Jesse Lecy

Arizona State University

TEXT AS DATA IN R

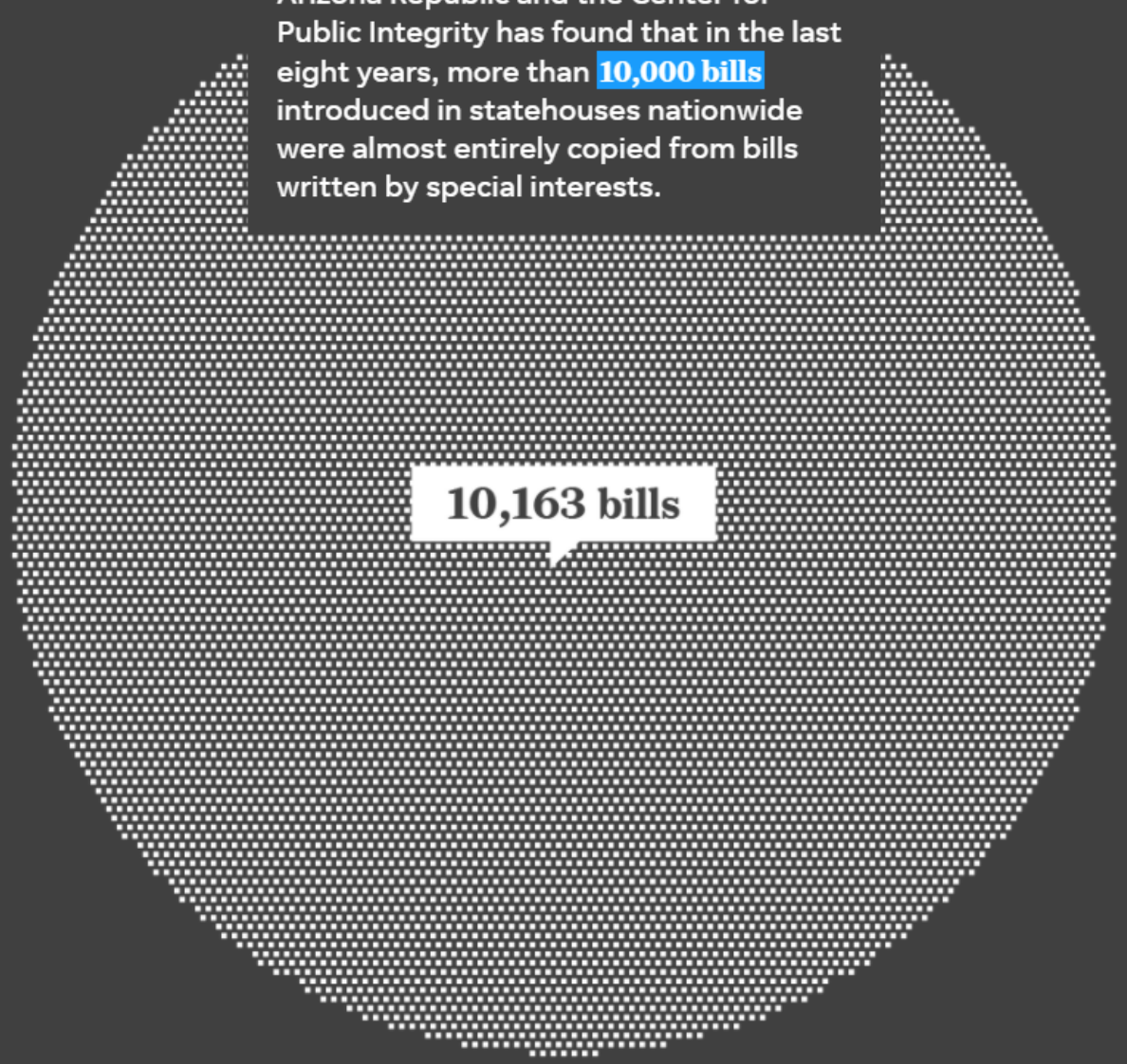
Model bill

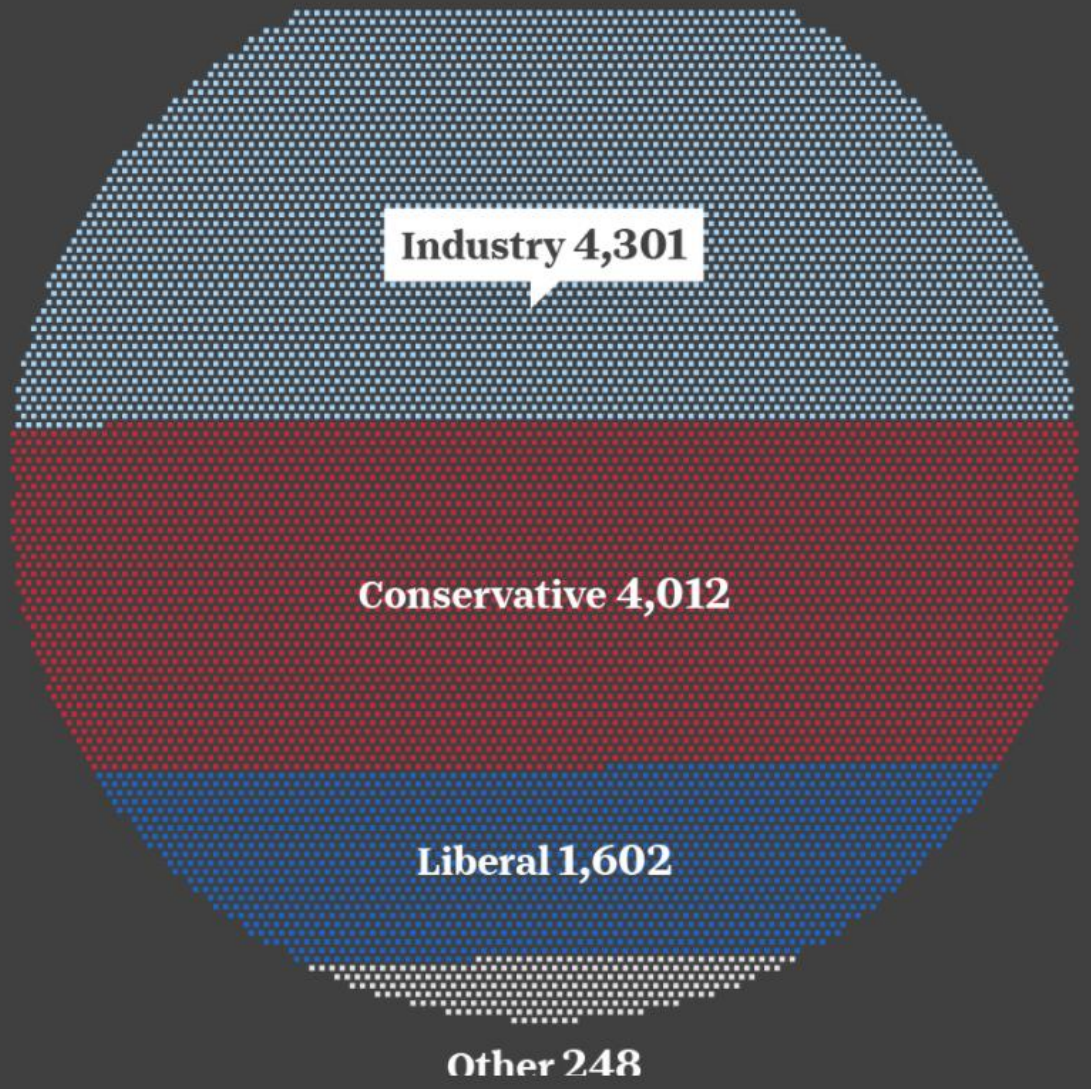
Liberties rights and privileges granted under the us and state constitutions including but not limited to due process freedom of religion speech or press and any right of privacy or marriage as specifically defined by the constitution of this state a contract or contractual provision if severable which provides for the choice of a law legal code or system to govern some or all of the disputes between the parties adjudicated by a court of law or by an arbitration panel arising from the contract mutually agreed upon shall violate the public policy of this state and be void and unenforceable if the law legal code or system chosen includes or incorporates any substantive or procedural law as applied to the dispute at issue that would not grant the parties the same fundamental liberties rights and privileges granted under the us and state constitutions including but not limited to due process freedom of religion speech or press and any right of privacy or marriage as specifically defined by the constitution of this state a a contract or contractual provision if severable which provides for a jurisdiction for purposes of granting the courts or arbitration panels in personam jurisdiction over the parties to

Model Bill

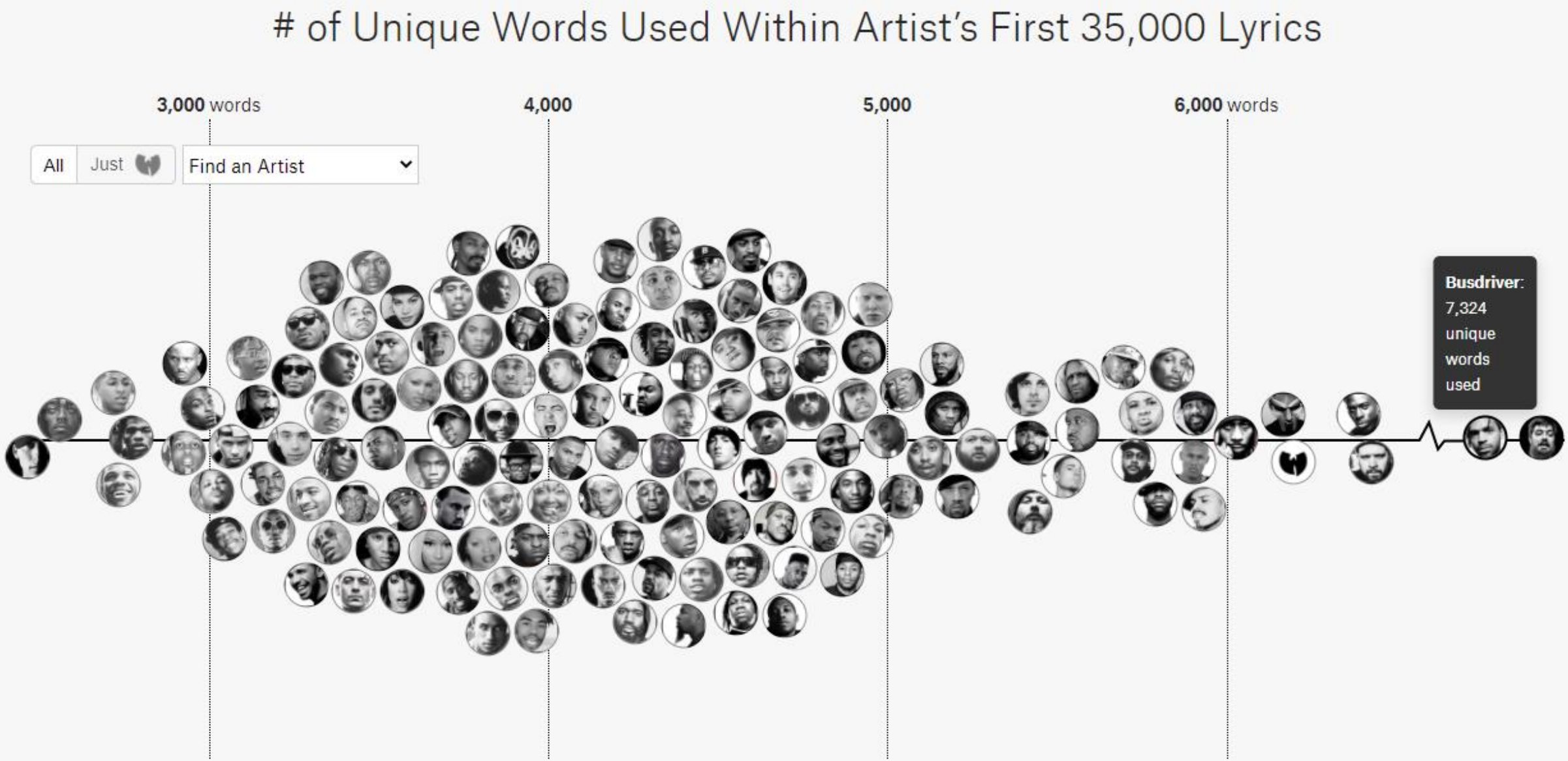
Attorneys fees and expenses to prepare and file the asbestos trust claim identified in the defendants motion exceed the plaintiffs reasonably anticipated recovery from the trust if the court determines that there is a sufficient basis for the plaintiff to file the asbestos trust claim identified by a defendant the court shall order the plaintiff to file the asbestos trust claim and shall stay the asbestos action until the plaintiff files the asbestos trust claim and provides all parties with all trust claims materials no later than thirty days before trial if the court determinesthat the plaintiffs expenses or attorneys fees and expenses to prepare and file the asbestos trust claim identified in the defendants motion exceed the plaintiffs reasonably anticipated recovery from the asbestos trust the court shall stay the asbestos action until the plaintiff files with the court and provides all parties with a verified statement of the plaintiffs history of exposure usage or other connection to asbestos covered by the asbestos trust not less than thirty days prior to trial in an asbestos action the court shall enter into the record a trust claims document that identifies each claim the plaintiff has made against an asbestos trust section valuation of asbestos trust claims

An investigation by USA TODAY, The Arizona Republic and the Center for Public Integrity has found that in the last eight years, more than **10,000 bills** introduced in statehouses nationwide were almost entirely copied from bills written by special interests.





The Largest Vocabulary In Hip Hop



David Robinson

Principal Data Scientist at Heap, works in R and Python.

