new working title:
Repetition effects in a sequential Monte Carlo sampler

Observations about SMC for Graph Partitioning

Sarah Cannon, Daryl DeFord, and Moon Duchin

December 20, 2023

1 Introduction

In this note, we consider the structure of the SMC (Sequential Monte Carlo) method developed by McCartan–Imai for sampling partitions of a fixed graph G into a given number k of connected subgraphs with nearly equal total node weight [9]. SMC operates by fixing a sample size S and a number of pieces k, then creating a top generation of partial plans by marking off one connected subgraph of G with roughly $1/k$ of the node weight (to be thought of as a "district") in each plan. In the next generation, S agents sample with replacement from those partial plans according to a weighting function, marking off a second subgraph of about the same size. This process continues for k generations until the entire graph is marked. We will summarize the salient combinatorial features of this scheme in a *descendancy diagram* (Figure 1), showing only the paths that connect from the bottom layer to the top. The nodes involved in these paths are called *active* or *activated*, we will count *top-level ancestors*, which are the active nodes in the highest layer from which all bottom nodes are descended. The final generation consists of S complete partitions.

SMC samples face a certain amount of characteristic redundancy. As the authors note, "because the SMC algorithm involves repeated resampling with replacement, for a finite number of samples it can suffer from particle system collapse [8], where many of the sampled plans share a small number of common districts (which originate as common ancestor particles in the SMC scheme)." One aim of this note is to study this concentration of ancestry in combinatorial terms.

The second aim is to understand the tradeoffs faced by users in the choice of sampler. Citing McCartan–Imai again: "Like MCMC algorithms, the SMC algorithm generates samples which approximate the target distribution arbitrarily well as the sample size increases." Our second set of questions explores the convergence guarantees available with each method and considers how those guarantees relate to the production of various data artifacts used in redistricting analysis, like histograms and boxplots.

We note that SMC for redistricting is already in widespread use in courts of law, and that the repetition of districts has been flagged as a limitation that undermines statistical claims about the sample. For instance, Kristopher Tapp produced an affidavit in New York state Senate litigation in which he attempted a replication of another expert's SMC ensemble of 5000 maps, and found that a certain set of 31 districts (covering about half of the state) appeared identically in over 64% of the sample.[1] In New Mexico state Senate litigation, a defense brief described the SMC method as being "plagued with duplicate simulations"; as a prophylactic measure, a defense expert cosmetically altered his SMC sample by perturbing the boundaries of districts so that he could that claim no districts were duplicated.[2] It is clear that rigorous attention to the issues around duplication and the consequences for statistical interpretation would be helpful in the field.

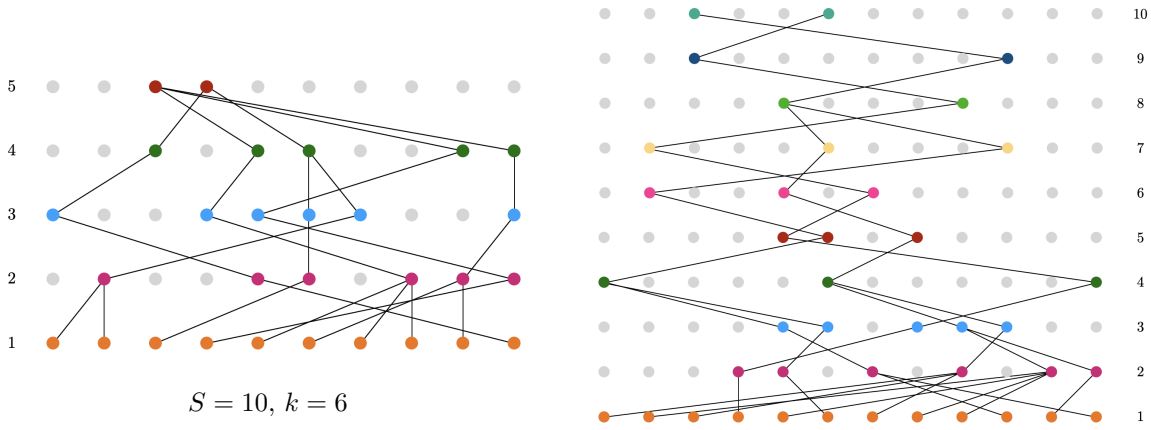
[1] Tapp further notes that while an ensemble of 5000 63-district maps can have up to 315,000 distinct districts, his replication ensemble had only 12,319, so that each district was repeated an average of 1360 times. He calls this "a head-turning level of redundancy." [2]

[2] "Dr. Chen's implementation of the MCMC version of an SMC algorithm [sic] did not result in any duplicated maps. [Exh. D, Dep. ST 54:17–55:17 (falsely testifying that Dr. Chen's simulations contain duplicates), 136:6–136:20 (correcting his mistaken testimony)]." Meanwhile the opposing expert opined in deposition that "Duplicates happen all the time... So it doesn't bother me, unless it gets extreme to where you end up having, like, 20 maps." [1]

1.1 Motivating questions

To understand the SMC algorithm for k districts, we will begin by studying *uniform descendancy diagrams* with levels (also called *layers* or *generations*) labeled $i = 1$ to $k-1$ from bottom to top, each of width S, in which each node chooses a parent uniformly at random from the generation above. Nodes in the top layer (indexed $k-1$) represent partial plans with a single initial district marked, and those at level i represent districting plans with $k-i$ districts marked. At level 1, $k-1$ districts are marked, which determines the kth and final district and amounts to specifying a complete plan. A node in a descendancy diagram is *active* if it has a descendent in the bottom layer. Calculations using descendancy diagrams will omit all non-active nodes, because they have no role in the final sample constructed by SMC.

