Automated Construction and Analysis of Political Networks via Open Government and Media Sources

Diego Garcia-Olano, Marta Arias, and Josep Lluís Larriba Pey[1]

Universitat Politècnica de Cataluyna (UPC) DAMA & LARCA groups

Abstract. We present a tool, www.whoyouelect.com, to generate real world political networks from user provided lists of politicians and news sites. Additional output includes visualizations, interactive tools and maps that allow a user to better understand the politicians and their surrounding environments as portrayed by the media. As a case study, we construct a comprehensive list of current Texas politicians, select news sites that convey a spectrum of political viewpoints covering Texas politics, and examine the results. We propose a "Combined" co-occurrence distance metric to better reflect the relationship between two entities. A topic modeling technique is also proposed as a novel, automated way of labeling communities that exist within a politician's "extended" network.

Keywords: political science, network science, text mining, open data

1 Introduction

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 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[ElE]. 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.

1.1 Problem description and background

There exists a vast literature on the use of network analysis to study important social and political phenomena [FJ1,FJ2,KHJ,MT1,LA1]. 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[EJ1]. In it, the authors perform text mining, partially by hand and partially automated, on a single book about the subject ”Los SeÑora's del Narco” by Anabel Hernandez and then derive a visual network from the actors and links discovered. This co-occurrence based network provides a simple and concise high-level view of the actors and entities involved in the text. This general idea of combining text mining and network science to both summarize content while allowing for the discovery of interesting relationships can be applied on 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 would be of great use for voters, journalists and researchers alike.

1.2 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 ”WhoYouElect.com” that could take one or many candidate names as input, along with a list of online new sources, and then retrieve all the articles pertaining to the candidates from them. 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, commonly referred to as an ego network, that allows a user to view the entities (politicians, businesses, etc.) most associated with a candidate along with the articles and context in which they co-occurred, and an ”extended” world view that is a global view that shows links between a politician and associated entities, and also the links between those entities themselves; thus allowing for the detection of communities and other traditional network analysis measures.

1.3 Description of Case Study: Texas Politics

In 2015. the Texas State Congress, composed of the House of Representatives and Senate, has 181 members, 31 senators and 150 representatives. On the federal level within the US Congress, Texas has 38 total members, 2 senators and 36 representatives. 27 other State Level Elected Officials are also studied including the Governor, Lieutenant Governor, etc. In total we have 246 Texas politicians composed of 74 Democrats and 172 Republicans in our study.

1.4 Organization of Document

The organization of this document is as follows. Chapter 2 will cover related works and additional background. Chapter 3 will describe the system structure and methods by which we automatically gather, store and process articles for politicians. Chapter 4 focuses on presenting the network analysis tools developed and available at WhoYouElect.com. Finally Chapter 5 contains conclusions.

2 Related Work

The types of graphs we will be constructing are undirected heterogeneous networks with weighted edges. Heterogeneous means the graph will contain different node types. In our case, nodes can be 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 [MR1,SY1]. 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.

For the moment we are only considering one "co-reference" relationship type, meaning one possible edge between each pair of nodes whose weight will be based on various distance metrics and as such do not find ourselves in the multiplex context which is more adept for studying complex networks. For an extensive examination of the field, uses, and visualization tools see [KM1,DM1]. Adapting the system to include more than one edge type by considering other features in addition to co-reference or perhaps leveraging linked datasets could be interesting. 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 [LY1] 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 [SW1]. 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 [CJ1,WC1]. An overview of Latent Dirichlet Allocation (LDA) or alternatively Latent Semantic Indexing (LSI) to produce topic models is found in [ND1]. 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.

There exist a good deal of prior work that has focused on deriving information networks from unstructured news and other web texts[PB1,MT1,TF1,ND1]. Two introductions and overviews on the process of deriving networks from text may be found [DJ1,HJ1]. 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”[MT1], while others rely on organizations such as the European Media Monitor [PB1] 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 MITs 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 weakness. 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[TF1,ND1]. Similarly, [MT1] 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 Yahoos 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 publicly 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 [PB1,MT1] or topical modeling [TF1,ND1] to infer relationships of interest. The textual context presented in [MT1] 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 [CR1] and evaluated them against a small hand-crafted Policy Network. Although interesting this approach does not scale as an evaluation approach. The works of [TH1,PB2] like [PB1] 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 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.

