Simulated Annealing for Graph Bisection

Mark Jerrum [*] Gregory Sorkin [†]

April 16, 1993

Abstract

We resolve in the affirmative a question of Boppana and Bui: whether simulated annealing can, with high probability and in polynomial time, find the optimal bisection of a random graph in $\mathcal{G}_{npr}$ when $p - r = \Theta(n^{\Delta-2})$ for $\Delta \leq 2$. (The random graph model $\mathcal{G}_{npr}$ specifies a "planted" bisection of density r, separating two $n/2$-vertex subsets of slightly higher density p.) We show that simulated "annealing" at an appropriate fixed temperature (i.e., the Metropolis algorithm) finds the unique smallest bisection in $O(n^{2+\varepsilon})$ steps with very high probability, provided $\Delta > 11/6$. (By using a slightly modified neighborhood structure, the number of steps can be reduced to $O(n^{1+\varepsilon})$.) We leave open the question of whether annealing is effective for Δ in the range $3/2 < \Delta \leq 11/6$, whose lower limit represents the threshold at which the planted bisection becomes lost amongst other random small bisections. It remains open whether hillclimbing (i.e., annealing at temperature 0) solves the same problem.

1 Introduction

Simulated annealing [15] is a popular heuristic technique for approximately solving hard problems in combinatorial optimization. The method works by simulating a stochastic process on the space of feasible solutions, in which transitions between states are governed by a rule that favors states with lower cost or "energy". A parameter known as "temperature" controls the extent to which lower energy states are favored: at low temperature equilibrium, low energy states occur with much higher probability than high energy ones, while at high temperatures the difference is less marked. To this extent, a low temperature is beneficial; however there is a tension between this effect and a detrimental one,

[*]Department of Computer Science, University of Edinburgh, The King's Buildings, Edinburgh EH9 3JZ, UK. The work described here was supported by grant GR/F 90363 of the UK Science and Engineering Research Council and by Esprit Working Group "RAND", and was partly done while the author was visiting the NEC Research Institute at Princeton NJ, USA.

[†]IBM Thomas J. Watson Research Center, J2-C27 Hawthorne, PO Box 704, Yorktown Heights, NY 10598, USA. This work was partly done while the author was visiting the University of Edinburgh under the above SERC grant.

namely that equilibrium typically takes much longer to attain at low temperatures than at high ones.

In simulated annealing proper, the temperature parameter is varied with time in an attempt to exploit this tension. The temperature starts high and is gradually reduced with time according to a specified "cooling schedule", in the belief that different "features" of the solution will "crystallize out" at different temperatures. Unfortunately, no combinatorial optimization problem that has been subjected to rigorous theoretical analysis has exhibited this phenomenon: those problems that can be solved efficiently by simulated annealing, can be solved just as effectively by "annealing" at a single carefully selected temperature. (Sorkin has exhibited fractal-like "energy landscapes" that require a non-trivial cooling schedule for fast convergence to a low energy state [20], but these landscapes do not correspond to a "naturally defined" optimization problem.) A rigorous demonstration that annealing is provably beneficial for some natural optimization problem would, in the authors' view, rate as a significant theoretical advance.

Simulated annealing at constant temperature is called the *Metropolis algorithm*, and the stochastic process (actually a Markov chain) on which it is based the *Metropolis process*. The theoretical results so far obtained relate to the Metropolis algorithm rather than simulated annealing as such. The results are few in number and can be briefly summarized. The most extensively studied problem is that of finding a maximum matching in a graph. Jerrum and Sinclair [14, 13] (see also Frieze [9] and Motwani [17]) showed that the Metropolis algorithm performs well on random graphs, even though — as Sasaki and Hajek [19] had previously demonstrated — pathological instances exist on which the algorithm fails badly. Jerrum considered the problem of finding a maximum clique in a random graph [12], and showed (in rough terms) that the Metropolis algorithm is no more effective than hillclimbing (i.e., simulated annealing at temperature 0). Aside from these results, Sasaki [18] has used a "density of states" argument to construct instances of various optimization problems that cause the Metropolis algorithm to run for exponential time.

By analyzing the performance of the Metropolis algorithm — and more generally simulated annealing — in the context of various simple, paradigmatic optimization problems, we may hope to learn to recognize when these approaches are likely to be effective, and to understand why they work when they do. In this paper we extend the study of the Metropolis algorithm to the problem of graph bisection, i.e., finding a partition of the vertex set of an undirected graph into two equal-sized sets so that the number of crossing edges is minimized. The problem is known to be NP-complete [11, 10], so we cannot expect good performance in the worst case. Instead, we study the effectiveness of the Metropolis algorithm on "random" instances of the problem, and show that its performance is surprisingly good. Thus, this paper and [12] can be viewed as complementary. We remark in passing that the bisection problem in various random graph models has received considerable attention [5, 4, 6, 3].

To describe our positive result, we require some notation. For n even, a random graph $G = (V, E) \in \mathcal{G}_{npr}$ is one with $n/2$ *white* and $n/2$ *black* vertices, with edge probability p between pairs of vertices of the same color and edge probability r between pairs of opposite color. A *partition* or *bisection* π of G is a partition of V into sets L and R (thought of as "left" and "right") with $|L| = |R| = n/2$. The *cut-width* of a bisection is the number of cut edges: $|\{\{u,v\} \in E : u \in L \text{ and } v \in R\}|$; and the *bisection width* of G is the minimum over all bisections (L, R) of the cut-width of (L, R).

