Pennsylvania's Congressional districting is an outlier: Expert Report

Wesley Pegden

November 27, 2017

I am an Associate professor in the department of Mathematical Sciences at Carnegie Mellon University, where I have been a member of the faculty since 2013. I received my Ph.D. in Mathematics from Rutgers University in 2010 under the supervision of József Beck, and I am an expert on stochastic processes and discrete probability. My research is funded by the National Science Foundation and the Sloan Foundation. A current CV with a list of publications is attached as Exhibit A. A list of my publications with links to online manuscripts is also available at my website at `http://math.cmu.edu/~wes`.

In early 2017 I published a paper[1] which gave a new statistical test to demonstrate that a configuration is unusual from among a set of candidate configurations, with my coauthors Maria Chikina and Alan Frieze. Our paper, which was published in the *Proceedings of the National Academy of Sciences*, shows that a simple and intuitive procedure can demonstrate that a given configuration is an outlier in a rigorous statistical sense, and in that paper, we showed that our test can be used to demonstrate that a Congressional districting is gerrymandered. A copy of the paper is attached as Exhibit B. Probability is a central aspect of my research; as shown on my CV, I have published papers in several well-regarded journals in probability, including *Annals of Applied Probability* and *Random Structures and Algorithms*. Many of these papers are also collaborations between myself and Alan Frieze, one of my co-authors on the [CFP] paper.

Our PNAS paper (including preliminary analyses of Pennsylvania's Congressional districting) was published before I was contacted by the lawyers for the present case. I am being compensated at a rate of $250 per hour for my work on the current case.

1 Executive Summary

I was asked to analyze whether the current districting of Pennsylvania is an outlier with respect to partisan bias (as opposed to having partisan bias which might be typical of districtings of Pennsylvania, given its political geography). In my analysis, I find that the present Congressional districting of Pennsylvania is indeed a gross outlier with respect to partisan bias, among the set of all possible districtings of Pennsylvania.

My analysis (as in [CFP]) works by testing whether the partisan bias in the current districting is fragile, in the sense that it evaporates when many random small changes are made to the districting. I find that when I begin from the current districting and make roughly 1 trillion of these changes in succession, the districting quickly develops less partisan bias. In particular, the current districting of Pennsylvania exhibits more partisan bias than roughly 99.999999% of districtings encountered in such a sequence of small changes, demonstrating that the current Congressional districting was very carefully crafted to ensure a Republican advantage.

Our analysis goes even further than this, however. As discussed in Section 4, our theorem proved in [CFP] establishes that it is mathematically impossible for the political geography of a state to cause such a result. That is: while political geography might conceivably interact with traditional districting

[1] M. Chikina, A. Frieze, W. Pegden. Assessing significance in a Markov Chain without mixing, in *Proceedings of the National Academy of Sciences* **114** 2860–2864, hereafter [CFP]. PNAS is the official journal of the National Academy of Sciences, and one of the most cited journals across all fields of science. (Articles in PNAS are peer-reviewed.)

criteria to create a situation where typical districtings of a state are biased in favor of one party, it is mathematically impossible for the political geography of a state to interact with traditional districting criteria to create a situation where typical districtings of a state quickly exhibit a *fragile* partisan bias, which quickly evaporates when small changes are made. Quantitatively, the [CFP] theorem tells us that more than 99.99% of the possible Congressional districtings of Pennsylvania would pass our gerrymandering test, showing in a mathematically rigorous way that the present districting was an extremely careful choice made to maximize partisan advantage.

In particular, I find that **Pennsylvania's Congressional districting is a gross outlier with respect to partisan bias** in a way that is **mathematically impossible to be caused by political geography and the traditional districting criteria** I consider.

2 Topic of Expert Report

Election results from Pennsylvania show that Republicans have enjoyed a strong advantage in Congressional elections in Pennsylvania, winning 13 out of 18 seats even in years when Democrats win a majority of statewide Congressional votes.

Though striking, this fact alone does not necessarily mean that Pennsylvania's districting was drawn to give Republicans advantage; *a priori*, it could conceivably be the case that traditional redistricting goals and Pennsylvania's unique political geography could interact to produce a Republican advantage even with an unbiased districting process.

To address this question, petitioners' counsel asked me to analyze whether the Republican advantage in the current Congressional districting of Pennsylvania could be a consequence of nonpartisan factors such as the political geography of the state. In particular, my analysis addresses the question: is the current districting a typical member of the set of possible districtings of Pennsylvania, with respect to its partisan bias? Or is it a gross outlier? We will see, in fact, that my analysis shows that the current Congressional districting of Pennsylvania is more unusual than the vast majority of districtings with respect to partisan bias.

3 A conservative notion of gerrymandering

For the purposes of this expert report, my analysis is predicated on a very conservative definition of what constitutes a gerrymandered districting of a state. In particular, I will **not** call a districting gerrymandered simply because it is more favorable to one party than to the other in terms of the number of seats it leads to for each party. I will **not** even call a districting gerrymandered simply because there were alternative plans available to the mapmakers with less partisan bias which they chose not to use. Instead, my analysis calls a districting gerrymandered only if it passes the following Tests (T1),(T2),(T3):

(T1) The districting has a partisan bias for one party;

(T2) Small random changes to the districting rapidly decrease the partisan bias of the districting, demonstrating that the districting was carefully crafted; and,

(T3) The overwhelming majority of all alternative districtings of the state exhibit (T1), (T2) less than the districting in question.

In particular, when I report that Pennsylvania's 2011 Congressional districting is gerrymandered, I mean not only that there is a partisan advantage for Republicans and that districtings with less partisan bias were available to mapmakers, but indeed that among the entire set of available districtings of Pennsylvania, the districting chosen by the mapmakers was an extreme outlier with respect to partisan bias, in a statistically rigorous way.

To make Test (T3) precise, I have a model for what would constitute a valid Congressional districting of Pennsylvania. For example, it is reasonable to expect that mapmakers drawing the Congressional districting of Pennsylvania want a districting with the following Properties:

(P1) **The districting consists of 18 contiguous districts**.

(P2) **The districting has equipopulous districts**. For example, I can require that the populations of districts differ by less than 2%.

(P3) **The districting has reasonably shaped ("compact") districts**. There are various ways proposed in the literature to quantify this; for example, one can simply impose an upper allowable limit on the total perimeter of all 18 districts.

Specifying constraints such as these determines a **"bag of districtings"** which are candidate districtings of the state. I find that the current districting of the state is an extreme outlier with respect to partisan bias in the sense of (T3), compared with the set of all districtings in this bag.

Moreover, I show that this finding is robust to exactly how I define the bag of districtings. For example: I can define a bag where populations of districts differ by $< 1\%$ instead of $< 2\%$ in implementing property (P2). I define bags of districtings with alternative metrics for implementing property (P3). I can impose additional constraints on the bag of districtings to align with other hypothetical redistricting criteria to ensure, for example, that:

(P4) **The districting does not divide any counties** not divided by the current map of Pennsylvania.

(P5) **The districting includes the current District 2, a Majority-Minority district, intact**, in case it was drawn to comply with the Voting Rights Act.

I show that my finding of the extreme outlier status of Pennsylvania is robust to all of these various choices one might make in defining the bag of districtings.

It is important to note that, for all of these choices I consider for how to define the bag of districtings, my parameters are chosen so that the 2011 districting meets all of corresponding requirements under consideration. In particular, my goal is not to compare the current districting to other "better" districtings which satisfy stricter requirements on the shapes of the districts, etc. Instead, my test assumes the geometric properties of the current districting are reasonable, and compares the districting to the other possible districtings of Pennsylvania with the same properties[2].

Note on population constraint

As we see in Property (P2), my method does not enforce 0% population difference on districts in the comparison districtings. My method does not simulate the results of elections for hypothetical elections at the per-person level, as direct voter preference data is not available at sufficient granularity. In particular, note that the Census does *not* ask individuals for political preference information. Note that this same limitation faces mapmakers who might try to draw a favorable districting for their party; a practical approach is to first use the available data to draw a "coarse" map with the desired properties, and then make small, negligible changes to the map to satisfy the population constraint. Using a population threshold of around 2% is reasonable, because

- A 2% threshold is small in magnitude compared to estimates of the actual error in the Census. In particular, in the 2010 Decennial Census in question, an estimated 3.3% of counts were erroneous

[2] I am not asserting that I consider the geometry of the current map reasonable. However, my analysis accepts the geometric properties of the present districting, to show that even compared to its geometrically similar peers, the map is an extreme outlier.

(e.g., double-counts), while an estimated 5.3% of individuals went uncounted[3], and these errors are correlated with demographic and geographic factors.

- The small population variation in my comparison districtings cannot account for the extreme outlier status I encounter. For example, in my tests, my measure of partisan bias for a districting decreases by a factor of two or more after the sequence of swaps are made, not just by a few percent. This means that even if the maps found by my method after many changes were altered to have equal (up to 1 person) populations, they would still exhibit less partisan bias than the initial maps.
- The particular threshold that I use does not affect our outcome. In particular, if using a 0% threshold would be very different from using a 2% threshold, then I should already see signs of trouble when using a 1% threshold, which is not the case.

4 Mechanics of the method

My goal is to make a statistical comparison of the current Congressional districting of Pennsylvania to the other members of the "bag of districtings" used for comparison. The simplest way to do this would be to simply choose many random districtings from the bag of districtings, to evaluate whether typical districtings of Pennsylvania exhibit less partisan bias than the the 2011 districting. However, there is no general purpose algorithm known which can accomplish this task[4]. The significance of the **$\sqrt{\varepsilon}$-test** from [CFP] is that it gives a simple and elegant way around this problem[5].

For districtings, the test from [CFP] works like this:

1. We begin from the 2011 Congressional districting of Pennsylvania.
2. We randomly select a Census tract on the boundary of 2 districts. We check: if we swap which district this tract belongs to, will the districting still satisfy all the constraints on our bag of districtings? If so, we make the swap.
3. Using voter preference data, we evaluate the partisan bias of the new districting and record whether it is more or less biased than the 2011 districting.[6]
4. We repeat Steps 2 and 3 for n steps, for any fixed number n. We report that the 2011 districting was gerrymandered if the overwhelming majority of districtings encountered by the test exhibited less partisan bias than the 2011 districting.

In my tests I take n to be $2^{40} = 1,099,511,627,776 \approx 1$ trillion. (The test is equally valid for any n, and more powerful the larger I make n.) Because the tests run slower when using the 1% population constraint instead of the 2% population constraint, those runs are conducted with 2^{39} (just over $\frac{1}{2}$ trillion) steps. Our test can be run on a standard personal computer; information on how to obtain our software package and the necessary input data is given in Appendix A.

The results in the next section show that the 2011 districting is more biased than the overwhelming fraction of districtings encountered by the test. For this analysis, I carried out 8 runs of the [CFP] test

[3] H. Hogan, P.J. Cantwell, J.Devine, V.T. Mule Jr., and V. Velkoff. Quality and the 2010 Census, in *Population Research and Policy Review* **32** (2013) 637–662.

[4] In mathematical language, the problem is how to draw efficient random samples from a possibly slowly "mixing" Markov Chain; this is a general problem which occurs throughout scientific disciplines where Markov Chains are used, i.e., in protein folding in microbiology, in statistical physics, and in simulations of chemical reactions. In this context, a *Markov Chain* is a way of generating a random sample through a series of small changes.

[5] Note that the number of districtings in the comparison bag can be astronomical; larger than the number of elementary particles in the known universe, for example, so we cannot simply look at them one by one for a comparison.

[6] We do this using the Median vs. Mean test, which simply compares the medians and the means of the Republican/Democrat splits in each district. This is discussed in Section A.3. Our Results section also contains analyses showing the 2011 districting is an outlier with respect to anti-competitiveness, in addition to partisan bias; we quantify anti-competitiveness of districtings using the variance of the Republican/Democrat splits.

with various constraints, and I consistently find that it is worse than roughly 99.9999999% of encountered districtings. For one of my eight runs, I even find that **the initial (current) Congressional districting of Pennsylvania exhibits more partisan bias than every other of the more than 1 trillion districtings encountered by the test.** In other words, not only does Pennsylvania's 2011 districting exhibit a strong partisan bias (T1), but it satisfies (T2) to an extreme degree.

Even without applying the mathematical theorem from [CFP], **this gives strong intuitively clear evidence of intent to create partisan bias in the districting**: since the test shows that making small changes invariably makes the districting have less partisan bias, it is natural to conclude that the districting was carefully drawn to create partisan bias.

