

Proposal for a  
Thesis in the Field of  
Information Technology  
In Partial Fulfillment of the Requirements  
For a Master of Liberal Arts Degree

Harvard University

Extension School

July 28, 2009

Michael Tracey Zellmann

[tracey.zellmann](mailto:tracey.zellmann@gmail.com) gmail

Proposed Start Date: September 1, 2009

Anticipated Date of Graduation: TBD

Thesis Director: TBD

## I. Tentative Title

Creating and Visualizing Congressional Districts

## II. Abstract

My thesis addresses the problem of dividing states into voting districts for electing Representatives to the US Congress.

This is important, relevant and timely. We live in a democracy, and the votes of our citizens are accurately and legally counted. Those votes decide elections. However, for the US House of Representatives, there is a twist. "We like to think that voters choose their politicians-but in the redistricting process, politicians choose their voters." Levitt (2009)

Following the census every ten years, many State Legislatures have to draw new maps for the voting districts of Representatives to the Congress. The Constitution gives State Legislatures broad discretion over drawing those boundaries, notwithstanding decisions of the Supreme Court and legislation like the Voting Rights Act of 1965. The party in power can draw those boundaries to create an outcome that favors themselves. For example, the Democrats controlled the Texas Legislature in 1990. They created a district plan that took effect with the 1992 election. The Democrats won with a small plurality - less than 2%. However, when the Congressional results were tallied, Democrats had won 70% of the 30 seats - with only one of the contests closer than 10%. All those votes were legal, but some counted more than others did. Redistricting altered the effect of those votes in ways the Founding Fathers probably did not intend.

The census of 2010 will lead to a reapportionment. It is likely that 12 states will lose 14 seats in the Congress, and nine states will gain those 14 seats. Eleven of the 14 increases will be in states where Republicans control the legislature, with Texas gaining four. Following the realities of partisan politics, those legislatures will draw new districts that will likely elect Republicans in 2012, and tilt the House towards Republican control, with a ripple effect through Committee chairmanships. This will serve the purposes of the Republican Party - but is it fair to the voters?

The process for creating these districts is complex, controversial and not well understood by the public. The process is conducted out of the public eye with little press coverage. It is ripe for behind-the-scenes horse-trading. As long as the process continues in relative darkness, legislators will make redistricting decisions according to their own priorities. It certainly helps incumbents. The process may not be the only reason, but looking at Congressional elections from 1982 - 2004, 95% of incumbents seeking re-election won their districts.

Open Government in the Executive Branch through "Democratizing Data" is a major theme of the Obama administration. A Web-based application could open up the redistricting process. With tools to understand, explore and compare alternative plans, citizens could participate in the process. The result would better reflect the interests of the electorate, not just the Representatives. Part of my thesis will create an application that applies "Open Government" to Redistricting.

My thesis has four pieces:

- Explore the theoretical foundations and algorithms for subdividing the states - as a Set-Partitioning or Graph-Partitioning problem.
- Acquire and serve the geographic and census data needed for redistricting the states, along with election statistics. In addition to MA, I will focus on NY, FL, TX and CA - large states with complex demographics and changing populations.
- Implement and refine several algorithms; measure and compare results and performance; create valid plans and choose the best ones.
- Create two interactive visualizations.
  - A desktop visualization to refine plans - available as a Java Web Start application.
  - A browser-based visualization to let people understand and compare plans, seeing how the demographics and voting patterns change, and what it means for an individual voter.

### III. Thesis Project Description

#### A. Background of the Project

I put Redistricting into context: What does the Constitution require; What is Gerrymandering; Supreme Court decisions and the Voting Rights Act; Measurements and Definitions; How partisan legislatures draw districts to their advantage; 2010 census and the elections of 2012; Recent examples of Gerrymandering;

Note that this is not a Thesis in Government or History. Nor is it a legal brief. I try to cover some of the important facts to provide context. There are complex background issues when translating this real problem into a computer model. A reader can skim or even skip this section, depending on depth of interest. However, a reader should read the Measurements and Definitions section to understand terms used throughout the body of this proposal.

#### CONSTITUTIONAL REQUIREMENTS

The Constitution defines the election process. Article I, Section 2 states - the "House of Representatives shall be composed of members chosen every second year by the people of the several states," and that an "Enumeration will take place ... every ten years." Article V states that "each House will be the Judge of the Elections, Returns and Qualities of its own Members." This clause has made the Supreme Court reluctant to intervene. The Fourteenth Amendment, passed in 1868, states that "all persons, born in the US, or naturalized citizens," have the right to vote. Section 1 of this Amendment goes on to say, "No State shall make any law abridging the privileges of citizens nor deprive any citizen equal protection under the laws." This becomes the fundamental legal argument for voting rights. Additionally, Section 2 states "Representatives shall be apportioned among the several states according to their respective populations."

Working from these principles, after each decennial census the seats in the House of Representatives are apportioned among the States. For each State, the census reports the population by county, census tract, block-group and block. Each State, according to its own laws, creates a plan to elect its Representatives. Early in the Republic, Representatives were elected at large. Eventually, people realized this led to "majoritarianism" where local interests were overlooked. In the mid-nineteenth century

States adopted voting Districts, letting voters in each District elect their own Representative. As the country grew, the number of Representatives was increased until 1913, when the size was capped at 435. In defining its plan, each State assigns census tracts to voting districts in time for the next election. The plan is aligned with existing political boundaries - cities and towns, wards and precincts, particularly for counting votes.

### GERRYMANDERING

This process worked, but was subject to manipulation by the incumbent legislators who created the district plans. In 1812, the Boston newspapers named this "Gerrymandering" (after MA Governor Eldridge Gerry, Harvard 1762). Not a dictionary definition, but a realistic definition of Gerrymandering would be: "Gerrymandering is an abuse of the redistricting process to draw election district boundaries that give a significant unfair and undeserved vote count advantage in future elections to the majority political party, which controls the redistricting process, and to incumbent politicians of all political parties." Robbins (2007).

### SUPREME COURT DECISIONS AND THE VOTING RIGHTS ACT OF 1965

If these plans caused complaints, the courts, all the way to the US Supreme Court were slow or reluctant to get involved - treating this as a matter for the Legislative Branch - as stated in the Constitution. One thought was that if the voters did not like what the politicians were doing - they could vote them out. Eventually, people realized that those same legislators were creating the plans that protected them from the voters. In 1960, the issue became focused on race. The Supreme Court was faced with cases showing that districting was used to render Black votes ineffective, denying Black people their voice in government. This culminated in the landmark Voting Rights Act of 1965. Justice Douglass, in *Gray v. Saunders* 1963, wrote:

"The Equal Protection Clause of the Fourteenth Amendment ...(requires) political equality ... (which) ... can only mean one thing ... one person, one vote."

