Supplemental Appendix for: "Scaling Data from Multiple Sources"

Ted Enamorado* Gabriel López-Moctezuma† Marc Ratkovic‡

July 9, 2019

This supplemental appendix contains two sections. In the first, we describe the estimation procedure of our algorithm. The second presents additional results from our simulations and the main empirical application.

A MD2S Estimation

Estimation proceeds in two steps. First, we recover estimates of the shared and idiosyncratic subspaces. Second, all the subspaces are partialed out of the data matrix. The aforementioned steps are fit for each dimension q.

Initialize $\widehat{Z}^0_{(1)} = lsv\left(Y_{(1)}\right)$; $\widehat{Z}^0_{(2)} = lsv\left(Y_{(2)}\right)$; $\widehat{Z}^0_S = lsv\left(\widetilde{Y} + \widetilde{Y}^\top\right)$ where $\widetilde{Y} = Y_{(1)}Y_{(1)}^\top Y_{(2)}Y_{(2)}^\top$. Initialize $Y^0_{(m)} = Y_{(m)}$; where we denote as $lsv(A)$ a function that extracts the first left singular value of the matrix A.

1. Convergence in each subspace,

*Assistant Professor, Department of Political Science, University of North Carolina at Chapel Hill, Chapel Hill NC 27514. Email: ted.enamorado@gmail.com, URL: http://www.tedenamorado.com

†Assistant Professor, Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena CA 91125. Email: glmoctezuma@caltech.edu, URL: http://glmoctezuma.com

‡Assistant Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 608-658-9665, Email: ratkovic@princeton.edu, URL: http://www.princeton.edu/∼ratkovic

(a) Update $\widehat{Z}_{(m)j}$ for $m = 1$ then $m = 2$.

i. $\widehat{Z}^{(t)}_{m|S;q} = lsv\left(M(\widehat{Z}^{(t-1)}_{S})Y^{(t-1)}_{(m)}\right)$

ii. $\widehat{Z}^{(t)}_{m;q} = \frac{\left(H(X_{(m)})\widehat{Z}^{(t)}_{m|S;q}(1-\gamma_m)+M(X_{(m)})\widehat{Z}^{(t)}_{m|S;q}\gamma_m\right)}{||H(X_{(m)})\widehat{Z}^{(t)}_{m|S;q}(1-\gamma_m)+M(X_{(m)})\widehat{Z}^{(t)}_{m|S;q}\gamma_m||}$

where $X_{(m)}$ represents a the matrix of covariates informing the mth subspace and γ_m is estimated as the argmax of:

$$\frac{\left(H(X_{(m)})\widehat{Z}^{(t)}_{m|S;q}(1-\gamma)+M(X_{(m)})\widehat{Z}^{(t)}_{m|S;q}\gamma\right)^{\top}Y^{(t-1)}_{m}Y^{(t-1)\top}_{m}\left(H(X_{(m)})\widehat{Z}^{(t)}_{m|S;q}(1-\gamma)+M(X_{(m)})\widehat{Z}^{(t)}_{m|S;q}\gamma\right)}{||H(X_{(m)})\widehat{Z}^{(t)}_{m|S;q}(1-\gamma)+M(X_{(m)})\widehat{Z}^{(t)}_{m|S;q}\gamma||^2}$$

The parameter γ_m seeks to balance how much the covariates explain the data either by explaining $Z^{(t)}_{m|S;q}$ through the linear projection $H(X_{(m)})\widehat{Z}^{(t)}_{m|S;q}$ or via the residual $M(X_{(m)})\widehat{Z}^{(t)}_{m|S;q}$

(b) Update $\widehat{Z}^{(t)}_{S;q}$.

i. Update $\widehat{Z}^{(t)}_{S|1;q} = lsv\left(M(\widehat{Z}^{(t)}_{(1)})Y_{(1)}\right)$; $\widehat{Z}^{(t)}_{S|2;q} = lsv\left(M(\widehat{Z}^{(t)}_{(2)})Y_{(2)}\right)$

ii. Update $\widehat{Z}^{(t)}_{S|12;q} = (\widehat{Z}^{(t)}_{S|1;q}(1-\alpha_S) + \widehat{Z}^{(t)}_{S|2;q}\alpha_S)/||\widehat{Z}^{(t)}_{S|1;q}(1-\alpha_S) + \widehat{Z}^{(t)}_{S|2;q}\alpha_S||$

where X_S represents the matrix of covariates informing the shared subspace and α_S represents the weight w_1 in the proposition, which is estimated as the argmax of:

$$\frac{(\widehat{Z}^{(t)}_{S|1;q}(1-\alpha) + \widehat{Z}^{(t)}_{S|2;q}\alpha)^{\top}\widetilde{Y}^{(t-1)}(\widehat{Z}^{(t)}_{S|1;q}(1-\alpha) + \widehat{Z}^{(t)}_{S|2;q}\alpha)}{||\widehat{Z}^{(t)}_{S|1;q}(1-\alpha) + \widehat{Z}^{(t)}_{S|2;q}\alpha||^2}$$

iii. Update $\widehat{Z}^{(t)}_{S;q} = \frac{\left(H(X_S)\widehat{Z}^{(t)}_{S|12;q}(1-\gamma_S)+M(X_S)\widehat{Z}^{(t)}_{S|12;q}\gamma_S\right)}{||H(X_S)\widehat{Z}^{(t)}_{S|12;q}(1-\gamma_S)+M(X_S)\widehat{Z}^{(t)}_{S|12;q}\gamma_S||}$

where γ_S is estimated as the argmax of:

$$\frac{\left(H(X_S)\widehat{Z}^{(t)}_{S|12;q}(1-\gamma) + M(X_S)\widehat{Z}^{(t)}_{S|12;q}\gamma\right)^{\top}\widetilde{Y}^{(t-1)}\left(H(X_S)\widehat{Z}^{(t)}_{S|12;q}(1-\gamma) + M(X_S)\widehat{Z}^{(t)}_{S|12;q}\gamma\right)}{||H(X_S)\widehat{Z}^{(t)}_{S|12;q}(1-\gamma) + M(X_S)\widehat{Z}^{(t)}_{S|12;q}\gamma||^2}$$

We can give γ_S a mirror interpretation to the one given to γ_m, but now we respect to $Z^{(t)}_{S|12;q}$ and $\widetilde{Y}^{(t-1)}$.