If we write $D \in \mathcal{D}(S,k)$ for a specific diagram of this form, then let $A(D)$ be the number of active nodes at the top level (indexed $k-1$) in that diagram, and let $A(S,k)$ be the expected number of top-level ancestors over the uniform distribution on $\mathcal{D}(S,k)$. We consider the following questions.



$S = 10,\ k = 6$ $\qquad$ $S = 12,\ k = 11$

Figure 1: These two figures show structures we call *descendancy diagrams*. The bottom row is labeled as generation 1 in each case, increasing in index with each layer until generation $k-1$ at the top. Each of these two diagrams has $A(D) = 2$, meaning that there are two top-level ancestors from which all members of the bottom generation are descended.

Question 1 (Ancestor extinction)**.** *As a function of S and k, what is the distribution of the number of top-level ancestors $A(D)$ in a uniform descendancy diagram? Give bounds or asymptotics for the expectation $A(S,k)$.*

For node j at level i, let $d(i,j)$ be the number of descendants it has in the bottom level.

Question 2 (Extent of redundancy)**.** *For each fraction $0 < \varphi < 1$, what is the distribution over $\mathcal{D}$ of*

$$F(D,\varphi) = \min\{i : \exists\ j \text{ with } d(i,j) > \varphi \cdot S\}$$

in a uniform descendancy diagram? (i.e., the layer at which a single ancestor accounts for a given fraction φ of the final generation of nodes)

Then we consider the introduction of non-uniform weights.

Question 3 (Weighting)**.** *How do ancestry concentration and redundancy get more severe as the weighting factors deviate from uniform?*

Next, we broaden the scope and consider the overall quality of the sample of size S that consists of members of the final generation.

Question 4 (Convergence guarantees)**.** *What are the convergence guarantees for the sampling distribution obtained from the weak SMC Central Limit Theorem (Prop 4.2), and how do they differ from guarantees for ergodic Markov chain samplers?*

Question 5 (Boxplots)**.** *How do the convergence guarantees relate to the production of histograms, boxplots, and other percentile summary statistics?*

Acknowledgments

We thank Peter Winkler and Chris Hoffman for helpful conversations and pointers to the literature. We are grateful to Peter Rock for his excellent work conducting SMC experiments to support this project. We also thank Cory McCartan for generously sharing his time to explain SMC for redistricting, and the Redist code. This material is based upon work supported by the National Science Foundation under Grant No. DMS-1928930 and by the Alfred P. Sloan Foundation under grant G-2021-16778, while the authors were in residence at the Simons Laufer Mathematical Sciences Institute (formerly MSRI) in Berkeley, California, during the Fall 2023 semester. SC is also supported in part by NSF CCF-2104795; MD by NSF DMS-2005512.

2 Structure of descendancy diagrams

2.1 Setup

We start with some simple observations about this model. When we take a uniform descendancy diagram with two layers (corresponding to an SMC process with three districts), Question 1 is a rephrasing of the classic birthday problem from probability. As k grows, the generalized birthday problem also has a combinatorial interpretation as a sequential balls-in-bins model (see [11]) or via random coagulations (see [3]). Even in the language of ancestry, this has been studied in the context of genetic drift as the *Wright-Fisher Model* (see [5]). It is well known that the probability that two individuals' lineages remain distinct for at least i levels is $(1-1/S)^i$ and the probability that ℓ lineages remain (pairwise) distinct for at least i levels is $(1-1/S)^i(1-2/S)^i\dots(1-(\ell-1)/S)^i$. After renormalizing and sending $S\to\infty$, the time of the first (pairwise) coalescence among ℓ distinct lineages approaches an exponential distribution with rate $\ell(\ell-1)/(2S)$. There is a rich literature considering variations of this model, using it to design Markov chains, and extending it to infinite S.

Our setting differs slightly from the Wright-Fisher model by only considering lineages that extend to the bottom layer and discarding the others—that is, in the language developed above, we only track active nodes.

Lemma 1 (One-step probabilities)**.** *If a given generation i has $1\le t\le S$ active nodes, then the expected number of ancestors in the generation immediately above (generation $i+1$) is $S-S(1-\frac{1}{S})^t$. The probability that there are exactly v activated nodes in generation $i+1$ when there are t activated nodes in generation i is $P(v,t,S)=\binom{S}{v}\sum_{i=0}^{v}(-1)^{v-i}\binom{v}{i}\left(\frac{i}{S}\right)^t$.*

Proof. In generation $i+1$, let I_j be an indicator variable representing node j being chosen at least once. For any individual $1\le j\le S$ we have $\mathbb{P}[I_j=1]=1-(1-\frac{1}{S})^t$. Then the number of activated nodes is $\sum_{j=1}^{S}I_j$, and by linearity of expectation its expected value is $S-S(1-\frac{1}{S})^t$, as desired.

The second statement is a birthday problem variant. For each set of $\binom{S}{v}$ parents the probability that all edges end up in that set is $\left(\frac{v}{S}\right)^t$ and applying inclusion/exclusion to account for versions that don't select every element in that set gives the desired result. □

With this lemma we can can compute the expected value across two or more generations exactly, but the formula does not give much insight, so we omit it.

We can reformulate the problem as a Markov chain on the states $1,2,3,\dots,s$ representing the number of activated nodes at a given layer. This is an absorbing Markov chain with absorbing state 1, with transition probabilities given by $M_{i,j}=\begin{cases}P(i,j,s) & i\ge j\\ 0 & i<j\end{cases}$, forming a lower-triangular transition matrix.