3 Automated Construction of Graphs

In our case study, we constructed the graphs for 246 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 code for constructing graphs discussed here is available on github[1]. The general outline of the process for constructing graphs follows.

3.1 Adding Politicians from Open Government Sources

We leveraged the Sunlight Foundations Openstates.org API[2]to obtain a list of both active and inactive members of Texas congress returned as JSON and saved them locally. Federal Representatives are obtained 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 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.
We obtained a list of non Congressional, state wide elected officials (Governor, Lieutenant Governor, etc.) making use of the Secretary of State of Texas website[3]. Finally, in order to get federal representatives and senators for the state of Texas, we leveraged the GovTrack.us API[4]. 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.

[1] http://github.com/diegoolano/whoyouelect/tree/master/generate_network
[2] http://openstates.org/api/v1/legislators/?state=tx&active=true
[3] http://www.sos.state.tx.us/elections/voter/elected.shtml
[4] https://www.govtrack.us/api/v2/role?current=true&state=TX

3.2 Adding News Sources

Now that the entities have been added to our database, we add the news sources for use in our case study. A subset of Texas newspapers with the highest circulation, the Dallas Morning News, Houston Chronicle, and the Austin American Statesman, representing the spectrum of conservative, centrist and progressive areas within Texas were selected along with two sites, the Texas Observer and Texas Tribune which focus on Texas politics and issues. In addition, the New York Times was selected to provide an outside context.

3.3 Data Acquisition by Template Modification

For each desired source we need to gather articles pertaining to our politicians list. Weve created two template web scrapers, one based on the Python package BeautifulSoup which is fast, but doesn't handle pages rendered with javascript well and another based on the Web.Selenium Python package which is a headless browser that works in all cases, but is slower. Each template version is a folder containing two files, one file 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, and another 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. This procedure of template editing, one folder per source, is straightforward for a web developer with experience to setup and test however simplifying this procedure requires considerable effort via inferring structure of pages probabilistically and could lead to potential loss of information.

3.4 Running & Storing the Web Search Results for Active Entities

This script then takes the article JSON results and 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 MongoDB. Currently the system is tailored for articles in English, but also works for Spanish and any languages supported by MITIE[5], an open source Named Entity Recognition engine that given a document identifies possible named entities and tags them as ”Organization”, ”Location”, ”Person” or ”Miscellaneous”.

In addition to these tag types, during processing we look for and include two additional 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.

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

Although it could be read as such, we do not mean to imply that Politicians are not People. The Bill entity type refers to legislation and we use simple naive 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.

3.5 Processing Article Results per Active Entity

Following the completion of the above, the script [6] which is the heart of the processing step of the retrieved articles for a given entity is called. The script queries MongoDB to gather all of the articles found for the person, pre-filters out "sports" and uninteresting results, and then processes the remaining articles 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. If an invalid date is seen, assign a default 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 trivial, but in actuality is of critical importance because we are dealing with uncertain input.

3. **Run named entity recognition** over the sentences to return the named entities and tags per sentence and verify that the candidate we are searching for appears. This step also identifies the cases where a pay wall exists for a given article and the returned article contains only the first few lines of the actual article that does not include the name of the candidate himself in which case we skip the article entirely.

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"

[6] https://github.com/diegoolano/whoyouelect/tree/generate_network/get_articles_for_person_then_find_and_save_relations.py

entities discovered in the article.

5. **Filter entities further.**[7] This optional step utilizes a user created look up table of "uninteresting" entities to be excluded from the overall graph. For instance, in our case the entities for "Texas Legislature" and "Congress" and other entities which provide little distinguishing capacity are excluded.

6. **Check if an article is a candidate listing** or something similar in which case discard it as noise. 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. Even in the case where the listing was a list of candidates for offices or winners of statewide races, it would naively create a great deal of spurious relationships where they didnt exist.

7. **Discovery of relationships**[8]. At this point, the system goes through each found instance of the politician being processed and for each, gathers all the entities that occur on the same sentence as it, within 3 sentences of it, or farther than three sentences away and stores the results in a matrix.