2. After convergence in the previous step, update $Y^{(t)}_{(m)} = M([\widehat{Z}_{S;q}, \widehat{Z}_{(m);q}])Y^{(t-1)}_{(m)}M([\widehat{W}_{(m);q}, \widehat{B}_{(m);q}])$ where $\widehat{W}_{(m);q} = Y^{\top}_{(m)}\widehat{Z}_{S;q}$ and $\widehat{B}_{(m);q} = Y^{\top}_{(m)}\widehat{Z}_{(m);q}$, both normed to have length one.

B Supplemental Results

B.1 Simulations: Bridging

As noted in the main text, our method can be used to combine information coming from different datasets. One such instance of combining data is bridging across different actors. To assess the performance of the proposed method along that direction, we conduct an additional simulation study. The main difference with the other simulations is that the actors across datasets are allowed to differ and only common items between datasets are used to jointly scale the actors (bridging).[1]

The simulation setup is going to be quite similar. Again, the observed data consist of matrices $Y_{(1)}$ and $Y_{(2)}$ with N rows (common items) and K_1 (number of actors in dataset 1) and K_2 (number of actors in dataset 2) columns respectively. N is varied along $\{50, 500, 1000\}$. For K_1 we have two scenarios. The first one we call it "balanced" as K_1 is set in $\{20, 40, 200\}$ and K_2 is chosen such that the ratio K_1:K_2 is equal to 2:3. In the second case, we varied K_1 along $\{10, 20, 100\}$ and K_2 is chosen such that the ratio K_1:K_2 is equal to 1:4. The latter is to mimic a situation where the number of actors in one dataset is significantly smaller if compared to the number of actors in the other dataset. As before, the data are generated according to equations (18) and (19) in the main text. All simulations were run 1,000 times.

As shown in Figure 1, MD2S does a remarkable job in terms of recovering the true ordering for each actor (W_m). Similarly, the correlation between the true scaling and the estimates obtained from MD2S is almost perfect as both the number of actors and the number of items used to bridge increases. The latter is true even in the lopsided case. In table 1 we compare MD2S to multidimensional scaling (MDS) as implemented in the R-package `smacof`. For MDS we bridge estimates by pooling datasets.[2] As table 1

[1] We thank an anonymous reviewer for suggesting the inclusion of this simulation exercise.

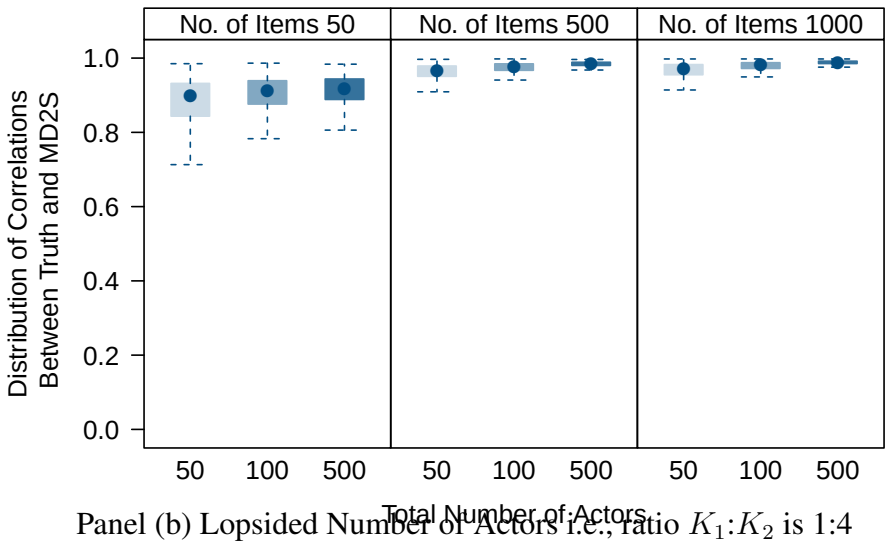
[2] MDS is a scaling method designed to take either continuous or discrete valued values as inputs. Due to the continuous nature of values each observation in our simulated datasets take, we use MDS as a reference point.

shows, MD2S recovers the estimates that correlate highly with true actors' subspace and its precision grows with the number of common items to bridge. In contrast, MDS does not recover a the true actors' subspace and its performance decreases as the number of actors increases.

We believe these are promising results to motivate additional tests of the bridging capabilities of MD2S – which offers a built-in approach when there are common items across datasets. A main advantage of MD2S over previous alternatives is that pooling the datasets is not needed. For MD2S bridging is a consequence of assuming that common items between datasets exist and that these items connect the scaled positions of the actors. Furthermore, through the dimension weights and the idiosyncratic subspaces, MD2S provides a more flexible approach to address the "constant behavior" assumption which requires that actors across datasets face the same concerns when referring to a particular issue.[3] However, having common items to bridge might not be an possibility in some situations, making bridging an impossible task in our framework. We leave for future research a thorough evaluation of the bridging properties of MD2S.

[3]See the work of Lewis and Tausanovitch (2015) for a literature review of the bridging literature and a formal set of tests for the "constant behavior" assumption across other scaling models.