The Voting Rights Act of 1965 outlawed historic practices used in the South to deny Black people the vote - literacy tests and the poll tax, for instance. It also required States to take affirmative steps to improve minority opportunities. Several States took that to mean districts should be created that made it likely that Blacks *would* be elected. The Act also required regions that had a history of voting discrimination - 9 States, mostly in the South, along with regions in six other States, surprisingly including Northern States like NY, VT and CA, to get "pre-clearance" from the Justice Department for any changes to their voting arrangements. For all others, it took a challenge from a citizen before the courts were involved. The Voting Rights Act has been renewed, but many people believe that these issues are behind us and things like pre-clearance are no longer appropriate. The Supreme Court will probably review the constitutionality of elements of the Voting Rights Act like pre-clearance in the future.

These decisions changed parts of the redistricting process. States now tried to create "minority/majority" districts - where a minority group could have a majority within a district and likely elect their own representative. Judgments that are more objective were needed on the fairness of district plans. In the past, States had used "traditional districting practices" - but now, they also had to evaluate population equality, contiguity and compactness. Those terms needed definition.

## MEASUREMENTS AND DEFINITIONS

Several terms are used frequently and should be carefully defined: Political and Census boundaries; Traditional districting practices, population equality, contiguity, and compactness.

States are divided into counties, and then into cities or towns. Cities and towns are divided into **precincts** - where each citizen casts their vote. Cities may also have **wards**, which group several precincts. The Census counts people by counties within each State. Counties are divided into census **tracts** - which do not cross county boundaries. Tracts are further divided into **block-groups** and ultimately census **blocks** - the smallest counting area for the census. Since census tracts may cross city boundaries, census and election reporting may not align exactly.

Except for racial equality, the courts now expect each state to use its own **traditional districting practices**. These may vary by state, but can include following political or natural boundaries, avoiding contests between incumbents and grouping voters with specific shared political interests. The resulting Districts should have equal populations, as much as is practical. If they are irregularly shaped or are not compact, there should be a good reason. Unfortunately, even for these terms, standard criteria are lacking. States had used local political boundaries - such as counties, cities and towns, along with natural boundaries as well as voting precincts and wards to create voting districts. They also made general reference to compactness and contiguity, but without being specific. Now they had to consider census tracts, block-groups and blocks.

At first, the Courts expected Districts to be reasonably equal. Over time, with cases continuing to arise, they started to insist that populations be as equal as practical. There may even be multiple ways to measure equality. For consistency, throughout the rest of this paper, I will adopt the definition provided in Minnesota Senate (2000). Population **equality** is the range from the smallest District to the largest District, expressed as a percentage of the "perfect" District. For example, MA in 2000 had a population of 6,349,097 and was apportioned 10 seats. Each District should have a population of 634,910. The largest District has 636,554, the smallest 633,846, giving a range of 0.42%. A deviation is also defined - the average of the absolute value of the difference between each District's population and the "perfect" District, expressed as a percent. For MA, that is 0.10%.

**Contiguity** requires that all the tracts or cities within a District have neighbors in that District. Neighbors might share just one point, but usually they share at least one edge.

**Compactness** is another District measurement. There are different ways to measure compactness, and no one, particularly the Supreme Court, has established a consistent approach. Altman (1998) describes and compares 30 different definitions. If the same district is ranked differently by different methods, comparisons will depend on which method is used. Consider whether to measure area or population. You can look at a map and recognize an irregularly shaped District that might indicate Gerrymandering. However, the border might be following a natural feature, like a river or mountain ridge. Area-compactness can be measured by the ratio of the area of the District to the area of the smallest circumscribed circle. Population-compactness can be measured by moment-of-inertia: the average distance between each pair of census tracts, weighted by population. Both rational approaches, but they can give different results for the same district. If the district had large, sparsely populated, irregularly

shaped tracts on the perimeter and densely populated concentrations in the center, the area approach would evaluate this district irregular, while the population approach would evaluate it as compact.

Two states - Colorado and Iowa - require compactness in their plans. They generally want District boundaries to follow rectangles. They have regular borders and their internal boundaries are mostly parallels of longitude and latitude - e.g. aerial photos of cornfields in Iowa. For most other states, compactness does not come easily.

There was hope that compactness could be a “silver bullet” in the redistricting process. Computers could create districts that were equal and as compact as possible. Automation would remove politics from the process. It does not look like that will be possible. Automation may create good plans, but they will need refinement to meet targets. Even then, depending on the approach, different algorithms will produce different plans. Some people argue that an automated process would have to be unpredictable. Otherwise, partisans would choose whether to support a process based on the expected outcome.

Notwithstanding the emphasis on equality, census counts are only estimates. Even the estimate is only true for “Census Day” - April 1 of the year ending in 10. The Supreme Court recognized that absolute adherence to census counts might be misguided in deliberations over a Georgia plan. By the time the case was considered, five years had passed. Georgia was a fast-growing state and much of the data had changed.

In summary, Districts should have equal populations, as much as is practical, be contiguous and compact.

#### HYPOTHETICAL EXAMPLE OF GERRYMANDERING

An example may help illustrate how Gerrymandering works. Gerrymandering uses techniques called packing and cracking to amplify or dilute the effectiveness of individual votes. The following example is adapted from Levitt (2008). Imagine a State with 80 cities and four Districts. Overall, the State is evenly divided between Republican and Democratic voters, although there are geographic concentrations within the State, as shown in the figure below. Cities with a plurality of registered Democrats are blue, while those with a Republican plurality are red. Particularly for local elections, voters tend to vote according to their registration. As shown, a Democratic-controlled legislature could create a plan that would likely capture 75% of the seats. Conversely, a Republican-controlled legislature could create a plan capturing 75% of the seats. As long as this did not show any evidence of racial bias, it would be legal.

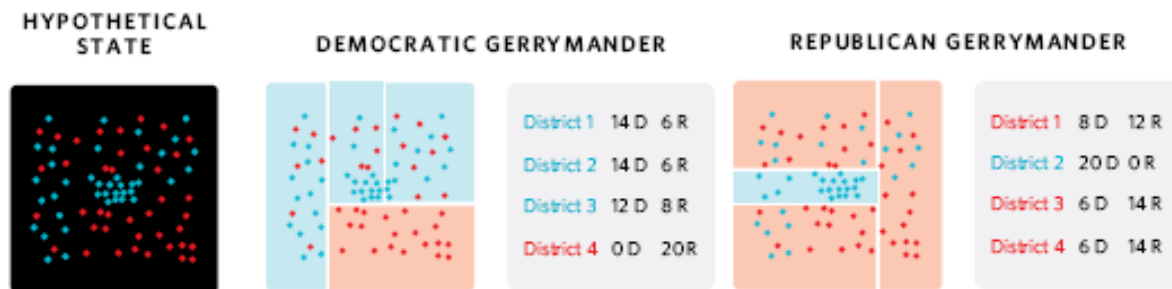


Figure 1. Hypothetical state that can be divided into districts to favor Republican or Democratic outcomes.

