Data Mining the Underlying Trust in the US Congress

Xiaoxiao Wu

Collaborators: **Anna Scaglione, Hoi-To Wai**

School of Electrical, Computer and Energy Engineering, Arizona State University, USA

Acknowledgement: NSF CCF-1011811.

A GlobalSIP Presentation, December 2016.

Social Signal Processing

- There is an increasing interest in mining human relationships and action patterns — integrating engineering science and social science.

Fig. Data Mining Human Actions and Minds.

(Left) From en.wikipedia.org (Right) From www.expertsglobe.com.

Understanding social network systems

- Investigate Congressman's relationship in the US Congress social network.

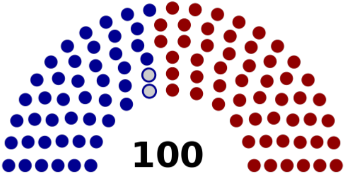
- *Opinion dynamic* model over the social network tells the trust between any Congressmen.

Data Mining the US Congress

- Our research target: the US Senate voting data (http://www.govtrack.us).
- What we can acquire from the dataset?
  - The contents of all the bills.
  - Roll call voting results.
  - Side information– the ideology and leadership scores.
- Our purpose: analyze the underlying trust information in the Senate.
  - Trust between any two Senators.
  - Trust between different groups: e.g., parties, genders, religions, states.

The US Senate

- Basic facts of the Senate from the 114th Congress:
  - There are 54 Republican Senators (red dots), 44 Democrat Senators (blue dots) and 2 Independent Senators (gray dots).
  - The Senators vote for the approval or disapproval of the bill.
  - The bills are sponsored by a group of congress members, and put forward to the Senate by the Committees.



- We consider data from the year 2015 and the portion of 2016 until March, 24th.

Prior Work

Statistical Perspective

- NOMINATE statistical procedure [Hare2014]
- Bayesian procedure [Clinton2004]
- Ideal point theory [Clinton2012]

Social Science Perspective

- History [Poole2000],
- Partisan [Fenno1982]

Machine Learning Perspective

Training and Prediction: PCA, MultiScheme, J48, RandomTree, RandomForest, ect.

An Overview of Our Methodology

- **The trust matrix**: The Senate social network is associated with a trust matrix $\overline{\boldsymbol{W}} \in \mathbb{R}_{+}^{N \times N}$: $[\overline{\boldsymbol{W}}]_{ij}$ denotes the trust strength of Senator i on Senator j. Final goal: estimate $\overline{\boldsymbol{W}}$.
- **Bill clustering and opinion estimation**: Senators vote for bills → group bills to clusters → extract Senators' opinion, i.e., a belief valued between 0 and 1 telling the probability of "YEA" on each cluster.
- **The DeGroot model**: We model the voting process as outcomes of underlying opinion dynamics that shape the beliefs of the Senators.

The Discuss-then-Vote Model with Stubborn Nodes

- Let $\boldsymbol{x}(0; k) \in \mathbb{R}^N$ be Senators' initial opinion on cluster k, i.e, the belief. As the discussion goes on, this opinion may be influenced by others.

- An underlying opinion dynamics as the DeGroot's model [DeG74]

$$\boldsymbol{x}(t; k) = \overline{\boldsymbol{W}} \boldsymbol{x}(t-1; k), \quad t = 1, 2, ...\infty, \tag{1}$$

  where t – underlying time index, $\boldsymbol{x}(\infty; k)$ – opinion for the steady state.

- Our previous work [Wai15] tells that this model is not informative if $\overline{\boldsymbol{W}}$ is stochastic and irreducible – they will go consensus. With stubborn nodes, the steady state opinion will shaped by the stubborn nodes.

- Stubborn Senators: whose ideologies are far left or far right, do not change their own opinions and always try to influence other nodes' opinion. Ideology scores [Hare2014, Tau12].

