Title: Extracting insights from the shape of complex data using topology

Authors: P.Y. Lum[1,*], G. Singh[1], A. Lehman[1], T. Ishkanov[1], M. Vejdemo-Johansson[2], M. Alagappan[1], J. Carlsson[3], G. Carlsson[1,4,*]

Affiliations:

[1]Ayasdi Inc., Palo Alto, CA.

[2]School of Computer Science, Jack Cole Building, North Haugh, St. Andrews KY16 9SX, Scotland, United Kingdom

[3]Industrial and Systems Engineering, University of Minnesota, 111 Church St. SE, Minneapolis, MN 55455, USA

[4]Department of Mathematics, Stanford University, Stanford, CA, 94305, USA

*Correspondence to: Gunnar Carlsson, Department of Mathematics, Stanford University, Stanford, CA, 94305, USA. gunnar@math.stanford.edu (1-650-723-2224). Pek Yee Lum, Ayasdi Inc., Palo Alto, CA94303, USA. pek@ayasdi.com (1-206-794-4097)

Supplementary Materials:

Figures S1-S4

Table S1

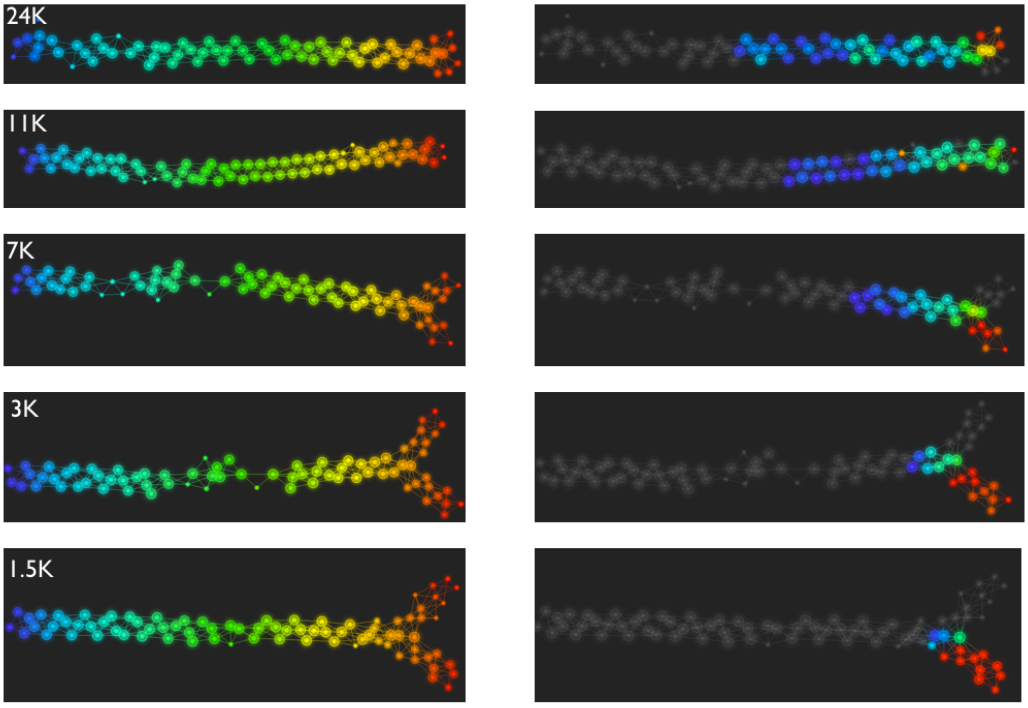

Fig. S1. Shape of the data becomes more distinct as the analysis columns are restricted to the top varying genes. Left panels: 24K: all the genes on the microarray were used in the analysis; 11K: 10,731 top most varying genes were used in the analysis; 7K: 6.688 top most varying genes were used in the analysis; 3K: 3212 top most varying genes were used in the analysis; 1.5K: 1553 top most varying genes were used in the analysis. Graphs are colored by the L-infinity centrality values. Red: high; Blue: low. Right panels: The same corresponding graphs to the left panels are now colored by the data points from the bottom flare of the 1.5K panel. In every panel on the

right, the color depicts the percentage of the data points coming from the left panel bottom flare in each node, red being 100% and grey being 0%. The blue gradient indicates a mixture of data points that came from the bottom flare from the 1.5K graph. This shows that as the number of genes increases, the data points from the bottom flare of the 1.5K graph becomes less distinct and more mixed with the other non-flare data points.