## THE 2010 CENSUS AND NATIONAL ELECTIONS IN 2012

The country has changed since the 2000 census. The count in 2010 will show growth and a continued shift from the Northeast to the South and West. As the figure below shows, 12 states will probably lose 14 seats, with nine other states gaining those 14. Howard Dean, as chairman of the Democratic National Committee, pursued a "50-State" strategy, pushing the Democrats to contest every State. One reason was a hope to control much of the upcoming Redistricting. The Democrats now control 27 of the 50 State legislatures. However, if we look carefully at the States where increases will take place, Republicans control most of them. Following the realities of partisan politics, Republicans will create plans that elect Republicans. The 2012 elections will likely increase Republican seats and tilt the House of Representatives toward Republican control. This will be legal, but may exaggerate the votes of some people in some districts - transformed through the distortion of partisan redistricting.

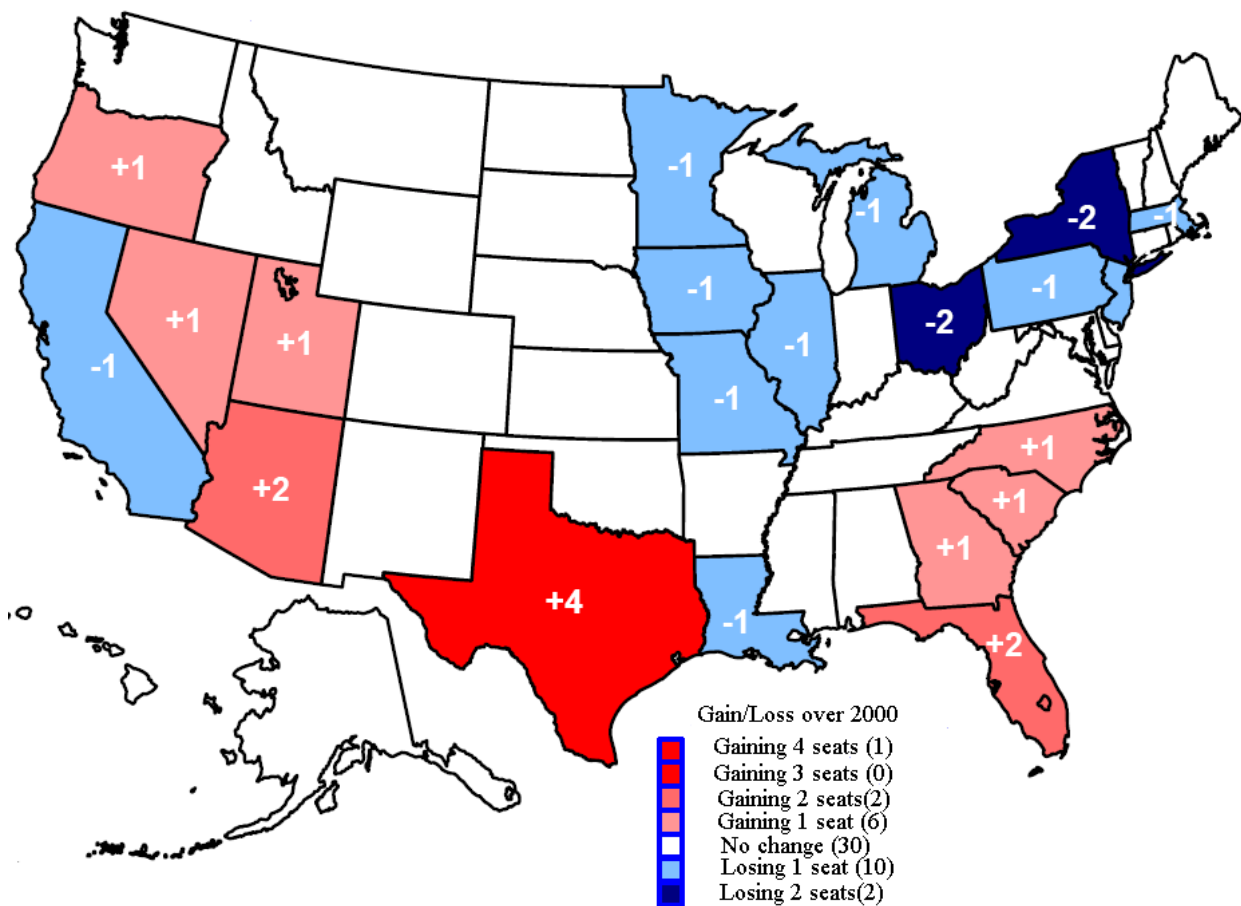


Fig 2. US Map showing states gaining and losing seats in the House of Representatives, based on 2010 census.

Source <http://www.polidata.org/census/st007nca.pdf>

## RECENT EXAMPLES OF GERRYMANDERING

### RACIAL DISCRIMINATION

In 1960, Tennessee redrew its district plans. Specifically, it rearranged the boundaries of Tuskegee so that 400 black voters were no longer within the city limits. There remained only five black voters in all of Tuskegee. Before this, the federal courts had been reluctant to intervene in elections, other than assuring that everyone could vote and those votes were counted. However, the NAACP sued, claiming that these voters were effectively disenfranchised. This case pushed the Court to intervene in election practices.

The following are legal, in the nature of “rough and tumble” politics - adapted from Levitt(2008)

### POLITICIANS CHOOSE THEIR VOTERS

Following the 2000 census, even though the Democrats controlled the California legislature and the Governor’s mansion, the Republicans threatened to put redistricting on a referendum if the Democrats tried too much manipulation. If a plan had to go to the state court, the Democrats would probably lose. The California Supreme Court had six Republican appointees and one Democrat. The result was an internal compromise to try to keep all the incumbent seats safe. The Democratic Party paid Michael Brennan \$1.3 million to create the resulting plan. In addition, 32 Democratic members paid Brennan \$20K each to custom-design their districts for safety.

