Èdouard Lucas:

The theory of recurrent sequences is an inexhaustible mine which contains all the properties of numbers; by calculating the successive terms of such sequences, decomposing them into their prime factors and seeking out by experimentation the laws of appearance and reproduction of the prime numbers, one can advance in a systematic manner the study of the properties of numbers and their application to all branches of mathematics.

Preview

Geospatial Data, Markov Chains, and Political Redistricting

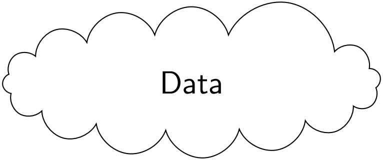
Daryl DeFord

MIT – CSAIL
Geometric Data Processing Group

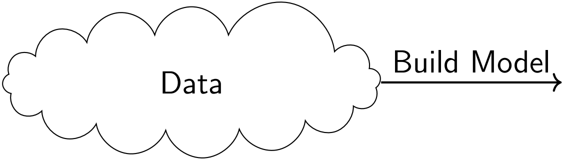
Department of Mathematics and Statistics
Washington State University
Pullman WA
January 21, 2020

Outline

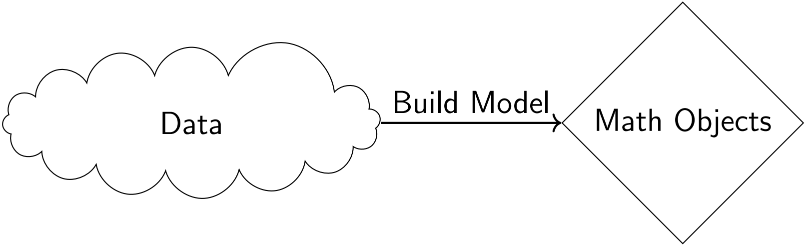
Philosophy



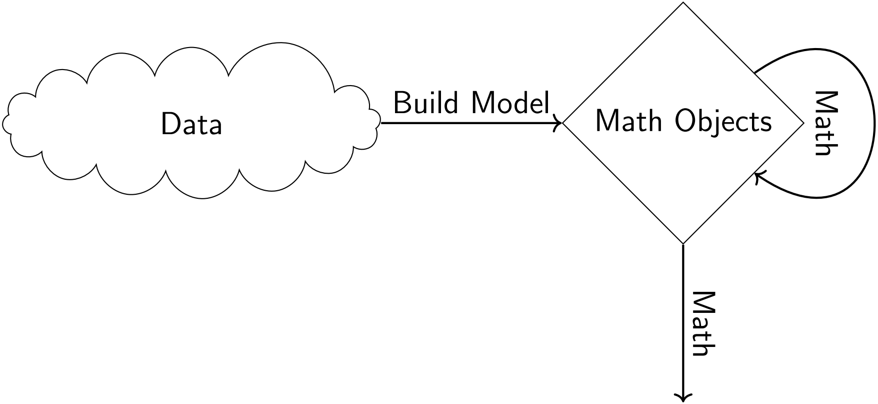
Philosophy



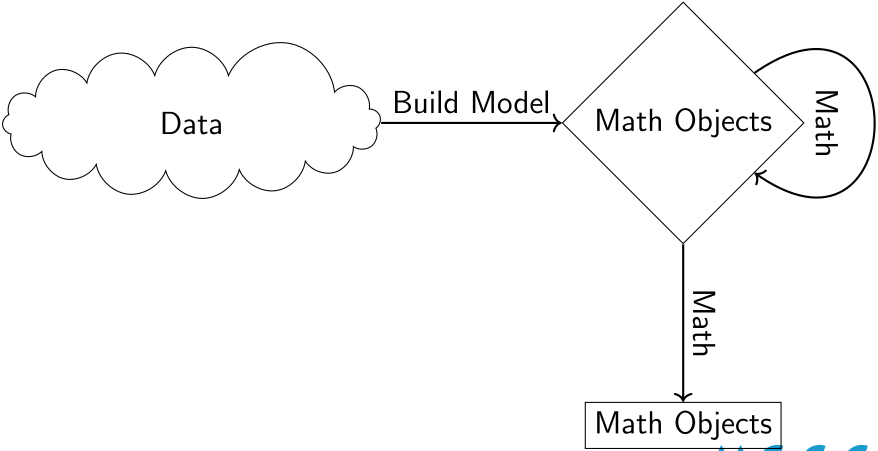
Philosophy



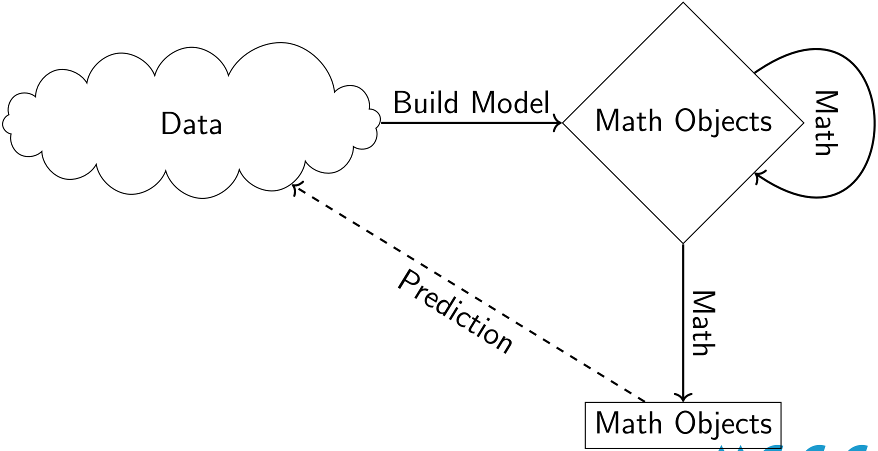
Philosophy



Philosophy



Philosophy


Data
Build Model
Math Objects
Math
Math
Prediction
Math Objects

MORAL #1:

MORAL #1:

Computational Redistricting is NOT a solved problem!

