CSE 416, Section 1

Project Discussion

2

Project Teams

- All teams are formed
- 2 3-person teams (will have a lower use case target)

Group	Group Members
To The Moon 🚀 🙂	Ahmed Khan, Robert Leung, Brian Zhu, and Kenny Hu
Brewers	David Zhao, Matthew Ng, Peter Hwang, and Jonathan Tsui
Cardinals	Michelle Huang, Tina La, Vivian Lam, and Lisa Zheng
Diamondbacks	Gary Jiang, Hui Chen, Jimmy Lin, and Timothy Shi
Dodgers	Wendy Hu, Michael Liang, Meng Ting Liu, and Steven Zou
Giants	Joseph Wilson, Matt Jones, Kenneth Rose, and Usman Shafiq
Marlins	Xin Wang, Jessica Chan, Suheyla Iyimaya, Ertugrul Eryildiz
Le Monke	Joey Leung, Edwin Ma, Harlam Lee, and Georgy

3

Session Objectives

- Understand issues and terminology used in US congressional redistricting and voting analysis
- Understand some data requirements to support analysis of redistricting
- Understand relationship between redistricting and graph partitioning

We will explore the project functionality in more detail in the next 1-3 class sessions

4

Project Background

- Project based on
  - Spring 2019 – explore feasibility of integrating demographic data
  - Fall 2019 – analyze demographic voting patterns
  - Fall 2020 - use SBU SeaWulf computer to generate random congressional districts
- Lessons learned from previous projects
  - Robust set of data available
  - Feasible measures districting quality
  - Graph algorithms for the generation of viable congressional districts
  - Supercomputer integration for computationally difficult tasks

5 Why This is an Important and Interesting Topic

- Very current
- Lots of interesting CS concepts and technologies
  - Multiple languages
  - Multiple computers
  - Database
  - Algorithms
- Realistic software development project

Courts & Law

Supreme Court says federal courts don't have a role in deciding partisan gerrymandering claims

Thursday's 5-to-4 decision was written by Chief Justice John G. Roberts Jr. and joined by the court's other conservatives. (Sarah L. Voisin/The Washington Post)

By Robert Barnes
June 27

6 Background Info

- Every state has a number of congressional districts proportional to the state population
- Population is recalculated after a US Census. District boundaries must be recalculated if the number of representatives change or population shifts
- District boundaries can also change due to court decisions (e.g., Pennsylvania, North Carolina, etc.)

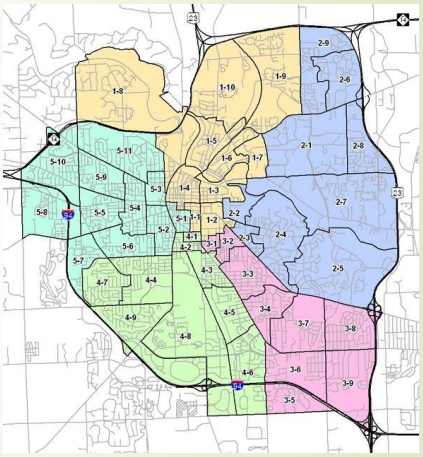
Population in districts must be almost equal

Most states will be redistricted following the 2020 census

US Census is performed every 10 years

7 Precincts (sometimes known as Wards)

- Lowest level voting division
- Contained in one polling place
- Data usually available for voting totals



Sometimes multiple precincts share a polling place

8 System Background

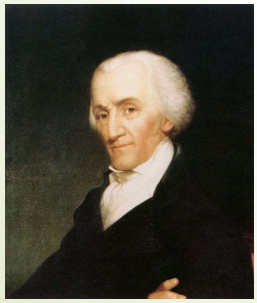
- Current redistricting approach leads to many unfair practices (Gerrymandering)
- Some US states have a history of denying equal voting to some minority groups (e.g., African American)
- Recent approaches involve "packing" minorities into a small number of districts, thereby minimizing their overall representation