DeGroot Model with Stubborn Nodes

- We partition the states vector and the transition matrix as

$$\boldsymbol{x}(t;k) = \begin{pmatrix} \boldsymbol{z}(t;k) \\ \boldsymbol{y}(t;k) \end{pmatrix}, \ \boldsymbol{z}(t;k) = \begin{pmatrix} \boldsymbol{s}_R(t;k) \\ \boldsymbol{s}_D(t;k) \end{pmatrix}, \ \boldsymbol{y}(t;k) = \begin{pmatrix} \boldsymbol{r}_R(t;k) \\ \boldsymbol{r}_D(t;k) \\ \boldsymbol{i}(t;k) \end{pmatrix},$$

$$\overline{\boldsymbol{W}} = \begin{array}{c} \\ R_s \\ D_s \\ R_n \\ D_n \\ F \end{array} \begin{array}{c} \begin{array}{ccccc} R_s & D_s & R_n & D_n & F \end{array} \\ \left[\begin{array}{c|c|c|c|c} \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{B}_1 & \mathbf{B}_4 & \mathbf{D}_1 & \mathbf{D}_4 & \mathbf{F}_1 \\ \hline \mathbf{B}_2 & \mathbf{B}_5 & \mathbf{D}_2 & \mathbf{D}_5 & \mathbf{F}_2 \\ \hline \mathbf{B}_3 & \mathbf{B}_6 & \mathbf{D}_3 & \mathbf{D}_6 & \mathbf{F}_3 \end{array}\right] \end{array} \qquad (2)$$

E.g., $\mathbf{B}_4$ is the normal Republicans' trust on stubborn Democrats.

Steady States for Stubborn Nodes and Normal Nodes

- The stubborn nodes can help facilitate us tracking the opinion dynamics:

$$\boldsymbol{y}(\infty;k) = \lim_{t\to\infty}\sum_{s=0}^{t}\boldsymbol{D}^{t-s}\boldsymbol{B}\boldsymbol{z}^{k}(0) + \boldsymbol{D}^{t}\boldsymbol{z}(0;k) + \boldsymbol{n}(k). \qquad (3)$$

$$= (\boldsymbol{I} - \boldsymbol{D})^{-1}\boldsymbol{B}\boldsymbol{z}(\infty;k) + \boldsymbol{n}(k).$$

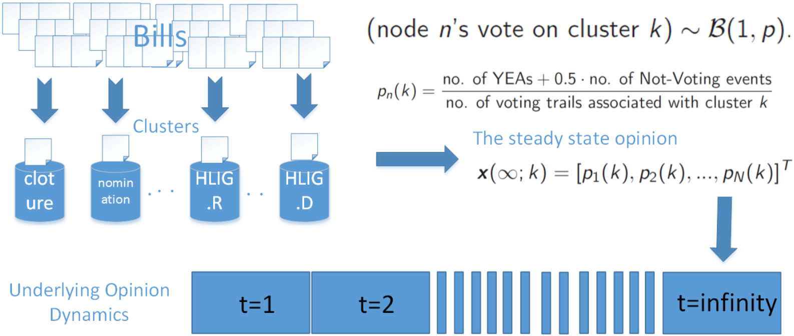
where $\boldsymbol{n}(k)$ is the Gaussian noise coming from the modeling noise and

$$\boldsymbol{B} = \left[\begin{array}{c|c} \mathbf{B}_1 & \mathbf{B}_4 \\ \hline \mathbf{B}_2 & \mathbf{B}_5 \\ \hline \mathbf{B}_3 & \mathbf{B}_6 \end{array}\right],\ \boldsymbol{D} = \left[\begin{array}{c|c|c} \mathbf{D}_1 & \mathbf{D}_4 & \mathbf{F}_1 \\ \hline \mathbf{D}_2 & \mathbf{D}_5 & \mathbf{F}_2 \\ \hline \mathbf{D}_3 & \mathbf{D}_6 & \mathbf{F}_3 \end{array}\right]. \qquad (4)$$

- The key observation underlying the opinion dynamics: If we know the steady states $\boldsymbol{y}(\infty;k)$ and $\boldsymbol{z}(\infty;k)$, we can estimate $\boldsymbol{B}$ and $\boldsymbol{D}$ from (3).

