High-Impact Lobbying

Joseph Decker, Kevin Kauffman, Maria Markhelyuk, Mark Takagi, and Zachary Taylor

The Problem - Business Value

- The Problem: Lobbying takes a lot of effort and money.

- Try to better understand politicians’ voting records.
  - Use machine learning techniques.
  - Maybe we can target politicians better to maximize our lobbying efforts.

- Each team member undertook a different analysis to look at the problem from a different angle.

Data Work

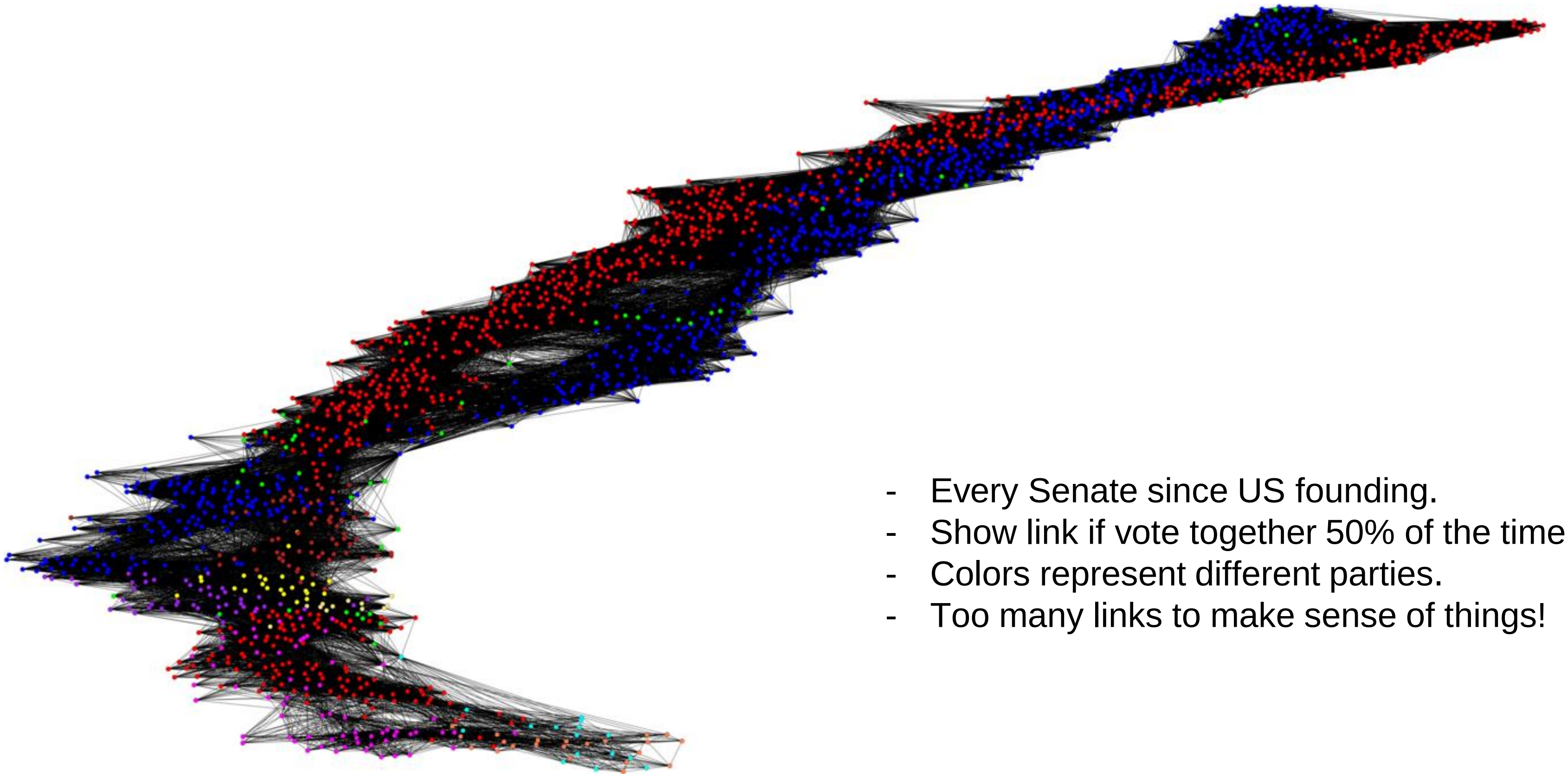
Data from voteview.com/dwnl.htm

```
1.  Congress Number
2.  ICPSR ID Number:  5 digit code assigned by the ICPSR as
                      corrected by Howard Rosenthal and myself.
3.  State Code:  2 digit ICPSR State Code.
4.  Congressional District Number (0 if Senate)
5.  State Name
6.  Party Code:  100 = Dem., 200 = Repub. (See PARTY3.DAT)
7.  Occupancy:  ICPSR Occupancy Code -- 0=only occupant; 1=1st occupant; 2=2nd occupant; etc.
8.  Last Means of Attaining Office:  ICPSR Attain-Office Code -- 1=general election;
             2=special election; 3=elected by state legislature; 5=appointed
9.  Name
.0 - to the number of roll calls + 10:  Roll Call Data --
                              0=not a member, 1=Yea, 2=Paired Yea, 3=Announced Yea,
                                              4=Announced Nay, 5=Paired Nay, 6=Nay,
                              7=Present (some Congresses, also not used some Congresses),
                              8=Present (some Congresses, also not used some Congresses),
                              9=Not Voting
```

```
 1  Federalist
 9  Jefferson Republican
10  Anti-Federalist
11  Jefferson Democrat
13  Democrat-Republican
22  Adams
25  National Republican
26  Anti Masonic
29  Whig
34  Whig and Democrat
37  Constitutional Unionist
```