Panel (a) Balanced Number of Actors i.e., ratio K_1:K_2 is 2:3



Panel (b) Lopsided Number of Actors i.e., ratio K_1:K_2 is 1:4

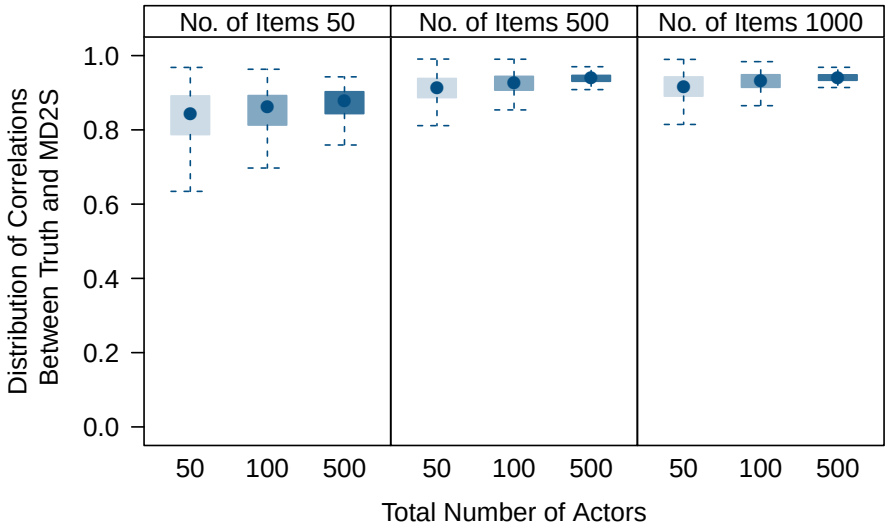


Figure 1: **Correlation between the True Actors' Subspaces (W_m) and their Corresponding MD2S estimates.** Number of common items used to bridge datasets is ($N \in \{50, 500, 1000\}$) and the number of distinct actors in each dataset is set to $(K_1, K_2) \in \{(20, 30), (40, 60), (200, 300)\}$ in panel (a) and to $(K_1, K_2) \in \{(10, 40), (20, 80), (100, 400)\}$. The total number of actors represents the sum of K_1 and K_2. The y-axis ranges from 0 to 1 and measures the correlation between the true and estimated values across 1000 simulations per combination of N and K_1. As in the simulations presented in the main text, MD2S does a remarkable job in terms of recovering the true ordering for each actor. In addition, MD2S improves its performance in the number of actors and the number of items used to bridge.

	MD2S			MDS		
	Mean	2.5%	97.5%	Mean	2.5%	97.5%
50 Items						
50 Actors	0.88	0.68	0.97	0.27	0.02	0.69
100 Actors	0.90	0.74	0.97	0.21	0.01	0.56
500 Actors	0.91	0.79	0.97	0.16	0.01	0.43
500 Items						
50 Actors	0.96	0.90	0.99	0.26	0.01	0.65
100 Actors	0.97	0.94	0.99	0.19	0.00	0.51
500 Actors	0.98	0.97	0.99	0.11	0.00	0.33
1000 Items						
50 Actors	0.97	0.91	0.99	0.28	0.01	0.63
100 Actors	0.98	0.95	0.99	0.22	0.01	0.52
500 Actors	0.99	0.98	1.00	0.11	0.00	0.34

Table 1: **Correlation between the True Actors' Subspaces (W_m) and their Corresponding MD2S and MDS estimates.** Number of common items used to bridge datasets is ($N \in \{50, 500, 1000\}$) and the number of distinct actors in each dataset is set to $(K_1, K_2) \in \{(20, 30), (40, 60), (200, 300)\}$. The total number of actors represents the sum of K_1 and K_2. As in the simulations presented in the main text, MD2S performs well in terms of recovering the true ordering for each actor. In the case of MDS, it does not perform as well and as the number of actors increases its performance decreases.

B.2 Combining Senate Rollcall and Text Data.

B.2.1 Estimates from MD2S without Senator-level Covariates

Figure (2) through (4) reproduce the results presented in the main text, with the sole difference that covariates are not used to inform each of the subspaces obtained from MD2S. The shared subspace and the idiosyncratic subspace for speech are substantively the same regardless whether we include covariates or not. The ranking obtained from the idiosyncratic subspace informed by the roll call data is slightly different when we omit covariates from the estimation stage. Again, we have conservative senators like DeMint, Lee, Toomey, Paul, and Risch, on one extreme, while the remaining Senators from the Republican party

are located on the other. Thus, the idiosyncratic subspace obtained from vote data, again, reveals a party divide among Republicans.

