AUTOMATED CONSTRUCTION & ANALYSIS OF POLITICAL NETWORKS VIA OPEN GOVERNMENT & MEDIA SOURCES

DIEGO GARCIA-OLANO
MARTA ARIAS
JOSEP LLUÍS LARRIBA PEY

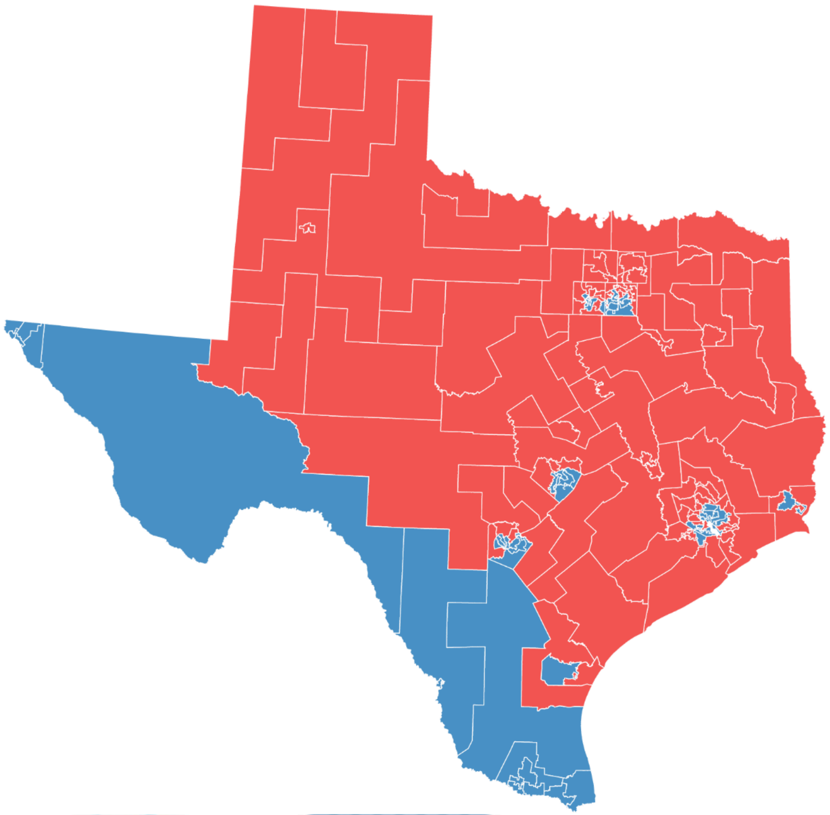
for ECML PKDD 2016 So Good Workshop

LARCA

DAMA-UPC. DATA MANAGEMENT
UNIVERSITAT POLITÈCNICA DE CATALUNYA

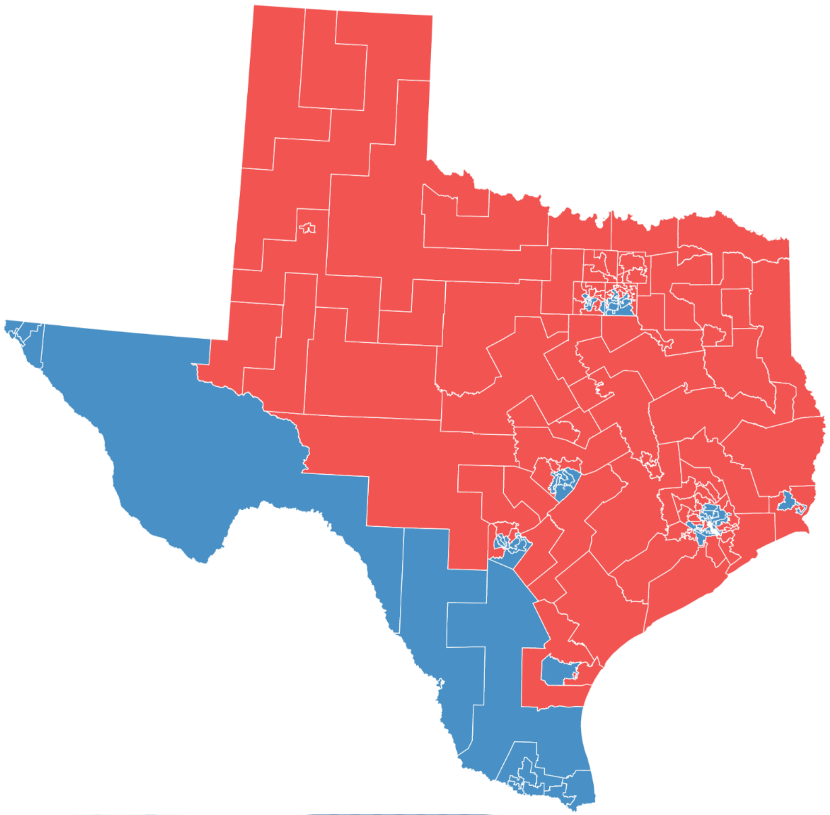
ECML–PKDD 2016 RIVA DEL GARDA

THEY MYSTERY THAT IS TEXAS



Population: 27 Million People

THE MYSTERY THAT IS TEXAS POLITICS



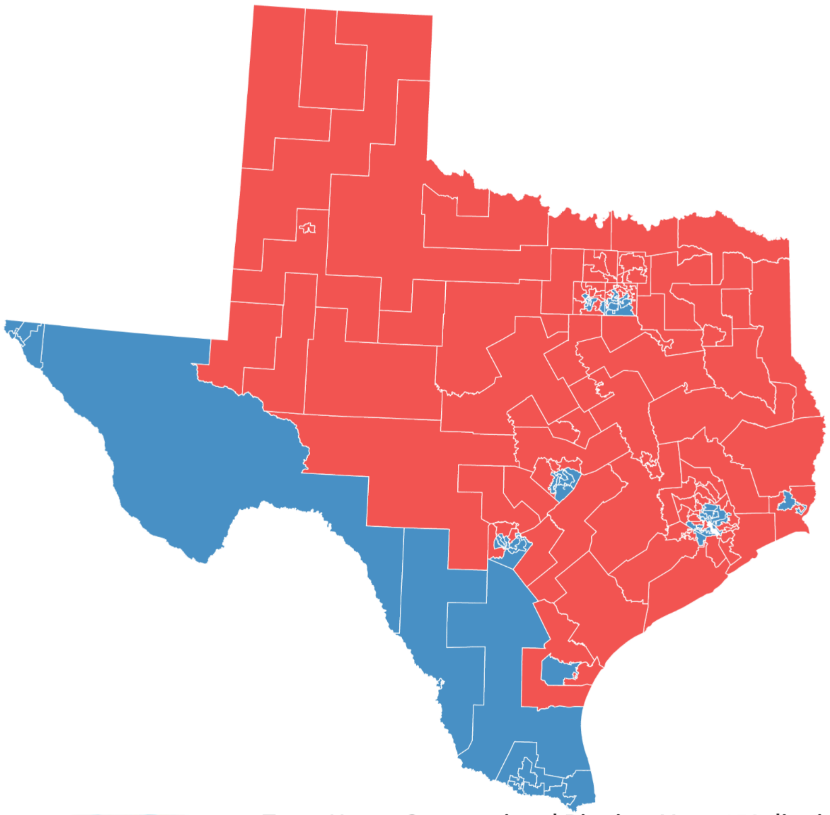
Population: 27 Million People

State & Federal Politicians: **250**

composed of:

- **State Congress**: 150 Representatives & 31 State Senators
- **U.S. Congress**: 36 Representatives & 2 Senators
- **State Executive officials**:
  Governor, Lieutenant Governor, Speaker of the House, Attorney General, Comissioners, etc
- **State Judges**: 9 Supreme Court & 9 Court of Criminal Appeals

THE MYSTERY THAT IS TEXAS VOTER TURNOUT



Population: 27 Million People

State & Federal Politicians: 250

composed of:

- **State Congress**: 150 Representatives & 31 State Senators
- **U.S. Congress**: 36 Representatives & 2 Senators
- **State Executive officials**:
  Governor, Lieutenant Governor, Speaker of the House, Attorney General, Comissioners, etc