Email
Twitter
Github
Stack Overflow

Subscribe

Your email

Subscribe to this blog

Recommended

- R Bloggers
- RStudio Blog
- R4Stats
- Simply Statistics
- Upfront

Who wrote the anti-Trump New York Times op-ed? Using tidytext to find document similarity

Like a lot of people, I was intrigued by "I Am Part of the Resistance Inside the Trump Administration", an anonymous New York Times op-ed written by a "senior official in the Trump administration". And like many data scientists, I was curious about what role text mining could play.

This is a useful opportunity to demonstrate how to use the tidytext package that Julia Silge and I developed, and in particular to apply three methods:

- Using TF-IDF to find words specific to each document (examined in more detail in Chapter 3 of our book)
- Using widyr to compute pairwise cosine similarity
- How to make similarity interpretable by breaking it down by word

Since my goal is R education more than it is political analysis, I show all the code in the post.

Even in the less than 24 hours since the article was posted, I'm far from the first to run text analysis on it. In particular Mike Kearney has shared a great R analysis on GitHub (which in particular pointed me towards CSPAN's cabinet Twitter list), and Kanishka Misra has done some exciting work here.

text as data

raw text

Attorneys fees and expenses to prepare and file the asbestos trust claim identified in the defendants motion exceed the plaintiffs reasonably anticipated recovery from the trust if the court determines that there is a sufficient basis for the plaintiff to file the asbestos trust claim identified by a defendant the court shall order the plaintiff to file the asbestos trust claim and shall stay the asbestos action until the plaintiff files the asbestos trust claim and provides all parties with all trust claims materials no later than thirty days before trial if the court determinesthat the plaintiffs expenses or attorneys fees and expenses to prepare and file the asbestos trust claim identified in the defendants motion exceed the plaintiffs reasonably anticipated recovery from the asbestos trust the court shall stay the asbestos action until the plaintiff files with the court and provides all parties with a verified statement of the plaintiffs history of exposure usage or other connection to asbestos covered by the asbestos trust not less than thirty days prior to trial in an asbestos action the court shall enter into the record a trust claims document that identifies each claim the plaintiff has made against an asbestos trust section valuation of asbestos trust claims

Step 1 – Get Some Text

LegiScan Search

View Top 50 Searches

State:
Kansas
Select area of search.

Bill Number:
Find an exact bill number.

Full Text Search:
Search bill text and data. [help]

Search Reset

LegiScan Info

- 2021 Schedules
- Governor Deadlines
- Effective Dates
- LegiScan API
- Weekly Datasets
- Documentation
- Search Help
- Legislation 101
- On Bill Numbers
- State Support

LegiScan Trends

View Top 50 National

National Legislative Datasets

Please note you must be logged in to download the datasets. Registration is free. Signup here.

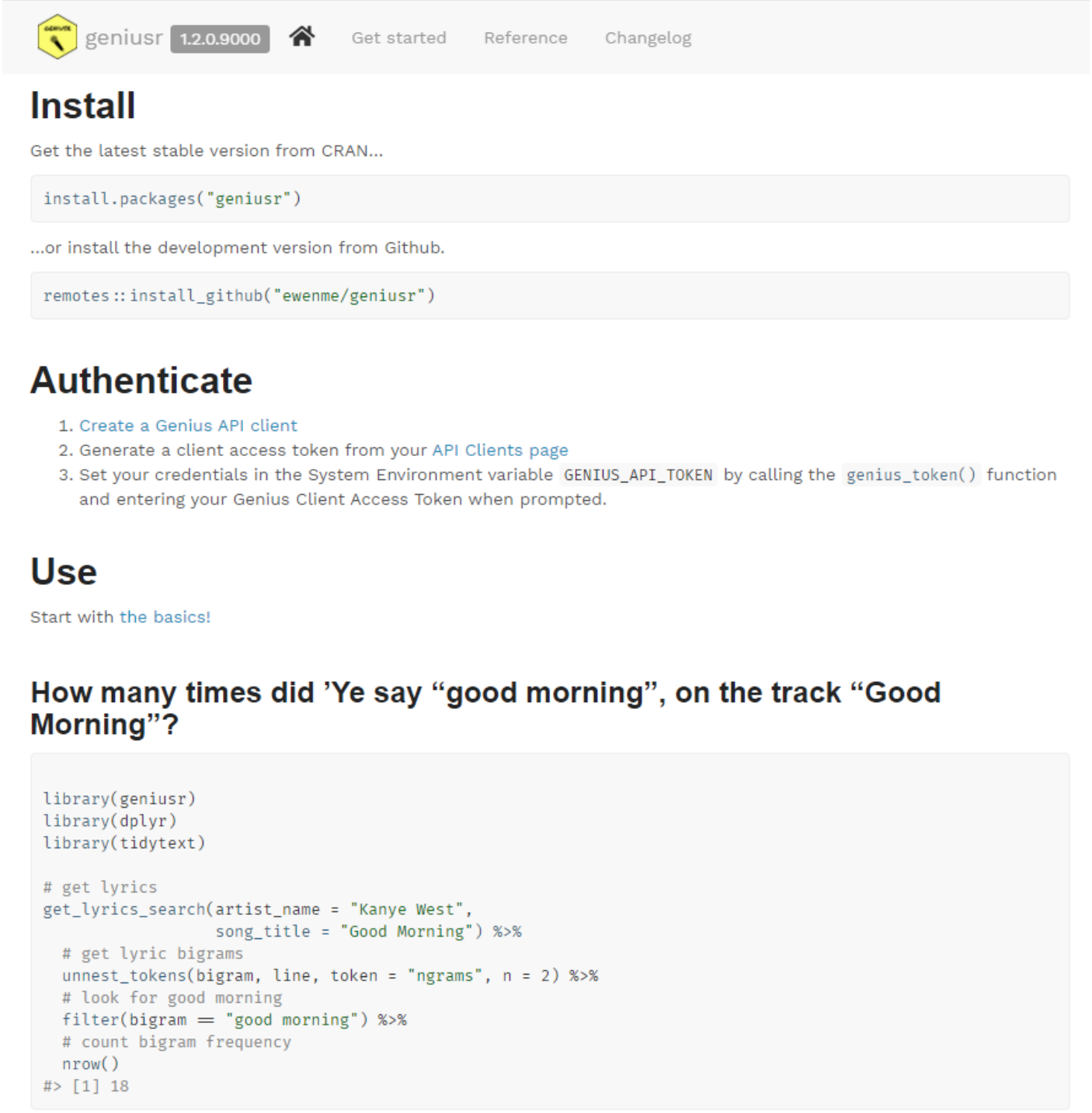
2021 AK AL AR AZ CA CO CT DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS MT NC ND NE NH NJ
NY OH OK OR PA RI SC SD TN TX UT VA VT WA WI WV WY DC US

Weekly snapshots of session data are created each Sunday morning with updated information on an as-needed basis. Provided in simple comma-separated values files for general bill data, or the most complete for LegiScan API JSON payloads. Current year data is below, for a complete historical list of session files select an individual state above.

Inquire for more details on subscription data services, including near real-time remote replication of the national database, alternative licensing terms, or snapshots of the 250GB text training corpus.

- LegiScan API Info - Register for LegiScan API Key
- LegiScan API Client Source - Turnkey Bulk/Pull/Push API Client & Database
- LegiScan API Client Docs - Commented Source Browser
- LegiScan API Manual - API Hooks and Data Structures

State	Year	Session	Modified	Exported	API JSON	CSV Basic
Alabama	2021	Regular Session	2021-02-11	2021-02-21	**JSON** 1.1 MB	**CSV** 155 KB
Alaska	2021-2022	32nd Legislature	2021-02-19	2021-02-21	**JSON** 321 KB	**CSV** 41 KB
Arizona	2021	Fifty-fifth Legislature 1st Regular	2021-02-19	2021-02-21	**JSON** 3.8 MB	**CSV** 375 KB
Arkansas	2021	93rd General Assembly	2021-02-10	2021-02-21	**JSON** 1.6 MB	**CSV** 239 KB
California	2021-2022	Regular Session	2021-02-19	2021-02-21	**JSON** 2.9 MB	**CSV** 257 KB
Colorado	2021	Regular Session	2021-02-19	2021-02-21	**JSON** 487 KB	**CSV** 62 KB



R Package geniusr: API for Genius platform

Genius celebrates More Than The Music—the lyrics, the stories behind the songs, and the creative connections that meaningfully drive culture.

We champion curiosity and believe that everyone has music knowledge to share: insights, intel, and musings that make us more informed, engaged music lovers. As the world's biggest music encyclopedia with a passionate community of more than two million contributors, Genius is a destination for artists, creatives, and superfans to discuss and deconstruct all things music.

Through our original content, technology, and live & virtual experiences, Genius spotlights the artists who are shaping music culture across every genre and musical discipline, sharing the stories behind their creativity and craft in their own words.

David Robinson

Principal Data Scientist at Heap, works in R and Python.

Email
Twitter
Github
Stack Overflow

Downloading data

The harder step is getting a set of documents representing "senior officials". An imperfect but fast approach is to collect text from their Twitter accounts. (If you find an interesting dataset of, say, government FOIA documents, I recommend you try extending this analysis!)

We can look at a combination of two (overlapping) Twitter lists containing administration staff members:

- CSPAN's list of Cabinet accounts
- digiphile's list of White House staff

"This results in a set of 136,501 tweets from 69 Twitter handles"

```
library(rtweet)

cabinet_accounts <- lists_members(owner_user = "CSPAN", slug = "the-cabinet")
staff <- lists_members(slug = "white-house-staff", owner = "digiphile")

# Find unique screen names from either account
accounts <- unique(c(cabinet_accounts$screen_name, staff$screen_name))

# Download ~3200 from each account
tweets <- map_df(accounts, get_timeline, n = 3200)
```

This results in a set of 136,501 from 69 Twitter handles. There's certainly no guarantee that the op-ed writer is among these Twitter accounts (or, if they are, that they even write their tweets themselves). But it still serves as an interesting case study of text analysis. How do we find the tweets with the closest use of language?

text as data

raw text

Attorneys fees and expenses to prepare and file the asbestos trust claim identified in the defendants motion exceed the plaintiffs reasonably anticipated recovery from the trust if the court determines that there is a sufficient basis for the plaintiff to file the asbestos trust claim identified by a defendant the court shall order the plaintiff to file the asbestos trust claim and shall stay the asbestos action until the plaintiff files the asbestos trust claim and provides all parties with all trust claims materials no later than thirty days before trial if the court determinesthat the plaintiffs expenses or attorneys fees and expenses to prepare and file the asbestos trust claim identified in the defendants motion exceed the plaintiffs reasonably anticipated recovery from the asbestos trust the court shall stay the asbestos action until the plaintiff files with the court and provides all parties with a verified statement of the plaintiffs history of exposure usage or other connection to asbestos covered by the asbestos trust not less than thirty days prior to trial in an asbestos action the court shall enter into the record a trust claims document that identifies each claim the plaintiff has made against an asbestos trust section valuation of asbestos trust claims

pre-processing

1. Remove punctuation
2. Remove “stop” words
3. Identify compound constructs
4. Stem words

Step 2: Standardize the data

Unit of Analysis: Nonprofit Mission Statements

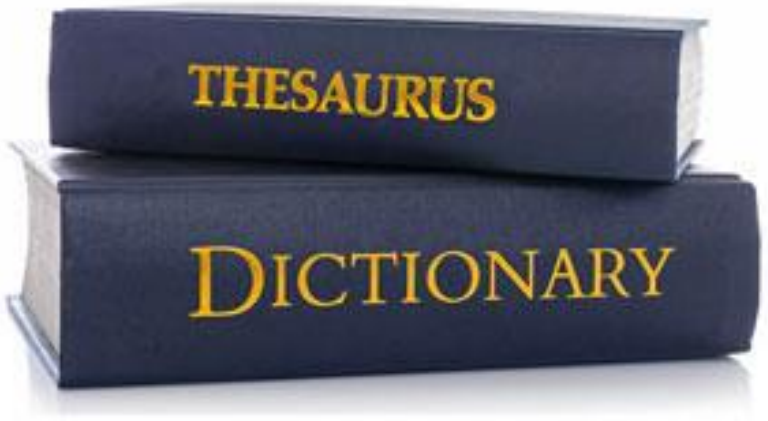
The corporation's specific purpose is to supports affordable housing, community development and economic development of the city and county of San Francisco's economically disadvantaged individuals and communities, by lending to, investing in, and directly acquiring such affordable housing and related community development real estate assets.

~~the~~ corporation specific purpose ~~is to~~ support AFFORDABLE_HOUSING, community development ~~and~~ ECONOMIC_DEVELOPMENT ~~of the~~ city and county of SAN_FRANCISCO economically disadvantaged individuals and communities by lending ~~to~~ investing ~~in and~~ directly acquiring ~~such~~ AFFORDABLE_HOUSING ~~and~~ related community development REAL_ESTATE assets

1. Remove punctuation
2. Delete words with little information value (“stop” words)
3. Identify compound constructs (n-grams)

STEMMING

LEND	RELATE
LEND~~ing~~	RELATE~~d~~



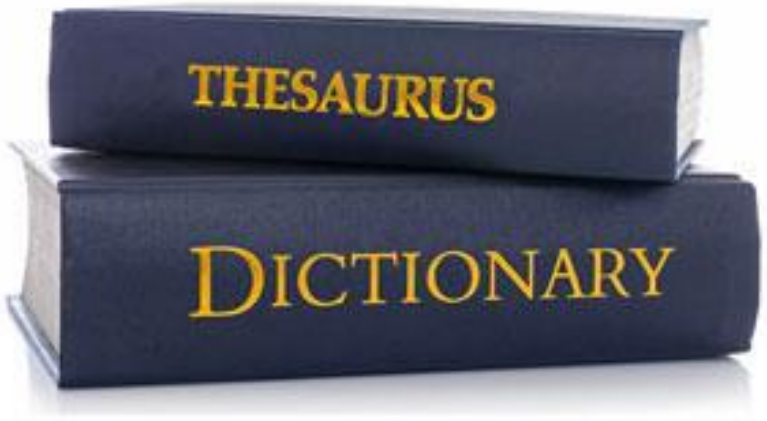
DISAMBIGUATION

George W. Bush

George Bush Jr.