### ELIMINATE INCUMBENTS

After 2000, the Republicans controlled the Virginia legislature. As part of their new plan, they targeted the Democratic minority leader, Richard Cranwell - a 29-year veteran of the state legislature. They carved his house, along with 20 neighbors out of his district and placed it in the district of another Democrat - Chip Woodrum - a 22-year veteran. Rather than run against Woodrum, Cranwell chose to resign.

### ELIMINATE CHALLENGERS

In the 2000 Democratic primary in Brooklyn NY, Hakeem Jeffries challenged long-time incumbent Roger Green - winning an impressive 40% of the vote, and setting the stage for a real electoral contest. In the meantime, however, the sitting legislators, including Green, rearranged the district, removing Jeffries’ house from that district. Green won, basically unopposed, with 95% of the vote.

### PACKING PARTISANS

In 1991, Texas Democrats mapped the 6th district - suburban Dallas - to contain as many Republicans as possible, making it easier for Democrats to control the neighboring districts. In the left figure below, the white area is the District containing the Republican plurality. To the right is the current districting of Dallas, continuing to show irregular boundaries.

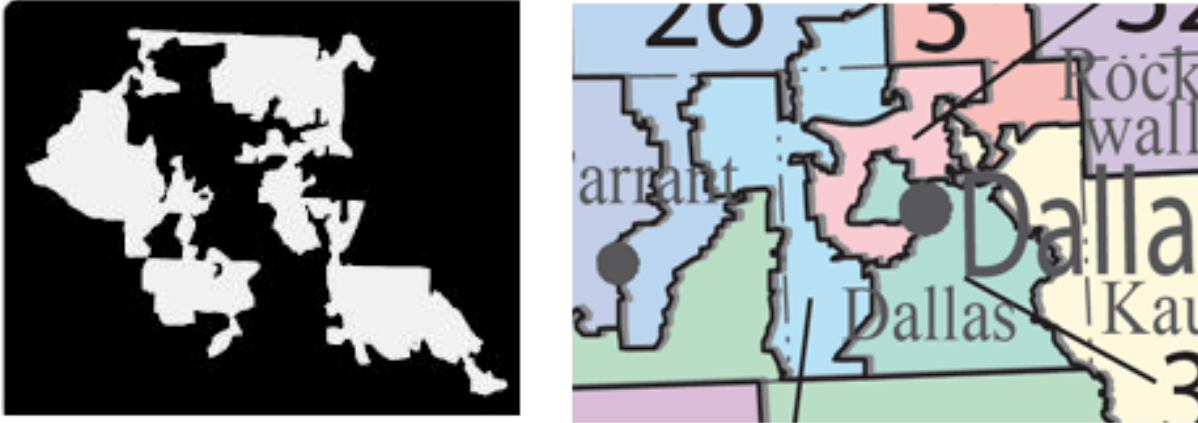


Fig. 3. Map of Congressional District around Dallas. 1992 on left. 2008 on right.

## B. Existing Work

I discuss the work that already applies computer technology to this problem: survey the literature and available programs that produce district plans; provide examples of plans created by existing programs; explore the theoretical foundations of the problem as an example of graph partitioning.

There is a large body of work in this area, dating back over 40 years. Nagel (1965) wrote about a program on an IBM 7090 that created a district plan for Illinois. He carefully described how to set up the punched-card data - and how to layout the 132-column printout. Redistricting helped define GIS (Geographic Information Systems). By now, there are several commercial software systems and a few open-source projects developed by interested people with a facility for programming and GIS that create and measure plans. I am focusing on the open-source work, where the algorithms are accessible to me. The intended users of these applications are government and GIS professionals or people deeply familiar with computer systems. They are not meant to help citizens understand the problems. Computer-minded people receive them enthusiastically, but they may fail to reach the general population.

### LITERATURE AND IMPLEMENTATIONS

Dr. Michael McDonald and Dr. Micah Altman have written an open-source application - BARD (Better Automated Redistricting) that is an extension of the R language, and described in Altman (2009) strongly based on Altman's Doctoral dissertation *Districting Principles and Democratic Representation*, Altman (1996). It uses a variety of algorithms and heuristics.

Brian Olson - Olson (2009) - has written an open-source application in C++ and created plans for all the US states. He uses clustering and hill climbing to minimize the average distance between any person and the center of their district.

George Clark - Clark (2004) - has developed an application in Pascal to group neighboring census tracts as long as the groups do not exceed the population limits. Compactness is the total of the perimeters of all districts. The program examines all the tracts on borders between Districts for swaps that can improve the compactness. The program seeks the most compact arrangement that keeps District populations within 1% of each other.

Span, Gulotta & Kane (2007) - They describe two methods - Recursive Splitting and Moment of Inertia.

Recursive Splitting – A function will split a region into an even ( $2n$ ) or odd ( $2n+1$ ) number of sub-regions. A least-squares line is fitted to the tract centers. The program establishes a perpendicular with that line at a point that divides the region into the correct ratio of population, either  $n:n$  or  $n:n+1$ . The recursion ends when  $n$  equals 1.

Moment of Inertia – Uses  $k$ -means clustering, with  $k$  the number of districts, minimizing the moment of inertia. The distance function is the average distance between each pair of people in a district.

Tobler (1986) describes a method that distorts a map to make the area represent the population. Then the districting task becomes dividing the state into regions of equal area.

Chandrasekharam (1993) describes a graph-partitioning approach used in VLSI design, which also uses a Genetic Algorithm.

Crincione (2000) describes four related algorithms. The basic one begins by randomly picking a tract as the starting point for the first district. It proceeds by picking randomly from the neighbors of the district, until the population of the emerging district meets the desired level. A district is thrown out if it is out of range. The next district begins with a tract randomly selected from the neighbors of existing districts.

The Center for Range Voting (2008) has created a shortest split-line algorithm and applied it to all the states. It is similar to the recursive splitting approach above.

Graph-partitioning is used in other areas, and some of those approaches may be applicable here. Service regions are divided to get better on-site response times. Circuits are divided to place components on different printed circuit-cards with minimum interconnect. Kernighan (1969) wrote his Ph.D. thesis on applying graph-partitioning to programs to reduce memory paging. Lately, similar techniques are being applied to parallel processing. To divide a job among parallel processors, model the job as a graph with steps as vertices and dependencies as edges. Partition the graph into sub-graphs of near-equal size with minimum edge connections and the separate steps can proceed with minimum waiting for other steps. Henderson (2008) from Sandia Labs has an implementation for parallel processing called Chaco. I will try to apply some of these techniques to redistricting.