- **State Judges**: 9 Supreme Court & 9 Court of Criminal Appeals

Voter Turnout: 20% for 2015 Governor election.

Texas House Congressional Districts Map. 150 districts total
http://www.whoyouelect.com/texas/texas-house-map.html


WHO YOU ELECT.com
TX 2015
1-3

WhoYouElect.com

1. list of Politicians
2. list of relevant news sites

Construct & Display the networks around these politicians

Case study:

1. 246 currently active Texas elected officials
2. 6 news sites that cover Texas politics

   the Austin American Statesman, the Dallas Morning News, the Houston Chronicle, the New York Times, the Texas Observer and the Texas Tribune

PRESENTATION OUTLINE

- Introduction
- Overview of WhoYouElect.com
- Automated Construction of Networks & Prior Work
    1. Ego Network of a Politician
    2. Extended Network of a Politician
    3. Automated Summarization of Communities via Topic Modeling
- Conclusions & Future Work

Case Study Results, Networks, Maps, Code: http://whoyouelect.com/texas
Slides: http://www.whoyouelect.com/texas/ecmlpkdd-sogood/

OVERVIEW OF WHOYOUELECT.COM

http://www.whoyouelect.com/texas

EDDIE RODRIGUEZ
Democratic State Representative
District: 51

WHO IS EDDIE RODRIGUEZ?

Table Of Contents View:

http://www.whoyouelect.com/texas/table-of-contents.html

1. EGO "INNER" NETWORK OF A POLITICIAN

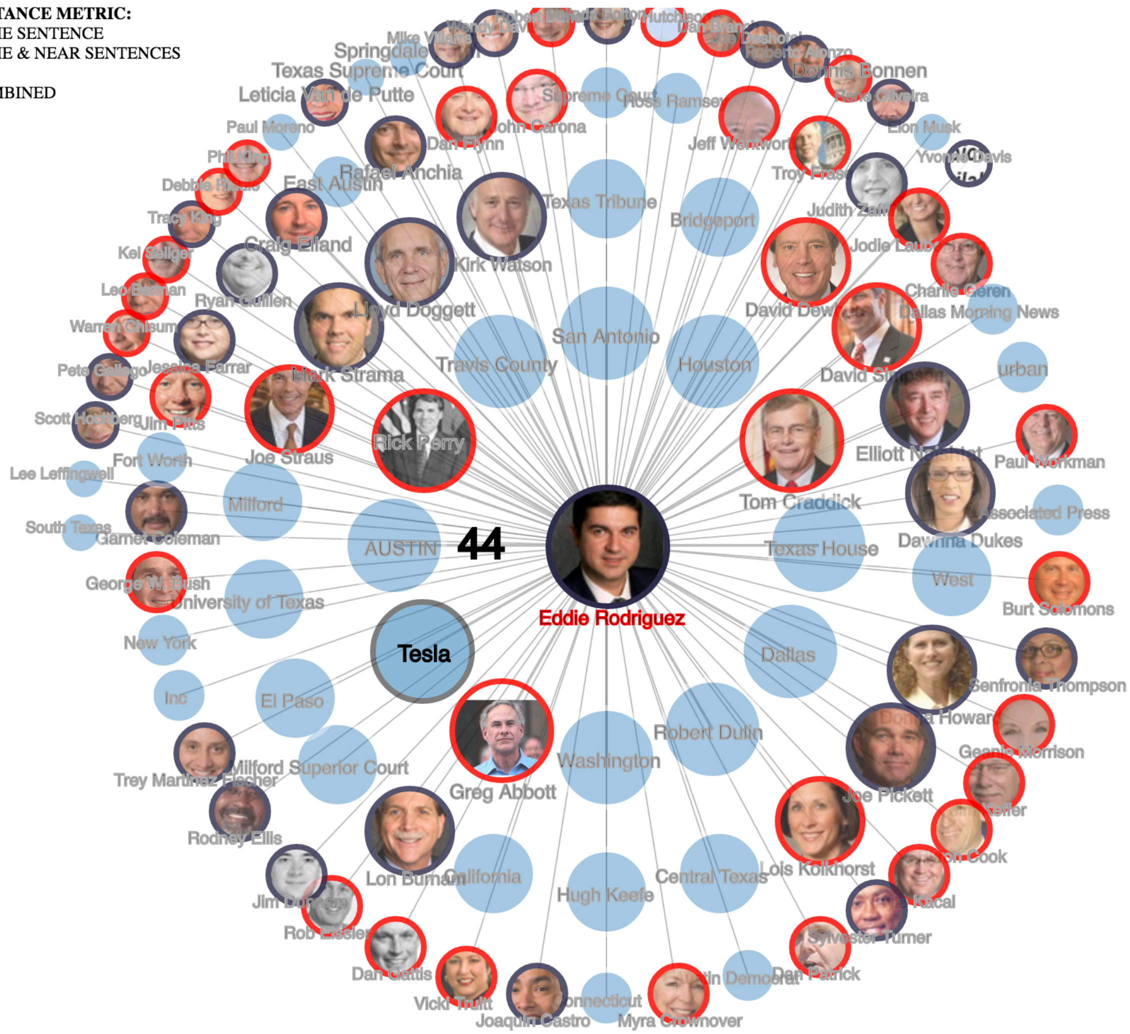
http://www.whoyouelect.com/texas/explorer-view.html?show=15&s=Eddie Rodriguez

or just click on "Inner Network" from the Table of Contents view

URL parameters

s	name of Politician
show	only show edges with weight greater than or equal to **show**
dm	(optional) which distance metric to use. possible values: ss, sn, all, comb . corresponding to Same Sentence, Same or Near, All Co-Occurences, Proposed Combined Metric. Defaults to "All"
near_co	(optional) Near Sentence Coefficient in calculation for Combined Metric
same_art_co	(optional) Same Article Coefficient in calculation for Combined Metric
from	(optional) date from which to include articles found. expected format: YYYY-MM-DD
to	(optional) end date for inclusion of articles. expected format: YYYY-MM-DD . ex: 2008-07-01
exclude	(optional) which news sources to exclude. possible values: AAS, DMN, HC, NYT, TXOB, TXTRB . use commas to exclude multiple

DISTANCE METRIC:
SAME SENTENCE
SAME & NEAR SENTENCES
ALL
COMBINED



TX 2015

SEARCH RESULTS:
FOR "EDDIE RODRIGUEZ" FOUND:
362 ARTICLES FROM
6 DIFFERENT SOURCES AND
6226 ENTITIES DISCOVERED
[LOCATION: 882, POLITICIAN: 276, MISC: 411, ORGANIZATION: 1945, PERSON: 2625, BILL: 87]
OF WHICH 106 ARE CURRENTLY SHOWING

FILTER BY DATE:
FROM | TO | APPLY DATE FILTER

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

APPLY NEW SOURCE FILTER

Showing nodes with more than **15** associations and hiding full details.