We note that a graph in $\mathcal{G}_{n,\frac{1}{2},\frac{1}{2}}$ is simply a random graph on n vertices with edge probability $\frac{1}{2}$, and that the cut-width of a random bisection has expectation $\frac{1}{8}n^2$ and variance $\Theta(n^2)$ as well. It can be shown that with high probability the bisection width is $\frac{1}{8}n^2 - \Theta(n^{3/2})$. On the other hand if $p = \frac{1}{2}$ but $p - r = n^{\Delta-2}$, with $3/2 < \Delta < 2$, the natural partition (with L the set of black vertices and R the set of white vertices) has expected width $\frac{1}{4}rn^2 = \frac{1}{8}n^2 - \frac{1}{4}(p-r)n^2 = \frac{1}{8}n^2 - \frac{1}{4}n^{\Delta}$. It can be shown [4] that with high probability this is in fact the minimum bisection.

The Metropolis process on bisections is described in detail in Section 2, but, in brief, the allowed transitions take a pair of vertices, one from each side of a bisection, and interchange them. The main new result is contained in Theorem 6.1: informally, this theorem states that with overwhelming probability the Metropolis process will converge to the unique minimum bisection in time about $O(n^2)$, provided $\Delta > 11/6$. The proof uses (in Claim 5.2) a novel argument concerning perturbed Markov chains that in a sense extends the investigations of Azar, Broder, Karlin, Linial, and Phillips [1] into biased random walks.

Boppana [3] and Bui [5] consider a $\mathcal{G}_{nmb}$ model, where graphs are chosen uniformly from the set of n-vertex graphs with m edges and bisection width b. This is roughly equivalent to the $\mathcal{G}_{npr}$ model, with $b = \frac{1}{4}rn^2$ and $m = \frac{1}{4}(p+r)n^2$. In order to compare the two models, define the *deficiency* of a graph to be the difference between its bisection width and the expected cut-width of a random bisection. For a random graph in $\mathcal{G}_{nmb}$ the deficiency is close to $\frac{1}{2}m - b$, whereas for the $\mathcal{G}_{npr}$ model it is typically $\frac{1}{8}n^{\Delta} \pm O(n)$. Boppana presents an algorithm which with probability $1 - O(n^{-1})$ finds the minimum bisection for graphs in $\mathcal{G}_{nmb}$ with $\frac{1}{2}m - b \geq \frac{5}{2}\sqrt{mn \log n}$, and sets as an open problem the average-case performance of simulated annealing (as well as the Kernighan-Lin and greedy heuristics).

The range of applicability of Boppana's algorithm is exceptionally wide, taking in graphs whose deficiency is close to the threshold where the planted bisection becomes lost amongst other random small bisections. In this regard, our result compares unfavorably, since the expected deficiency is required to be substantially higher: $O(n^{11/6+\varepsilon})$ against $\Omega(n^{3/2}\sqrt{\log n})$ for dense graphs. Nevetherless, it is perhaps surprising that an algorithm as simple as that of Metropolis can provably deal with deficiencies of order n^{Δ} with $\Delta < 2$. It is worth observing that Boppana's algorithm is by no means simple to implement, requiring as it does the Ellipsoid Method as a subroutine, whereas the Metropolis process is trivial. The Metropolis algorithm for minimum bisections is also fast: faster even than Theorem 6.1 would suggest. As will be apparent later, the algorithm burns up almost all its alloted time in a rather mindless way after the minimum bisection has already essentially been found; by slightly modifying the choice of perturbations the execution time could easily be reduced to $O(n^{1+\varepsilon})$.

For sparse graphs ($p \leq n^{-1/6}$) our results become vacuous. This is broadly in line with a result of Bui [5] to the effect that hillclimbing (randomized greedy, or the Metropolis algorithm at temperature 0) is ineffective for very sparse graphs, and cannot even discover the obvious minimum bisection existing between a pair of $\frac{1}{2}n$-vertex random cubic graphs. Bui also conjectured that simulated annealing, and maybe even hillclimbing, would be effective for dense random graphs, and this is now seen to be correct, at least for simulated annealing. The performance of hillclimbing on dense random graphs remains unanalyzed.

2 Annealing Model and Key Notation

2.1 Annealing algorithm

An annealing algorithm is begun in a random state. Moves are *generated* or *proposed* at random, and may be *accepted* (in which case the current state is updated) or *rejected* (in which case the state is left unchanged). A move which would increase the cost by a (positive or negative) quantity Δcost is accepted with probability $A(\Delta\text{cost})$.

We consider the most obvious application of annealing to the graph bisection problem, in which

- a *state* is a bisection of G into sets L and R,
- the *cost* (or "energy") $\text{cost}(L, R)$ of a state (L, R) is its cut-width, and
- a *move* is the exchange of a vertex in L with a vertex in R.

We take as the acceptance function

$$A(\Delta\text{cost}) = \frac{\exp(-\Delta\text{cost}/T)}{1+\exp(-\Delta\text{cost}/T)} = \frac{1}{1+\exp(\Delta\text{cost}/T)},$$

where T is the *temperature* parameter for the annealing. This acceptance function is common, albeit not as common as $A(\Delta\text{cost}) = \min\{1, \exp(-\Delta\text{cost}/T)\}$ whose nonuniformity complicates the analysis. Both acceptance rules define an ergodic Markov chain on bisections, and both share the stationary distribution in which the probability of (L, R) is proportional to $\exp(-\text{cost}(L, R)/T)$. (The stationary distribution gives the limiting probability of being at state (L, R) as time $t \to \infty$.)