The expectation we seek is $A(S,k) = \begin{bmatrix}0 & 0 & \cdots & 1\end{bmatrix} M^{k-2} \begin{bmatrix}1\\2\\\vdots\\S\end{bmatrix}$. For example, when $S = 3$, we have $M = \begin{bmatrix}1 & 0 & 0\\1/3 & 2/3 & 0\\1/9 & 6/9 & 2/9\end{bmatrix}$ and for diagrams with two layers ($k = 3$) we have that the three nodes have an expected $19/9 = 2.111...$ parents.

2.2 Limiting behavior

First, observe that for fixed S, we have $\lim_{k\to\infty} A(S,k) = 1$. This follows directly from the Markov chain interpretation of the problem, because it is a non-increasing sequence of positive integers with a positive probability of strict decrease at each step while the value is greater than 1.

Given S, construct a sequence of coefficients as follows: $a_{S,1} = 1$; and $a_{S,i+1} = 1 - \left(1 - \frac{1}{S}\right)^{(a_{S,i})S}$. Where S is understood to be fixed we will write simply $a_1 = S$, $a_{i+1} = 1 - \left(1 - \frac{1}{S}\right)^{a_i S}$. By Lemma 1, these approximate the share of active nodes at level i. We will get a rigorous upper bound below.

Note that as S gets large, $\left(1 - \frac{1}{S}\right)^S$ rapidly converges to $1/e$ from below. With this in mind, we can further approximate the a_i with a second sequence given by $b_1 = 1$; and $b_{i+1} = 1 - \frac{1}{e}^{b_i}$, which is more likely to have a useful generating function, if an analytic description is desired. Table 1 shows the b_i to be good approximations $a_{S,i}$ for large S.

i	1	2	3	4	5	6	7	8	9	10	11
$a_{10,i}$	1	0.6513	0.4965	0.4073	0.3490	0.3077	0.2769	0.253	0.234	0.2185	0.2056
$a_{100,i}$	1	0.6340	0.4712	0.3772	0.3155	0.2718	0.2390	0.2135	0.1056	0.1931	0.1625
$a_{1000,i}$	1	0.6323	0.4688	0.3744	0.3124	0.2684	0.2355	0.2099	0.1895	0.1727	0.1587
$a_{5000,i}$	1	0.6322	0.4686	0.3741	0.3121	0.2682	0.2352	0.2096	0.1891	0.1723	0.1583
b_i	1	0.6321	0.4685	0.3741	0.3121	0.2681	0.2352	0.2095	0.1890	0.1723	0.1582

Table 1: Values of $a_{S,i}$ for $S \in \{10, 100, 1000\}$ and $1 \le i \le 11$. As S grows, the $a_{S,i}$ and b_i get close.

Lemma 2. *For fixed $S > 1$ and $a_i = a_{i,S}$, we have $\lim_{i\to\infty} a_i = \frac{1}{S}$.*

Proof. First, we show by induction that $a_i \ge 1/S$ for all i. This is clearly true for $a_1 = 1$. If $a_i \ge 1/S$, then $a_i S \ge 1$ and

$$a_{i+1} = 1 - \left(1 - \frac{1}{S}\right)^{a_i S} \ge 1 - \left(1 - \frac{1}{S}\right)^{1} = \frac{1}{S}.$$

We will also need the following fact, which follows from the inequality $1 + c < e^c$ for $c = 1/(S-1) \ne 0$:

$$-\left(1 - \frac{1}{S}\right) \ln\left(1 - \frac{1}{S}\right)^S < 1. \tag{1}$$

As we know $a_{i+1} = 1 - \left(1 - \frac{1}{S}\right)^{a_i S}$, we will focus on the function $f(x) = 1 - \left(1 - \frac{1}{S}\right)^{xS}$. We begin by noting $f(\frac{1}{S}) = \frac{1}{S}$. We also see that

$$f'(x) = -\left(1 - \frac{1}{S}\right)^{Sx} \ln\left(1 - \frac{1}{S}\right)^S$$

Equation (1) tells us that $f'(\frac{1}{S}) < 1$. For $x \ge \frac{1}{S}$, the values we are interested in, this slope is always positive (because $\ln\left(1 - \frac{1}{S}\right)^S$ is negative) and is strictly decreasing in x. This implies for $x > \frac{1}{S}$, $f(x)$ is strictly bounded above by the line tangent to it at $1/S$:

$$f(x) < f\left(\frac{1}{S}\right) + f'\left(\frac{1}{S}\right)\left(x - \frac{1}{S}\right) = \frac{1}{S} + f'\left(\frac{1}{S}\right)\left(x - \frac{1}{S}\right)$$

First, we will show the sequence of a_i's is decreasing. Using $f'(\frac{1}{S}) < 1$, we see

$$a_{i+1} = f(a_i) < \frac{1}{S} + f'\left(\frac{1}{S}\right)\left(a_i - \frac{1}{S}\right) < \frac{1}{S} + a_i - \frac{1}{S} = a_i$$

As the sequence of a_i's is bounded below by $\frac{1}{S}$ and strictly decreasing, its limit must exist.