Show More Nodes | Show Less Nodes | Show Full Details

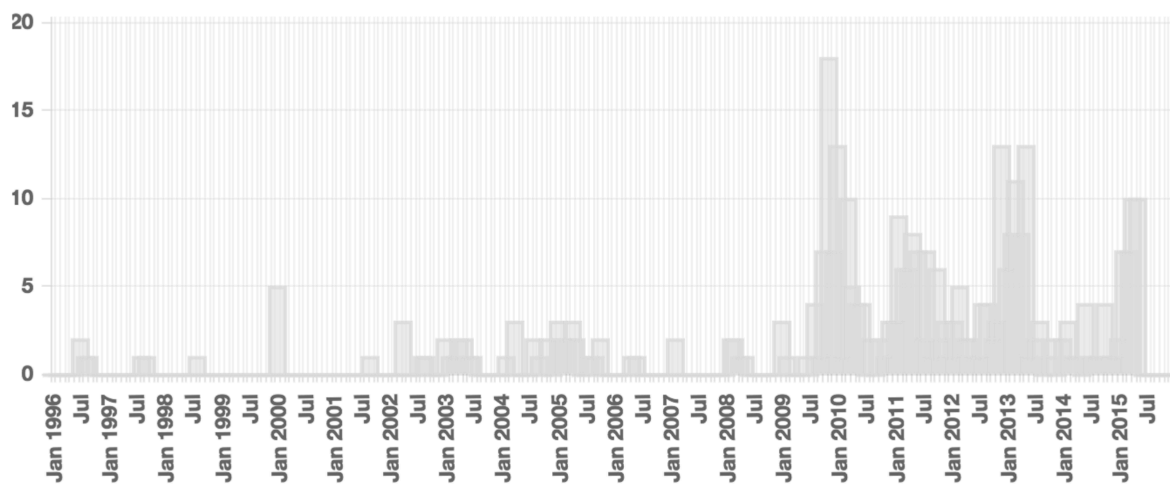
ABOUT THE PROJECT | ABOUT US | SHOW STATS | ARTICLE URLS

FILTER BY: POLITICIAN (REPUBLICAN | DEMOCRAT) | ORGANIZATION | PERSON | LOCATION | BILL | MISC | ALL

Top Associated same sentence | Top Associated same sentence/near | Top Associated same sentence/near/article | Top Associated same article only (not near) | Top Associated combined metric | Summary comparison

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	urban - 8
Mark Strama - 14	Mexican American Legislative Caucus - 3	Paul Moreno - 4	Texas House - 6		
Lon Burnam - 12		Lee Leffingwell - 4	Milford - 6		

2. EXTENDED VIEW OF A POLITICIAN

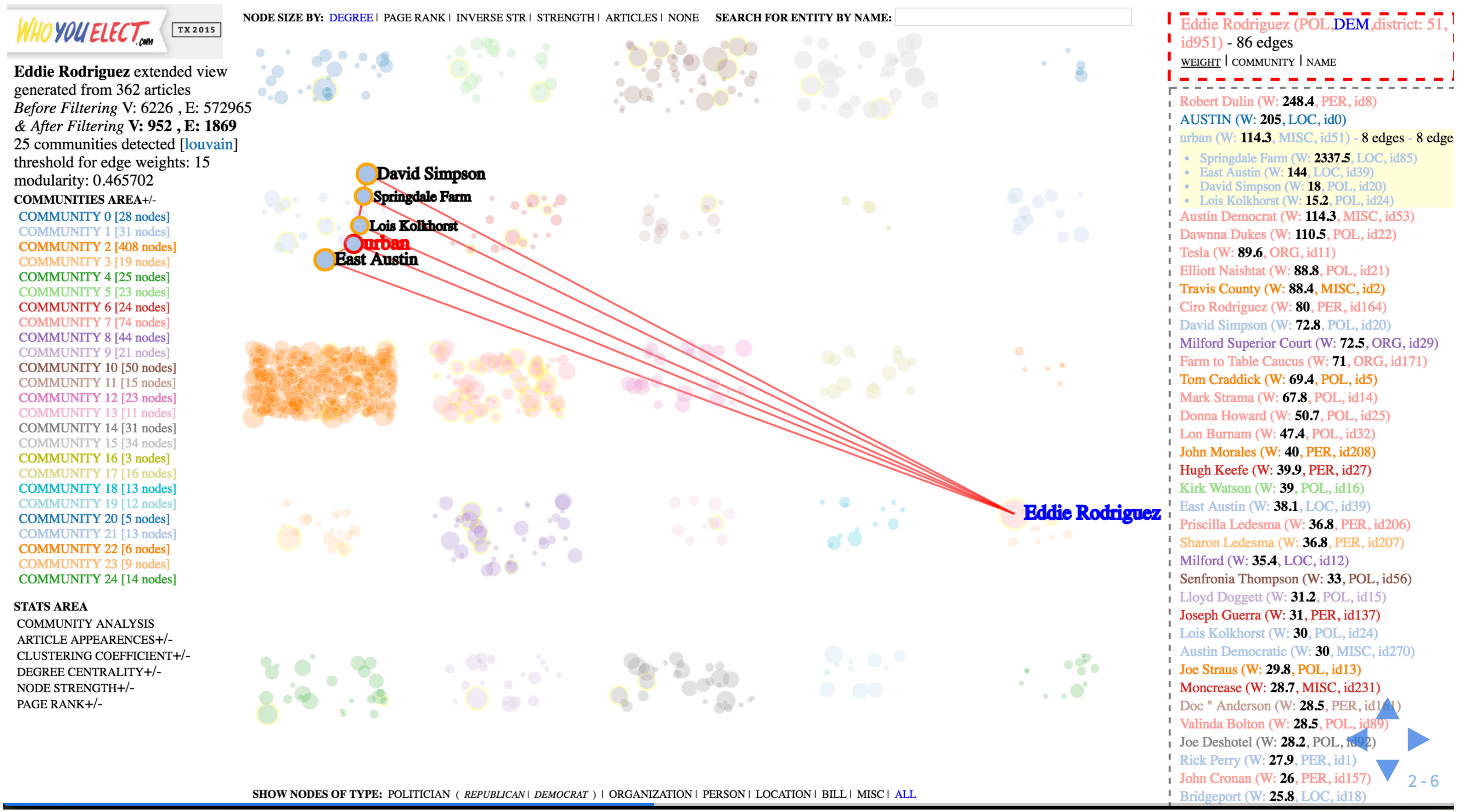
http://www.whoyouelect.com/texas/communities-from-ncol.html?cl=25&t=15&s=Eddie Rodriguez