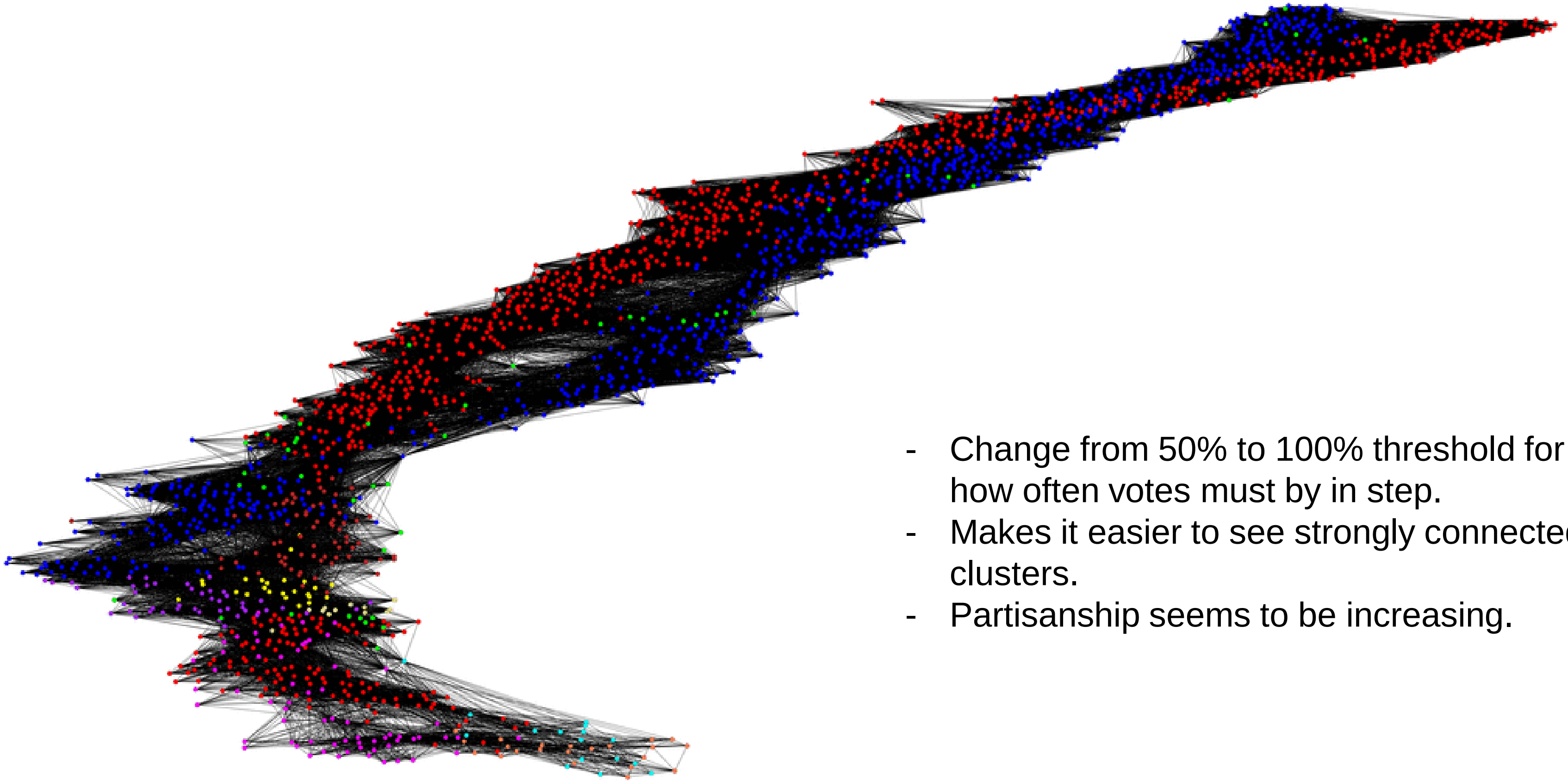
```
1139991199 0USA     100  OBAMA
999999999999699999999199999999999999999999699999999961999999
999999999999699969999991996919999999999969969999996999999999
999999999999699999996999999969999999999996999999999199999999
999969666999966966969969619969669969969969699999991999999999
999961999999999999999999699969999999999999999699999999969999996
999999999999999999969999999999999999999199199999969999999999
999999999999999999919999999999999999999999996969999999119199999
61969969999999999999999996996999999991619
1132030041 1ALABAMA 20011BONNER
111619116116611111116116111611111666111111111116111111611111116
119999999999999999911116161666666616116661161111116111116111111
669661166661916111611111166161111111166611111666666666666611
000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000
1132137641 1ALABAMA 20022BYRNE
000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000
111611111111666166666666161611191116111111616166666661111661
611111111116111116161661161111116666111111661166611116116666 6
116666166111116161111666661661111111111111161619999999999999
1661661111111111116111116116111111111111
```