All these approaches are interesting and informative to me, but I believe they do not connect with the general public. They ignore city boundaries and set a population equality target inferior to existing plans. As a result, the plans look academic and detached. Below is an example of Brian Olson's results for Massachusetts and the Moment of Inertia – Clustering approach for New York.

SOME RESULTS FROM EXISTING WORK

Olson - Massachusetts



Fig. 4. Congressional Districts in MA. Current plan on left. Alternate plan generated by Brian Olson's software on right.

Moment of Inertia - New York

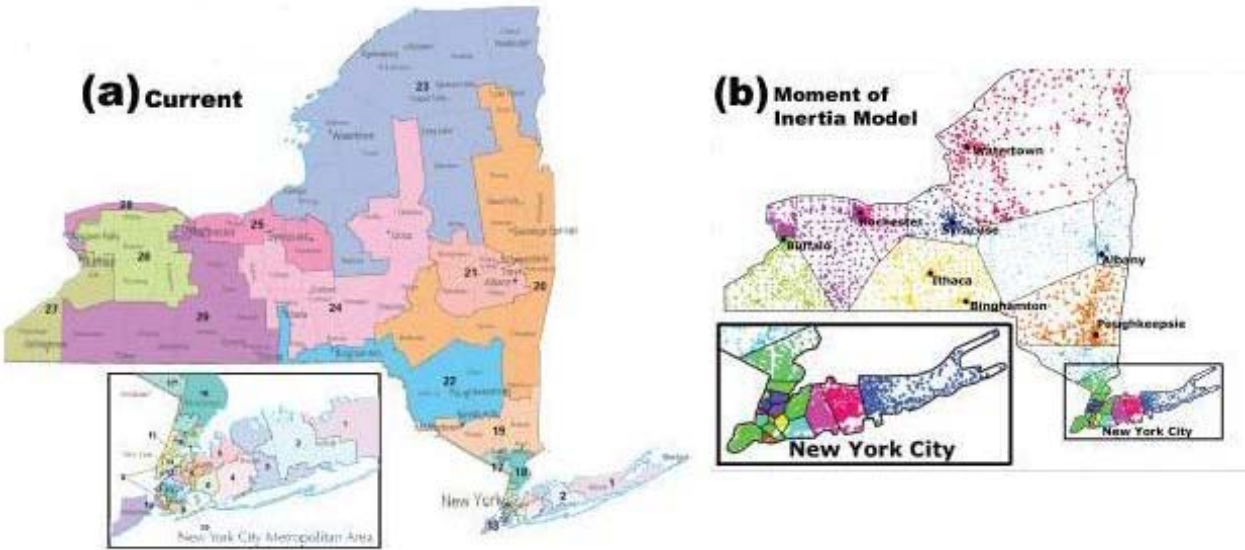


Fig. 5. Congressional Districts in NY. Current Plan on left. Alternate included in Spann (2008).

These results are typical and I believe they do not connect with what people are familiar with, and this will prevent their adoption. They remind me of Warren Buffet's admonition - "Beware of geeks bearing formulas." From the resulting maps, it is difficult to see what happens to your own city. The solutions ignore existing city boundaries and typically have population equality inferior to existing plans. For example, Olson's results for MA have an equality measure of .97% while the current plan measure is .42% (smaller is better). The Moment of Inertia method considers an equality measure of 2% good enough. Clark expects equality of 1%, but when applied to AZ could not do better than 5%. Altman

advises that clustering will not produce equal population districts and the districts may not be contiguous. The resulting plans need manual refinement.

#### THEORETICAL FOUNDATIONS: GRAPH-PARTITIONING, COMPLEXITY AND HEURISTICS

Dividing a state into districts can be considered a graph-partitioning problem. Census tracts are vertices, connected to neighboring tracts with edges. Graph-partitioning is NP-Complete - Skiena (2008). The solution space is large and increases exponentially with the number of census tracts. Eliminating branches that lead to non-contiguous solutions or ones that are outside of the population tolerance can narrow the search for a solution. However, all the applications in use generate solutions that depend on heuristics and are only locally optimal. If there were  $n$  tracts and we want to divide them into  $r$  districts, an exhaustive search would evaluate a large number of possible solutions:

$$S(n, r) = \frac{1}{r!} \sum_{i=0}^r (-1)^i \binom{r}{i} (r-i)^n$$

### C. How My Work Will be Different

I discuss how my work will be different from existing approaches: Reach a target audience of non-programmers with a web application; Produce comparisons of various algorithms subdividing several states; Use KML as an open-source format for geographic information; Minimize city-splitting as each state is divided; Provide a visualization tool for refining plans; Visualize the resulting plans in an interactive and informative way, using overlays on Google maps and Google Earth.

#### TARGET AUDIENCE

My thesis creates an application intended for users who are not programmers or GIS professionals. Through a web interface, interested citizens should be able to explore, compare and evaluate alternative plans. For the covered states, I will provide several alternative plans to compare with the existing plan. Users can also modify and refine those plans. Plans can be exported or imported to different Redistricting applications.

#### COMPARE PERFORMANCE OF DIFFERENT ALGORITHMS

After acquiring and arranging the geographic and population data in a consistent manner, and implementing several of the algorithms, I will run experiments, seeing how the different approaches perform against a wide range of data. I will profile the algorithms and make focused improvements. I will present the resulting data as a lab report. The algorithm work-bench will be written in Java with documented code, installation and operation instructions to enable another Java programmer to duplicate the results. I will also provide clear instructions on how to create plans for additional states. The work-bench will provide common services to the algorithms - functions like distance between points, neighbors of a tract, whether a region of tracts is connected, determining the area and boundary of a group of tracts.

## OPEN DATA FORMAT

My application will use open formats for the data. KML is the XML extension that Google uses for Google Earth and Google maps - based on the technology they acquired from Keyhole. This will let my application share the contributions made to Open-GIS from a worldwide base of developers. I will import and export district plans, providing an XML schema for the protocol. Already, several people who have looked at my initial visualization would like to use it for their plans, but there is no standard format.

## CITY SPLITTING

The current implementations do not recognize existing city boundaries. They simply create a grouping of census tracts. Most of the real plans I have seen use many city boundaries. I believe citizens will relate better to plans that have recognizable city boundaries. I also believe people in the same city should have the same representative. It may not always be possible to achieve population equality without splitting some cities. I plan to measure "city-splitting" and try to achieve good equality and good compactness with minimum city splitting.

## REFINING PLANS

Most implementations use algorithms to automate the process completely. However, there are limits to how good these solutions can be - particularly given the intractable nature of the problem. I will include a visualization tool that helps refine the plan. The display will show the district map with an overview of the population equality and compactness measurements, indicating the biggest contributors to those scores. The display will highlight the areas on the borders between districts, indicating the relative populations. That will guide the user to make swaps to improve population equality.