The theorem from [CFP] allows me to translate this intuition into a rigorous statement about how unusual the districting is in the whole bag of possible alternatives. In other words, it gives a formula which can be used to deduce Test (T3) for a districting when it strongly satisfies Test (T2). In particular, the [CFP] theorem, in the districting case, says the following:

$\sqrt{\varepsilon}$ test:

- Suppose that we have run the above test with Pennsylvania's 2011 Congressional districting as the initial districting, and that we have observed that only an ε fraction of districtings we encounter in our test have partisan bias as strong as the 2011 districting. (For example, perhaps $\varepsilon = .000000005$, which means that the initial districting had more partisan bias than 99.9999995% of districtings.)
- The theorem from [CFP] says that among all possible districtings in the bag of alternatives, a randomly chosen districting would perform this badly at most $p = \sqrt{2\varepsilon}$ of the time. For $\varepsilon = .00000000005$, for example, we would conclude that a randomly chosen districting could have probability at most

$$p = \sqrt{2 \times .00000000005} = .00001 = 00.001\%$$

of appearing as biased[7].

One way of interpreting the point of the theorem is as follows: as mentioned in Section 2, it is possible for political geography to make a state more favorable to one party or the other. (For example, Democrats, clustered in cities, could conceivably "waste" more votes even for districtings drawn without bias.) This means that in principle, if one only looks at election outcomes under the districting in question without considering how alternative districtings behave, political geography might conceivably give a false impression that a districting was drawn with bias, whereas really it was not.

The same is not true for the "small changes" test I perform in this analysis, as a consequence of the [CFP] theorem. In particular, the [CFP] theorem tells us that it is mathematically impossible for a state's political geography to inherently produce partisan bias that evaporates quickly when small random changes are made to the state's districting. In other words, when a districting strongly satisfies Test (T2), then it must also satisfy Test (T3), regardless of the political geography of the state. Thus, **political geography cannot fool my analysis into calling a districting an outlier.**

5 Results

Each row shows the results of the test for various conditions on the bag of districtings. In the *compactness measure* column, "Avg. P.P" indicates that the average of the (inverse) Polsby-Popper compactness values for the districts was constrained, while "perimeter" indicates that the total perimeter of all districts was constrained. These choices are discussed in Section A.1.

This table presents our analysis for two properties of the districting: our analysis for partisan bias, and an additional analysis for the anti-competitiveness of the districting. These two tests are done exactly the

[7] Note that this is just an upper bound; the true probability is likely to be even lower. To estimate it directly, however, would require a method of directly choosing random districtings from the bag.

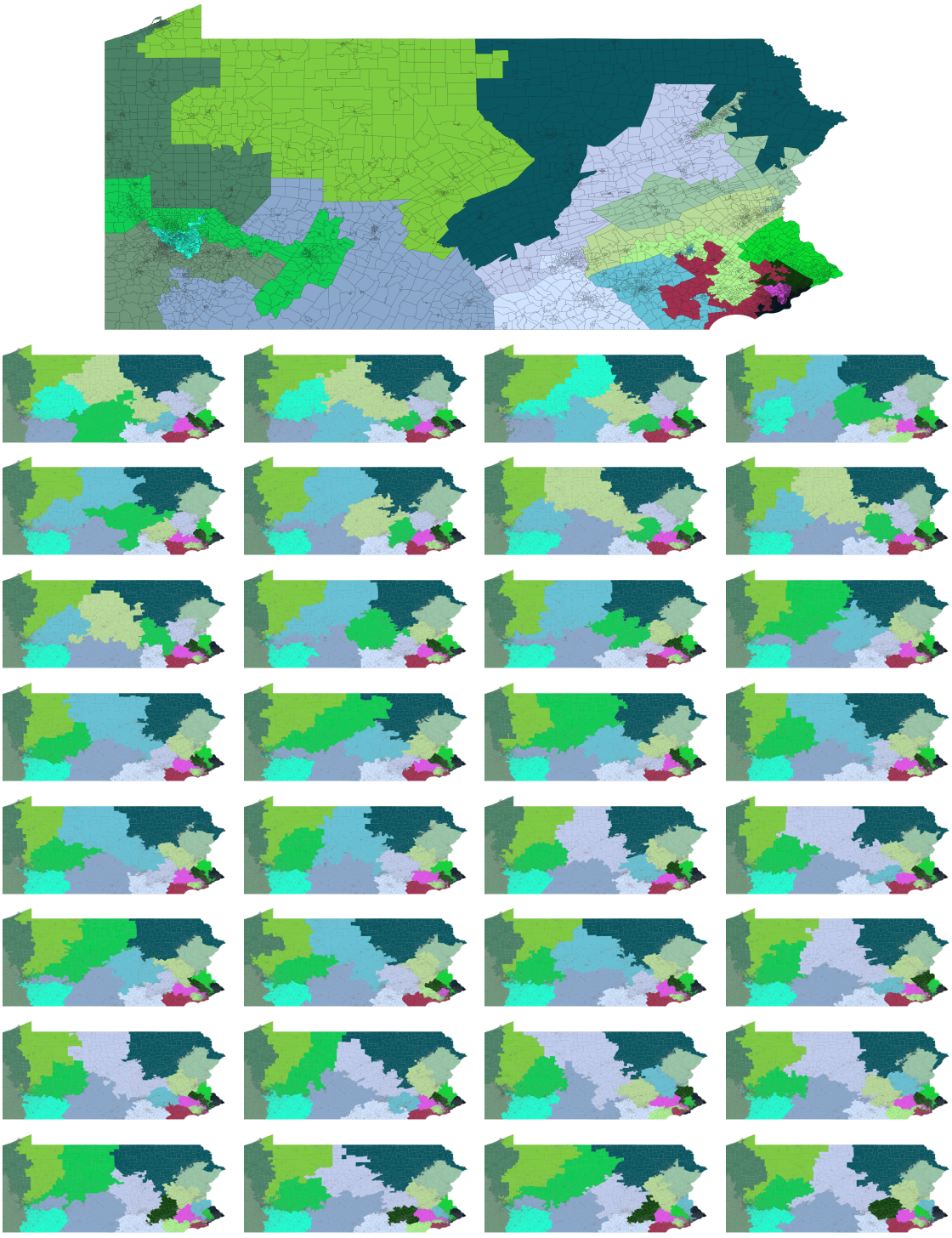

Figure 1: Examples of maps encountered in the small changes test. The large map at the top is the initial (current) districting. These maps are from the run corresponding to the 2nd row of our results table. In particular, only geometry and population are constrained. To produce these maps, our algorithm simply saved a map every $10 \cdot 2^{31} = 21,474,836,480 \approx 20$ billion steps.

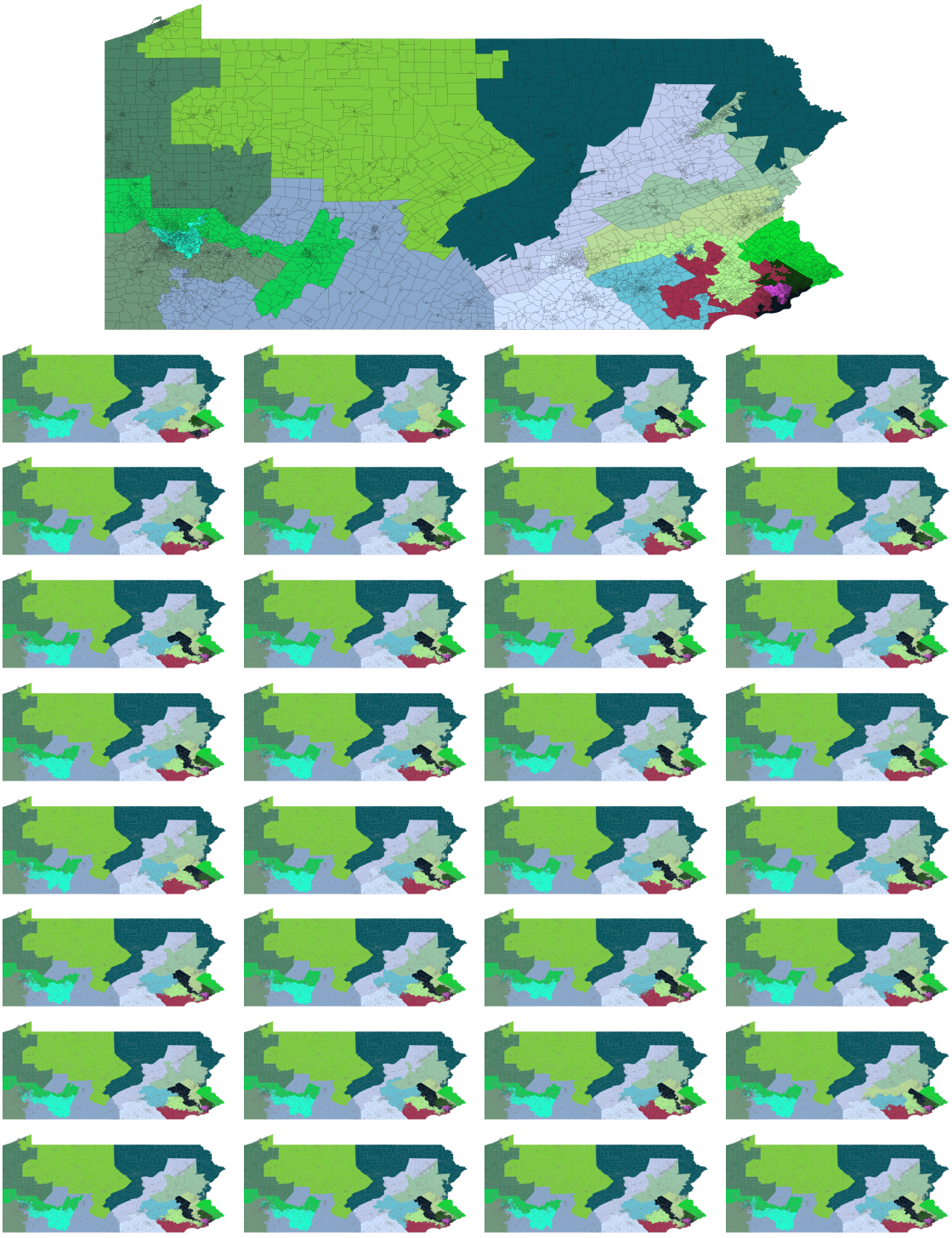

Figure 2: These examples are from the 6th row of our results table. In particular, careful inspection shows that precincts in District 2 remain assigned to District 2 in these maps, and that several rural districts experience few changes since large portions of their boundaries are county boundaries which this run is required to preserve. Again, examples were taken every $10 \cdot 2^{31}$ steps.

same way, except that for the partisan bias test, districtings encountered by the test are evaluated using the Median vs Mean test, while for the competitiveness test, districtings are evaluated according to the variance of the number of Democrats / Republicans in each district. These measures of partisan bias and competitiveness are discussed in Section A.3.

For each test, the ε column is the fraction of districtings encountered in the run of the test for that row which exhibited as much partisan bias (or anti-competitiveness) as the initial (current) districting, and the p column, computed as above as $\sqrt{2\varepsilon}$, is the statistical significance of the observation; in other words, it is the probability that chance alone could lead to the bias measured by the test in that case (regardless of the political geography of the state)[8].

				partisan bias		anti-competitiveness	
population threshold	compactness measure	preserve counties?	freeze dist. 2?	ε-outlier at $\varepsilon =$	significant at $p =$	ε-outlier at $\varepsilon =$	significant at $p =$
2%	perimeter	No	No	.00000000058	.000034	.000000031	.00025
2%	Avg. P.P	No	No	.00000000057	.000034	.000000000051	.000011
2%	perimeter	Yes	No	.0000000013	.000051	.000000032	.00025
2%	avg. P.P.	Yes	No	.00000000000017	.00000058	.00000000042	.000029
2%	perimeter	Yes	Yes	.00000000050	.000032	.0000000000049	.0000032
2%	avg. P.P.	Yes	Yes	.00000000097	.000045	.0000000000048	.0000031
1%	perimeter	Yes	Yes	.00000000038	.000028	.0000000000099	.0000045
1%	avg. P.P.	Yes	Yes	.00000000053	.000033	.0000000000096	.0000044

For example, the first row of the table indicates that when I used a 2% bound for the population constraint, constrained the compactness of districts using the total perimeter of the districting, ignored the preservation of counties and did not freeze the Majority-Minority district #2, I found that only a .00000000058 fraction of districtings encountered in the test showed as much partisan bias as the current districting, an observation which our theorem shows can happen with probability at most 00.0034% for a typical districting, regardless of the political geography of the state.