President Bush

→ GW_BUSH



DISAMBIGUATION

New York City → PLACE

THING ← New York University

New Yorker → STATE OF MIND

STATE ← New York

Step 3: Wrangle the data

raw text

Attorneys fees and expenses to prepare and file the asbestos trust claim identified in the defendants motion exceed the plaintiffs reasonably anticipated recovery from the trust if the court determines that there is a sufficient basis for the plaintiff to file the asbestos trust claim identified by a defendant the court shall order the plaintiff to file the asbestos trust claim and shall stay the asbestos action until the plaintiff files the asbestos trust claim and provides all parties with all trust claims materials no later than thirty days before trial if the court determinesthat the plaintiffs expenses or attorneys fees and expenses to prepare and file the asbestos trust claim identified in the defendants motion exceed the plaintiffs reasonably anticipated recovery from the asbestos trust the court shall stay the asbestos action until the plaintiff files with the court and provides all parties with a verified statement of the plaintiffs history of exposure usage or other connection to asbestos covered by the asbestos trust not less than thirty days prior to trial in an asbestos action the court shall enter into the record a trust claims document that identifies each claim the plaintiff has made against an asbestos trust section valuation of asbestos trust claims

pre-processing

transformation

- Word frequency table
- Document Frequency Matrix
- Co-occurrence Matrix
- Word vector machines

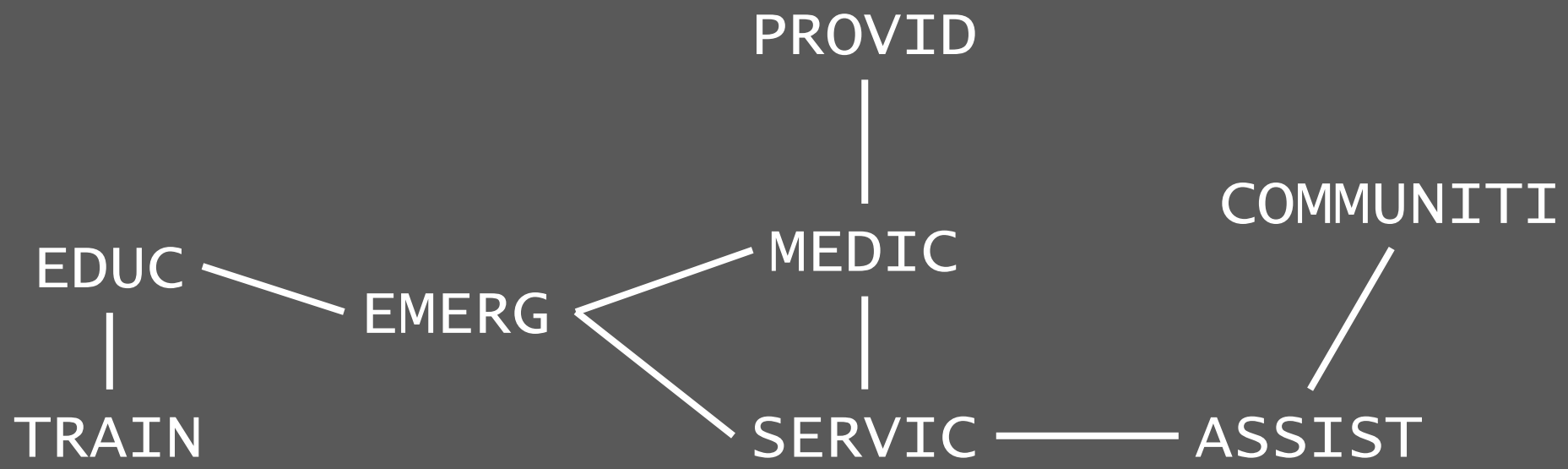
Raw text

To educate, train and assist in providing emergency medical service for the community.

Tokenized text

```
"EDUC" "TRAIN" "ASSIST"
"PROVID" "EMERG" "MEDIC"
"SERVIC" "COMMUNITI"
```

Semantic Networks



Document Frequency Matrix (DFM)

Terms	Documents													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14
abnormalities	0	0	0	0	0	0	0	1	0	1	0	0	0	0
age	1	0	0	0	0	0	0	0	0	0	0	1	0	0
behavior	0	0	0	0	1	1	0	0	0	0	0	0	0	0
blood	0	0	0	0	0	0	0	1	0	0	1	0	0	0
close	0	0	0	0	0	0	1	0	0	0	1	0	0	0
culture	1	1	0	0	0	0	0	1	1	0	0	0	0	0
depressed	1	0	1	1	1	0	0	0	0	0	0	0	0	0
discharge	1	1	0	0	0	1	0	0	0	0	0	0	0	0
disease	0	0	0	0	0	0	0	0	1	0	1	0	0	0
fast	0	0	0	0	0	0	0	0	0	1	0	1	1	1
generation	0	0	0	0	0	0	0	0	1	0	0	0	1	0
oestrogen	0	0	1	1	0	0	0	0	0	0	0	0	0	0
patients	1	1	0	1	0	0	0	1	0	0	0	0	0	0
pressure	0	0	0	0	0	0	0	0	0	0	1	0	0	1
rats	0	0	0	0	0	0	0	0	0	0	0	0	1	1
respect	0	0	0	0	0	0	0	1	0	0	0	1	0	0
rise	0	0	0	1	0	0	0	0	0	0	0	0	0	1
study	1	0	1	0	0	0	0	0	1	0	0	0	0	0

text as data

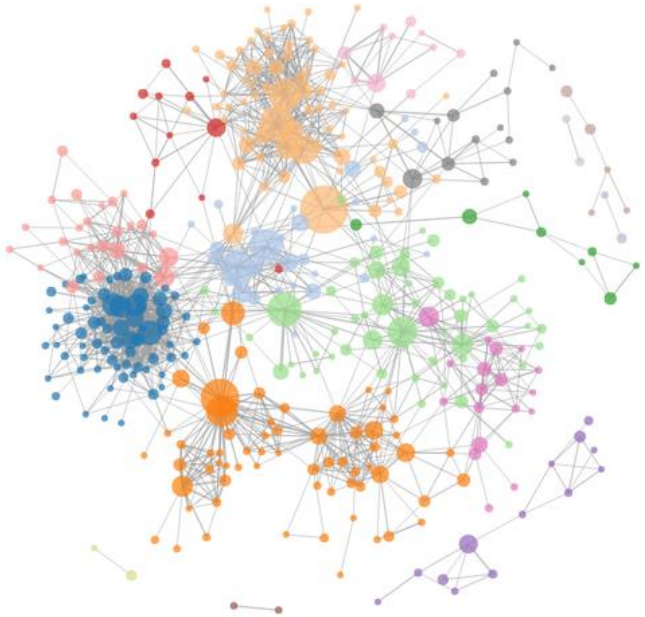
raw text

Attorneys fees and expenses to prepare and file the asbestos trust claim identified in the defendants motion exceed the plaintiffs reasonably anticipated recovery from the trust if the court determines that there is a sufficient basis for the plaintiff to file the asbestos trust claim identified by a defendant the court shall order the plaintiff to file the asbestos trust claim and shall stay the asbestos action until the plaintiff files the asbestos trust claim and provides all parties with all trust claims materials no later than thirty days before trial if the court determinesthat the plaintiffs expenses or attorneys fees and expenses to prepare and file the asbestos trust claim identified in the defendants motion exceed the plaintiffs reasonably anticipated recovery from the asbestos trust the court shall stay the asbestos action until the plaintiff files with the court and provides all parties with a verified statement of the plaintiffs history of exposure usage or other connection to asbestos covered by the asbestos trust not less than thirty days prior to trial in an asbestos action the court shall enter into the record a trust claims document that identifies each claim the plaintiff has made against an asbestos trust section valuation of asbestos trust claims

pre-process

transform

analyze

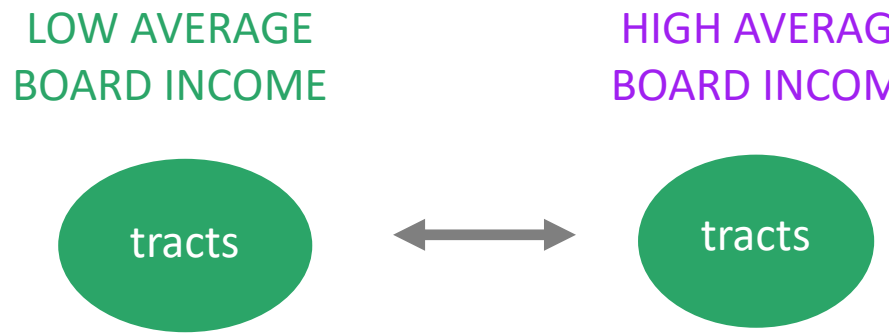


- Content analysis
- Sentiment analysis
- Semantic network analysis
- Bayesian classifiers

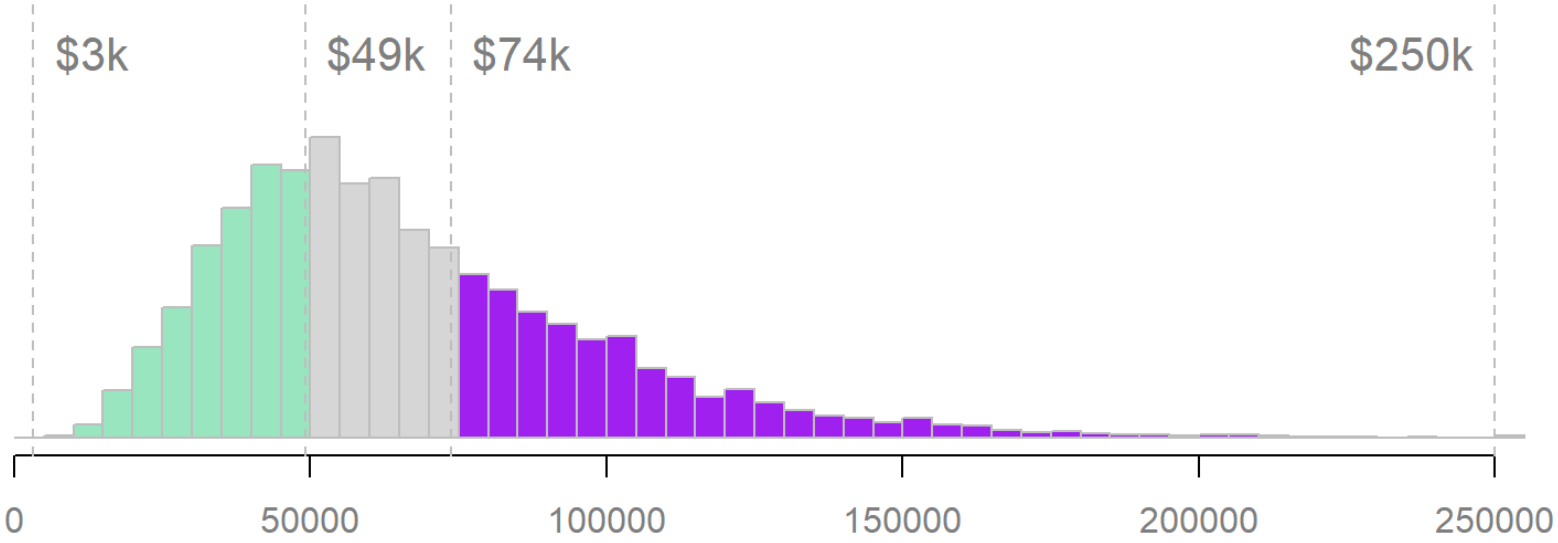
Board Influence on Mission:

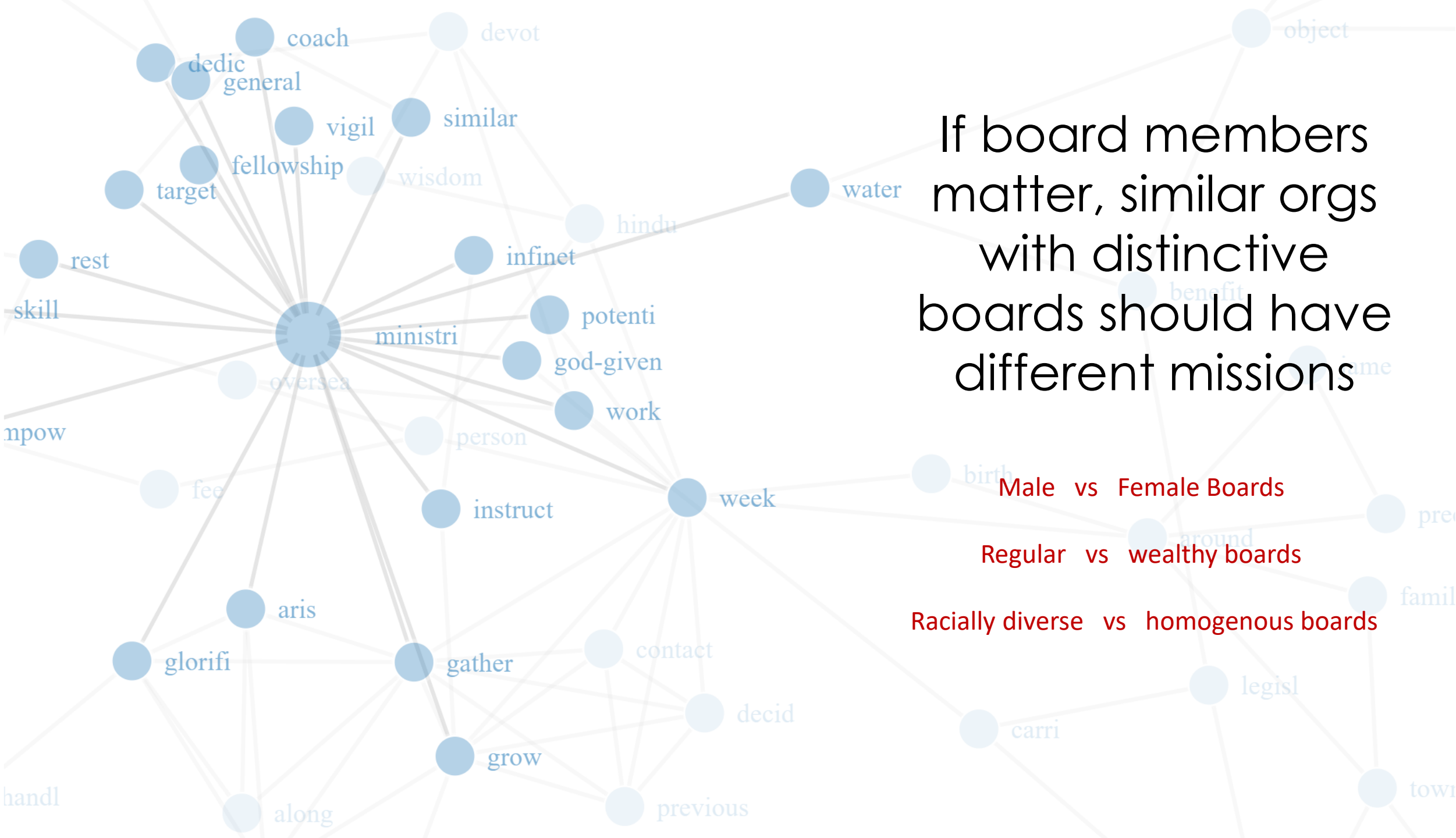
HOLD CONSTANT
nonprofit subsector and community income status

VARY the board traits of the nonprofit



ARTS nonprofits located in low-income communities





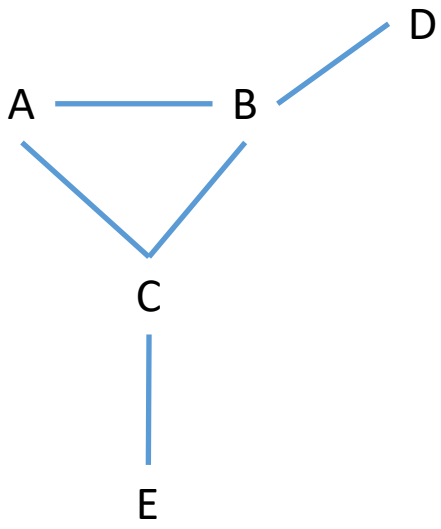
If board members matter, similar orgs with distinctive boards should have different missions

Male vs Female Boards

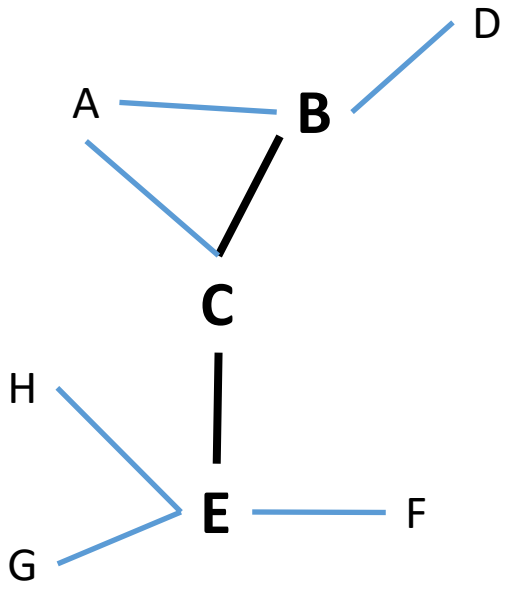
Regular vs wealthy boards

Racially diverse vs homogenous boards

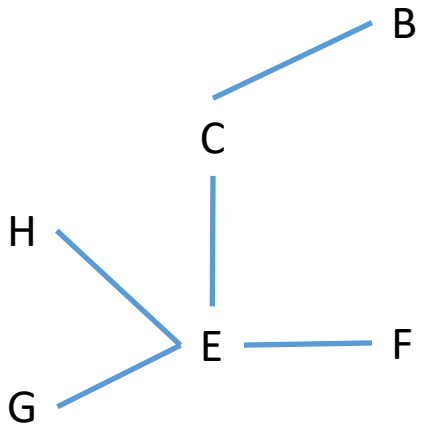
Mission Statement 1



Union (all statements) and **Intersection**



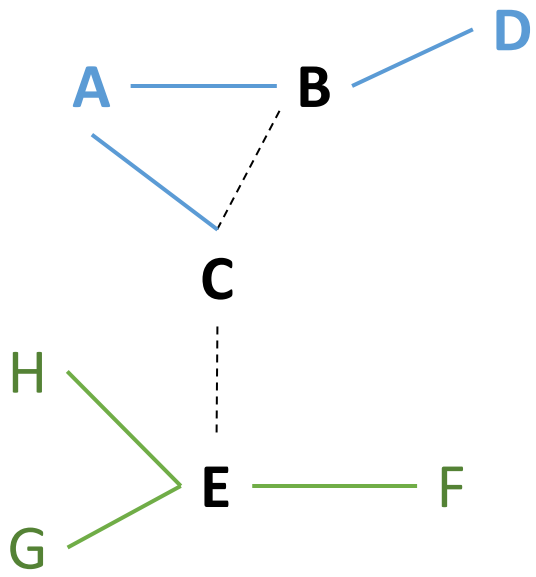
Mission Statement 2



Analyzing Missions by Types of Nonprofits

Mission Statement Components Unique to Org 1:

A-B
A-C
B-D



H-E
G-E
E-F

Mission Statement Components Unique to Org 2

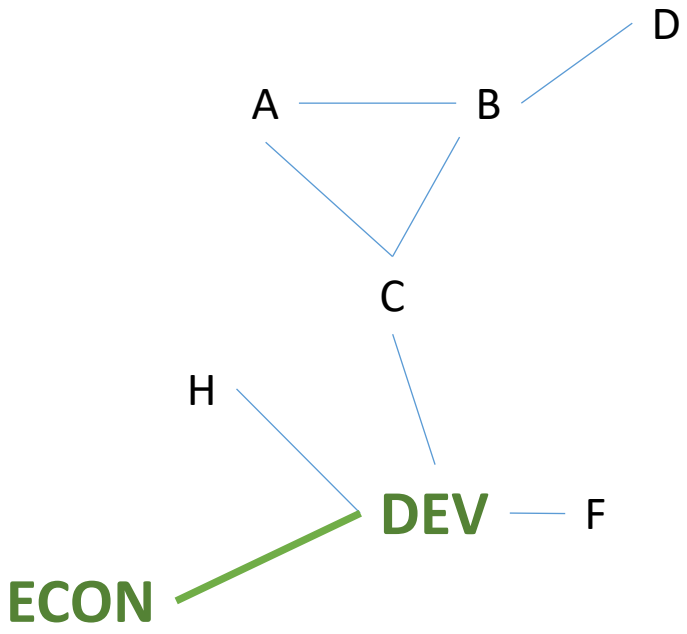
Doesn't work well with dense weighted graphs!

Data structure of a weighted network:

Freq ALL	Freq GROUP	Term 1	Term 2
10	**7**	**econ**	**dev**
7	4	self	reliance
5	3	dev	con
5	2	globla	econ
4	2	local	econ
4	1	soc	econ
3	2	econ	socialism
3	3	finance	global
3	2	global	finance
3	2	global	impsm
3	1	impsm	global
3	1	impsm	invasion

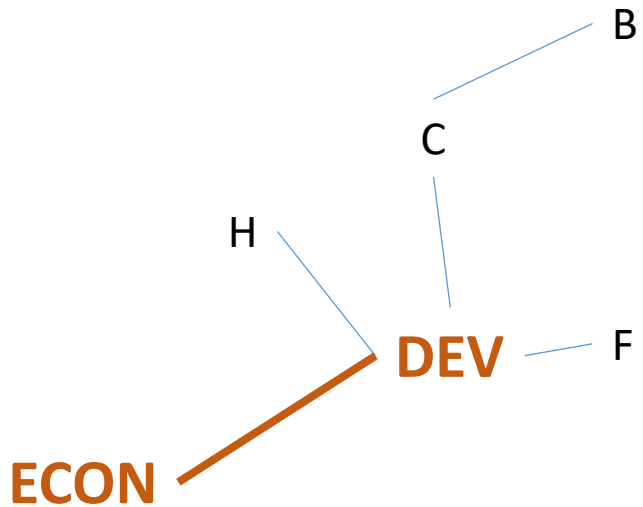
Is it significant that **economic development** was mentioned **7 times** by a specific type of organization?

ALL MISSION STATEMENTS



N=10

SUB-GROUP

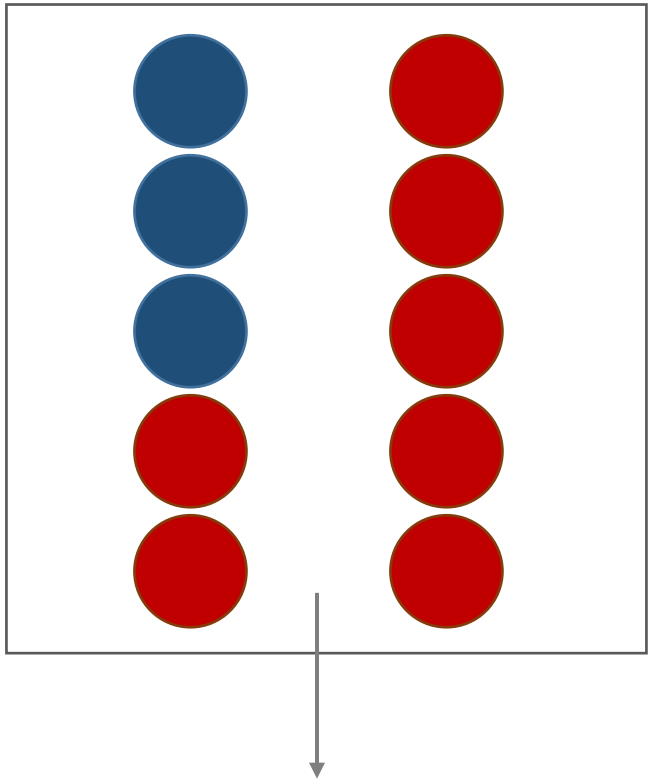


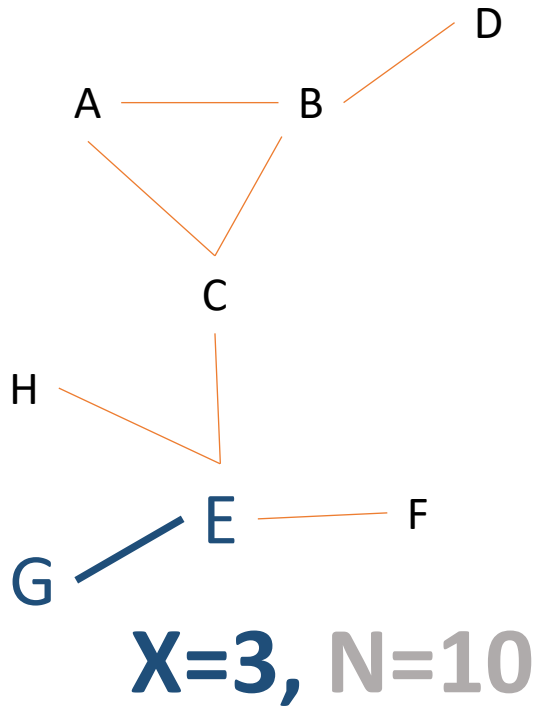
n=7

How often will a **random sample** of dyads from the **full weighted network** produce the **observed number** of "statements" (semantic network ties) in a group?

Is it significant that **G-E occurs 2 times in the sample**?

What is the probability of selecting **2 blue balls** from a **sample of 5 balls**?