☹

MORAL #2:

MORAL #2:

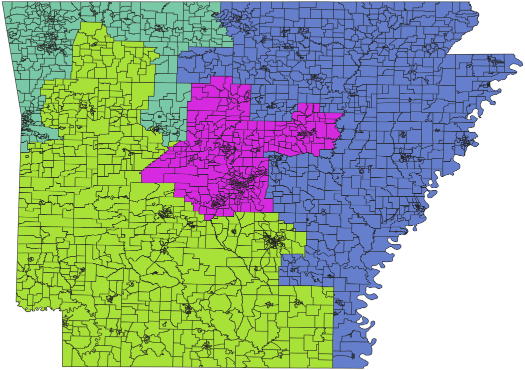
Computational Redistricting is NOT a solved problem!

MORAL #2:

Computational Redistricting is NOT a solved problem!

☺

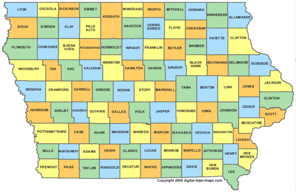
What is a district?

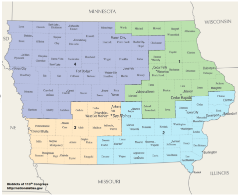


Permissible Districting Plans

- Contiguity
- Population Balance
- Compactness
- Communities of Interest
- Municipal Boundaries
- Competitiveness/Responsiveness/Symmetry
- Incumbency Protection
- ...

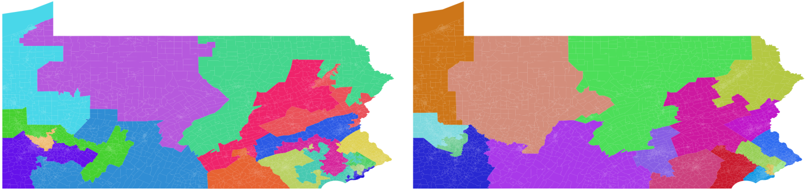
Example: Iowa





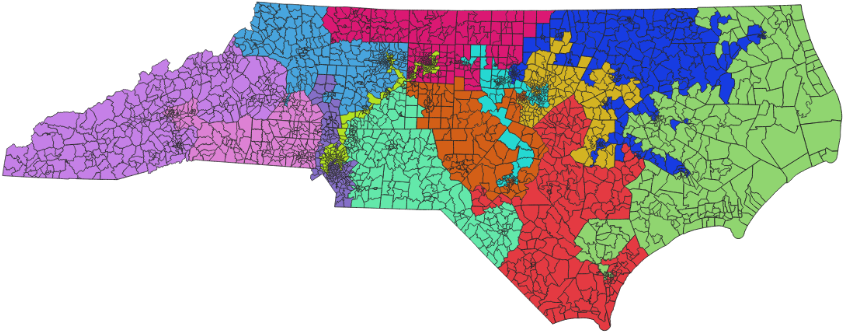
- 4 Congressional Districts, 100 House Districts, 50 Senate Districts
- House districts nest into Senate districts
- Congressional districts made out of counties
- Independent committee with legislative approval
- No partisan data allowed

Example: Pennsylvania



- 18 Congressional Districts, 203 House Districts, 50 Senate Districts
- Zero–balanced population
- Legislature draws congressional districts - subcommittee draws legislative districts
- Partisan considerations allowed

Ugly Shapes



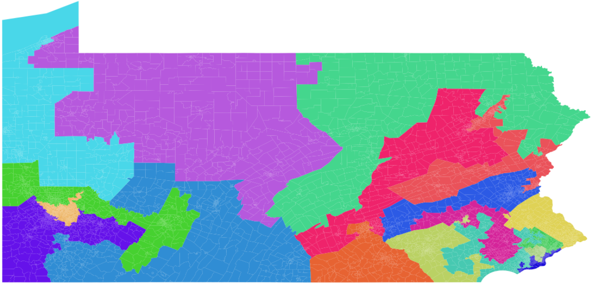
Ugly Shapes

NC12 #1

NC12 #2

NC12 #12

Ugly Shapes



Measurement Problems

Theorem (Bar-Natan, Najt, and Schutzman 2019[1])

There is no local homeomorphism from the globe to the plane that preserves your favorite compactness measure.

Problem (Barnes and Solomon 2018[2])

Geographic Compactness scores can be distorted by:

- *Data resolution*
- *Map projection*
- *State borders and coastline*
- *Topography*
- *...*

[1] The Gerrymandering Jumble: Map Projections Permute Districts' Compactness Scores, arXiv:1905.03173

[2] Gerrymandering and Compactness: Implementation Flexibility and Abuse, Political Analysis, to appear 2020.

Multiresolution Measures

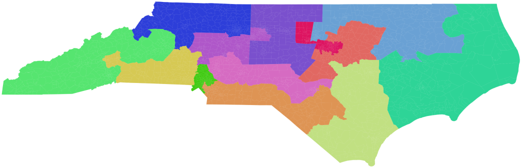
Theorem (D., Lavenant, Schutzman, and Solomon 2019[1])

The total variation relaxation of the isoperimetric profile:

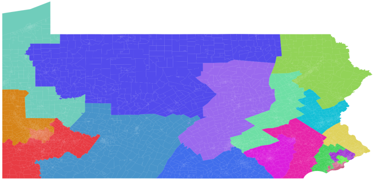
1. *satisfies an isoperimetric inequality,*
2. *is the lower convex envelope of the original profile,*
3. *and admits a distinguished family of efficiently computable solutions that recover the Cheeger set and take at most three values for every t.*

[1] Total Variation Isoperimetric Profiles, SIAM J. Appl. Algebra Geometry, 3(4), 585-613, 2019.

