Spring 5-2010

Regionalization via Network-Constrained Clustering

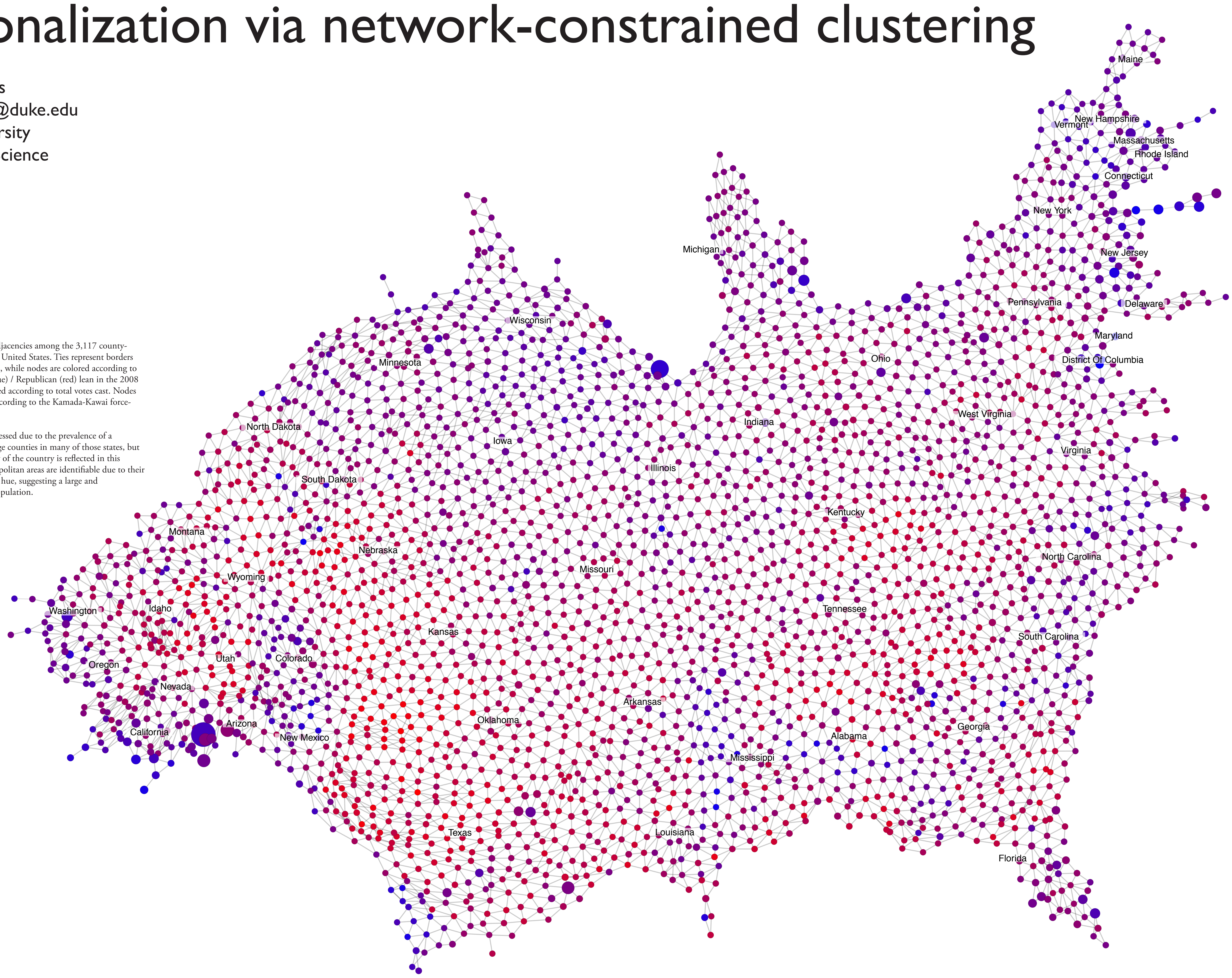
David Sparks
Duke University, d.sparks@duke.edu