Our finding is that Pennsylvania is dramatically gerrymandered; **its current Congressional districting is an extreme outlier among the set of possible alternatives, in a way that it is insensitive to how precisely I define the set of alternatives.**

A Technical details

In this appendix I discuss some technical details regarding the preparation of our results. For a more precise account of our methods, the precise mathematical statement of the theorem I am employing, and its mathematical proof, I point the reader to our paper [CFP].

The software package implementing our test can be downloaded from the following URL:

`http://math.cmu.edu/~wes/files/markovchain.tgz`

[8] It is actually possible for the ε column to be slightly smaller than 1 in 1 trillion, as happens with the 4th row. This is because we run the test until 2^{40} swaps have succeeded, which in general, takes a bit more than 2^{40} steps; in particular, the same map may serve as a comparison multiple times in a sequence. This consideration of repeated maps is necessary to ensure that each map is given equal weight for the purpose of comparison. (In technical terms, it ensures that the *uniform distribution* is a *stationary distribution* for our *Markov Chain*.) For the technical details and reasons behind this phenomenon, we refer the reader to [CFP].

This package contains both the code and input files necessary to run our tests (and generate the maps shown in Figures 1 and 2 of this report). It includes a README file which describes the installation and use of the package in a standard Linux environment. The input files distributed with the package are derived (as described in [CFP]) from the 2010 Pennsylvania data from the Harvard Election Data Archive, available here:

`http://hdl.handle.net/1902.1/16389`

A.1 Compactness measure

Various precise metrics have been proposed to quantify the compactness of a given district mathematically. One of the simplest and most commonly used metrics is the Polsby-Popper metric, which simply considers the ratio of the area of the district to the square of its perimeter. (This is sensible because for "nice" shapes such as circles and squares, the area grows roughly as the square of the perimeter, up to constant factors.)

In particular, mathematically, one computes the Polsby-Popper compactness as

$$C_D = \frac{4\pi A_D}{P_D^2},$$

where A_D and P_D are the area and perimeter of the district D. Here 4π is a normalizing constant which simply ensures that the the maximum value of this measure is 1 (which would be achieved only by a perfectly circular district). All other shapes have compactness between 0 and 1, and smaller values indicate more "contorted" shapes. The inverse of the P.P. compactness is a number between 1 and ∞, with higher numbers indicating more contorted shapes.

In our results table, rows whose compactness measure is indicated as "avg. P.P." had their compactness constrained by a threshold on the average of the inverse of the Polsby-Popper metric of the 18 districts in the districting; this value is approximately 156.4 for the initial districting, and the threshold for these runs was set at 160.

Rows whose compactness measure is indicated as "perimeter" had their compactness constrained by a threshold on the total perimeter of the 18 districts in the districting. This is approximately 121.2 for the initial districting, and our threshold was set at 125[9].

A.2 Voter preference

To assess the partisan bias of a given districting, it is necessary to have an estimate of voter preferences in each "Census tract" from which districtings are assembled.

As a proxy for partisan bias I use the election results from the 2010 Pennsylvania senate race between Pat Toomey and Joe Sestak. This race has several characteristics which make it an excellent proxy for voter preference:

- It was a statewide race;
- there was no incumbent in the race; and,
- it was among the most recent data available to the mapmakers when drawing the currently contested districting.

Of course no proxy for voter preference is perfect; however, any imperfect relationship between this proxy and true voter preference only *decreases* the sensitivity of our test. In particular, the fact that we have only imperfect proxies for true voter preference makes it *harder* to detect gerrymandering, not *easier*, since my analysis will only call a districting gerrymandered when it is carefully crafted relative to the voter preference proxy I am using.

[9] The units for these values are in the coordinate system of the Census shapefiles, and correspond to roughly $100km$.

A.3 Metrics for partisan bias and competitiveness

Our test can be applied using any standard metrics to evaluate districtings. To evaluate the partisan bias of districtings, I used the Median vs. Mean test (called the *Symmetry Vote Bias* test by McDonald and Best[10]), a metric which has been used to evaluate partisan advantage in districting since the 19th century[11].

The Median vs. Mean test can be carried out very simply for any districting as follows. First one lists the fraction of each district which are expected to vote for Republicans. For example X_1 is this fraction for district 1, X_2 for district 2, and so on. In particular, the *mean* of $X_1, \ldots, X_{18}$ is just the overall fraction of the state which we expect to vote for Republicans (assuming each district has roughly the same number of voters).

The Median vs. Mean test simply returns the difference between the median and mean of these numbers. If the difference between the median and the mean is positive, this indicates an advantage for the Republicans, and if the difference is negative, it indicates an advantage for the Democrats. (If the X_i's represented Democrats' shares of votes, this relationship would just be reversed.)

To give just a rough intuition for the motivation for the metric, suppose that the median of the X_i's is 50%; this means that Republicans are winning half the seats (since 50% is the threshold for them to win an election). Now if the *mean* of the X_i's is much smaller than 50%, it means they are winning half the seats even though they have a small minority of the total *votes*, since the mean of the X_i's is just the overall level of support for Republicans.

I quantify the competitiveness of districtings by simply measuring the *variance* of the X_i's, computed as the average of the squares of the X_i, minus the square of the average:

$$\frac{X_1^2 + X_2^2 + \cdots + X_{18}^2}{18} - \left(\frac{X_1 + X_2 + \cdots + X_{18}}{18}\right)^2.$$

Put differently, the variance of the X_i's is just the square of the standard deviation of the X_i's, and so this is just a measure of how far the X_i's are from their mean. For example, when the variance is high, it means that there are districts that are significantly more Republican than the statewide average, and districts that are significantly more Democratic than the statewide average. For districtings where the variance is especially large compared with alternative districtings, this means that there are especially anti-competitive districts. In short, high variance means anti-competitive districtings, while low variance means competitive districtings[12].

I hereby certify that the foregoing statements are true and correct to the best of my knowledge, information, and belief. This verification is made subject to the penalties of 18 Pa.C.S. §4904 relating to unsworn falsification to authorities.

Wesley Pegden
11/27/17

[10] M.D. McDonald and R.E. Best. Unfair Partisan Gerrymanders in Politics and Law: A Diagnostic Applied to Six Cases, in *Election Law Journal* **14** (2015) 312–330.

[11] F.Y. Edgeworth. Miscellaneous applications of the Calculus of Probabilities, in the *Journal of the Royal Statistical Society* **60** (1897) 681–698.

[12] See for example, the section "Partisan Outcomes by Congressional District and State", in S.M. Theriault, *Party polarization in Congress*. Cambridge University Press, 2008. His use of the standard deviation is equivalent to our use of the variance.

Exhibit A

Wesley Pegden

CONTACT INFORMATION

Department of Mathematical Sciences
Carnegie Mellon University
Pittsburgh, PA 15217

office: 412 268 9782
cell: 412 708 3772
wes@math.cmu.edu
http://math.cmu.edu/~wes

CURRENT POSITION

Carnegie Mellon University, Pittsburgh, PA

Associate Professor, 2017–present
Assistant Professor, 2013–2017

POSTDOCTORAL

Courant Institute, New York, NY

NSF Postdoctoral Fellow, 2010–2013

EDUCATION

Rutgers University, New Brunswick, NJ

Ph.D., May 2010.
Advisor: József Beck
Thesis: "Graphs, games and geometry".

Budapest Semesters in Mathematics, Budapest, Hungary: 2004–2005

University of Chicago: 2001–2004.
BA in Mathematics, with Honors.

GRANTS, FELLOWSHIPS, AND AWARDS

Kavli Fellow
NSF Grant DMS-1700365 (2017–2020)
Sloan Fellowship (2016–2018)
NSF Grant DMS-1363136 (2014–2017)
NSF Postdoctoral Research Fellowship (2010–2013)
Torrey Fellow (Rutgers, 2005–2007)

PUBLICATIONS

Assessing significance in a Markov chain without mixing, with M. Chikina and A. Frieze.
Proceedings of the National Academy of Sciences **114** (2017) 2860-2864.

Looking for vertex number one, with A. Frieze.
Annals of Applied Probability **27** (2017) 582-630.

The Apollonian structure of integer superharmonic matrices, with L. Levine and C. Smart.
Annals of Mathematics **186** (2017) 1-67.

Traveling in randomly embedded random graphs, with A. Frieze.
RANDOM 2017. Preprint at http://arxiv.org/abs/1411.6596

Apollonian structure in the Abelian Sandpile, with L. Levine and C. Smart.
GAFA **26** (2016) 306-336.

Separating subadditive Euclidean functionals, with A. Frieze.
STOC 2016. Journal version is *Random Structures & Algorithms* **51** (2017) 375-403.

Between 2- and 3-colorability, with A. Frieze.
Electronic Journal of Combinatorics **22** #P1.34 (2015).

Walker-Breaker games, with L. Espig, A. Frieze, and M. Krivelevich.
SIAM J. Discrete Math. **29** (2015).

The topology of competitively constructed graphs, with A. Frieze.
The Electronic Journal of Combinatorics **21** #P2.26.

Convergence of the Abelian Sandpile, with Charles K. Smart.
Duke Mathematical Journal **162** (2013) 627–642.

An extension of the Moser-Tardos algorithmic Local Lemma.
SIAM J. Discrete Math. **28** (2014)

Critical graphs without triangles: an optimum density construction.
Combinatorica **33** (2013) 495–512

The Lefthanded Local Lemma characterizes chordal dependency graphs.
Random Structures & Algorithms **41** (2012) 546–556

Highly nonrepetitive sequences: winning strategies from the Local Lemma.
Random Structures & Algorithms **38** (2011)

Sets resilient to erosion.
Advances in Geometry **11** (2011) 201–224.

The Hales-Jewett number is exponential, with J. Beck and S. Vijay.
Analytic Number Theory: Essays in Honour of Klaus Roth (Eds: W.W.L. Chen, W.T. Gowers, H. Halberstam, W.M. Schmidt, R.C. Vaughan), Cambridge University Press 2009.

A finite goal set in the plane which is not a winner.
Discrete Mathematics **308** (2008) 6546–6551.

Distance Sequences in Locally Infinite Vertex-Transitive Digraphs.
Combinatorica **26** (2006) 577–585.

Preprints

A partisan districting protocol with provably nonpartisan outcomes, with A. D. Procaccia, D. Yu.
Submitted. Preprint at `https://arxiv.org/abs/1710.08781`.

Stability of patterns in the Abelian sandpile, with C. Smart.
Submitted. Preprint at `https://arxiv.org/abs/1708.09432`.

Online purchasing under uncertainty, with A. Frieze.
Random Structures & Algorithms (Accepted). Preprint at `http://arxiv.org/abs/1605.06072`

Diffusion limited aggregation in the Boolean lattice, with A. Frieze.
Submitted. Preprint at `https://arxiv.org/abs/1705.00692`.

Constraining the clustering transition for colorings of sparse random graphs, with M. Anastos, A. Frieze.
Submitted. Preprint at `https://arxiv.org/abs/1705.07944`

Minors of a random binary matroid, with C. Cooper and A. Frieze.

Submitted. Preprint at https://arxiv.org/abs/1612.02084

On the distribution of the minimum weight clique, with A. Frieze and G. Sorkin. Submitted. Preprint at http://arxiv.org/abs/1606.04925

Scalefree hardness of average-case Euclidean TSP approximation, with A. Frieze. Submitted. Preprint at http://arxiv.org/abs/1604.04549

Coverage in Popular Media

The amazing, autotuning sandpile. Article in *Nautilus* magazine (2015) by J. Ellenberg. http://nautil.us/issue/23/dominoes/the-amazing-autotuning-sandpile.

Das Wahnsinnsamt, Sandhäufchen und Apollonische Dreiecke. Article in *Spektrum der Wissenschaft* (German-language edition of Scientific American) by C. Pöppe, pages 67–71, August 2015.

Piling on and on and on.... Article and podcast interview by M. Breen, at http://www.ams.org/samplings/mathmoments/mm117-sandpiles-podcast (2015).

On sandpiles. Coverage in *AMS Math in the media* column, by Allyn Jackson (April 2015).

Math and the gerrymander. Coverage in *AMS Math in the media* column, by Tony Phillips (March 2017).