Partisan Imbalance

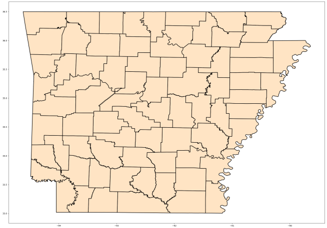


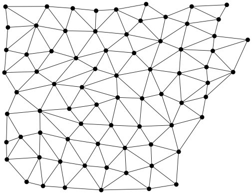
NC16



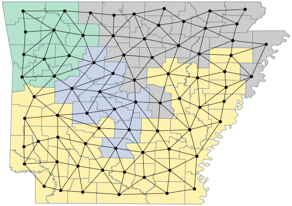
PA TS-Proposed

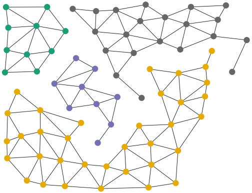
Discrete Partitioning





Discrete Partitioning





Mathematical Formulation

Given a (connected, planar) graph $G = (V, E)$:

- A k-**partition** $P = \{V_1, V_2, \ldots, V_k\}$ of G is a collection of disjoint subsets $V_i \subseteq V$ whose union is V. The full set of k–partitions of G will be denoted $\mathcal{P}_k(G)$.
- A partition P is **connected** if the subgraph induced by V_i is connected for all i.
- A partition P is ε-**balanced** if $\mu(1-\varepsilon) \leq |V_i| \leq \mu(1+\varepsilon)$ for all i where μ is the mean of the $|V_i|$'s
- The (context dependent) collection of constraints will be denoted with a function $\mathcal{C}_\theta : \mathcal{P}_k(G) \mapsto \{\texttt{True}, \texttt{False}\}$. The set of permissible partitions will be $\mathcal{C}_\theta(G)$.

Why analyze?

- Court cases
  - Detecting gerrymandering
  - Evaluating proposed remedies
- Reform Efforts
  - Establishing baselines
  - Potential impacts of new rules
- Commissions and plan evaluation
  - Unintentional gerrymandering [1]
  - Full space of plans

[1] with apologies to J. Chen and J. Rodden, *Unintentional Gerrymandering: Political Geography and Electoral Bias in Legislatures*, Quarterly Journal of Political Science, 8, 239–269, 2013.

Abstracted Problem Instances

Problem

Given a fixed G *and metric of interest* $f : \mathcal{P}(G) \mapsto \mathbb{R}^n$.

1. *Given a partition* P*, is it a statistical outlier*[a] *with respect to* f*?*
2. *Given* $\mathcal{C}_\theta$ *and* $\mathcal{C}_{\theta'}$ *how do the distributions* $f\left(\mathcal{C}_\theta(G)\right)$ *and* $f\left(\mathcal{C}_{\theta'}(G)\right)$ *compare?*

---

[a] gerrymander

Abstracted Problem Instances

Problem

Given a fixed G and metric of interest $f : \mathcal{P}(G) \mapsto \mathbb{R}^n$.

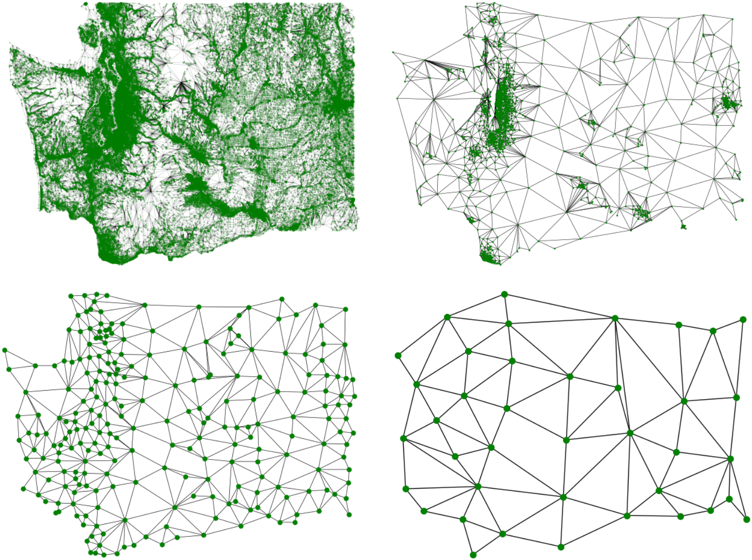
1. *Given a partition P, is it a statistical outlier[a] with respect to f?*
2. *Given $\mathcal{C}_\theta$ and $\mathcal{C}_{\theta'}$ how do the distributions $f\left(\mathcal{C}_\theta(G)\right)$ and $f\left(\mathcal{C}_{\theta'}(G)\right)$ compare?*

[a] gerrymander

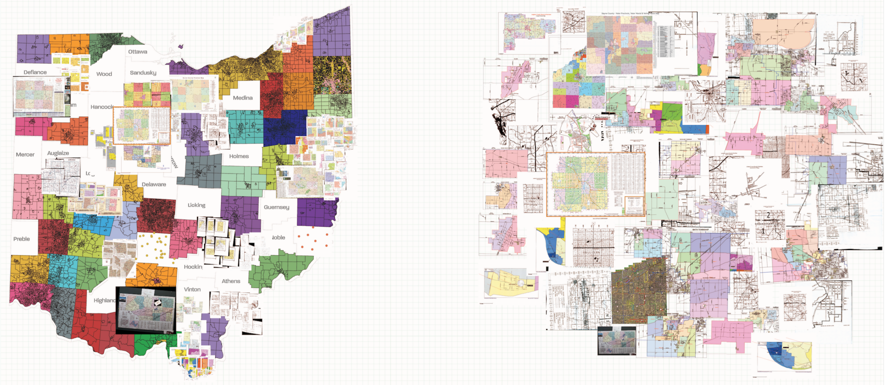
Solution?

Draw (many) samples from (distributions over) $\mathcal{C}_\theta(G)$!