## VISUALIZE RESULTS, GOOGLE MAPS/EARTH

The current implementations do not give ordinary citizens a good picture of the results. I will generate KML files for the plans, letting them be displayed and explored as overlays on Google Earth or Google Maps. The plans show up on a familiar map. Users can explore how the plans differ for their specific city or street address.

I will provide several basic plans, along with the existing plan and any available draft plans. The user selects two plans to explore and compare. I will show those plans side-by-side, with comparative measures of population equality and compactness. The overall demographics of the state are shown - age-mix, racial-mix, income and household value distributions. Selecting a district in either plan will enlarge that district and show the specific demographics. Choosing from a list of all cities in the state displays the appropriate district in both plans, along with demographics. With those tools, a citizen can see how the plans differ and what they might mean for their own vote. The demographics describe the people with whom their vote is pooled. Where voter registration and election statistics are available, I will show election history and how those results might have changed with the different district plans. As part of CS E64 Spring 2009, I built an interactive visualization tool with similar capabilities, using a manually generated redistricting plan for Massachusetts - screen shot below.

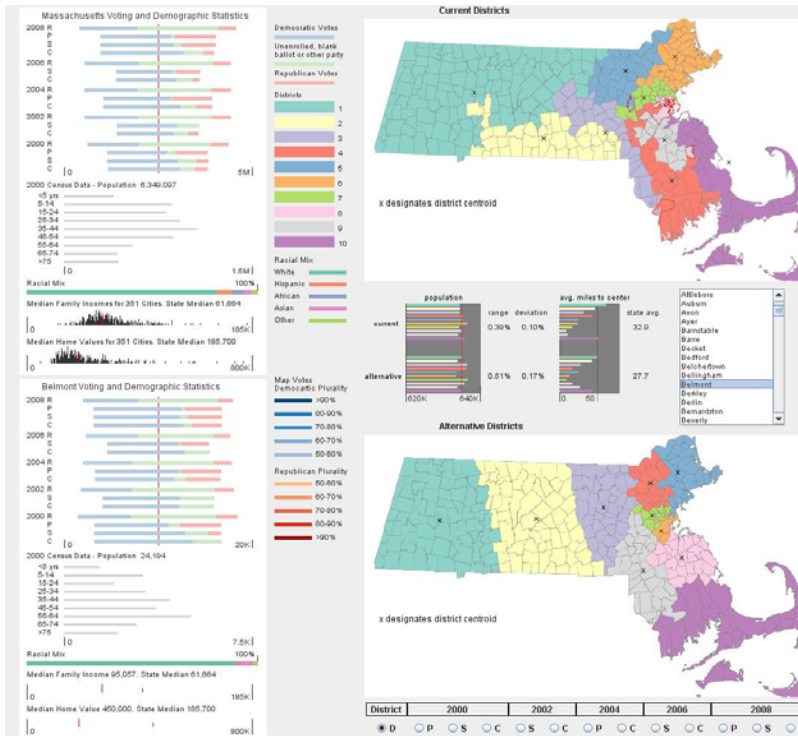


Fig. 6. Screen Shot of interactive visualization of current and alternate districting plan for MA.

## D. Description of Components

I describe the software elements of my project: Supply the data; Environment to run and measure algorithms; Implement and improve algorithms; Port existing implementations to my environment; Generate best plans; Visualize those plans across the web; Implementation and operation.

### GEOGRAPHIC AND POPULATION DATA STORE

The first task is to acquire the information - the boundaries of the census tracts in each state, along with the population and demographic data, and if possible, election statistics - voter registrations and election results. I need to assemble the data in a consistent manner for easy access by the algorithms. The geographic information is available from the Census Bureau as TIGER files (Topographically Integrated Geographic Encoding and Reference) using the ArcInfo format developed by ESRI (formerly Earth Sciences Research, Inc). I will convert this to KML format. The population count and demographic information is available from the Census Bureau Fact-Finder service. Voter registration and election results come from the Secretary of State for each State. I did obtain that information for MA, but it was time-consuming. I may not be as successful with the other states.

### ALGORITHM WORKBENCH

I will create a framework to conduct experiments using the algorithms against a range of data. I will assemble the results over various conditions, as well as profiling performance and making improvements. I will publish the results as a lab report. There are numerous functions needed by the several algorithms - e.g. distance between points, gathering cities or tracts into a larger region, writing

the KML for a region boundary. I will provide those as common functions. This should make the actual algorithm implementations simpler.

#### IMPLEMENTATION OF SEVERAL ALGORITHMS

I will implement several of the algorithms from the literature, probably including the Moment-of-Inertia and Recursive Splitting from Span (2008), Area Morphing described by Tobler (1973) and the Graph Partitioning described by Chandrasekharam (1993). To illustrate the recursive splitting, the following two figures show the first and last step, applied to Massachusetts. This approach still needs refinement.

MA has ten districts. The first step splits MA into two parts of equal population. Each black dot is the areal center of a census tract. There are 1361 tracts in Massachusetts with mean population of 4665 and std. dev. of 1845. Draw a line AB, the least-squares fit to the census tract centers. Choose a line CD, perpendicular to AB which splits the state so that the total population of census tracts to the left of AB equals the total of those to the right of AB. I would do this by projecting the centers onto the line AB, creating an ordering by population, and then seek the pair of centers that straddled the desired ratio. The granularity of tract sizes will limit how exact the split is. These two regions are then split in a ratio of 3:2. We repeat this process until all the regions are of size one. Note that this only produces a single solution.

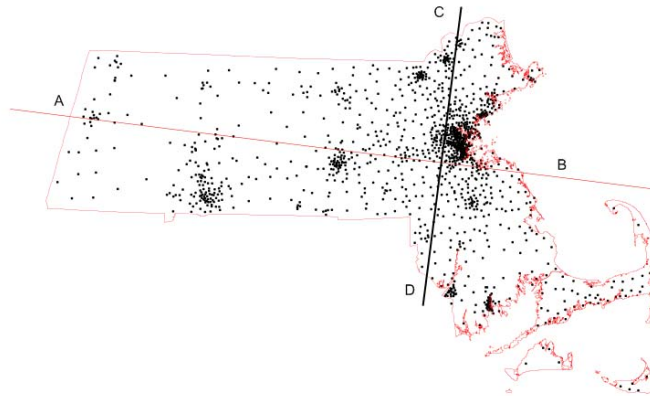


Fig. 7. First step in recursive-splitting algorithm applied to map of MA census tracts.