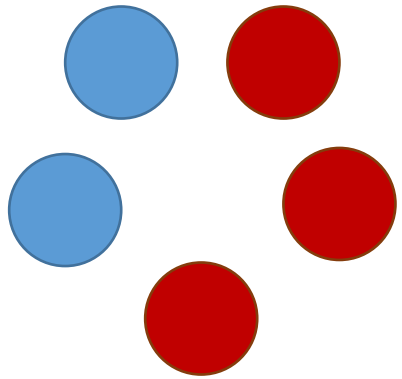


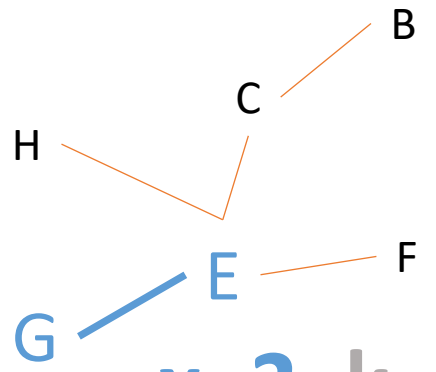


X=3, N=10

$$\Pr(\,blue = 2 \mid n = 5\,) = \frac{\binom{3}{2}\binom{7}{3}}{\binom{10}{5}}$$

$$= 0.42$$





x=2, k=5

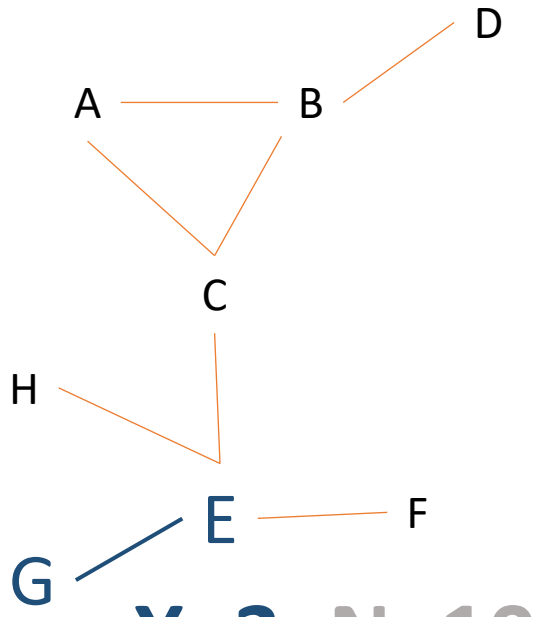
Generalized:

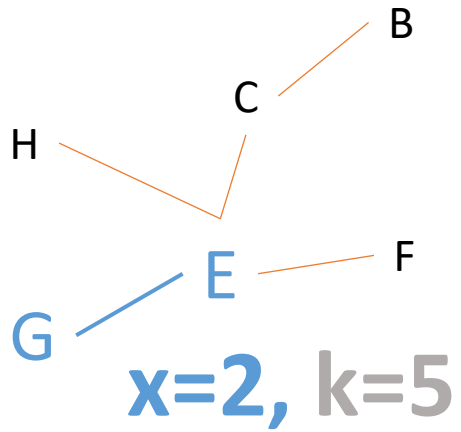
$$\Pr(\mathit{StatementCount} = x \mid \mathit{sample} = k) = \frac{\binom{X}{x}\binom{N-X}{k-x}}{\binom{N}{k}}$$

$Where\ X = the\ number\ of\ time\ a\ statement\ appears$

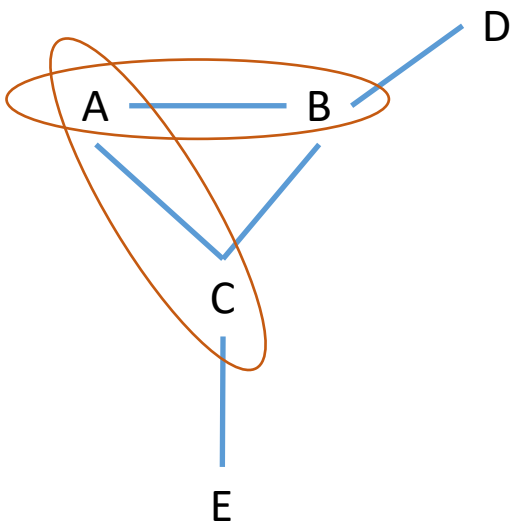
$N = total\ number\ of\ statements$

$k = number\ of\ statements\ in\ a\ specific\ period\ or\ group$

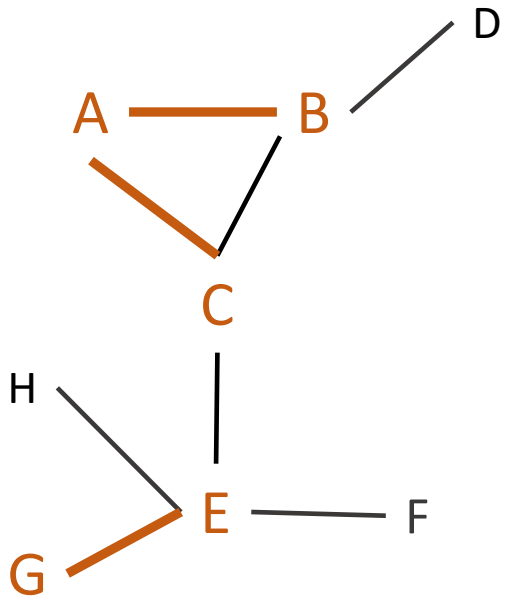




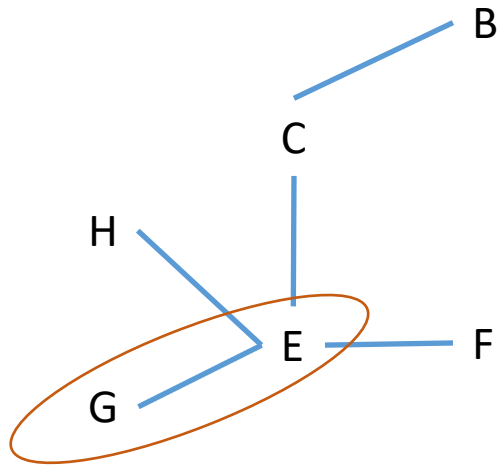
Mission Statement 1



Distinct Statements

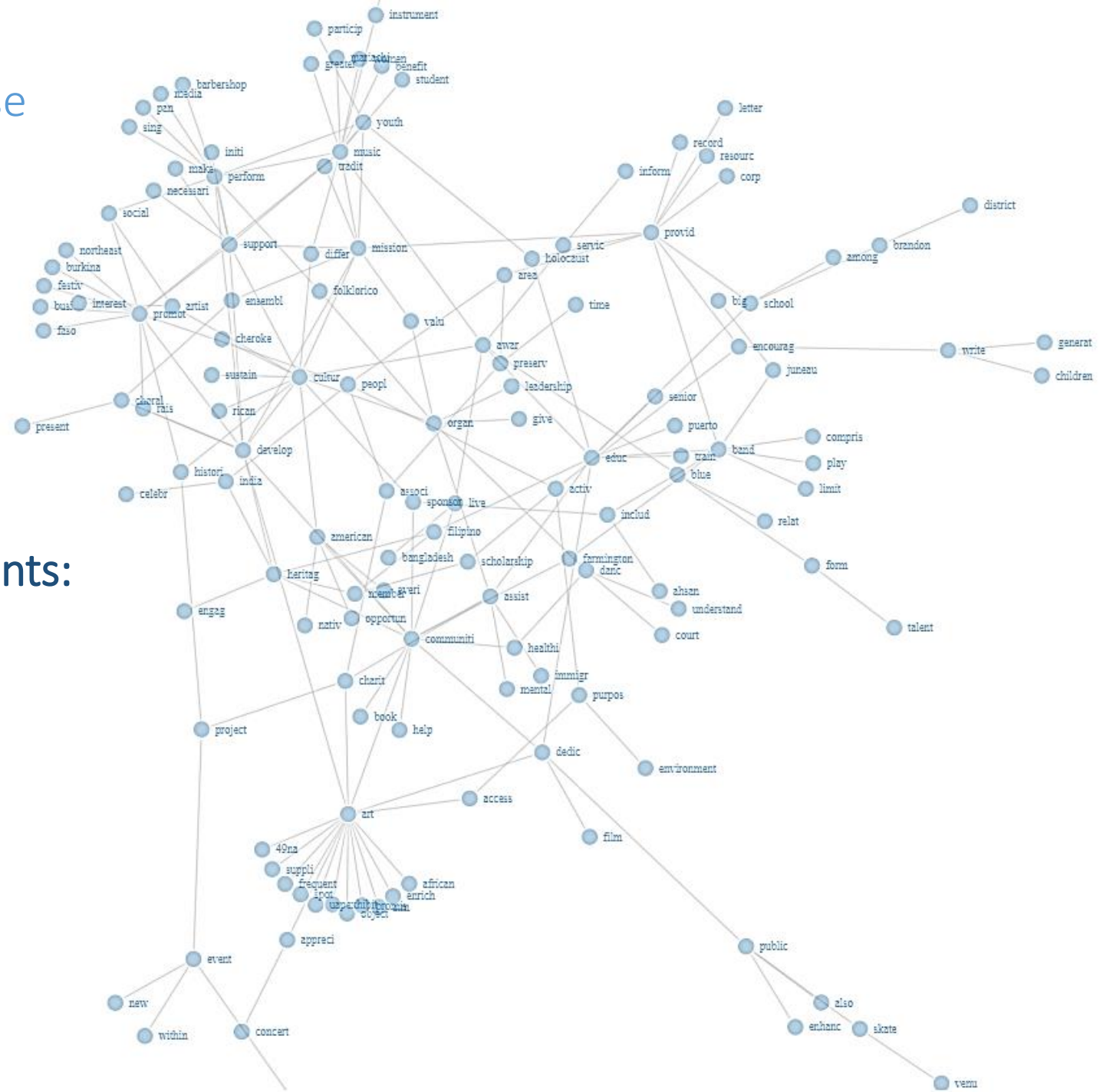


Mission Statement 2



Noise reduction: unique “statements” that occur more often than would be predicted by chance

Only distinct edges retained those with observed frequencies that would occur less than 5% of the time by chance



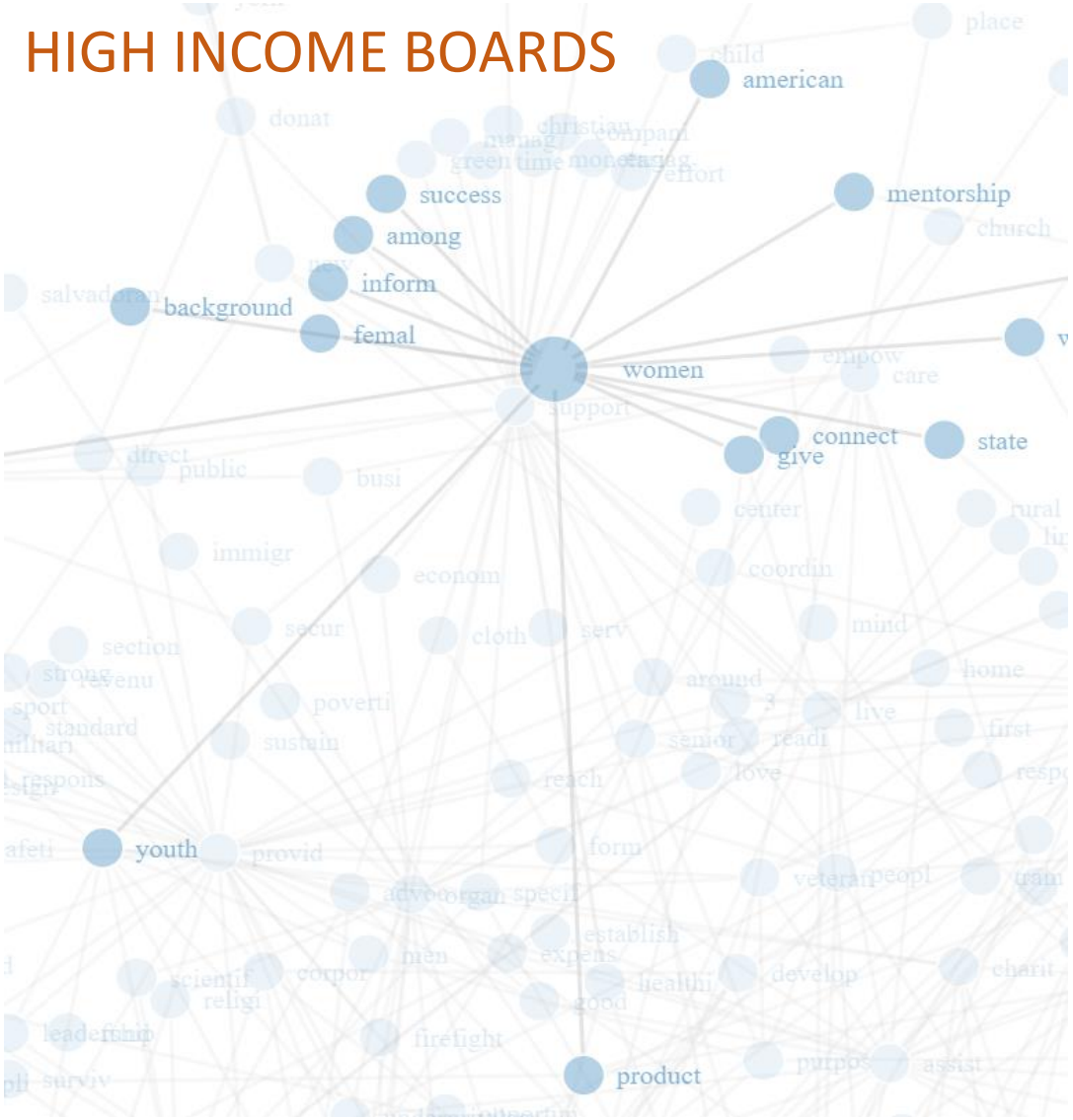
Each Semantic Network Represents:

COMMUNITY: LOW INCOME

SUBSECTOR: HEALTHCARE

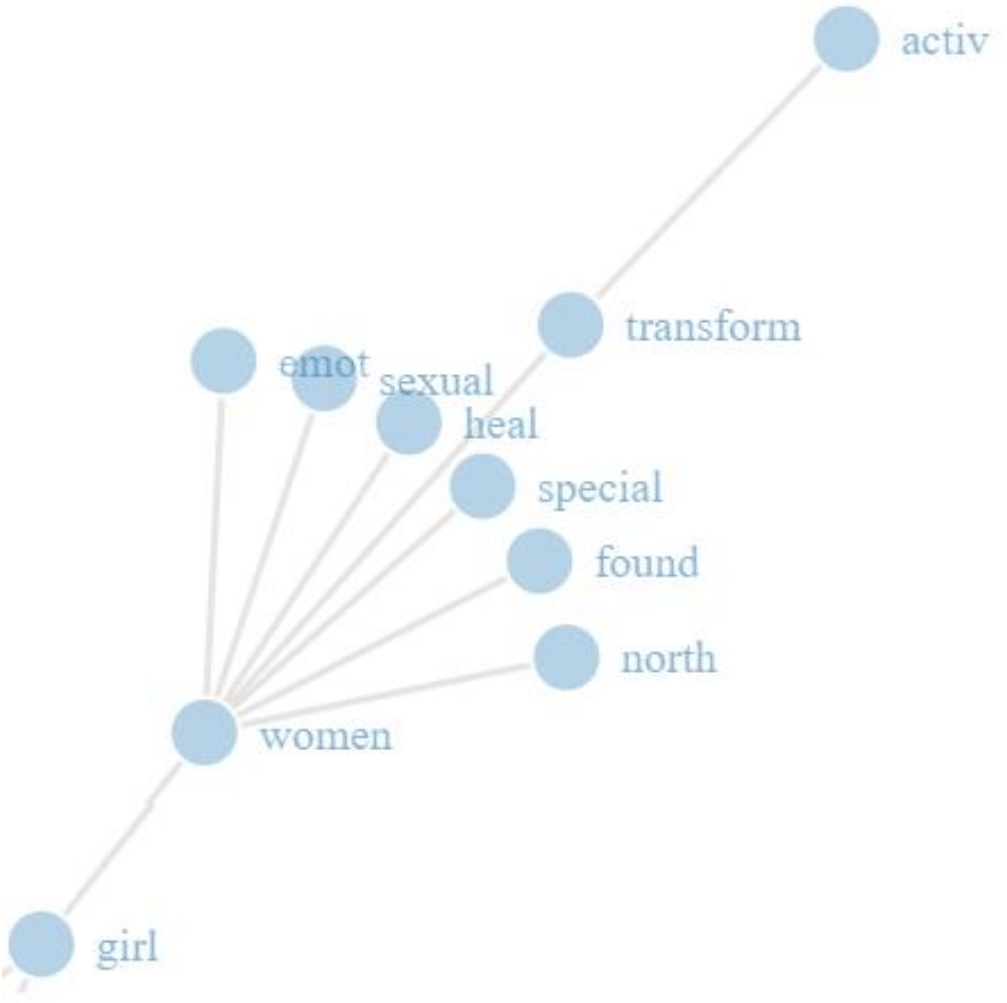
BOARD STATUS: LOW INCOME

HIGH INCOME BOARDS



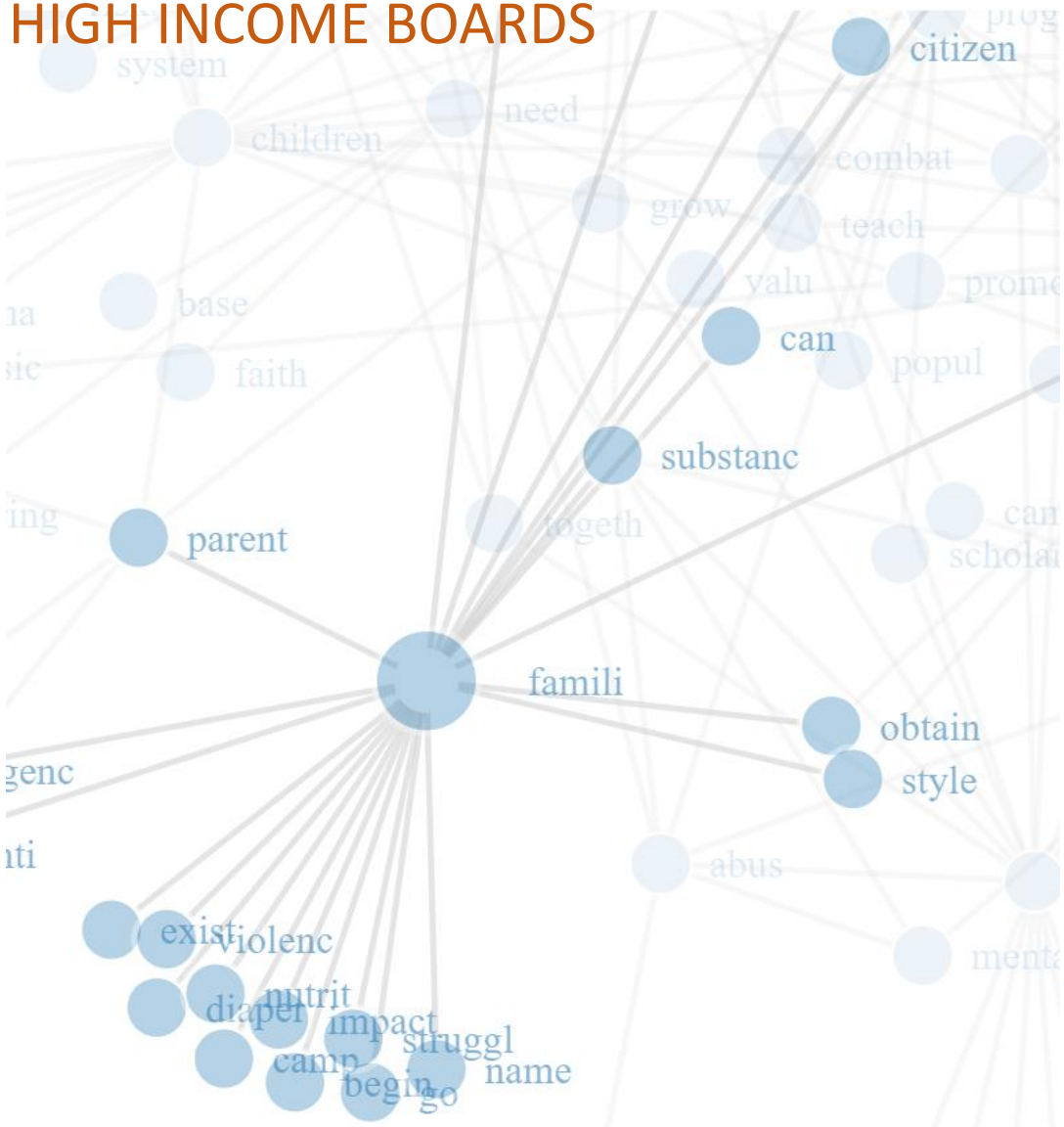
HUMAN SERVICES: LOW INCOME NHOODS

LOW INCOME BOARDS



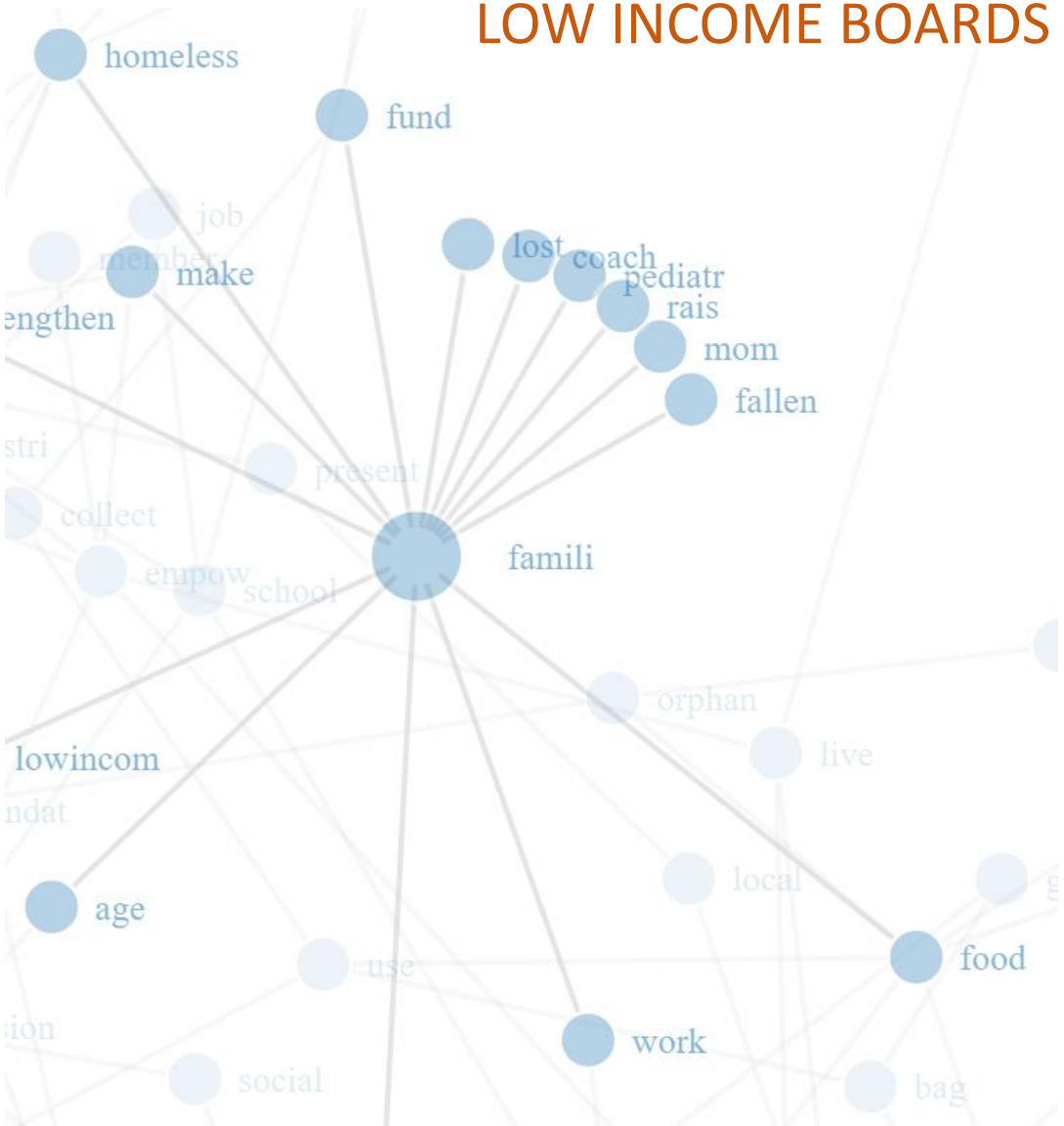
HUMAN SERVICES: LOW INCOME NHOODS

HIGH INCOME BOARDS



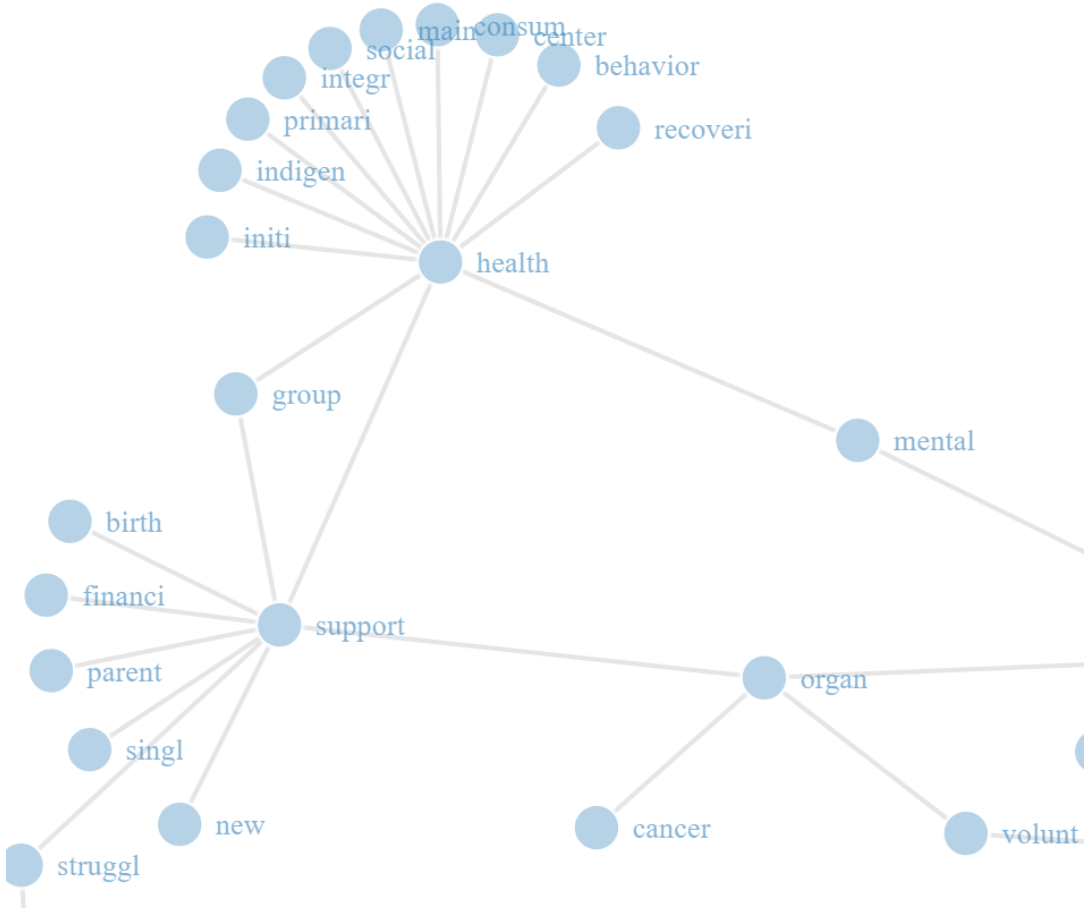
HUMAN SERVICES: LOW INCOME NHOODS

LOW INCOME BOARDS



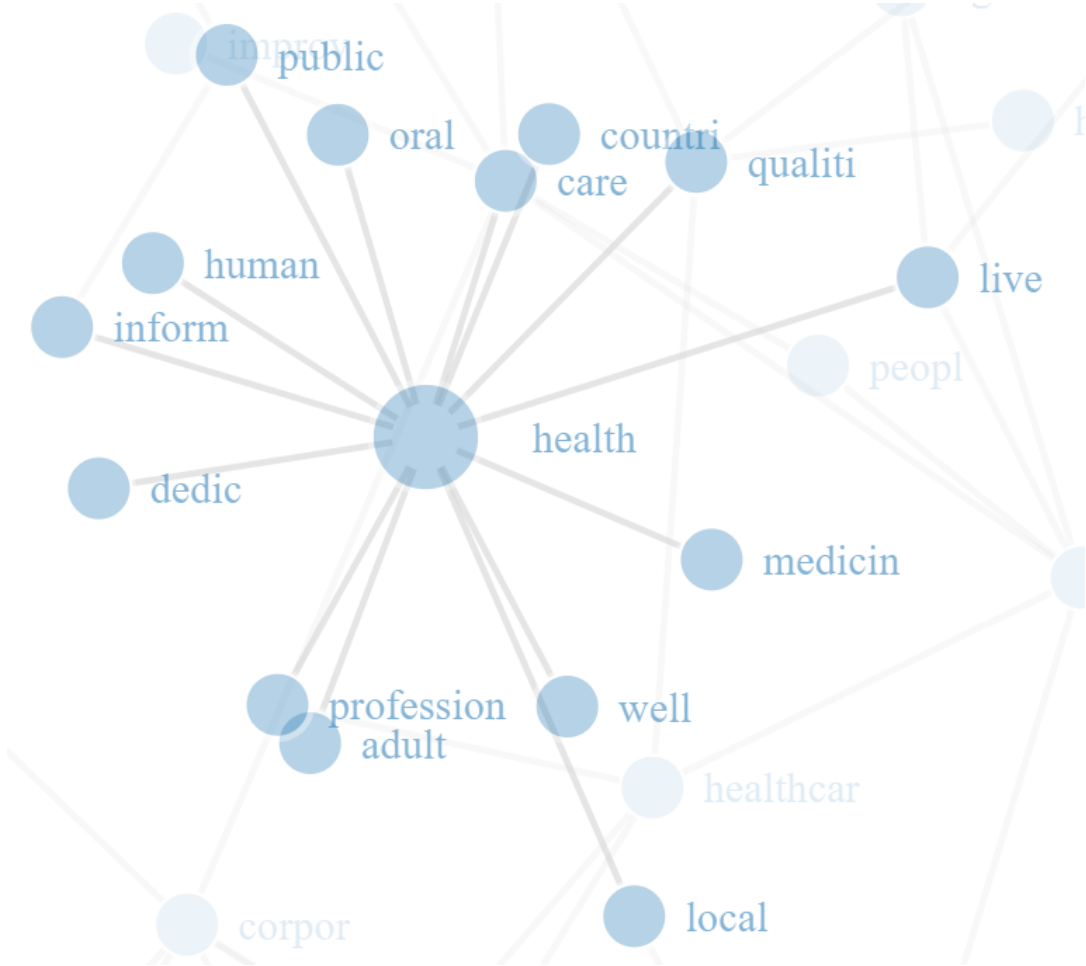
HUMAN SERVICES: LOW INCOME NHOODS

HIGH INCOME BOARDS



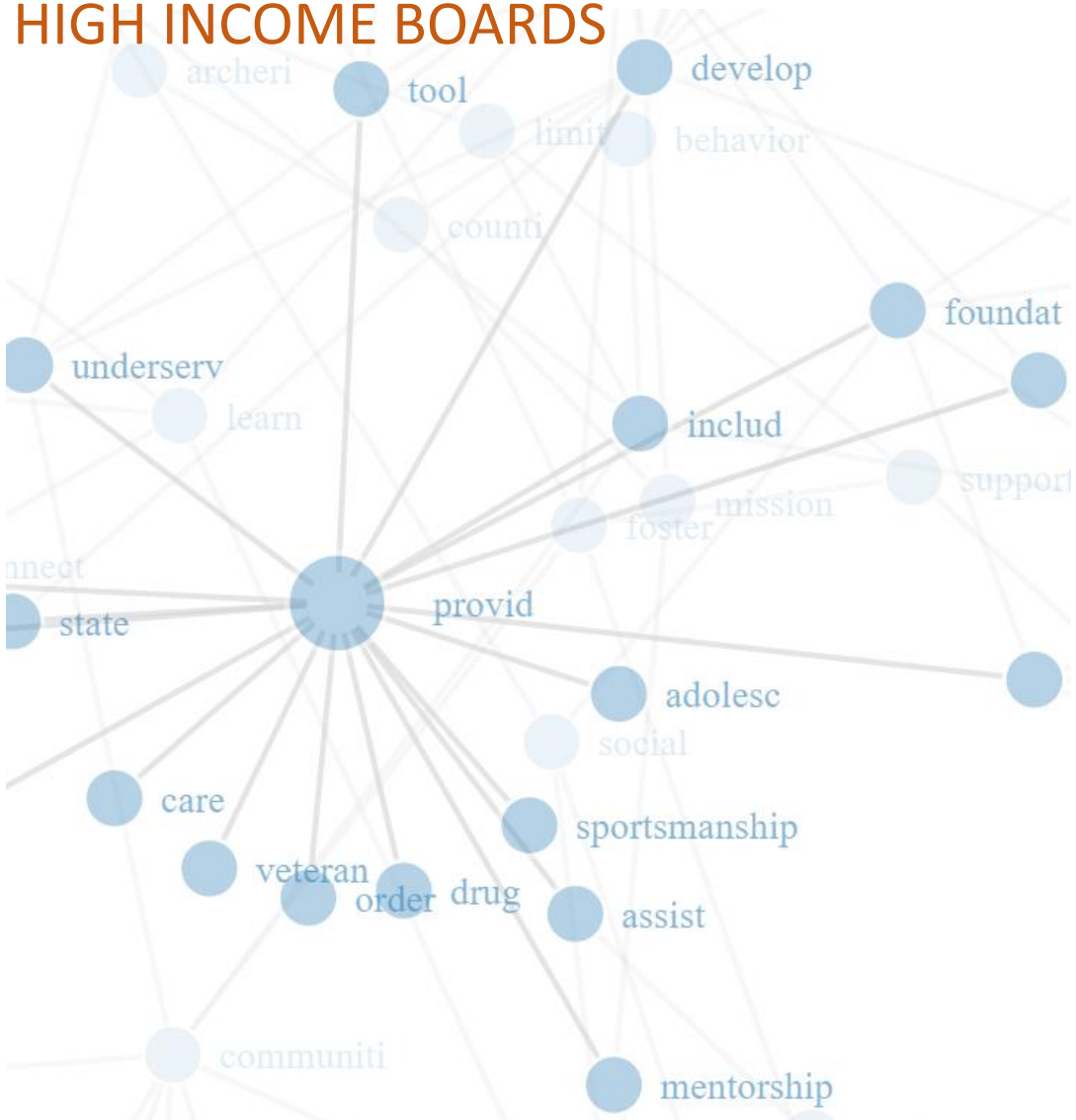
HEALTHCARE: LOW INCOME NHOODS

LOW INCOME BOARDS



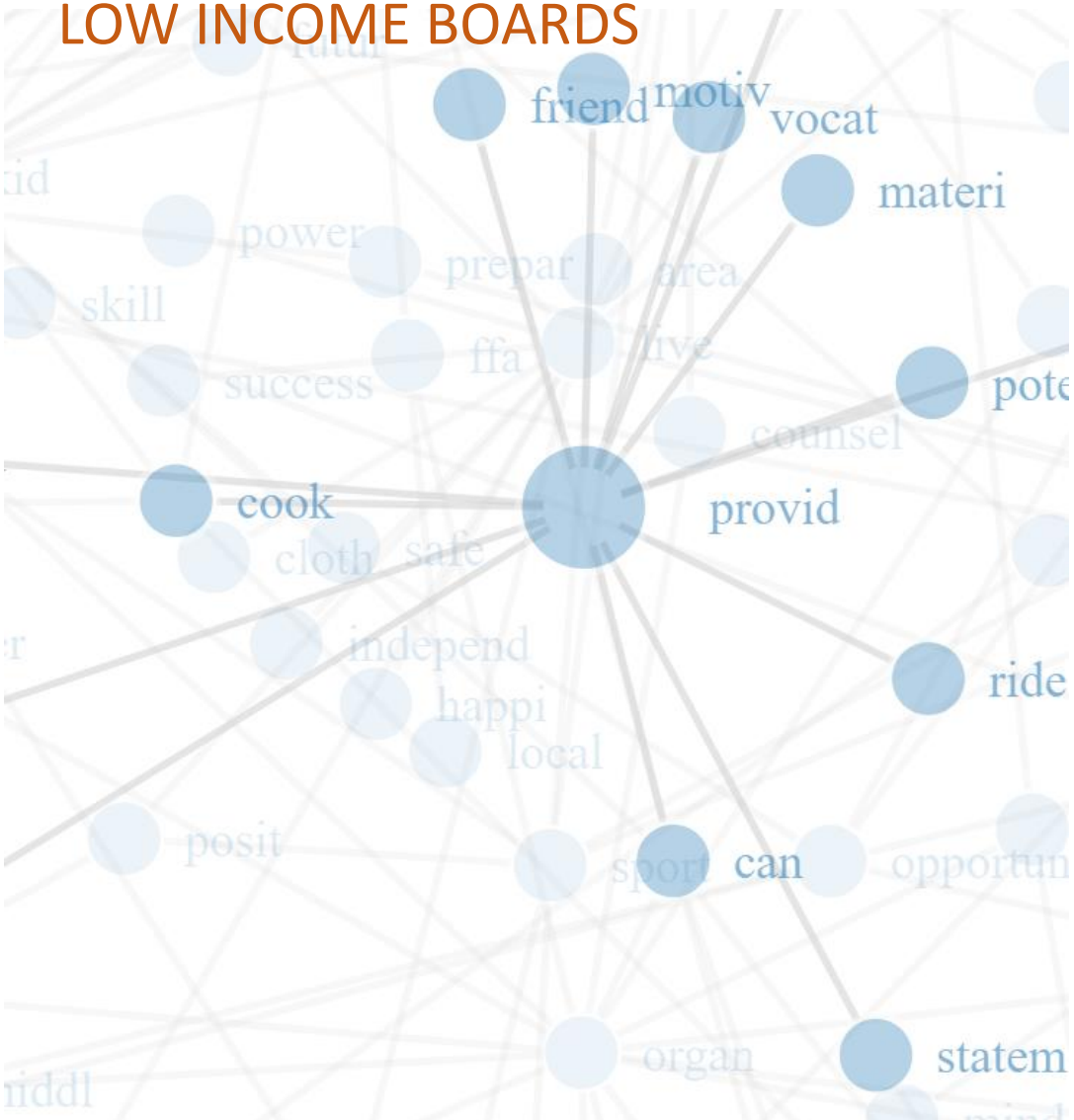
HEALTHCARE: LOW INCOME NHOODS

HIGH INCOME BOARDS



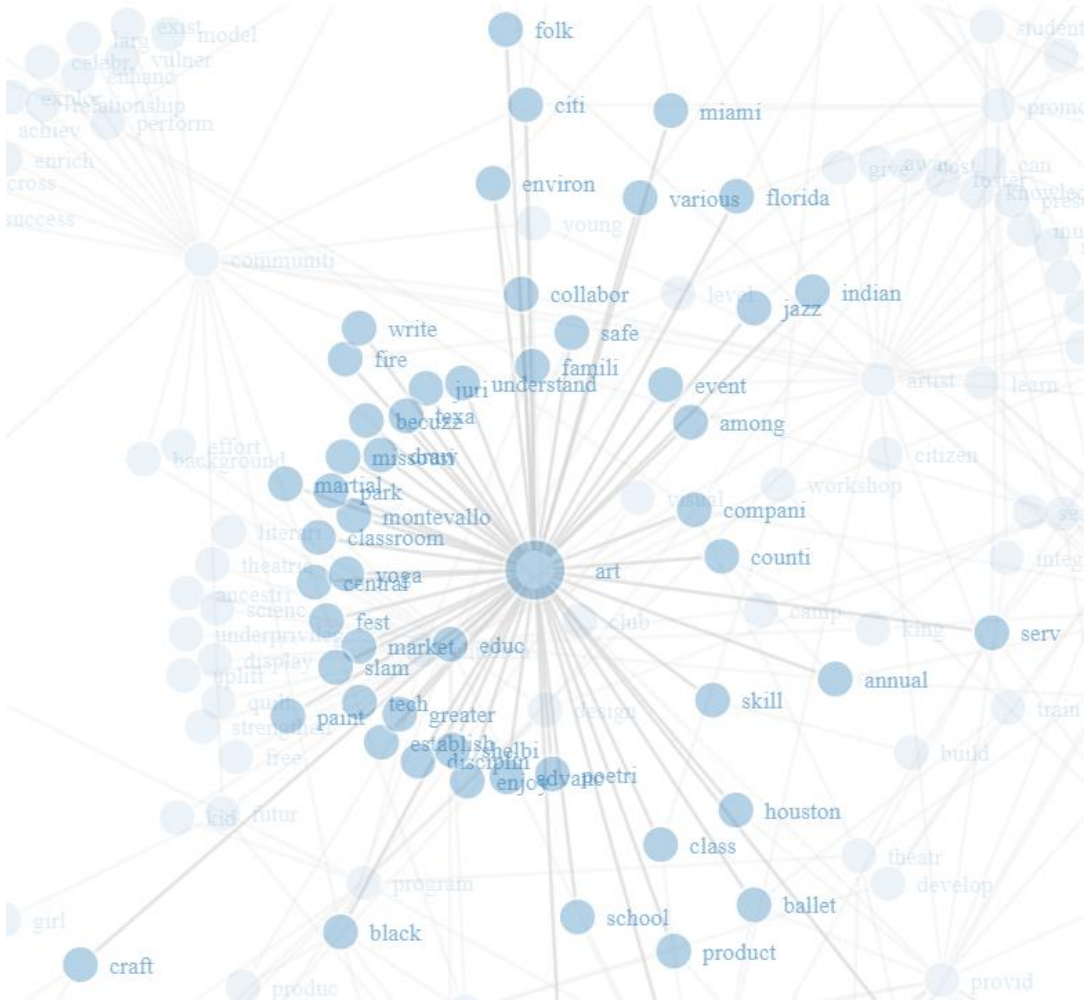
YOUTH: LOW INCOME NHOOD

LOW INCOME BOARDS



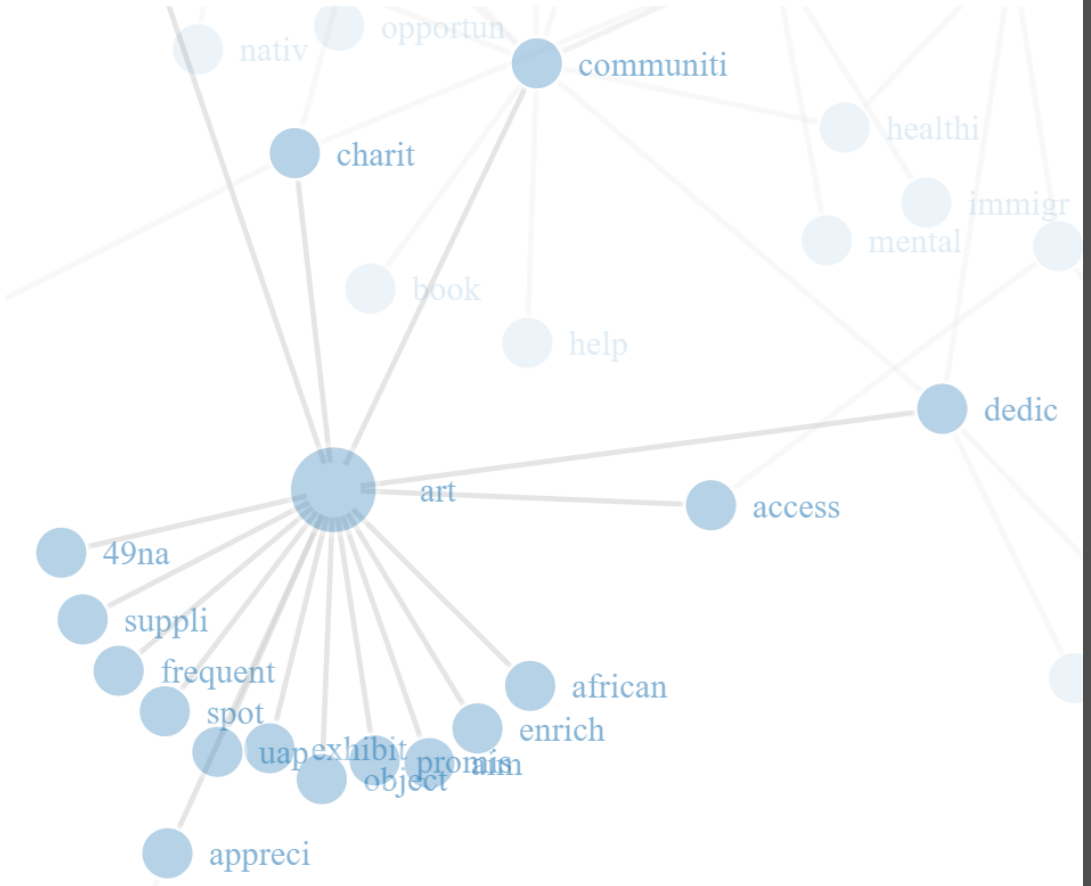
YOUTH: LOW INCOME NHOOD

HIGH INCOME BOARDS



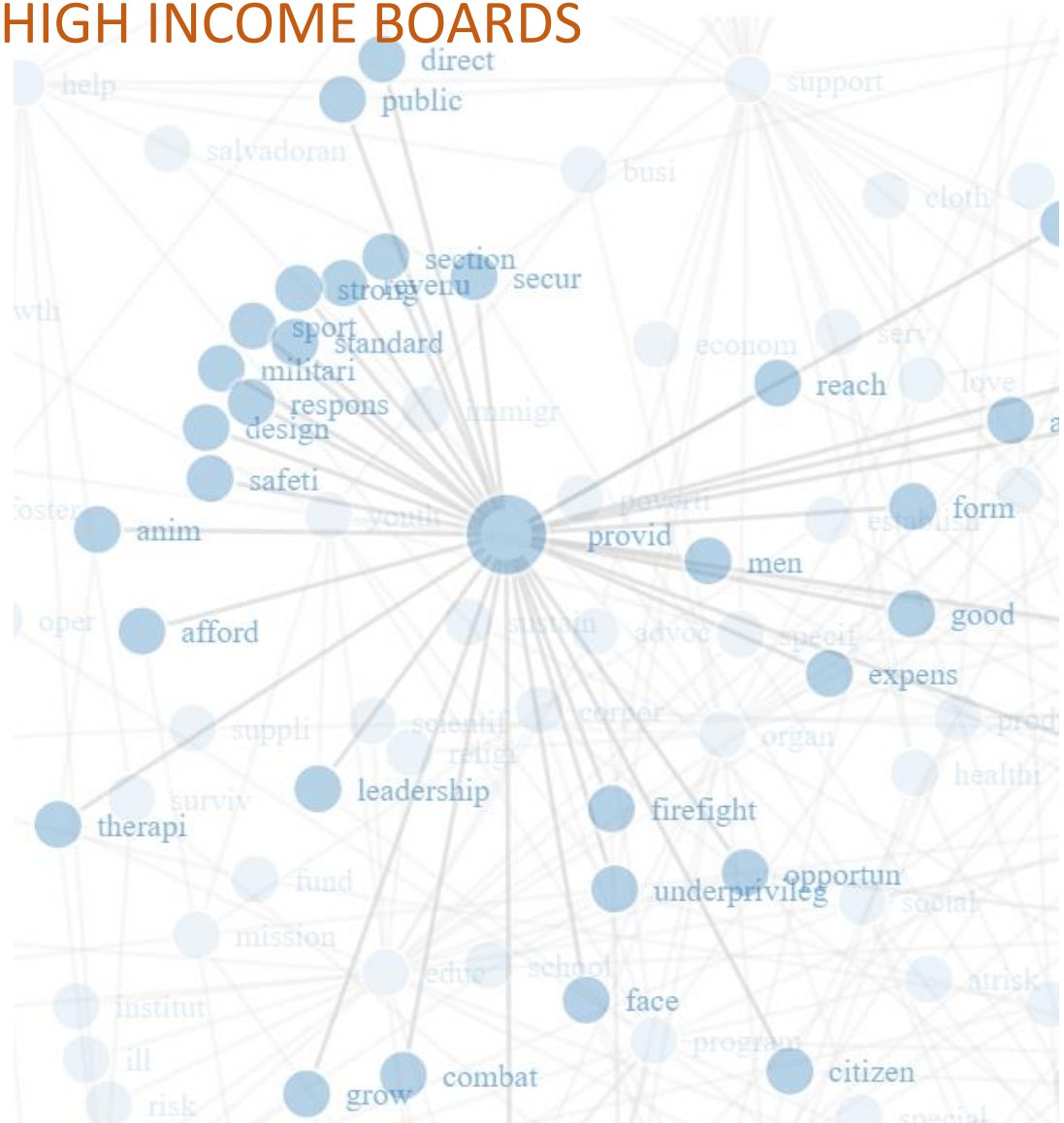