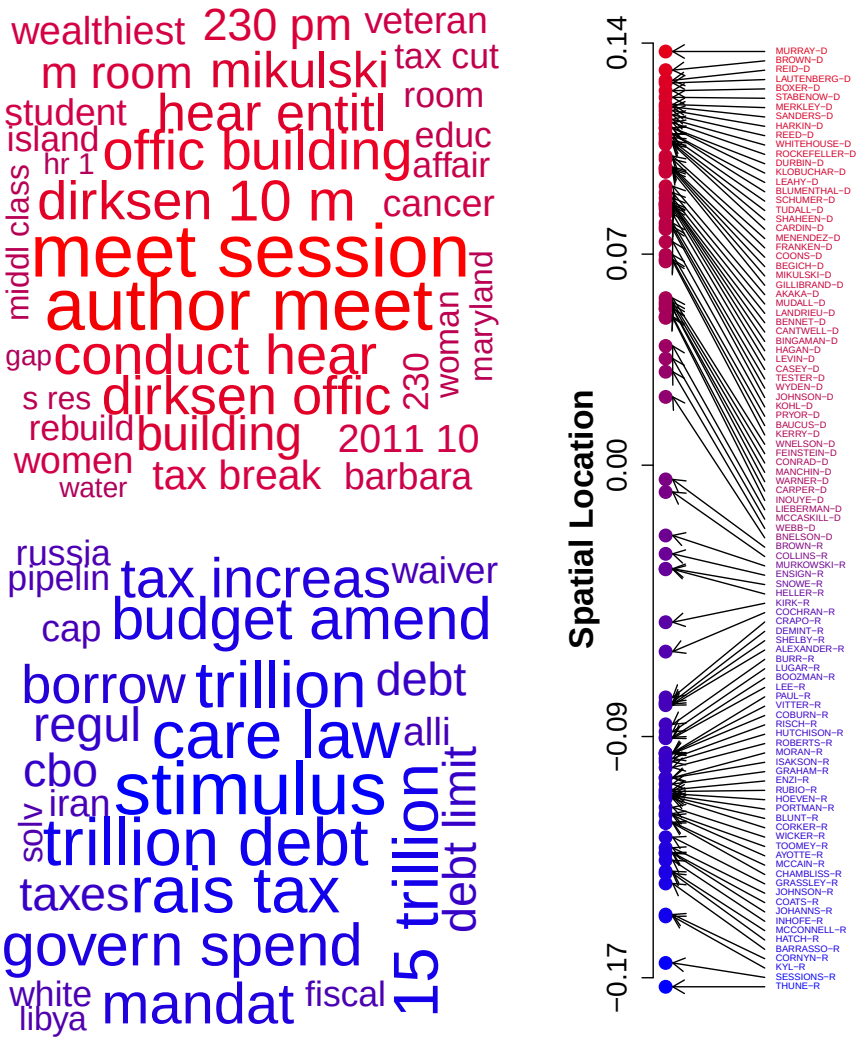


Figure 2: **Shared Subspace Locations Estimated via MD2S for the Members of the 112th U.S. Senate.**

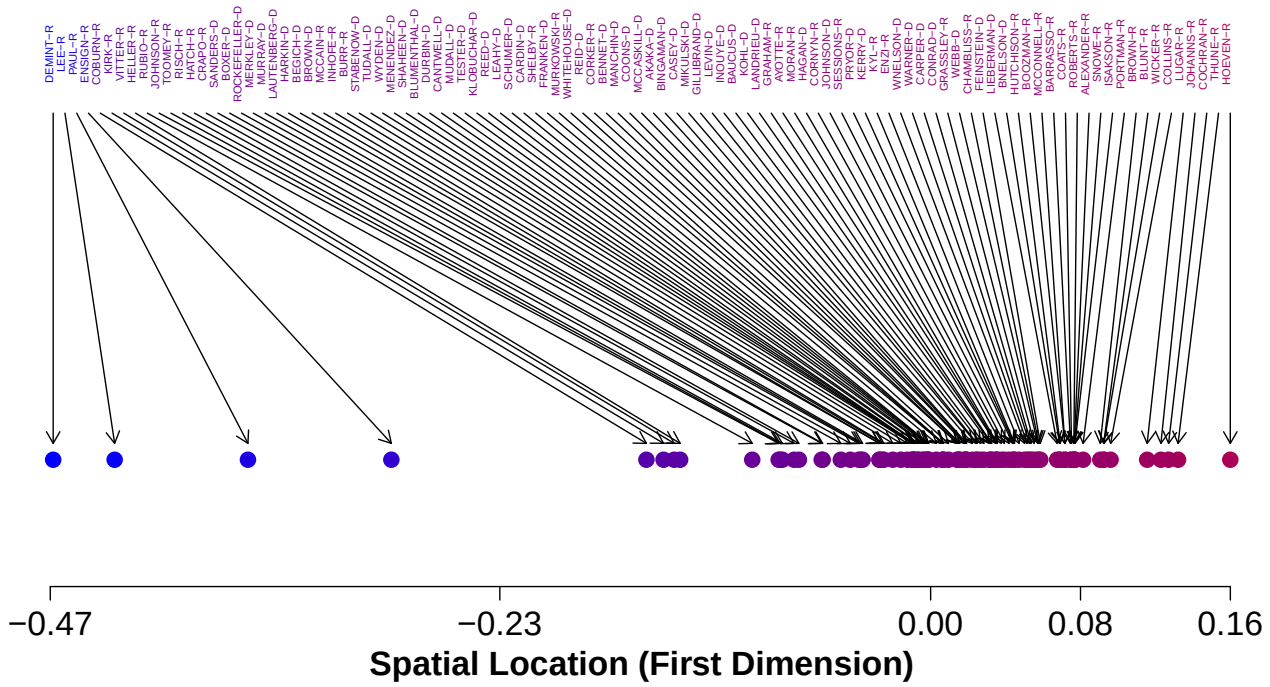


Figure 3: **Idiosyncratic Vote Subspace Locations Estimated via MD2S for the Members of the 112th U.S. Senate.**