Suppose, for the sake of contradiction, that the limit of the a_i's is strictly greater than $1/S$, that is, it is $1/S+\alpha$ for some $\alpha > 0$. This means for all $\varepsilon > 0$, there is some sufficiently large i such that $a_i < 1/S+\alpha+\varepsilon$. Choose ε such that $0 < \varepsilon < \alpha(1 - f'(\frac{1}{S}))/f'(\frac{1}{S})$. This is possible to do because $\alpha > 0$ and $f'(\frac{1}{S}) < 1$. Note this choice of ε means $f'(\frac{1}{S})(\alpha + \varepsilon) < \alpha$. It follows that:

$$a_{i+1} = f(a_i) < \frac{1}{S} + f'\left(\frac{1}{S}\right)\left(a_i - \frac{1}{S}\right) < \frac{1}{S} + f'\left(\frac{1}{S}\right)(\alpha + \varepsilon) < \frac{1}{S} + \alpha.$$

As this is a monotone decreasing sequence and we assumed its limit was $\frac{1}{S} + \alpha$, it is impossible to have $a_{i+1} < \frac{1}{S} + \alpha$, giving a contradiction. Therefore it must be the case that the limit of this sequence is $\frac{1}{S}$, as claimed. □

Proposition 3. $A(S, k) \le a_{k-1}S$.

Proof. Let X_i be a random variable denoting the number of active nodes at level i (those that have descendants in level 1). Thus $X_1 \equiv S$. We are trying to get bounds on $\mathbb{E}[X_{k-1}] = A(S, k)$, the expected number of active nodes in the top level of a k-district descendancy diagram, which has $k - 1$ levels.

We will prove by induction that $\mathbb{E}[X_i] \le a_i S$, which suffices to prove the proposition. When $i = 1$, we have $\mathbb{E}[X_1] = a_1 S = S$, and the statement is true. Fix $i \ge 1$, and suppose $\mathbb{E}[X_i] \le a_i S$. By Lemma 1 and linearity of expectation, we have

$$\mathbb{E}[X_{i+1} \mid X_i] = S - S\left(1 - \frac{1}{S}\right)^{X_i}.$$

We will use the Law of Total Expectation ($\mathbb{E}[X_{i+1}] = \mathbb{E}[\mathbb{E}[X_{i+1} \mid X_i]]$) and Jensen's Inequality (for $c > 0$, $\mathbb{E}[c^X] \ge c^{\mathbb{E}[X]}$). We get

$$\mathbb{E}[X_{i+1}] = \mathbb{E}[\mathbb{E}[X_{i+1} \mid X_i]] = \mathbb{E}\left[S - S\left(1 - \frac{1}{S}\right)^{X_i}\right] = S - S\mathbb{E}\left[\left(1 - \frac{1}{S}\right)^{X_i}\right] \le S - S\left(1 - \frac{1}{S}\right)^{\mathbb{E}[X_i]}$$

$$\le S - S\left(1 - \frac{1}{S}\right)^{a_i S} = a_{i+1}S.$$

□

Remark 4. *While we only show $a_{k-1}S$ is an upper bound, for large k it is tight: in the limit as $k \to \infty$, it matches the trivial lower bound $\mathbb{E}[X_k] \ge 1$, which holds because there is at least one active node at each level. The empirical results, such as those presented in Figure 2, suggest the much stronger asymptotic $A(S, k) \sim a_{k-1}S$ as $S, k \to \infty$.*

Remark 5. *The case of square diagrams is a natural one to consider. Numerical results suggest that $A(S, S)$ limits to a constant slightly greater than 2. Subsequently, once there are two active nodes in a population of S, it takes an expected S more steps for those to collide, leaving a single common ancestor. This suggests that when $k \approx 2S$, we expect the ancestry to collapse to a single node. Compare Table 3.*

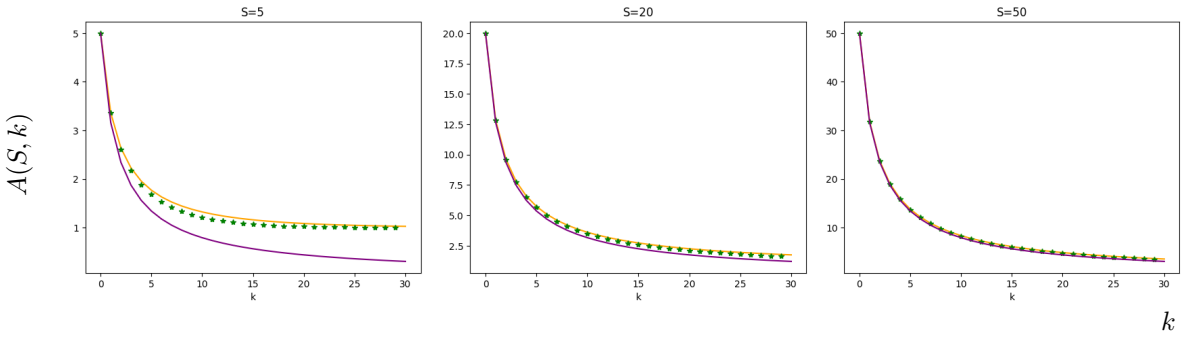


Figure 2: If the distribution of weights is uniform, these plots show the expected number of top-level nodes with at least one descendant at the bottom as k grows, for $S = 5, 20, 50$. The horizontal axis is k in each plot and the vertical axis is the expected number of top-level ancestors. Green stars are precise outputs from the Markov chain expression, compared to the $a_k S$ values in orange and the $b_k S$ values in purple (each interpolated by a curve).

