEXT VIEW

Figure 4.7 shows the result when the node for “Tesla” is selected. The right hand side shows that 44 co-occurrences were discovered between Eddie Rodriguez and Tesla in 12 different articles, 3 from the Austin American Statesman (AAS), 3 from the Houston Chronicle (HC), 2 from the Dallas Morning News (DMN) and 4 from the Texas Tribune (TXR). Any of the news source headers can be clicked to display only the articles from that source. The articles are listed in reverse chronological order with the article title being a link to the original article that is colored to match its sources color.

Relationship between **Eddie Rodriguez** and **Tesla**
44 co-occurrences found in 12 articles.
(AAS:3 | HC:3 | DMN:2 | TXR:4 | TOB:0 | NYT:0 | ALL:12)

Tesla

Letters to the editor feb 25 2015
2015-02-24 | AUSTIN AUSTIN STATESMAN
20 article, "New legislation aims to let **Tesla** sell directly to Texas consumers." *Bribing public officials is just so ho-hum. As Representative **Eddie Rodriguez**, D-Austin, says, "It helps to have more of a presence at the Capitol."*

As Representative **Eddie Rodriguez**, D-Austin, says, "It helps to have more of a presence at the Capitol." *Presence noted from **Tesla**, who recently gave \$133,000 to lawmakers.*

same article

New legislation aims to let tesla sell directly to
2015-02-19 | AUSTIN AUSTIN STATESMAN
Measures filed by state Representative **Eddie Rodriguez**, D-Austin, and Representative Charles "Doc" Anderson, R-Waco, among others, would allow **Tesla** to sell directly to buyers through its own dealerships.

Tesla resumed its fight with Texas' auto dealerships Thursday, with a posse of state lawmakers filing bills that would let the electric car company sell directly to Texas consumers. Measures filed by state Representative **Eddie Rodriguez**, D-Austin, and Representative Charles "Doc" Anderson, R-Waco, among others, would allow **Tesla** to sell directly to buyers through its own dealerships.

California-based **Tesla** currently operates "galleries" in Austin and Dallas, but customers cannot actually test drive the cars there, or buy them. Customers must order cars online that can cost more than \$100,000. If passed, the bills, which have been introduced in both the state House and Senate, would mean **Tesla** would no longer have to rely on the middlemen it argues are hampering its business. "This is a free-market issue, and it's a popular one," **Rodriguez** said.

If passed, the bills, which have been introduced in both the state House

Fig. 4.7 View of Co-occurrences between Eddie Rodriguez and Tesla in articles processed from the Austin American Statesman

4.4.3 TOP ASSOCIATED DISTANCE METRICS

Figure 4.8 shows the results of clicking on “SHOW STATS” from the landing page or from the right hand side view for when the central Eddie Rodriguez node is clicked. The “Top Associated” tabs along the top of the window each show what are the resulting most associated entities by entity type

if different distance metrics are used. For instance, figure 4.8 has the Top Associated “same sentences” tab currently selected, hence highlighted in red, and as such we see a ranked list of the entities which have the most “same sentence” co-occurrences with Eddie Rodriguez separated by entity type (Politicians, Organizations, Person, Location, Bills, Misc).

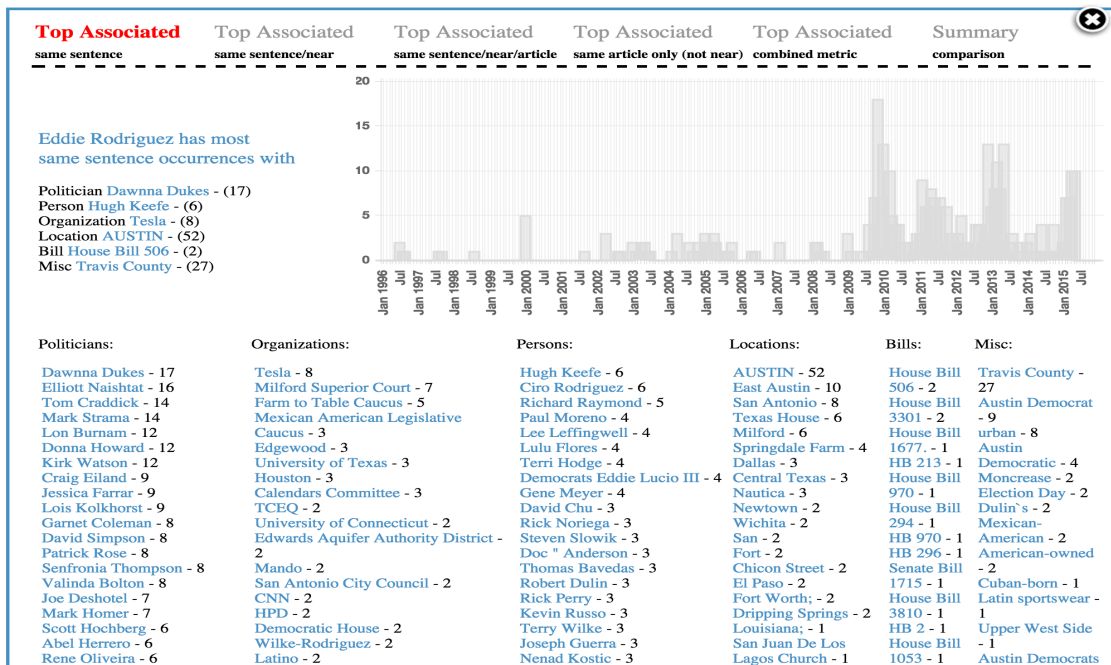


Fig. 4.8 “Show Stats” Screen for Texas Representative Eddie Rodriguez

We observe that “Dawna Dukes” is the Politician with the most same sentence occurrences with Eddie Rodriguez where on the Landing Page visualization, which uses the third Top Associated metric; “Tom Craddick” is the Politician with the most same article co-occurrences with Eddie Rodriguez. The fourth Top Associated metric “same article only (not near)” refers to entities that occur the most at a distance from the main politician being studied. This listing can be used as a sort of specialized, local term frequency – inverse document frequency (TF-IDF) measure because it allows the user to observe which entities occur the most at a distance from the politician of study, and from that, it can be inferred that the strength of the relationship is lessened.

The final Top Associated “combined” metric is a proposed combination of the first three top metrics with an additional ratio term which penalizes relations with high “far” distance co-occurrence counts with respect to their same sentence and near ones and is discussed in detail section 4.4.5.

4.4.4 ARTICLE STATISTICS TEMPORAL VIEW

In addition to the ranked listings, we also see an interactive time line of the distribution of the article dates retrieved. Figure 4.9 shows what happens when a month is clicked on. Additionally, the months can be navigated via the “prior” and “next” links in the gray popup. This view allows us to visually assess whether there are periods of interest based on peaks in articles retrieved and the landing page “date range” filters can be used to focus on this time period only. Future work of interest could include trying to see whether communities found in the extended view visualization correlate with “events”, i.e. periods of time with increased article output.

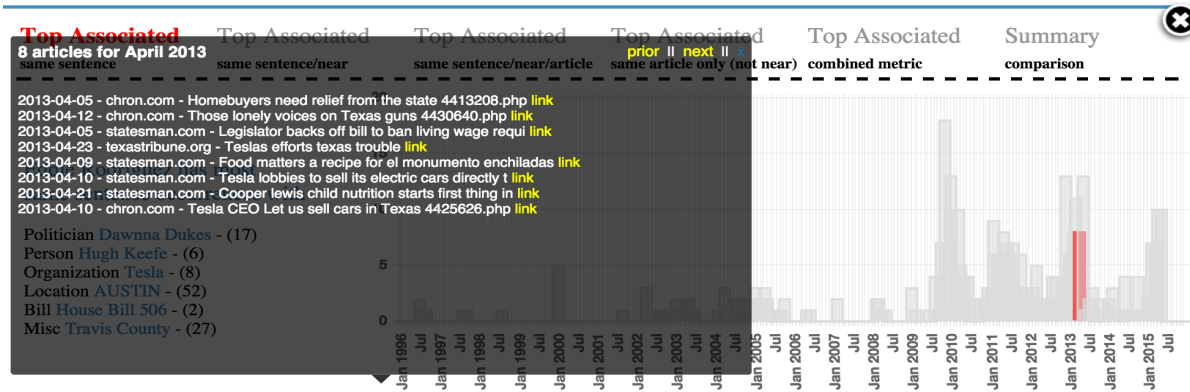


Fig. 4.9 “Show Stats” Screen for Texas Representative Eddie Rodriguez

4.4.5 “COMBINED” METRIC & OTHER DISTANCE METRICS SUMMARY COMPARISON VIEW

The final “Summary combined results” tab along the top of the window that pops up when “SHOW STATS” is clicked on from the landing page is of particular interest in that it allows a user to compare the ranked lists that each metric returns with respect to a particular entity type. The proposed “Combined” metric that appears in the final column is defined as:

$$weight = (same\ sent + .5 * near\ sent + .1\ same\ article) * boosting$$

$$where\ boosting = combined\ co-occurrences / same\ article\ only\ co-occurrences$$

To clarify, “same sent” refers to the number in column 1, “near sent” to column 2 – column 1, “same article” to column 4, and “combined co-occurrences” to column 3. The coefficients associated with penalizing “near sentences” and “same article” co-occurrences (.5 and .1 respectively above) could and should be improved by having a person with domain knowledge, a political scientist specializing in Texas Politics in this instance, view the ranked results for each entity type of various, different politician graphs and then reorder the results of each if necessary thereby assessing their accuracy. It would then be a relatively straightforward process to use these newly labeled rankings to update the coefficients to produce results that more reflect the opinions of domain experts.

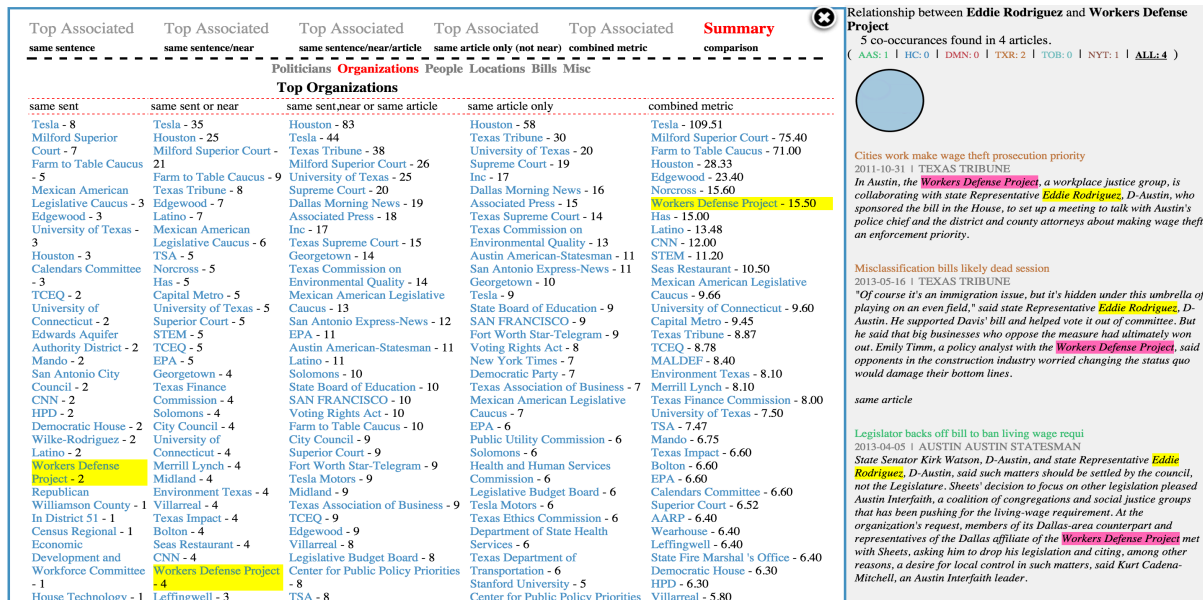


Fig. 4.10 “Show Stats” Screen for Texas Representative Eddie Rodriguez

This process in itself is still subjective and could theoretically lead to over fitting if the training set is too small or is very different in behavior than the test data. The current coefficients were chosen by intuition to reflect near sentence occurrences being half as important as same sentence matches, and same article co-occurrences being a tenth as important. Its probably more likely that near sentence co-occurrences are more important and should be given a higher number, perhaps .75 and that same article ones should be slightly less important, perhaps a coefficient of .08, but that is left for future work.

These results could be improved by additionally taking into account the number of sentences and unique entities in which co-occurrences are detected. For instance, in articles that have large numbers of entities relative to the number of sentences, and where it is possible that the article in itself is a listing of election results or candidates running for an office, it may be better to exclude “same article” relationships formed from the document since they will be largely noise.

As a convenience method for showing what sort of rankings would entail from using different coefficient values, the visualization allows for user to pass them in, via the URL parameters “near_co” and “same_art_co”. For instance, the url:

“explorer-view.html?s=Eddie Rodriguez&near_co=0.75&same_art_co=0.08”

would show the Individual Star view for Eddie Rodriguez with the “Combined” metric using .75 and .08 for the new coefficients of near and same article only co-occurrences. Figure 4.10a shows the top revised results for Organizations most associated with Eddie Rodriguez using the above coefficients.

Top Associated same sentence/near/article	Top Associated same article only (not near)	Top Associated combined metric	Summary comparison
Houston - 83	Houston - 58	Tesla - 141.63	
Tesla - 44	Texas Tribune - 30	Milford Superior Court - 93.08	
Texas Tribune - 38	University of Texas - 20	Farm to Table Caucus - 80.80	
Milford Superior Court - 26	Supreme Court - 19	Houston - 34.55	
University of Texas - 25	Inc - 17	Edgewood - 27.72	
Supreme Court - 20	Dallas Morning News - 16	Norcross - 22.98	
Dallas Morning News - 19	Associated Press - 15	Has - 20.00	
Associated Press - 18	Texas Supreme Court - 14	Workers Defense Project - 17.90	
Inc - 17	Texas Commission on	Latino - 16.69	
Texas Supreme Court - 15	Environmental Quality - 13	Seas Restaurant - 15.40	
Georgetown - 14	Austin American-Statesman - 11	STEM - 14.56	
Texas Commission on	San Antonio Express-News - 11	CNN - 14.00	
Environmental Quality - 14	Georgetown - 10	Capital Metro - 13.69	
Mexican American Legislative	Tesla - 9	Texas Finance Commission - 12.00	
Caucus - 13	State Board of Education - 9	University of Connecticut - 10.98	

Fig. 4.10a “Show Stats” Screen for Texas Representative Eddie Rodriguez with different coefficients

When a user scrolls over or selects an entity from any of the lists in the “Summary comparison” tab, that entity is additionally highlighted in the other columns. For instance, Figure 4.10 shows the ranked position of the Workers Defense Project, an Organization entity type, among the difference metrics. As we can see, the Workers Defense Project currently does not appear among the upper ranked Organization entities for the third metric, which again is what the landing page visualization defaults to using, but we do see that it ranks highly according to our new metric since it has 2 same sentence occurrences, 4 – 2 = 2 near sentence occurrences, and relatively few distant occurrences (which is not visible in the figure, but is 1). Thus our new metric is computed as:

$$weight = (2 + .5*2 + .1*1) * (2 + 2 + 1) / (1) = 3.1 * 5 = 15.5$$

If we then click on the Workers Defense Project (see APPENDIX F for full results) we see that in fact there is a strong direct relationship between the two as shown by this quote from an article from Texas Tribune¹ in October 31, 2011.

¹ <http://www.texastribune.org/2011/10/31/cities-work-make-wage-theft-prosecution-priority/>

*“In Austin, **the Workers Defense Project**, a workplace justice group, **is collaborating with state Representative Eddie Rodriguez, D-Austin**, who sponsored the bill in the House, to set up a meeting to talk with Austin’s police chief and the district and county attorneys about making wage theft an enforcement priority”*

4.5 EXTENDED VIEW WITH COMMUNITY DETECTION

When a user clicks on “Larger Network” for a given politician in the Table of Contents page, the data for the extended view network created during step 3.6 is displayed in an interactive webpage. This view is of the undirected graph with edges weighted according to the “Combined” metric described in section 4.4.5 that was created during the processing of the politician being studied. While viewing the graph, it is important to keep in mind that the edges are weighted by this “Combined” metric and do not just reflect the straight co-occurrence of two entities. This graph is much larger than the prior individual star view that could have at most $N-1$ edges displaying, where N is the number of entity nodes. This full graph on the other hand could possibly have $N*(N-1)/2$ edges if it is fully connected, i.e. if all nodes have connections with all other nodes. For this reason, in order to derive meaningful insight into the network it is necessary to be able to search for “communities” amongst the nodes and to filter out edges based on the weight, i.e. “importance”, of an edge between two entities. The idea of detecting communities in a network is similar conceptually to that of clustering in multivariate analysis and machine learning, and refers to a densely connected group of nodes that are well separated from the rest of the network. More formally and commonly, the definition of a community entails that the number of intra-community edges amongst the nodes of a single community be greater than the number of inter-community edges. There are a vast number of detection algorithms¹ that can be used, but for our case we decided to go with the Louvain method[14] since it works on weighted graphs, provides a hierarchy of clusters, and is fast to run even on large graphs. We leverage an open-source JavaScript implementation of the method and D3.js to produce visualizations such as that of Figure 4.11 that shows the network formed by articles obtained and processed for Eddie Rodriguez. The graph in its entirety has 6226 unique entities/nodes and 572,965 edges.

Our tool does community detection automatically, but allows the user to pass in a parameter to specify the number of communities they want the system to find and display, similar in fashion to selecting the “k” number of clusters to discover in k-means clustering. Selecting fewer communities to find and display leads to higher computational and time cost, and vice versa, more communities to find is less computational and time intensive.

Additionally, the modularity of the overall communities is calculated and presented in the upper left hand side of the visualization. Modularity is a function that given a partition of nodes, tells you how good the community structure of that partition is [36] This value, between 0 and 1, measures the strength of division of the communities of the network and as such can be used as a sort of quality measure for the communities. Specifically, it is a weighted sum over all of the communities where the total number of inter-community edges is calculated for each node and then the expected number of edges under a random graph setting is subtracted from it. For more information on modularity and a concise overview of community detection in graphs in general see [35,37].

The tool also provides the ability to pass in a “threshold” parameter that removes edges whose weight is less than the threshold from the graph, and after which any nodes that no longer have edges are also filtered. In this case, setting a low weight threshold parameter is more computationally and time intensive, and results in more edges and nodes to visually portray which can make the

¹ “Community structure in networks” by Arias & Ferrer-i-Cancho provides a good overview of community detection concepts & different methods. <http://www.cs.upc.edu/~csn/lab/session5.pdf>

visualization very congested. Conversely it does allow for the teasing out of more detailed relationships.

In Figure 4.11 the community number parameter was set to 25 and the weight threshold was set to 15, which reduced the size of the graph dramatically to 952 nodes and 1869 edges. Community assignments are listed in the left hand side, and each community section can be expanded or collapsed, and the listing of nodes within each community can be selected therein or from the graph directly. Once a node is selected, its contents and edges are shown in the right hand side area. In Figure 4.11, the node for Eddie Rodriguez has been selected and dragged to a position for greater visibility. The right hand side shows he has 86 edges with weight greater than the threshold of 15 passed in, and the nodes associated with those edges are listed beneath him in descending order based on weight. The entity for “Robert Dulin” of type PERSON for instance has the highest “Combined” metric weight of 248.4, which means that he is the entity associated with Eddie Rodriguez and is thus listed at the top. The list of entities can also be ranked by “community” with the nodes within each community then ranked by weight or simply in alphabetical order by first name.