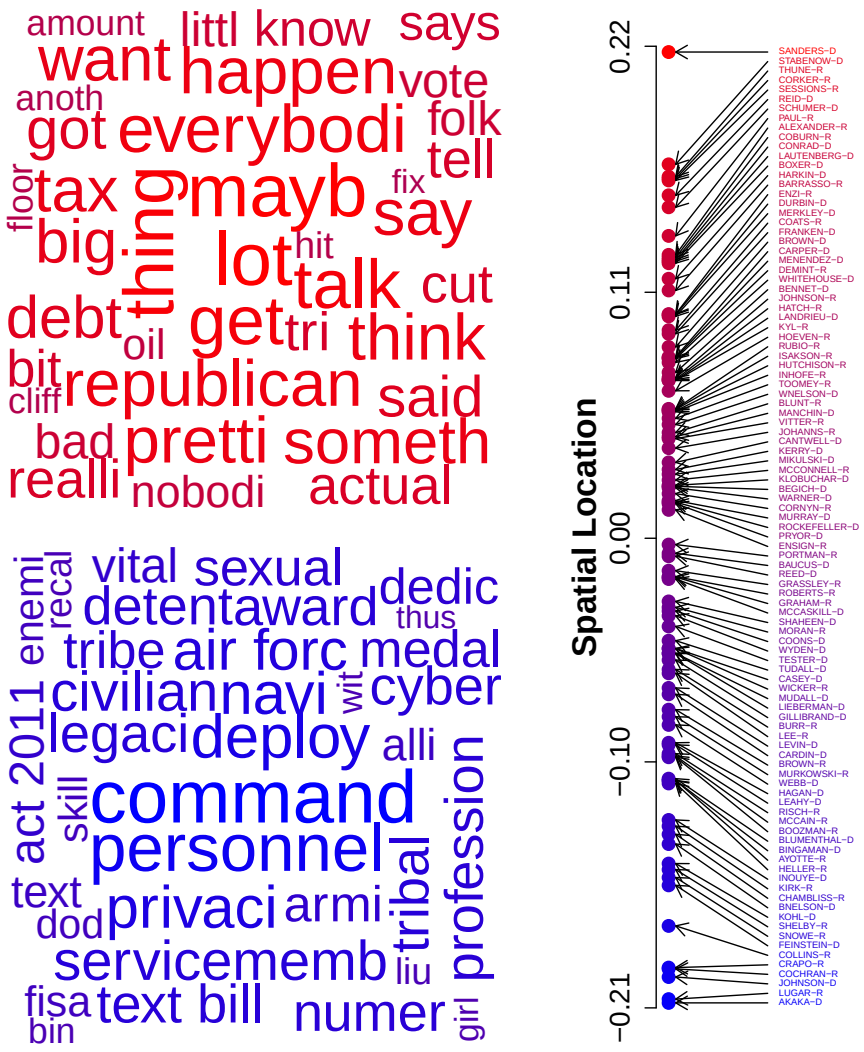


Figure 4: **Idiosyncratic Word Subspace Locations Estimated via MD2S for the Members of the 112th U.S. Senate. First Dimension.**

B.2.2 Different Levels of Sparsity

In this subsection we present evidence that our method is robust to several pre-processing steps when working with text data, namely whether we use unigrams and bigrams, and how we trim sparse terms from the term-document matrix (with low, medium, and high levels of sparsity in the term document matrix). As shown in table **??** across the six settings ($\{unigram, bigram\} \times \{low, med, hi\}$), the correlations across the first-dimension estimates are no lower than 0.9 for the shared subspace, 0.97 for the vote idiosyncratic subspace, and 0.86 in the word idiosyncratic subspace.

		Unigrams and Bigrams		
		Levels of Sparsity		
		(High, Medium)	(High, Low)	(Medium, Low)
Shared subspace	correlation	0.98	0.91	0.97
Word subspace	correlation	0.95	0.86	0.97
Vote subspace	correlation	0.99	0.97	0.98
		Unigrams only		
		Levels of Sparsity		
		(High, Medium)	(High, Low)	(Medium, Low)
Shared subspace	correlation	0.99	0.95	0.98
Word subspace	correlation	0.96	0.87	0.97
Vote subspace	correlation	0.99	0.96	0.98

Table 2: Correlations Between Subspaces (1st Dimension) for Different Levels of Sparsity across the Document-Term Matrix. *Low sparsity* is equal to removing terms that are not used by at least 20 senators for a total of 1852 terms. *Medium sparsity* is equal to removing terms that are not used by at least 30 senators for a total of 2616 and *High sparsity* is equal to removing terms that are not used by at least 40 senators for a total of 3622 terms.

B.2.3 Correlations Across Different Scaling Methods

Table (3) and figure (5) present the overall and within-party correlation across different scaling methods, respectively. First, all scaling methods separate Senators in two clusters (by party). Second, both both MD2S and SFA recover the same set of moderate and extreme senators within each party, consistent with the results of IDEAL and DW-Nominate. Third, jointly incorporating votes and floor speech introduces some variation in senators' rakings within party. For instance, republican Senators such as Sessions and

Kyl are found to be more extreme when combining vote and text data, than their scaled locations using only votes.

	Votes Only:			Votes and Words:		
	DW-Nominate	IDEAL	MDS	SFA	MD2S	INDSCAL
Votes Only:						
DW-Nominate	1.00					
IDEAL	0.99	1.00				
MDS	0.98	0.98	1.00			
Votes and Words:						
SFA	0.93	0.93	0.95	1.00		
MD2S	0.94	0.95	0.96	0.97	1.00	
INDSCAL	0.48	0.47	0.43	0.30	0.38	1.00

Table 3: Correlations Across Different Scaling Methods for the 112th U.S. Senate. **Votes Only** refers to scaling methods that use roll call data, while **Votes and Words** refers to those scaling methods that use both roll call and speech data. The correlations presented here are calculated using the first dimension recovered by each method. In the case of MDS and INDSCAL we have reduced the roll call data to a dissimilarity matrix using a renormalized version of the Manhattan distance (L1 norm) between each pair of rows. The normalization factor is the number of columns in the roll call data, thus our our dissimilarity measure tells us the share of times two legislators vote in the same way. Similarly, for INDSCAL (Votes and Words) we have reduced the vote and speech data to two dissimilarity matrices (one per data set) using the same renormalized distance between rows. In the case of MD2S, we use the shared subspace. The table shows that traditional methods to scale the U.S. Congress, like IDEAL, DW-Nominate and MDS, correlate almost perfectly with SFA, and MD2S. In contrast, if we use votes and speech data, INDSCAL recovers a subspace does not correlate as well with the other alternatives.