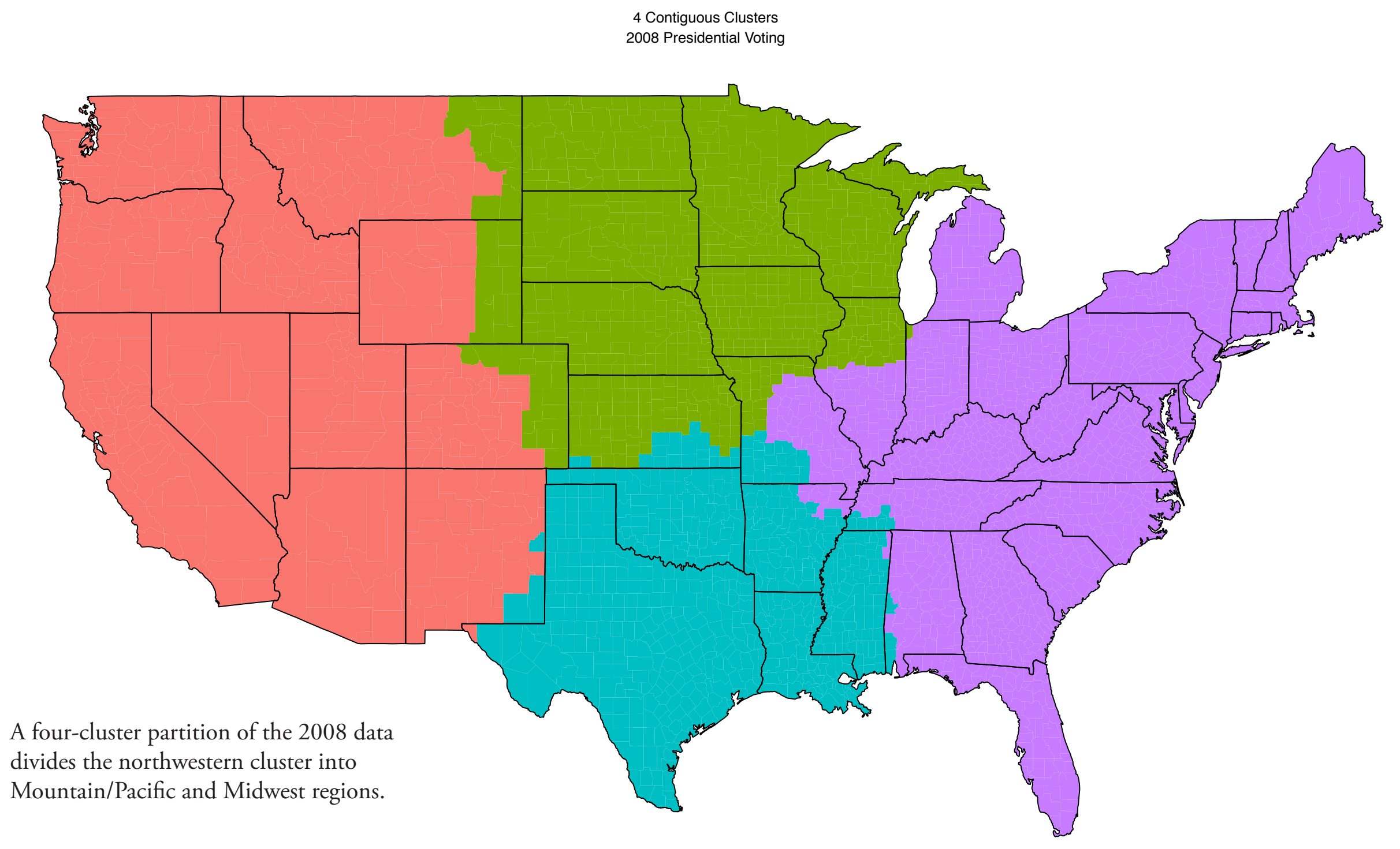
Regionalization via network-constrained clustering