2.2 Graph model and parameterization

So that the graph can be bisected exactly, it is natural to restrict the number of vertices to be even. In fact, we shall further restrict it to be a multiple of 4, so that henceforth we consider random graphs drawn from $\mathcal{G}_{4n,p,r}$. This restriction is immaterial to the results but conveniently makes the *imbalance* – a key quantity in our analysis – an integral value, The *imbalance* of any partitioning of $G \in \mathcal{G}_{4n,p,r}$ is the value $k \geq 0$ for which there are $n - k$ white vertices in one partition (which without loss of generality we call partition L) and $n + k$ in the other (partition R). (L and R are labeled arbitrarily for $k = 0$.) A given annealing move will either leave k unchanged, or will increase or decrease it by 1.

Note that a typical random partition has $k = O(n^{1/2})$, while the canonical partition (which with high probability is the minimum) is uniquely characterized by $k = n$. As we will see, the cost of a state is equal to a function of k plus an error term which is small (with high probability). Thus the sequence $k(t)$ of imbalances after t attempted moves can be used to study the progress of annealing. That is, while the transition probabilities $k \mapsto k-1$, $k \mapsto k$, and $k \mapsto k+1$ depend on the full state π, under appropriate conditions they can be approximated from k alone, making $k(t)$ a tractable "near Markov chain".

2.3 Randomness and move generation

Throughout this paper the phrase *with overwhelming probability* will mean "with probability $1 - O(\exp(-n^c))$ for some constant $c > 0$".

The full probability space under consideration is the cross product of the space of random graphs, the random proposed moves over all steps, and the random coin tosses at all steps. We will write $\mathbf{E}_{\mathcal{G},\mathcal{M}}$ to denote the expectation of some quantity over random graphs and moves, $\Pr_{\mathcal{M}}$ to denote a probability over random moves alone, and so forth.

Finally, in Sections 4 and 5 it is presumed that the sequence of all proposed moves is specified in advance: this must be done in such a way that the acceptance or rejection of a previous move does not jeopardize the legality or uniform randomness of later moves. The following means of specifying moves satisfies this, and is a convenient concrete model for implementation of the algorithm. Consider the collections L and R to be ordered lists of vertices rather than mere sets, so that a state $\pi = (L, R)$ is a permutation. (Conversely, for a state specified by the partition π, L and R will refer to the first and last $2n$ vertices.) Let the move (i, j) indicate the exchange of L_i and R_j: the i'th vertex in L and the j'th vertex in R. In annealing, moves will be generated by choosing i and j uniformly at random from $\{1 \dots 2n\}$.

3 Typical Transition Probabilities

The transition $k \mapsto k+1$ occurs iff a move taking k to $k+1$ is both proposed and accepted.

Claim 3.1 *For any graph G and any partition π with imbalance k, the probability that a proposed move would increase k is*

$$\Pr_{\mathcal{M}}[\mathit{propose}\, k \mapsto k+1] \;=\; \frac{1}{4} - \frac{k}{2n} + \frac{k^2}{4n^2}. \tag{1}$$

Proof A proposed move (i, j) increases k if L_i is white, and R_j is black, which has probability $((n-k)/2n)^2$. ∎

For a given k-bisection π, and k-increasing move M, let $X^+ = X^+(G)$ be the random variable representing the cost increase that occurs when move M is applied to the partitioning π of the graph G. Recall that $A(x) = (1 + \exp(x/T))^{-1}$ is the acceptance function. Note that $A(X^+)$ is the probability that a given k-increasing move is accepted.

Claim 3.2 *Suppose $T \geq \sqrt{n}$. Then $\mathbf{E}_{\mathcal{G}} A(X^+) \geq \frac{1}{2} + \Omega(\min\{k(p-r)/T,\, 1\})$, where the expectation is over random graphs $G \in \mathcal{G}_{4n,p,r}$.*

Proof The distribution of X^+ can be expressed as the sum and difference of $8n-4$ Bernoulli random variables, corresponding to the $8n-4$ potential edges affected by the specified move:

$$X^+ = 2(n-k-1)\cdot B_p + 2(n+k)\cdot B_r - 2(n+k)\cdot B_p - 2(n-k-1)\cdot B_r \qquad (2)$$

where the equality sign is to be interpreted as "has the same distribution as", and the notation $t \cdot B_p$ stands for the sum of t Bernoulli r.v.s each with expectation p. Introduce a new r.v. X^0 with distribution

$$X^0 = 2n\cdot B_p + 2(n-1)\cdot B_r - 2n\cdot B_p - 2(n-1)\cdot B_r.$$

Observe that the distribution of X^0 is symmetric about the origin, while the function $A(x) - \frac{1}{2}$ is antisymmetric, and hence $\mathbf{E}A(X^0) = \frac{1}{2}$. Our strategy is to estimate $\mathbf{E}A(X^+)$ by comparing the distributions of X^+ and X^0, which are obviously close for small k.