A look at the US Senate



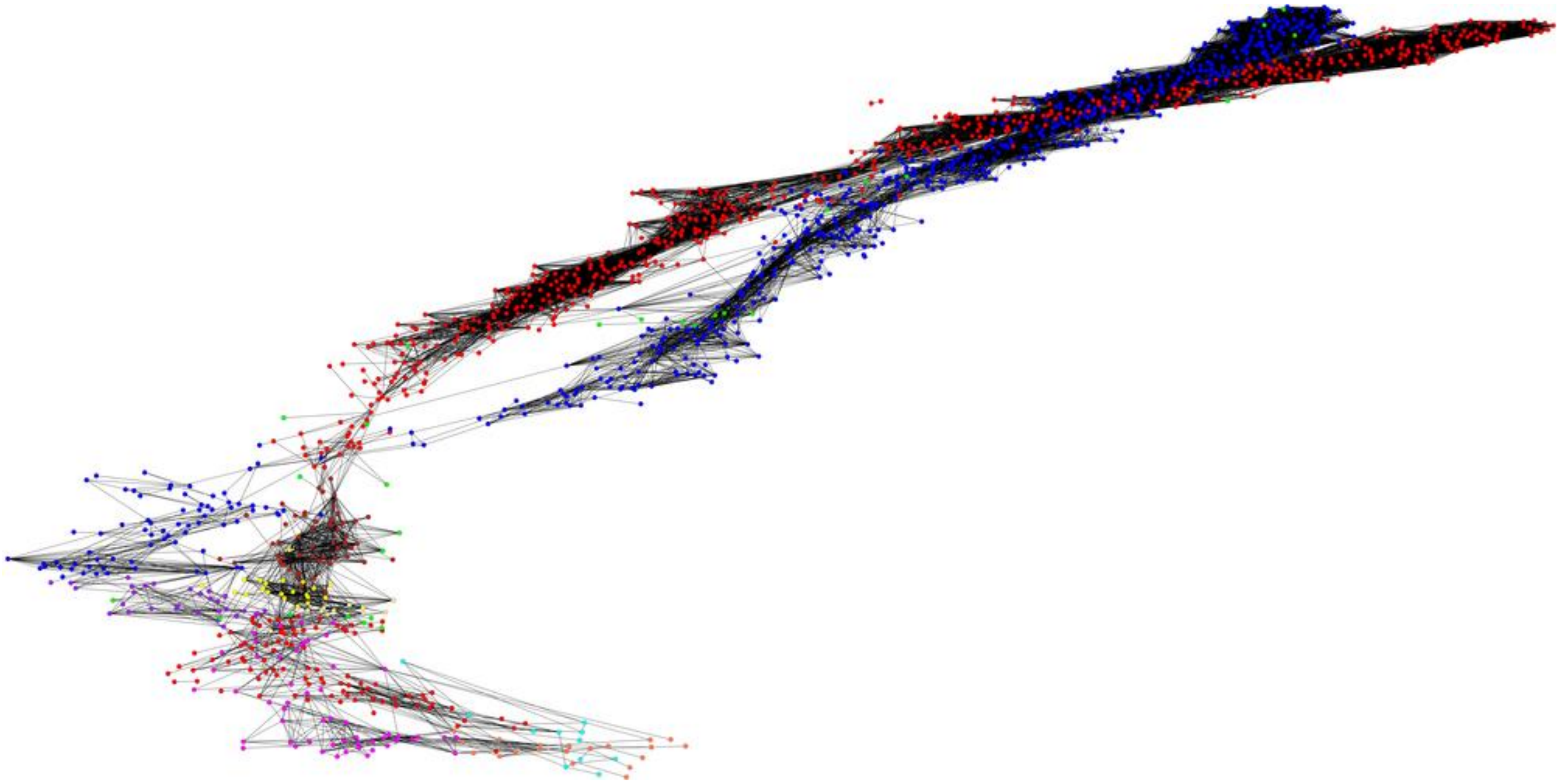
- Every Senate since US founding.
- Show link if vote together 50% of the time.
- Colors represent different parties.
- Too many links to make sense of things!