David Sparks
d.sparks@duke.edu
Duke University
Political Science

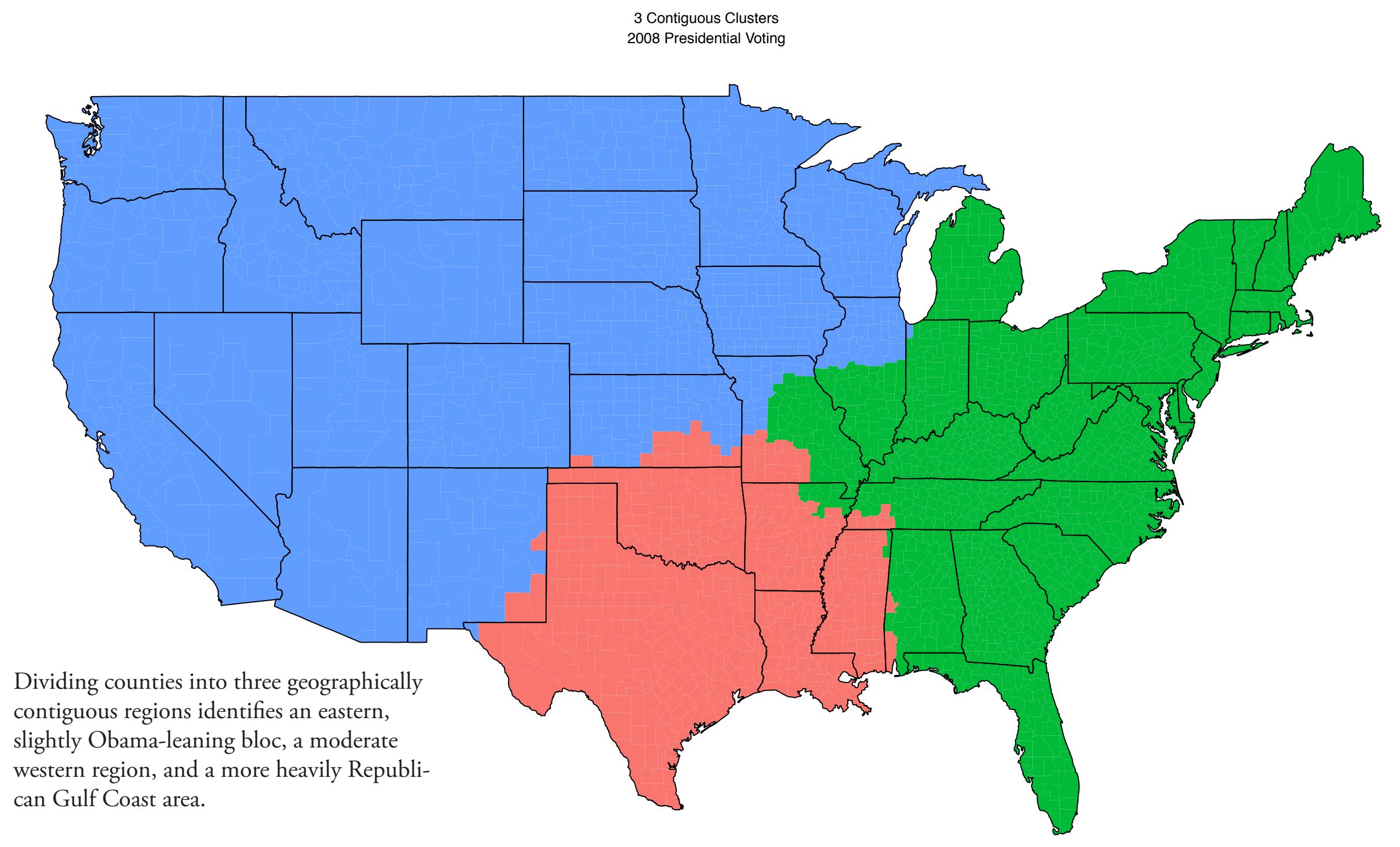
This network graph depicts adjacencies among the 3,117 county-equivalents in the continental United States. Ties represent borders between neighboring counties, while nodes are colored according to each county's Democratic (blue) / Republican (red) lean in the 2008 presidential election, and scaled according to total votes cast. Nodes are positioned in the graph according to the Kamada-Kawai force-based algorithm.

The southwest appears compressed due to the prevalence of a relatively small number of large counties in many of those states, but the overall political geography of the country is reflected in this county network. Many metropolitan areas are identifiable due to their relatively large size and bluish hue, suggesting a large and Democratic-leaning voting population.