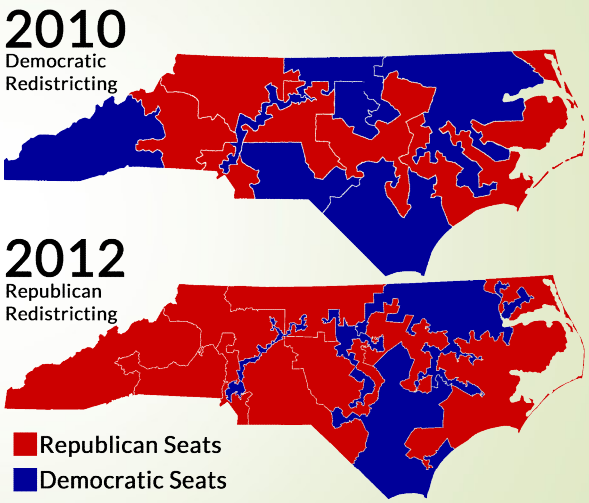
"I propose that we draw the maps to give a partisan advantage to 10 Republicans and three Democrats because I do not believe it's possible to draw a map with 11 Republicans and two Democrats." – Chairman of NC House redistricting committee

9

What is a Gerrymander?

- Refers to a voting district that resembles a salamander
- Named after Elbridge Gerry, 5th VP of US





10

Why is Gerrymandering a Hot Topic?

- Gerrymandering is a practice intended to establish an advantage for a particular party or group by manipulating district boundaries
- Usually features "packing" (packing maximum number of opposing party voters in to a handful of districts) and "cracking" (split opposing party voters into many districts)
- Occurring in the US since 1812
- Used aggressively in 2010, resulting in congressional dysfunction

Definition from Wikipedia

11

Consequences of Current Gerrymandering

- Many congressional seats are not competitive
- Members of congress are more concerned with a primary battle than an election battle
- Republicans and Democrats represent their party's position more than the wishes of their constituents
- Extremes of each party dominate, instead of the middle

Congressional Gridlock

12

Districting Techniques

- Approaches
  - Optimal districting – generate the graph partitioning (i.e., Congressional districts) that optimize an objective function (search in N-Dimensional space for an optimal solution)
  - Evaluate an enacted district plan to determine if it is "fair" (compare a districting plan with random samples taken from all possible districting plans)
- Note: Historically unfair districtings
  - Underrepresent the voting importance of minority groups (e.g., African-Americans)
  - Underrepresent the voters of a party not responsible for redistricting

Typical fairness concerns are racial fairness and political fairness

13

Algorithmic Formulation

- Both Congressional districting approaches are formulated as a graph partitioning problem
  - Each node in the graph is a voting district (or sometimes a census block)
  - Nodes are connected (i.e., share an edge) if the corresponding districts are contiguous
  - Graph (around 10,000 nodes) is partitioned into n sub-graphs (usually 5-50) such that each partition satisfies constraints (e.g., connected) and optimizes measures of quality (e.g., compactness)
- Optimization problem is NP-hard, and must be solved through computational search
- Random districting approach requires a large number of iterations to ensure randomness

Some algorithms use census blocks as nodes

14

Optimal Districting - Typical Approach

- Steps
  - Start with an initial solution referred to as a seeding (e.g., existing Congressional districts)
  - Move nodes in the graph (i.e., voting districts) to reconfigure the graph partition in a way to optimize an objective function
  - Stop when objective function does not improve
- Issues
  - Solution space not well understood
  - May not find the global optimum
  - Some measures not legally defined