* Edges weighted according to the proposed "Combined" metric!
* Placement of nodes/communities is calculated to maximize seperation and clarity

or just click on "Extended Network" from the Table of Contents view

URL parameters

s	name of Politician
t	only show edges with weight greater than or equal to threshhold **t**. Defaults to 15.
cl	number of communities to discover in network. Defaults to 25.


WHO YOU ELECT .com
TX 2015
NODE SIZE BY: DEGREE | PAGE RANK | INVERSE STR | STRENGTH | ARTICLES | NONE
SEARCH FOR ENTITY BY NAME:
Eddie Rodriguez extended view
generated from 362 articles
Before Filtering V: 6226 , E: 572965
& After Filtering V: 952 , E: 1869
25 communities detected [louvain]
threshold for edge weights: 15
modularity: 0.465702
COMMUNITIES AREA+/-
COMMUNITY 0 [28 nodes]
COMMUNITY 1 [31 nodes]
COMMUNITY 2 [408 nodes]
COMMUNITY 3 [19 nodes]
COMMUNITY 4 [25 nodes]
COMMUNITY 5 [23 nodes]
COMMUNITY 6 [24 nodes]
COMMUNITY 7 [74 nodes]
COMMUNITY 8 [44 nodes]
COMMUNITY 9 [21 nodes]
COMMUNITY 10 [50 nodes]
COMMUNITY 11 [15 nodes]
COMMUNITY 12 [23 nodes]
COMMUNITY 13 [11 nodes]
COMMUNITY 14 [31 nodes]
COMMUNITY 15 [34 nodes]
COMMUNITY 16 [3 nodes]
COMMUNITY 17 [16 nodes]
COMMUNITY 18 [13 nodes]
COMMUNITY 19 [12 nodes]
COMMUNITY 20 [5 nodes]
COMMUNITY 21 [13 nodes]
COMMUNITY 22 [6 nodes]
COMMUNITY 23 [9 nodes]
COMMUNITY 24 [14 nodes]
STATS AREA
COMMUNITY ANALYSIS
ARTICLE APPEARENCES+/-
CLUSTERING COEFFICIENT+/-
DEGREE CENTRALITY+/-
NODE STRENGTH+/-
PAGE RANK+/-
David Simpson
Springdale Farm
Lois Kolkhorst
urban
East Austin
Eddie Rodriguez
SHOW NODES OF TYPE: POLITICIAN (REPUBLICAN | DEMOCRAT) | ORGANIZATION | PERSON | LOCATION | BILL | MISC | ALL
Eddie Rodriguez (POL,DEM,district: 51, id951) - 86 edges
WEIGHT | COMMUNITY | NAME
Robert Dulin (W: 248.4, PER, id8)
AUSTIN (W: 205, LOC, id0)
urban (W: 114.3, MISC, id51) - 8 edges - 8 edge
Springdale Farm (W: 2337.5, LOC, id85)
East Austin (W: 144, LOC, id39)
David Simpson (W: 18, POL, id20)
Lois Kolkhorst (W: 15.2, POL, id24)
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)
2 - 6

3. AUTOMATED SUMMARIZATION OF COMMUNITIES

1. Treat the articles of a given community collectively as a single corpus.

2. Run an initial TF-IDF procedure to filter terms and reduce noise, and

3. followed by Latent Dirichlet Allocation (LDA) to derive topics.