Study: Math proves Pennsylvania's congressional districts 'almost certainly' gerrymandered. Story in *Philly Voice* by Daniel Craig (March 1, 2017).

Where Allegheny County's Harrisburg delegation stands on redistricting reform. Story in *The Incline* by Sarah Anne Hughes (March 23, 2017).

Groups sue Pennsylvania over congressional district map, citing gerrymandering. Story in the *Pittsburgh Post-Gazette* by Chris Potter (June 15, 2017).

Cake-cutting game theory trick could stop gerrymandering. Story in the *New Scientist* by Timothy Revell (November 1, 2017).

Talks

Ohio State Discrete Math Seminar, April 20, 2017, at Ohio State University.

Atlanta Lecture Series in Combinatorics and Graph Theory XVIII, October 22-23, 2016, at Emory University.

Princeton Discrete Mathematics Seminar, October 13, 2016, at Princeton University.

[*Conference*]*STOC 2016*, June 19 2016, in Boston, MA.

Princeton Discrete Mathematics Seminar, March 10, 2016, at Princeton University.

University of Chicago Theory Seminar, October 20, 2015 at the University of Chicago.

CMU CS Theory Seminar, May 14, 2015.

[*Colloquium*] University of Geneva, March 5 2015.

Ohio State Discrete Math Seminar, November 6 2014, at Ohio State University.

[*Conference*] *SIAM DM14, Special Session on Combinatorics and Statistical Mechanics*, June 18 2014 in Minneapolis, MN (2 talks)

Princeton Discrete Math Seminar, March 13 2014, at Princeton University.

AIM workshop: Generalizations of chip-firing and the critical group, July 2013 at AIM.

[*Conference*] *Special Session on Combinatorics and Classical Integrability* at the AMS Spring Eastern Sectional Meeting, April, 2013 at Boston College.

[*Colloquium*] University of Illinois at Urbana-Champaign, January 30, 2013.

[*Colloquium*] CMU, January 16, 2013.

[*Colloquium*] University of Illinois at Chicago, December 5, 2012.

Cornell Workshop on Sandpiles and Number Theory, October 2012 at Cornell University in Ithaca, NY.

MIT Combinatorics Seminar, April 27, 2012, at MIT.

UPenn seminar on Combinatorics and Probability, February 21, 2012, at the University of Pennsylvania.

Rutgers Discrete Math Seminar, February 7, 2012, at Rutgers University in New Brunswick.

Princeton Discrete Math Seminar, September 27, 2012 at Princeton University.

Probabilistic Combinatorics Mini-symposium of SIAM DM12, June 19, 2012 in Halifax, Nova Scotia.

[*Conference*] *the 15th conference on Random Structures & Algorithms*, May 24, 2012, at Emory University.

Columbia Discrete Math Seminar, February 14, 2012 at Columbia University in New York, NY.

Rutgers Discrete Math Seminar, February 1, 2011 at Rutgers University in New Brunswick.

New York Number Theory Seminar, November 4, 2010.

Columbia Discrete Math Seminar, October 27, 2009 at Columbia University in New York, NY.

Princeton Discrete Math Seminar, October 22, 2009 at Princeton University.

The 14th International Conference on Random Structures and Algorithms, in Poznań, Poland, August 2009.

[*Conference*] *Special Session on Probabilistic and Extremal Combinatorics*, at the 2009 AMS Spring Sectional Meeting, UIUC in Urbana-Champaign, IL.

Rutgers Discrete Mathematics Seminar, April 28 at Rutgers University in New Brunswick.

[*Conference*] *National AMS meeting*, in Washington, DC, January 2009.

Rutgers Experimental Mathematics Seminar, February 7 at Rutgers University in New Brunswick.

Princeton Discrete Mathematics Seminar, in December 2007 at Princeton University.

[*Conference*] *Workshop on Extremal Combinatorics*, Alfred Renyi Institute of Mathematics, in Budapest, Hungary, June 2007.

Exhibit B

Assessing significance in a Markov chain without mixing

Maria Chikina[a], Alan Frieze[b], and Wesley Pegden[b,1]

[a]Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA 15213; and [b]Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved January 24, 2017 (received for review October 21, 2016)

We present a statistical test to detect that a presented state of a reversible Markov chain was not chosen from a stationary distribution. In particular, given a value function for the states of the Markov chain, we would like to show rigorously that the presented state is an outlier with respect to the values, by establishing a *p* value under the null hypothesis that it was chosen from a stationary distribution of the chain. A simple heuristic used in practice is to sample ranks of states from long random trajectories on the Markov chain and compare these with the rank of the presented state; if the presented state is a 0.1% outlier compared with the sampled ranks (its rank is in the bottom 0.1% of sampled ranks), then this observation should correspond to a *p* value of 0.001. This significance is not rigorous, however, without good bounds on the mixing time of the Markov chain. Our test is the following: Given the presented state in the Markov chain, take a random walk from the presented state for any number of steps. We prove that observing that the presented state is an ε-outlier on the walk is significant at $p=\sqrt{2\varepsilon}$ under the null hypothesis that the state was chosen from a stationary distribution. We assume nothing about the Markov chain beyond reversibility and show that significance at $p\approx\sqrt{\varepsilon}$ is best possible in general. We illustrate the use of our test with a potential application to the rigorous detection of gerrymandering in Congressional districting.

Markov chain | mixing time | gerrymandering | outlier | *p* value

The essential problem in statistics is to bound the probability of a surprising observation under a null hypothesis that observations are being drawn from some unbiased probability distribution. This calculation can fail to be straightforward for a number of reasons. On the one hand, defining the way in which the outcome is surprising requires care; for example, intricate techniques have been developed to allow sophisticated analysis of cases where multiple hypotheses are being tested. On the other hand, the correct choice of the unbiased distribution implied by the null hypothesis is often not immediately clear; classical tools like the t test are often applied by making simplifying assumptions about the distribution in such cases. If the distribution is well-defined but is not be amenable to mathematical analysis, a p value can still be calculated using bootstrapping if test samples can be drawn from the distribution.

A third way for p value calculations to be nontrivial occurs when the observation is surprising in a simple way and the null hypothesis distribution is known but where there is no simple algorithm to draw samples from this distribution. In these cases, the best candidate method to sample from the null hypothesis is often through a Markov chain, which essentially takes a long random walk on the possible values of the distribution. Under suitable conditions, theorems are available that guarantee that the chain converges to its stationary distribution, allowing a random sample to be drawn from a distribution quantifiably close to the target distribution. This principle has given rise to diverse applications of Markov chains, including to simulations of chemical reactions, Markov chain Monte Carlo statistical methods, protein folding, and statistical physics models.

A persistent problem in applications of Markov chains is the often unknown rate at which the chain converges with the stationary distribution (1, 2). It is rare to have rigorous results on the mixing time of a real-world Markov chain, which means that, in practice, sampling is performed by running a Markov chain for a "long time" and hoping that sufficient mixing has occurred. In some applications, such as in simulations of the Potts model from statistical physics, practitioners have developed modified Markov chains in the hopes of achieving faster convergence (3), but such algorithms have still been shown to have exponential mixing times in many settings (4–6).

In this article, we are concerned with the problem of assessing statistical significance in a Markov chain without requiring results on the mixing time of the chain or indeed, any special structure at all in the chain beyond reversibility. Formally, we consider a reversible Markov chain $\mathcal{M}$ on a state space Σ, which has an associated label function $\omega:\Sigma\to\Re$. (The definition of Markov chain is recalled at the end of this section.) The labels constitute auxiliary information and are not assumed to have any relationship to the transition probabilities of $\mathcal{M}$. We would like to show that a presented state σ_0 is unusual for states drawn from a stationary distribution π. If we have good bounds on the mixing time of $\mathcal{M}$, then we can simply sample from a distribution of $\omega(\pi)$ and use bootstrapping to obtain a rigorous p value for the significance of the smallness of the label of σ_0. However, such bounds are rarely available.

We propose the following simple and rigorous test to detect that σ_0 is unusual relative to states chosen randomly according to π, which does not require bounds on the mixing rate of $\mathcal{M}$.

The $\sqrt{\varepsilon}$ test. Observe a trajectory $\sigma_0,\sigma_1,\sigma_2\ldots,\sigma_k$ from the state σ_0 for any fixed k. The event that $\omega(\sigma_0)$ is an ε-outlier among $\omega(\sigma_0),\ldots,\omega(\sigma_k)$ is significant at $p=\sqrt{2\varepsilon}$ under the null hypothesis that $\sigma_0\sim\pi$.

Here, we say that a real number α_0 is an ε-outlier among $\alpha_0,\alpha_2,\ldots,\alpha_k$ if there are, at most, $\varepsilon(k+1)$ indices i for which

Significance

Markov chains are simple mathematical objects that can be used to generate random samples from a probability space by taking a random walk on elements of the space. Unfortunately, in applications, it is often unknown how long a chain must be run to generate good samples, and in practice, the time required is often simply too long. This difficulty can preclude the possibility of using Markov chains to make rigorous statistical claims in many cases. We develop a rigorous statistical test for Markov chains which can avoid this problem, and apply it to the problem of detecting bias in Congressional districting.

Author contributions: M.C., A.F., and W.P. performed research and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: wes@math.cmu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1617540114/-/DCSupplemental.

$\alpha_i \leq \alpha_0$. In particular, note for the $\sqrt{\varepsilon}$ test that the only relevant feature of the label function is the ranking that it imposes on the elements of Σ. In *SI Text*, we consider the statistical power of the test and show that the relationship $p \approx \sqrt{\varepsilon}$ is best possible. We leave as an open question whether the constant $\sqrt{2}$ can be improved.

Roughly speaking, this kind of test is possible, because a reversible Markov chain cannot have many local outliers (Fig. 1). Rigorously, the validity of the test is a consequence of the following theorem.

Theorem 1.1. Let $\mathcal{M} = X_0, X_1, \ldots$ be a reversible Markov chain with a stationary distribution π, and suppose the states of $\mathcal{M}$ have real-valued labels. If $X_0 \sim \pi$, then for any fixed k, the probability that the label of X_0 is an ε-outlier from among the list of labels observed in the trajectory $X_0, X_1, X_2, \ldots, X_k$ is, at most, $\sqrt{2\varepsilon}$.

We emphasize that Theorem 1.1 makes no assumptions on the structure of the Markov chain beyond reversibility. In particular, it applies even if the chain is not irreducible (in other words, even if the state space is not connected), although in this case, the chain will never mix.

In Detecting Bias in Political Districting, we apply the test to Markov chains generating random political districting for which no results on rapid mixing exist. In particular, we show that, for various simple choices of constraints on what constitutes a "valid" Congressional districting (e.g., that the districts are contiguous and satisfy certain geometric constraints), the current Congressional districting of Pennsylvania is significantly biased under the null hypothesis of a districting chosen at random from the set of valid districting. (We obtain p values between $\approx 2.5 \cdot 10^{-4}$ and $\approx 8.1 \cdot 10^{-7}$ for the constraints that we considered.)

One hypothetical application of the $\sqrt{\varepsilon}$ test is the possibility of rigorously showing that a chain is not mixed. In particular, suppose that Research Group 1 has run a reversible Markov chain for n_1 steps and believes that this was sufficient to mix the chain. Research Group 2 runs the chain for another n_2 steps, producing a trajectory of total length $n_1 + n_2$, and notices that a property of interest changes in these n_2 additional steps. Heuristically, this observation suggests that n_1 steps were not sufficient to mix the chain, and the $\sqrt{\varepsilon}$ test quantifies this reasoning rigorously. For this application, however, we must allow X_0 to be distributed not exactly as the stationary distribution π but as some distribution π' with total variation distance to π that is small, as is the scenario for a "mixed" Markov chain. In *SI Text*, we give a version of Theorem 1.1, which applies in this scenario.

One area of research related to this manuscript concerns methods for perfect sampling from Markov chains. Beginning with the Coupling from the Past (CFTP) algorithm of Propp and Wilson (7, 8) and several extensions (9, 10), these techniques are designed to allow sampling of states exactly from the stationary distribution π without having rigorous bounds on the mixing time of the chain. Compared with the $\sqrt{\varepsilon}$ test, perfect sampling techniques have the disadvantages that they require the Markov chain to possess a certain structure for the method to be implementable and that the time that it takes to generate each perfect sample is unbounded. Moreover, although perfect sampling methods do not require rigorous bounds on mixing times to work, they will not run efficiently on a slowly mixing chain. The point is that for a chain that has the right structure and that actually mixes quickly (despite an absence of a rigorous bound on the mixing time), algorithms like CFTP can be used to rigorously generate perfect samples. However, the $\sqrt{\varepsilon}$ test applies to any reversible Markov chain, regardless of the structure, and has running time k chosen by the user. Importantly, it is quite possible that the test can detect bias in a sample even when k is much smaller than the mixing time of the chain, which seems to be the case in the districting example discussed in Detecting Bias in Political Districting. Of course, unlike perfect sampling methods, the $\sqrt{\varepsilon}$ test can only be used to show that a given sample is not chosen from π; it does not give a way for generating samples from π.