8. **Verification and construction of relationships**. We take the matrix from the prior step and construct the relationships for each instance of the main politician found with other entities. All the relationships that have been created with respect to the main politician will be used later to construct the "star" individual view graph. The system constructs the relationships between all the other non-main politician entities for the "extended" network if the initial script was called with the "include_larger" flag.

9. **Save intermediate results**. We now have a results object 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 doesnt 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

[7] The optional filter entities, verification/construction of relationships and saving to intermediate object is done mostly in https://github.com/diegoolano/whoyouelect/tree/generate_network/verify_and_save_relations.py.

[8] Discovery of relations functionality is predominantly in https://github.com/diegoolano/whoyouelect/tree/generate_network/entity_funcs.py

took a large amount of time; ”large” meaning anything over 5 seconds. It was observed that these articles were ones where the *number of candidate instances found multiplied by the number of entities and the number of sentences* was found to be usually greater than a certain threshold (8000 is the threshold we ended up using). Thus we developed an additional function specifically optimized to these larger instances by pre-computing the hashes for lookups.

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

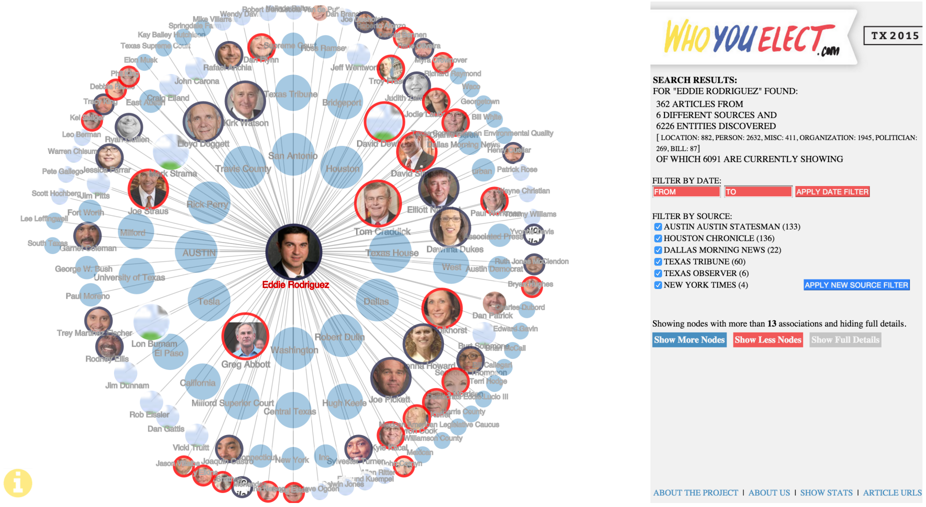
Once all articles have been processed, we save the resulting article metrics to our global metrics file and then save our global result set and some other variables into a python ”pickle” file, essentially an internal tar file for python that we compress to save space. This later step was done largely 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 Graphs

4 Overview of Who You Elect Visualization Tools

4.1 Individual ”Star” Ego Network View

When ”Inner Network” is selected from the ”Table of Contents” view, the politicians derived ego network 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 of whether the co-occurrences were in the same sentence, near sentences or farther. The following figure shows the graph produced for Democratic State Representative of District 51, Eddie Rodriguez, who we will be using as a running example. 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.
Clicking on the center node, brings up the information associated with Eddy Rodriguez including name, party, position, district number, wiki description (if found), a map of the politician’s district and a list of the entities he is most associated with. Clicking on any non-central node, shows a list of the articles in which that node and Eddy Rodriguez appear (including date, article title, and URL) which maybe filter by media source, and highlights the sentences where they co-occur together in that article. Along the bottom and left hand side of



WHO YOU ELECT.com TX 2015

SEARCH RESULTS:
FOR "EDDIE RODRIGUEZ" FOUND:
362 ARTICLES FROM
6 DIFFERENT SOURCES AND
6226 ENTITIES DISCOVERED
[LOCATION: 882, PERSON: 2632, MISC: 411, ORGANIZATION: 1945, POLITICIAN: 269, BILL: 87]
OF WHICH 6091 ARE CURRENTLY SHOWING