Labeling, Clustering and Bernoulli Sampling

How can we obtain $\boldsymbol{y}(\infty; k)$ and $\boldsymbol{z}(\infty; k)$?



$$(\text{node } n\text{'s vote on cluster } k) \sim \mathcal{B}(1, p).$$

$$p_n(k) = \frac{\text{no. of YEAs} + 0.5 \cdot \text{no. of Not-Voting events}}{\text{no. of voting trails associated with cluster } k}$$

The steady state opinion

$$\boldsymbol{x}(\infty; k) = [p_1(k), p_2(k), ..., p_N(k)]^T$$

We cluster bills by the committee and the ideology of the first sponsor of the bill, e.g., (Judiciary, Republican).

Step 3: The Linear Regression

How can we identify the system?

- Form an N by K opinion matrix by:

$$\begin{aligned} \boldsymbol{X}(\infty) &\triangleq [\boldsymbol{x}(\infty;1), \boldsymbol{x}(\infty;2), ..., \boldsymbol{x}(\infty;K)] \\ &= [\boldsymbol{Z}(\infty); \boldsymbol{Y}(\infty)], \end{aligned} \tag{5}$$

- We solve a *regularized* least square problem to obtain the trust matrix among all the Senators; i.e.,

$$\begin{aligned} \min_{\boldsymbol{B} \geq 0, \boldsymbol{D} \geq 0} \quad & \rho \|(\boldsymbol{I} - \boldsymbol{D})\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{Z}\|_F^2 + \|[\boldsymbol{B}_2, \boldsymbol{B}_4, \boldsymbol{D}_2, \boldsymbol{D}_4]\|_1 \\ \text{s.t.} \quad & (4) \text{ satisfied}, \ (\boldsymbol{B} + \boldsymbol{D})\mathbf{1} = \mathbf{1}, \ \text{diag}(\boldsymbol{D}) = \boldsymbol{\ell}, \end{aligned} \tag{6}$$

  where $\boldsymbol{\ell}$ is a self-trust *prior* of the normal nodes and ρ is a prescribed constant for tuning the regularization.

Data Analysis for the Congress Members

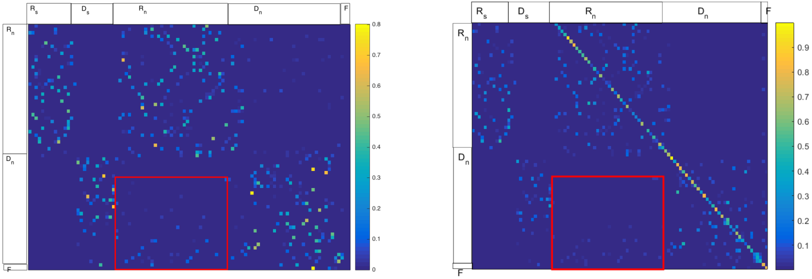


Figure 1: Trust matrix $[\boldsymbol{B}, \boldsymbol{D}]$ with $\mathrm{diag}(\boldsymbol{D}) = \mathbf{0}$ (left) and $\mathrm{diag}(\boldsymbol{D}) = \hat{\boldsymbol{\ell}}$ with $\hat{\boldsymbol{\ell}}$ being the leadership scores in [Tau12] (right).

- Different self-trusts give similar relative trust strengths.
- Bright pixels in the red box imply that Democrats agree more with Republicans (vs. Republicans to Democrats) for an approval of a bill.

Data Analysis for the Congress Member: Cont's

We can determine according to our model who trusts whom more, and help devise political strategies