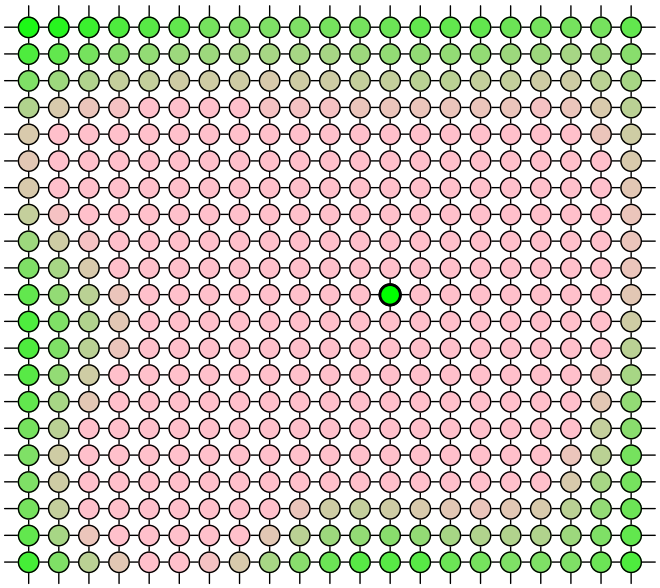

Fig. 1. This schematic illustrates a region of a potentially much larger Markov chain with a very simple structure; from each state seen here, a jump is made with equal probability to each of four neighboring states. Colors from green to pink represent labels from small to large, respectively. It is impossible to know from this local region alone whether the highlighted green state has unusually small label in this chain overall. However, to an unusual degree, this state is a local outlier. The $\sqrt{\varepsilon}$ test is based on the fact that no reversible Markov chain can have too many local outliers.

Definitions

We remind the reader that a Markov chain is a discrete time random process; at each step, the chain jumps to a new state, which only depends on the previous state. Formally, a Markov chain $\mathcal{M}$ on a state space Σ is a sequence $\mathcal{M} = X_0, X_1, X_2, \ldots$ of random variables taking values in Σ (which correspond to states that may be occupied at each step), such that, for any $\sigma, \sigma_0, \ldots, \sigma_{n-1} \in \Sigma$,

$$\begin{aligned}\mathbf{Pr}(X_n = \sigma | X_0 = \sigma_0, X_1 = \sigma_1, \ldots, X_{n-1} = \sigma_{n-1}) \\ = \mathbf{Pr}(X_1 = \sigma | X_0 = \sigma_{n-1}).\end{aligned}$$

Note that a Markov chain is completely described by the distribution of X_0 and the transition probabilities $\mathbf{Pr}(X_1 = \sigma_1 | X_0 = \sigma_0)$ for all pairs $\sigma_0, \sigma_1 \in \Sigma$. Terminology is often abused, so that the Markov chain refers only to the ensemble of transition probabilities, regardless of the choice of distribution for X_0.

With this abuse of terminology, a stationary distribution for the Markov chain is a distribution π, such that $X_0 \sim \pi$ implies that $X_1 \sim \pi$ and therefore, that $X_i \sim \pi$ for all i. When the distribution of X_0 is a stationary distribution, the Markov chain $X_0, X_1, \ldots$ is said to be stationary. A stationary chain is said to be reversible if, for all i, k, the sequence of random variables $(X_i, X_{i+1}, \ldots, X_{i+k})$ is identical in distribution to the sequence $(X_{i+k}, X_{i+k-1}, \ldots, X_i)$. Finally, a chain is reducible if there is a pair of states σ_0, σ_1, such that σ_1 is inaccessible from σ_0 via legal transitions and irreducible otherwise.

A simple example of a Markov chain is a random walk on a directed graph beginning from an initial vertex X_0 chosen from some distribution. Here, Σ is the vertex set of the directed graph. If we are allowed to label the directed edges with positive reals

and if the probability of traveling along an arc is proportional to the label of the arc (among those leaving the present vertex), then any Markov chain has such a representation, because the transition probability $\mathbf{Pr}(X_1 = \sigma_1 | X_0 = \sigma_0)$ can be taken as the label of the arc from σ_0 to σ_1. Finally, if the graph is undirected, the corresponding Markov chain is reversible.

Detecting Bias in Political Districting

A central feature of American democracy is the selection of Congressional districts in which local elections are held to directly elect national representatives. Because a separate election is held in each district, the proportions of party affiliations of the slate of representatives elected in a state do not always match the proportions of statewide votes cast for each party. In practice, large deviations from this seemingly desirable target do occur.

Various tests have been proposed to detect "gerrymandering" of districting, in which a district is drawn in such a way as to bias the resulting slate of representatives toward one party, which can be accomplished by concentrating voters of the unfavored party in a few districts. One class of methods to detect gerrymandering concerns heuristic "smell tests," which judge whether districting seems generally reasonable in its statistical properties (11, 12). For example, such tests may frown on districting in which difference between the mean and median votes on district by district basis is unusually large (13).

The simplest statistical smell test, of course, is whether the party affiliation of the elected slate of representatives is close in proportion to the party affiliations of votes for representatives. Many states have failed this simple test spectacularly, such as in Pennsylvania; in 2012, 48.77% of votes were cast for Republican representatives and 50.20% of votes were cast for Democrat representatives in an election that resulted in a slate of 13 Republican representatives and 5 Democrat representatives.

Heuristic statistical tests such as these all suffer from lack of rigor, however, because of the fact that the statistical properties of "typical" districting are not rigorously characterized. For example, it has been shown (14) that Democrats may be at a natural disadvantage when drawing electoral maps, even when no bias is at play, because Democrat voters are often highly geographically concentrated in urban areas. Particularly problematic is that the degree of geographic clustering of partisans is highly variable from state to state: what looks like gerrymandered districting in one state may be a natural consequence of geography in another.

Some work has been done in which the properties of valid districting are defined (which may be required to have roughly equal populations among districts, districts with reasonable boundaries, etc.), so that the characteristics of a given districting can be compared with what would be typical for valid districting of the state in question, by using computers to generate random districting (15, 16); there is discussion of this in ref. 13. However, much of this work has relied on heuristic sampling procedures, which do not have the property of selecting districting with equal probability (and more generally, distributions that are not well-characterized), undermining rigorous statistical claims about the properties of typical districts.

In an attempt to establish a rigorous framework for this kind of approach, several groups (17–19) have used Markov chains to sample random valid districting for the purpose of such comparisons. Like many other applications of real-world Markov chains, however, these methods suffer from the completely unknown mixing time of the chains in question. Indeed, no work has even established that the Markov chains are irreducible (in the case of districting, irreducibility means that any valid districting can be reached from any other by a legal sequence of steps), even if valid districting was only required to consist of contiguous districts of roughly equal populations. Additionally, indeed, for very restrictive notions of what constitutes valid districting, irreducibility certainly fails.

As a straightforward application of the $\sqrt{\varepsilon}$ test, we can achieve rigorous p values in Markov models of political districting, despite the lack of bounds on mixing times of the chains. In particular, for all choices of the constraints on valid districting that we tested, the $\sqrt{\varepsilon}$ test showed that the current Congressional districting of Pennsylvania is an outlier at significance thresholds ranging from $p \approx 2.5 \cdot 10^{-4}$ to $p \approx 8.1 \cdot 10^{-7}$. Detailed results of these runs are in *SI Text*.

A key advantage of the Markov chain approach to gerrymandering is that it rests on a rigorous framework, namely comparing the actual districting of a state with typical (i.e., random) districting from a well-defined set of valid districting. The rigor of the approach thus depends on the availability of a precise definition of what constitutes valid districting; in principle and in practice, the best choice of definition is a legal question. Although some work on Markov chains for redistricting (in particular, ref. 19) has aimed to account for complex constraints on valid districting, our main goal in this manuscript is to illustrate the application of the $\sqrt{\varepsilon}$ test. In particular, we have erred on the side of using relatively simple sets of constraints on valid districting in our Markov chains, while checking that our significance results are not highly sensitive to the parameters that we use. However, our test immediately gives a way of putting the work, such as that in ref. 19, on a rigorous statistical footing.

The full description of the Markov chain that we use in this work is given in *SI Text*, but its basic structure is as follows: Pennsylvania is divided into roughly 9,000 census blocks. (These blocks can be seen on close inspection of Fig. 2.) We define a division of these blocks into 18 districts to be a valid districting of Pennsylvania if districts differ in population by less than 2%, are contiguous, are simply connected (districts do not contain holes), and are "compact" in ways that we discuss in *SI Text*; roughly, this final condition prohibits districts with extremely contorted structure. The state space of the Markov chain is the set of valid districting of the state, and one step of the Markov chain

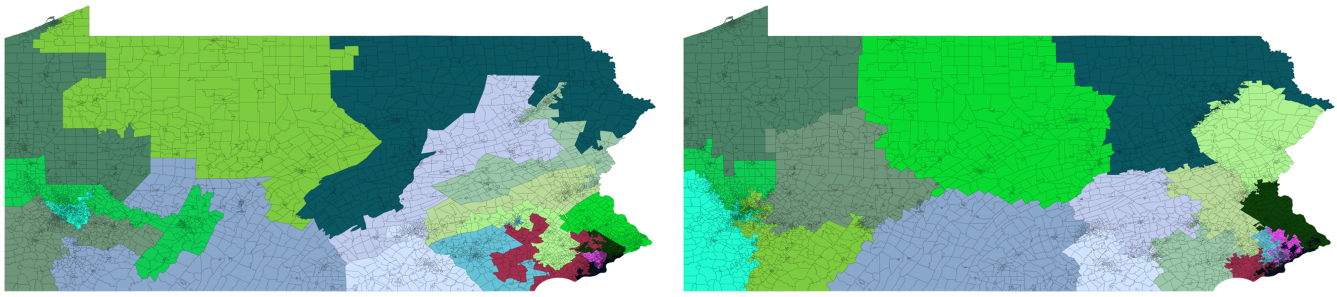

Fig. 2. (*Left*) The current districting of Pennsylvania. (*Right*) Districting produced by the Markov chain after 2^{40} steps. (Detailed parameters for this run are given in *SI Text*.)

consists of randomly swapping a precinct on the boundary of a district to a neighboring district if the result is still a valid districting. As we discuss in *SI Text*, the chain is adjusted slightly to ensure that the uniform distribution on valid districting is indeed a stationary distribution for the chain. Observe that this Markov chain has a potentially huge state space; if the only constraint on valid districting was that the districts have roughly equal population, there would be 10^{10000} or so valid districtings. Although contiguity and especially, compactness are severe restrictions that will decrease this number substantially, it seems difficult to compute effective upper bounds on the number of resulting valid districtings, and certainly, it is still enormous. Impressively, these considerations are all immaterial to our very general method.

Applying the $\sqrt{\varepsilon}$ test involves the choice of a label function $\omega(\sigma)$, which assigns a real number to each districting. We have conducted runs using two label functions: ω_{var} is the (negative) variance of the proportion of Democrats in each district of the districting (as measured by 2012 presidential votes), and ω_{MM} is the difference between the median and mean of the proportions of Democrats in each district; ω_{MM} is motivated by the fact that this metric has a long history of use in gerrymandering and is directly tied to the goals of gerrymandering, whereas the use of the variance is motivated by the fact that it can change quickly with small changes in districtings. These two choices are discussed further in *SI Text*, but an important point is that our use of these label functions is not based on an assumption that small values of ω_{var} or ω_{MM} directly imply gerrymandering. Instead, because Theorem 1.1 is valid for any fixed label function, these labels are tools used to show significance, which are chosen because they are simple and natural functions on vectors that can be quickly computed, seem likely to be different for typical versus gerrymandered districtings, and have the potential to change relatively quickly with small changes in districtings. For the various notions of valid districtings that we considered, the $\sqrt{\varepsilon}$ test showed significance at p values in the range from 10^{-4} to 10^{-5} for the ω_{MM} label function and the range from 10^{-4} to 10^{-7} for the ω_{var} label function (see Fig. S1 and Table S1).