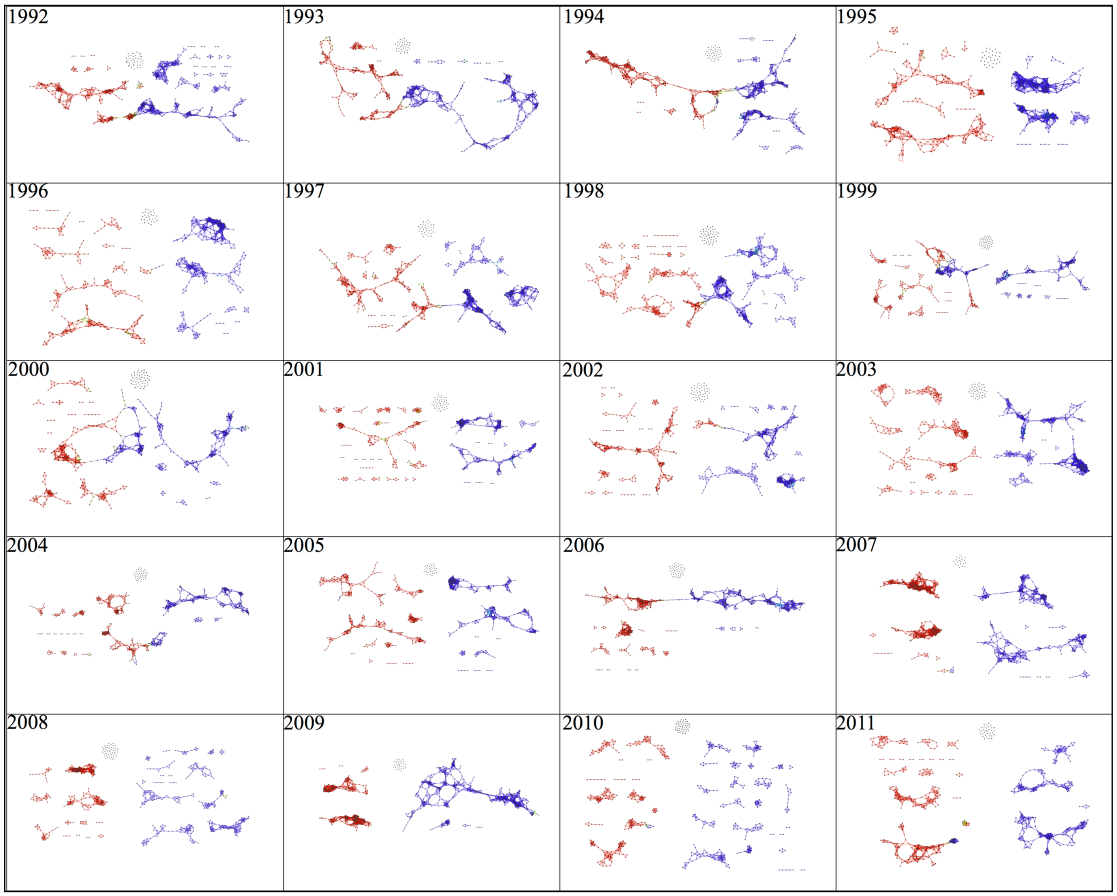

Fig. S2. Voting behavior networks from 1992 to 2011. Color: Red nodes: Republican; Blue nodes: Democrats