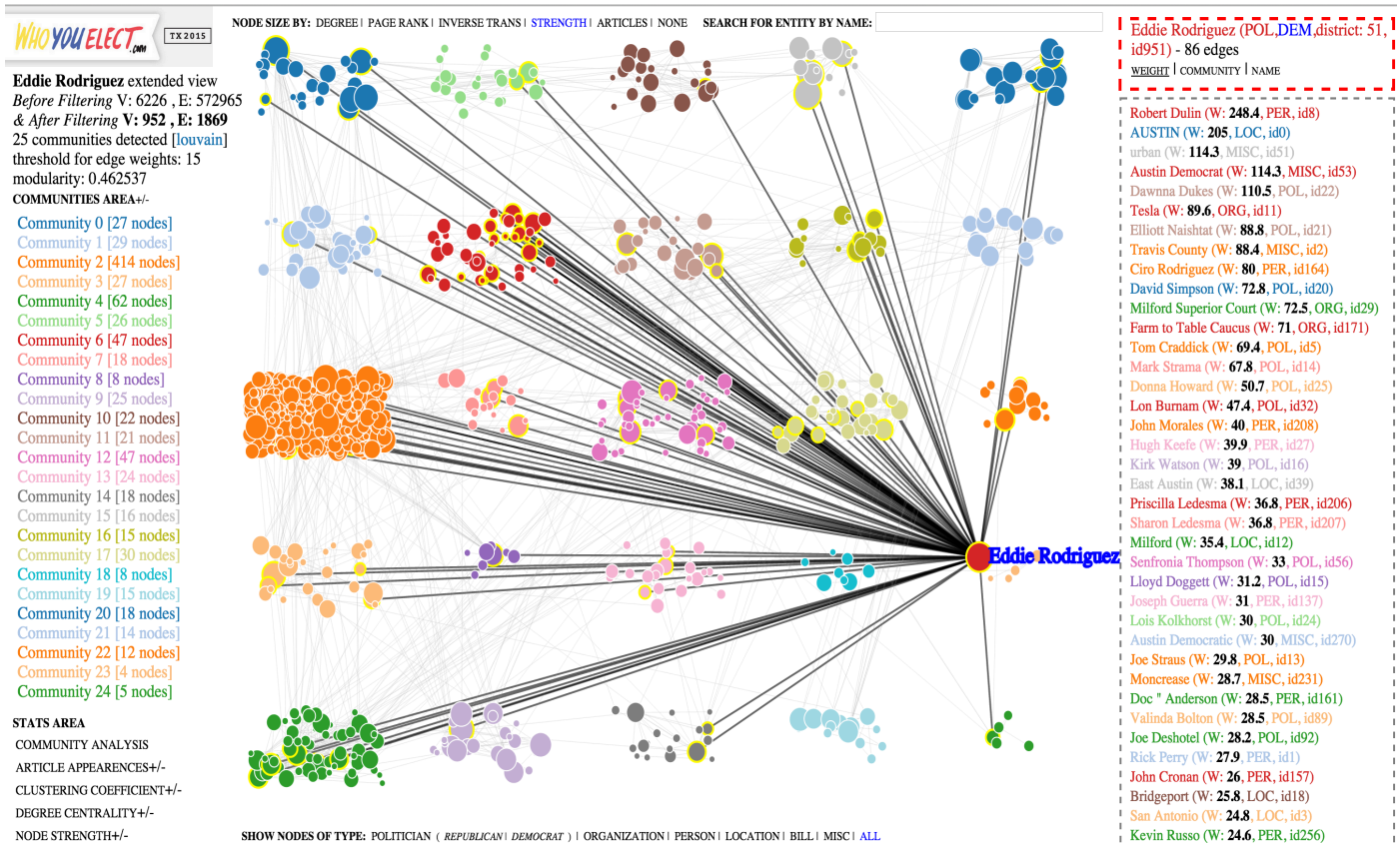


Fig. 4.11 “Extended View” with Communities Screen for Eddie Rodriguez with Nodes Weighed By Strength.

4.5.1 ENTITY-ENTITY GRAPH EXPANSION VIEW

Additionally, when a node is selected, the list of associated nodes in the right hand column can be selected in which case a reduced view of the graph is displayed showing the link between that node and the first node selected along with any other nodes they both have edges with. Figure 4.12 shows what happens when the entity “urban” of type MISC is selected from the prior results set for Eddie Rodriguez.

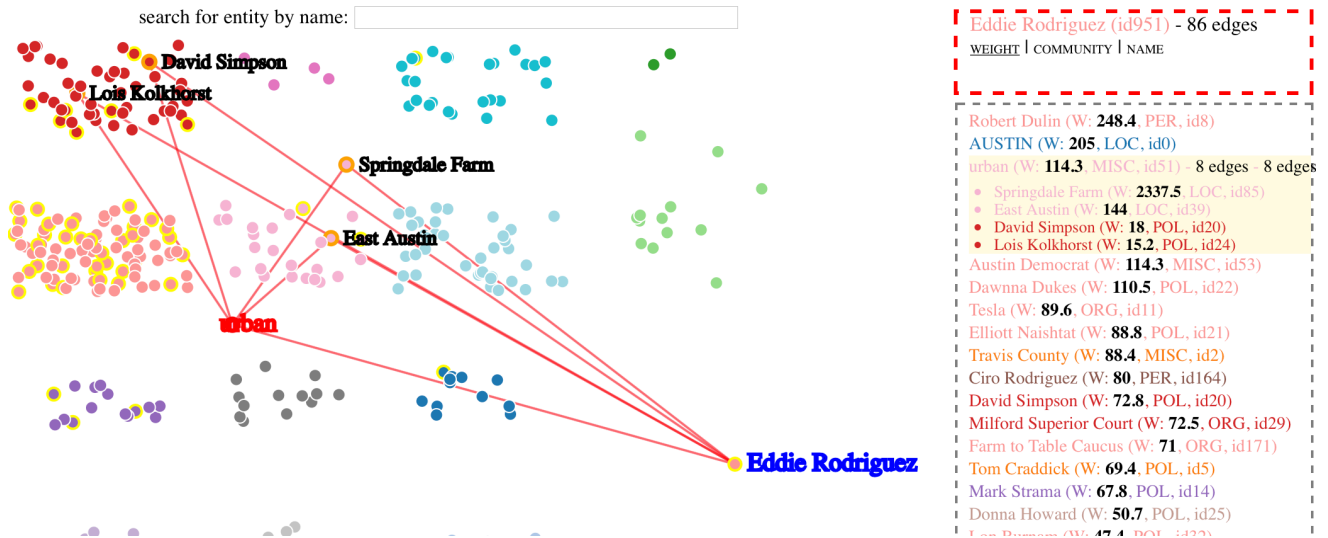


Fig. 4.12 “Extended View” with Eddie Rodriguez and then “urban” selected from right menu.

In it, we can see that there are 4 other nodes associated with Eddie and “urban”; the locations “Springdale Farm”, and “East Austin”, and the politicians “David Simpson” and “Lois Kolkhorst”. As it turns Eddie Rodriguez, David Simpson, and Lois Kolkhorst are all part of the “Farm to Table Caucus” which appears a few entities down in the right hand side list, and that “Springdale Farm” is an “urban” farm located in the “urban” neighborhood of “East Austin”. This relationship is made even clearer if one selects the “urban” node directly from the graph as seen in Figure 4.13.

The only new entities in this grouping are the newspaper that reported the most on these entities, the Texas “Tribune” of entity type ORGANIZATION, and Paula Foore, one of the owners of the “Springdale Farm”. This information was discovered simply by clicking on “Farm To Table” and “urban” in the prior Star view as seen in Fig 4.14.

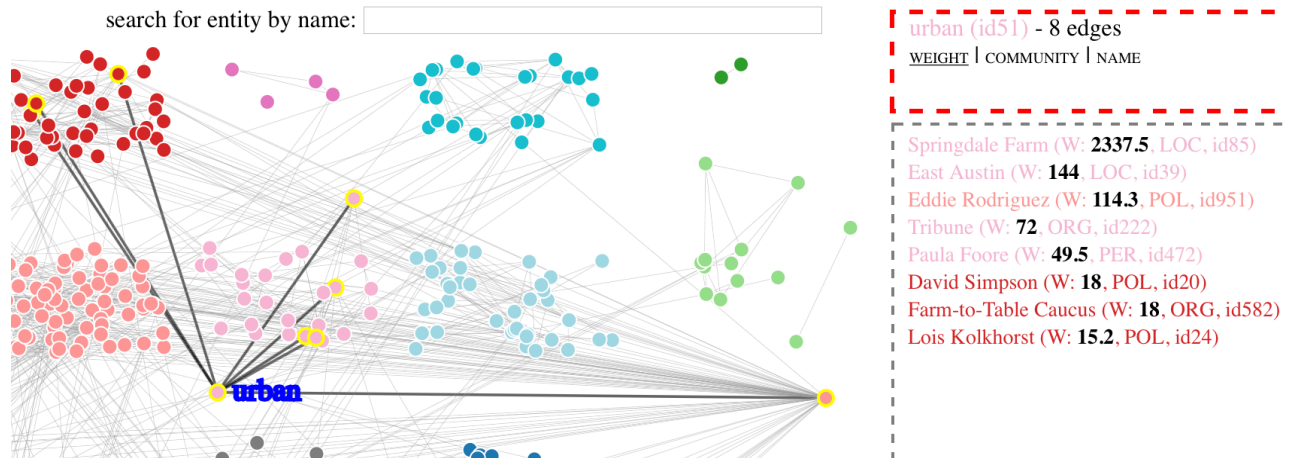


Fig. 4.13 “Extended View” with “urban” selected

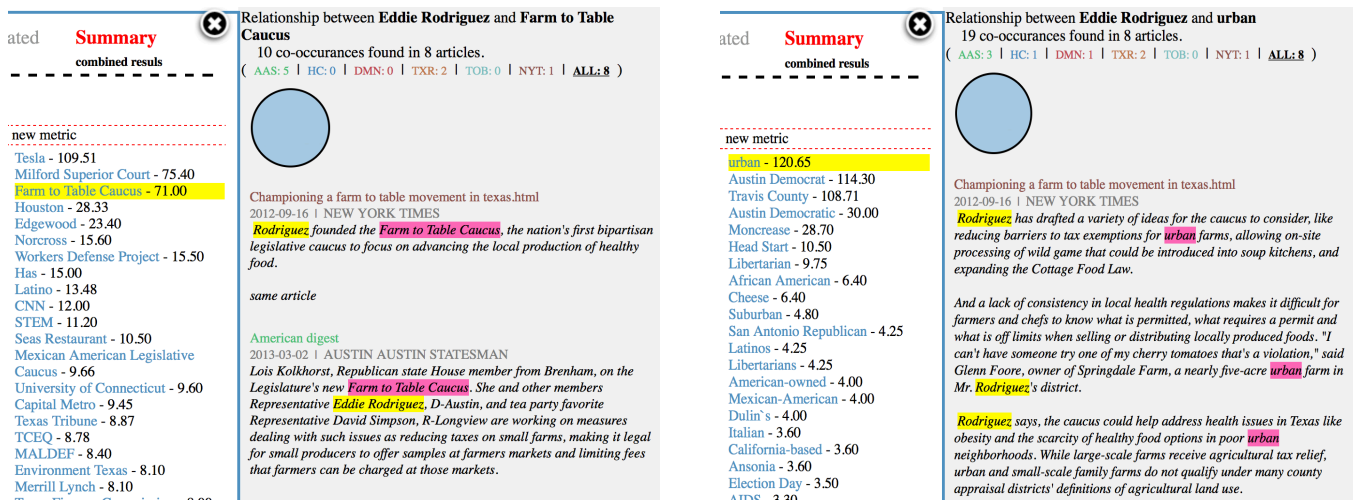


Fig. 4.14 “Individual Star” Views of Eddie Rodriguez with “Farm To Table” & “urban” respectively selected

To emphasize the importance of the community number and threshold parameters that a user may pass into the tool, Figure 4.15 shows the results of the prior example when the threshold parameter is passed in with a value 3 as opposed to 15, which allows for many more edges and nodes to remain. This time when the “urban” entity is selected we see many more entities than we did before including the Farm to Table Caucus itself, and the other owner of the Springdale Farm, Glen Foore.

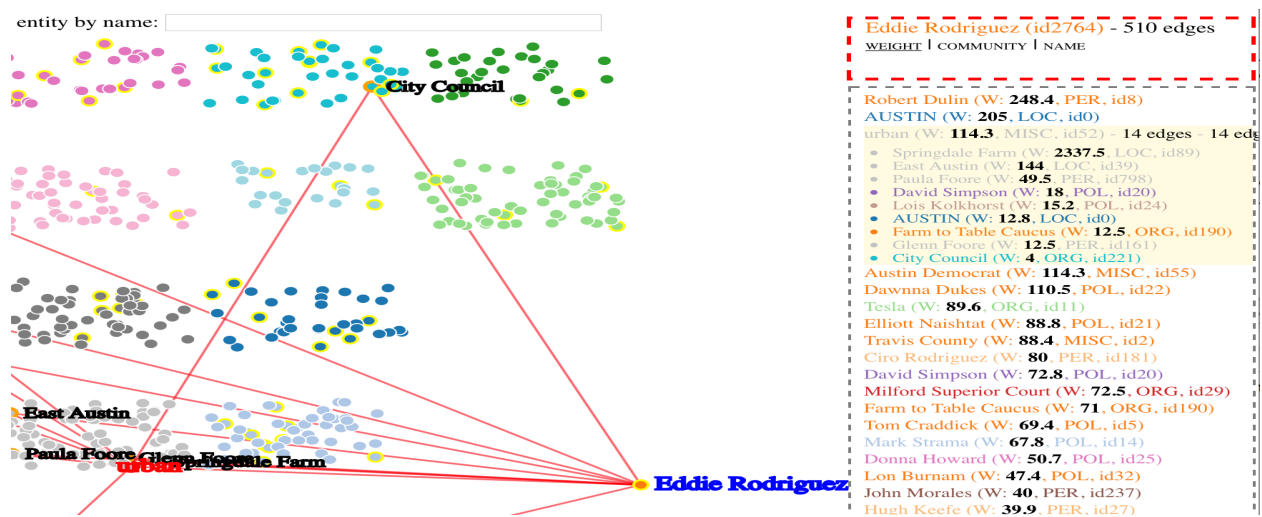


Fig. 4.15 “Individual Star” Views of Eddie Rodriguez with threshold parameter lowered to 3

Additionally, in order to facilitate searching for specific entities by name, the “search entity by name” input box at the top of the tool begins to auto-complete and show entities once three characters have been entered. Figure 4.16 shows the results when “Far” has been entered into the search box. Upon selecting one of the results, the node is displayed on the main graph, its community is opened in the left hand side “Communities Area”, and its information and edges are displayed in the right panel.



Fig. 4.16 Autocomplete functionality while searching for entities

4.5.2 COMMUNITY EXPANSION VIEW

In addition to these features, a user can expand any of the communities on the left hand side of the Extended View to see and interact with the listing of entities within it. For instance, figure 4.17 shows the view for community 13 expanded.

Here we can see many of the same entities from the prior example, “East Austin”, “Springdale Farm”, “Paula Foore”, her husband “Glen Foore”, “urban”, and “Tribune”. “Foundation School Fund” and “Travis Central” have clear links to that grouping, however we also see many entities that seem to belong within two other general groupings found within the community. There is a definite grouping found around House Bill 3839, House Bill 3916, House Joint Resolution, NSA, National Security Agency, Joe Deshotel and most of the other politicians of the community, and there is another around LULAC, Luis Vera and Natasha Rosofsky, Eddie Rodriguez’s wife who is only mentioned once in 362 articles.

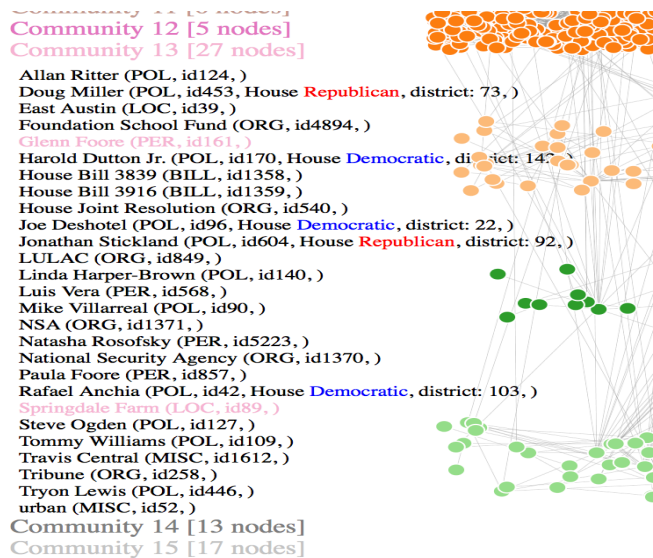


Fig. 4.17 Community Listing

This finding of groups within the community, as is often in the case of the art of cluster analysis, means that the allowance of more clusters could help separate these out, at the expense of having more clusters to analyze and a less cohesive overall summarization of content.

4.5.3 ADDITIONAL CENTRALITY MEASURE TOOLS AND VISUALIZATION OPTIONS

In addition to the aforementioned methods with which to inspect communities and nodes, the system also includes other mechanisms whose goal is to help find and highlight potentially interesting relationships to aid the end user’s data exploration. These appear under the “Stats Area” header in the left hand side, and include “Article Appearances”, “Clustering Coefficient”, “Degree Centrality”, “Node Strength”, “Page Rank”, and “Inverse Articles Strength”.

The “Article Appearances” mechanism simply produces an ordered list of nodes in decreasing order by how many articles they have appeared. The first listing is always of the main politician being studied (“Eddie Rodriguez” in our case) and shows how many articles were found over all. The rest of the listing highlights the “famous” entities, i.e. those who get mentioned in many articles. In Eddie’s case, the top listings after his own name are mainly the largest cities/counties in Texas “Austin, Houston, San Antonio, Dallas, Travis County” followed by the biggest politicians.

The next mechanism is the “Clustering Coefficient” list where by the system calculates the clustering coefficient, also called transitive property, for each node. This is a common technique in network

analysis that shows what percentage of a node’s direct neighbors are also connected. An analogy to understand it better is that the clustering coefficient of a person measures how many of their friends are also friends with one another. This value is high if your friends are also generally friends with one another, and low if not. Figure 4.18 shows the list generated for the extended view of Eddie Rodriguez. It is ordered first from highest to lowest transitive value, and then entities with the same transitive value are ordered in descending fashion based on the degree, i.e., the number of edges, for the entity. Additionally, the results are filtered so that only unique groupings appear in the list and results with a value of 1, which means a node had only one edge, are also filtered as they are generally uninteresting while searching for groupings.

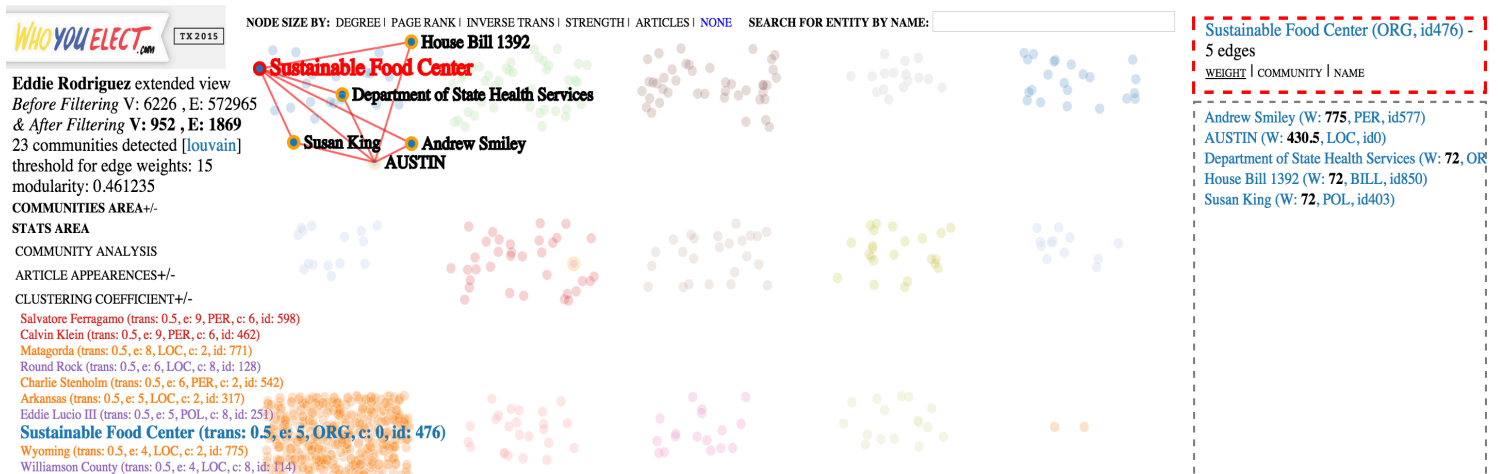


Fig. 4.18 Results from automated Clustering Coefficient analysis for “Sustainable Food Center”

In the result set for Eddie Rodriguez, we see about 125 initial results with the highest transitivity value possible, which imply the detection of perfectly connected groups. In Figure 4.18, “Sustainable Food Center” has been selected from the list and we see they all pertain to the same community already, but more so that they revolve around a specific topic or story. By using the individual star view again for Eddie Rodriguez and selecting “Sustainable Food Center” shown in figure 4.19, we observe that the two entities co-occurred 3 times in three different articles, one each from the Austin American Statesmen, Houston Chronicle, and Texas Tribune respectively from late 2012 to mid 2013. From them we see that Andrew Smiley, who has the edge with most weight, is the deputy director of the Sustainable Food Center located in Austin (2nd highest weight) and that Representative Susan King (identified correctly as a politician) authored House Bill 1392 which would ensure consistency by requiring the Department of State Health Services to provide written responses to regulatory inquiries within 30 days. As a caveat, after seeing two articles with the same title in Fig 4.19, we observed that the Houston Chronicle ran the Texas Tribune’s story on the same day though the system does not take this into account.