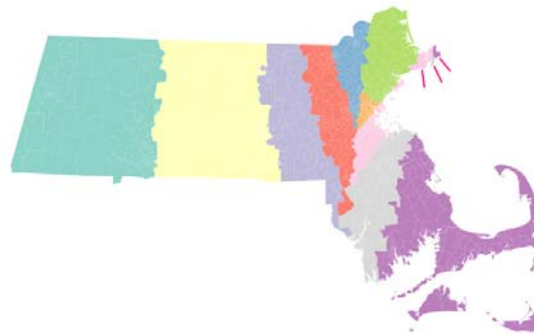


Fig. 8. Result of recursive splitting, showing boundaries of census tracts. Note that three small tracts on the North Shore have been put in Districts on the South Shore, forming Districts that are not contiguous.

This is my first implementation of this algorithm and has some problems with the convex shape of Massachusetts Bay - causing three regions on the North Shore (marked with red lines) to be included in Districts based on the South Shore. The resulting Districts are not contiguous. These two figures also show how I plan to use visualizations as I implement and evaluate the algorithms. It is much easier to see the problem with this recursive split picture than reading through a list of census tract ids.

#### PORT EXISTING IMPLEMENTATIONS

I plan to implement the algorithms used in the open-source solutions. I have the code, but I am not familiar with the languages. This includes BARD of Altman (2009), Olson (2009) and Clark (2004).

#### FROM THE BEST, CREATE DISTRICT PLANS FOR SEVERAL STATES

Given the range of approaches, I can pick the better ones for each State. I may be able to feed those plans into a final Genetic Algorithm to see how far the process can proceed. I could shift to a finer level of granularity - census block-groups or even blocks along the district borders for a final swap to improve population equality. Finally, there will be a range of plans with measurements available for selection and manual refinement.

#### DESKTOP INTERACTIVE VISUALIZATION FOR REFINING PLANS

The first visualization helps algorithm development and refinement. It will show the plan as a map, with information available on all the cities and tracts along the borders of each district. The user can try swaps and trades of tracts between districts, seeking a better arrangement. I will explore how to make that easy for the user and what to move back into the algorithms.

#### WEB-BASED INTERACTIVE VISUALIZATION TO LET VOTERS EXPLORE, UNDERSTAND AND COMPARE PLANS - HOW DO DEMOGRAPHICS AND VOTING PATTERNS CHANGE? WHAT DOES IT MEAN FOR A SPECIFIC VOTER?

The second visualization will show plans to public citizens. They can choose plans to compare, exploring the demographics of alternate district arrangements, as well as retrospective projections for voting results (for those elections where I have data) - would any results be different? A citizen can also ask about their specific city or address: "What districts am I placed in and how do they compare with each other or my current district?"

#### INSTALLATION AND OPERATION

I will provide instructions so a Java programmer can install and operate the algorithm workbench and create additional algorithms. I will describe the specific steps required to develop plans for an additional state. I will provide schemas for importing and exporting plans, as well as importing voter registration and election statistics.

## E. Technology Choices

I describe my technology choices: data format and storage; language for algorithms; displaying visualizations; web-technologies for graphics.

### DATA FORMAT AND STORAGE.

Traditionally, the geographic data is stored in ArcInfo format. There are individual edges – with a lat/long at each end. Each edge has a left and right region that it separates. Edges move clock-wise around regions. Edges are only stored once and a region is just a list of edges. This optimizes relational data base design. Unfortunately, this also locks everyone into ESRI's closed architecture, and you need their products to view or manipulate the information. Google has developed Google Maps and Google Earth using KML, a different, open format that extends XML. Each region is a polygon defined as a linear ring. There is a String of coordinates that describes the border – space-separated tuples, where each tuple is a comma-separated list of longitude and latitude, expressed as signed decimal-degrees, along with an elevation. Their implementation scales well, is widely used, and only requires a browser plug-in to view. There is an open-source tool, shp2kml, that converts the formats, and despite a few quirks, I have been able to use it. For the districting phase, I just need the district id, the center, the area, the neighboring ids and the population. For the visualization, I need the entire boundary.

The census demographic data is available from the American Fact Finder web site, part of the US Census Bureau. A relational database can import the data.

### LANGUAGE FOR ALGORITHM IMPLEMENTATION

For the partitioning algorithms, my major technology choice is which language to use. I am most familiar with Java and its libraries, followed by Perl, Ruby and Lisp. Python is also a serious possibility. The output is a list of census tract id / district number. I have already written several genetic algorithms using LISP, so I do not foresee problems there. I can represent each solution as an array of integers of size equal to the number of tracts, with each entry indicating to which district that indexed tract belongs. The hard part will be generating populations of valid candidates.

### DISPLAYING VISUALIZATIONS

Once the districting plans are complete, there are several choices for displaying them and offering an interactive visualization. My current application uses Swing and that is certainly a possibility, although it may not be easy enough for a broad population. In addition, the maps do not have any context – you cannot see any other detail. A better choice may be Google Earth or Google Maps. They would run in browsers and therefore, be widely available. They both use KML and avoid the problems of a closed format. Google Earth is bit more complex, but offers a wider range of zoom levels, while Google Maps only has 15 discrete zoom levels. My current Swing application uses Affine Transforms to fill the available window completely with whatever size map is in view – the entire state, or any one of the districts. Therefore, it would be better to use Google Earth. The design will need two concurrent browser sessions, one for the current map and one for the alternative. JavaScript can handle the interaction. I think this will be better overall, but if it is too complex or takes too long, I can use the Swing approach.

## WEB GRAPHICS

If I go with the browser design, I also have to create the visual charts. The Swing application uses the 2D-API for creating the charts. HTML5 now supports a Canvas, so that would work, but Microsoft has stated that they do not plan to support this capability. Therefore, an alternative is using PNG images generated on the server.

## IV. Work Plan

### A. Assumptions, Risks and Alternatives

Since there already exist working versions of these algorithms that successfully partition the states, I think the risk is minimal for creating one myself. I cannot predict whether I can create a plan that has population equality as good as current plans, with better compactness and measurably less city-splitting. Thoroughly exploring that may be the heart of the thesis.

For the public facing part of the application, there are two big risks: having a design that people will find meaningful and useful; acquiring enough election data to illustrate how the districts might have changed previous election results.