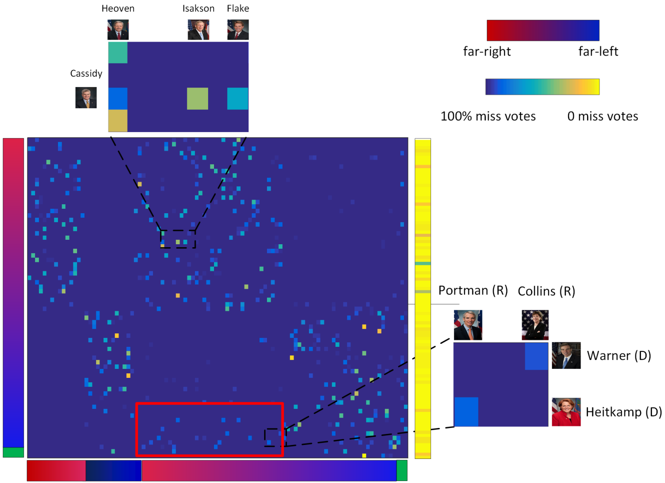


Figure 2: Trust matrix $[\boldsymbol{B}, \boldsymbol{D}]$ with $\mathrm{diag}(\boldsymbol{D}) = \mathbf{0}$.

Impact Factor and Influence Factor

After thresholding, we can define impact factor and influence factor

- Impact Factor: the number of non-zeros entries in each column of the trust matrix $\tilde{\boldsymbol{W}} = [\boldsymbol{B}, \boldsymbol{D}]$:

  $$\text{Impact Factor}_i = \|\tilde{\boldsymbol{w}}_i\|_0, \ \text{ where } \ \tilde{\boldsymbol{w}}_i \ \text{ is the } i\text{th column of } \tilde{\boldsymbol{W}}.$$

  It represents the number of Senators one Senator has an influence on.

- Influence Factor: we sum up all the non-zeros entries in each column in the trust matrix $\tilde{\boldsymbol{W}} = [\boldsymbol{B}, \boldsymbol{D}]$:

  $$\text{Influence Factor}_i = \|\tilde{\boldsymbol{w}}_i\|_1, \ \text{ where } \ \tilde{\boldsymbol{w}}_i \ \text{ is the } i\text{th column of } \tilde{\boldsymbol{W}}.$$

  The total amount of trust one Senator has influence on other Senators.

Top Impact Factor and Influence Factor

Table 1: Influence Factor Top 7

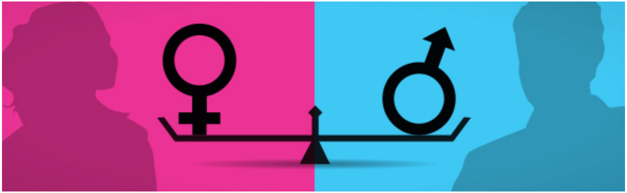
Senator	Total Trust
Debbie Stabenow (MI/D)	2.448
Michael Bennet (CO/D)	1.828
Richard Blumenthal (CT/D)	1.754
Kelly Ayotte (NH/R)	1.653
Thomas Carper (DE/D)	1.603
Patty Murray (WA/D)	1.482
Timothy Kaine (VA/D)	1.397

Table 2: Selected High Impact Factor Senators

Senator	Total Number
Susan Collins (ME/R)	12 (Top 1)
Kelly Ayotte (NH/R)	11
Jefferson Sessions (AL/R)	10
Timothy Kaine (VA/D)	9

Women versus Men

Question: who plays a more important role in the US Senate, Men or Women?



- Our numerical results tell the average (impact Factor/influence Factor) for different genders are

  Women: 6.733/0.960; Men: 6.141/0.678.

- Women are more influential in the US Senate.

Predictions

Table 3: Prediction Accuracy for Different Machine Learning Algorithms [FD2001]. Train data:test data = 75% : 25%. We train (estimate) the model then use stubborn nodes' votes in the test data as the input for the test.

Machine Learning Algorithm	Prediction Accuracy
Logistic	74.14%
MultiScheme	83.62%
Random Tree	87.07%
Simple Logistic	88.79%
J-48	88.79%
KStar	89.66%
Decision Table	89.66%
Random Forest	93.96%
Our Model	92.24%

Only our model can provide the trust information.

Conclusions

- We learned Senators' relationship by analyzing their voting record based on a DeGroot Model.
- The role of stubborn Senators is important. They control the opinion of the system.
- The trust information we mined from the dataset is useful for devising political strategy.
- Our main contribution: the methodology for data mining the Senators' trust.