Animated Data Visualization

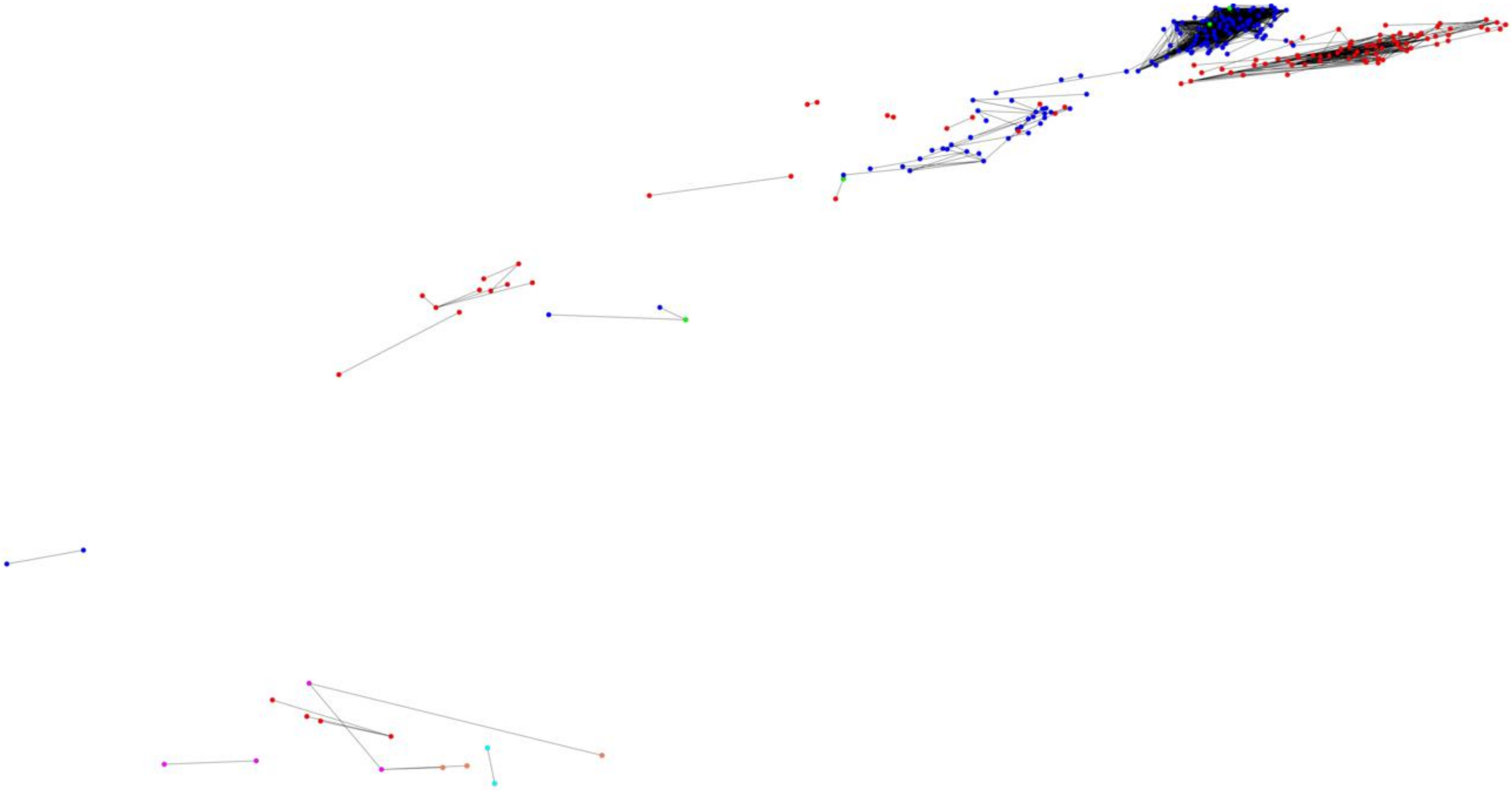


- Change from 50% to 100% threshold for how often votes must by in step.
- Makes it easier to see strongly connected clusters.
- Partisanship seems to be increasing.

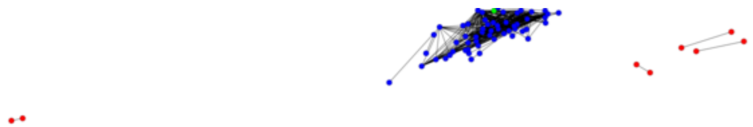
Animated Data Visualization - Vote Together 75%