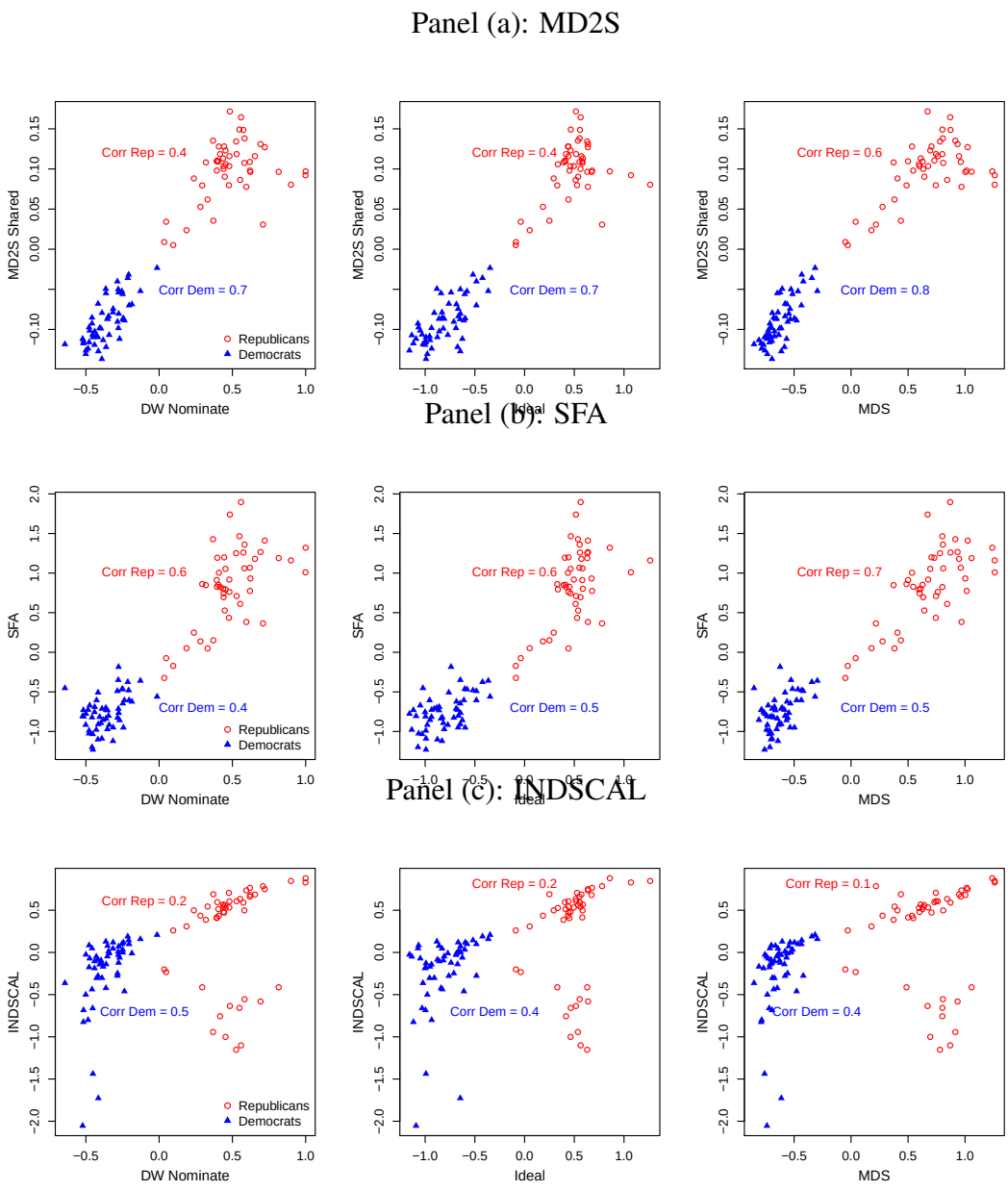


Figure 5: **Scaled Locations and Partisanship.** The three panels above present the results comparing MD2S, SFA, and INDSCAL to traditional scaling methods such as DW Nominate, Ideal, and MDS. As the three panels show, MD2S recovers a shared subspace that overall correlate well with scaling methods that use only vote data. Not only that, MD2S separates Senators' in two clusters (partisanship). SFA also performs well when compared to traditional methods and INDSCAL produces estimates that are not as highly correlated with the aforementioned traditional approaches – producing at least three clusters of Senators.

B.2.4 Bootstrap Estimates

Figure (6) through (8) present the 95% percentile confidence intervals obtained via the non-parametric bootstrap (5000 replications) described in Section 3.5.