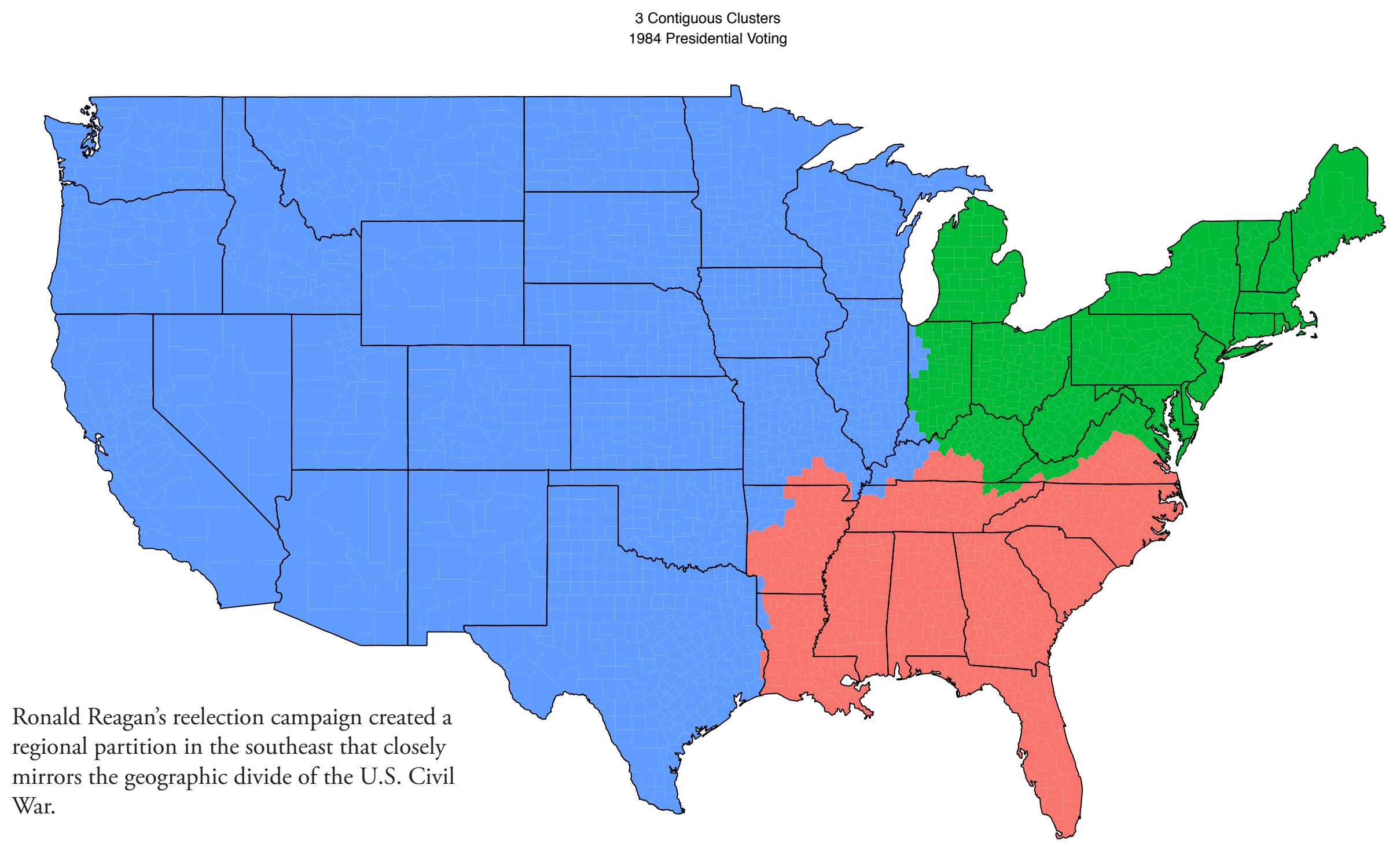



4 Contiguous Clusters
2008 Presidential Voting

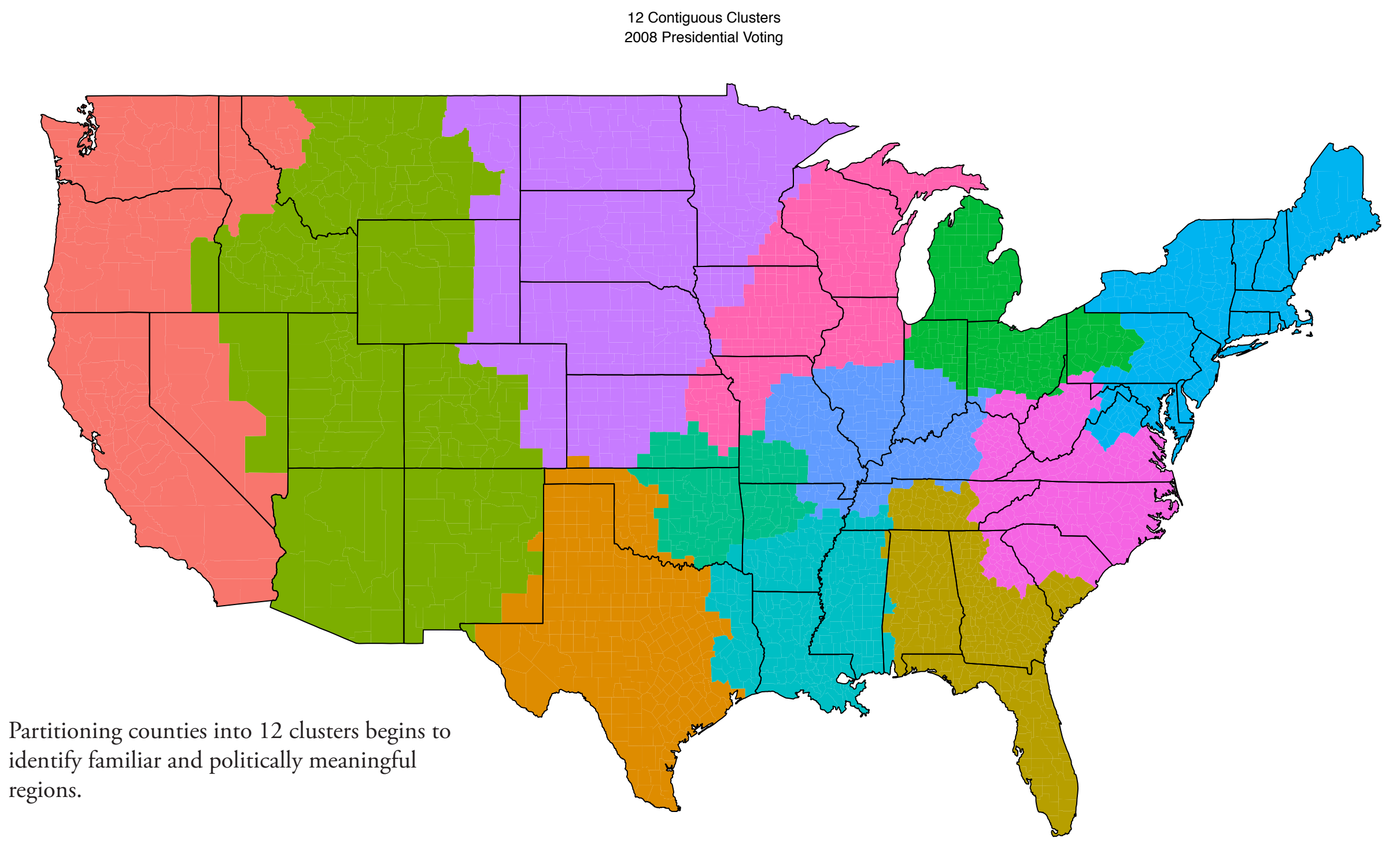
A four-cluster partition of the 2008 data divides the northwestern cluster into Mountain/Pacific and Midwest regions.