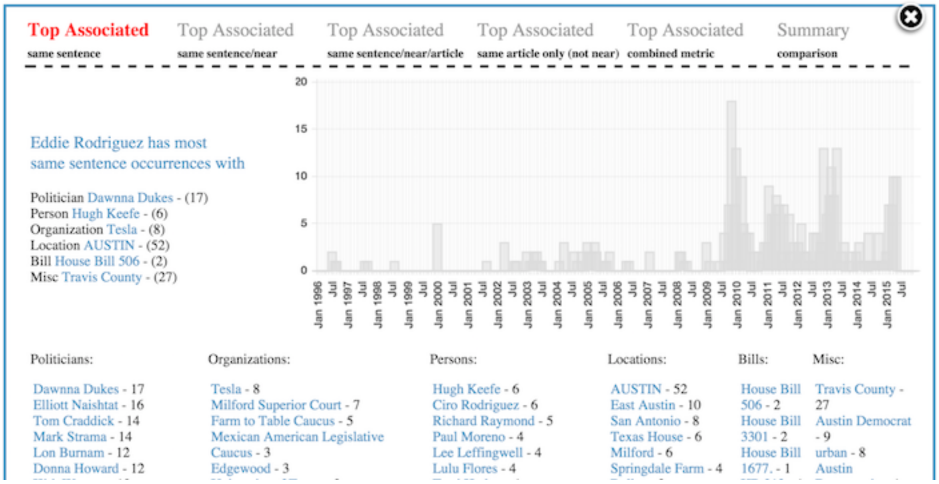
FILTER BY DATE:
FROM | TO | APPLY DATE FILTER

FILTER BY SOURCE:
☑ AUSTIN AUSTIN STATESMAN (133)
☑ HOUSTON CHRONICLE (136)
☑ DALLAS MORNING NEWS (22)
☑ TEXAS TRIBUNE (60)
☑ TEXAS OBSERVER (6)
☑ NEW YORK TIMES (4)
APPLY NEW SOURCE FILTER

Showing nodes with more than **13** associations and hiding full details.

Fig. 1. Individual Star Network Landing Page for Representative Eddie Rodriguez

the screen, the user can filter the results to only show Politicians or Organizations, etc and also change the distance metric used for calculating edge weights. 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. The functionality for "Show Stats" is of particular interest and when clicked provides both a time line histogram of how many articles were published by month for the central politician along with a comprehensive view of their most associated entities under different distance metrics. 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, the prior figure has the Top Associated "same sentences" tab currently selected, 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). We observe that "Dawna Dukes" is the Politician with the most same sentence occurrences with Eddie Rodriguez. The 2nd and 3rd tabs refer to distance metrics which additionally take into account co-occurrences within 3 sentences ("near" in the 2nd case) and farther (co-occurrence within the same article, but farther than 3 sentences away in the 3rd case) 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 combina-



Eddie Rodriguez has most same sentence occurrences with

Politician Dawnna Dukes - (17)
Person Hugh Keefe - (6)
Organization Tesla - (8)
Location AUSTIN - (52)
Bill House Bill 506 - (2)
Misc Travis County - (27)

Politicians:	Organizations:	Persons:	Locations:	Bills:	Misc:
Dawnna Dukes - 17	Tesla - 8	Hugh Keefe - 6	AUSTIN - 52	House Bill 506 - 2	Travis County - 27
Elliott Naishtat - 16	Milford Superior Court - 7	Ciro Rodriguez - 6	East Austin - 10	House Bill 3301 - 2	Austin Democrat - 9
Tom Craddick - 14	Farm to Table Caucus - 5	Richard Raymond - 5	San Antonio - 8	House Bill 1677. - 1	urban - 8
Mark Strama - 14	Mexican American Legislative Caucus - 3	Paul Moreno - 4	Texas House - 6		Austin
Lon Burnam - 12	Edgewood - 3	Lee Leffingwell - 4	Milford - 6		
Donna Howard - 12		Lulu Flores - 4	Springdale Farm - 4		

Fig. 2. ”Show Stats” Screen for Texas Representative Eddie Rodriguez

tion 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. The proposed ”Combined” metric is defined as:

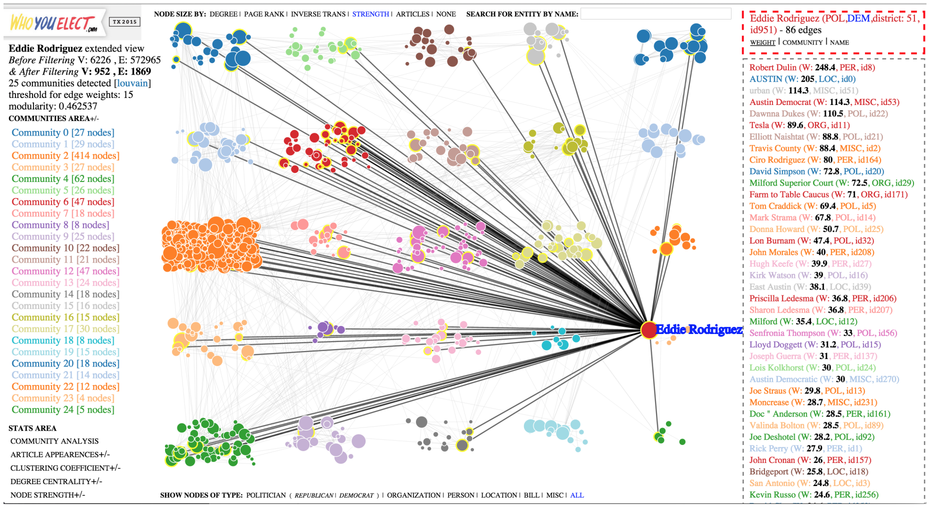
weight = (same sent + .5 * near sent + .1 same article) * boosting

where ***boosting = combined co-occurrences / same article co-occurrences***

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.

4.2 Extended View with Communities

The 2nd generated network is a global view of the entities and is an undirected graph with edges weighted according to the ”Combined” metric described already. The prior individual ”star” view could have at most N-1 edges displaying, where N is the number of entity nodes, whereas this ”extended” 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 algorithms that can be used, but for our case we decided to go with the Louvain method[BV1] 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 in the following figure that shows the network formed by articles obtained and processed for Eddie Rodriguez. A possibly better approach would be to implement a temporal based community detection approach with soft membership rules such as [LYR1] to explicitly account for the evolution of communities over time. The system defaults to searching for 25 communities and uses a link threshold of 15 to only display relationships above that amount, though both quantities are set by URL variables which may be set by the user. The Extended view spatially positions


WHO YOU ELECT.com
TX 2015
NODE SIZE BY: DEGREE | PAGE RANK | INVERSE TRANS | STRENGTH | ARTICLES | NONE
SEARCH FOR ENTITY BY NAME:
COMMUNITIES AREA+/-
STATS AREA
COMMUNITY ANALYSIS
ARTICLE APPEARENCES+/-
CLUSTERING COEFFICIENT+/-
DEGREE CENTRALITY+/-
NODE STRENGTH+/-
Eddie Rodriguez
SHOW NODES OF TYPE: POLITICIAN (REPUBLICAN | DEMOCRAT) | ORGANIZATION | PERSON | LOCATION | BILL | MISC | ALL
WEIGHT | COMMUNITY | NAME

Eddie Rodriguez extended view
Before Filtering V: 6226 , E: 572965
& After Filtering **V: 952 , E: 1869**
25 communities detected [louvain]
threshold for edge weights: 15
modularity: 0.462537

Community 0 [27 nodes]
Community 1 [29 nodes]
Community 2 [414 nodes]
Community 3 [27 nodes]
Community 4 [62 nodes]
Community 5 [26 nodes]
Community 6 [47 nodes]
Community 7 [18 nodes]
Community 8 [8 nodes]
Community 9 [25 nodes]
Community 10 [22 nodes]
Community 11 [21 nodes]
Community 12 [47 nodes]
Community 13 [24 nodes]
Community 14 [18 nodes]
Community 15 [16 nodes]
Community 16 [15 nodes]
Community 17 [30 nodes]
Community 18 [8 nodes]
Community 19 [15 nodes]
Community 20 [18 nodes]
Community 21 [14 nodes]
Community 22 [12 nodes]
Community 23 [4 nodes]
Community 24 [5 nodes]