	PA Congress $k = 18$	NM House $k = 42$	CA Congress $k = 52$	NY House $k = 63$	PA House $k = 203$
$a_{5000,k-1}$	0.1066	0.0466	0.0377	0.0313	0.0099
b_{k-1}	0.1065	0.0464	0.0376	0.0312	0.0098
predicted $A(5000, k)$	533.1	232.8	188.7	156.3	49.4
empirical $A(D, k)$	276.5	43.8	$*$	8.13	$*$

one day of computing time was not adequate to complete runs in the cases marked $*$

Table 2: For several realistic-sized problems, we consider the expected share of top-level ancestors that survive the full diagram if the weighting is uniform, using $5000 \cdot a_{5000,k-1}$ as an estimate for $A(5000, k)$. We include $k = 18$ (PA Congress 2010), $k = 42$ (NM state House), $k = 52$ (CA Congress 2020), $k = 63$ (NY state House), and $k = 203$ (PA state House). For example, if 5000 maps are sampled for California congressional districts with uniform weighting, we expect there to be a subset of 188.7 initial districts with at least one appearing in every map, so that each one is present an expected 26.5 times in the final sample. However, empirical runs of SMC run on the actual state geography (averaged over 1000 trials) show far more pronounced collapse. Reasons for this include non-uniform weights in the diagram, as discussed below.

2.3 Non-uniform weights

Above we assumed that each active node chooses its parent uniformly at random. In practice, this is not the case; the weights depend on graph properties of the partitions. The weight on node j at level i is given by $w_i^{(j)} = \frac{\tau\left(G_i^{(j)}\right)^{\rho-1}}{\left|\partial G_i^{(j)}\right|}$, where the numerator is the product of the number of spanning trees in the pieces of the partial plan $G_i^{(j)}$, raised to a power, and the denominator is the size of the edge cut. When $\rho \neq 1$ these weight factors will give wildly different probability of selection to plans based on their compactness, due to τ values for districts that can easily differ by 10^{100} in realistic problems.[3] Even at $\rho = 1$, plans with longer boundaries will be weighted down. These intergenerational weights thus present a compactness bias of some kind pulling away from uniformity for any choice of parameters.

Next we show that non-uniform weights exacerbate the sample repetition. Recall that X_i was a random variable denoting the number of active nodes at level i, where $X_1 \equiv S$ and parents are chosen uniformly

[3] In particular, $\tau > e^N$ for many planar graphs on N vertices (for instance $\tau \sim e^{1.6N}$ for triangular lattices), and Congressional districts might typically contain $N = 500$ precincts. This is one reason that validating on a 6×6 grid with 6 districts is inadequate to see salient effects of scale: in that setting, τ values for individual districts can only differ by a factor of 15. See related discussion in [4, §5.1].

at random at each level. We now set up our second model by fixing some non-uniform distribution over S nodes at each level $1, \ldots, k-2$ for the selection of parents. Fixing those distributions, we initialize $Y_1 \equiv S$, and let Y_i be a random variable denoting the number of active nodes at level i with the specified parent selection probabilities.

Lemma 6 (Uniform descendancy minimizes ancestor collapse)**.** *For $i \geq 2$,*

$$\mathbb{E}[Y_i \mid Y_{i-1} = a] \leq \mathbb{E}[X_i \mid X_{i-1} = a].$$

Proof. Consider the random variable Y_i. For $j \in \{1, 2, \ldots, S\}$, let p_j denote the probability that an individual at level $i-1$ chooses j as their parent in level i (note the p_j's may also vary with i). This proof uses Hölder's Inequality, which states that for $p, q \in [1, \infty)$ satisfying $1/p + 1/q = 1$ and any two vectors $\mathbf{u}$ and $\mathbf{v}$, $\|\langle \mathbf{u}, \mathbf{v} \rangle\|_1 \leq \|\mathbf{u}\|_p \|\mathbf{v}\|_q$. We apply this to the vectors $\mathbf{u}$ where $u_j = (1-p_j)/(S-1)$ and $\mathbf{v} = (1, 1, 1, \ldots, 1)$. Note

$$\|\langle \mathbf{u}, \mathbf{v} \rangle\|_1 = \|\mathbf{u}\|_1 = \sum_{j=1}^{S} \frac{1-p_j}{S-1} = 1.$$

Using $p = a$ and $q = a/(a-1)$, we see that

$$\|\mathbf{u}\|_a = \left(\sum_{j=1}^{S} \left(\frac{1-p_j}{S-1} \right)^a \right)^{1/a} = \left(\frac{1}{(S-1)^a} \sum_{j=1}^{S} (1-p_j)^a \right)^{1/a}$$

$$\|\mathbf{v}\|_{a/(a-1)} = \left(\sum_{j=1}^{S} 1^{a/(a-1)} \right)^{(a-1)/a} = S^{(a-1)/a}$$

Putting this together with Hölder's inequality, we see that

$$\left(\frac{1}{(S-1)^a} \sum_{j=1}^{S} (1-p_j)^a \right)^{1/a} S^{(a-1)/a} \geq 1 \qquad \text{and} \qquad \sum_{j=1}^{S} (1-p_j)^a \geq \frac{(S-1)^a}{S^{a-1}} = S \left(1 - \frac{1}{S} \right)^a.$$

We now see that

$$\begin{aligned} \mathbb{E}[Y_i \mid Y_{i-1} = a] = \sum_{j=1}^{S} (1 - (1-p_j)^a) &= S - \sum_{j=1}^{S} (1-p_j)^a \\ &\leq S - S \left(1 - \frac{1}{S} \right)^a = S \left(1 - \left(1 - \frac{1}{s} \right)^a \right) = \mathbb{E}[X_i \mid X_{i-1} = a] \end{aligned}$$

This completes the proof. □

We find empirically that the weights at each level have a distribution shaped like the one shown in Figure 3, which was drawn from a run on New Mexico state Senate districts. In New Mexico, it was common to see max-to-median weight ratios of 10 within a generation, and max-to-min ratios of 30, even with $\rho = 1$. In New York's Senate districts, these ratios were commonly 30 and 2000, respectively. As we will see, skews of this kind will tend to significantly increase the collision rate.