**We can weigh articles by their relative importance in the community.

TOPIC MODELING TO SUMMARIZE A SINGLE COMMUNITY

```
library("topicmodels")
library("tm")
k = 5                                                        #num of communities to look for
highend = 4000                                               #want less than this many terms
lowend = 2000                                                #want more than this many terms
tsvfile = "eddie_rodriguez-articles.tsv"                                #community article texts
articlesForCommunity <- read.csv("lda/community-articles-lda.csv") #comm article meta data
colnames(articlesForCommunity) <- c("url","date","entities")
ngramtype <- 1
weighbyentity = T
article_cutoff = 1
runTopicModelingOnCommunity(tsvfile,k,highend,lowend,
        articlesForCommunity,ngramtype,article_cutoff,weighbyentity)

runTopicModelingOnCommunity <- function( ... ){
        # .... load tsv to get article texts and put in JSS_papers list
        corpus <- Corpus(VectorSource(sapply(JSS_papers[, "text"], remove_HTML_markup)))
        JSS_dtm <- DocumentTermMatrix(corpus, control = list(stopwords = TRUE,
                    minWordLength = 3, removeNumbers = TRUE, removePunctuation = TRUE))
        term_tfidf <- tapply( JSS_dtm$v/row_sums(JSS_dtm)[JSS_dtm$i], JSS_dtm$j, mean)
                                        * log2( nDocs(JSS_dtm) / col_sums(JSS_dtm > 0))
        cutoffvals <- get_cutoffval_and_type(term_tfidf,highend,lowend)
        # .... filter corpus by cutoff vals ...
        jss_TM <- LDA(JSS_dtm, k = k, control = list(seed = SEED))
        Topic <- topics(jss_TM)
        Terms <- terms(jss_TM)
        # ... generate most frequent topics of the articles and the terms for each one
}
```

SHOW COMMUNITY: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 | OVERALL

Community 0 with 17 entities SPANNING 170 articles

CONDUCTANCE = 37/(2*52 + 37) = **0.2624** -- fraction of edges leaving the communities. (smaller is better)
EXPANSION = 37/89 = **0.4157** -- nr of edges per node leaving the community

L = LOCAL TO COMMUNITY
G = GLOBAL

ENTITIES INFO+/-

Name	Type	Party	Level	Position	District	G Degree	L Degree	G Page Rank	G Transitivity	G Strength	L Strength	L Betweeness	Articles
Andrew Smiley	PER					5	5	0	0.5	1766	1766	0	2
AUSTIN	LOC					34	10	100	0.033	2810.2	1582.6	63	155
Sustainable Food Center	ORG					5	5	0	0.5	1421.5	1421.5	0	3
House Bill 1392	BILL					7	5	0	0.357	648	504	0	1
Department of State Health Services	ORG					7	5	0	0.357	648	504	0	7
Susan King	POL	Republican	statewide-active	Representative	71	7	5	0	0.357	607.5	463.5	0	4
Uncle Billy	PER					2	2	0	0	180	180	12	1
Kel Seliger	POL	Republican	statewide-active	Senator	31								
Central Texas	LOC												

SHOW COMMUNITY: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Community 0 with 17 entities SPANNING 170 articles

CONDUCTANCE = 37/(2*52 + 37) = **0.2624** -- fract
EXPANSION = 37/89 = **0.4157** -- nr of edges per nod

ENTITIES INFO+/-

ARTICLES INFO+/-

URL	DATE	NUM TIMES IN COMMUNITY
www.texastribune.org/2013/05/30/bipartisan-caucus-lays-groundwork-food-movement/	2013-05-30	6
www.texastribune.org/2005/01/17/easy-as-pie/	2005-01-17	4
www.texastribune.org/2013/05/10/guest-column-early-look-legislative-partisanship/	2013-05-10	4
www.texastribune.org/2010/03/01/one-question-remains/	2010-03-01	4
www.texastribune.org/2005/04/04/snake-eyes-1/	2005-04-04	3
www.texastribune.org/2008/04/14/how-it-all-came-out/	2008-04-14	3

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

LARCA

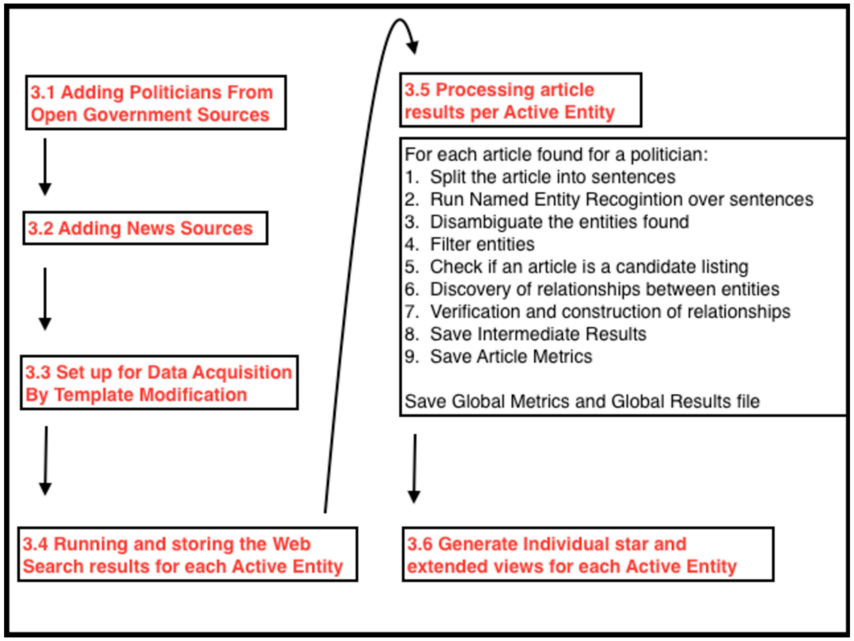
AUTOMATED CONSTRUCTION OF NETWORKS & PRIOR WORK

PRIOR WORKS

- 1. Time intensive, hand crafted networks
- 2. Article lookup systems with an impressive, though limited, breadth of sources
- 3. Paid-for search engine results, only processing first twenty results and only using text present in Yahoo's search result listings .