As noted earlier, the $\sqrt{\varepsilon}$ test can easily be used with more complicated Markov chains that capture more intricate definitions of the set of valid districtings. For example, the current districting of Pennsylvania splits fewer rural counties than the districting in Fig. 2, *Right*, and the number of county splits is one of many metrics for valid districtings considered by the Markov chains developed in ref. 19. Indeed, our test will be of particular value in cases where complex notions of what constitute valid districting slow the chain to make the heuristic mixing assumption particularly questionable. Regarding mixing time, even our chain with relatively weak constraints on the districtings (and very fast running time in implementation) seems to mix too slowly to sample π, even heuristically; in Fig. 2, we see that several districts still seem to have not left their general position from the initial districting, even after 2^{40} steps.

On the same note, it should also be kept in mind that, although our result gives a method to rigorously disprove that a given districting is unbiased—e.g., to show that the districting is unusual among districtings X_0 distributed according to the stationary distribution π—it does so without giving a method to sample from the stationary distribution. In particular, our method cannot answer the question of how many seats Republicans and Democrats should have in a typical districting of Pennsylvania, because we are still not mixing the chain. Instead, Theorem 1.1 has given us a way to disprove $X_0 \sim \pi$ without sampling π.

Proof of Theorem 1.1

We let π denote any stationary distribution for $\mathcal{M}$ and suppose that the initial state X_0 is distributed as $X_0 \sim \pi$, so that in fact, $X_i \sim \pi$ for all i. We say σ_j is ℓ-small among $\sigma_0, \dots, \sigma_k$ if there are, at most, ℓ indices $i \neq j$ among $0, \dots, k$, such that the label of σ_i is, at most, the label of σ_j. In particular, σ_j is 0-small among $\sigma_0, \sigma_1, \dots, \sigma_k$ when its label is the unique minimum label, and we encourage readers to focus on this $\ell = 0$ case in their first reading of the proof.

For $0 \leq j \leq k$, we define

$$\rho^k_{j,\ell} := \mathbf{Pr}\left(X_j \text{ is } \ell\text{-small among } X_0, \dots, X_k\right)$$

$$\rho^k_{j,\ell}(\sigma) := \mathbf{Pr}(X_j \text{ is } \ell\text{-small among } X_0, \dots, X_k \mid X_j = \sigma).$$

Observe that, because $X_s \sim \pi$ for all s, we also have that

$$\rho^k_{j,\ell}(\sigma) = \mathbf{Pr}\left(X_{s+j} \text{ is } \ell\text{-small among } X_s, \dots, X_{s+k} \mid X_{s+j} = \sigma\right). \quad [1]$$

We begin by noting two easy facts.

Observation 4.1.

$$\rho^k_{j,\ell}(\sigma) = \rho^k_{k-j,\ell}(\sigma).$$

Proof. Because $\mathcal{M} = X_0, X_1, \dots$ is stationary and reversible, the probability that $(X_0, \dots, X_k) = (\sigma_0, \dots, \sigma_k)$ is equal to the probability that $(X_0, \dots, X_k) = (\sigma_k, \dots, \sigma_0)$ for any fixed sequence $(\sigma_0, \dots, \sigma_k)$. Thus, any sequence $(\sigma_0, \dots, \sigma_k)$ for which $\sigma_j = \sigma$ and σ_j is a ℓ-small corresponds to an equiprobable sequence $(\sigma_k, \dots, \sigma_0)$, for which $\sigma_{k-j} = \sigma$ and σ_{k-j} is ℓ-small. □

Observation 4.2.

$$\rho^k_{j,2\ell}(\sigma) \geq \rho^j_{j,\ell}(\sigma) \cdot \rho^{k-j}_{0,\ell}(\sigma).$$

Proof. Consider the events that X_j is an ℓ-small among $X_0, \dots, X_j$ and among $X_j, \dots, X_k$. These events are conditionally independent when conditioning on the value of $X_j = \sigma$, and $\rho^j_{j,\ell}(\sigma)$ gives the probability of the first of these events, whereas applying Eq. **1** with $s = j$ gives that $\rho^{k-j}_{0,\ell}(\sigma)$ gives the probability of the second event.

Finally, when both of these events happen, we have that X_j is 2ℓ-small among $X_0, \dots, X_k$. □

We can now deduce that

$$\rho^k_{j,2\ell}(\sigma) \geq \rho^j_{j,\ell}(\sigma) \cdot \rho^{k-j}_{0,\ell}(\sigma) = \rho^j_{0,\ell}(\sigma) \cdot \rho^{k-j}_{0,\ell}(\sigma) \geq \left(\rho^k_{0,\ell}(\sigma)\right)^2. \quad [2]$$

Indeed, the first inequality follows from Observation 4.2, the equality follows from Observation 4.1, and the final inequality follows from the fact that $\rho^k_{j,\ell}(\sigma)$ is monotone nonincreasing in k for fixed j, ℓ, σ.

Observe now that $\rho^k_{j,\ell} = E\,\rho^k_{j,\ell}(X_j)$, where the expectation is taken over the random choice of $X_j \sim \pi$.

Thus, taking expectations in Eq. **2**, we find that

$$\rho^k_{j,2\ell} = \mathbf{E}\rho^k_{j,2\ell}(\sigma) \geq \mathbf{E}\left(\left(\rho^k_{0,\ell}(\sigma)\right)^2\right) \geq \left(\mathbf{E}\rho^k_{0,\ell}(\sigma)\right)^2 = \left(\rho^k_{0,\ell}\right)^2, \quad [3]$$

where the second of the two inequalities is the Cauchy–Schwartz inequality.

For the final step in the proof, we sum the left- and right-hand sides of Eq. **3** to obtain

$$\sum_{j=0}^{k} \rho^k_{j,2\ell} \geq (k+1)\left(\rho^k_{0,\ell}\right)^2.$$

If we let ξ_j $(0 \le i \le k)$ be the indicator variable that is one whenever X_j is 2ℓ-small among $X_0, \ldots, X_k$, then $\sum_{j=0}^{k} \xi_j$ is the number of 2ℓ-small terms, which is always, at most, $2\ell+1$. Therefore, linearity of expectation gives that

$$2\ell + 1 \ge (k+1)(\rho_{0,\ell}^{k})^2, \qquad [4]$$

giving that

$$\rho_{0,\ell}^{k} \le \sqrt{\frac{2\ell+1}{k+1}}. \qquad [5]$$

Theorem 1.1 follows, because if X_i is an ε-outlier among $X_0, \ldots, X_k$, then X_i is necessarily ℓ-small among $X_0, \ldots, X_k$ for $\ell = \lfloor \varepsilon(k+1) - 1 \rfloor \le \varepsilon(k+1) - 1$, and then, we have $2\ell + 1 \le 2\varepsilon(k+1) - 1 \le 2\varepsilon(k+1)$. □

ACKNOWLEDGMENTS. We thank John Nagle, Danny Sleator, and Dan Zuckerman for helpful conversations. M.C. was supported, in part, by NIH Grants 1R03MH10900901A1 and U545U54HG00854003. A.F. was supported, in part, by National Science Foundation (NSF) Grants DMS1362785 and CCF1522984 and Simons Foundation Grant 333329. W.P. was supported, in part, by NSF Grant DMS-1363136 and the Sloan Foundation.

1. Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. *Stat Sci* 7(4):457–472.
2. Gelman A, Rubin DB (1992) A single series from the Gibbs sampler provides a false sense of security. *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*, eds Bernardo JM, Berger JO, Dawid AP, Smith AFM (Clarendon, Gloucestershire, UK), pp 625–631.
3. Swendsen RH, Wang JS (1987) Nonuniversal critical dynamics in Monte Carlo simulations. *Phys Rev Lett* 58(2):86–88.
4. Borgs C, Chayes J, Tetali P (2012) Tight bounds for mixing of the Swendsen–Wang algorithm at the Potts transition point. *Probab Theory Relat Fields* 152(3-4):509–557.
5. Cooper C, Frieze A (1999) Mixing properties of the Swendsen-Wang process on classes of graphs. *Random Struct Algorithms* 15(3-4):242–261.
6. Gore VK, Jerrum MR (1999) The Swendsen–Wang process does not always mix rapidly. *J Stat Phys* 97(1-2):67–86.
7. Propp JG, Wilson DB (1996) Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Struct Algorithms* 9(1-2):223–252.
8. Propp J, Wilson D (1998) Coupling from the past: A user's guide. *Microsurveys in Discrete Probability*, eds Aldous DJ, Propp J (American Mathematical Society, Providence, RI), Vol 41, pp 181–192.
9. Fill JA (1997) An interruptible algorithm for perfect sampling via Markov chains. Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing (Association for Computing Machinery, New York), pp 688–695.
10. Huber M (2004) Perfect sampling using bounding chains. *Ann Appl Probab* 14(2): 734–753.
11. Wang SS-H (2016) Three tests for practical evaluation of partisan gerrymandering (December 28, 2015). *Stanford Law Rev* 68:1263–1321.
12. Nagle JF (2015) Measures of partisan bias for legislating fair elections. *Elect Law J* 14(4):346–360.
13. McDonald MD, Best RE (2015) Unfair partisan gerrymanders in politics and law: A diagnostic applied to six cases. *Elect Law J* 14(4):312–330.
14. Chen J, Rodden J (2013) Unintentional gerrymandering: Political geography and electoral bias in legislatures. *Quart J Polit Sci* 8(3):239–269.
15. Cirincione C, Darling TA, O'Rourke TG (2000) Assessing South Carolina's 1990s Congressional districting. *Polit Geogr* 19(2):189–211.
16. Rogerson PA, Yang Z (1999) The effects of spatial population distributions and political districting on minority representation. *Soc Sci Comput Rev* 17(1):27–39.
17. Fifield B, Higgins M, Imai K, Tarr A (2015) *A New Automated Redistricting Simulator Using Markov Chain Monte Carlo. Working Paper. Technical Report*. Available at imai.princeton.edu/research/files/redist.pdf. Accessed October 21, 2016.
18. Chenyun Wu L, Xiaotian Dou J, Frieze A, Miller D, Sleator D (2015) Impartial redistricting: A Markov chain approach. arXiv:1510.03247.
19. Vaughn C, Bangia S, Dou B, Guo S, Mattingly J (2016) *Quantifying Gerrymandering@Duke*. Available at https://services.math.duke.edu/projects/gerrymandering. Accessed October 21, 2016.
20. Ansolabehere S, Palmer M, Lee A (2014) *Precinct-Level Election Data, Harvard Dataverse, v1*. Available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/21919. Accessed October 21, 2016.
21. Edgeworth FY (1897) Miscellaneous applications of the calculus of probabilities. *J R Stat Soc* 60(3):681–698.
22. Levin DA, Peres Y, Wilmer EL (2009) *Markov Chains and Mixing Times* (American Mathematical Society, Providence, RI).
23. Aldous D, Fill J (2002) *Reversible Markov Chains and Random Walks on Graphs*. Available at https://www.stat.berkeley.edu/∼aldous/RWG/book.html. Accessed October 21, 2016.
24. Frieze A, Karoński M (2015) *Introduction to Random Graphs* (Cambridge Univ Press, Cambridge, UK).

Supporting Information

Chikina et al. 10.1073/pnas.1617540114

SI Text

S1. Precinct Data. Precinct-level voting data and associated shape files were obtained from the Harvard Election Data Archive (projects.iq.harvard.edu/eda/home) (20). The data for Pennsylvania contain 9,256 precincts. The data were altered in two ways: 258 precincts that were contained within another precinct were merged and 79 precincts that were not contiguous were split into continuous areas, with voting and population data distributed proportional to the area. The result was a set of 9,060 precincts. All geometry calculations and manipulations were accomplished in R with "maptools," "rgeos," and "BARD" R packages. The final input to the Markov chain is a set of precincts with corresponding areas, neighbor precincts, the length of the perimeter shared with each neighbor, voting data from 2012, and the current Congressional district to which the precinct belongs.

S2. Valid Districting. We restrict our attention to districtings satisfying four restrictions, each of which we describe here.

S2.1. Contiguous districts. A valid districting must have the property that each of its districts is contiguous. In particular, two precincts are considered adjacent if the length of their shared perimeter is nonzero (in particular, precincts meeting only at a point are not adjacent), and a district is contiguous if any pair of precincts is joined by a sequence of consecutively adjacent pairs.