Eddie Rodriguez (POL,DEM,district: 51, id951) - 86 edges

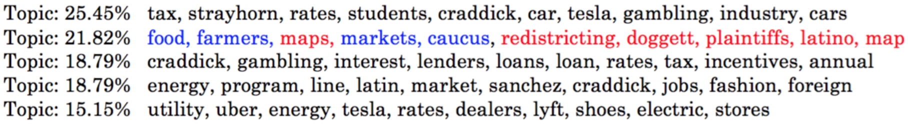
Robert Dulin (W: **248.4**, PER, id8)
AUSTIN (W: **205**, LOC, id0)
urban (W: **114.3**, MISC, id51)
Austin Democrat (W: **114.3**, MISC, id53)
Dawnna Dukes (W: **110.5**, POL, id22)
Tesla (W: **89.6**, ORG, id11)
Elliott Naishtat (W: **88.8**, POL, id21)
Travis County (W: **88.4**, MISC, id2)
Ciro Rodriguez (W: **80**, PER, id164)
David Simpson (W: **72.8**, POL, id20)
Milford Superior Court (W: **72.5**, ORG, id29)
Farm to Table Caucus (W: **71**, ORG, id171)
Tom Craddick (W: **69.4**, POL, id5)
Mark Strama (W: **67.8**, POL, id14)
Donna Howard (W: **50.7**, POL, id25)
Lon Burnam (W: **47.4**, POL, id32)
John Morales (W: **40**, PER, id208)
Hugh Keefe (W: **39.9**, PER, id27)
Kirk Watson (W: **39**, POL, id16)
East Austin (W: **38.1**, LOC, id39)
Priscilla Ledesma (W: **36.8**, PER, id206)
Sharon Ledesma (W: **36.8**, PER, id207)
Milford (W: **35.4**, LOC, id12)
Senfronia Thompson (W: **33**, POL, id56)
Lloyd Doggett (W: **31.2**, POL, id15)
Joseph Guerra (W: **31**, PER, id137)
Lois Kolkhorst (W: **30**, POL, id24)
Austin Democratic (W: **30**, MISC, id270)
Joe Straus (W: **29.8**, POL, id13)
Moncrease (W: **28.7**, MISC, id231)
Doc " Anderson (W: **28.5**, PER, id161)
Valinda Bolton (W: **28.5**, POL, id89)
Joe Deshotel (W: **28.2**, POL, id92)
Rick Perry (W: **27.9**, PER, id1)
John Cronan (W: **26**, PER, id157)
Bridgeport (W: **25.8**, LOC, id18)
San Antonio (W: **24.8**, LOC, id3)
Kevin Russo (W: **24.6**, PER, id256)

Fig. 3. Extended Network for Texas Representative Eddie Rodriguez

nodes within the same community together and assigns them the same color. Selecting a node, displays a list of most related nodes and their types and weights in the right hand column while also visually only displaying the nodes and links which are connected to the selected node. Along the top of the visualization the user may choose to size nodes by various common network metrics such as "Degree", "Page Rank", "Inverse Transitivity", "Strength", "Number of Articles" or "None" to keep them the same size. Along the left hand side, the communities are listed and may be expanded to see the nodes within a community or for more detail, the user may select "Community Analysis" which pulls up a detailed report of the articles and nodes, and network measures that define each community. The system also includes other mechanisms with the goal of helping find and highlight potentially interesting relationships to aid the end users 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”. Clicking on any of these headers displays an ordered list of the nodes with the highest number of ”Article Appearances”, etc.

4.3 Automated Summarization of Communities

If we select the Community Analysis link for Eddy Rodriguez , we can see the details pertaining to a specific community detected within the context of his retrieved and processed articles. The information there provides details into who the central figures are within a community and additionally lists 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 an initial TF-IDF procedure to filter terms and reduce noise, followed by performing Latent Dirichlet Allocation to derive topics. 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. 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 the following table we show the results of applying this technique to a community of Eddy Rodriguez’s using 1-grams with the initial amount of topics being set to 5. The initial topics number is set low because if the number of communities is 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



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

Fig. 4. Automated Community Analysis

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 Austins 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 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, and is left for future work.

5 Conclusions

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 with data obtained from various open sources in order to facilitate verified politician meta-data and provide some structure along with 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. The proof-of-concept use of topic modeling for labeling specific communities within a politicians 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.

6 Acknowledgements

We thank the anonymous reviewers for their helpful comments. Research at UPC was partially funded by SGR2014-890 (MACDA) project of the Generalitat de Catalunya and MINECO project APCOM (TIN2014-57226-P).