We construct a model for X^+ and X^0, on a common sample space, that gives the correct distribution for each r.v., even though X^+ and X^0 will (intentionally) be highly correlated. Express X^+ as the sum $X^+ = \sum_{i=1}^{4n-2} X_i^+ - \sum_{i=4n-1}^{8n-4} X_i^+$ of independent 0,1 r.v.s $\{X_i^+\}$, where the expectation of individual variables is suggested by expression (2). (Thus the first $2(n-k-1)$ have expectation p, the next $2(n+k)$ have expectation r, and so on.) Let $X^0 = \sum_{i=1}^{4n-2} X_i^0 - \sum_{i=4n-1}^{8n-4} X_i^0$ be a similar expression for X^0. For pairs (X_i^+, X_i^0) containing variables of the same type (i.e., both having expectation p, or both r), arrange for the variables to be perfectly correlated, i.e., $X_i^+ = X_i^0$. For the remaining $4(k+1)$ pairs (e.g., where X_i^+ has expectation p and X_i^0 has expectation r) arrange for the p-variable to dominate the r-variable (e.g., $X_i^+ \geq X_i^0$); this can be achieved by setting $\Pr(X_i^+ = X_i^0 = 0) = 1-p$, $\Pr(X_i^+ = 1 \wedge X_i^0 = 0) = p - r$, and $\Pr(X_i^+ = X_i^0 = 1) = r$. This construction yields the following sharply concentrated distribution for the difference of X^0 and X^+:

$$X^0 - X^+ = (4k+2)\cdot B_{p-r} \geq 0.$$

It follows by symmetry that $\Pr\left(X^0 - X^+ \geq 2k(p-r)\right) \geq \frac{1}{2}$.

From usual estimates for the binomial distribution, $\Pr(|X^0| \leq 4\sqrt{n}\,) \geq \frac{3}{4}$, and hence, with probability at least $\frac{1}{4}$ the events $X^0 - X^+ \geq 2k(p-r)$ and $|X^0| \leq 4\sqrt{n}$ occur simultaneously. (This is so, irrespective of any probabilistic dependency between the events.) Thus, since $T \geq \sqrt{n}$, there is a probability of at least $\frac{1}{4}$ that $|X^0/T| \leq 4$ and $X^0/T - X^+/T \geq 2k(p-r)/T$ hold simultaneously. Note that the conjunction of these two events implies $A(X^+) - A(X^0) = \Omega(\min\{k(p-r)/T, 1\})$, since in any bounded range, the "sigmoid" $(1+\mathrm{e}^x)^{-1}$ has negative gradient uniformly bounded away from 0. It follows that

$$\begin{aligned}
\mathbf{E}A(X^+) - \tfrac{1}{2} &= \mathbf{E}A(X^+) - \mathbf{E}A(X^0) \\
&= \mathbf{E}(A(X^+) - A(X^0)) \\
&\geq \Omega(\min\{k(p-r)/T,\, 1\}) \qquad (3)
\end{aligned}$$

where we have used the fact that $X^0 - X^+$ is non-negative. (Naturally, the expectations here are with respect to the sample space constructed earlier.) The result follows immediately. ■

For a given partition π with imbalance k, let $f^+ = f^+(G, \pi)$ be a random variable denoting the probability that the new partition after one move has imbalance $k+1$. Let $f^- = f^-(G, \pi)$ be the corresponding r.v. for k-decreasing moves.

Corollary 3.3 *Suppose $T \geq \sqrt{n}$. Then for a partition π with imbalance k,*

$$\mathbf{E}_{\mathcal{G}} f^+(G,\pi) \;\geq\; \tfrac{1}{8} + \Omega\left(\min\{k(p-r)/T, 1\}\right) - O(k/n)$$

and

$$\mathbf{E}_{\mathcal{G}} f^-(G,\pi) \;\leq\; \tfrac{1}{8} - \Omega\left(\min\{k(p-r)/T, 1\}\right) + O(k/n)$$

where the expectation is over random graphs $G \in \mathcal{G}_{4n,p,r}$.

Proof Follows from Claim 3.1, Claim 3.2, and symmetry. ■

To show that f^+ is concentrated near its expected value we employ a "bounded difference" inequality. Specifically we use a generalization of Azuma's inequality in a form due to McDiarmid [16] (see also Bollobás [2]):

Theorem 3.4 (Azuma) *Let $X_1, \ldots, X_n$ be independent random variables, with X_k taking values in a set A_k for each k. Suppose that the (measurable) function $f : \prod A_k \to \mathbf{R}$ satisfies $|f(x) - f(x')| \leq c_k$ whenever the vectors x and x' differ only in the k'th coordinate. Let Y be the random variable $f(X_1, \ldots, X_n)$. Then for any $t > 0$,* $\Pr[\,|Y - \mathbf{E}Y| \geq t] \leq 2\exp(-2t^2/\sum c_k^2)$.

Definition 3.5 *Call a pair (G,π) λ-*atypical *if either $|f^+(G,\pi) - \mathbf{E}_{\mathcal{G}} f^+(G,\pi)| > \lambda$ or $|f^-(G,\pi) - \mathbf{E}_{\mathcal{G}} f^-(G,\pi)| > \lambda$.*

Claim 3.6 (atypicality) *For any partition π and any $\lambda > 0$, the probability, over random $G \in \mathcal{G}_{4n,p,r}$, that (G,π) is λ-atypical is at most* $4\exp(-4\lambda^2 T^2)$.