Table 3 shows an simulated comparison of $F(D, \varphi)$ between the case of uniform weights and a simple non-uniform setup where one node at each layer is 100 times more likely to be selected than each of the others (that is, the weights are $100 : 1 : 1 \cdots : 1$). As we would expect given Lemma 6, the ancestor collapse is significantly accelerated in the non-uniform case. But even this significantly understates the actual repetition in SMC samples; recall Tapp's expert affidavit finding roughly that $F(D, .6) < 31$ for a $S = 5000$ sample in New York. This is partly because the graph partition step itself can boost repetition; if partial progress has created a hard-to-split remainder, this creates yet another scenario in which a generation may be filled out with repeats.

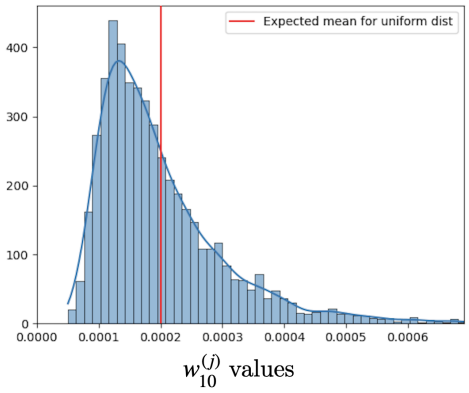


Figure 3: Truncation of a long-tailed histogram of weights in a descendancy diagram on state Senate districts in New Mexico ($k = 42, i = 10, S = 5000$, with compactness parameter $\rho = 1$). If weights were uniform, the distribution of weights would be concentrated at the red line. Instead, when drawing the 33rd district in this SMC process, some 32-district partial plans are over 100 times likelier than others to be chosen. This trial was run with the McCartan–Imai Redist package using default settings for SMC.

First level with a "mega-ancestor" accounting for φ share of the final nodes

	$S = 10$	100	1000	5000
$\varphi = .01$	—	—	3.8	15.4
.1	—	5.5	51.1	256.3
.25	2.5	18.7	188.3	957.8
.5	5.5	60.2	622.2	3187.3
.75	14.6	144.7	1433.5	7092.4
1	17.9	201.9	2065.2	9966.2

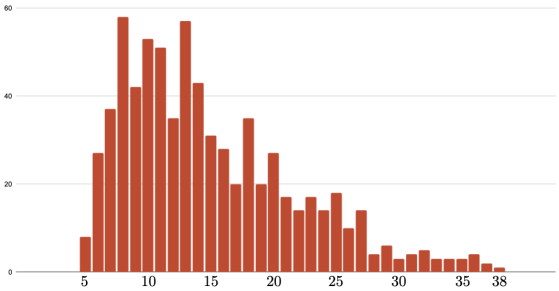
Uniform weights

	$S = 10$	100	1000	5000
$\varphi = .01$	—	—	2.0	2.0
.1	—	2.0	4.6	78.4
.25	2.0	2.0	19.7	320.1
.5	2.0	2.8	65.8	1032.7
.75	2.0	5.3	151.7	2401.9
1	2.7	11.1	232.3	3533.2

$100 : 1 : 1 \cdots : 1$ weights

	NM $k = 42$ $S = 5000$	NY $k = 63$ $S = 5000$
$\varphi = .01$	3.2 (100%)	2.9 (100%)
.1	8.8 (100%)	6.9 (100%)
.25	14.9 (71.4%)	11.9 (99%)
.5	17.0 (23.5%)	18.8 (74%)

Actual runs



Observed $F(D, .25)$ for New Mexico runs

Table 3: Recall that $F(D, \varphi)$ reports the lowest level at which some node is an ancestor to φ share of the bottom generation. (This is vacuous if $\varphi S \leq 1$.) The tables show estimated expectations for $F(D, \varphi)$ with uniform weights and with stylized non-uniform weights—each cell value is obtained by averaging over 1000 trials. Reading across the bottom row, for instance, confirms that $A(S, 2S) \approx 1$. Collapse is exacerbated significantly in the non-uniform case. Comparison of the highlighted cells shows that the empirical repetition is far greater than what is predicted by the stylized combinatorial models. For instance, 714 out of 1000 trials in New Mexico had a node serving as ancestor to 25% of the final generation; the average first occurrence of that node was at level 14.9, with a spike at level 8. The histogram shows that it was quite common for an SMC run to produce outputs for which one-quarter of plans shared 30 or more identical districts ($F(D, .25) \leq 13$ with $k = 42$).

3 Convergence guarantees and the interpretation of samples

In [9], the main convergence result that is presented is a weak central limit theorem stated as follows.

Proposition 7 (McCartan–Imai Prop 4.2)**.** *Let $\pi_S = \sum_{j=1}^{S} w^{(j)} \delta_{[\xi(j)]}$ be the weighted particle approximation generated by [their SMC Algorithm]. Then for all measurable h on unlabeled plans, as $S \to \infty$,*

$$\sqrt{S}\left(\mathbb{E}_{\pi_S}\left[h\left([\xi]\right)\right] - \mathbb{E}_{\pi}\left[h\left([\xi]\right)\right]\right) \xrightarrow{d} \mathcal{N}(0, V_{\mathrm{SMC}}(h))$$

for some asymptotic variance $V_{\mathrm{SMC}}(h)$.