Supplementary Numerical Results

The Relative Trust Matrices

Definition

Given a tuple of trust matrices $(\boldsymbol{B}, \boldsymbol{D})$ *and an* N_n*-dimensional vector* $\mathbf{0} \leq \mathbf{c} < \mathbf{1}$*, define the map* $\mathrm{rel}_{\mathbf{c}} : \mathbb{R}^{N_n \times N} \to \mathbb{R}^{N_n \times N}$ *such that* $\mathrm{rel}_{\mathbf{c}}(\boldsymbol{B}, \boldsymbol{D}) = (\boldsymbol{B}', \boldsymbol{D}')$ *with*

$$\boldsymbol{B}' = (\boldsymbol{I} - \mathrm{Diag}(\boldsymbol{c}))\boldsymbol{\Lambda}_s^{-1}\boldsymbol{B},\ \mathrm{diag}(\boldsymbol{D}') = \boldsymbol{c},$$
$$\mathrm{off}(\boldsymbol{D}') = (\boldsymbol{I} - \mathrm{Diag}(\boldsymbol{c}))\boldsymbol{\Lambda}_s^{-1}\mathrm{off}(\boldsymbol{D}),\ \boldsymbol{\Lambda}_s = \boldsymbol{I} - \mathrm{Diag}(\boldsymbol{D})$$

We call $(\boldsymbol{B}', \boldsymbol{D}')$ *the relative trust matrices resulting from* $(\boldsymbol{B}, \boldsymbol{D})$.

Lemma

Consider two networks with transition matrices parameterized by the pair of tuples $(\boldsymbol{B}, \boldsymbol{D})$ *and* $(\tilde{\boldsymbol{B}}, \tilde{\boldsymbol{D}})$*. Suppose that* $\mathrm{rel}_{\mathbf{c}}(\boldsymbol{B}, \boldsymbol{D}) = \mathrm{rel}_{\mathbf{c}}(\tilde{\boldsymbol{B}}, \tilde{\boldsymbol{D}})$*, i.e., the two tuples give the same relative trust matrices, then the following holds:*

$$(\boldsymbol{I} - \boldsymbol{D})^{-1}\boldsymbol{B} = (\boldsymbol{I} - \tilde{\boldsymbol{D}})^{-1}\tilde{\boldsymbol{B}}. \tag{7}$$

How the Self-trust Influence Relative Trust Matrices

- The assumption implies that the system $(\boldsymbol{I} - \boldsymbol{D})\boldsymbol{Y} = \boldsymbol{BZ}$ admits a unique solution as long as ℓ is fixed.

- Let $(\boldsymbol{B}, \boldsymbol{D})$ be an optimal solution to (6) solved with $\rho > 0$ and self-trust prior ℓ, and $(\tilde{\boldsymbol{B}}, \tilde{\boldsymbol{D}})$ be an optimal solution to (6) solved with self-trust $\tilde{\ell}$ and $\tilde{\rho}$ (with $\tilde{\ell} \neq \ell$). Then, we have $(\boldsymbol{B}, \boldsymbol{D}) = \mathrm{rel}_{\boldsymbol{\ell}}(\tilde{\boldsymbol{B}}, \tilde{\boldsymbol{D}})$.

- Only relative trust matrices are concerned: $\frac{B_{ij}}{B_{ij'}} = \frac{\tilde{B}_{ij}}{\tilde{B}_{ij'}}$, $\frac{\mathrm{off}(D)_{ij}}{\mathrm{off}(D)_{ij'}} = \frac{\mathrm{off}(\tilde{D})_{ij}}{\mathrm{off}(\tilde{D})_{ij'}}$.
  1) the connectivity of the network
  2) the relative trust strengths

Verification of Proposition 1

We generate the graph that corresponds to $\boldsymbol{D}$ as an Erdos-Renyi graph with connectivity $p = 0.1$, and the bipartite graph corresponding to $\boldsymbol{B}$ is a random left-regular graph with left-degree $d = 6$ [Wai15]. $[\boldsymbol{B}, \boldsymbol{D}]\mathbf{1} = \mathbf{1}$. We also assume that the support of $\boldsymbol{B}$ is given, to avoid further tuning on ρ.