Animated Data Visualization - Vote Together 90%



Animated Data Visualization - Vote Together 95%

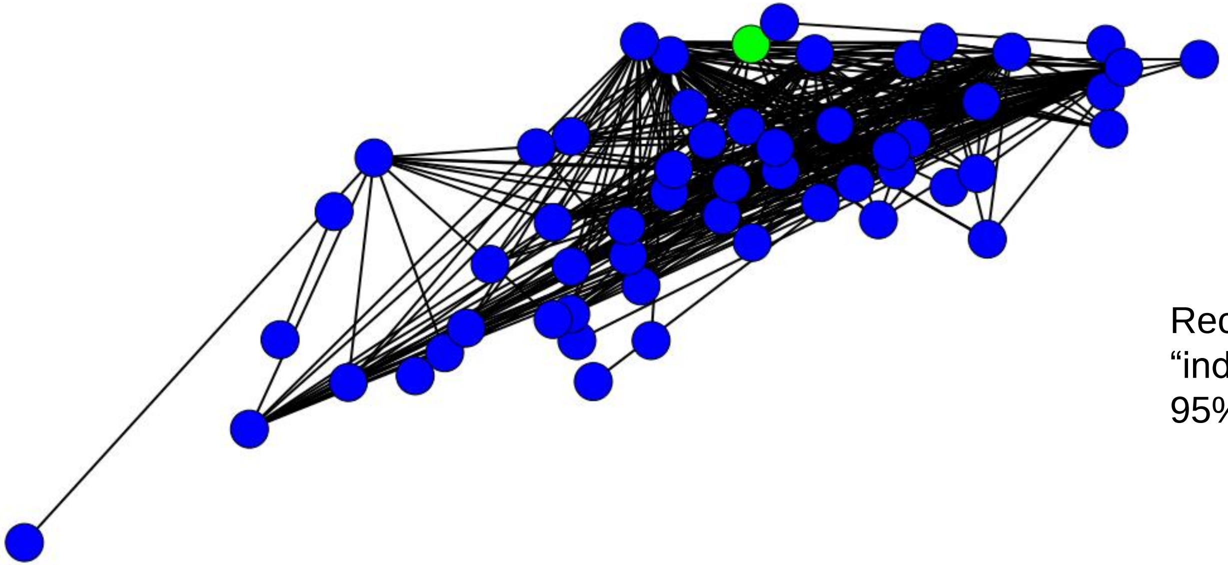


Implied Lobbying Strategy

- We can identify groups that nearly always vote in concert.

- Target highly connected (influential) individuals in such a group.

conflict is high.



Recent democrats and one "independent" that vote together 95% of the time.

Clustering In Spark

Clustering Algorithms, especially on large data-sets can see great speedup when executed in a parallel environment

Take well-known clustering algorithm which is slow on large data-sets and re-implement in spark to take advantage of high-performance clusters (K-medioids)

Generate output of actual clusters/cluster centers

K-Medioids

like K-means, but suitable for a generalized distance metric (select a vertex as the center of a cluster rather than an arbitrary point in space)

Also like K-means, can have poor runtime on large datasets

Only a heuristic solution: depending on randomized selection of initial medioids, may produce different clusterings. Is prone to getting caught in non-optimal local minima

Metric: sum of cosine distance from a medioid to all points in the cluster

Can be used to see which senators have influence at a variety of levels (use small K for broad influence, large K for fine grained influence)

Performance

Step 1: Assign a node to the nearest medioid.

Normally Θ(n*k)

With enough hardware, this can approach O(k * log(n))

Θ(k) for a single vertex to find the medioid it’s closest to

O(k * log(n)) to coalesce the individual nodes into their cluster

Step 2: find the new optimal medioid for each cluster

Normally O(n^2)

Reduced to O(n + k * log(n))

O(n) for each vertex to calculate its metric if it were the new medioid of the cluster

Experiment

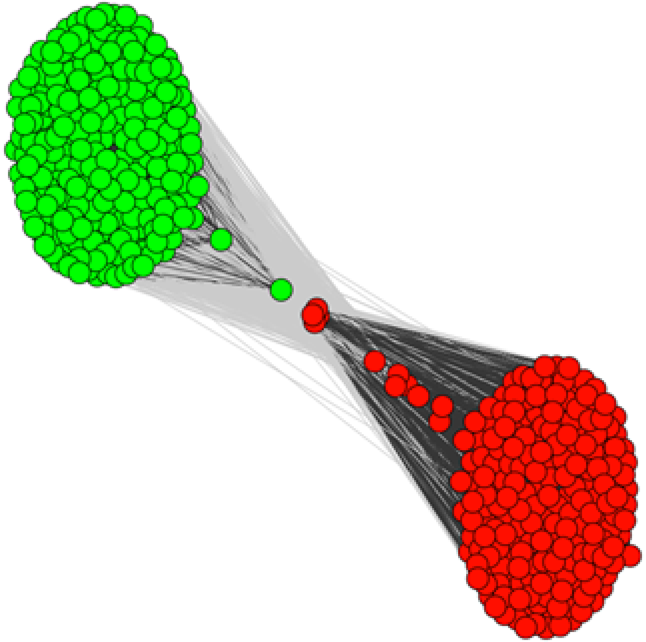
- Analyzed senate data from last 30 years
  - preprocessing the graph outside of spark was the limiting factor!
- Ran 100 iterations on a variety of cluster sizes
- Small number of senators were at the center of most clusters regardless of Cluster size
  - McConnell and Mikulski 1-2 across cluster counts (center of Rep and Dem parties!)
- Some senators only show up in smaller groups
  - Cardin #3 when k=4, but does not appear when k=2
- Some senators only show up when the groups are large