As the authors note, this is convergence in probability rather than almost sure convergence. This style of convergence result does not rule out the sample repetition described here, and it offers interestingly different guarantees from those provided by Markov chain samples.

Theorem 8 (Markov chain ergodic theorem[4])**.** *If M is the transition matrix for an ergodic Markov chain on a finite state space Ω, then there exists a unique steady state π so that $\pi M = \pi$, and for an arbitrary probability vector w on Ω we have*

$$\|wM^k - \pi\| \to 0.$$

A *sample path* starts from an arbitrary basepoint $y_1 \in \Omega$ and successively selects a sequence of neighbors according to the transition probabilities in M to form a multiset $\Lambda_S = \{y_1, y_2, \ldots, y_S\}$ (which we can call a sample or an *ensemble*). (This is a multiset because it may happen that $y_i = y_j$, so the state is recorded with multiplicity.) The associated sample measure can be denoted $\mu_S = \frac{1}{S}\sum_{j=1}^{S} \delta_{y_i}$.

Because Theorem 8 gives convergence in total variation, it means that large samples give a probability to every state $x \in \Omega$ with a frequency approaching $\pi(x)$. That is, for a.e. sample path,

$$\frac{1}{S}\left|\left\{1 \le i \le S : y_i = x\right\}\right| \to \pi(x).$$

From this, it follows that for a summary statistic $h : \Omega \to \mathbb{R}$, we have convergence in distribution $h_*\mu_S \to h_*\pi$ (and therefore pointwise convergence). Thus, a histogram of observed values of $h(y_i)$ over a Markov chain ensemble is guaranteed (almost surely) to approach the π-weighted histogram on $h(x)$ over $x \in \Omega$.

To illustrate the limited guarantees of Proposition 7, consider the following construction.

Example 9 (Total repetition sampler)**.** *Given a distribution π on a state space Ω, let the* total repetition sampler TRS *be defined as follows: a sample of size S is constructed by taking one draw $x \sim \pi$ and returning x with multiplicity S. For instance, if π is a target distribution on districting plans, μ draws a single π-distributed plan and repeats S identical copies of the full plan. The associated measure is just the Dirac measure $\nu_S = \frac{1}{S} \cdot S\delta_x = \delta_x$.*

We can see that this measure satisfies the conclusion of Proposition 7, because $\mathbb{E}_{\nu_S}(h) = \mathbb{E}_{\pi}(h)$ for any h, so the difference equals the zero function (i.e., the normal centered at zero with variance $V_{\mathrm{TRS}} = 0$). However, enlarging S does not improve the sample, and any histogram formed by a single large sample from this process will have a single spike and will not approach the true shape of $h_*(\pi)$ as S grows.[5] To compensate for the repetition from TRS and attempt to recover the shape of a distribution, we would have to take multiple separate samples rather than enlarging a single sample.

The authors are clearly aware that SMC faces challenges of this kind, because even in a small validation example (dividing the 6×6 grid into $k = 6$ districts, [9, Fig 4]) they have averaged 24 independent runs to obtain their estimates rather than enlarging the size of a single sample. In other published work [10], the same authors and collaborators present ensembles of 5000 maps for all 50 states, and appear to do so by combining subsamples from two SMC runs.

It is not clear what balance of the sample size S and the number of separate runs would constitute best practices for users of SMC.[6]

[4] See [6, Thm 5.6.6, p313] or [7, Thm 4.9, p52].

[5] We suspect that this is not the case for McCartan–Imai's SMC process—though the central limit theorem does not suffice to guarantee it, it seems likely that for large enough S, the SMC empirical measure π_S may still resemble π as a distribution. However, the asymptotic result has not been proved, and on the practical sie it seems that it would require samples far larger than the McCartan–Imai implementation can generate.

[6] For the illustrative example TRS, the ideal structure would be many samples of size 1. For SMC, this would fail, because a large S is already needed for the reweighting of the final sample to take the target distribution into account at all.

4 Discussion

Since spanning tree methods were introduced in redistricting around 2018, many authors who need a single exemplar of a plan (such as to serve as the starting point for a Markov chain) obtain one by recursively partitioning a tree down to districts. One way to view SMC is that it constructs the whole sample by running this seeding process many times. Employing the structure of a descendancy diagram allows the use of weights that help control the properties of the sample, but at the cost of introducing significant combinatorial redundancy.

Repetition and interpretation. SMC is sometimes claimed to produce nearly independent samples from arbitrary distributions.[7] However, district repetition can create massive dependencies, with many plans being identical on large regions. Numerous factors that are present in real use cases—non-uniform weights stemming from compactness bias, and restricted choices in the decision tree caused by the connection topology of the state—can contribute to very severe repetition, which is visible in plots. The repetition undermines the use of an SMC ensemble to infer the shape of a distribution of summary statistics, such as in histograms and boxplots. A highly redundant sample does not allow for reliable outlier analysis because its percentile statistics can be far from those of the target distribution. The stated central limit theorem does not justify the accuracy of histograms drawn from one SMC sample; if citing this theorem, we would require a large number of large samples to be confident of accurate histograms and outlier claims.

