An Extreme-Scale Computational Approach to Redistricting Optimization

Wendy K. Tam Cho, Yan Liu, and Shaowen Wang (PI)

CyberGIS Center for Advanced Digital and Spatial Studies (CyberGIS Center)
CyberInfrastructure and Geospatial Information Laboratory (CIGI Laboratory)
Department of Geography and Geographic Information Science
Department of Computer Science
Department of Urban and Regional Planning
Graduate School of Library and Information Science
National Center for Supercomputing Applications (NCSA)
University of Illinois at Urbana-Champaign

NCSA Blue Waters Symposium
for Petascale Science and Beyond
May 12-15, 2014
Gerrymandering
Zoning Analytics

• Partitioning a group of indivisible geographic units into a smaller number of districts
  
  o Objectives and constraints
    ▪ Contiguity, competitiveness, equal population, preservation of communities of interest and local political subdivisions, minority districts
  
  o Computational complexity
    ▪ Number of possible solutions
      • Stirling number of the second kind: $S(n, k)$
      • Example: $S(55, 6) = 8.7 \times 10^{39}$
    ▪ Computationally intractable
      • NP-hard
Collaborative Work by Hao Hu, Tao Lin, Yan Liu, Luis F. Rodriguez, and Shaowen Wang
Exact Algorithms vs. Heuristics

- **Exact algorithms**
  - Guarantee to find an optimal solution
  - Methods
    - Branch & bound
    - Branch & cut
    - Etc.
  - Computationally intractable

- **Heuristics**
  - Algorithms that produce optimal or near-optimal solutions within a reasonable amount of time
  - Population-based heuristics
    - *Genetic algorithm*
    - Swarm
    - Ant colony
    - Etc.
Genetic Algorithm (GA)

• **Principles**
  - Evolutionary process
    - “survival of the fittest”
    - Iterative algorithm
  - Solution population: a diverse set of initial solutions
  - GA operators
    - Selection, crossover, mutation, replacement
  - Stopping criteria
    - Solution quality
    - Time or the number of iterations

• **Spatial GA operators**
  - Solution generation
  - Crossover
  - Mutation
Spatial GA Operators – Feasible Solution Generation

Seeding

Expansion

Completion
Parallel GA

1  2  3  4
5  6  7  8
9 10 11 12
13 14 15 16

e Export
i Export
Challenges for PGA

- Scalability to a large number of processor cores
  - Solution migration
    - Migration interval, rate
    - Traditional implementation: global barrier
  - Communication cost increases significantly when using a large number of cores

- Synchronous vs. Asynchronous
  - Goals of asynchronous migration
    - Mimic the natural behavior of GA
    - Increase the overlapping of computing and communication
  - Breaking the global barrier
    - Asynchronous migration
    - Buffer handling is inevitable
Asynchronous Migration

- **Operators**
  - export, import, inject
## PGA Parameters

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population size per deme</td>
<td>100</td>
</tr>
<tr>
<td>Initial population generation</td>
<td>Random with feasibility improvement or constraint-based improvement</td>
</tr>
<tr>
<td>Selection</td>
<td>Binary tournament</td>
</tr>
<tr>
<td>Crossover</td>
<td>1-point. Probability: 0.8</td>
</tr>
<tr>
<td>Mutation</td>
<td>1-item mutation. Probability: 0.2</td>
</tr>
<tr>
<td>Replacement</td>
<td>Replacing the unfittest or worst</td>
</tr>
<tr>
<td>Elitism</td>
<td>Yes</td>
</tr>
<tr>
<td>Stopping rules</td>
<td>No solution improvement, bounded solution quality reached, or fixed number of iterations</td>
</tr>
<tr>
<td>Connectivity $d$</td>
<td>4</td>
</tr>
<tr>
<td>Migration rate $r$</td>
<td>2</td>
</tr>
<tr>
<td>Export interval $M_{expt}$</td>
<td>100</td>
</tr>
<tr>
<td>Import interval $M_{impt}$</td>
<td>50</td>
</tr>
<tr>
<td>Probability of holding</td>
<td>1/20 (the probability to export when no better solution found during a previous export interval)</td>
</tr>
<tr>
<td>Sending buffer size $K_{sendbuf}$</td>
<td>20 solutions. Actual memory requirement is $(20 \times n \times 4 + buffer_overhead)$ bytes</td>
</tr>
<tr>
<td>Import pool size $K_{impt}$</td>
<td>80 solutions. Actual memory requirement is $(80 \times n \times 4)$</td>
</tr>
</tbody>
</table>
Case Study
Computational Experiments

• Problem size
  o 4140 voting districts
  o Number of districts: 8

• Data
  o Input: shape file; rook and queen neighborhood files
  o Output: shape file; each core outputs one

• Number of cores
  o 32768, 65536, 131072
• Measurement: time taken to achieve different solution bounds (the smaller, the better)
• Using 131072 cores led to significant improvement
An Example Solution

- Found by the run using 131072 cores
- Competitiveness as objective
  - $R / (R + D)$
  - 0.499998
- Constraint: population deviation
Solution Quality Improvement

![Bar chart showing fitness value improvement percentage for different numbers of processors.]