S2.2. Simply connected districts. A valid districting must have the property that each of its districts is simply connected. Roughly speaking, this constraint means the district cannot have a "hole." Precisely, a district is simply connected if, for any circuit of precincts in the district, all precincts in the region bounded by the circuit also are in the district.

Apart from aesthetic reasons for insisting that districtings satisfy this condition, there is also a practical reason: it is easier to have a fast local check for contiguity when one is also maintaining that districtings are simply connected.

S2.3. Small population difference. According to the "one person, one vote" doctrine, Congressional districts for a state are required to be roughly equal in population. In the current districting of Pennsylvania, for example, the maximum difference in district population from the average population is roughly 1%. Our chain can use different tolerances for population difference between districts and the average, and the tolerances used in the runs below are indicated.

S2.4. Compactness. If districtings were drawn randomly with only the first three requirements, the result would be districtings in which districts have very complicated, fractal-like structure (because most districtings have this property). The final requirement on valid districtings prevents this by ensuring that the districts in the districting have a reasonably nice shape. This requirement on district shape is commonly termed "compactness" and explicitly required of Congressional districts by the Pennsylvania Constitution.

Although compactness of districts does not have a precise legal definition, various precise metrics have been proposed to quantify the compactness of a given district mathematically. One of the simplest and most commonly used metrics is the Polsby–Popper metric, which defines the compactness of a district as

$$C_D = \frac{4\pi A_D}{P_D^2},$$

where A_D and P_D are the area and perimeter of the district D, respectively. Note that the maximum value of this measure is one, which is achieved only by the disk as a result of the isoperimetric inequality. All other shapes have compactness between zero and one, and smaller values indicate more "contorted" shapes.

Perhaps the most straightforward use of this compactness measure is to enforce some threshold on compactness and require valid districtings to have districts with compactness that is above that lower bound. (For consistency with our other metrics, we actually impose an upper bound on the reciprocal $1/C_D$ of the Polsby–Popper compactness C_D of each district D.) In Table S1, this metric is the L^∞ compactness metric.

One drawback of using this method is that the current districting of Pennsylvania has a few districts that have very low compactness values (they are much stranger looking than the other districts). Applying this restriction will allow all 18 districts to be as bad as the threshold chosen, so that, in particular, we will be sampling districtings from space in which all 18 districts may be as bad as the worst district in the current plan. In fact, because there are more noncompact regions than compact ones, one expects that, in typical such districting, all 18 districts would be nearly as bad as the currently worst example.

To address this issue and also, to show the robustness of our finding for the districting question, we also consider some alternate restrictions on the districting, which measure how reasonable the districting as a whole is with regard to compactness. For example, one simple measure of this is to have a threshold for the maximum allowable sum

$$\frac{1}{C_1} + \cdots + \frac{1}{C_{18}}$$

of the inverse compactness values of 18 districts. This metric is the L^1 metric in Table S1. Similarly, we could have a threshold for the maximum allowable sum of squares

$$\frac{1}{C_1^2} + \cdots + \frac{1}{C_{18}^2}.$$

This metric is the L^2 metric in Table S1. Finally, we can have a simple condition that simply ensures that the total perimeter

$$P_1 + \cdots + P_{18}$$

is less than some threshold.

S2.5. Other possible constraints. It is possible to imagine many other reasonable constraints on valid districtings. For example, the Pennsylvania Constitution currently requires of districtings for the Pennsylvania Senate and Pennsylvania House of Representatives that, unless absolutely necessary, no county, city, incorporated town, borough, township, or ward shall be divided in forming either a senatorial or representative district.

There is no similar requirement for US Congressional districts in Pennsylvania, which is what we consider here, but it is still a reasonable constraint to consider.

There are also interesting legal questions about the extent to which majority–minority districts (in which an ethnic minority is an electoral majority) are either required to be intentionally created or forbidden to be intentionally created. On the one hand, the US Supreme Court ruled in Thornburg v. Gingles (1986) that, in certain cases, a geographically concentrated minority population is entitled to a Congressional district in which it constitutes a majority. On the other hand, in several US Supreme Court cases [Shaw v. Reno (1993), Miller v. Johnson (1995), and Bush v. Vera (1996)], Congressional districtings were thrown out, because they contained intentionally drawn majority–minority districts that were deemed to be a "racial gerrymander." In any case, we have not attempted to answer the question of whether or

how the existence of majority–minority districts should be quantified. (We suspect that the unbiased procedure of drawing a random districting is probably acceptable under current majority–minority district requirements, but in any case, our main intent is to show the application of the $\sqrt{\varepsilon}$ test.)

Importantly, we emphasize that any constraint on districtings that can be precisely defined (i.e., by giving an algorithm that can identify whether a districting satisfies the constraint) can be used in the Markov chain setting in principle.

S3. The Markov Chain. The Markov chain $\mathcal{M}$ that we use has as its state space Σ, the space of all valid districtings (with 18 districts) of Pennsylvania. Note that there is no simple way to enumerate these, and there is an enormous number of them.

A simple way to define a Markov chain on this state space is to transition as follows.

i) From the current state, determine the set S of all pairs (ρ, D), where ρ is a precinct in some district D_ρ, and $D \neq D_\rho$ is a district that is adjacent to ρ. Let N_S denote the size of this set.
ii) From S, choose one pair (ρ_0, D_0) uniformly at random.
iii) Change the district membership of ρ_0 from D_{ρ_0} to D_0 if the resulting district is still valid.

Let the Markov chain with these transition rules be denoted by $\mathcal{M}_0$. This chain is a perfectly fine reversible Markov chain to which our theorem applies, but the uniform distribution on valid districtings is not stationary for $\mathcal{M}_0$; therefore, we cannot use $\mathcal{M}_0$ to make comparisons between a presented districting and a uniformly random valid districting.

A simple way to make the uniform distribution stationary is to "regularize" the chain (that is, to modify the chain so that the number of legal steps from any state is equal). ($\mathcal{M}_0$ is not already regular, because the number of precincts on the boundaries of districts will vary from districting to districting.) We do this by adding loops to each possible state. In particular, using a theoretical maximum $N_{\max}$ on the possible size of N_S for any district, we modify the transition rules as follows.

i) From the current state, determine the set S of all pairs (ρ, D), where ρ is a precinct in some district D_ρ, and $D \neq D_\rho$ is a district that is adjacent to ρ. Let N_S denote the size of this set.
ii) With probability $1 - \frac{N_S}{N_{\max}}$, remain in the current state for this step. With probability $\frac{N_S}{N_{\max}}$, continue as follows.
iii) From S, choose one pair (ρ_0, D_0) uniformly at random.
iv) Change the district membership of ρ_0 from D_{ρ_0} to D_0 if the resulting district is still valid. If it is not, remain in the current district for this set.

In particular, with this modification, each state has exactly $N_{\max}$ possible transitions, which are each equally likely; many of these transitions are loops back to the same state. (Some of these loops arise from step *ii*, but some also arise when the if condition in step *iv* fails.)

S4. The Label Function. In principle, any label function ω could be used in the application of the $\sqrt{\varepsilon}$ test; note that Theorem 1.1 places no restrictions on ω. Thus, when we choose which label function to use, we are making a choice based on what is likely to achieve good significance rather than what is valid statistical reasoning (subject to the caveat discussed below). To choose a label function that was likely to allow good statistical power, we want to have a function that is

i) likely very different for a gerrymandered districting compared with a typical districting and
ii) sensitive enough that small changes in the districting might be detected in the label function.

Although the role of the first condition in achieving outlier status is immediately obvious, the second property is also crucial to detecting significance with our test, which makes use of trajectories that may be quite small compared with the mixing time of the chain. For the $\sqrt{\varepsilon}$ test to succeed at showing significance, it is not enough for the presented state σ_0 to actually be an outlier against π with respect to ω; this outlier status must also be detectable on trajectories of the fixed length k, which may well be too small to mix the chain. This second property discourages the use of "coarse-grained" label functions, such as the number of seats of 18 that the Democrats would hold with the districting in question, because many swaps would be needed to shift a representative from one party to another.

We considered two label functions for our experiments (each selected with the above desired properties in mind) to show the robustness of our framework. The first label function ω_{var} that we used is simply the negative of the variance in the proportions of Democrat voters among the districts. Thus, given a districting σ, $\omega_{\text{var}}(\sigma)$ is calculated as

$$\omega_{\text{var}}(\sigma) = -\left(\frac{\delta_1^2 + \delta_2^2 + \cdots + \delta_{18}^2}{18} - \left(\frac{\delta_1 + \delta_2 + \cdots + \delta_{18}}{18}\right)^2\right),$$

where for each $i = 1, \ldots, 18$, δ_i is the fraction of voters in that district that voted for the Democrat presidential candidate in 2012. We suspect that the variance is a good label function from the standpoint of the first characteristic listed above but a great label function from the standpoint of the second characteristic. Recall that, in practice, gerrymandering is accomplished by packing the voters of one party into a few districts, in which they make up an overwhelming majority. This technique, naturally, results in a high-variance vector of party proportions in the districts. However, high-variance districtings can exist that do not favor either party (note, for example, that the variance is symmetric with respect to Democrats and Republicans, ignoring third-party affiliations). Thus, for a districting that is biased against π because of a partisan gerrymander to "stand out" as an outlier, it must have especially high variance. In particular, statistical significance will be weaker than it might be for a label function that is more strongly correlated with partisan gerrymandering. However, ω_{var} can detect very small changes in the districting, because essentially, every swap will either increase or decrease the variance. Indeed, for the run of the chain corresponding to the L^∞ constraint (SI Text, Runs of the Chain), $\omega_{\text{var}}(X_0)$ was strictly greater than $\omega_{\text{var}}(X_i)$ for the entire trajectory ($1 \leq i \leq 2^{40}$). That is, for the L^∞ constraint, the current districting of Pennsylvania was the absolute worst districting seen according to ω_{var} among the more than 1 trillion districtings on the trajectory.

The second label function that we considered is calculated simply as the difference between the median and the mean of the ratios $\delta_1, \ldots, \delta_{18}$. This simple metric, called the "Symmetry Vote Bias" by McDonald and Best (13) and denoted as ω_{MM} by us, is closely tied to the goal of partisan gerrymandering. As a simple illustration of the connection, we consider the case where the median of the ratios $\delta_1, \ldots, \delta_{18}$ is close to $\frac{1}{2}$. In this case, the mean of the δ_i tracks the fraction of votes that the reference party wins to win one-half of the seats. Thus, a positive Symmetry Vote Bias corresponds to an advantage for the reference party, whereas a negative Symmetry Vote Bias corresponds to a disadvantage. The use of the Symmetry Vote Bias in evaluating districtings dates at least to the 19th century work of Edgeworth (21). These features make it an excellent candidate from the standpoint of our first criterion: gerrymandering is very likely to be reflected in outlier values of ω_{MM}.

However, ω_{MM} is a rather slow-changing function compared with ω_{var}. Indeed, observe that, in the calculation

$$\text{Symmetry Vote Bias} = \text{median} - \text{mean},$$

the mean is essentially fixed, so that changes in ω_{MM} depend on changes in the median. In initial changes to the districting, only changes to the 2 districtings giving rise to the median (2 because 18 is even) can have a significant impact on ω_{MM}. (However, changes to any district directly affect ω_{var}.)

It is likely possible to make better choices for the label function ω to achieve better significance. For example, the metric B_G described by Nagle (12) seems likely to be excellent from the standpoints of conditions *i* and *ii* simultaneously. However, we have restricted ourselves to the simple choices of ω_{var} and ω_{MM} to clearly show our method and make it obvious that we have not tried many labeling functions before finding some that worked (in which case, a multiple hypothesis test would be required).

One point to keep in mind is that, often when applying the $\sqrt{\varepsilon}$ test—including in this application to gerrymandering—we will wish to apply the test and thus, need to define a label function after the presented state σ_0 is already known. In these cases, care must be taken to choose a "canonical" label function ω, so that there is no concern that ω was carefully crafted in response to σ_0 (in this case, a multiple hypothesis correction would be required for the various possible ω values that could have been crafted depending on σ_0); ω_{var} and ω_{MM} are excellent choices from this perspective: the variance is a common and natural function on vectors, and the Symmetry Vote Bias has an established history in the evaluation of gerrymandering (and in particular, a history that predates the present districting of Pennsylvania).

S5. Runs of the Chain. In Table S1, we give the results of eight runs of the chain under various conditions. Each run was for $k = 2^{40}$ steps. Code and input data for our Markov chain are available at the website of W.P. (math.cmu.edu/∼wes).