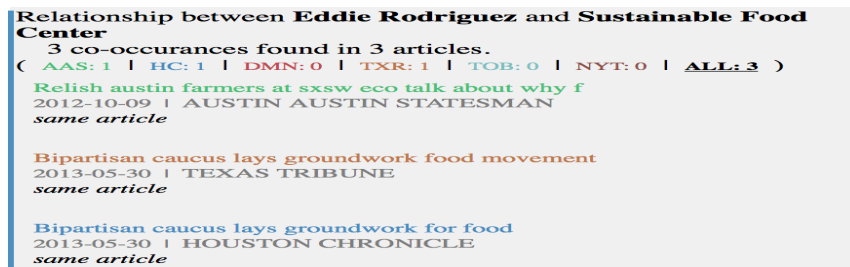


Fig. 4.19 “Sustainable Food Center” Results

The Clustering Coefficient mechanism thus does a good job of gathering articles of the same topic together independent of whether there are direct mentions of the entities composing the groups within the articles and generally does a good job of discovering well connected groupings within

communities. For instance, in the above example, only Eddie Rodriguez and Sustainable Food Center occur in the articles gathered while the others occur in only some. For future work, it would be good to further take into account the number of articles associated with each grouping from the Clustering Coefficient list to better identify, filter and rank more highly topics/groupings with more articles backing them, though this addition could be included at least implicitly by adjusting the proposed metric by which the edges in the extended view are weighted to factor in the number of articles as well.

The next mechanism by which an end user can also discover more easily interesting relationships is through the Degree Centrality listing. This mechanism simply provides a way to see which are the nodes with the most edges. As expected because the context of all articles should include him, Eddie Rodriguez has the highest degree centrality value of 86. In the next position is Lance Gooden, with entity type politician and currently inactive, who shows up in only one article with Eddie Rodriguez! This article¹ is in essence a listing of committee assignments and thus is packed with entities found and relationships created during the processing step. Figure 4.20 shows Lance Gooden being selected from the Degree Centrality list and then Scott Turner being selected from the right hand side. Scott Turner was chosen because of the number of node edges he has in common with Lance Gooden; a characteristic discovered by noticing that most of the entities in the right hand side when hovered over made reference to him. In the figure we can see the names of many committees, “General Investigating and Ethics”, “Homeland Security & Public Safety” etc., but also notice that the named-entity-recognition processing, namely steps 3 and 4 in Section 3.5, erroneously joined them with politician names. Many of these entities are placed in the largest and “noisiest” community that contains 412 of the 952 total nodes after filtering. Future work will account for this in both the processing steps above, but also by allowing a site administrator the ability to remove articles determined to be “noise”. The removal of this node or better yet, the correct processing of it, would improve the community detection and analysis greatly.

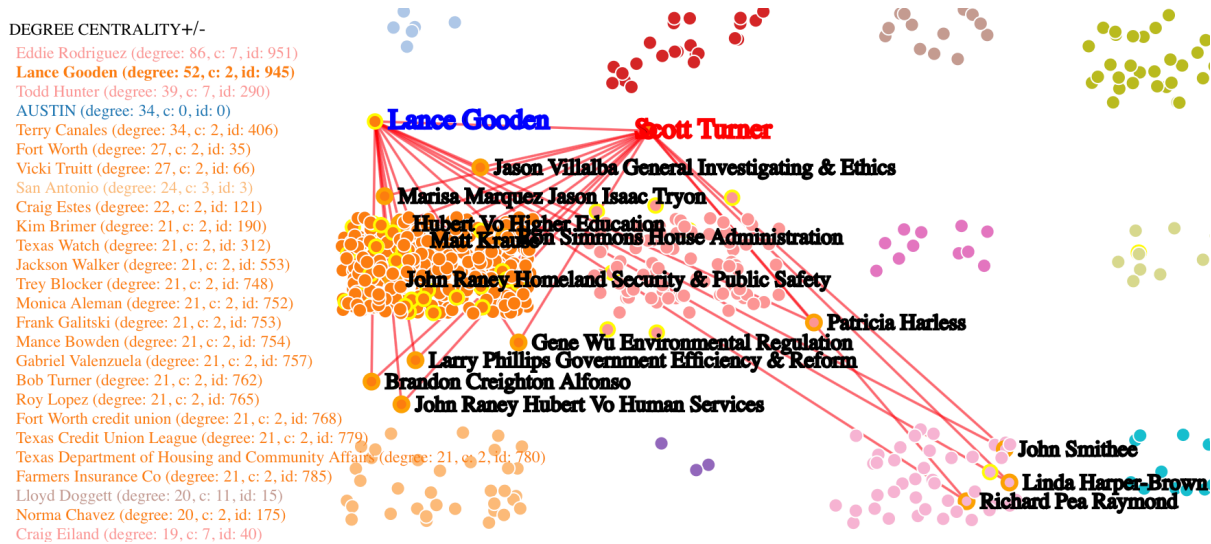


Fig. 4.20 Degree Centrality stumbling upon Committees

If instead of looking at the noisy example of Lance Gooden, which was good in fact because it allowed us to easily identify noise/issues, we look at the next entity of the Degree Centrality list, we see Todd Hunter. Looking at the Committees View page, outlined in Section 4.3, we see he is the chair of the important Calendars committee, of which Eddie Rodriguez is a member, and which in turn explains his high degree.

¹ <http://www.texastribune.org/2013/01/31/strauss-makes-house-committee-assignments/>

The next mechanism in the list is “Node Strength”, which is similar to Degree Centrality except that instead of just being the number of edges a node has; it is the sum of the edges of a node. Since our “Combined” metric is such that stronger connections between people have higher weights, this allows a user to find entities that have combined high weighted edges. Figure 4.21 shows the listing for Node Strength from Eddie Rodriguez’s extended view graph. We observe that the highest rank goes to Joe Strauss and additionally show the grouping that is formed by hovering over his most associated edge from the right hand list, Public Safety Committee. As the Speaker of the House of Representatives, Joe Strauss has the ability to set Committee appointments and as such its quite reasonable that he would be highly associated with the Public Safety Committee, the Technology, Economic Development and Workforce Committee and Natural Resources one.



Fig. 4.21 Node Strength showing the Speaker of the House

Page Rank is the final mechanism listing. This one is based on the paradigm that Google uses to rank web pages where in a page is important if it is pointed to by other important pages. It is a general measure for reputation, trust, and prestige. Figure 4.22 shows the resulting list of top Page Rank entities within the context of Eddie Rodriguez articles and a well-known environmental nonprofit organization, the Sierra Club, is one of the top entries.

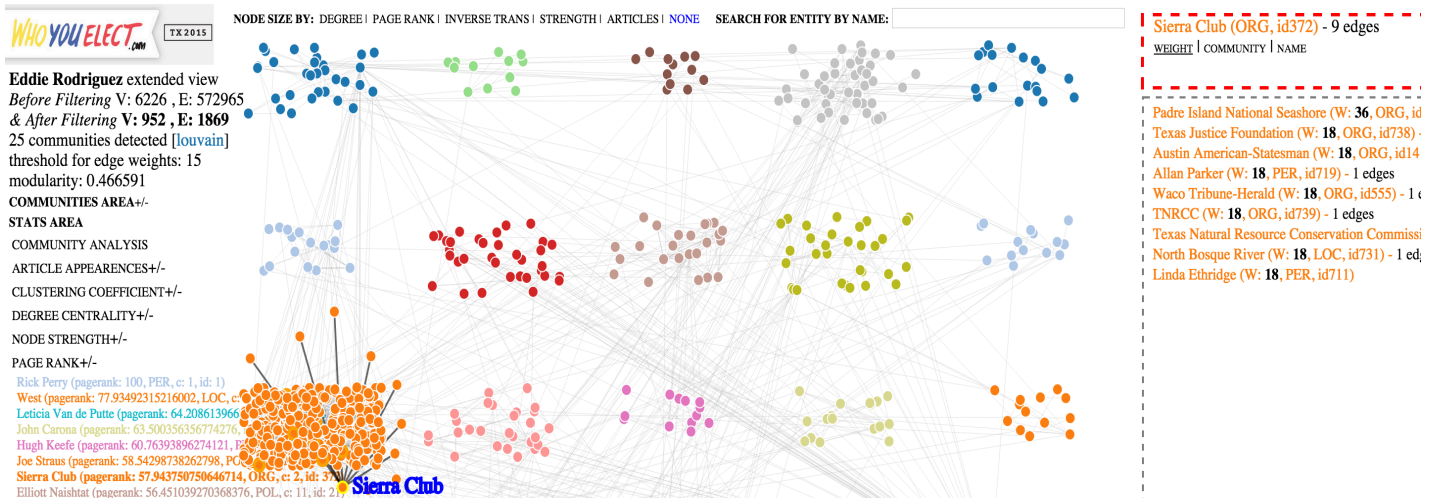


Fig. 4.22 Node Strength showing the Speaker of the House

This is of particular interest because it fits in with the theme of “agriculture” that has been already observed in other entities associated with Eddie Rodriguez, but because the Sierra Club has no same sentence or near sentence co-occurrences with Eddie Rodriguez, and hence had a low weighted edge with him, it had not yet been observed.

In addition to the mechanisms for automatic discovery of interesting relationships just mentioned, the user also has the options to change the size of nodes based on their degree, page rank, article appearances, strength, or the inverse of the node's strength. This last one is to facilitate the search for nodes that are closer to the edge threshold. This allows for finding relationships that are still possibly important, but do not automatically shoot up to the surface because of others having more strength. It is helpful when the number of nodes on the graph is quite large; though additionally this can be achieved manually on the right hand side when a node is selected by changing the ordering to be from decreasing to increasing.

4.5.4 COMMUNITY ANALYSIS VIEW

The final area of comment is regarding the "Community Analysis" section that appears in the left hand side. It is the most detailed in terms of the information it provides the user, but for that reason is also the least user friendly since it expects some degree of familiarity with network analysis jargon. When clicked, a popup window appears with analytics for each community along with an "Overall" view that allows for comparing communities side by side. The window may be resized and dragged to whatever location on the screen for ease of use.

Figure 4.23 shows the "Entities Info" analysis view for the first community. In it we can see that the community is composed of 17 entities spanning 170 articles and that the community has a conductance value of .2624 and an expansion value of .4157. These final two values are quality measures that give us an idea of how well the community is defined. Conductance is the fraction of edges leaving a single community and smaller values are better since that means most of the edges are internal to the community. Expansion is a related measure which is the number of edges per node leaving the community; again the lower the better. The Entities Info table provides in depth information about each entity in the community selected including name, entity type, political party,

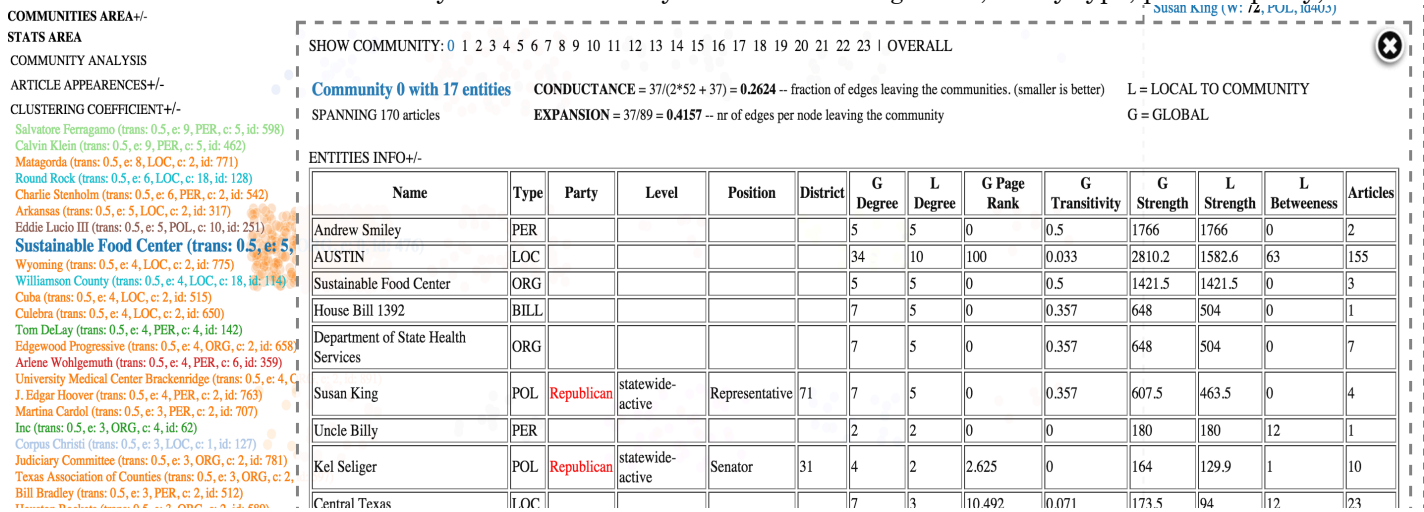


Fig. 4.23 Community Analysis View with Entities Info Panel Open for Community 0 sorted by L Strength.

level, position, district, global degree, local degree, global page rank, global transitivity, global strength, local strength, local betweenness and the number of articles the entity appears in. Political party, level (federal, state-active, state-inactive), position, and district are only displayed if the entity is of type "POL". "Global" in this context means graph wide, whereas "local" means local with respect to the nodes in the community. The "betweenness" of a node is the number of shortest-paths in the network that pass through it. It gives an idea of nodes near the center of a community though who are not the central hubs, since degree more appropriately determines that. Local strength and degree help us ascertain important nodes in the community that can then be assessed using the Individual Star page to look up texts pertaining to important nodes. The results in Figure 4.23 line

up with the Clustering Coefficient results explained near the beginning of Section 4.5.3 and displayed in Figure 4.18.

Another way to ascertain the “story” or “stories” behind a community is to simply look at the articles most associated with it. To do so we look at the “Articles Info” section located below the Entities Info table. The results for community 0’s articles are shown in Figure 4.24 and confirm the “agriculture” them from before, which is not to say there aren’t more “important stories” to be derived from the information presented; this just happens to be the most obvious.

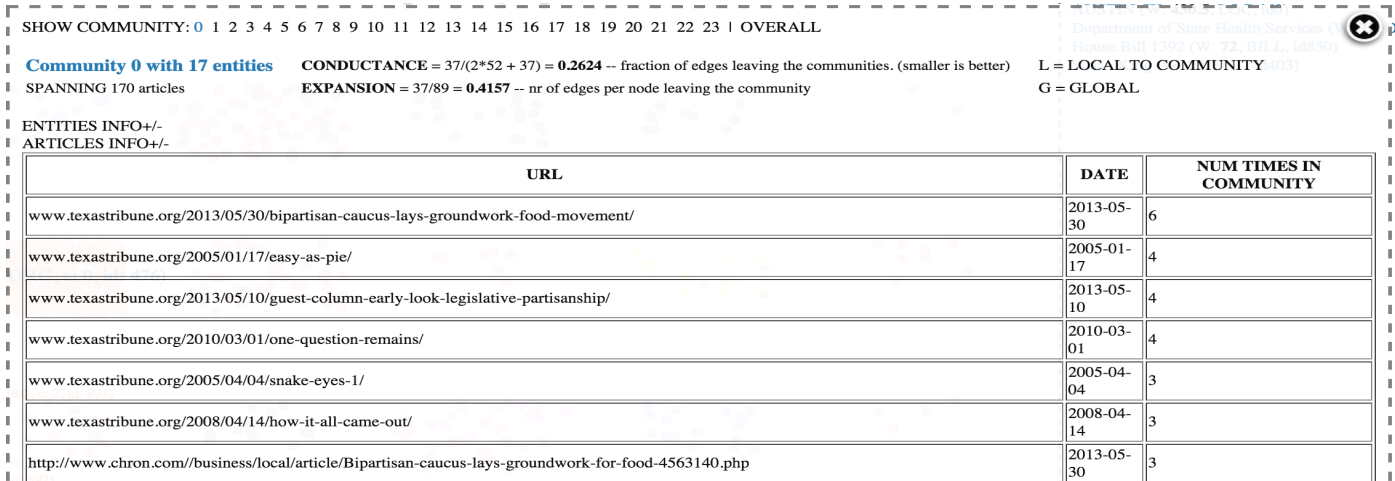


Fig. 4.24 Community Analysis View with Articles Info Panel Open and sorted by # of nodes the article appears in.

Taking into account the temporal aspect of when articles are published is important as well especially when a given politician has been in politics for a relatively long period of time since different communities may pertain to entities found largely in articles from a particular period of time. It should also be understood though that most online sources do not offer free access to articles published before the mid nineties as in order to do so the publisher had to manually digitize the content in some way.

Finally, the “Overall” view allows us to see statistics on the communities side by side as seen in Figure 4.25. We see each communities Node count, Internal Edges Count, Conductance Value, Expansion Value, Articles Covered, and information regarding when those articles were published. We can see which are the communities that are better defined by finding those with lower conductance and expansion values; which communities are larger based on nodes and which are composed of entities largely from a given time period.

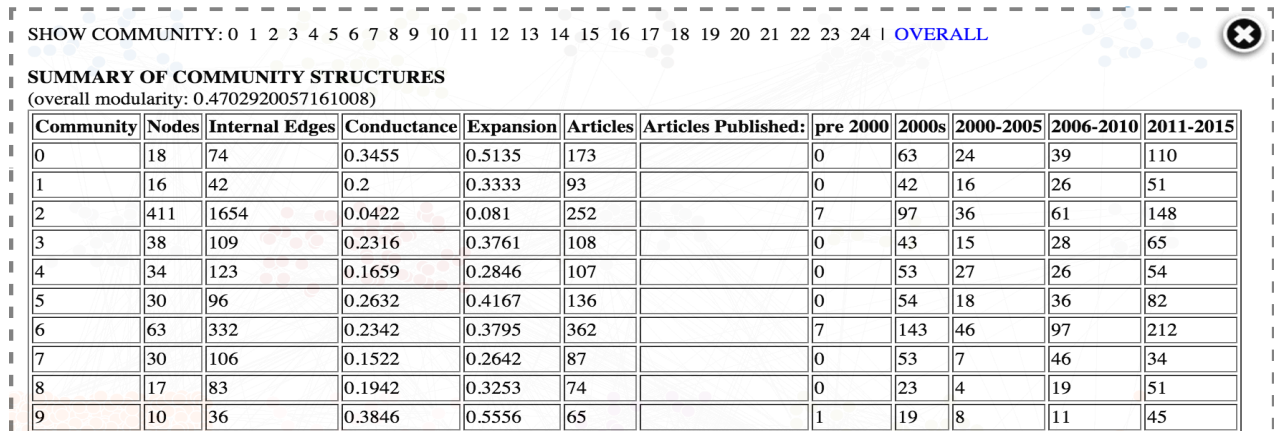


Fig. 4.25 Community Analysis Overall View To Compare Communities

4.6 MEDIA ANALYSIS TOOLS

In addition to the aforementioned tools that help users analyze politicians specifically from the vantage point of the people, bills, locations, and organizations with which they are associated in news articles produced from a set of news sources, the following tools allow the study of the news sources themselves. In our case study, the Austin American Statesman, the Dallas Morning News, the Houston Chronicle, the New York Times, the Texas Observer, and Texas Tribune were the sources used to find articles for the politicians studied.