Proof Equating G with the collection $\{\chi(u,v) : \{u,v\} \in V^{(2)}\}$ of indicator variables for the edges, $f^+(G,\pi)$ is a function of these independent r.v.s: it is the sum over k-increasing moves (i,j) that the move is both proposed and accepted.

Varying one $\chi(u,v)$ while leaving all other χ's fixed, the probability of *proposing* a particular move (i,j) taking k to $k+1$ is unchanged at $1/(2n)^2$. The probability of *accepting* (i,j) changes only for the $4n$ values of i,j for which $\{i,j\} \cap \{u,v\} \neq \emptyset$. Even then the move's cost change X is altered by at most 1, and the derivative of the acceptance function A is bounded in absolute value by $1/4T$, so the acceptance probability $A(X)$ changes by at most $1/4T$. Thus changing a single $\chi(u,v)$ changes f^+ by at most $c(u,v) = \frac{1}{(2n)^2} \cdot 4n \cdot \frac{1}{4T} = \frac{1}{4nT}$; summing over all $\binom{4n}{2}$ pairs $\{u,v\}$ gives $\sum c(u,v)^2 \leq 1/2T^2$, whereupon Azuma's inequality yields the claim. ■

4 Progress of Annealing: Sketch

In this section we begin by proving that for any $\varepsilon > 0$, $p - r = n^{-1/6+2\varepsilon}$, $T = n^{1/2+\varepsilon}$, $\lambda = n^{-1/6}$, and any $t = O(n^{2/3+\varepsilon})$, in the course of t moves it is almost certain that a λ-atypical state is never encountered. Presuming that this is so, the probabilities for k increasing (or decreasing or remaining the same) are closely bounded by Corollary 3.3.

We go on to present an intuitive justification that with these parameter values, and assuming an initial imbalance of $k = \Omega(\sqrt{n}\,)$, the pessimistic values for these bounds on the transition probabilities suffice to drive k from $\Omega(n^{1/2})$ to $\Omega(n^{2/3})$ within t moves. For now we simply note that this is a comparatively huge imbalance, and do not consider completion to $k = n$.

In later sections we will tighten up what is essentially the same reasoning. First, if we do not presume that $k = \Omega(\sqrt{n}\,)$ initially, getting to such a k calls for both a greater number of moves and a larger T (and consequently a smaller λ), requiring a more sophisticated argument to show that no atypical state is likely to be encountered (Claim 5.2). Then, it is proved that the pessimistic values for the probabilities of k increasing, decreasing, or unchanging, when interpreted as the transition probabilities for a Markov chain (which k is not), do indeed yield pessimistic bounds for k (Theorem 6.2). Finally, from the point $k = \Omega(n)$ at which these arguments leave off, different arguments show that after about n^2 additional moves, with high probability k is exactly n (Subsection 6.3).

Claim 4.1 *Let $\lambda = n^{-1/6}$ and $T = n^{1/2+\varepsilon}$, and let t be any time bound $t = O(n^{2/3+\varepsilon})$. For a random $G \in \mathcal{G}_{4n,p,r}$ and any initial partition $\pi(0)$: with overwhelming probability, the Metropolis process run at temperature T for t moves encounters no λ-atypical partition of G.*

Proof In addition to $\pi(0)$, let the sequence of $t = O(n^{2/3+\varepsilon})$ proposed moves be given. For any such sequence, over all possible choices of accepting or rejecting each move, there are 2^τ states possible at time τ, for a set of $2^{t+1} - 1$ states possible over all t moves. By Claim 3.6, for a random graph G the *expected* number of λ-atypical states in this set is at most $2^{t+1} \cdot 4\exp(-4\lambda^2 T^2)$, which for the given parameter values is $\exp(O(n^{2/3+\varepsilon}) - 4n^{2/3+2\varepsilon}) = \exp(-\Omega(n^{2/3+2\varepsilon}))$.

It follows that it is exponentially unlikely that *any* λ-atypical state exists among these $2^{t+1} - 1$ states, and in particular that there is unlikely to be one among the $t+1$ states in the actual "trajectory" where the Metropolis rule determines the acceptances and rejections. ∎

Presuming now that no λ-atypical states are encountered, by Corollary 3.3 and the definition of λ-atypicality, at every step,

$$
\begin{aligned}
f^+ &\geq \tfrac{1}{8} + \Omega(\min\{k(p-r)/T, 1\}) - O(k/n) - \lambda \\
&= \tfrac{1}{8} + \Omega(\min\{kn^{-2/3+\varepsilon}, 1\}) - O(k/n) - O(n^{-1/6})
\end{aligned}
$$

which for $\Omega(n^{1/2}) \leq k \leq O(n^{2/3})$ is $f^+ = \frac{1}{8} + \Omega(\min\{kn^{-2/3+\varepsilon}, 1\})$. Symmetrically, $f^- = \frac{1}{8} - \Omega(\min\{kn^{-2/3+\varepsilon}, 1\})$.

Commencing with $k = \Omega(n^{1/2})$, the *bias* $f^+ - f^- = \Omega(\min\{kn^{-2/3+\varepsilon}, 1\}) = \Omega(n^{-1/6+\varepsilon})$ can naively be expected, over $n^{2/3}$ moves, to contribute an imbalance of $\Omega(n^{1/2+\varepsilon})$. Over the same time, as the transition probabilities are all approximately $\frac{1}{8}$, the number of k-increasing moves has variance $\Theta(n^{2/3})$ and is exponentially unlikely to differ by more than $O(n^{1/2})$ from its nominal value; the same is true for k-decreasing moves; and so with exponential likelihood the final imbalance k will be $\Omega(n^{1/2+\varepsilon}) \pm O(n^{1/2}) = \Omega(n^{1/2+\varepsilon})$.