15

Probabilistic Districting Analysis

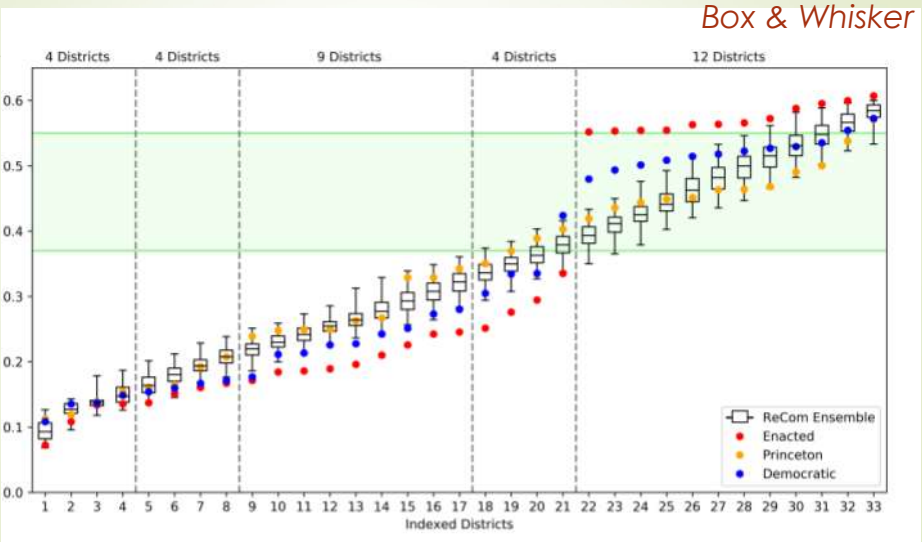
- MCMC (Markov Chain, Monte Carlo)
- Steps
  - Start with a seed districting
  - Randomly move nodes (or groups of notes) until the resulting districting is independent of the seed
  - Repeat the above step until a sufficient number (usually 1,000-10,000) of random districtings are generated
  - Rearrange the districts in each generated districting in the order of a parameter to be tested (e.g., African-American population)
  - Compare enacted districting to generated probability distribution

16

Enacted District Evaluation – Typical Approach

Box & Whisker Plot

Random districts are compared with enacted districts



17

Spring 2021 Project

- Analyze and compare the districtings generated in a MCMC approach, based on user requirements
- Combines the best of both approaches (optimal districting and enacted districting analysis) to
  - Calculate the objective function for each generated districting – better understand the search space
  - Do parametric analysis of terms in the objective function

Computation of this analysis requires an interesting combination of programming (multi-processors), computer science (e.g., graph algorithms), and applied math

18

High Level View of the Project

- Build a robust system to
  - Display a current districting plan for a state
  - Generate "random" sets of districting jobs for multiple states on a high-performance computer
  - Store each such districting job in a DB for analysis
  - Calculate the objective function (including all the relevant measures for each districting)
  - Subset the random districtings, if needed (e.g., population equality threshold, compactness, VRA compliance)
  - Compare and analyze measures
  - Display concise and detailed results of the measures
  - Display interesting data combinations to the user

19

Analysis of Random Districtings

- Examples
  - Determine districting with the best scores of the objective function
  - Compare geometric similarity of selected districtings (e.g., among 10 best objective function scores)
  - Fix one or more measures in the objective function and compare the sensitivity for other measures
  - Compare aggregate results for multiple districting jobs

20

System Display

- Any generated districting might be displayed
- Districts and precinct boundaries are displayed
- Setup of a run defines the constraints in districting (e.g., compactness)
- The characteristics of a large number of random districting plans are displayed

21

Project Requirements Analysis

- You will generate detailed requirements (use cases)
- Requirements will evolve over the first 6 weeks of the project
  - Top-Level functional requirements provided in first 3 weeks of class sessions
  - You will develop detailed requirements based on top-level requirements
  - Requirements aggregated into a master use-case list
  - Near-final set of use cases by early March

22

What Data is Needed?

- Geospatial boundary data
  - Precincts
  - Existing Congressional districts
  - Cities/counties
  - Census tracts (including demographic data)
- Election results data
- Population data
- And more

Multiple data sources can be used to measure the party preference of a precinct (congressional vote, presidential vote, registration, etc.)