<table>
<thead>
<tr>
<th>Number of processors</th>
<th>Fitness value improvement (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2048</td>
<td>0.60%</td>
</tr>
<tr>
<td>4096</td>
<td>0.64%</td>
</tr>
<tr>
<td>8192</td>
<td>0.84%</td>
</tr>
<tr>
<td>16384</td>
<td>1.09%</td>
</tr>
</tbody>
</table>
Number of New Solutions

Data: Minnesota VTD (4140 units)

Number of new solutions

Number of cores

- 32768
- 65536
- 131072
What is GIS?

- Systems
- Science
- Services
- Society
- Synthesis
- Geo and spatial are special

Image source:
CyberGIS for What and Whom?

CyberGIS Toolkit

GISolve

CyberGIS Gateway
NSF SI2-SSI: CyberGIS Project
$4.43 million, Year: 2010-2015

**Principal Investigator**
- Shaowen Wang

**Co-Principal Investigators**
- Luc Anselin
- Budhendra Bhaduri
- Timothy Nyerges
- Nancy Wilkins-Diehr

**Senior Personnel**
- Michael Goodchild
- Sergio Rey
- Marc Snir
- E. Lynn Usery

**Chair of the Science Advisory Committee**
- Michael Goodchild

**Project Manager**
- Anand Padmanabhan

**Project Staff**
- ASU: Wenwen Li and Rob Pahle
- ORNL: Ranga Raju Vatsavai
- SDSC: Choonhan Youn
- UIUC: Yan Liu and Anand Padmanabhan
- USGS: Michael Finn and David Mattli
- Graduate and undergraduate students

**Industrial Partner: Esri**
- Steve Kopp and Dawn Wright
CyberGIS Communities

• **Science and Technology Communities**
  - Advanced cyberinfrastructure
  - Climate change impact assessment
  - Emergency management
  - Geographic information science
  - Geography and spatial sciences
  - Geosciences
  - Social sciences
  - Etc.

• **User Communities**
  - Biologists
  - Geographers
  - Geoscientists
  - Social scientists
  - General public
  - Broad GIS users
  - Etc.
Image source: http://blogs.esri.com/esri/arcgis/2013/10/01/what-is-cybergis/
CyberGIS Toolkit

CyberGIS Toolkit is a suite of loosely coupled open-source geospatial software components that provide computationally scalable spatial analysis and modeling capabilities enabled by advanced cyberinfrastructure. CyberGIS Toolkit represents a deep approach to CyberGIS software integration research and development and is one of the three key pillars of the CyberGIS software environment, along with CyberGIS Gateway and GISolve middleware. The integration approach to building CyberGIS Toolkit is focused on developing and leveraging innovative computational strategies needed to solve computing- and data-intensive geospatial problems by exploiting high-end cyberinfrastructure resources such as supercomputing resources provided by the NSF Extreme Science and Engineering Discovery Environment (XSEDE) and high-throughput computing resources on the Open Science Grid (OSG).

A rigorous process of software engineering and computational intensity analysis is applied to integrate an identified software component into the toolkit, including software building, testing, packaging, scalability and performance analysis, and deployment. This process includes three major steps:

1. Local build and test by software contributors and developers using continuous integration software or services such as Travis CI;
2. Continuous integration testing, portability testing, small-scale scalability testing on the National Middleware Initiative (NDI) build and test facility; and
3. XSEDE-based evaluation and testing of software performance, scalability, and portability. By leveraging the high-performance computing expertise in the integration team of the NSF CyberGIS Project, large-scale problem-solving tests are conducted on various supercomputing environments on XSEDE to identify potential computational bottlenecks and achieve maximum problem-solving capabilities of each software installation.

Initial Software Components
- Curriculum and pedagogy
- Partnerships
- Open
Acknowledgments

- National Science Foundation
  - Blue Waters
  - BCS-0846655
  - OCI-1047916
  - XSEDE SES070004
Thanks!

- Comments / Questions?
- Email: shaowen@illinois.edu