Reweighting. The SMC process leverages importance sampling by constructing an initial sample through a descendancy diagram and then reweighting according to the target distribution π only when the sample is complete. Before the final reweighting, the sample is approximately distributed by the spanning tree distribution τ on partitions—but somewhat distorted by repetition, labeling bias, and other artifacts of the construction.[8] The authors intend to use this method to target arbitrary distributions $\pi(\xi) \propto e^{-J(\xi)}\tau(\xi)^{\rho}$, but, in particular, the energy functional J is never used until the descendancy diagram is complete; partitions that are never encountered by the tree-based generation process cannot be rescued by reweighting. Thus attempts to target a general π with SMC will fare no better than applying one-shot reweighting to previously known methods to sample from τ. The accuracy depends on emitting a large and diverse sample from the descendancy process. The Redist implementation of SMC struggles to produce sample sizes beyond the tens of thousands on practical problems, which means that many legally relevant events will never be observed—whereas Markov chain methods for sampling from τ can be run to billions of steps.

District-level properties. Courts have often expressed an interest in the presence of individual districts with particular properties. For instance, in the litigation challenging the Pennsylvania Congressional plan, the court strongly discouraged the splitting of Pittsburgh (and the special master was said to treat it as a "disqualifying feature" of a plan).[9] In the New Mexico legal challenge, the parties to litigation debated whether it was disqualifying if any district "contains more than 60% of the state's active oil wells."[10] Ensembles constructed from small numbers of SMC runs, including the ALARM ensembles published in *Scientific Data* [10] which are made by combining two runs, are unsuited for estimating the frequency with which district-level properties occur.

Too many districts. Large numbers of districts (large k) exacerbate all of the problems described here. Validation efforts have only been conducted for $k \leq 6$, and the authors themselves have used subdivision to make the problem more tractable—the ALARM 50-state project breaks up Texas ($k = 38$), Florida ($k = 27$), and California ($k = 52$) into three or more ad hoc pieces, samples them separately, and combines to make

[7] For instance, this is explicit in the Redist documentation at https://perma.cc/YV37-JZNR.

[8] Here, *labeling bias* refers to the fact that the descendancy process produces plans which implicitly carry sequential labels on the districts, whereas the target is a distribution on unlabeled plans. This is addressed with another sampled and approximated factor ψ, described in [9, §4.4.2].

[9] *Carter v. Chapman* (2022), see https://www.pacourts.us/assets/opinions/Supreme/out/J-20-2022mo.pdf?cb=1.

[10] *Republican Party vs. Oliver* (2023), see nmlegis.gov links, particularly *Plaintiff's Opposed Motion to Exclude Expert Report & Expert Testimony of Dr. Jowei Chen* [1].

Congressional plans.[11] Ensembles that have been modularized in this way have an unknown relationship to the target distribution on the full state.

All of these observations counsel caution in using SMC on full-scale problems, especially if attempting to target distributions that differ from the spanning tree distribution τ or with more than (say) 20 districts. When there are many districts, the combinatorics of the descendancy diagram create severe repetition; on the other hand, when there are few districts with many units, the weights can be highly non-uniform, which will likewise boost repetition. Minimizing the effects of repetition would call for far larger sample sizes than is currently possible.

References

[1] Legislative defendants' response to plaintiffs' proposed motion to exclude. Fifth Judicial District of New Mexico. *Republican Party of New Mexico v. Maggie Toulouse Oliver* (2022). https://www.nmlegis.gov/Redistricting2021/Litigation%20Docs/292%20-%20September%2025,%202023%20Plaintiff's%20Opposed%20Motion%20to%20Exclude%20Expert%20Report%20&%20Expert%20Testimony%20of%20Dr.%20Jowei%20Chen.pdf.

[2] Second Affidavit of Dr. Kristopher R. Tapp, PhD. Supreme Court of the State of New York. *Harkenrider v. Hochul* (2022). https://perma.cc/29X3-59CX.

[3] Jean Bertoin. *Random Fragmentation and Coagulation Processes.* Cambridge University Press, 2006.

[4] Daryl DeFord, Moon Duchin, and Justin Solomon. Recombination: A Family of Markov Chains for Redistricting. *Harvard Data Science Review*, 3(1), March 2021. https://hdsr.mitpress.mit.edu/pub/1ds8ptxu.

[5] Rick Durrett. *Probability Models for DNA Sequence Evolution.* Springer, second edition, 2008.

[6] Rick Durrett. *Probability: Theory and Examples, 5th ed.* Cambridge University Press, 2019.

[7] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov chains and mixing times.* American Mathematical Society, Providence, RI, 2009. With a chapter by James G. Propp and David B. Wilson.

[8] Jun S. Liu, Rong Chen, and Tanya Logvinenko. A theoretical framework for sequential importance sampling with resampling. In *Sequential Monte Carlo methods in practice*, Stat. Eng. Inf. Sci., pages 225–246. Springer, New York, 2001.

[9] Cory McCartan and Kosuke Imai. Sequential Monte Carlo for sampling balanced and compact redistricting plans. *The Annals of Applied Statistics*, 17(4):3300 – 3323, 2023.

[10] Cory McCartan, Christopher T. Kenny, Tyler Simko, George Garcia III, Kevin Wang, Melissa Wu, Shiro Kuriwaki, and Kosuke Imai. Simulated redistricting plans for the analysis and evaluation of redistricting in the United States. *Scientific Data*, 9:689, 2022.

[11] Martin Raab and Angelika Steger. "Balls into Bins" — A simple and tight analysis. In Michael Luby, José D. P. Rolim, and Maria Serna, editors, *Randomization and Approximation Techniques in Computer Science*, pages 159–170, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg.

[11] See Texas, Florida, and California README files from the ALARM 50-state project.