Another $n^{2/3}$ moves starting from this point would by the same reasoning produce $k = \Omega(n^{1/2+2\varepsilon})$, with $\Theta(1/\varepsilon)$ repetitions yielding $k = \Omega(n^{2/3})$. The total number of moves used, $\Theta(1/\varepsilon) \cdot n^{2/3}$, is within the allotment $t = O(n^{2/3+\varepsilon})$.

Again, this merely provides intuition for how k grows: we will take it up again formally after showing that λ-atypical states can be avoided under the more rigorous parameter values that will be required.

5 An Improved Argument

The previous section's proof that no atypical state is encountered could be thought of as an adversary argument: even an adversary with freedom to accept or reject each move cannot guide the process to a partition π for which (G, π) is atypical.

To allow extension of the result to longer annealing runs with more exacting parameter values, in this section we make a roughly analogous argument which capitalizes on the fact that the adversary actually has limited freedom. Specifically, for almost all $G \in \mathcal{G}_{4n,p,r}$, and every bisection π, the transition probabilities from π must "on average" lie within a small range about their expectation over G. If the adversary is confined to choosing a probability within this range and then respecting a fair coin toss, there is little chance it can alter the outcome. Thus in t moves it makes far fewer than t decisions, allowing a larger t than previously.

For a given graph G, let $P(\pi, \pi') = P_G(\pi, \pi')$ be the probability that from state π the transition π' is made (generated and accepted), and let $\widehat{P}(\pi, \pi') = \mathbf{E}_{\mathcal{G}} P(\pi, \pi')$. For a given G and π, let $\|P - \widehat{P}\| = \|P(\pi, \cdot), \widehat{P}(\pi, \cdot)\| = \sum_{\pi'} |P(\pi, \pi') - \widehat{P}(\pi, \pi')|$. Call a pair (G, π) "rough" if $\|P - \widehat{P}\|$ exceeds some threshold, which we set to be $O(\sqrt{n}/T)$.

Claim 5.1 *For any temperature $T > 0$, with overwhelming probability over $G \in \mathcal{G}_{4n,p,r}$, for all π, $\|P - \widehat{P}\| = O(\sqrt{n}/T)$.*

Proof Let X' be the cost increase associated with a possible transition $\pi \mapsto \pi'$, for $\pi' \neq \pi$. First,

$$\begin{aligned}
\mathbf{E}_{\mathcal{G}} |P(\pi, \pi') - \mathbf{E}_{\mathcal{G}} P(\pi, \pi')| &= \mathbf{E}_{\mathcal{G}} \sqrt{[P(\pi, \pi') - \mathbf{E}_{\mathcal{G}} P(\pi, \pi')]^2} \\
&\leq \sqrt{\mathbf{E}_{\mathcal{G}} [P(\pi, \pi') - \mathbf{E}_{\mathcal{G}} P(\pi, \pi')]^2} \\
&= \frac{1}{4n^2} \sqrt{\mathbf{E}_{\mathcal{G}} [A(X') - \mathbf{E}_{\mathcal{G}} A(X')]^2}
\end{aligned} \tag{4}$$

$$
\begin{aligned}
&\le \frac{1}{4n^2}\sqrt{\mathbf{E}_{\mathcal{G}}[A(X') - A(\mathbf{E}_{\mathcal{G}}X')]^2} && (5)\\
&\le \frac{1}{4n^2}\frac{1}{4T}\sqrt{\mathbf{E}_{\mathcal{G}}[X' - \mathbf{E}_{\mathcal{G}}X']^2} && (6)\\
&= \frac{1}{16n^2T}\sqrt{\mathrm{Var}_{\mathcal{G}}X'}\\
&= \frac{1}{16n^2T}O(\sqrt{n}\,). && (7)
\end{aligned}
$$

Inequality (4) is from Jensen's inequality and concavity of square root [8, eq. V.8.6]; (5) holds because $\xi = \mathbf{E}X$ minimizes $\mathbf{E}(X-\xi)^2$; (6) because the absolute value of the gradient of A is bounded by $1/4T$; and (7) follows by regarding X' as a sum and difference of Bernoulli r.v.s (see eq. (2)).

It follows that

$$
\begin{aligned}
\mathbf{E}_{\mathcal{G}}\|P - \widehat{P}\| &= \sum_{\pi'} \mathbf{E}_{\mathcal{G}}|P(\pi,\pi') - \widehat{P}(\pi,\pi')|\\
&\le 2\sum_{\pi'\ne\pi} \mathbf{E}_{\mathcal{G}}|P(\pi,\pi') - \widehat{P}(\pi,\pi')|\\
&= O(\sqrt{n}/T). && (8)
\end{aligned}
$$

Next we claim that for any π, and any $c > 0$,

$$
\Pr_{\mathcal{G}}\left(\Big|\, \|P(\pi,\cdot\,) - \widehat{P}(\pi,\cdot\,)\| - \mathbf{E}_{\mathcal{G}}\|P(\pi,\cdot\,) - \widehat{P}(\pi,\cdot\,)\|\,\Big| \ge \frac{c\sqrt{n}}{T}\right) \le 2\exp(-c^2 n). \qquad (9)
$$