3 Contiguous Clusters
2008 Presidential Voting

Dividing counties into three geographically contiguous regions identifies an eastern, slightly Obama-leaning bloc, a moderate western region, and a more heavily Republican Gulf Coast area.


3 Contiguous Clusters
1984 Presidential Voting

Ronald Reagan's reelection campaign created a regional partition in the southeast that closely mirrors the geographic divide of the U.S. Civil War.


12 Contiguous Clusters
2008 Presidential Voting

Partitioning counties into 12 clusters begins to identify familiar and politically meaningful regions.

Constrained clustering is a family of classification techniques that generalize familiar clustering algorithms to allow the imposition of structural constraints over the partitioning of observations into clusters. Here, I apply network-constrained clustering to historical county electoral data to identify regions of political preference within the continental United States.

Network-constrained clustering operates on a dissimilarity matrix computed on any observations of interest, multiplied element-wise by an adjacency matrix representing connections between those observations. The clustering algorithm then interprets any off-diagonal zero elements as though that pair of observations is, essentially, infinitely dissimilar. Under this constraint, hierarchical clustering methods generate distinct communities / regions / eras / contiguous clusters within the set of observations.

Using the percentage of votes cast in each county for Democratic, Republican, and other candidates in the 2008 presidential election, I performed hierarchical agglomerative cluster analysis, generating three unconstrained clusters as depicted in the map below center.

These clusters give a sense of candidate preference by county, but do a poor job of conveying a sense of geographic bases of partisanship.

Constrained clustering, as seen in the three-cluster graph above center, gives a much clearer picture of partisan geographic tendencies, and localized bases of candidate support.

Each of the other maps visualizes constrained clusters of electoral data, illustrating differences derived from changing the parameter governing the number of clusters, and across time.

The key advantage of network constraint is that it allows consideration of both measured variables and network position in identifying interesting clusters or communities within the network.