4.6.1 MEDIA ANALYSIS TABLE VIEWS

Figure 4.26 shows the top results obtained after all of the articles discovered for the 247 politicians amongst our news sources were processed. This table shows for which politicians the system obtained the most articles and how those articles are distributed amongst our 6 news sources. For instance, we see the Greg Abbott, the current Republican Governor of Texas, had 4200 articles processed and of those 1558 were from the Austin American Statesman, 88 from the Dallas Morning News, 780 from the Houston Chronicle, 449 from the New York Times, 81 from the Texas Observer and 1224 from the Texas Tribune. The next in the list include both of Texas' federal senators, the Texas Speaker of the House, and the Lieutenant Governor of the State which is to be expected. Similarly Figure 4.47 shows the politicians for whom the least number of articles were processed, and they are composed of mostly junior State Representatives from largely rural districts and a judge elected in 2014 to the Texas Court of Criminal Appeals; none have appeared in the New York Times.






Who YOU ELECT. MEDIA ANALYSIS												
	NAME	LEVEL	POSITION	PARTY	DISTRICT	ARTICLES	AAS succes	DMN succes	HC succes	NYT succes	TXOB succes	TXTR succes
	GREG ABBOTT	STATEWIDE-ELECTED	GOVERNOR	REPUBLICAN		4200	1558	88	780	449	81	1224
	TED CRUZ	FEDERAL	SENATOR	REPUBLICAN		3618	639	88	693	817	65	1316
	JOE STRAUS	STATEWIDE	REPRESENTATIVE	REPUBLICAN	121	2678	622	96	829	81	72	978
	JOHN CORNYN	FEDERAL	SENATOR	REPUBLICAN		2621	498	94	784	465	86	694
	DAN PATRICK	STATEWIDE-ELECTED	LIEUTENANT GOVERNOR	REPUBLICAN		1796	640	86	756	127	84	103

Fig. 4.26 Media Analysis Table Sorted By Articles Successfully Processed. Top 5 Showing.






	DEWAYNE BURNS	STATEWIDE	REPRESENTATIVE	REPUBLICAN	58	6	0	1	0	0	2	3
	FOUR PRICE	STATEWIDE	REPRESENTATIVE	REPUBLICAN	87	5	0	2	1	0	0	2
	KEVIN YEARY	STATEWIDE-ELECTED	JUDGE COURT OF CRIMINAL APPEALS PL 4	REPUBLICAN		4	0	0	2	0	0	2
	BROOKS LANDGRAF	STATEWIDE	REPRESENTATIVE	REPUBLICAN	81	2	0	1	1	0	0	0
	J.D. SHEFFIELD	STATEWIDE	REPRESENTATIVE	REPUBLICAN	59	0	0	0	0	0	0	0

Fig. 4.27 Media Analysis Table Sorted By Articles Successfully Processed. Bottom 5 Showing.

In addition to statistics on the breakdown of articles that were successfully processed by the system, the table shown in Figure 4.28 also continues to the right and shows how many articles were “skipped” by each source (shown in light red), and then goes further to show the breakdown of why they were skipped in the remaining five colored sections.

These sections show articles returned from the internal site search engine for each news source that were not processed because they:

- were sports articles (green),
- were a duplicate result that was already processed in the same session (tan),
- returned a reference to a link with no associated text possibly due to a subscription paywall which does not allow access to the article without a paid subscription or registration (purple),
- returned an article that did not contain an exact reference to the politician used in the search query which could also be due to paywalls though not necessarily (marsh green)
- were determined to be a list, possibly a candidate listing, high school scholarship winners listing, etc that were skipped to avoid introducing noise into the system.

AAS skip	DMN skip	HC skip	NYT skip	TXOB skip	TXTR skip	AAS sport	DMN sport	HC sport	NYT sport	TXOB sport	TXTR sport	AAS dupe	DMN dupe	HC dupe	NYT dupe	TXOB dupe	TXTR dupe	NAME	AAS empty	DMN empty	HC empty	NYT empty	TXOB empty	TXTR empty	AAS nofour	DMN nofour	HC nofour	NYT nofour	TXOB nofour	TXTR nofour	AAS list	DMN list	HC list	NYT list
1836	10	112	87	5	11	7	1	2	0	0	0	3	0	18	2	0	0	GREG ABBOTT	1294	0	0	83	0	0	531	9	92	2	5	11	1	0	0	0
3355	6	148	191	8	179	8	0	2	0	0	0	2	0	22	0	0	0	TED CRUZ	3089	0	0	171	0	0	256	6	124	20	8	179	0	0	0	0
421	4	95	8	7	124	4	0	7	0	0	0	0	0	2	0	0	0	JOE STRAUS	216	0	0	8	0	0	201	4	84	0	7	124	0	0	2	0
850	5	108	434	4	88	1	0	0	0	0	0	0	0	10	6	0	0	JOHN CORNYN	698	0	0	425	0	0	151	5	98	3	4	88	0	0	0	0
941	14	81	22	3	1	118	8	4	13	0	0	1	0	3	1	0	0	DAN PATRICK	499	0	0	8	0	0	323	6	74	0	3	1	0	0	0	0

Fig. 4.28 Media Analysis Table Continued

Generally speaking, articles skipped for the New York Times and Austin American Statesmen were due to paywalls. Articles skipped for the Dallas Morning News, Texas Observer, and Texas Tribune were due to the politician not being explicitly found in the article text returned. It should be noted that the internal search engines for both the Dallas Morning News and Texas Observer return a maximum of 100 articles per query, and as such are less represented as a whole. Finally articles skipped for the Houston Chronicle were mostly due to duplicate results being returned by the site’s internal search engine though it also contained many sports articles that were skipped; the Austin American Statesman, New Dallas Morning News and New York Times additionally contained sports articles that were skipped, but to a much lesser extent.

Another view that allows for a more concise grouping of the top politicians as covered by the news sources in our case study is provided in Figure 4.29. In addition to providing a side-by-side ranked list, this “News Source Centric” view allows the user to hover over a politician to easily see how that person ranks in other news sources. In the figure it is easy to see how State Senator John Whitmire, whose district is in Houston, fairs among the different sources in relative article accounts. “Relative” here means that it is important to look at actual article counts since one table’s article counts may be low compared with another. For instance, there were more processed John Whitmire articles for the Texas Tribune than the Dallas Morning News. Abbreviations used in the level and position columns are provided at the bottom of the tool and all columns are sortable.

MEDIA ANALYSIS TOP POLITICIANS BY SOURCE																							
AUSTIN AMERICAN STATESMEN				DALLAS MORNING NEWS				HOUSTON CHRONICLE				NEW YORK TIMES				TEXAS OBSERVER				TEXAS TRIBUNE			
name	level	position	AAS	name	level	position	DMN	name	level	position	HC	name	level	position	NYT	name	level	position	TJOB	name	level	position	TXTR
GREG ABBOTT	SE	G	1558	RODNEY ELLIS	S	S	98	TOM CRADDICK	S	R	897	TED CRUZ	F	S	817	TOM CRADDICK	S	R	91	TED CRUZ	F	S	1316
KIRK WATSON	S	S	671	JOE STRAUS	S	R	96	GARNET COLEMAN	S	R	885	JOHN CORNYN	F	S	465	JOHN CORNYN	F	S	86	GREG ABBOTT	SE	G	1244
DAN PATRICK	SE	G	640	JOHN WHITMIRE	S	S	96	JOHN WHITMIRE	S	S	870	GREG ABBOTT	SE	G	449	DAN PATRICK	SE	G	84	JOE STRAUS	S	R	978
TED CRUZ	F	S	639	KIRK WATSON	S	S	96	SYLVIA GARCIA	S	S	841	JEB HENSARLING	F	R	187	GREG ABBOTT	SE	G	81	JOHN CORNYN	F	S	694
JOE STRAUS	S	R	622	MICHAEL MCCAUL	F	R	96	SYLVESTER TURNER	S	R	839	MICHAEL MCCAUL	F	R	159	RODNEY ELLIS	S	S	74	TOM CRADDICK	S	R	505
JOHN CORNYN	F	S	498	MARC VEASEY	F	R	95	JOE STRAUS	S	R	829	DAN PATRICK	SE	G	127	JOE STRAUS	S	R	72	RODNEY ELLIS	S	S	421
LLOYD DOGGETT	F	R	403	JANE NELSON	S	S	94	JOHN CULBERSON	F	R	829	LAMAR SMITH	F	R	104	JANE NELSON	S	S	71	KEN PAXTON	SE	AG	377
JOHN WHITMIRE	S	S	348	JEB HENSARLING	F	R	94	RODNEY ELLIS	S	S	824	PETE SESSIONS	F	R	90	DONNA CAMPBELL	S	S	68	JANE NELSON	S	S	361
JOHN CARTER	F	R	314	JOHN CORNYN	F	S	94	CAROL ALVARADO	S	R	822	JOE STRAUS	S	R	81	JOHN WHITMIRE	S	S	66	JUDITH ZAFFIRINI	S	S	330
MICHAEL MCCAUL	F	R	302	GREG ABBOTT	SE	G	88	JOHN CORNYN	F	S	784	LOUIE GOHMERT	F	R	66	TED CRUZ	F	S	65	KEL SELIGER	S	S	308
LAMAR SMITH	F	R	245	SAM JOHNSON	F	R	88	GREG ABBOTT	SE	G	780	KEN PAXTON	SE	AG	63	KEN PAXTON	SE	AG	60	TREY MARTINEZ	S	R	259
DONNA HOWARD	S	R	187	TED CRUZ	F	S	88	MICHAEL MCCAUL	F	R	780	GEORGE P. BUSH	SE	CGLO	60	KIRK WATSON	S	S	60	SYLVESTER TURNER	S	R	236
JANE NELSON	S	S	182	ERIC JOHNSON	S	R	87	KEVIN BRADY	F	R	778	KEVIN BRADY	F	R	54	SYLVESTER TURNER	S	R	57	HENRY CUELLAR	F	R	234
TOM CRADDICK	S	R	180	GLENN HEGAR	SE	CPA	87	LAMAR SMITH	F	R	758	HENRY CUELLAR	F	R	50	ROYCE WEST	S	S	53	JOHN WHITMIRE	S	S	230
TROY FRASER	S	S	175	MICHAEL BURGESS	F	R	87	DAN PATRICK	SE	G	756	JOAQUIN CASTRO	F	R	44	SID MILLER	SE	COA	46	KIRK WATSON	S	S	229
RODNEY ELLIS	S	S	171	DAN PATRICK	SE	G	86	TED POE	F	R	743	JOE BARTON	F	R	44	JUDITH ZAFFIRINI	S	S	41	TROY FRASER	S	S	227
JUDITH ZAFFIRINI	S	S	148	JEFF LEACH	S	R	85	SHEILA JACKSON LEE	F	R	704	JOHN WHITMIRE	S	S	38	TROY FRASER	S	S	41	GARNET COLEMAN	S	R	216
EDDIE RODRIGUEZ	S	R	134	PETE SESSIONS	F	R	85	TED CRUZ	F	S	693	SHEILA JACKSON LEE	F	R	35	GARNET COLEMAN	S	R	36	DAVID SIMPSON	S	R	210

Fig. 4.29 Media Analysis News Source Centric Table View

4.6.2 MEDIA ANALYSIS HEAT MAPS OF TEXAS HOUSE, SENATE AND FEDERAL DISTRICTS

In addition to the table view, there are heat maps showing the number of articles processed from different news sources for the districts pertaining to Texas House of Representatives, Texas Senate, and the Federal House. Figure 4.30 shows the map for the Texas House and in it the Texas

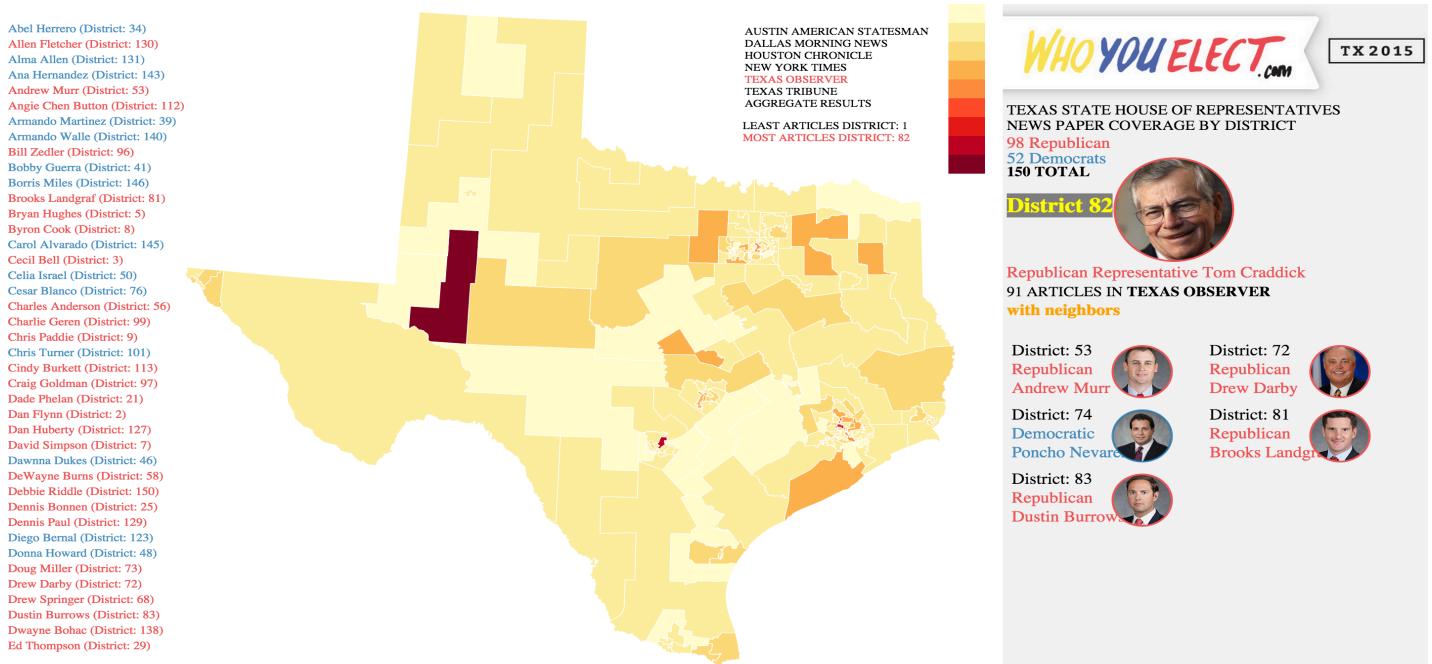


Fig. 4.30 Heat Map of Articles Published by the Texas Observer on the Texas House Representatives by District

Observer has been selected. The sources that can be selected are in the top upper right area next to a heat scale which means districts with the least articles published about their representative are

colored light yellow and those districts with the most articles are dark red. It is important to note that these heating color values are relative to the maximum articles published for a Texas Representative by the Texas Observer and not the aggregate whole of articles published at the State Representative level. The districts whose Representatives had the least number and the most number of articles published about them are automatically shown immediately below the sources. In this case Tom Craddick of District 82 had the most articles with 897, visible immediately below his picture, and District 58 with Representative J.D. Sheffield (not shown) had the least articles, zero in the case though to be fair there was a two way between himself and Dewayne Burns. It should be remembered though that there are 150 Texas House Districts as opposed to 31 Texas Senate Districts and as such the amount of media coverage for them as a whole is generally less than it is for Texas, and similarly Federal Senators.

Figure 4.31 shows the Heat Map of articles published by the Dallas Mornings on the Texas Senate. We can see that the districts pertaining to and near the Dallas metropolitan area in the north eastern area of the state are more well represented than districts farther away from it.

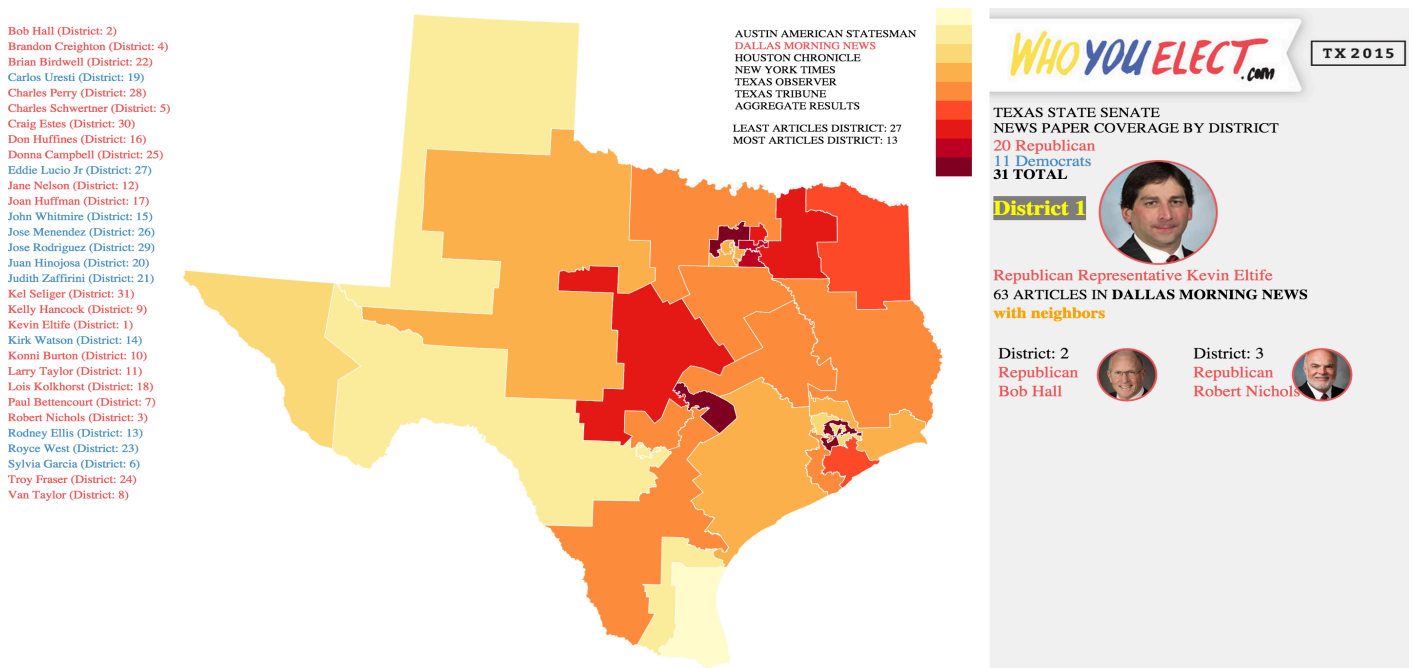


Fig. 4.31 Heat Map of Articles Published by the Dallas Morning News on the Texas Senate by District

Figure 4.32 shows the Heat Map of articles published by the Houston Chronicle pertaining to Texas members of the Federal U.S. House of Representatives. In it we can again validate the importance of locality in terms of media coverage as the area around Houston is darkly shaded, and the Federal Representative with the most articles published about them, Republican John Culberson with 829 articles, represents a district within Houston. In both Figure 4.31 and 4.30 we can see that the area around the state capitol of Austin in central Texas has considerable shading as expected.

Additionally there is an “Aggregate Results” and an “Aggregate Results Scaled” link among the list of sources that allows the user to view the news source article distributions as a collective whole, giving an overall average perspective of media coverage for the Texas House, Senate and Federal House. The Aggregate Results is a lump sum of all the sources per district, whereas the scaled version takes into account the differences in total amounts of articles obtained for each source. For instance, for the Federal House Representatives map, the sums of the articles produced by each news source overall is as follows: AAS= 1962, DMN= 1751, **HC= 13012**, NYT= 1191, TXOB= 363 & TXTR= 3111.