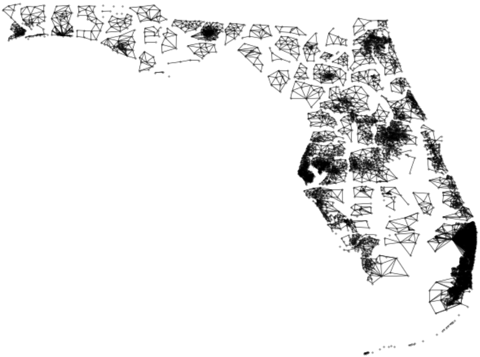
Census Data



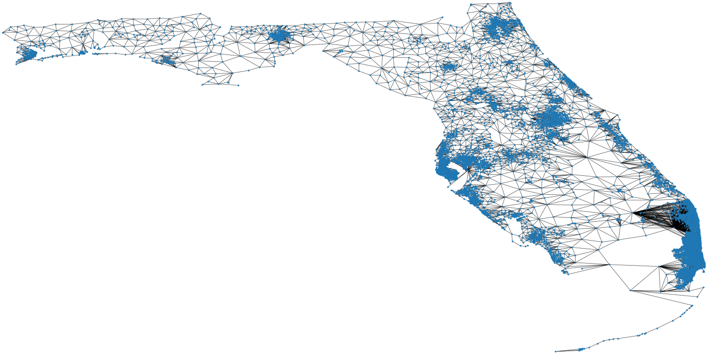
Voting Data (precincts)



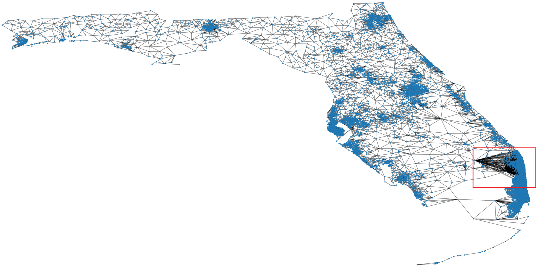
Florida



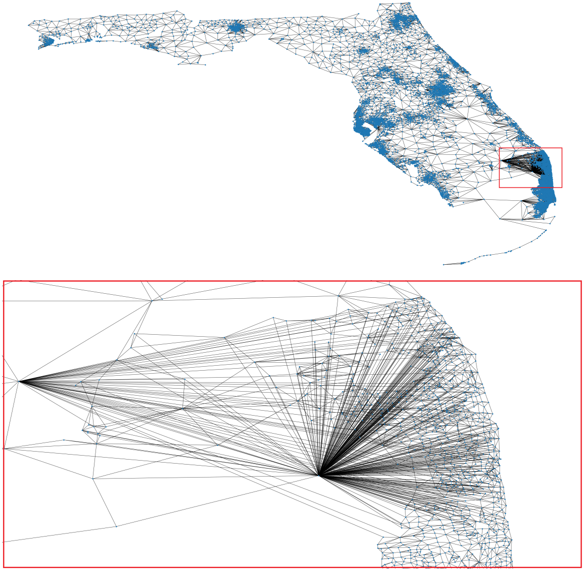
Florida



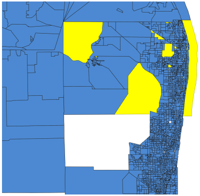
Florida

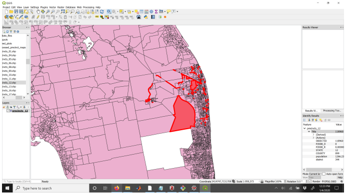


Florida

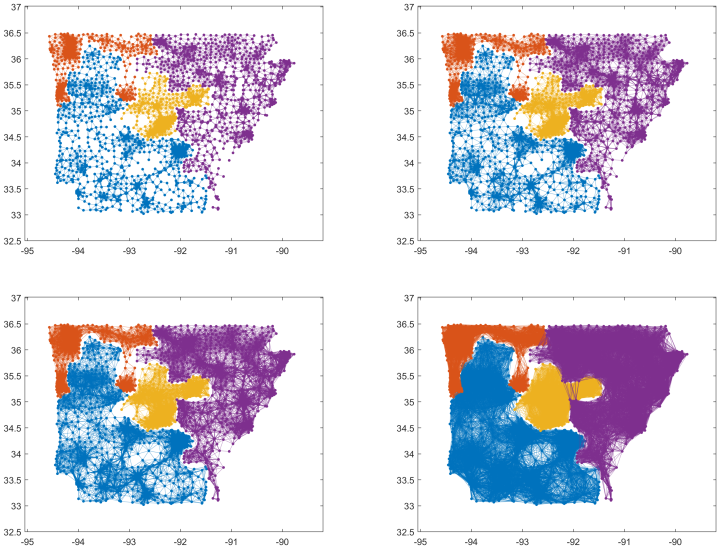


Florida





Topological Data Analysis

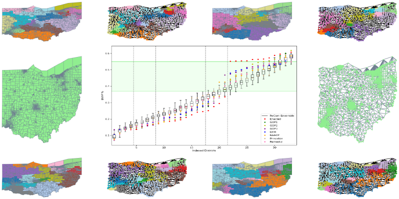


Sample Data Projects:

- Impacts of Modified Areal Unit Problem (MAUP)
- Understand the network structure of census/precinct data
- Efficient database design for boundary changes over time
- Automated tools for district projection, island connections, compactness optimization, etc.

Ensemble Analysis

- The wide variety in rules applied to districting problems (even in the same state) means that any single measure of gerrymandering will be insufficient/exploitable
- Instead we want to compare to large ensembles of other feasible plans.
- This allows us to understand the impacts of the underlying political and demographic geography on a wide collection of metrics.



AR Outlier Example

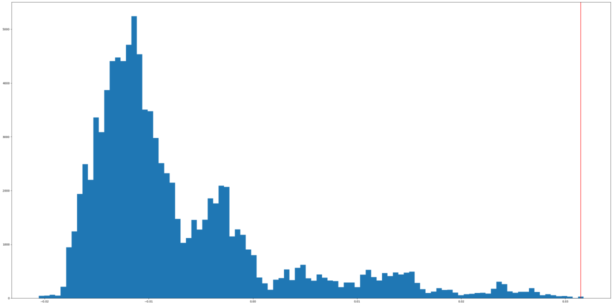


Figure: Mean–Median score using senate 2016 election data on 1,000,000 plans.