As usual this follows from applying the method of bounded differences, in this case to $\|P(\pi,\cdot\,) - \widehat{P}(\pi,\cdot\,)\|$, viewed as a function of G. Changing any edge indicator variable $\chi(u,v)$ does not affect any $\widehat{P}(\pi,\cdot\,)$. It changes $P(\pi,\pi')$ for only $4n$ states π', and then — for $\pi' \ne \pi$ — by at most $\frac{1}{4n^2}\cdot\frac{1}{4T}$: the probability of generating the state times the maximum possible change in the probability of accepting it. It therefore changes $P(\pi,\pi)$ by at most the same total amount $\frac{1}{4nT}$. So each $\chi(u,v)$ causes a change to $\|P(\pi,\cdot\,) - \widehat{P}(\pi,\cdot\,)\|$ of $c(u,v) \le 1/2nT$, resulting in $\sum c(u,v)^2 \le 2T^{-2}$, whence (9) follows by Azuma's inequality.

Combining equations (8) and (9), for any π,

$$
\Pr_{\mathcal{G}}\left(\|P(\pi,\cdot\,) - \widehat{P}(\pi,\cdot\,)\| \ge O\left(\frac{\sqrt{n}}{T}\right) + \frac{c\sqrt{n}}{T}\right) \;\le\; 2\exp(-c^2 n).
$$

The probability over $G \in \mathcal{G}_{4n,p,r}$ that there is any rough π is at most the expected number of π's which are rough, and since there are exactly $\binom{4n}{2n}$ possible values of π, this is $\exp\left(O(4n) - c^2 n\right) = \exp\left(-\Omega(n)\right)$ for c sufficiently large.

That is, all but a fraction $\exp\left(-\Omega(n)\right)$ of graphs $G \in \mathcal{G}_{4n,p,r}$ have no rough partitions. ∎

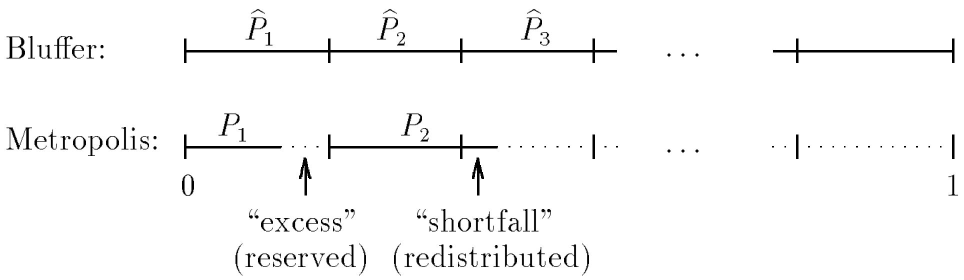

Figure 1: Sample space for Bluffer and Metropolis

Claim 5.2 *For any temperature* $0 < T \le \mathrm{poly}(n)$ *and any tolerance* λ*, the Metropolis process run at temperature* T *for* $t = \Theta(\lambda^2 T^3/\sqrt{n}\log^2 n)$ *steps encounters a* λ*-atypical state with probability at most* $\exp(-\Omega(\lambda^2 T^2/\log^2 n))$.

Proof Acting in parallel with the Metropolis process, we introduce a second random process, the *bluffer*, which should shadow it closely without knowledge of the graph G.

The bluffer is defined as follows. Given the current state $\hat{\pi}$, the next state $\hat{\pi}'$ is chosen as π_i with probability $\widehat{P}_i = \widehat{P}(\hat{\pi}, \pi_i)$. This will be implemented by partitioning $[0, 1)$ into intervals of length $\widehat{P}_i$, and selecting $\hat{\pi}' = \pi_i$ when the random "coin toss" α falls into the i'th interval. This "bluffer's mapping" taking $\hat{\pi}$ to $\hat{\pi}'$ is applied to $\hat{\pi}(\tau)$ to yield $\hat{\pi}(\tau+1)$.

We modify the bluffer's procedure according to a *correction tape*: a set of $O(t\sqrt{n}/T)$ times τ_j and associated moves $m(\tau_j)$. For any time τ_j, $\hat{\pi}(\tau_j)$ will be the state obtained from $\hat{\pi}(\tau_j - 1)$ by applying move $m(\tau_j)$, rather than the state obtained by applying the bluffer's mapping. Note that an initial state $\hat{\pi}(0)$, a random sequence $(\alpha(\tau))_{\tau=1}^{t}$, and a correction tape, completely determine the bluffer's sequence $(\hat{\pi}(\tau))_{\tau=1}^{t}$.

Base a trajectory of the Metropolis process on the same sample space of α's as used for the bluffer, as follows. For any $\pi = \hat{\pi}$, take as a starting point the same partitioning of $[0, 1)$ into intervals of length $\widehat{P}_i$. Now — given the graph G — imagine overlaying each of these intervals with an interval of length $P_i = P_G(\pi, \pi_i)$, matching up the left ends. (Refer to Figure 1.) When $P_i < \widehat{P}_i$, the P_i-long interval is "labeled" π_i, and the "excess" interval of length $\widehat{P}_i - P_i$ is reserved. When $P_i > \widehat{P}_i$, the label π_i is applied to the interval $\widehat{P}_i$, giving a "shortfall" of $P_i - \widehat{P}_i$. Because the total of the excesses is equal to the total of the shortfalls, we end by arbitrarily distributing the labels that fell short into the reserved excess spaces. By construction, each π_i labels a collection of intervals of total length P_i, so $\pi' = \pi_i$ is chosen with the correct Metropolis probability $P_G(\pi, \pi_i)$.