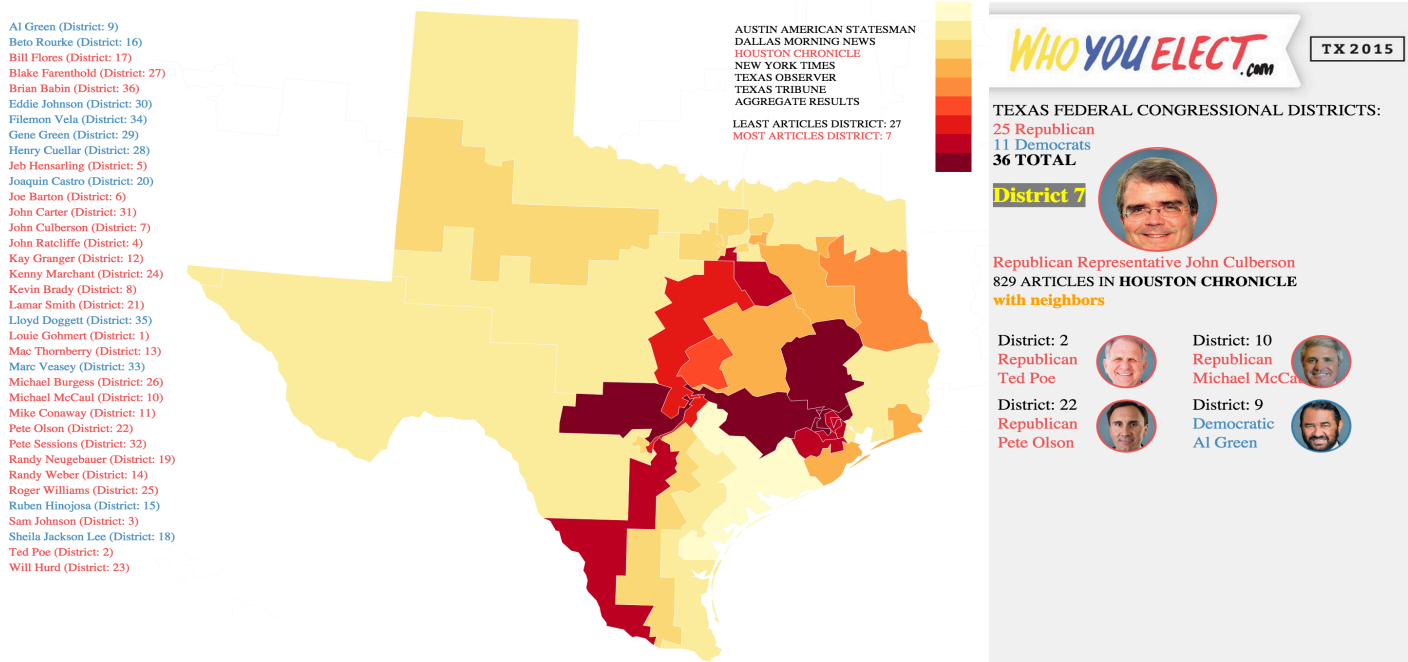


Fig. 4.32 Heat Map of Articles Published by the Houston Chronicle on the US House of Representatives by District

From these numbers we can see that the articles and subsequently the districts covered by the Houston Chronicle (HC) will dominate the aggregate lump view since 13012 articles pertaining to Federal Representatives were found there while the next highest count comes for the Texas Tribune with 3111 articles. Thus a possible solution for this disparity is to scale everything down to the least represented news source, the Texas Observer in the case, or conversely scale everything up to the most represented one, the Houston Chronicle, which is what the system does. Figure 4.33 shows the aggregate lump and aggregate scaled results for the Federal House of Representatives side by side. In it we can see some of the mass has been moved from the areas near Austin and Houston in the left “lump sum” version to areas near Dallas in the right “scaled” version.

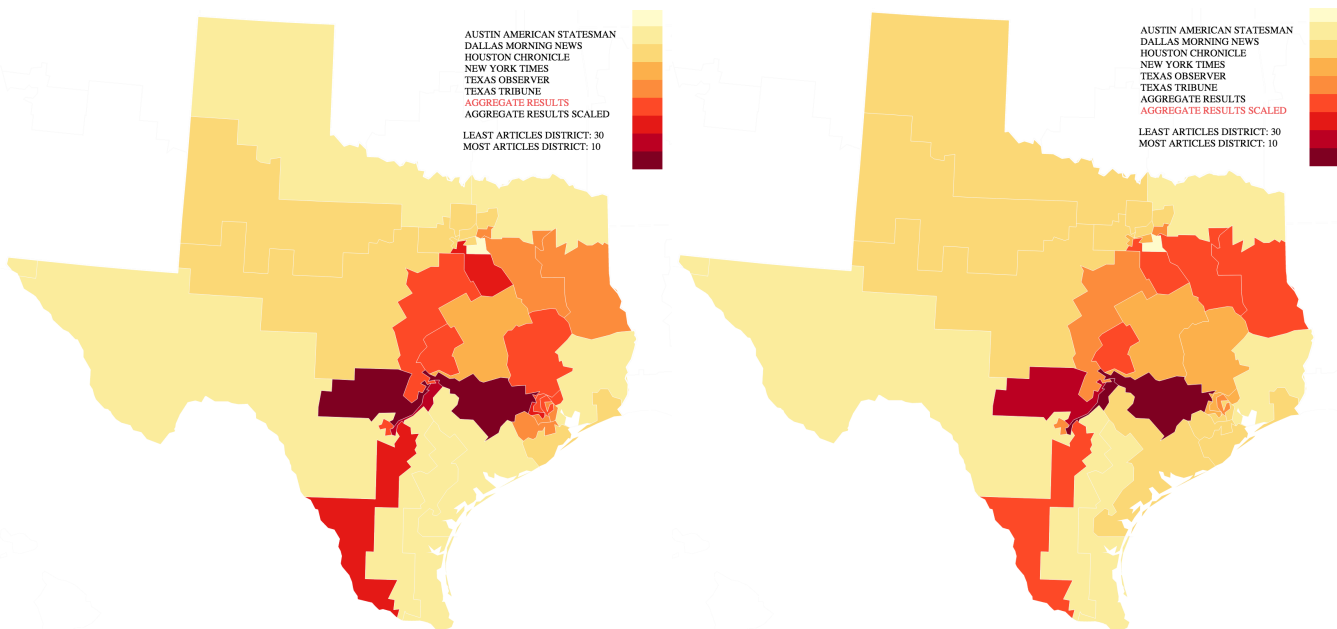


Fig. 4.33 Aggregate Lump (left) and Aggregate Scaled (right) maps of the US House of Representatives by District

See APPENDICES J, K, and L for all of the possible news source – congressional body combinations.

As a final technical note, it should be clarified that the Media tools presented were not entirely automatically generated whereas the Network Visualizations for each Politician are. The only thing necessary to automatically utilize the media tools is simply a small script to handle the merging of a JSON file containing the final “politician processing” system statistics along with another JSON file containing the politician’s meta data; their district, party affiliation, image URL, etc. This procedure generates the data file used by the tools, and was handled here via a small R script.

5. Additional Analyses

5.1 AUTOMATED SUMMARIZATION OF POLITICIANS

In an effort to get an overall picture of the political landscape of Texas given the data we obtained for our list of politicians, two approaches, one network centric and one based on text analysis through information retrieval and topic modeling techniques were considered. The former approach involves merging the “Extended” graphs for all the 247 politicians in our case study, and then running traditional network analysis on the large merged graph to determine the importance of politicians, organizations and other entities via various centrality measures. Although this method would inarguably produce interesting insight into the political landscape as a whole, it would do little in terms of providing a simple summary of the issues and topics surrounding each politician. To clarify, the network-based approach would show the important connections between the different entities of our graph, but as it is still entity-based, since it was derived from an entity co-occurrence based metric, it would not explicitly extract the issues central to a politician. This information could be inferred of course from the graph by seeing for instance that a politician was highly connected to a particular organization that advocated a particular issue or by seeing a politician was connected closely to a given bill involving an issue. A negative consequence of this approach however is that if an entity from an article was not recognized correctly by our NER solution during processing, a connection to it will not be established. This fault also exists largely independent of how many articles that entity appears in since it is expected that if the NER did not identify properly in one article, it is probable that it doesn’t in another. NER tools can be pre-trained of course so this can be improved. However, and more importantly, if an article contains just the thoughts of a politician on a given issue, but contains no explicit mentions to an organization, politician, location, bill or other entity, that information will be lost in the graph. For this reason and for time considerations, this merging network based approach is left for future work.

The second approach considered is based on textual analysis of the articles in which a politician occurs. By considering the combined text of the articles a politician occurs in as a single corpus, we can create a document term frequency matrix where each row is a particular article and the columns represent the terms (ie, single words, bigrams, trigrams) that occur in the corpus. We can then use this to calculate the TF-IDF (term frequency – inverse document frequency) of the corpus. This value gives a measure of how often a word appears through out the corpus with respect to the number of distinct articles it appears in. We can then use this TF-IDF value to filter out terms that appear all the time and provide little information or inversely those that occur vary rarely and could be noise. The art of deciding what stays in and what is filtered is very application specific and in this case was assessed by calculating a cutoff point based on a range of how many terms we desired appear overall in the filtered corpus. Then, using this refined corpus, we can run Latent Dirichlet Allocation [LDA 38] to uncover the “topics” (ie, issues, latent concepts) expressed within the articles for a single politician. The core assumption of LDA is that words in documents are generated by a mixture of topics with each topic itself being a distribution of words. In this algorithm, the number of topics is fixed initially and after it is run, each article in the corpus of a politician is assigned a topic. The topics themselves are composed of words that are ranked by those words with a higher probability of occurring. Because we have how many documents were assigned a given topic, we have the most frequent topics and can use these as a general summary of the issues surrounding a politician.

We ran LDA using 20 topics over each politician's articles set for single words, bi-grams (pairs of words) and tri-grams separately, and gathered the results in a searchable tool shown in Figures 5.0 and 5.1. The number of topics to begin with is imprecise and based on seeing how the results performed for different topic values for a handful of cases normally used in the literature (5, 10, 20, etc.) The work in [16] is particularly interesting because it discovers the number of topics to use without an initial user given one. In the end, the optimal value is specific to each politician in question, and has a lot to do with the analysis one wants to present. As long as a sufficient number of topics were used, it had a less overall effect than the thresholds used in determining which terms to keep in the refined corpus. Figure 5.0 shows the results of searching for the topic "farms"

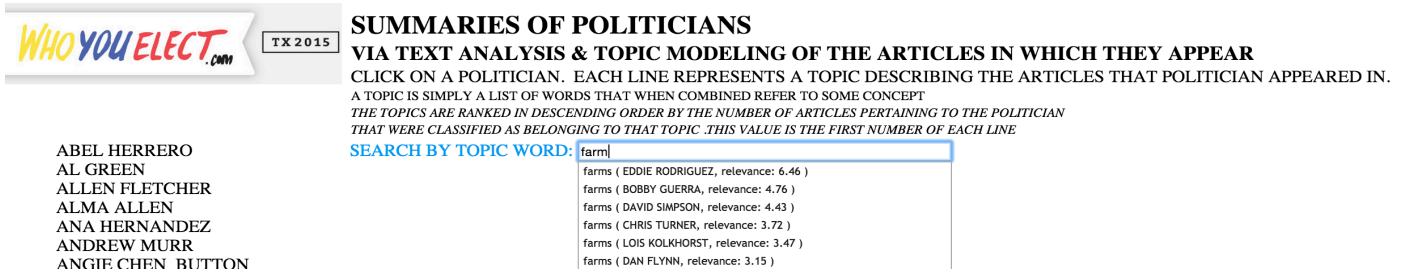


Fig. 5.0 Searching for Farms in Texas in single term case

In the figure, we can see a list of the politicians most relevant to the search query in question and see that Eddie Rodriguez has the highest relevance score. Upon selecting his name we see the all of the 20 topics found for Eddie Rodriguez along with the top 10 words for each, as shown in Figure 5.1a. The relevance score here is simply the percentage of articles for a politician that were assigned a topic with a term matching the query. In the figure we see the third topic, with relevance score 6.46 and highlighted in light red, matched the query and contains the words "food, markets, farms, regulations, milk" which all align with our prior examples showing Eddie's involvement in agriculture and farmer's markets issues. The relevance score means that 6.46% of the articles Representative Rodriguez occurred in were assigned this topic.

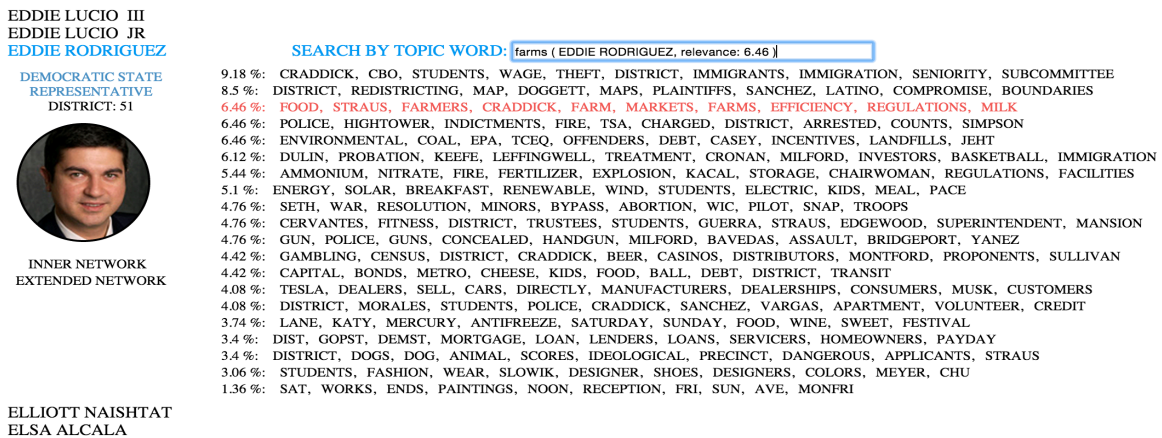


Fig. 5.1a Finding a Politician whose articles in the news generally pertain to a specific topic using single words

The above results were obtained using the "single words" version of tool that pertains to the LDA run using single word (i.e. 1-grams). Figure 5.1b shows the results using the bi-gram version of the tool. In it we see that "farms" is no longer in any of the topics, but the very first result which 15.65 of the documents involving Eddie Rodriguez are labeled with, includes the terms "Farmers Markets, Springdale Farm, Raw Milk, Cottage Food". Whereas the first version is good at pulling out "broad" single-word themes that are each generally cohesive as a concept, the bi-gram version extracts proper nouns, with topics that are less cohesive/interpretable, and as such should be regarded as method of creating a word cloud of issues around someone. Both methods still are quite noisy, highly sensitive

to the parameters used when filtering, and the process of deciphering topics, i.e. “labeling” them, is a manual process that often requires prior knowledge of the domain space, Texas politics in our case. Because we have the article URLs, and their associated metadata, that the topics are assigned to, we can present them to the user to assist in the understanding and labeling of topics (not shown).

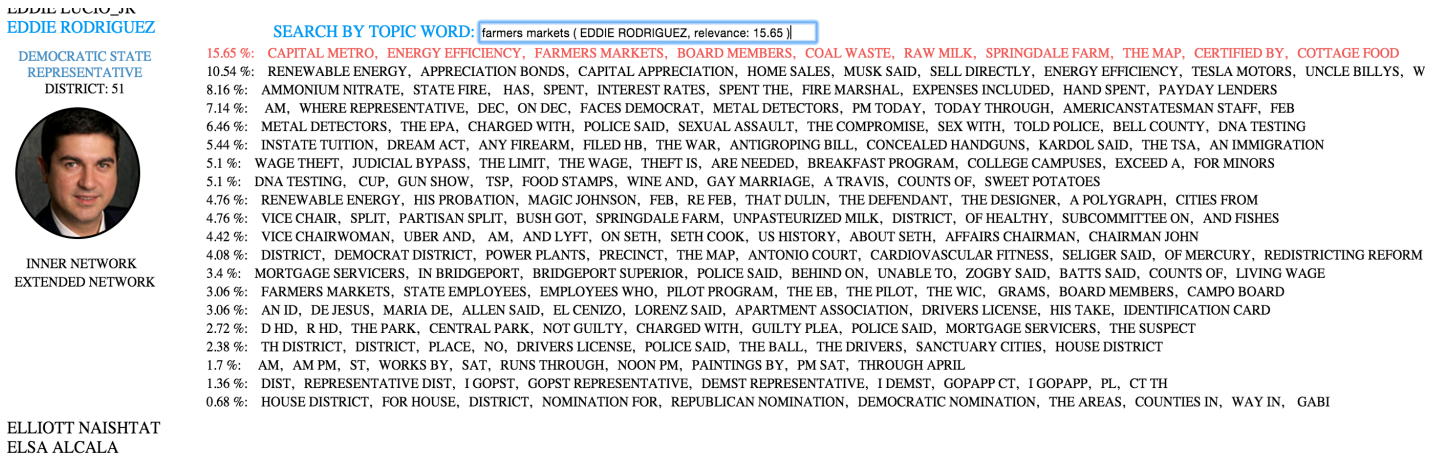


Fig. 5.1b Finding a Politician whose articles in the news generally pertain to a specific topic using bi-grams

The tri-gram topics (not shown) were more noisy, difficult to interpret and computationally more costly than the other two, providing little value outside of identification of 3-word proper nouns. Possible better approaches include trying to combine the three versions into one or instead focus on “phrase mining” techniques such as those shown in [17].

5.2 AUTOMATED SUMMARIZATION OF COMMUNITIES

If we return to Figures 4.24 and 4.25 presented in Section 4.5.4, we can see the details pertaining to a specific community detected within the context of the articles retrieved and processed for Eddie Rodriguez. The information provided in the figures provides details into who the central figures are within that community and additionally, the articles that are most prevalent within it. It would be useful however to have a way of automatically providing a description of the community at a higher level in order to give a more easily digestible global perspective of it. In that way we can then label all communities and allow the end user an additional perspective into the summaries as a whole.

One way to do that is by treating the articles of a given community collectively as a single corpus. We can then analyze the corpus using the same procedure we use to “summarize politicians” described in the prior section; namely an initial TF-IDF procedure to filter terms and reduce noise, followed by performing Latent Dirichlet Allocation to derive topics. The main difference here though is that since we know how many entities from a community occurred in each article found in the community, we can weigh these articles by their relative importance. For instance, in Figure 4.25, the article¹ including the most entities from the community will be weighed by a factor of 6. Additionally, we consider those articles with only one entity from the community as noise and exclude them from the corpus. As a proof of concept in Table 5.1 we show the results of applying this technique to the community from Figures 4.24 and 4.25 using single word, and the initial amount of topics being set to 5. The initial topics number is set low because if the number of communities is set

- Topic: 25.45% tax, strayhorn, rates, students, craddick, car, tesla, gambling, industry, cars
- Topic: 21.82% food, farmers, maps, markets, caucus, redistricting, doggett, plaintiffs, latino, map
- Topic: 18.79% craddick, gambling, interest, lenders, loans, loan, rates, tax, incentives, annual
- Topic: 18.79% energy, program, line, latin, market, sanchez, craddick, jobs, fashion, foreign
- Topic: 15.15% utility, uber, energy, tesla, rates, dealers, lyft, shoes, electric, stores

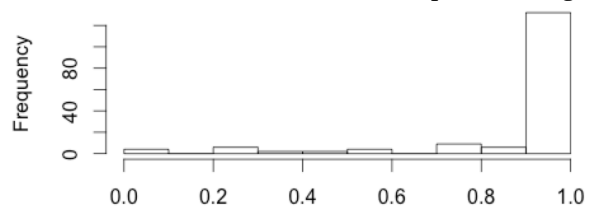
Table 5.1 Summarizing a Community Via Topic Modeling with topics composed of single terms

¹ <http://www.texastribune.org/2013/05/30/bipartisan-caucus-lays-groundwork-food-movement/>

sufficiently high, we would hope that each community would encapsulate at most a few topics though this again varies per community and is based on the number of entities, articles, expansion/conductance values and the overall modularity of the communities

We observe that due to the relatively low number of topics there is some overlap of concepts as highlighted in the second “single-words” topic in the above table, where the blue words refer to “agriculture” terms (farmers market, farm to table caucus) and the red refer to “redistricting” terms associated with articles discussing a lawsuit involving the re-drawing of district maps that would change U.S. Rep. Lloyd Doggett of Austin’s district to include Latino neighborhoods of San Antonio. Additionally for this proof of concept, we are not taking advantage of the stochastic nature of the results returned by LDA, which are in fact the likelihoods of a document belonging to any particular topic. The results above use just the most likely topic for assignment of documents, and as such lose the additional information provided in the posterior values of the model. This information will change the results quite a bit if a sizeable number of the articles have more than one topic with high probability.