Clusters in iGraph

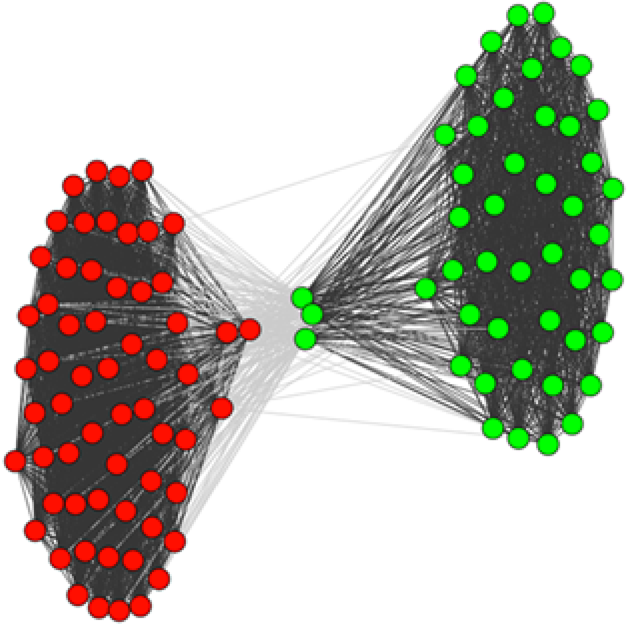
- Perform cluster analysis individually on each of the previous 13 houses and senates, dating back to 1990
- Fast Greedy Modularity Optimization
  - Determine clusters based on optimal modularity
  - Good for large datasets, such as house data - other implementations proved to be too slow
- Capture cluster size and modularity for each

Example Clusters

113th House

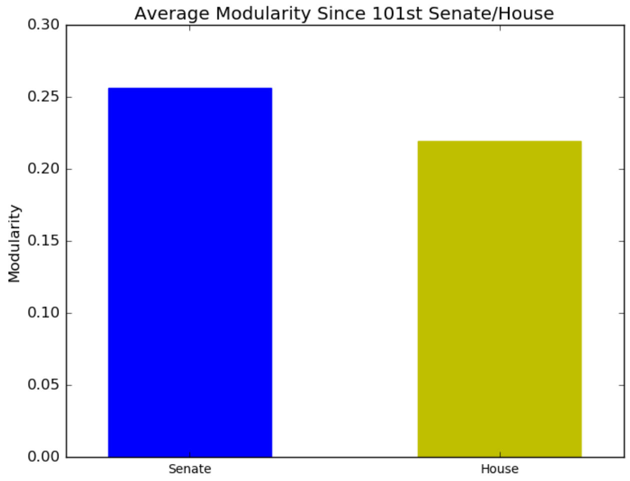


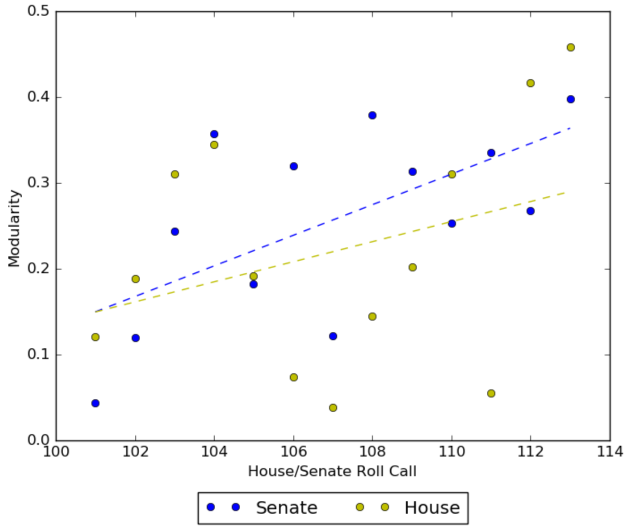
113th Senate



Results

- Pushing lobbying efforts towards House of Reps may be more effective than the Senate
  - House of Reps produced higher number of clusters, and lower modularity
  - Senate is more tightly coupled, and would be harder to influence an individual
- Interesting to note the trend over the past 25 years, showing modularity increasing, likely harder to influence individuals





A Comparison of Community Detection Algorithms

Fast Greedy Algorithm

Each node belongs to a separate community initially, and nodes are merged iteratively. A merge is performed only if it leads to the maximum increase in modularity. Merging stops once modularity can't be improved any further.

Walktrap Algorithm

Short random walks of 3 - 5 steps (depending on step parameter) are performed and the results are used to merge separate communities from the bottom up (as in Fast Greedy). Modularity is used to choose where to cut the resulting dendrogram.

Leading Eigenvector Algorithm

Initially, all the nodes belong to one community. The graph is iteratively divided into two clusters such that the division results in a significant increase in modularity. A modularity matrix is computed and the corresponding eigenvector is used to determine the split.

A Comparison of Community Detection Algorithms

Data:

Roll Call data from the 100^{th} to the 113^{th} U.S. Congress
(Same as in previous experiment to provide point of comparison.)