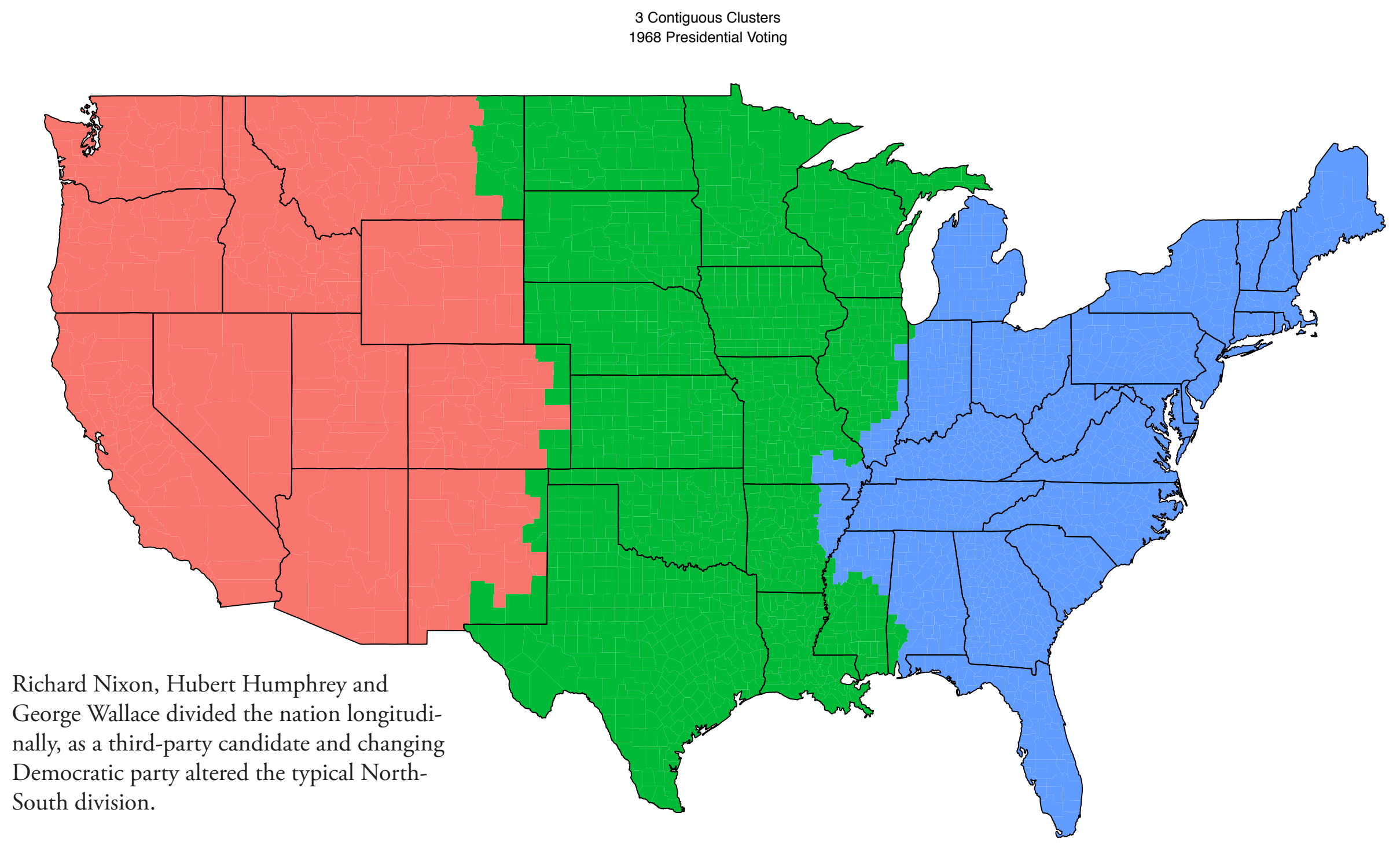
References

Ferligoj, Anuška and Vladimir Batagelj. 1982. "Clustering with Relational Constraint." *Psychometrika* 47: 413-426.

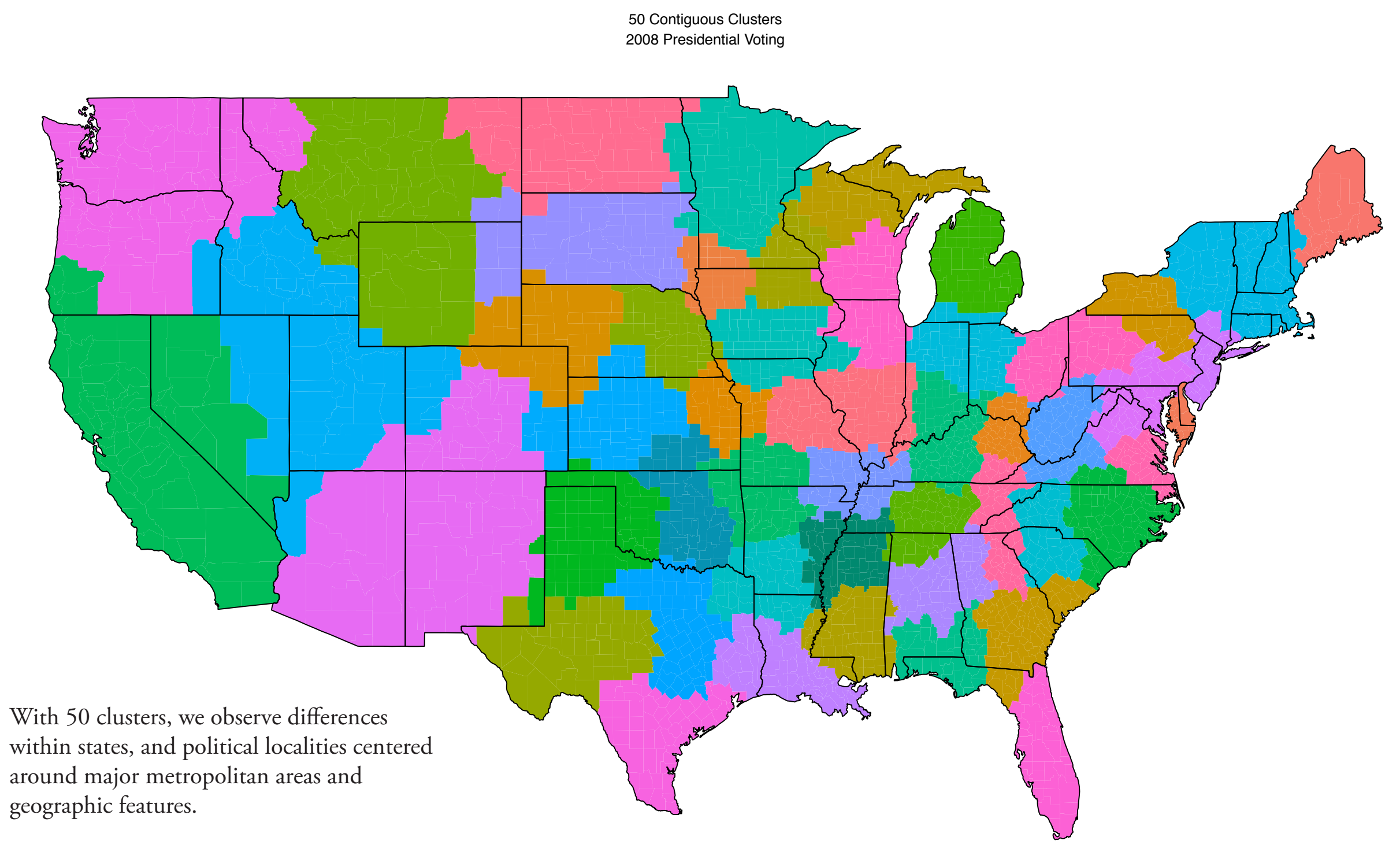
Murtagh, Fionn D. 1985. "A Survey of Algorithms for Contiguity-Constrained Clustering and Related Problems." *The Computer Journal* 28: 82-88.