To illustrate the point, in the above example case, the community contains 170 articles and we calculate the difference between the most probable topic and the next most for each document to produce the histogram in Figure 5.2. Here we see that most of the documents have a clear topic assignment, but this may not always be the case. Future work will account for it.



difference between max topic probability and second
 Fig. 5.2 Topic Assignment Verification for Example Case

5.3 MEDIA CENTRIC RESULTS OF CASE STUDY

The distribution of articles that were downloaded, successfully processed and skipped considering all the news sources and politicians collectively are presented in Figure 5.3. In it we see that on average 800 articles were downloaded for the 247 politicians we are studying, and of those about 370 were processed successfully on average per politician while 429 were skipped for reasons explained in Section 4.6.0. These numbers illustrate the volatile nature of using unstructured text data for analysis and why preprocessing, filtering, and verifying data quality is so vital to systems such as our own.

In order to delve more into where successfully processed articles were obtained, Figure 5.4 shows the distribution of processed articles by news source. In it, we see the relatively low numbers of articles coming from the Texas Observer and Dallas Morning News as previously mentioned, but we also see some outliers occurring in the Austin American Statesman, Texas Tribune and New York Times which pertain to Governor Abbott and US Senator Ted Cruz.

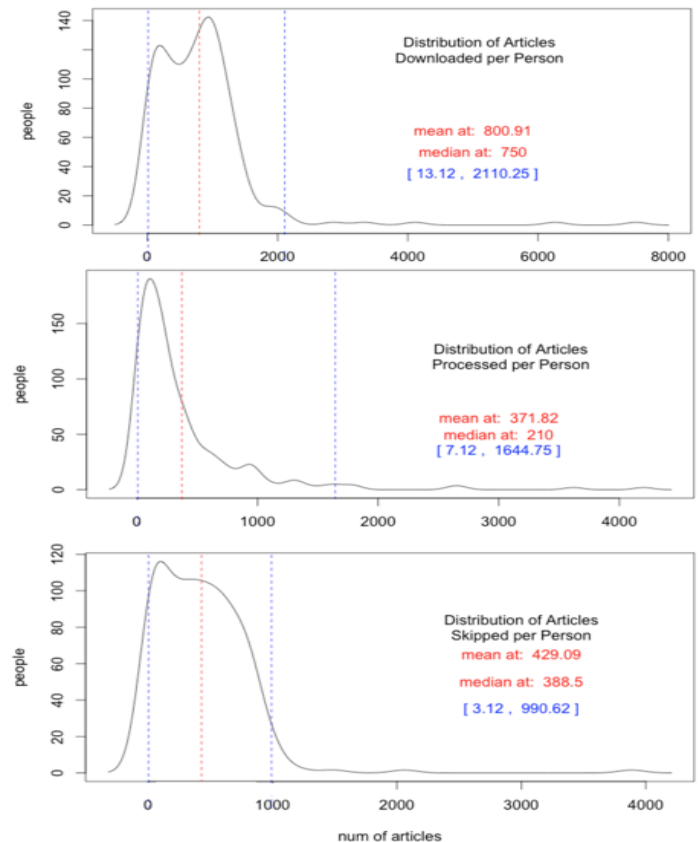


Fig. 5.3 Distributions of Articles for Case Study

To go further into the average and more appropriately the median use cases for the sources, since none of the distributions pictured in Figure 5.4 are remotely normally distributed, we look at some summary statistics for each of the news sources in Figure 5.5.

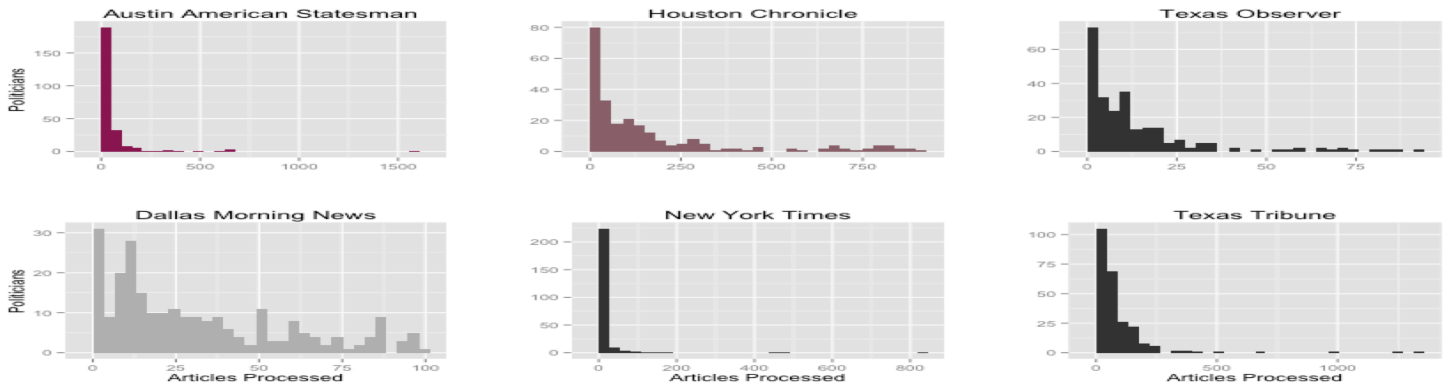


Fig. 5.4 Distributions of Articles Processed By News Source

Figure 5.5 doesn't provide much in terms of new insight except interestingly that the median articles processed per user are higher in the Dallas Morning News than expected

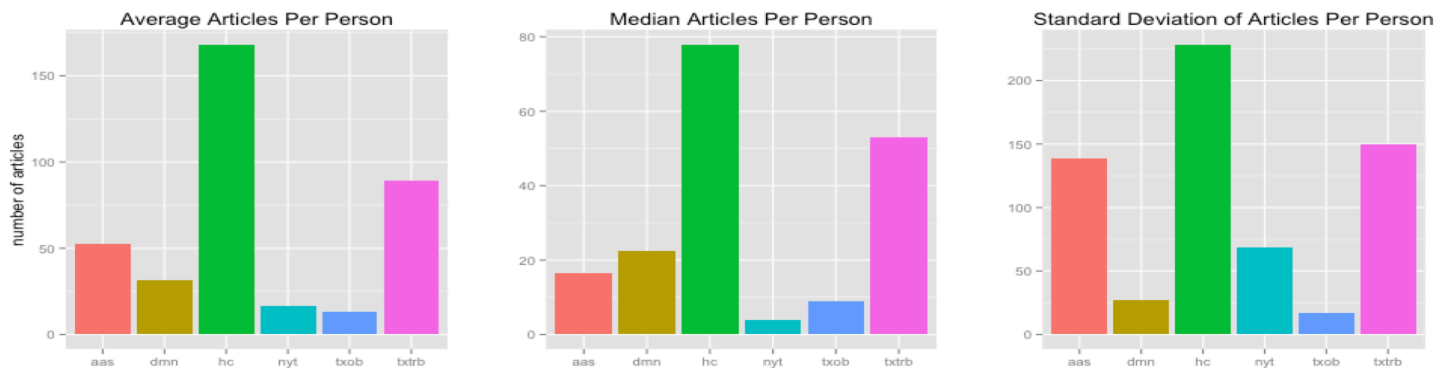


Fig. 5.5 Summary Statistics By News Source (in order: AAS, DMN, HC, NYT, TXOB, TXTRB)

In order to crudely determine if there is any bias in terms of reporting on Republicans versus Democrats as a whole among the different bodies we are studying we look at Figures 5.4 and 5.5.

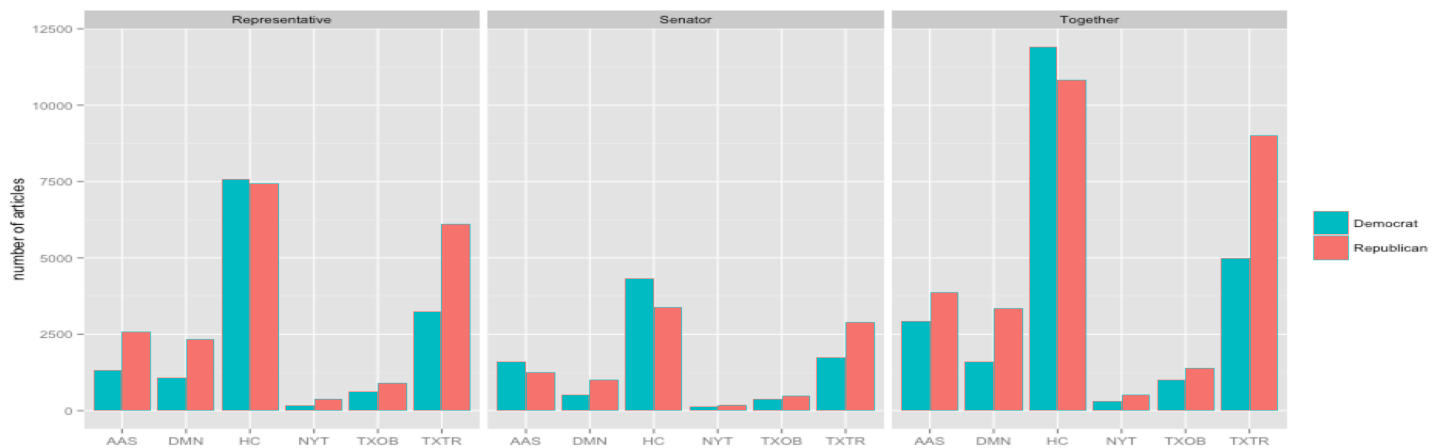


Fig. 5.4 Articles Processed For Texas House Representatives, Texas Senators, and the Texas Congress as a whole

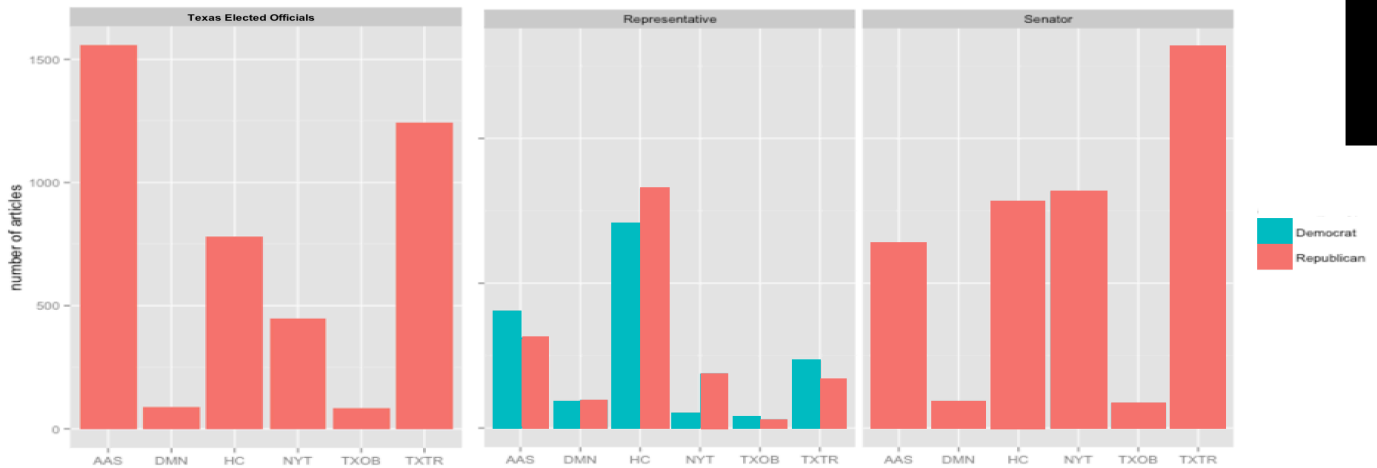


Fig. 5.5 Articles Processed For Texas “Elected Officials”, Federal Representatives, and Federal Senators

Figure 5.5 shows that we can not look for the existence on reporting differences by political affiliation for Federal Senators or the Texas “Elected Officials” group (outlined in Section 1.4) because they are all of the same political party. Among the 4 remaining bodies, nothing stands out as particularly biased though the Texas Tribune does tend to report more on Republican Texas House and Senate members as a whole. Given there are almost 50% more Republicans than Democrats in those legislative bodies, this does not seem out of the realm of being reasonable.

For the reader interested in seeing how each individual politician’s news coverage was distributed, a side-by-side comparison of all relative article distributions is found in APPENDIX H AND I. Each one additionally contains a subsequent “scaled” version that takes into account the news sources “total articles” disparity mentioned before.

6. Conclusions & Future work

In this work we presented a tool that generates real world political networks from user provided lists of politicians and news sites. The downloaded and processed article data for each politician were enriched them with data obtained from various open sources in order to facilitate verified politician meta-data and provide some structure to the unstructured article texts. In addition to the newly created networks, various visualizations, and tools that allow for the exploration of a politician and their environment were automatically generated. As of now the maps are derived from open-source geographic shape files, so making maps for new studies and areas consists in largely just finding or constructing the appropriate shape file and then simply using them instead of the Texas based shape files. We showed that the proposed “Combined” co-occurrence distance metric better determines the strength of the relationship between two actors in a graph as compared with the other more traditional metrics used in the literature. Additionally, the automated summarization tools created to extract topics and issues characterizing individual politicians are effective, but still quite noisy with room for improvement. The proof-of-concept use of topic modeling for labeling specific communities within a politician’s “extended” network is interesting and warrants further exploration and development. We have left as future work the performing of an extensive statistical study of the obtained graphs and media results, but have provided tools that allow a user to access these results individually or collectively through the images presented in APPENDIX H,I, J & K.

This tool can be utilized by researchers to easily construct and tailor real world graphs of their exact choosing instead of handing constructing them, resorting to using synthetic data, or paying for access

to large data sets. Additionally, this system is based on entirely open data and technology, mostly python and javascript. It does not require massive computational power or storage capabilities, all content was downloaded and processed locally on a single laptop, so NGOs and researchers without access to or technical expertise in big data can easily use it as well. The tool can be potentially used by journalists to discover new stories to write about it and serves as a simple and trustworthy mechanism for voter education. As with any system, it is only as good as the data provided to it so it is important for users to spend time thinking about the sources to include.

On a personal note, I had no idea of who many of these politicians were before creating these tools, but now by using the tools, particularly the individual star network, extended network and politician summarization ones, I am able to quickly gain a fairly good idea of any politicians I looked into, provided that sufficient articles were processed for them. That is very powerful and quite useful especially when attempting to understand a politician's history since given the pace of today's 24 hour news cycle, we largely only focus on what's immediately in front of this.

FUTURE WORK

In addition to the aforementioned future planned work in the paper, there is quite a bit of planned and desired future work including:

- Incorporating twitter information and tweets for politicians and other entities. Some politician's metadata already contain their handles, though it is lacking for many state level politicians.
- Incorporating data relating to State and Federal Bills from APIs available through OpenStates.org, and GovTrack.us to improve Bill coverage, especially for those not mentioned much in news articles, and to allow for better study of how politicians are connected through the bills they vote on.
- Incorporating campaign-funding data from OpenSecrets.org APIs to allow study of lobbying.
- Geographically calculating overlap of different Congressional body districts (Texas House, Texas Senate & Federal Congress), keeping in mind they evolve over time. Use as ground truth.
- Leveraging inactive politicians data we've already obtained from APIs, and noting that district maps change over time (especially on federal level) so need to evolve that overlap
- Leveraging Google Civic Information API for inclusion of Local and Citywide information
- Determining whether an article is predominantly about the politician being searched or whether it is in fact an article composed of many separate stories in which case the former's inferred relationships should hold more weight.
- Continuing development of automated filtering and merging tools to help with disambiguation of entities. As it stands now, the system downloaded articles for Eddie Rodriguez, the Texas House Representative, Eddie Rodriguez Jr., a Superior Court Judge of Connecticut, Eddie Rodriguez, a Cuban catcher of the New York Yankees, and Eddie Rodriguez, a fashion designer/businessman based in Miami. There are many relatively simple fixes to this issue; particularly the use of context words (location, political party, position, etc) to aid in resolution and Eddie Rodriguez was chosen specifically to illustrate this effect.
- Automating construction of and use of alias lists for gathering, merging and validating results
- Assessing the use of Stochastic Block Modeling and other community detections to allow for probabilistic community assignment and overlapping communities.
- Adopting use of multiplex paradigm by introduction of additional link types ("neighboring districts", "author of bill", "member of committee", etc.) for more robust network analysis.
- Incorporating event detection for periods of increased articles published for a politician and additionally using historic Wikipedia page view data, when applicable, to aid.
- Determining whether relationships are positive/negative/neutral, when applicable, through sentiment analysis and techniques such as co-sponsorship of bills, & tracking evolution over time.
- Adding public health, socio-economic, and voting history data by different district granularities.
- Allowing for zooming in and better placement of dense communities.
- Refactoring of NER solution for use with Catalan case study to show entirely non-English use. This will require some slight tweaks to the visualizations in particular for political affiliations.

- Creating comparison tool of network statistics for each individual extended networks to see how they compare against each other including article metadata and community information.
- Merging all extended views into a single large network for analysis.
- Improving web-scraping solution to simplify adoption of sources and refactoring of the manner in which text snippets are stored in the system for processing time and storage space improvement.
- Running a case study on a massive set of users and news sources to determine strength of scalability, and possible transfer of file storage and database solution.
- Developing mechanism for downloading, processing and adding new articles for existing politicians. This is an important, but rather straightforward step.

7. REFERENCES

- [1] Jesús Espinal-Enriquez, J. Mario Siqueiros-García, Rodrigo García-Herrera, and Sergio Antonio Alcalá-Corona. *A literature-based approach to a narco-network*. In *Social Informatics*, pages 97–101. Springer, 2014.
- [2] James H Fowler. Connecting the congress: *A study of cosponsorship networks*. *Political Analysis*, 14(4):456–487, 2006.
- [3] James H Fowler, Timothy R. Johnson, James F. Spriggs II, Sangick Jeon, Paul J. Wahlbeck, *Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court*, *Political Analysis* 15:324–346, 2007.
- [4] Justin H Kirkland. *The relational determinants of legislative outcomes: Strong and weak ties between legislators*. *The Journal of Politics*, 73(03):887–898, 2011.
- [5] Matt Thomas, Bo Pang, and Lillian Lee. *Get out the vote: Determining support or opposition from Congressional floor-debate transcripts*. Proceedings of EMNLP, pp. 327–335, 2006
- [6] Rokia Missaoui. *Tutorial on Mining Heterogeneous Information Networks*, EGC Toulouse, 2013.
- [7] Yizhou Sun, Jiawei Han: *Mining Heterogeneous Information Networks: Principles and Methodologies*. Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan & Claypool Publishers 2012.
- [8] Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, and Tamara Oellinger. *Building and Displaying Name Relations using Automatic Unsupervised Analysis of Newspaper Articles*. In Journées internationales d'Analyse statistique des Données Textuelles JADT 8es, 2006.
- [9] Theodosios Moschopoulos, Elias Iosif, Leeda Demetropoulou, Alexandros Potamianos, and Shrikanth Shri Narayanan. *Toward the automatic extraction of policy networks using web links and documents*. *Knowledge and Data Engineering, IEEE Transactions on*, 25(10):2404–2417, 2013
- [10] Fangbo Tao, George Brova, Jiawei Han, Heng Ji, Chi Wang, Brandon Norick, Ahmed El-Kishky, Jialu Liu, Xiang Ren, Yizhou Sun. *NewsNetExplorer: Automatic Construction and Exploration of News Information Networks*. SIGMOD 2014, June 22–27, 2014
- [11] David Newman, Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. *Analyzing entities and topics in news articles using statistical topic models*. In *Intelligence and Security Informatics*, pages 93–104. Springer, 2006.
- [12] David Newman, Sarvnaz Karimi, Lawrence Cavedon. *External Evaluation of Topic Models*, Proceedings of the 14th Australian Document Computing Symposium, 2009.
- [13] Elect Project, <http://www.electproject.org/2014g>
- [14] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, *Fast unfolding of communities in large networks*. In *J. Stat. Mech.* and arXiv:0803.0476 (2008)
- [15] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, Mason A. Porter. *Multilayer networks* In *Journal of Complex Networks* (2014) 2, 203–271
- [16] Jonathan Chang, Jordan Boyd Gaber, David M Blei, *Connections between the Lines: Augmenting Social Networks with Text*. KDD 2009.
- [17] Chi Wang, Marina Danilevsky, Nihit Desai, Yinan Zhang, Phuong Nguyen, Thrivikrama Taula, Jiawei Han. *A Phrase Mining Framework for Recursive Construction of a Topical Hierarchy*. KDD 2013.
- [18] Marina Danilevsky, Chi Wang, Nihit Desai, Jiawei Han, *Entity Role Discovery in Hierarchical Topical Communities*. KDDD 2013.
- [19] Keith Henderson, Brian Gallagher, Tina Eliassi-Rad, Hanghang Tong, *RoIX: Structural Role Extraction & Mining in Large Graphs* KDD'12, 2012,
- [20] Yang Li, Chi Wang, Fangqiu Han, Jiawei Han, Dan Roth, Xifeng Yan. *Mining Evidences for Named Entity Disambiguation*, KDD 2013.
- [21] Wei Shen, Jiawei Han, Jianyong Wang, *A Probabilistic Model for Linking Named Entities in Web Text with Heterogeneous Information Networks*. SIGMOD'14
- [22] Jana Diesner, *From Words to Networks: Extraction and Analysis of Semantic Network Data from Text Data* Tutorial presented at the Semantic Network Analysis Workshop at St. Petersburg State University, May 2013 .
- [23] Alejandro Llorente, Manuel Garcia-Herranz, Manuel Cebrian, Esteban Moro. *Social media fingerprints of unemployment*. arXiv:1411.3140v2 [physics.soc-ph] 19 Nov 2014
- [24] Manlio De Domenico, Mason A. Porter, Alex Arenas, *MuxViz: a tool for multilayer analysis and visualization of networks*. In *Journal of Complex Networks*. (2014)
- [25] Jiawei Han, Chi Wang and Ahmed El-Kishky. *Bringing Structure to Text*. KDD 14, 2014