Table 4: Mean Square Error under Different Self-trust

	noiseless	$\sigma^2 = -3\text{dB}$	$\sigma^2 = 0\text{dB}$
MSE under $\boldsymbol{\ell} = \text{rand}(N, 1)$	0	1.29	1.62
MSE under $\boldsymbol{\ell} = \tilde{\boldsymbol{\ell}}$	0	1.29	1.62

- When the observation is noiseless, $\boldsymbol{B}, \boldsymbol{D}$ can be correctly identified and no matter how we choose $\boldsymbol{\ell}$, the relative trust matrices are almost the same;
- If the the observation is noisy, the estimated matrix would be subject to error, whose MSE is almost the same for different self-trust priors.

Data Analysis for the Congress Member: Cont's

- Analysis on Senator Marco Rubio:

Table 5: Senator Marco Rubio's Trust on Others

M. Enzi (WY/R)	T. Cruz (TX/R)	D. Perdue (GA/R)	B. Sanders (VT/I)
0.185	0.597	0.083	0.136

Table 6: Senator Marco Rubio's Influence on Others

D. Fischer (NE/R)	T. Cruz (TX/R)	D. Perdue (GA/R)	D. Sullivan (AK/R)
0.032	0.405	0.019	0.042
B. Casey (PA/D)	J. Manchin (WV/D)	B. Sanders (VT/I)	
0.017	0.083	0.152	

Data Analysis for the Congress Member

- Analysis on Senator Ted Cruz:

Table 7: Senator Ted Cruz's Trust on Others

D. Vitter (LA/R)	T. Scott (SC/R)	R. Paul (KY/R)
0.128	0.011	0.04
P. Toomey (PA/R)	M. Rubio (FL/R)	R. Shelby (AL/R)
0.120	0.405	0.296

Table 8: Senator Ted Cruz's Influence on Others

T. Scott (SC/R)	R. Paul (KY/R)	P. Toomey (PA/R)
0.065	0.113	0.07
M. Rubio (FL/R)	R. Shelby (AL/R)	D. Heller (NV/R)
0.60	0.150	0.0318

Senator Marco Rubio and Senator Ted Cruz trust each other a lot.

Data Analysis for the Congress Member: Cont's

Table 9: Senator Bernie Sanders's Trust on Others

R. Durbin (IL/D)	M. Rubio (FL/R)	R. Wyden (OR/D)	C. Booker (NJ/D)
0.411	0.152	0.148	0.289

Table 10: Senator Bernie Sanders's Influence on Others

R. Paul (KY/R)	M. Rubio (FL/R)	L. Graham (SC/R)	D. Sullivan (AL/R)
0.078	0.136	0.183	0.048
T. Baldwin (WI/D)	R. Wyden (OR/D)	C. Booker (NJ/D)	T. Udall (NM/D)
0.0393	0.0872	0.153	0.0319
E. Markey (MA/D)	C. McCaskill (MO/D)		
0.107	0.121		

It seems that Senator Bernie Sanders has a big impact in the Senate, since he has in total 0.995 influence on ten senators. But we could know that he trusts Senator Richard Durbin, Senator Marco Rubio, Senator Ron Wyden and Senator Cory Booker more than others in the Senate.

Data Analysis for the States

We can also analyze the trust between any two States by

- Assigning the stubborn States by the mean of the ideology scores of the Senators in the State.
- Obtaining the each State's belief on the theme of the cluster based on the voting results of all the Senators in the State.
- Estimating the trust matrix by solving the following linear regression

$$\begin{array}{ll} \min_{\boldsymbol{B} \geq 0, \boldsymbol{D} \geq 0} & \rho \|(\boldsymbol{I} - \boldsymbol{D})\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{Z}\|_F^2 + \|[\boldsymbol{B}, \boldsymbol{D}]\|_1 \\ \text{s.t.} & (\boldsymbol{B} + \boldsymbol{D})\mathbf{1} = \mathbf{1}, \ \mathrm{diag}(\boldsymbol{D}) = \mathbf{0}. \end{array} \tag{8}$$