Which ensembles?

Ensembles in Practice

- The appeal of an ensemble method is that you get to control the input data very carefully
- However, just because a particular type of data was not considered doesn't mean that the outcome is necessarily "fair"
- There are lots of "random" methods for constructing districting plans
- Most don't offer any control over the distribution that you are drawing from

MCMC for redistricting

1. Start with an initial plan
2. Propose a modification of the current plan
3. Accept using MH criterion
4. Repeat

MCMC for redistricting

1. Start with an initial plan
2. Propose a modification of the current plan
3. Accept using MH criterion
4. Repeat

Why?

MCMC for redistricting

1. Start with an initial plan
2. Propose a modification of the current plan
3. Accept using MH criterion
4. Repeat

Why?

- Control over sampling distribution
- Possibility of local sampling
- Ergodic Theorem

MCMC for redistricting

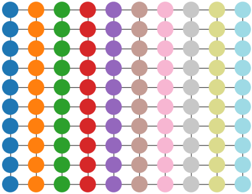
1. Start with an initial plan
2. **Propose a modification of the current plan**
3. Accept using MH criterion
4. Repeat

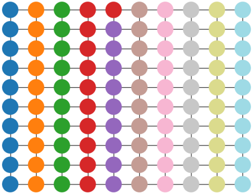
Why?

- Control over sampling distribution
- Possibility of local sampling
- Ergodic Theorem

Single Edge Flip Proposals

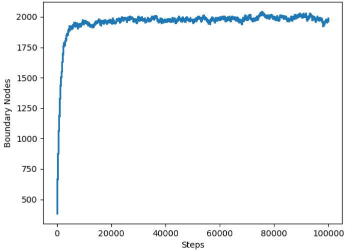
1. Uniformly choose a cut edge
2. Change one of the incident node assignments to the other

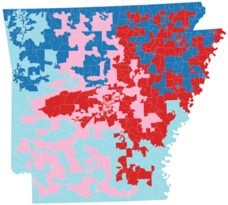




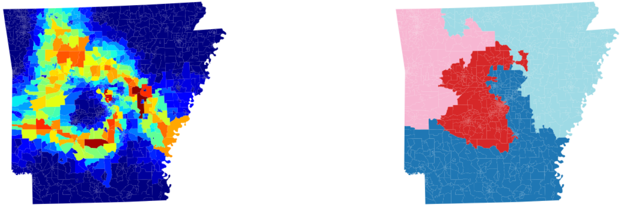
- Mattingly et al. (2017, 2018) Court cases in NC and WI.
- Pegden et al. Assessing significance in a Markov chain without mixing, PNAS, (2017). Court case in PA.

Unconstrained Flip





Constrained Flip



PA Flip Ensemble

Slowly Mixing Graph Families

Theorem (Najt, D., and Solomon 2019[1])

Let G be any connected graph. Then let $G^{(d)}$ be the graph obtained by replacing each edge by a doubled d-star. Then the flip walk on partitions of family of graphs $G^{(d)}_{d\geq 1}$ is slowly mixing, in the sense the Cheeger constant is decaying exponentially fast. More specifically:

$$H(Partition\ Graph(G^{(d)}) = O(2^{-d})$$

[1] Complexity and Geometry of Sampling Connected Graph Partitions, arXiv:1908.08881, (2019).

Slowly Mixing Graph Families

Theorem (Najt, D., and Solomon 2019[1])

Let G be any connected graph. Then let $G^{(d)}$ be the graph obtained by replacing each edge by a doubled d-star. Then the flip walk on partitions of family of graphs $G^{(d)}_{d \geq 1}$ is slowly mixing, in the sense the Cheeger constant is decaying exponentially fast. More specifically:

$$H(Partition\ Graph(G^{(d)}) = O(2^{-d})$$

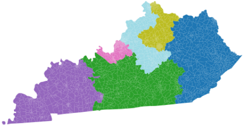
Remark

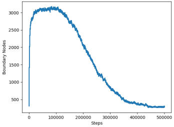
There are many similar constructions that give rise to equivalent mixing results.

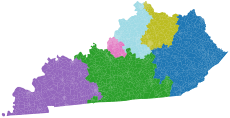
[1] Complexity and Geometry of Sampling Connected Graph Partitions, arXiv:1908.08881, (2019).

Slow Mixing Example

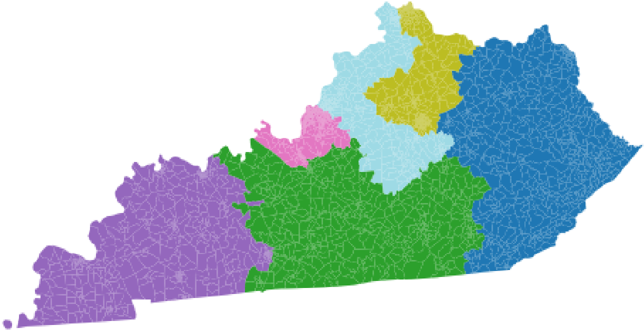
Annealing



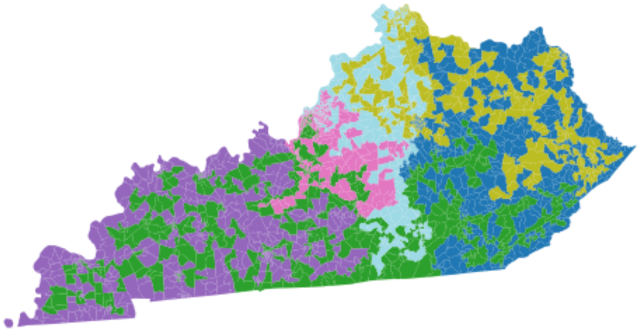

3000
2500
2000
1500
1000
500
Boundary Nodes
0
100000
200000
300000
400000
500000
Steps