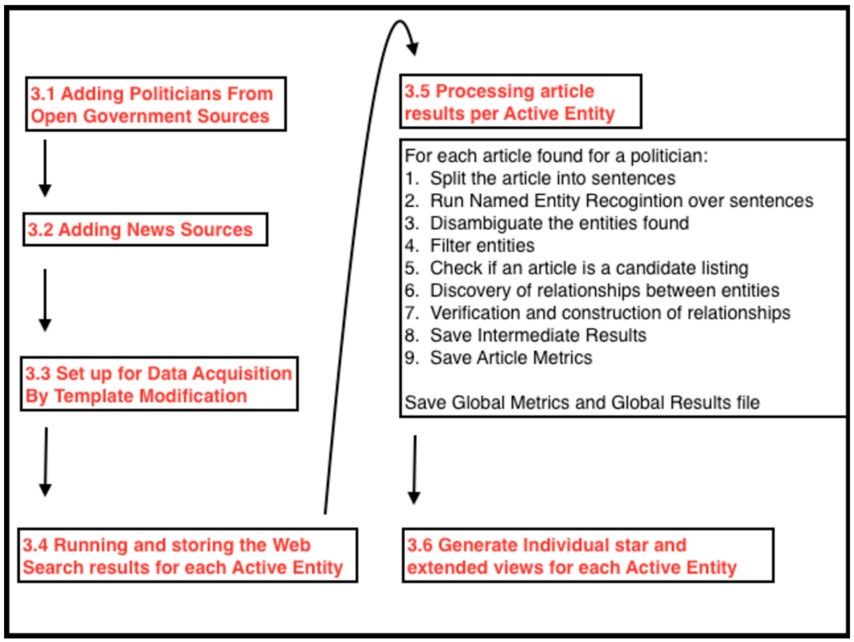
We were unable to find any work that leveraged the publically available search engines present in most news websites.

GENERAL OVERVIEW OF GRAPH CONSTRUCTION PROCESS



http://openstates.org/api/v1/legislators/?state=tx&active=true
https://www.govtrack.us/api/v2/role?current=true&state=TX
http://www.sos.state.tx.us/elections/voter/elected.shtml

GENERAL OVERVIEW OF GRAPH CONSTRUCTION PROCESS



http://openstates.org/api/v1/legislators/?state=tx&active=true
https://www.govtrack.us/api/v2/role?current=true&state=TX
http://www.sos.state.tx.us/elections/voter/elected.shtml

Languages, Libraries, and Databases

Python: general backend work
MongoDB: to store article texts and entity information
MITIE: MIT open source Named Entity Recognition tool
BeautifulSoup and Selenium libraries:
python webscrapers used in obtaining articles.
BS4 is for static web pages,
while Selenium, using the PhantomJs webdriver,
handles pages constructed by javascript dynamically
langdetect:
open source python library for language detection
D3.js: network visualizations & maps
jQuery UI: some frontend interactivity functionality
jLouvain: javascript Louvain community detection
html5 webworkers:
asynchronous, nonblocking JS load of data for graphs

3.1 ADDING POLITICIANS FROM OPEN GOV. SOURCES

- 1. Texas Congress data was obtained from OpenStates.org for both active and inactive members
- 2. Federal Congress data was obtained from GovTrack.us for current federal representatives
- 3. Other Texas state officials data was obtained from the Secretary of State of Texas website via a script

This gives us the metadata for all the politicians

3.2 ADDING NEWS SOURCES

- Dallas Morning News
- Houston Chronicle
- Austin American Statesman
- Texas Observer
- Texas Tribune
- New York Times

A reasonable mix of representative media sources on Texas politics.

3.3 SETUP FOR DATA ACQUISITION BY TEMPLATE MODIFICATION

Two template "web scraper" solutions are provider

Based on whether a news site renders its site content statically or dynamically via Javascript

1. The static version, based on BeautifulSoup
 - acquires data more quickly,
 - but can not handle dynamic content.

2. The dynamic version, based on Selenium/PhantomJS
 - can handle static or dynamic content,
 - but goes slower than the static solution;

Future work is planned to unify the approach into one template

3.4 RUNNING AND STORING THE WEB SEARCH RESULTS FOR EACH ACTIVE ENTITY

This step calls the webscraper template for a politican.
For each news source

- download the list of article urls returned from its internal search engine for that politician
- download full articles into JSON files
- do language detection for the text of the article before importing JSON into MongoDB.