Presidential general election. 2003. In *CQ voting and elections collection (Web site).* Washington: CQ Press.

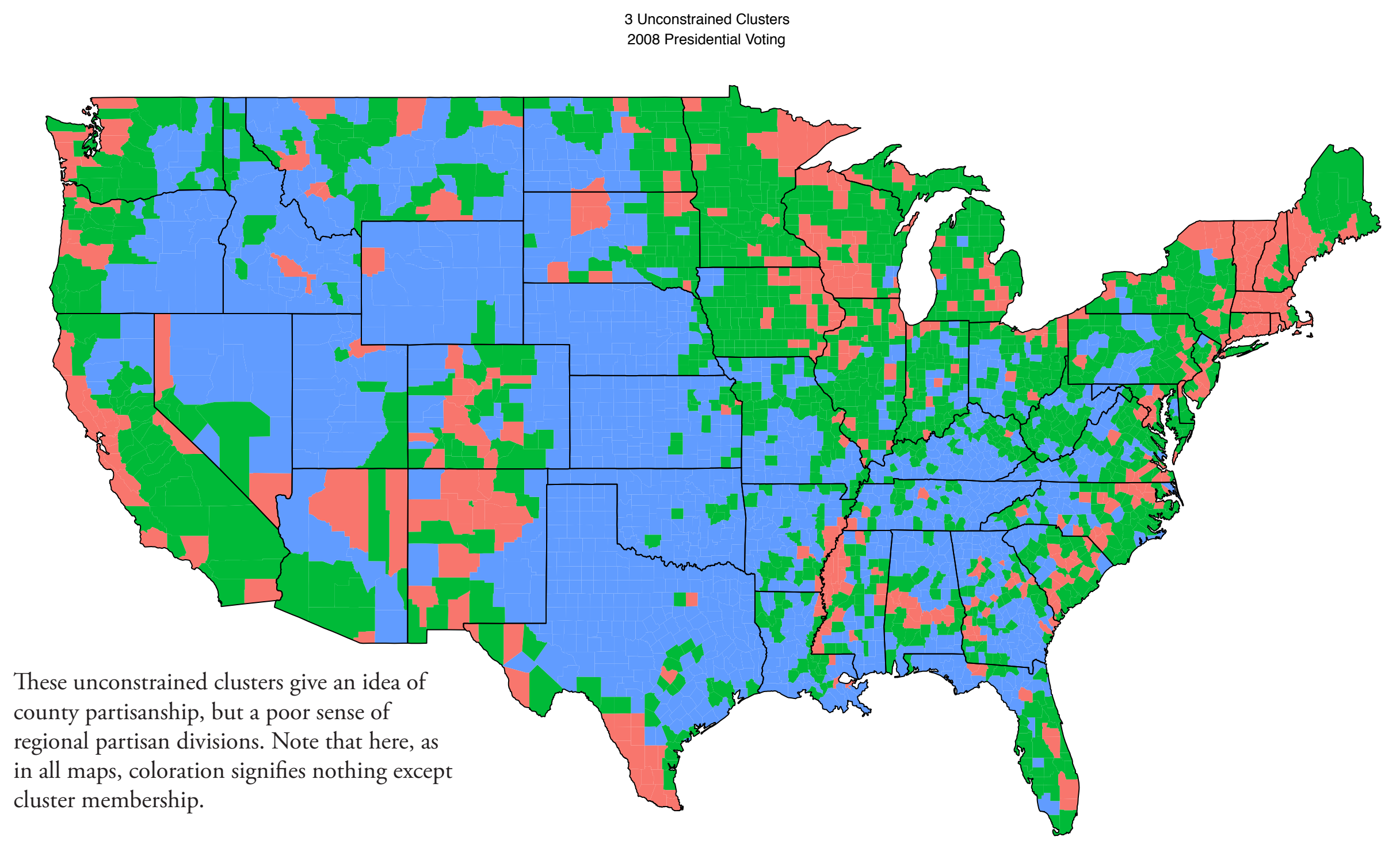
Recchia, Anthony. 2010. "Contiguity-Constrained Hierarchical Agglomerative Clustering Using SAS." *Journal of Statistical Software* 33.