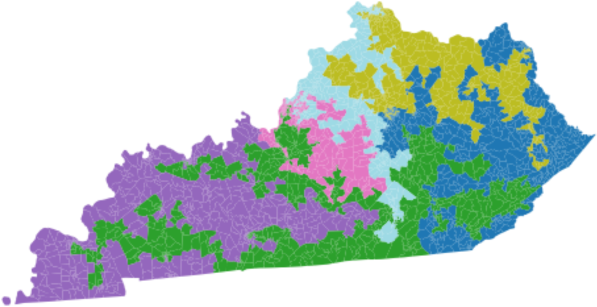
Starting Partition



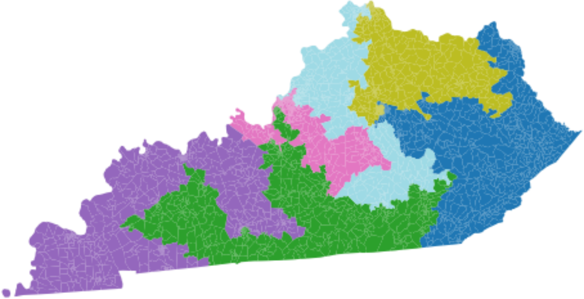
Generic Partition



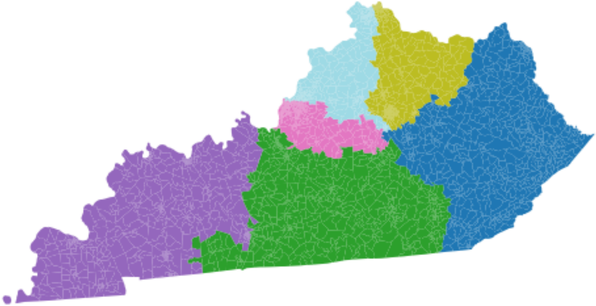
Annealing



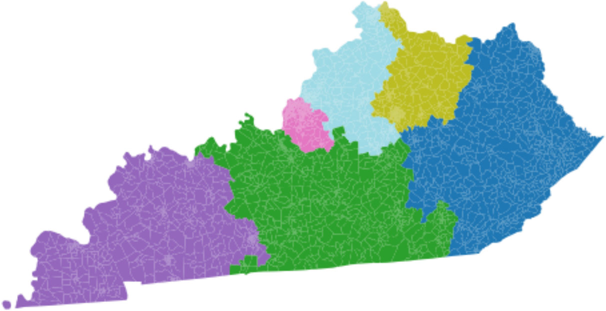
Annealing



Annealing



Annealing



Uniform Sampling of Contiguous Partitions

Theorem (Najt, D., and Solomon 2019)

Suppose that $\mathscr{C}$ *is the class of connected planar graphs and* $k \geq 2$. *If there is a polynomial time algorithm to sample uniformly from:*

- *the connected k-partitions of graphs in* $\mathscr{C}$,
- *or the connected, 0-balanced k-partitions of graphs in* $\mathscr{C}$.

then $RP = NP$.

Uniform Sampling of Contiguous Partitions

Theorem (Najt, D., and Solomon 2019)

Suppose that $\mathscr{C}$ is the class of connected planar graphs and $k \geq 2$. If there is a polynomial time algorithm to sample uniformly from:

- *the connected k-partitions of graphs in $\mathscr{C}$,*
- *or the connected, 0-balanced k-partitions of graphs in $\mathscr{C}$.*

then $RP = NP$.

Remark

This theorem has various interesting extensions, including:

- *Connectivity constraints on $\mathscr{C}$*
- *Degree bounds*
- *Distributions proportional to cut length*
- *TV distribution approximation*

Stronger Version Example

Theorem (D., Najt, and Solomon 2019)

Let $\mathscr{C}$ be the class of cubic, planar 3-connected graphs, with face degree bounded by $C = 60$. Let $\mu_x(G)$ be the probability measure on $P_k(G)$ such that a partition P is drawn with probability proportional to $x^{cut(P)}$. Fix some $x > 1/\sqrt{2}$, $\epsilon > 0$ and $\alpha < 1$. Suppose that there was an algorithm to sample from $P_2^{\epsilon}(G)$ according to a distribution $\nu(G)$, such that $||\nu_G - \mu_x(G)||_{TV} < \alpha$, which runs polynomial time on all $G \in \mathscr{C}$. Then $RP = NP$.

Proof Outline Sketch

Following technique of Jerrum, Valiant, and Vazirani[1].

1. Show that uniformly sampling simple cycles is hard on some class $\mathscr{C}$
    1. Choose a gadget that respects $\mathscr{C}$ and allows us to concentrate probability on long cycles
    2. Count the proportion of cycles as a function of length
    3. Reduce to Hamiltonian path on the graph class
2. Show closure of class under planar dual
3. Identify partitions with cut edges $\mapsto$ simple cycles (via planar duality)
4. Conclude that sampling partitions would allow you to sample from cycles which would allow you to find Hamiltonian cycles

□

[1] M. Jerrum, L. Valiant, and V. Vazirani, Random generation of combinatorial structures from a uniform distribution, Theoretical Computer Science, 43 (1986), 169–188.

Proof Sketch – Planar 2–Partitions

Still following technique of Jerrum, Valiant, and Vazirani.

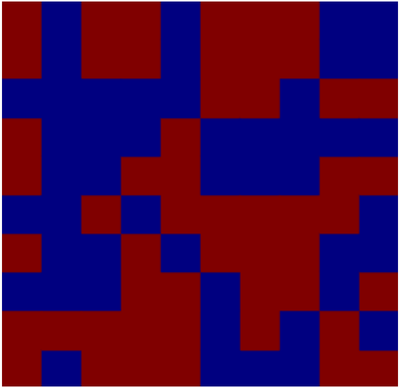
1. Let $\mathscr{C}$ be the planar connected graphs
   1. Replace the edges with chains of dipoles
   2. Hamiltonian hardness for $\mathscr{C}$ given by [1]
2. $\mathscr{C}$ closed under planar duals
3. Identify partitions with cut edges (via planar duality)

□

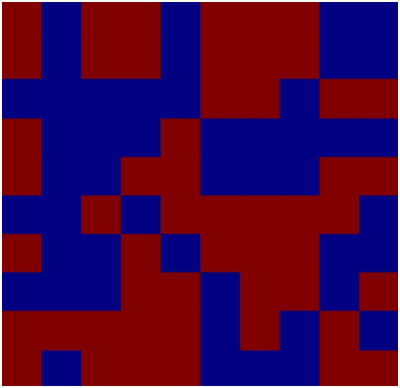
[1] M. Garey, D. Johnson, and R. Tarjan, The Planar Hamiltonian Circuit Problem is NP-Complete, SIAM Journal on Computing, 5, (1976), 704–714.