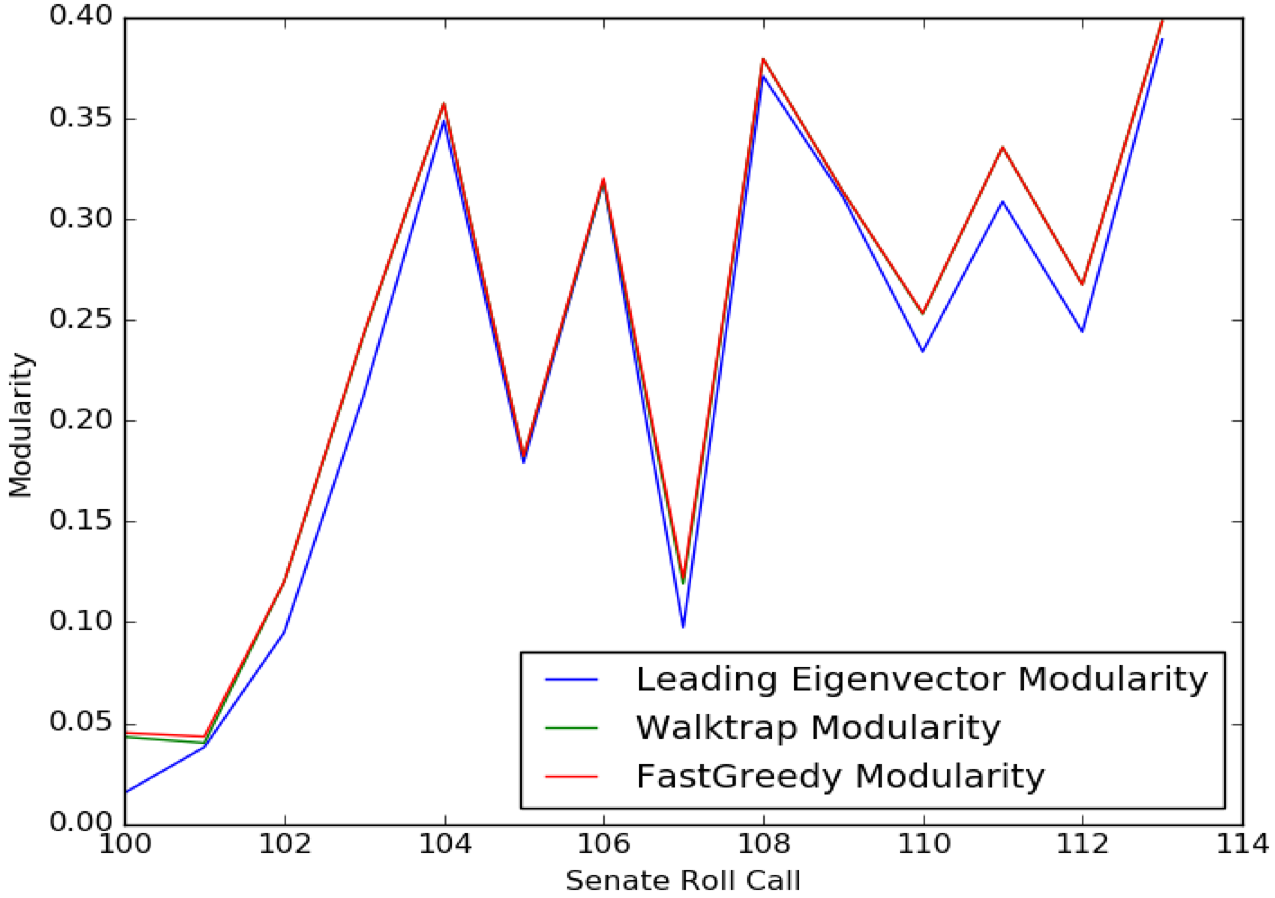
Analysis:

Performed Fast Greedy, Walktrap, and Leading Eigenvector community detection on each graph in the dataset using igraph. Compared the quality of the resulting partitions using the modularity scores obtained. (Using modularity since the actual structure of the communities we'd like to target for lobbying purposes is unknown and modularity provides an approximation of goodness.)

Results:

Walktrap and Fast Greedy produce nearly the same modularity values, while Leading Vector results in slightly lower modularity scores. Difficult to distinguish quality based on modularity, perhaps coverage or conductance metrics would result in more meaningful comparisons in future analyses.
(See next slide for plot.)

A Comparison of Community Detection Algorithms



Data - Digging Deeper

Can we tell what congress is passing bills on?

What additional features can be found if the corpus of bill text was available?