For the first concern, I plan to test my concepts – using my current application - on various members of the Massachusetts government, parties and population. However, I still may not be able to synthesize their feedback into a cohesive solution. Regarding the second concern, it was a surprise to me that election statistics were not available in electronic form from Massachusetts. To build the relevant data, I had to obtain the data in book form from the Secretary of State's office, scan them and process them with OCR. It was a daunting task, and not something I could do for the other, larger states. An initial search shows that some data does exist, but costs ~ \$200 for one year of New York State! I will have to keep looking. In the same vein, there are difficulties tying the data together when the plan splits cities. I processed Massachusetts's data at the city level, not at the precinct level. Even at the precinct level, I would need to tie that to the appropriate census tract. The alternative is giving each census tract in a city the same ratio of votes.

Election statistics adds depth to the tool and lets users explore how elections might have been different. At a minimum, I will try to get voter registration data.

Acquiring the geographic data, converting it to KML and loading it into the application with an appropriate data structure is a major task. So far, I have been successful with Massachusetts, although the KML locations were displaced. I may encounter unexpected difficulties with the other states. Additionally, since census tracts can overlap city boundaries, there are complications with mapping the districts onto a map of cities. This was not a large problem with my first application dealing with Massachusetts, since I did not do any city-splitting. To handle this, there is a query against the Fact Finder service - geo within geo - which returns tracts wholly or partially contained within a city or place. Presently, I have not been able to query multiple locations, so I will have to write screen-scraping routines to query all the cities in each state. This overall approach may not work. A counter-measure is to get data for each state sufficient to run a simple algorithm like Cirincione (2000), and display the results on Google Earth early in my schedule. If this does not work, my alternative is to use ESRI tools.

The other over-arching risk is time and the complexity of the whole application – along with possible interference from my other work. The initial work on algorithms and determining plans for several states will be solid, and even if I do not complete the interactive public visualization, that should still be a substantial and satisfactory thesis.

## B. Preliminary Schedule

Here is an initial estimate of the time-line. Some of these steps will be proceeding even before such a start. This may change if my exploration of the literature turns up better approaches.

Task	Checkpoint Deliverable	Duration - weeks
Read literature and existing code.	Algorithms chosen. Data requirements established.	4
Acquire geographic and census data. Storage strategy and KML conversion.	Google Earth display of each State and existing Congressional districts.	4
Algorithm workbench established. First algorithm working for each state - probably split-line.	Google Earth display of proposed district plans.	4
Implement other algorithms.	Google Earth display of proposed district plans.	6
Profile and improve algorithms.	Lab report on comparative performance.	4
Design refinement and web interface.	Illustrator / Photoshop mock-up.	4
Working refinement process.	Demonstrate edge swaps.	4
Working web-interface.	Demonstrate state plans and drill down.	4
Write Thesis	Document	6
Total		40

## V. Glossary

**Apportionment.** Based on the Census taken every ten years - each state's share of the 435 seats in the US House of Representatives is rearranged, to take account of shifting, growing and shrinking populations.

**ArcInfo.** A data format used by ESRI (Earth Sciences Research, Inc) to transfer data between their applications.

**Census Tract.** A defined neighborhood within which the Census Bureau counts and reports the population.

Congressional District. A geographic section of a state, which elects their own representative to the US Congress.

Gerrymandering. “to divide (a territorial unit) into election districts to give one political party an electoral majority in a large number of districts while concentrating the voting strength of the opposition in a s few districts as possible. This is done by arranging the boundaries to ‘pack’ the ‘other’ parties’ votes into as few districts as possible, or ‘diluting’ them by spreading them across as many districts as possible.”

Redistricting. Redrawing the boundaries within a state that define the Congressional districts. Usually done in response to a Census that increases or decreases the number of representatives that a state elects.

TIGER. Acronym for Topographically Integrated Geographic Encoding and Reference system - adopted by the US Census Bureau.

## VI. References

### A. Works Consulted

Altman, Micah, 1998. *Districting Principles and Democratic Representation*, Doctoral Thesis, California Institute of Technology

Altman, M. and P. MacDonald. Forthcoming. “BARD: Better Automated Redistricting” in Journal of Statistical Software. Available from <http://cran.r-project.org/web/packages/BARD/vignettes/bardJSS.pdf>

Center for Range Voting. 2008. Shortest split-line algorithm. Available from <http://rangevoting.org/SplitLR.html>

Cirincione, C., T. Darling and T. O’Rourke. “Assessing South Carolina’s 1990’s congressional districting.” *Political Geography* 19 (2000) 189-211.

Clark, G.L., 2004, *Stealing Our Votes*. Pittsburgh, PA. Dorrance Publishing Co.

Kernighan, B.W., “Some Graph Partitioning Problems Related to Program Segmentation.” Ph.D. Thesis, Princeton University, January 1969.

Levitt, Justin. 2009. “Drawing the Lines in Ohio: A Big Step Forward.” July 29, 2009. Retrieved from [http://www.brennancenter.org/blog/archives/drawing\\_the\\_lines\\_in\\_ohio\\_a\\_big\\_step\\_forward/](http://www.brennancenter.org/blog/archives/drawing_the_lines_in_ohio_a_big_step_forward/)

Levitt, Justin and Bethany Foster. 2008. *A Citizen’s Guide to Redistricting*. New York: Brennan Center for Justice at New York University

Nagel, Stuart S. 1965. “Simplified Bipartisan Computer Redistricting”, *Stanford Law Review*, 17, 1964-1965, pp 863-869

Minnesota Senate. 2000. “3. Equal Population” June 25, 2009. Retrieved from

<http://www.senate.leg.state.mn.us/departments/scr/redist/red2000/ch2equal.htm>

Robbins, Michael D. 2007. Gerrymander and the Need for Redistricting Reform. July 2, 2009. Retrieved from <http://www.fraudfactor.com/ffgerrymander.html>

Skiena, Steven S. 2008. *The Algorithm Design Manual*. London. Springer-Verlag.

## B. Works To Be Consulted

Chandrasekharam, R., S. Subhranian and S. Chaudhury. 1993. "Genetic Algorithm for Node Partitioning problem and application in VLSI design." *IEE Proceedings-Series E* 140(5): 255-260.

de Berg, Mark. 2008. *Computational Geometry*. Berlin. Springer-Verlag.

Stanton, Dennis and Dennis White. 1986. *Constructive Combinatorics*. New York: Springer-Verlag.

Tobler, W.R. 1973. "A Continuous Transformation Useful for Redistricting." In *Democratic Representation and Apportionment: Quantitative Methods, Measures, and Criteria*. New York: Annals of the New York Academy of Sciences.

## C. Software Implementations

BARD in R                      2009 Available from <http://cran.r-project.org/web/packages/BARD/index.html>

Brian Olson in C++            2009 Available from <http://code.google.com/p/redistrictier/>

George C. Clark in Pascal 2004 Available from GCLSr@aol.com