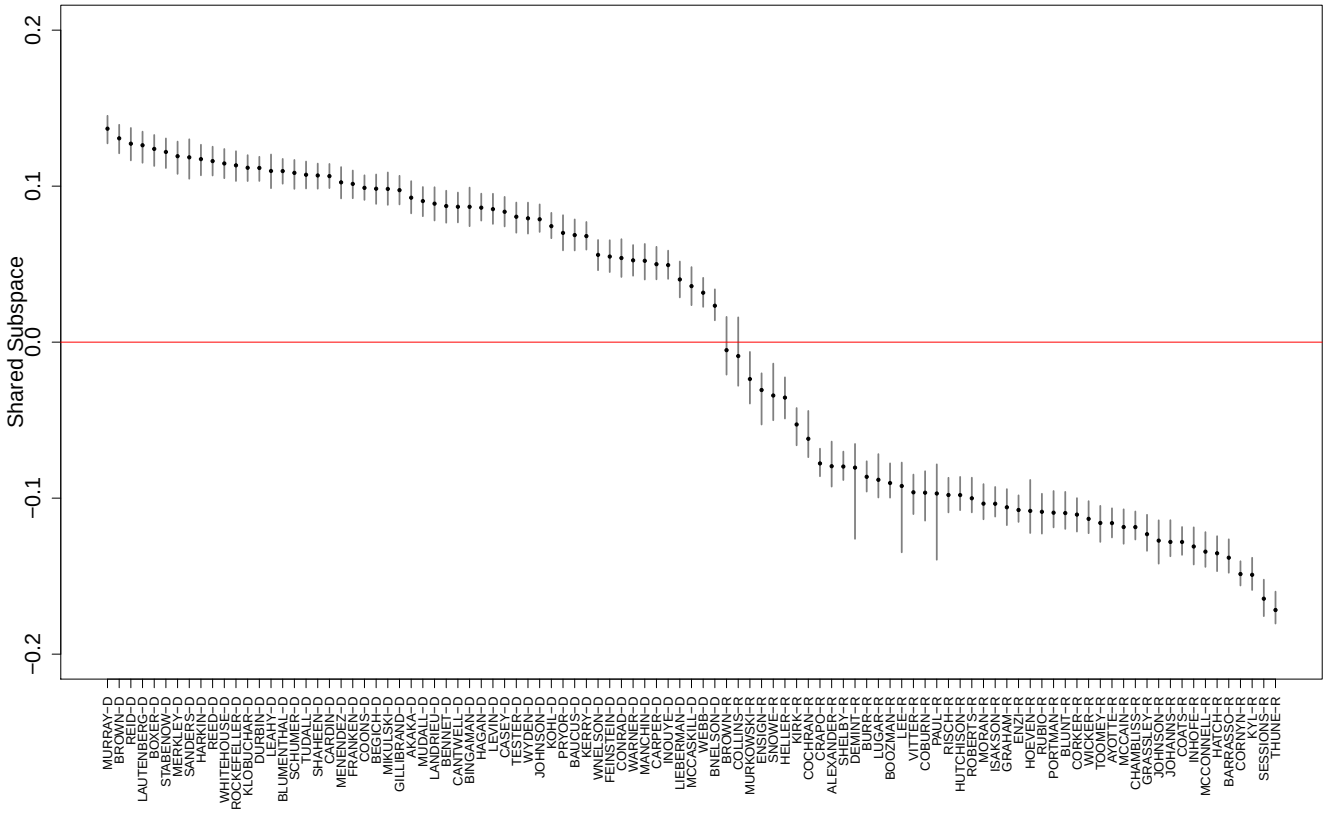


Figure 6: **Bootstrap Confidence Intervals for the Shared Subspace Locations Estimated via MD2S for the Members of the 112th U.S. Senate.**

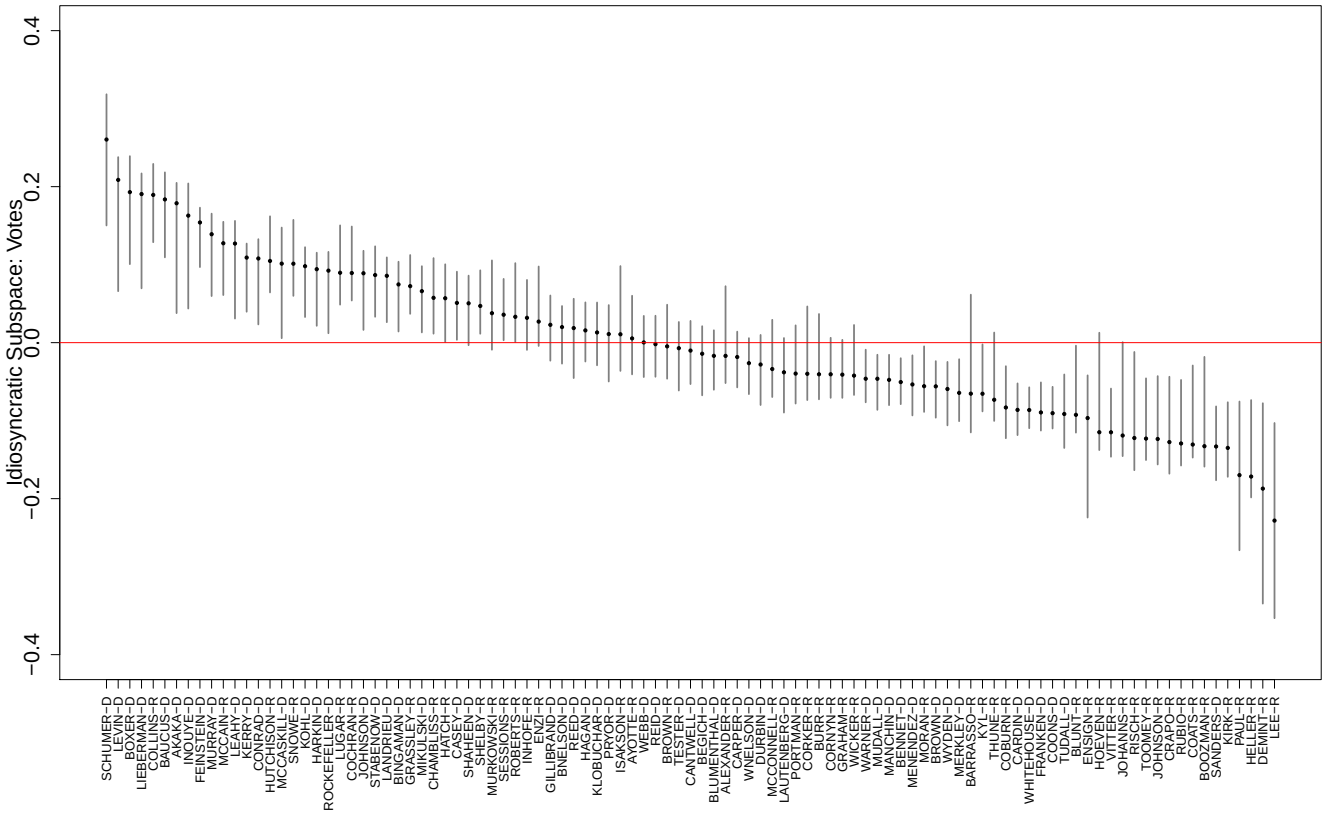


Figure 7: **Bootstrap Confidence Intervals for Idiosyncratic Vote Subspace Locations Estimated via MD2S for the Members of the 112th U.S. Senate.**