- [26] Francisco Andrés Rodríguez Drumond. Mining parliamentary data & news articles to find patterns of collaboration between politicians & third party actors. Masters Thesis. Universitat Politècnica de Catalunya 2015.
- [27] Hristo Tanev. *Unsupervised learning of social networks from a multiple-source news corpus*. Multisource, Multilingual Information Extraction And Summarization, page 33, 2007.
- [28] Bruno Pouliquen, Hristo Tanev, and Martin Atkinson. *Extracting and learning social networks out of multilingual news*. In Proceedings of the Social Networks and Application tools workshop (Skalica, Slovakia, September). Citeseer, 2008.
- [29] Rudi L Cilibrasi and Paul MB Vitanyi. *The google similarity distance*. Knowledge and Data Engineering, IEEE Transactions on, 19(3):370–383, 2007.
- [30] Matt Levin, *This is how efficiently Republicans have gerrymandered Texas congressional districts*. Houston Chronicle. May 6, 2015. <http://www.chron.com/news/politics/texas/article/This-is-how-badly-Republicans-have-gerrymandered-6246509.php>
- [31] *The State of Gerrymandering*. Available at <http://svds.com/gerrymandering/>
- [32] Russell C. Weaver. *Gerrymandering Politics Out of the Redistricting Process*: Berkeley Planning Journal. September 2012. Available at: <http://ced.berkeley.edu/bpj/2012/09/gerrymandering-politics-out-of-the-redistricting-process-toward-a-planning-revolution-in-redrawing-local-legislative-boundaries/>
- [33] John N. Friedman and Richard T. Holden. *Towards a Theory of Optimal Partisan Gerrymandering*. Harvard University, 2005. Available at: http://bcep.haas.berkeley.edu/papers/friedman_20080225.pdf
- [34] John Mackenzie, *Gerrymandering and Legislator Efficiency*. University Of Delaware. 2006. Available at: <http://www.udel.edu/johnmack/research/gerrymandering.pdf>
- [35] Marta Arias & R. Ferrer-i-Cancho. *Community structure in networks*. Universitat Politècnica de Catalunya. 2014. Available at: <https://www.cs.upc.edu/~CSN/slides/07communities.pdf>
- [36] Newman, M. E. J. (2006). "Modularity and community structure in networks". Proceedings of the National Academy of Sciences of the United States of America 103 (23): 8577–8696.
- [37] Fortunato, Santo. "Community detection in graphs." Physics Reports 486.3 (2010): 75-174.
- [38] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. The Journal of Machine Learning Research, 3:993–1022, 2003.

8. APPENDIX

APPENDIX A0:

The short comings of the EMM system (emm.newsbrief.eu) [8] is shown via the following sample. Using their online advanced search tool, searching for articles containing "Garnet Coleman" in the Houston Chronicle between May 2001 and May 2015, the system returns 6 results between 2014 and 2015, whereas doing a search directly against the Houston Chronicles internal search system returns 762. The system also does not contain many important papers for Texas politics (Austin American Statesmen, Texas Observer, Texas Tribune, etc.)

APPENDIX A1: sample API return results for a Senator using the OpenStates.org API

```
{
  "last_name": "Taylor",
  "updated_at": "2015-05-09 20:27:44",
  "nimsp_candidate_id": "112117",
  "full_name": "Larry Taylor",
  "+district_address": " 174 Calder Road, Suite 116\nLeague City, TX 77573\n(281) 338-0924",
  "first_name": "Larry",
  "middle_name": "",
  "district": "11",
  "id": "TXL000349",
  "state": "tx",
  "votesmart_id": "25471",
  "party": "Republican",
  "all_ids": [ "TXL000349", "TXL000167", "TXL000423" ],
  "leg_id": "TXL000349",
  "active": true,
  "transparencydata_id": "33194a0192b14e4fbecd8787da18b95b",
  "photo_url": "http://www.legdir.legis.state.tx.us/FlashCardDocs/images/Senate/small/A1030.jpg",
  "+capital_address": " P.O. Box 2910\nAustin, TX 78768\n(512) 463-0729",
  "url": "http://www.legdir.legis.state.tx.us/MemberInfo.aspx?Chamber=S&Code=A1030",
  "country": "us",
  "created_at": "2010-06-19 03:51:42",
  "level": "state",
}
```

```

    "nimsp_id": "4028",
    "chamber": "upper",
    "offices": [
      { "fax": null, "name": "Capitol address", "phone": "512-463-0111",
        "address": "P.O. Box 12068, Capitol Station\nAustin, TX 78711",
        "type": "capitol", "email": null },
      { "fax": null, "name": "District address", "phone": "281-485-9800",
        "address": "6117 Broadway, Suite 122\nPearland, TX 77581",
        "type": "district", "email": null }
    ],
    "suffixes": ""
  },
}

```

The fields in bold are what the system primarily leverages.

APPENDIX A2: sample API return results for a Representative using the GovTrack.us API

```

{
  "caucus": null,
  "congress_numbers": [
    114
  ],
  "current": true,
  "description": "Representative for Texas's 6th congressional district",
  "district": 6,
  "enddate": "2017-01-03",
  "id": 43268,
  "leadership_title": null,
  "party": "Republican",
  "person": {
    "bioguideid": "B000213",
    "birthday": "1949-09-15",
    "cspanid": 5248,
    "firstname": "Joe",
    "gender": "male",
    "gender_label": "Male",
    "id": 400018,
    "lastname": "Barton",
    "link": "https://www.govtrack.us/congress/members/joe_barton/400018",
    "middlename": "Linus",
    "name": "Rep. Joe Barton [R-TX6]",
    "namemod": "",
    "nickname": "",
    "osid": "N00005656",
    "pvsid": "27082",
    "sortname": "Barton, Joe (Rep.) [R-TX6]",
    "twitterid": "RepJoeBarton",
    "youtubeid": "repjoebarton"
  },
  "phone": "202-225-2002",
  "role_type": "representative",
  "role_type_label": "Representative",
  "senator_class": null,
  "senator_rank": null,
  "startdate": "2015-01-06",
  "state": "TX",
  "title": "Rep.",
  "title_long": "Representative",
  "website": "http://joebarton.house.gov"
}

```

APPENDIX B:

This editing consists of filling in variable values for the URL of the internal news site search mechanism, the location of the DIV that stores the search results, the name of the DIVs holding the title, the date, and the location of where a "next" page link would exist if any.

After a template has been filled in for each source, there exists a single folder within "generate_network/data/" for each news source where the results for our article search for each politician by news source will be placed. For instance the articles returned for the politician Eddie Rodriguez for news source N, will be stored in the folder Eddie_Rodriguez within the N folder. That politician folder will contain a JSON file, in this case "links-Eddie_Rodriguez.json" that contains the article URLs, dates, and titles associated with the internal search results from news source N. Additionally, the politician folder will also contain a JSON file for each article URL found in "links-Eddie_Rodriguez.json" that contains the title, url, date and article text found for that URL. The following is an explanation of the folder structure in generate_network/data/ that contains article results:

```
austinamericanstatesman/ - source based on beautifulsoup
dallasnewssearch/       - source based on phantomjs
delete_entity.py        - script to remove an entity from our source folders
houstonchronsearch/     - source based on beautifulsoup
jsons/                  - folder to store individual entity networks by generate_networks.py
numbers/                - folder that stores local counts of all articles by source for a given entity
nytimes/                - source based on phantomjs
pickles/                - folder that contains tmp data for use in case of need
                        - to restart a given entities network generation
show_article_numbers_for.py - script to get downloaded article counts for a given entity
template-beautifulsoup  - template version for beautiful soup
template-phantomjs     - template version for phantomjs
txobserver/            - source based on phantomjs
txtribune/              - source based on beautifulsoup
```

APPENDIX C: Sample JSON Relation between a politician POL1 and entity ENT1

```
{ 'term1':POL1name , 'term1_id':POL1id,
  'term2':ENT1name, 'term2_id':ENT1id,
  'type':'same sentence', 'text_snippet': TEXT,
  'sentence_num':sentn}
```

Here the **type** may be "same sentence", "near", or "same article". In the case that it is "near", there is also a subtype field to specify whether it occurred "prior" or "post" to the first term.

APPENDIX D: JSON object constructed after Verification Step 3.9

```
{'url':url,
 'entities': ent,
 'relations': relations,
 'date': art['date'],
 'mongoid':moid,
 'num_sentences':len(sentences)}
```

APPENDIX E: Truncated Sample Metrics For Garnet Coleman. The 0 index refers to English articles, while the 1 index refers to Spanish articles found. The entire set of metrics gathered during processing can be found in data/numbers/article-processing-results.json.

```
"Garnet Coleman": {
"0": {
  "firstpass": [{
    "articles": 1432,
    "completed": 1417,
    "processingtime": 2685.1298100000104,
    "totalsents": 63481,
```



```

"totalrels": 3282958,
"totalinsts": 2653,
"totalents": 43997,
"BILL": 6355,
"LOCATION": 7279,
"ORGANIZATION": 7385
"MISC": 6843,
"PERSON": 8064,
"politician": 0,
"Austin American Statesman": {
  "skipped": 0,
  "success": 104
},
"DALLAS MORNING NEWS": {
  "skipped": 0,
  "success": 52
},
"HOUSTON CHRONICLE": {
  "skipped": 0,
  "success": 959
},
"New York Times": {
  "skipped": 0,
  "success": 21
},
"TEXAS OBSERVER": {
  "skipped": 0,
  "success": 37
},
"Texas Tribune": {
  "skipped": 0,
  "success": 244
},
},
{
  "http://www.chron.com/news/houston-texas/article/Governor-decides-he-ll-fill-rest-of-seats-on-TSU-1611870.php": {
    "insts": 1,
    "processingtime": 1.0470820000000458,
    "rels": 33,
    "ents": 6,
    "result": "completed",
    "sents": 22
  },
},
.....
"1": {
  "firstpass": [
    {
      "articles": 15,
      "completed": 15
      "processingtime": 73.80452599999999,
      "totalsents": 635,
      "totalrels": 44276,
      "totalinsts": 29,
      "totalents": 771,

```

```

"BILL": 75,
"LOCATION": 75,
"MISC": 75,
"ORGANIZATION": 75,
"politician": 0,
"PERSON": 75,
"HOUSTON CHRONICLE": {
  "skipped": 0,
  "success": 15
},
},
{
  "http://www.chron.com/default/article/Sesi-n-legislativa-especial-culmina-con-nota-2078478.php": {
    "insts": 2,
    "processingtime": 7.31278999999995,
    "rels": 3562,
    "ents": 56,
    "result": "completed",
    "sents": 42
  },
  ...
}

```

APPENDIX F: WORKERS DEFENSE PROJECT – EDDIE RODRIGUEZ text results

Relationship between **Eddie Rodriguez** and **Workers Defense Project**

5 co-occurrences found in 4 articles.

(AAS: 1 | HC: 0 | DMN: 0 | TXR: 2 | TOB: 0 | NYT: 1 | ALL: 4)

Cities work make wage theft prosecution priority

2011-10-31 | TEXAS TRIBUNE

In Austin, the **Workers Defense Project**, a workplace justice group, is collaborating with state Representative **Eddie Rodriguez**, D-Austin, who sponsored the bill in the House, to set up a meeting to talk with Austin's police chief and the district and county attorneys about making wage theft an enforcement priority.

Misclassification bills likely dead session

2013-05-16 | TEXAS TRIBUNE

"Of course it's an immigration issue, but it's hidden under this umbrella of playing on an even field," said state Representative **Eddie Rodriguez**, D-Austin. He supported Davis' bill and helped vote it out of committee. But he said that big businesses who oppose the measure had ultimately won out. Emily Timm, a policy analyst with the **Workers Defense Project**, said opponents in the construction industry worried changing the status quo would damage their bottom lines.

same article

Legislator backs off bill to ban living wage requi

2013-04-05 | AUSTIN AUSTIN STATESMAN

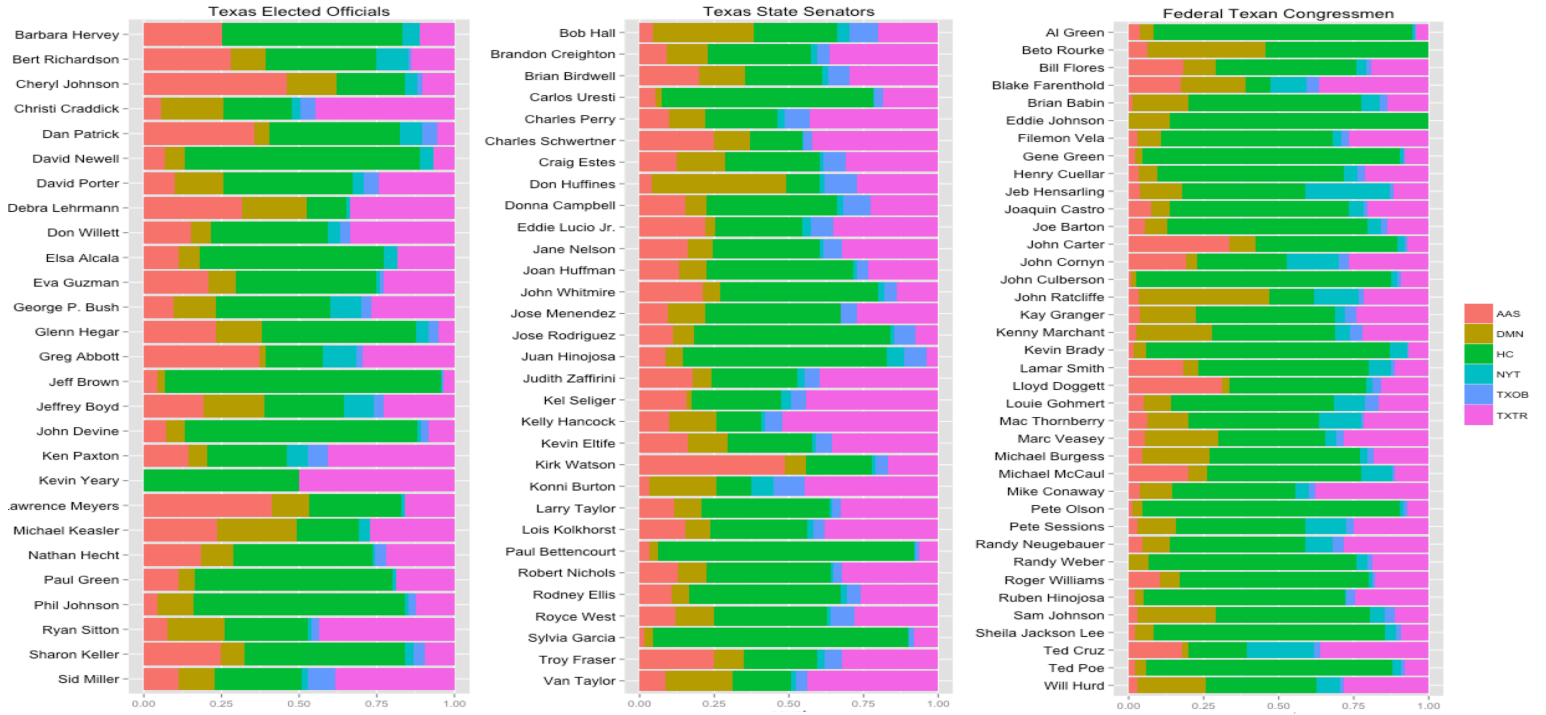
State Senator Kirk Watson, D-Austin, and state Representative **Eddie Rodriguez**, D-Austin, said such matters should be settled by the council, not the Legislature. Sheets' decision to focus on other legislation pleased Austin Interfaith, a coalition of congregations and social justice groups that has been pushing for the living-wage requirement. At the organization's request, members of its Dallas-area counterpart and representatives of the Dallas affiliate of the **Workers Defense Project** met with Sheets, asking him to drop his legislation and citing, among other reasons, a desire for local control in such matters, said Kurt Cadena-Mitchell, an Austin Interfaith leader.

Organizations work to enforce wage theft bill aimed at immigrants employers.html

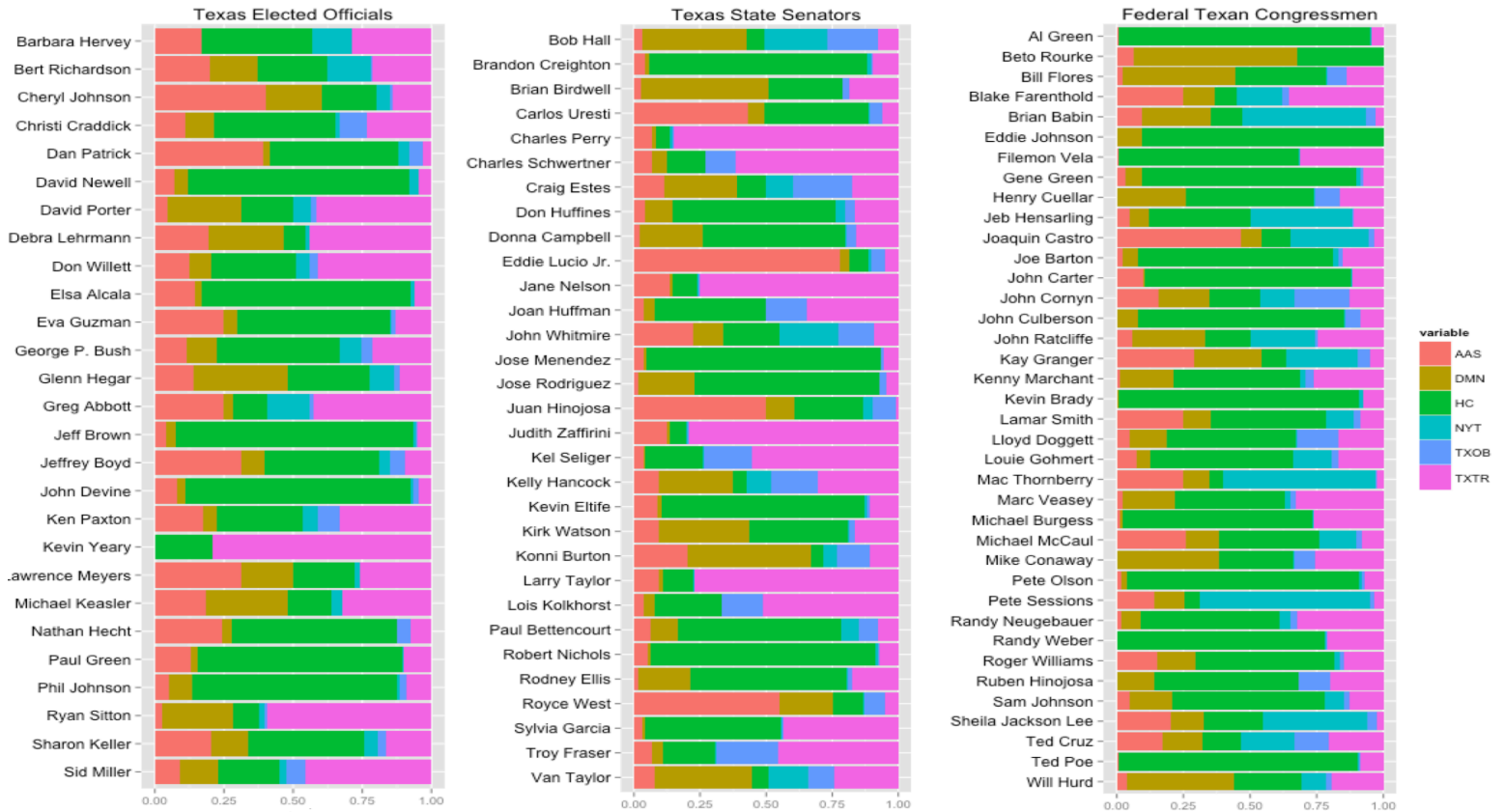
2011-10-30 | NEW YORK TIMES

In Austin, the **Workers Defense Project**, a workplace justice group, is collaborating with Representative **Eddie Rodriguez**, Democrat of Austin, who sponsored the bill in the House, to set up a meeting with Austin's