References

[ElE] Elect Project: http://www.electproject.org/2014g

[FJ1] James H. Fowler: Connecting the congress: A study of cosponsorship networks. Political Analysis, 14(4):456487, 2006

[FJ2] James H. Fowler, Timothy R. Johnson, James F. Spriggs II, Sangick Jeon, and Paul J. Wahlbeck: Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court. Political Analysis, 15:324346, 2007

[KHJ] Justin H Kirkland: The relational determinants of legislative outcomes: Strong and weak ties between legislators The Journal of Politics, 73(03):887898, 2011

[MT1] Matt Thomas, et. al: Get out the vote: Determining support or opposition from Congressional floor- debate transcripts. Proceedings of EMAIL, pp. 327335. 2006

[LA1] Alejandro Llorente, Manuel Garcia-Herranz, Manuel Cebrian, Esteban Moro: Social media fingerprints of unemployment arXiv:1411.3140v2 [physics.soc-ph] 2014

[EJ1] Jesús Espinal-Enriquez, Mario Siqueiros-Garca, et. al.: A literature-based approach to a narco-network. Social Informatics, pages 97101. Springer, 2014.

[MR1] Rokia Missaoui.: Mining Heterogeneous Information Nets, EGC Toulouse, 2013.

[SY1] Yizhou Sun, et a Mining Heterogeneous Information Networks. Synthesis Lectures on Data Mining & Knowledge Discovery. Morgan & Claypool Publishers 2012.

[KM1] Mikko Kivela, et. al.: Multilayer networks. Journal of Complex Networks, 2014.

[DM1] Manlio De Domenico, Mason A. Porter, Alex Arenas: MuxViz: a tool for multilayer analysis and visualization of networks. In Journal of Complex Networks. 2014

[LY1] Yang Li, et al.: Mining Evidences for Named Entity Disambiguation, KDD 2013

[SW1] Wei Shen, Jiawei Han, et al: A Probabilistic Model for Linking Named Entities in Web Text with Heterogeneous Information Networks. SIGMOD 2014

[CJ1] Jonathan Chang, Jordan Boyd Gaber, David M Blei: Connections between the Lines: Augmenting Social Networks with Text. KDD 2009.

[WC1] Chi Wang, Marina Danilevsky, et al.: A Phrase Mining Framework for Recursive Construction of a Topical Hierarchy. KDD 2013.

[ND1] David Newman, et al.: Analyzing entities and topics in news articles using statistical topic models. Intelligence and Security Informatics. Springer, 2006.

[PB1] Bruno Pouliquen, Ralf Steinberger, et al. : Building and Displaying Name Relations using Automatic Unsupervised Analysis of Newspaper Articles. In Journées internationales dAnalyse statistique des Donnees Textuelles. 8es, 2006.

[MT1] Theodosis Moschopoulos, Elias Iosif, et al.: Toward the automatic extraction of policy networks using web links and documents. Knowledge and Data Engineering, IEEE Transactions on, 25(10):24042417, 2013

[TF1] Fangbo Tao, George Brova, et al.: NewsNetExplorer: Automatic Construction and Exploration of News Information Networks. SIGMOD 2014, June 2227, 2014

[DJ1] Jana Diesner: Extraction and Analysis of Semantic Network Data from Text Data. Semantic Network Analysis Workshop, St. Petersburg State, May 2013.

[HJ1] Jiawei Han, Chi Wang, et al: Bringing Structure to Text. KDD 2014

[CR1] Rudi L Cilibrasi and Paul MB Vitanyi.: The google similarity distance. Knowledge and Data Engineering, IEEE Transactions on, 19(3):370383, 2007.

[TH1] Hristo Tanev: Unsupervised learning of social networks from a multiple-source news corpus. Multisource, Multilingual Info Extraction & Summarization, 2007.

[PB2] Bruno Pouliquen, et. al.: Extracting and learning social networks out of multilingual news. Social Networks and Application tools. Skalica,Slovakia, Sept 2008.

[BV1] Vincent D. Blondel, Jean-Loup Guillaume, et al: Fast unfolding of communities in large networks. In J. Stat. Mech. and arXiv:0803.0476, 2008

[LYR1] Yu-Ru Lin, et al: FacetNet: A Framework for Analyzing Communities and Their Evolutions in Dynamic Networks WWW 2008, April 2125, 2008