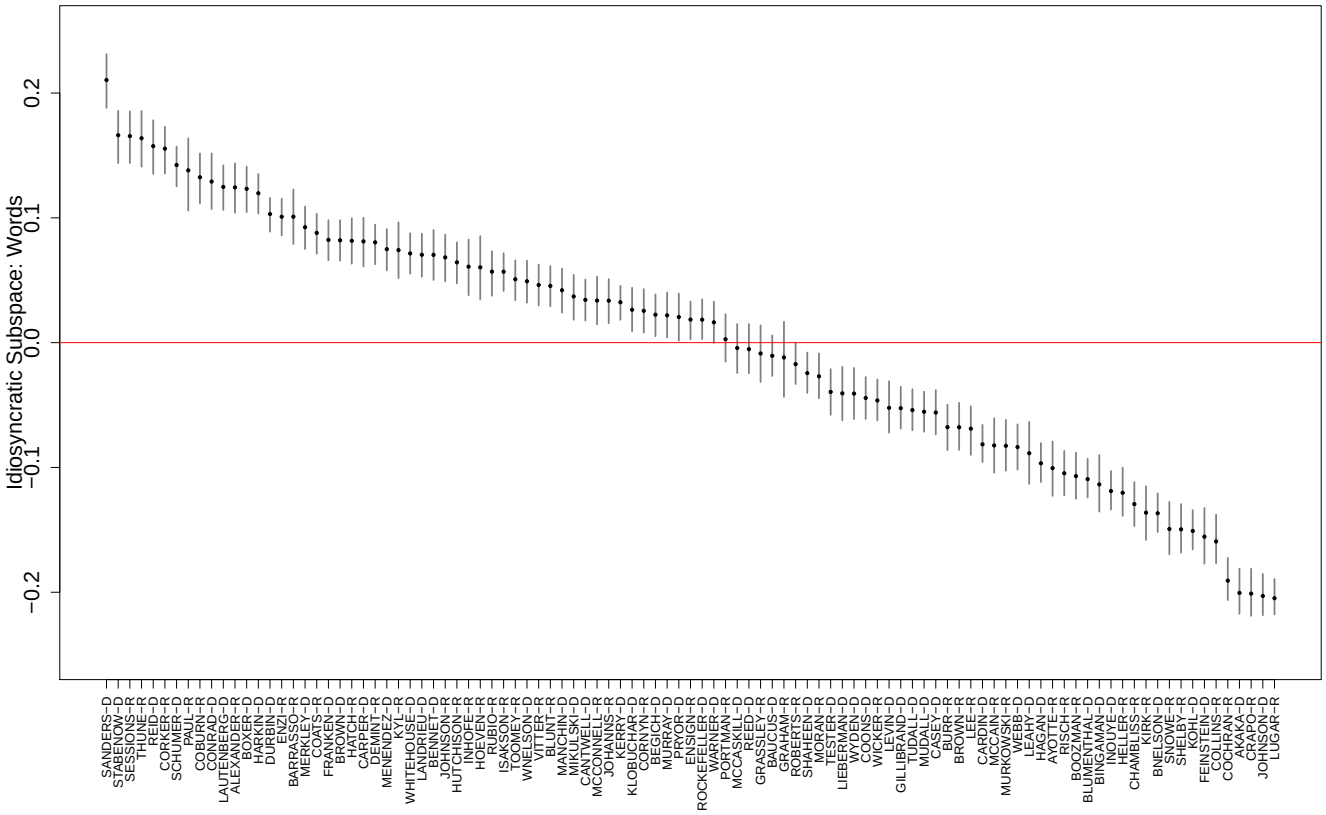


Figure 8: **Bootstrap Confidence Intervals for Idiosyncratic Word Subspace Locations Estimated via MD2S for the Members of the 112th U.S. Senate. First Dimension.**

B.2.5 Correlation of Subspaces with Covariates

Finally, we regress each subspace on a set of covariates for each senator. Table (4), presents the results discussed in Section 5 of the main text.

Table 4: Correlation with Covariates

	Dependent variable:		
	Shared	Votes	Words
	(1)	(2)	(3)
party	−0.187	−0.054	−0.018
	(0.007)	(0.004)	(0.020)
gender	0.022	0.066	−0.031
	(0.009)	(0.005)	(0.028)
seniority	−0.0003	0.003	−0.002
	(0.0004)	(0.0002)	(0.001)
membership	0.00001	0.0004	−0.006
	(0.003)	(0.002)	(0.009)
agricultural	0.006	−0.110	0.068
	(0.020)	(0.012)	(0.060)
economics	0.011	0.013	0.034
	(0.035)	(0.020)	(0.104)
security	−0.009	0.094	−0.067
	(0.011)	(0.006)	(0.032)
leadership	0.003	0.101	0.014
	(0.007)	(0.004)	(0.022)
Constant	0.086	−0.067	0.054
	(0.013)	(0.008)	(0.040)
Observations	101	101	101
R^2	0.901	0.967	0.119
Adjusted R^2	0.892	0.964	0.043
F Statistic (df = 8; 92)	104.322	336.981	1.559

Note: Standard errors in parentheses

References

Lewis, Jeffrey B. and Chris Tausanovitch. 2015. When Does Joint Scaling Allow For Direct Comparisons of Preferences? Technical report University of California, Los Angeles.