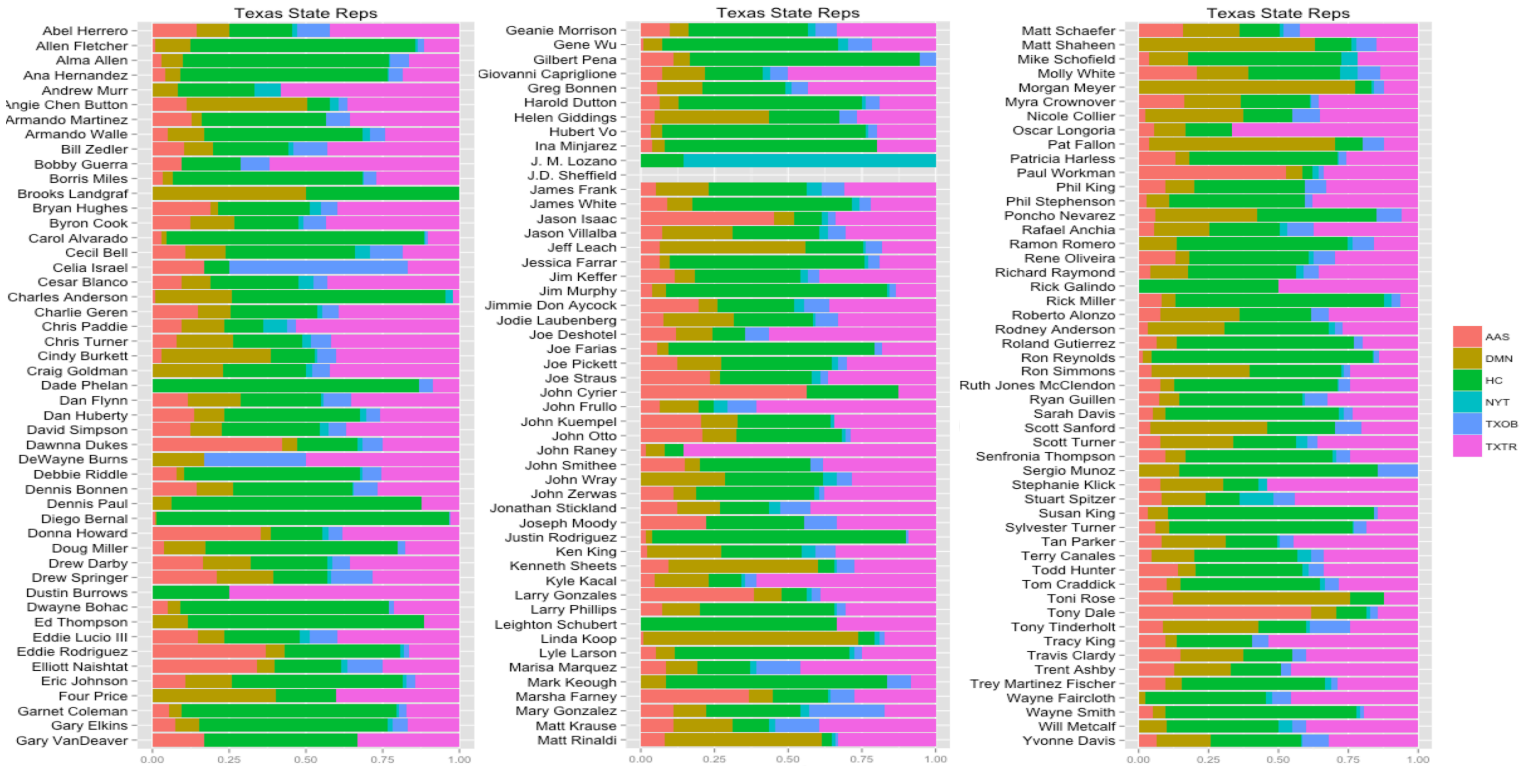
APPENDIX H: INDIVIDUAL POLITICIAN RELATIVE ARTICLE VALUES. TEXAS & US SENATE, SELECT TEXAS ELECTED OFFICIALS



APPENDIX H: INDIVIDUAL POLITICIAN SCALED ARTICLE VALUES. TEXAS & US SENATE, SELECT TEXAS ELECTED OFFICIALS



APPENDIX I: INDIVIDUAL POLITICIAN RELATIVE ARTICLE VALUES. TEXAS HOUSE OF REPRESENTATIVES 2015



APPENDIX I: INDIVIDUAL POLITICIAN SCALED ARTICLE VALUES. TEXAS HOUSE OF REPRESENTATIVES 2015

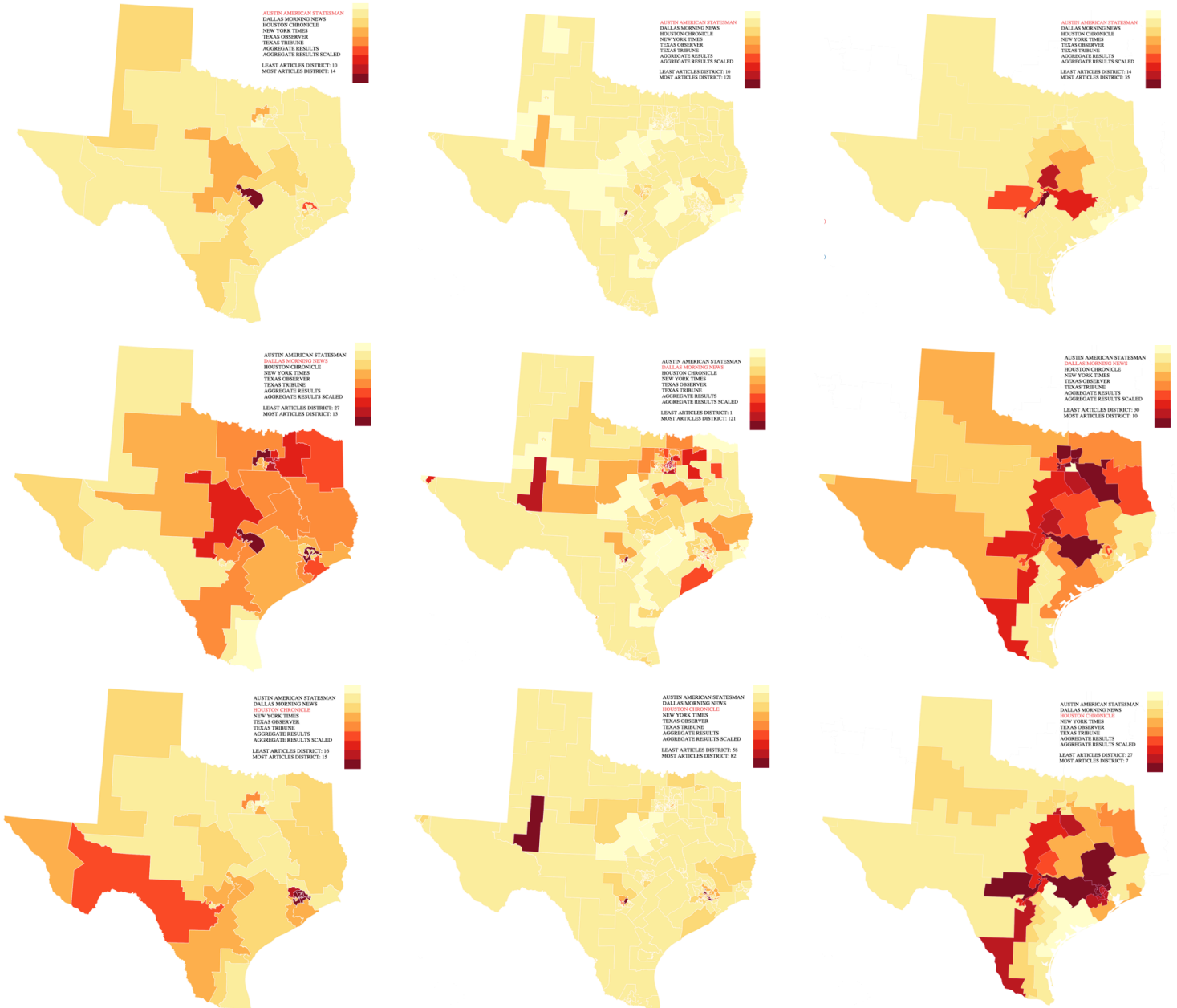


APPENDIX J: TEXAS SENATE, TEXAS HOUSE & FEDERAL ARTICLE DISTRIBUTIONS BY DISTRICT FOR AUSTIN AMERICAN STATESMAN, DALLAS MORNING NEWS, AND HOUSTON CHRONICLE. COLUMNS ARE LEGISLATIVE BODIES AND ROWS ARE NEWS SOURCES

i. Texas Senate District Maps By Source

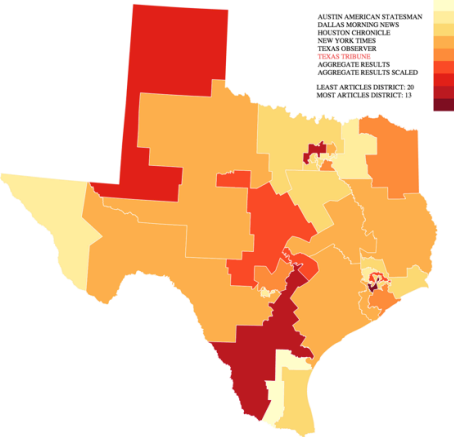
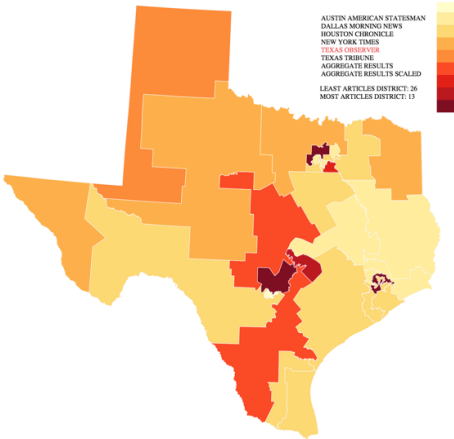
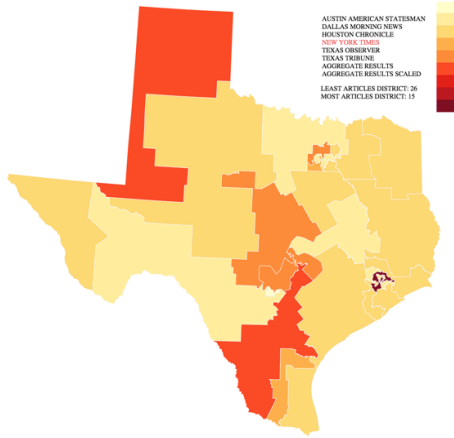
ii. Texas House District Maps By Source

iii. Federal Texas House Districts By Source

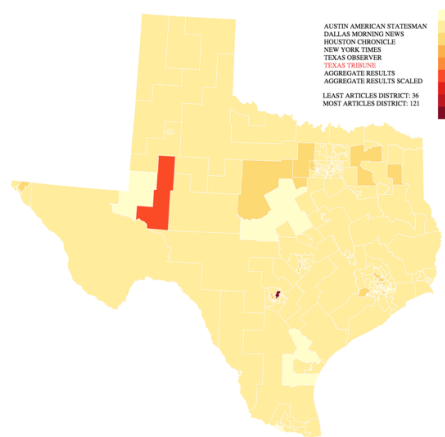
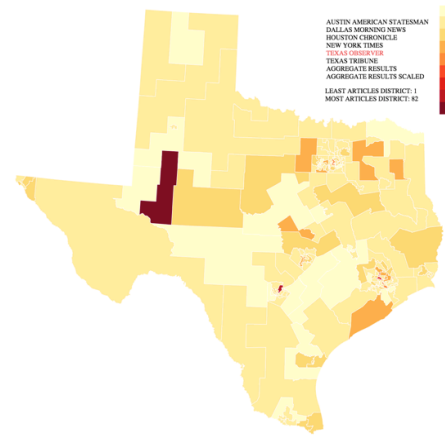
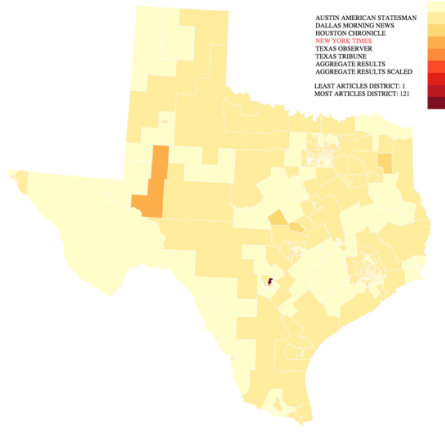


APPENDIX K: TEXAS SENATE, TEXAS HOUSE & FEDERAL ARTICLE DISTRIBUTIONS BY DISTRICT FOR THE NEW YORK TIMES, TEXAS OBSERVER AND TEXAS TRIBUNE. COLUMNS ARE LEGISLATIVE BODIES AND ROWS ARE NEWS SOURCES

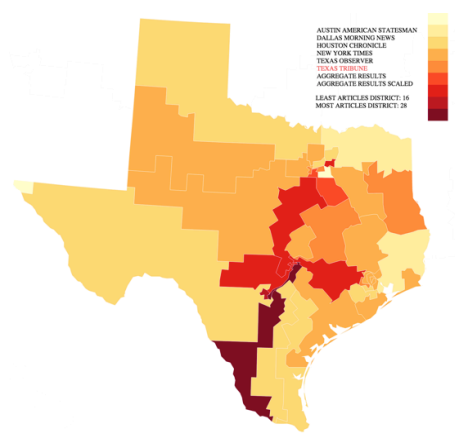
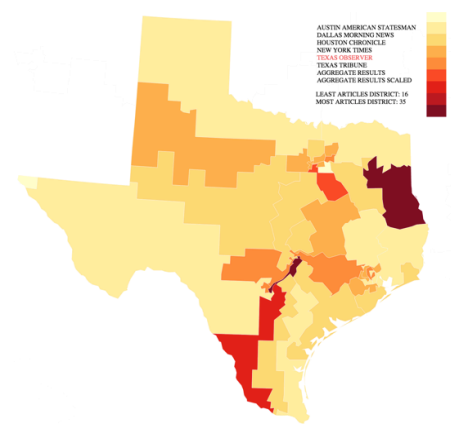
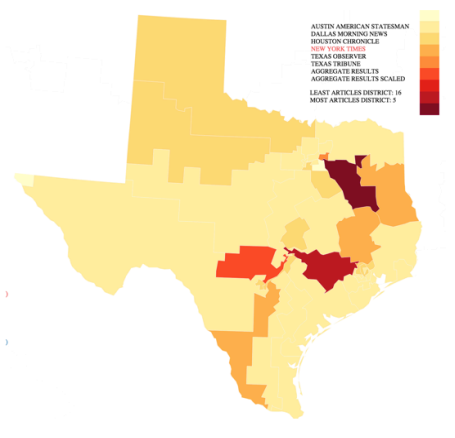
i. Texas Senate District Maps By Source



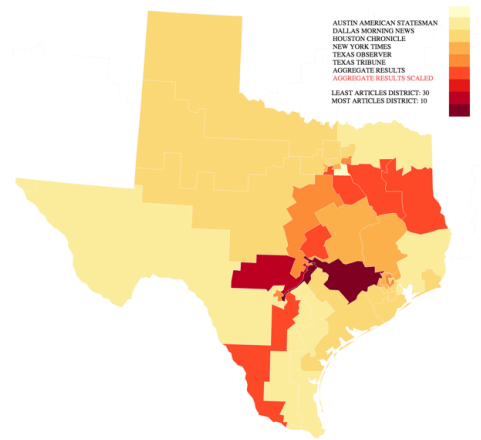
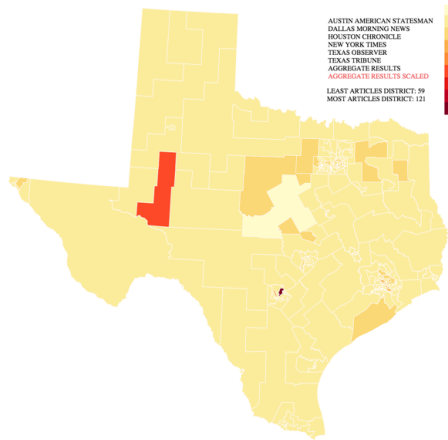
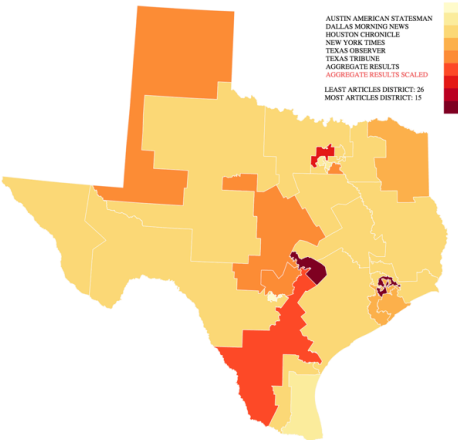
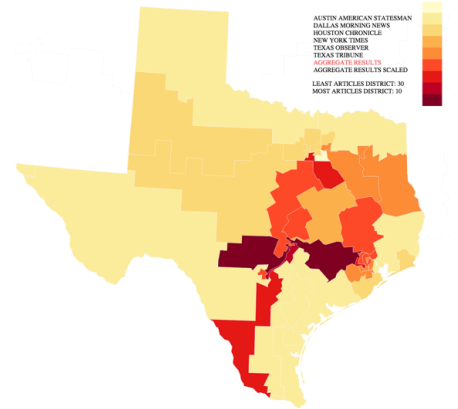
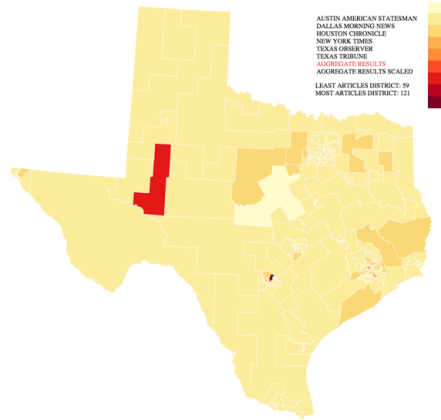
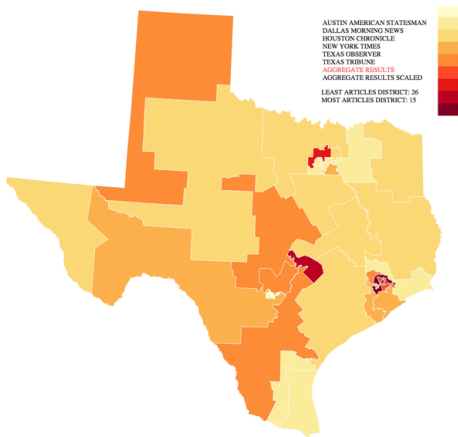
ii. Texas House District Maps By Source



iii. Federal Texas House Districts By Source



APPENDIX L: TEXAS SENATE, TEXAS HOUSE & FEDERAL ARTICLE DISTRIBUTIONS BY DISTRICT FOR AGGREGATE RESULTS AND AGGREGATE RESULTS SCALED. COLUMNS ARE LEGISLATIVE BODIES AND ROWS ARE NEWS SOURCES



i. Texas Senate District Maps By Source

ii. Texas House District Maps By Source

iii. Federal Texas House Districts By Source