23

State Selection

- Your system will run for 3 states of your choosing
- Maximum of 4 teams per state
- You can change any state during the semester as long as it is available
- Once the table is posted, send me an e-mail with your team's choices. Be sure to include alternate selections in case your first choice is already filled

State table posted soon

State	Team 1	Team 2	Team 3	Team 4
Alabama ^				
Arizona ^				
Arkansas + #				
California + * ^				
Colorado ^				
Florida + * ^				
Georgia * ^				
Illinois * ^				
Kentucky + #				
Louisiana # ^				
Maryland # ^				
Michigan *				
Mississippi ^				
New Jersey ^				
New York *				

24

Sources of Data

- Project Web site suggests many sources of data
- The most accurate data source is the originator
  - US Census Bureau
  - State Election Office
  - US Government repository of region borders

Harvard might be the best source now

- Easier sources of data (including some consolidation) are available
- Sometimes difficult to locate the best source of data

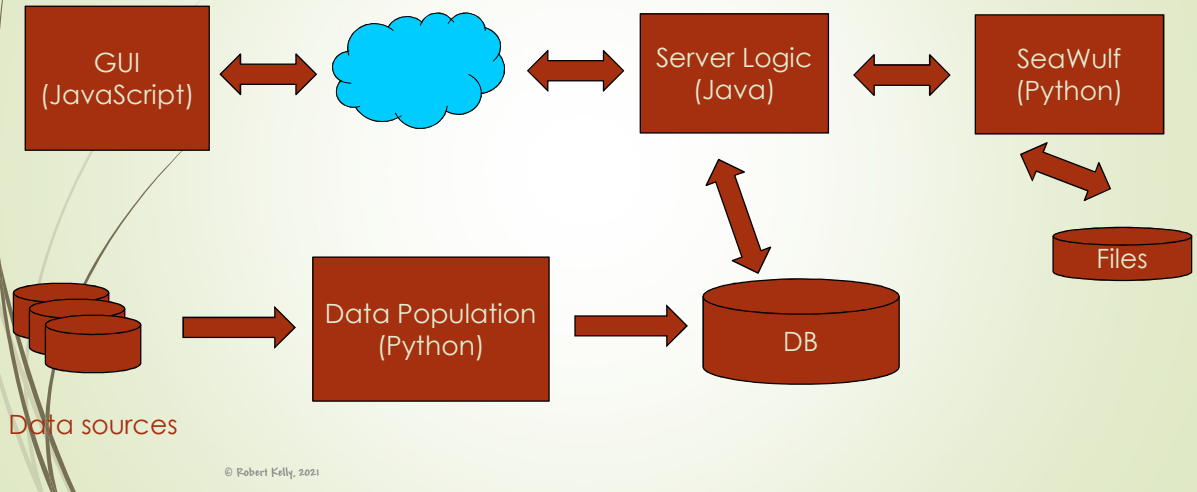
25

Algorithms

- Research papers cited for
  - measures and
  - graph partitioning
- MCMC code (Python) available to you at the MGGG code repository
- Algorithmic change requirements will be discussed in class

26

Top-Level System Architecture



27

What Skills Do You Need?

- Programming (Java, JavaScript, Python)
- Client/server interaction (e.g., Spring, JAX-RS)
- Graph algorithms (e.g., spanning tree)
- Thread programming
- Data serialization (migration of data –client/server/SeaWulf)
- Performance analysis (parallel speedup measurement)
- Map system integration
- Client data display (e.g., box and whisker plot)
- Client framework (e.g., React)
- DB
- And more

TAs were selected based on experience with these technologies

Free SW libraries are available for everything you need

Almost impossible for a team to have all these skills

28

Have You Satisfied the Objectives?

- Understand issues and terminology used in US congressional redistricting and voting analysis
- Understand some data requirements to support analysis of redistricting
- Understand relationship between redistricting and graph partitioning