Data Analysis for the States

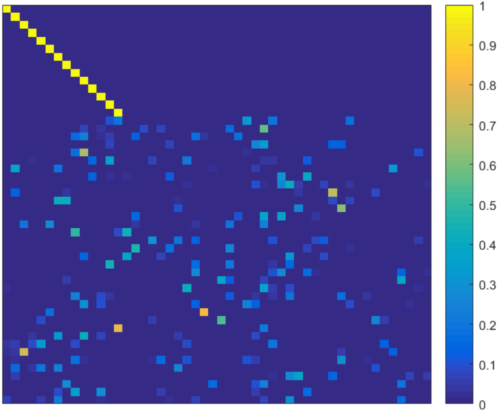


Figure 3: Trust matrix $\boldsymbol{W}$ when the normal States' self-trust is set to be zero.

State Impact Factor and State Influence Factor

- State Impact Factor: the number of States one State has an influence on.
- State Influence Factor: the amount of trust one has an influence on others.

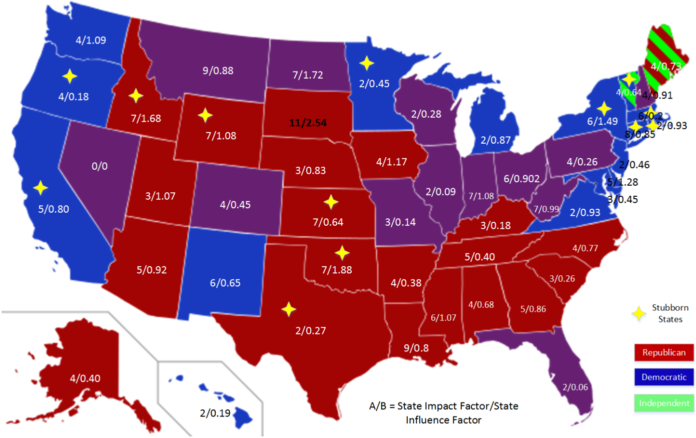


Figure 4: State Impact Factor and State Influence Factor

References

[Poole2000] K. T. Poole and H. Rosenthal, *Congress: A political-economic history of roll call voting*. Oxford University Press on Demand, 2000.

[Fenno1982] R. F. Fenno, *The United States senate: A bicameral perspective*. Aei Press, 1982, vol. 362.

[Hare2014] C. Hare and K. T. Poole, "The polarization of contemporary American politics," *Polity*, vol. 46, no. 3, pp. 411–429, 2014.

[Clinton2004] J. Clinton, S. Jackman, and D. Rivers, "The statistical analysis of roll call data," *American Political Science Review*, vol. 98, no. 02, pp. 355–370, 2004.

[Clinton2012] J. D. Clinton, "Using roll call estimates to test models of politics," *Annual Review of Political Science*, vol. 15, pp. 79–99, 2012.

[DeG74] M.H. DeGroot.
Reaching a consensus.
In *Journal of American Statistcal Association*, volume 69, pages 118–121, 1974.

[HP14] C. Hare and K. T. Poole.
The polarization of contemporary American politics.
Polity, 46(3):411–429, 2014.

[ideology] GovTrack.us.,
Ideology analysis of members of congress,
`https://www.govtrack.us/about/analysis`, 2013.

[leadership] ——,
Leadership analysis of members of congress,
`https://www.govtrack.us/about/analysis`, 2013.

[Tau12] J. Tauberer,
Observing the unobservable in the US congress,
Law Via the Internet, 2012.

[Wai15] Hoi-To Wai, Anna Scaglione, and Amir Leshem.
Active Sensing of Social Networks
IEEE Transactions on Signal and Information Processing over Networks (2015).

[FD2001] J. Friedman, T. Hastie, and R. Tibshirani,
The elements of statistical learning.
Springer series in statistics Springer, Berlin, 2001, vol. 1.