Other Partition Sampling Frameworks

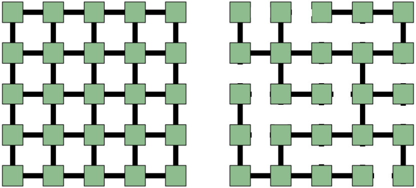
Other Partition Sampling Frameworks



Other Partition Sampling Frameworks

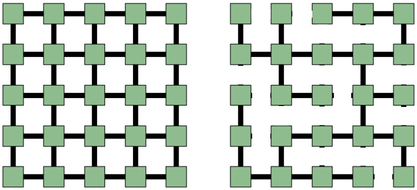


New Proposal: Spanning Trees



[1] ReCombination: A family of Markov chains for redistricting, with M. Duchin and J. Solomon, arXiv:1911.05725, (2019).

New Proposal: Spanning Trees

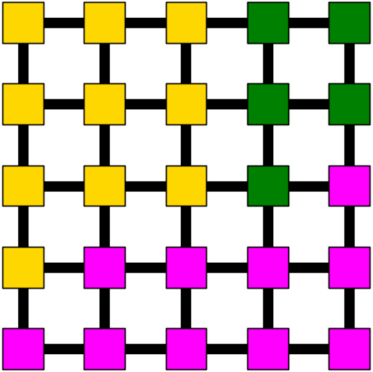


ReCombination [1]

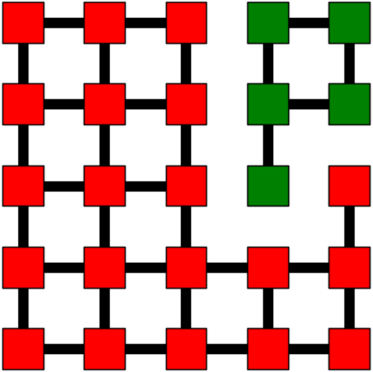
1. At each step, select two adjacent **districts**
2. Merge the subunits of those two districts
3. Draw a spanning tree for the new super–district
4. Delete an edge leaving two population balanced districts

[1] ReCombination: A family of Markov chains for redistricting, with M. Duchin and J. Solomon, arXiv:1911.05725, (2019).

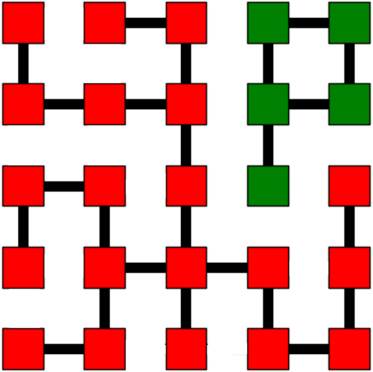
ReCombination Example



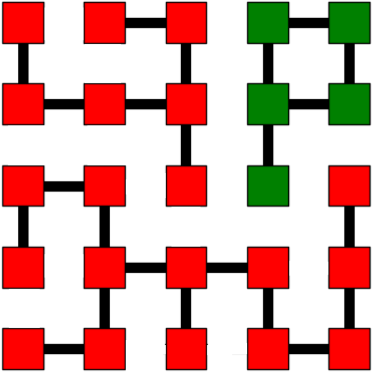
ReCombination Example



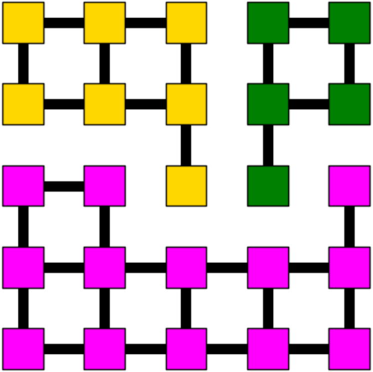
ReCombination Example



ReCombination Example



ReCombination Example



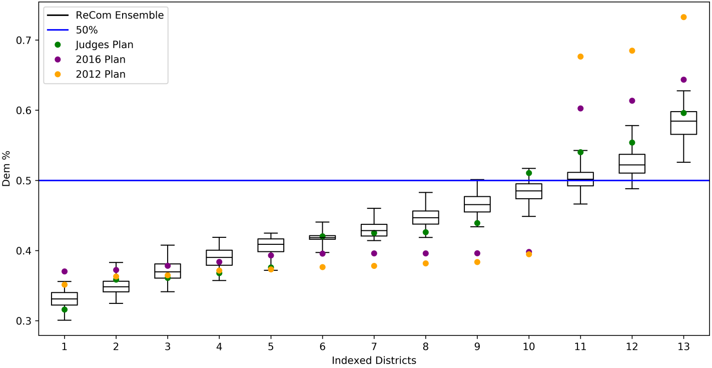
Tree Ensembles

PA Recombination Steps

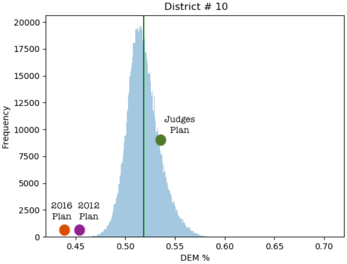
Sampling Future Work:

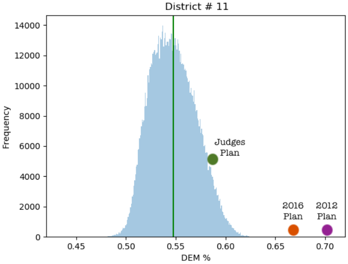
- Characterize steady state distributions
- Spanning tree analysis on grids
- Introduce new proposal methods
- Theoretical basis for mixing in node-aggregate marginals
- Applications of other bipartitioning methods
- Computations in expanded state spaces and lifted walks
- Metamandering analyses
- Applications to statistical physics and combinatorics

Outlier Example: NC

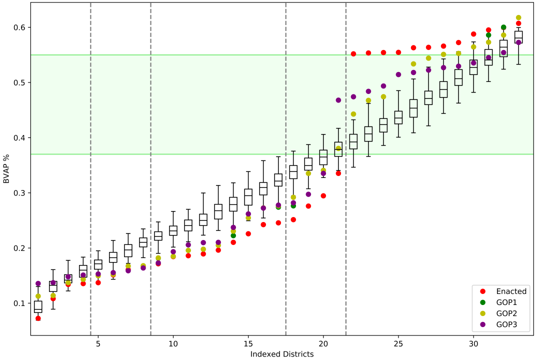


Outlier Example: NC

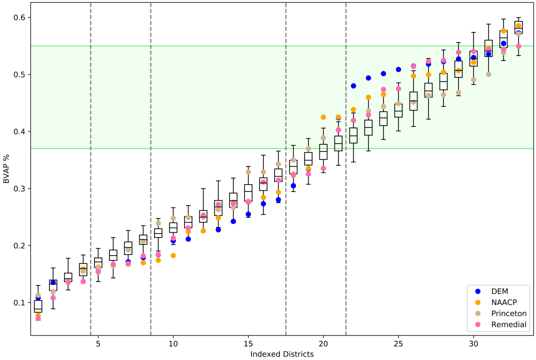




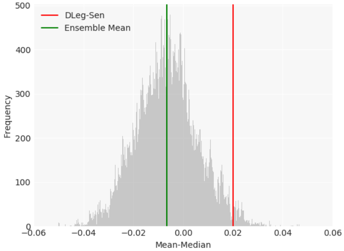
Outlier Example: VA



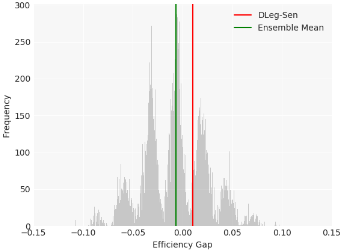
Outlier Example: VA



Baseline Example: VA

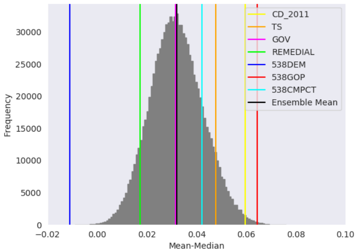


Mean–Median

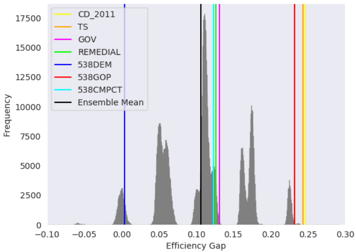


Efficiency Gap

Baseline Example: PA

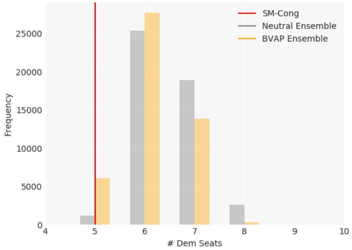


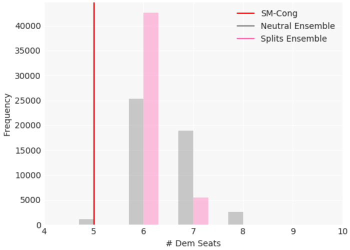
Mean–Median



Efficiency Gap

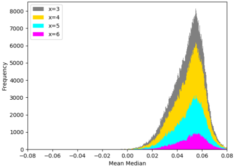
Reform Example: VA



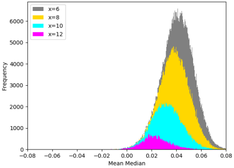


Redistricting Reform in Virginia: Districting Criteria in Context, with M. Duchin, Virginia Policy Review, 12(2), 120-146, (2019).

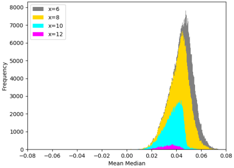
Reform Example: Competitiveness



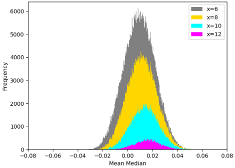
UT



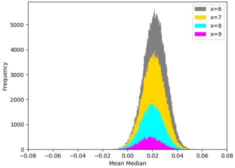
GA



WI



VA



MA

A Computational Approach to Measuring Vote Elasticity and Competitiveness, with M. Duchin and J. Solomon, mggg.org/competitiveness, (2019).

Future Applications:

- Baselines for entire country
- Coalition ecological inference
- Multi-member districts
- Ranked choice voting
- 2021 redistricting
- City council races
- New ballot measures
- District nesting

The End

Thanks!

Try it at home!

- Draw your own districts with **Districtr**
  - `https://districtr.org`
  - Easy to generate complete districting plans in browser or on a tablet
  - Measures district demographics and expected partisan performance
  - Identifies communities of interest
- Generate your own ensembles with **GerryChain**
  - `https://github.com/mggg/gerrychain`
  - Flexible, modular software for sampling graph partitions
  - Handles the geodata processing as well as the MCMC sampling
  - Templates to get started: `https://github.com/drdeford/GerryChain-Templates`
  - Detailed documentation: `http://people.csail.mit.edu/ddeford/GerryChain_Guide.pdf`
- Data is available for your favorite state!
  - Census dual graphs with demographic information:
  - `https://people.csail.mit.edu/ddeford/dual_graphs`
  - Precincts with electoral results
  - `https://github.com/mggg-states`