Using a separate BI process - load bill text

```
date,session,number,bill,question,result,description,yeatotal,naytotal
2013-01-03,1st,1,QUORUM,Call of the House,Passed,,0,0
2013-01-03,1st,2,,Election of the Speaker,Boehner,,,
2013-01-03,1st,3,H RES 5,On Motion to Table the Motion to Refer,Passed,Adopt
2013-01-03,1st,4,H RES 5,On Ordering the Previous Question,Passed,Adopting r
2013-01-03,1st,5,H RES 5,On Motion to Commit,Failed,Adopting rules for the C
2013-01-03,1st,6,H RES 5,On Agreeing to the Resolution,Passed,Adopting rules
2013-01-04,1st,7,H R 41,On Motion to Suspend the Rules and Pass,Passed,To te
Agency for carrying out the National Flood Insurance Program,354,67
2013-01-14,1st,8,H R 219,On Motion to Suspend the Rules and Pass,Passed,"To
purposes",403,0
2013-01-14,1st,9,JOURNAL,On Approving the Journal,Passed,,300,95
2013-01-14,1st,10,ADJOURN,On Motion to Adjourn,Failed,,4,397
2013-01-15,1st,11,H RES 23,On Ordering the Previous Question
the fiscal year ending September 30, 2013, and for other purposes",293,127
2013-01-15,1st,12,H RES 23,On Agreeing to the Resolution,Passed,"Providing
fiscal year ending September 30, 2013, and for other purposes",367,52
2013-01-15,1st,13,ADJOURN,On Motion to Adjourn,Failed,,0,419
2013-01-15,1st,14,H R 152,On Agreeing to the Amendment,Failed,,162,258
2013-01-15,1st,15,H R 152,On Agreeing to the Amendment,Agreed to,,327,91
2013-01-15,1st,16,H R 152,On Agreeing to the Amendment,Agreed to,,221,197
2013-01-15,1st,17,H R 152,On Agreeing to the Amendment,Failed,,206,214
2013-01-15,1st,18,H R 152,On Agreeing to the Amendment,Failed,,202,217
2013-01-15,1st,19,H R 152,On Agreeing to the Amendment,Agreed to,,216,205
2013-01-15,1st,20,H R 152,On Agreeing to the Amendment,Failed,,208,212
2013-01-15,1st,21,H R 152,On Agreeing to the Amendment,Agreed to,,223,198
```

```
[Congressional Bills 113th Congress]
[From the U.S. Government Printing Office]
[H.R. 3811 Referred in Senate (RFS)]

113th CONGRESS
  2d Session
                                H. R. 3811

_______________________________________________________________________

                    IN THE SENATE OF THE UNITED STATES

                             January 13, 2014

     Received; read twice and referred to the Committee on Health,
                   Education, Labor, and Pensions

_______________________________________________________________________

                                  AN ACT

   To require notification of individuals of breaches of personally
identifiable information through Exchanges under the Patient Protection
                       and Affordable Care Act.

    Be it enacted by the Senate and House of Representatives of the
United States of America in Congress assembled,

SECTION 1. SHORT TITLE.

    This Act may be cited as the ``Health Exchange Security and
Transparency Act of 2014''.

SEC. 2. NOTIFICATION OF INDIVIDUALS OF BREACHES OF PERSONALLY
              IDENTIFIABLE INFORMATION THROUGH PPACA EXCHANGES.

    Not later than two business days after the discovery of a breach of
security of any system maintained by an Exchange established under
section 1311 or 1321 of the Patient Protection and Affordable Care Act
(42 U.S.C. 18031, 18041) which is known to have resulted in personally
identifiable information of an individual being stolen or unlawfully
accessed, the Secretary of Health and Human Services shall provide
notice of such breach to each such individual.

            Passed the House of Representatives January 10, 2014.
```

Data Collection - Difficulties

Web Scraping - Difficulties getting the right content, still only about 40% correctly scraped

Large datasets - Spark handles multiple files easily!

Quality - words can get garbled in parsing i.e. xxiiiair

Incorporating with other graph data - difficult to match back but has possibility of mapping politicians to bill types

Collected approximately **19633023** words of bill text from 763 bills

Bills passed by congress - themes

Word	**Count**
defence	6513
fiscal	6943
funds	5359
security	3518
military	3137
budget	2328
energy	2103
housing	1948

Bills failed by congress - themes

Word	Count
land	2231
area	1148
water	1110
food	904
conservation	893
wilderness	879
river	605
farm	499

Word2Vec - Passed- Synonyms - health

```
synonyms = passed_congress_model\
          .findSynonyms('health', 40)

for word, cosine_distance in synonyms:
    print("{}: {}".format(word, cosine_distance))
```

marketing: 0.895839500546
competitions: 0.849055429027
economically: 0.848936713079
headstones: 0.84857053943
owned: 0.835485145274
correspondence: 0.829716005454
benchmarking: 0.821671952711
outreach: 0.820389391379
socially: 0.819080883186

Word2Vec - Failed Synonyms - health

priorities: 0.64083128422
systems: 0.63415806923
improve: 0.629074653369
communication: 0.62721723196
colleges: 0.622499133964
resource: 0.615561390819
products: 0.614292804891
upgrades: 0.605861872278
pest: 0.599340724698

References

- DW-Nominate - http://voteview.com/pdf/nomboot.pdf
- Fast Greedy Modular Optimization - http://arxiv.org/abs/cond-mat/0408187
- Community Detection Algorithms: a comparative evaluation on artificial and real-world networks - http://www.robots.ox.ac.uk/~yannis/psorakis-report1.pdf
- An Evalutation of Community Detection Algorithms on Large-Scale Email Traffic - http://www.syssec-project.eu/m/page-media/3/moradi-sea12.pdf
- A Comparison of Community Detection Algorithms on Artificial Networks - https://hal.archives-ouvertes.fr/hal-00633640/document