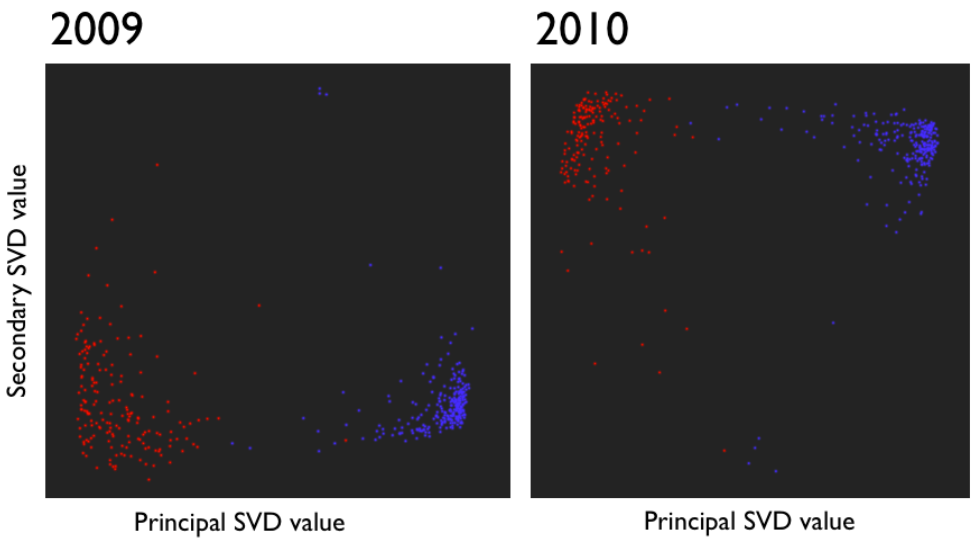

Fig. S3. PCA analysis of year 2009 (left panel) and 2010 (right panel).

Color: Red: Republican; Blue: Democrats

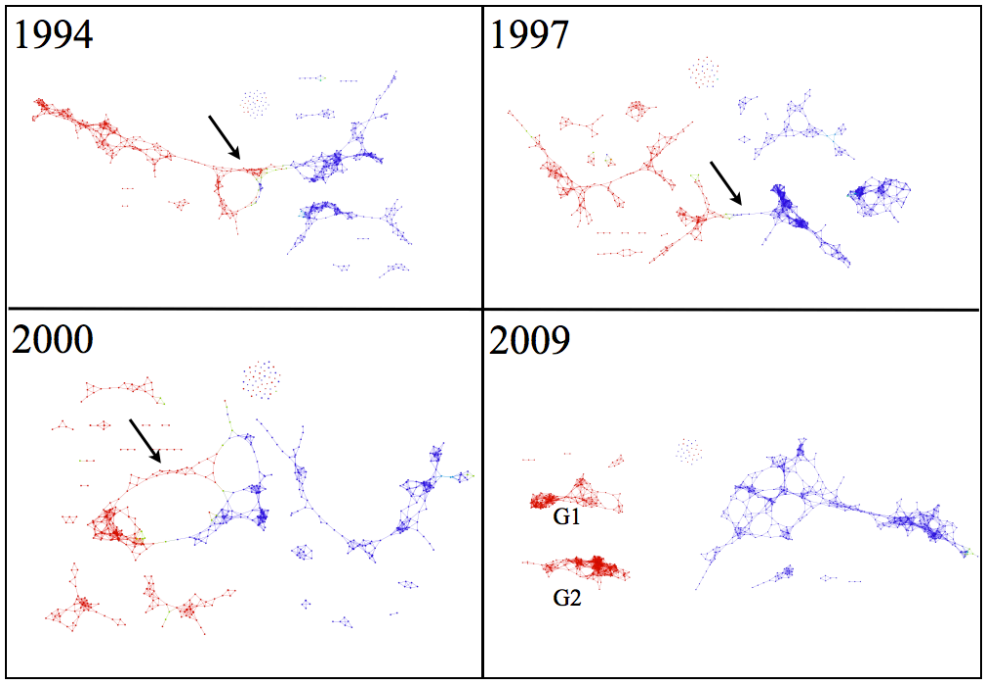

Fig. S4. Voting behavior networks for 1994, 1997, 2000 and 2009 highlighting the "Central Group". Some members of G2 in 2009 are the same members for the other years (arrow). Arrow points to the nodes with some of the same members of the "Central Group".

Table S1. Quantitative differences in expression levels of the genes in the KEGG chemokine pathway.

NKI All	NKI ESR-Non-survivors (lowERNS)	NKI ESR-survivors (lowERHS)	GSE 2304 All	GSE 2304 ESR-relapse	GSE 2304 ESR-Non-relapse
-.21	-.05	0.186	7.12	7.36	7.86
54%	77%	92%	56%	68%	84%

The numbers in the first row are the average levels of expressions of genes belonging to the chemokine pathway within the relevant groups. The second row gives these same levels reported as percentiles belonging to the collection of the chemokine expression levels across *all* the patients in the study to which the group belongs. Note that the lowERHS groups display very high chemokine expression levels, 92nd percentile in the case of NKI and 84th percentile in the case of GSE2304. The lowERNS groups are also consistently higher than the general population in the study, although to different degrees in the studies. The difference can perhaps be accounted for by the difference between the survival vs. non-relapse criteria in the two data sets.