3 Contiguous Clusters
1968 Presidential Voting

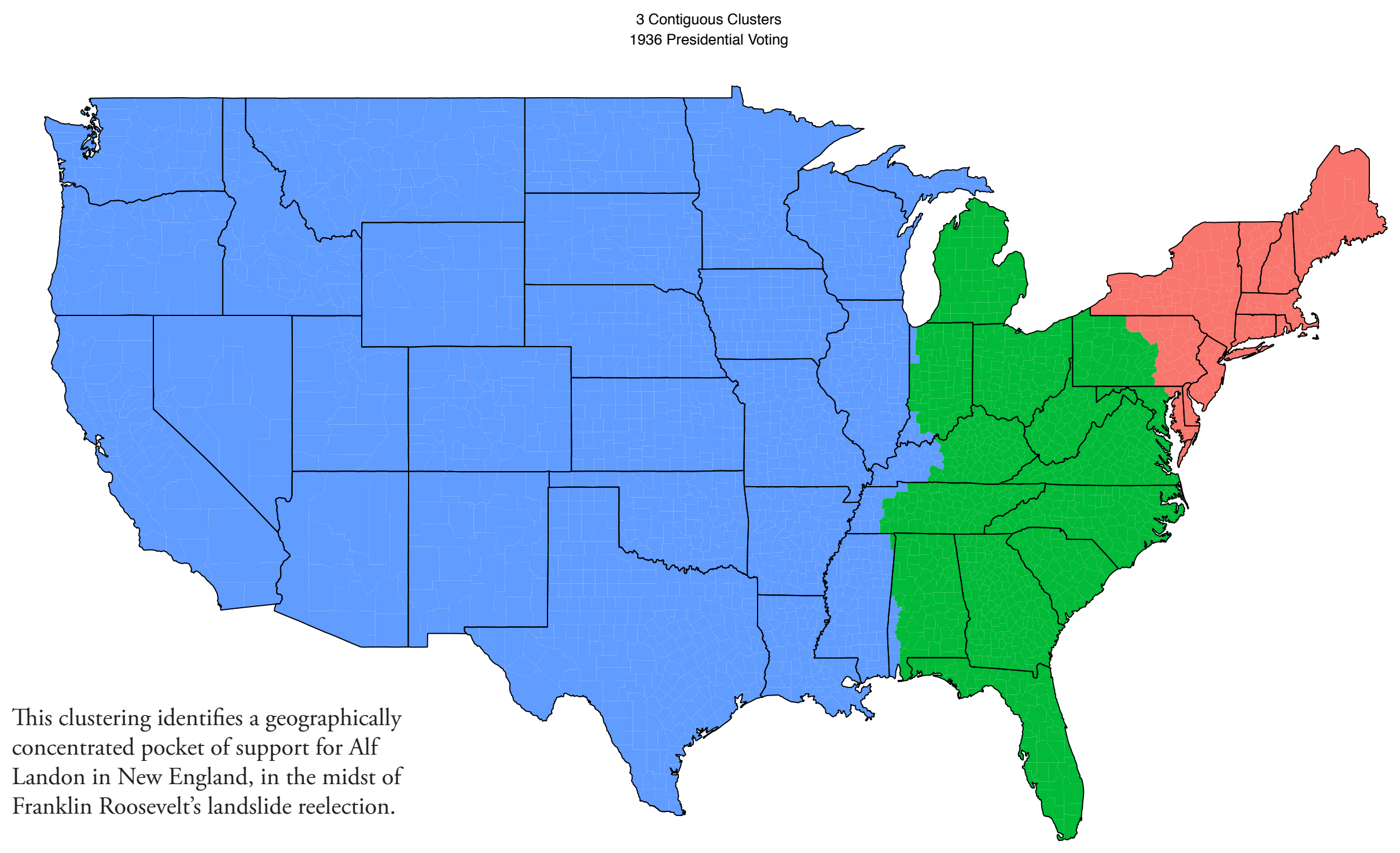
Richard Nixon, Hubert Humphrey and George Wallace divided the nation longitudinally, as a third-party candidate and changing Democratic party altered the typical North-South division.


50 Contiguous Clusters
2008 Presidential Voting

With 50 clusters, we observe differences within states, and political localities centered around major metropolitan areas and geographic features.


3 Unconstrained Clusters
2008 Presidential Voting

These unconstrained clusters give an idea of county partisanship, but a poor sense of regional partisan divisions. Note that here, as in all maps, coloration signifies nothing except cluster membership.


3 Contiguous Clusters
1936 Presidential Voting

This clustering identifies a geographically concentrated pocket of support for Alf Landon in New England, in the midst of Franklin Roosevelt's landslide reelection.