ARTS: LOW INCOME NHOOD

LOW INCOME BOARDS



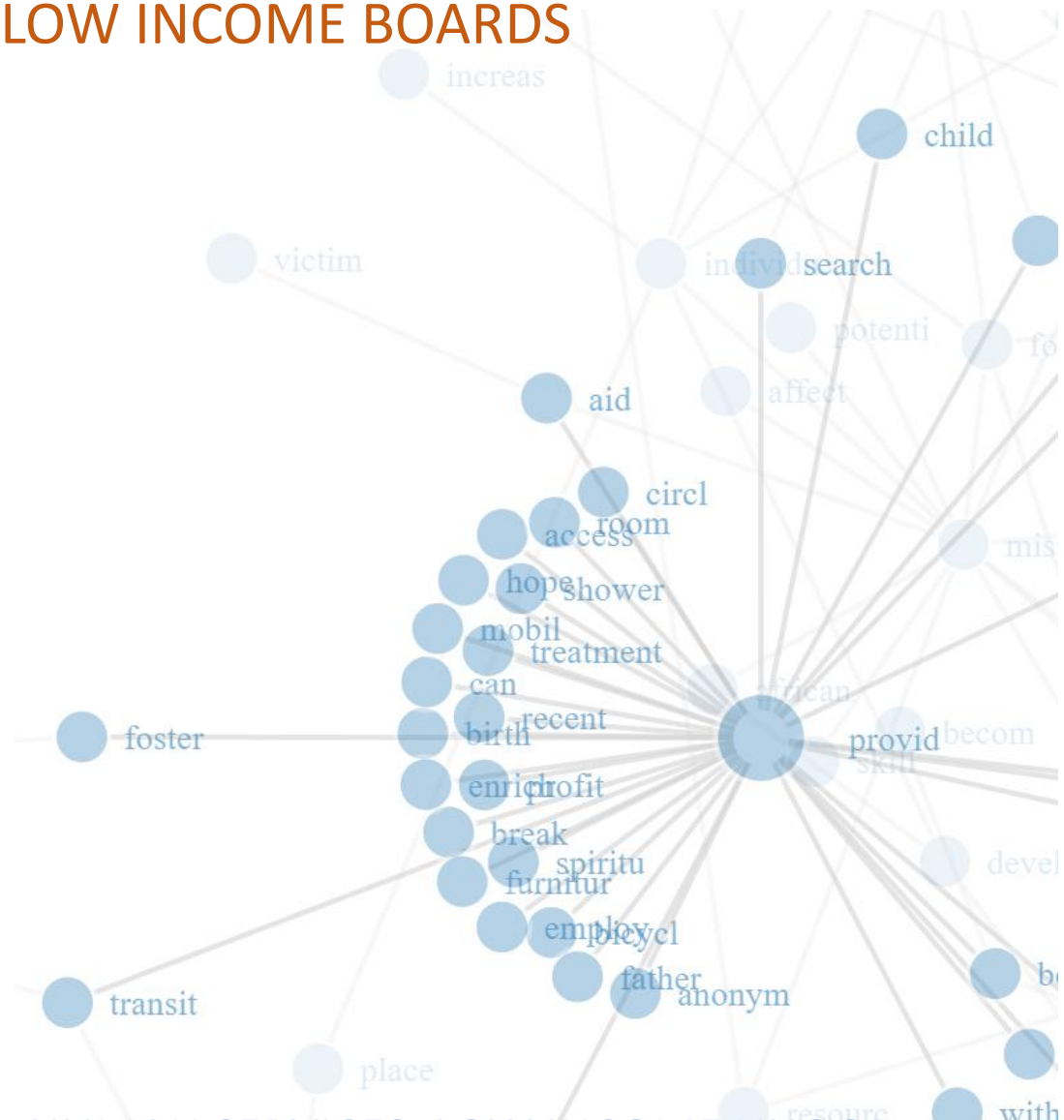
ARTS: LOW INCOME NHOOD

HIGH INCOME BOARDS



HUMAN SERVICES: LOW INCOME NHOOD

LOW INCOME BOARDS



HUMAN SERVICES: LOW INCOME NHOOD

Nonprofit Open Data Collective

THE ASPEN INSTITUTE | URBAN INSTITUTE | CHARITY NAVIGATOR Your Guide To Intelligent Giving

PROPUBLICA | open990 | GUIDESTAR

ASU Arizona State University | Carleton University | Syracuse University Maxwell School

Nonprofit Mission Classifiers

https://nonprofit-open-data-collective.github.io/machine_learning_mission_codes/

Methods Taxonomies Data R Packages Model Assessment About

Creating Machine Learning Classifiers for Nonprofit Mission Statements

Having clear taxonomies or categorical variables that describe nonprofit program activities makes many forms of data more theoretically meaningful and practially useful. They can be used to organize grants, examine collective impact from a set of programs, or find nonprofits with similar purposes.

This is a set of replication files and vignettes that demonstrate the task of using mission and program service accomplishment text from admistrative tax forms to predict mission activity codes such as the NTEE.

Creating accurate classifier models that are trained on large, readily available archives allow for the models to be used with custom text repositories such as grant data, reports, social media text, etc.

The goal of this project is to provide a robust set of test data, a reasonable set of text pre-processing steps, and examples of useful classifiers to lower the entry barriers for others that would like to engage with the work and offer some benchmarks for performance.

As such, we are following an open science model where all data and code used to produce this analysis is accessible and extensible through Creative Commons licensing.