3.5 PROCESSING ARTICLE RESULTS PER ACTIVE ENTITY

Take all the articles downloaded for a given politician, process and store them, and construct graphs for the politician:
1. the Ego "Inner" Network, and
2. the Extended Network View

3.5 PROCESSING ARTICLE RESULTS ... CONTINUED

Go through all the articles for a politician one by one and

- 1. filter out empty articles, sports articles, and articles not containing the politician's name explicitly.
- 2. Split article into sentences, and run the MITIE Named Entity Recognition library over each one. This finds "entities" in each sentence and gives each a tag of "person","location","organization" or "misc". Additionally check if "person" tags are "politicians" using our entities DB or whether any Congressional "bill"s exist in the sentence using a heuristic.
- 3. Run coreference resolution over all the entities found to get an additional dictionary of all distinct entities found in the article.
- 4. From this and the tagged sentences, we then find and store all co-occurences that occurred within the same sentence, within three sentences, or outside of that distance for all entities in the dictionary.
- 5. At this point, the article has been processed and we merge and save it locally.

3.6 CREATE NETWORK VIEWS

Using saved result objects and statistics, construct the Ego and Expanded network data files

CONCLUSIONS & FUTURE WORK

CONTRIBUTIONS

We presented ...

1. a tool that generates real world political networks from user provided lists of politicians and news sites.
2. enriched with data obtained from open sources to provide structure via verified politician meta-data.
3. the Ego "Inner" and "Extended" graph visualizations
4. a “Combined” distance metric to better assess the strength of relationships between actors in a graph.
5. a proof-of-concept use of topic modeling for labeling communities in a politician’s “extended” network (not shown)

Uses:

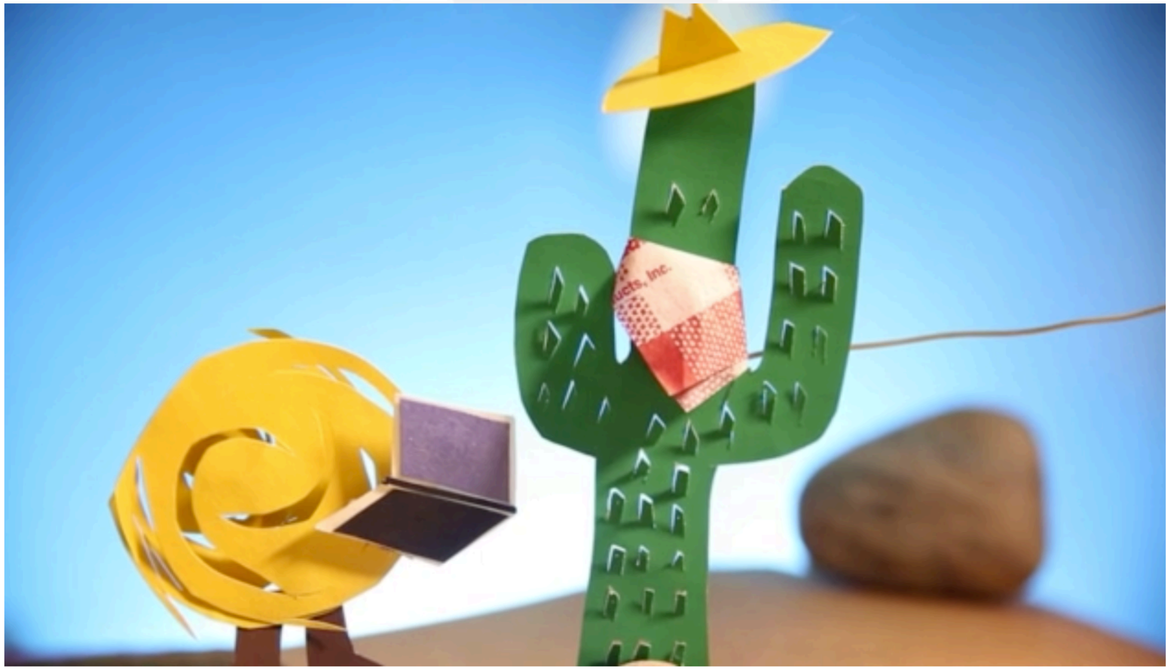
- voter education
- creating real world networks for study
- discovering potential news stories

FUTURE WORK

1. an extensive statistical study of the merged "extended" graphs obtained
2. incorporation of city & local APIs for better resolution of elected officials, campaign funding APIs for influence tracking, congressional bill APIs, public health, socio-economic, and voting history APIs
3. all articles aren't equal, and as such weighing article relations differently is very important.
4. better disambiguation of entities, use of alias lists, automated merging tools
5. simplification of webscraping solution & refactoring code to handle "parties" more generally for non-US cases
6. expanding NER solution to provide for more language handling (Catalan for instance)
7. refactoring text-snippet solution for better scalability.
8. developing mechanism for downloading, processing and adding new articles for existing politicians.
9. assessing use of multiplex paradigm by introducing additional link types (“neighboring districts”, “author of bill”, “member of committee”, etc.) for more robust network analysis.
10. leverage posteriors of LDA for better topic analysis, and similarly leverage stochaistic community detection methods
11. relationship labeling/role discovery incorporation (signed positive/negative edges when applicable)
12. temporarl community detection work/view

WHO YOU ELECT.com TX 2015

THANKS!



QUESTIONS ?

TELL YOUR TEXAS FRIENDS :)

Who You Elect Texas: whoyouelect.com/texas

Slides: whoyouelect.com/texas/ecml-pkdd2016-sogood

email: diegoolano@gmail.com web: diegoolano.com
github: github.com/diegoolano twitter: dgolano