Generally, after an initial "burn-in" period, we expect the chain to (almost) never again see states as unusual as the current districting of Pennsylvania, which means that we expect the test to show significance proportional to the inverse of the square root of the number of steps (i.e., the p value at 2^{42} steps should be one-half the p value at 2^{40} steps). In particular, for the L^1, L^2, and L^∞ constraints, these runs never revisited states as bad as the initial state after 2^{21} steps for the ω_{MM} label and after 2^6 steps for the ω_{var} label. Note that this agrees with our guess that ω_{var} had the potential to change more quickly than ω_{MM}. The perimeter constraint did revisit enough states as bad as the given state with respect to the ω_{var} label to adversely affect its p value with respect to the ω_{var} label. This observation may reflect our guess that the ω_{var} label is worse than the ω_{MM} label in terms of how easily it can distinguish gerrymandered districtings from random ones.

The parameters for the first row were used for Fig. 2.

One quick point is that, although we have experimented here with different compactness measures, there is no problem of multiple hypothesis correction to worry about, because every run that we encountered produces strong significance for the bias of the initial districting. The point of experimenting with the notion of compactness is to show that this is a robust framework and that the finding is unlikely to be sensitive to minor disagreements over the proper definition of the set of valid districtings.

S6. An Example Where $p \approx \sqrt{\varepsilon}$ Is Best Possible. It might be natural to suspect that observing ε-outlier status for σ on a random trajectory from σ is significant at something like $p \approx \varepsilon$ instead of the significance $p \approx \sqrt{\varepsilon}$ established by Theorem 1.1. However, because Theorem 1.1 places no demand on the mixing rate of $\mathcal{M}$, it should instead seem remarkable that any significance can be shown in general, and indeed, we show by example in this section that significance at $p \approx \sqrt{\varepsilon}$ is essentially best possible.

Let N be some large integer. We let $\mathcal{M}$ be the Markov chain where X_0 is distributed uniformly in $[0, 1, 2, \ldots, N-1]$, and for any $i \geq 1$, X_i is equal to $X_{i-1} + \zeta_i$ computed modulo N, where ζ_i is 1 or -1 with probability $1/2$. Note that the chain is stationary and reversible.

If N is chosen large relative to k, then with probability arbitrarily close to one the value of X_0 is at distance greater than k from zero (in both directions). Conditioning on this event, we have that X_0 is minimum among $X_0, \ldots, X_k$ if and only if all of the partial sums $\sum_{i=1}^{j} \zeta_i$ are positive. The probability of this event is just the probability that a k-step 1D random walk from the origin takes a first step to the right and does not return to the origin. The calculation of this probability is a classical problem in random walks, which can be solved using the reflection principle.

In particular, for k even, the probability is given by

$$\frac{1}{2^{k+1}}\binom{k}{k/2} \sim \frac{1}{\sqrt{2\pi k}}.$$

Because being the minimum of $X_0, \ldots, X_k$ corresponds to being an ε-outlier for $\varepsilon = 1/k + 1$, this example shows that the probability of being an ε-outlier can be as high as $\sqrt{\varepsilon/2\pi}$.

The best possible value of the constant in the $\sqrt{\varepsilon}$ test seems to be an interesting problem for future work.

S7. Notes on Statistical Power. The effectiveness of the $\sqrt{\varepsilon}$ test depends on the availability of a good choice for ω and the ability to run the test for long enough (in other words, choose k large enough) to detect that the presented state is a local outlier.

It is possible, however, to make a general statement about the power of the test when k is chosen large relative to the actual mixing time of the chain. Recall that one potential application of the test is in situations where the mixing time of the chain is actually accessible through reasonable computational resources, although this fact cannot be proved rigorously, because theoretical bounds on the mixing time are not available. In particular, we do know that the test is very likely to succeed when k is sufficiently large and $\omega(\sigma_0)$ is atypical.

Theorem S1. Let $\mathcal{M}$ be a reversible Markov chain on Σ, and let $\omega : \Sigma \to \mathbb{R}$ be arbitrary. Given σ_0, suppose that, for a random state $\sigma \sim \pi$, $\mathbf{Pr}(\omega(\sigma) \leq \omega(\sigma_0)) \leq \varepsilon$. Then, with probability at least

$$\rho \geq 1 - \left(1 + \frac{\varepsilon k}{10\tau_2}\right)\frac{1}{\sqrt{\pi_{\min}}}\exp\left(\frac{-\varepsilon^2 k}{20\tau_2}\right),$$

$\omega(\sigma)$ is an 2ε-outlier among $\omega(\sigma_0), \omega(\sigma_1), \ldots, \omega(\sigma_k)$, where $\sigma_0, \sigma_1, \ldots$ is a random trajectory starting from σ_0.

Here, τ_2 is the relaxation time for $\mathcal{M}$ defined as $1/(1-\lambda_2)$, where λ_2 is the second eigenvalue of $\mathcal{M}$. τ_2 is thus the inverse of the spectral gap for $\mathcal{M}$ and intimately related to the mixing time of $\mathcal{M}$ (22–24). The probability ρ in Theorem S1 converges exponentially quickly to 1 and moreover, is very close to one after k is large relative to τ_2. In particular, Theorem S1 shows that the $\sqrt{\varepsilon}$ test will work when the test is run for long enough. Of course, one strength of the $\sqrt{\varepsilon}$ test is that it can sometimes show bias, even when k is far too small to mix the chain, which is almost certainly the case for our application to gerrymandering. When these short-k runs are successful at detecting bias is, of course, dependent on the relationship of the presented state σ_0 and its local neighborhood in the chain.

Theorem S1 is an application of the following theorem of Gillman.

Theorem S2. Let $\mathcal{M} = X_0, X_1, \ldots$ be a reversible Markov chain on Σ, let $A \subset \Sigma$, and let $N_k(A)$ denote the number of visits to A among $X_0, \ldots, X_k$. Then, for any $\gamma > 0$,

$$\mathbf{Pr}(N_k(A)/n - \pi(A) > \gamma) \leq \left(1 + \frac{\gamma n}{10\tau_2}\right)\sqrt{\sum_{\sigma}\frac{\mathbf{Pr}(X_0 = \sigma)^2}{\pi(\sigma)}} \times \exp\left(\frac{-\gamma^2 n}{20\tau_2}\right).$$

Proof of Theorem S1. We apply Theorem S2, with A as the set of states $\sigma \in \Sigma$, such that $\omega(\sigma) \leq \omega(\sigma_0)$, $X_0 = \sigma_0$, and $\gamma = \varepsilon$. By assumption, $\pi(A) \leq \varepsilon$, and Theorem S2 gives that

$$\mathbf{Pr}(N_k(A)/k > 2\varepsilon) \leq \left(1 + \frac{\varepsilon k}{10\tau_2}\right)\sqrt{\frac{1}{\pi_{\min}}}\exp\left(\frac{-\varepsilon^2 k}{20\tau_2}\right).$$

□

S8. A Result for Small Variation Distance. In this section, we give a corollary of Theorem 1.1 that applies to the setting where X_0 is not distributed as a stationary distribution π but instead, is distributed with small total variation distance to π.

The total variation distance $||\rho_1 - \rho_2||_{TV}$ between probability distributions ρ_1, ρ_2 on a probability space Ω is defined to be

$$||\rho_1 - \rho_2||_{TV} := \sup_{E \subseteq \Omega} |\rho_1(E) - \rho_2(E)|. \qquad \textbf{[S1]}$$

Corollary S1. Let $\mathcal{M} = X_0, X_1, \ldots$ be a reversible Markov chain with a stationary distribution π, and suppose that the states of $\mathcal{M}$ have real-valued labels. If $||X_0 - \pi||_{TV} \leq \varepsilon_1$, then for any fixed k, the probability that the label of X_0 is an ε-outlier from among the list of labels observed in the trajectory $X_0, X_1, X_2, \ldots, X_k$ is, at most, $\sqrt{2\varepsilon} + \varepsilon_1$.

The theorem is intuitively clear; we provide a formal proof below for completeness.

Proof. If ρ_1, ρ_2, and τ are probability distributions, then we have that the product distributions (ρ_1, τ) and (ρ_2, τ) satisfy

$$||(\rho_1, \tau) - (\rho_2, \tau)||_{TV} = ||\rho_1 - \rho_2||_{\text{TV}}. \qquad \textbf{[S2]}$$

Our plan now is to split the randomness in the trajectory $X_0, \ldots, X_k$ of the Markov chain into two independent sources: the initial distribution is $X_0 \sim \rho$, and τ is the uniform distribution on sequences of length k of real numbers $r_1, r_2, \ldots, r_k$ in $[0, 1]$. We can view the distribution of the trajectory $X_0, X_1, \ldots, X_k$ as the product (ρ, τ) by using sequences of reals $r_1, \ldots, r_k$ to choose transitions in the chain; from $X_i = \sigma_i$, if there are L transitions possible, with probabilities $p_1, \ldots, p_L$. Then, we make the tth possible transition if $r_i \in [p_1 + \cdots + p_{t-1}, p_1 + \cdots + p_{t-1} + p_t)$.

Now we have from Eq. **S2** that, if $||\rho - \pi||_{TV} \leq \varepsilon_1$, then $||(\rho, \tau) - (\pi, \tau)||_{TV} \leq \varepsilon_1$. Therefore, any event that would happen with probability at most p for the sequence $X_0, \ldots, X_k$ when $X_0 \sim \pi$ must happen with probability at most $p + \varepsilon_1$ when $X_0 \sim \rho$, where $||\rho - \pi||_{TV} \leq \varepsilon_1$. The corollary follows. □

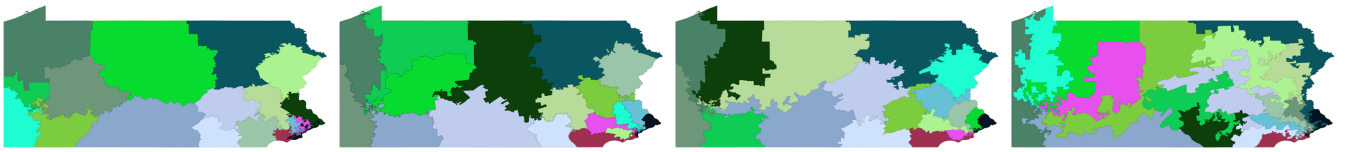

Fig. S1. The last state from each of the above runs of the chain (perimeter, L^1, L^2, and L^∞, respectively). Note that the L^∞ districting is quite ugly; with this notion of validity, every district among the 18 is allowed to be as noncompact as the worst district in the current Pennsylvania districting. The perimeter constraint produces a districting that appears clean at a large scale but allows rather messy city districts, because they contribute only moderately to the perimeter anyway. The L^1 and L^2 constraints are more balanced. Note that none of these districtings should be expected to be geometrically "nicer" than the current districting of Pennsylvania. Indeed, the point of our Markov chain framework is to compare the present districting of Pennsylvania with other "just as bad" districtings to observe that, even among this set, the present districting is atypical.

Table S1. Runs of the redistricting Markov chain with results of the $\sqrt{\varepsilon}$ test

Population threshold, %	Compactness measure	Compactness threshold	Initial value	(Steps) $k =$	Label function	ε-Outlier at $\varepsilon =$	Significant at $p =$
2	Perimeter	≤ 125	121.2 . . .	2^{40}	ω_{var}	$3.0974 \cdot 10^{-8}$	$2.4889 \cdot 10^{-4}$
					ω_{MM}	$5.7448 \cdot 10^{-10}$	$3.3896 \cdot 10^{-5}$
2	L^1	≤ 160	156.4 . . .	2^{40}	ω_{var}	$5.0123 \cdot 10^{-11}$	$1.0012 \cdot 10^{-5}$
					ω_{MM}	$5.6936 \cdot 10^{-10}$	$3.3745 \cdot 10^{-5}$
2	L^2	≤ 44	43.06 . . .	2^{40}	ω_{var}	$8.2249 \cdot 10^{-11}$	$1.2826 \cdot 10^{-5}$
					ω_{MM}	$6.8038 \cdot 10^{-10}$	$3.6888 \cdot 10^{-5}$
2	L^∞	≤ 25	24.73 . . .	2^{40}	ω_{var}	$3.3188 \cdot 10^{-13}$	$8.1472 \cdot 10^{-7}$
					ω_{MM}	$6.9485 \cdot 10^{-8}$	$3.7279 \cdot 10^{-4}$