Furthermore, if $\hat{\pi}(\tau) = \pi(\tau)$, the bluffer's uncorrected successor for $\hat{\pi}(\tau)$ and the Metropolis's $\pi(\tau+1)$ are equal with probability $1 - \frac{1}{2}\|P - \widehat{P}\|$, over random $\alpha(\tau+1) \in [0, 1)$: they are unequal only if $\alpha(\tau+1)$ falls into one of the "excess" regions, and the total size of these is $\frac{1}{2}\|P - \widehat{P}\|$.

We now claim that for any graph G which is not rough (which satisfies $\|P - \widehat{P}\| = O(\sqrt{n}/T)$ for all bisections π), from any initial state $\hat{\pi}(0) = \pi(0)$, with high probability over random sequences $\alpha(\tau)$, some correction tape makes $\hat{\pi}(\tau) = \pi(\tau)$ for all $0 \leq \tau \leq t$. To see this, we will "reverse-engineer" the correction tape: the "bad" times τ_i are those at which the bluffer's choice for the state following $\hat{\pi}(\tau_i - 1)$ — given $\hat{\pi}(\tau_i - 1) = \pi(\tau_i - 1)$ — fails to agree with the true Metropolis state $\pi(\tau_i)$; the suggested moves $m(\tau_i)$ are those required to take state $\hat{\pi}(\tau_i - 1)$ to state $\pi(\tau_i)$. It should be clear that this construction makes the corrected bluffer's sequence identical to the Metropolis trajectory. It remains to verify that the number of bad times is small, so that the correction tape is of acceptable brevity.

A given time τ is bad only if α falls into one of the "excess" intervals, which occurs with probability at most $\frac{1}{2}\|P - \widehat{P}\| \leq O(\sqrt{n}/T)$. Thus the number of bad times is distributed binomially as $B(t, O(\sqrt{n}/T))$. Over the space of coin tosses, the probability that the total number of bad moves is $O(t\sqrt{n}/T)$ is $1 - \exp(-\Omega(t\sqrt{n}/T))$, which by choice of t is $1 - \exp(-\Omega(\lambda^2 T^2/\log^2 n))$.

We now close the argument. For any initial state and with overwhelming probability over graphs and random coin tosses, for some correction tape, the sequence $(\hat{\pi}(\tau))_{\tau=1}^{t}$ produced by a bluffer (without access to G) is identical to the Metropolis trajectory $(\pi(\tau))_{\tau=1}^{t}$. Since there are only

$$\binom{t}{O(t\sqrt{n}/T)} = \exp(O(\sqrt{n}\, t \log t/T))$$

possibilities for the set of times at which the corrector can act, and

$$(4n^2)^{O(t\sqrt{n}/T)} = \exp(O(\sqrt{n}\, t \log n/T))$$

possibilities for how it can act at those times, there are only $\exp(O(t\sqrt{n}\log n/T))$ possible correction tapes. (Note that, by assumption, T is polynomial in n, and hence $\log t = O(\log n)$.)

Given an initial state $\pi(0)$ and the sequence of coin tosses $\alpha(\tau)$, run the bluffer in turn with all possible correction tapes, producing a set of $\exp(O(t\sqrt{n}\log n/T))$ states. Only now choose a random $G \in \mathcal{G}_{4n,p,r}$. Applying Claim 3.6, the probability that any state in the set is λ-atypical for G is at most $t \cdot \exp(O(t\sqrt{n}\log n/T)) \cdot 4\exp(-4\lambda^2 T^2)$, which is $\exp(-\Omega(\lambda^2 T^2))$, by choice of t. Since, with overwhelming probability, the set includes the Metropolis trajectory $\pi(\tau)$, it is overwhelmingly likely that no λ-atypical state is encountered by the Metropolis algorithm. ■

Specializing the parameters T, λ, and t appearing in Claim 5.2 we obtain the following.

Corollary 5.3 *Let $\varepsilon > 0$, $T = n^{5/6+\varepsilon}$, $\lambda = n^{-1/2}$, $t = \Theta(n^{1+2\varepsilon})$, and select $G \in \mathcal{G}_{4n,p,r}$. With overwhelming probability, the Metropolis process run at temperature T for t steps starting at an arbitrary initial state encounters no λ-atypical state.*

6 Progress of Annealing: Detailed Analysis

We now refine the argument sketched in Section 4 to demonstrate that the Metropolis process on bisections, when run at an appropriate temperature, converges rapidly to the optimal bisection with overwhelming probability.

Theorem 6.1 *Let $\varepsilon > 0$, $p - r = n^{-1/6+2\varepsilon}$, and $T = n^{5/6+\varepsilon}$, and select $G \in \mathcal{G}_{4n,p,r}$. At constant temperature T, the Metropolis process on bisections of G reaches the unique minimum bisection in time $O(n^{2+\varepsilon})$, with overwhelming probability.*

Proof The result follows immediately from Lemmas 6.3, 6.4, and 6.5, which will be established later in the section. ■

It should be noted that the constant implicit in the definition of "overwhelming probability" depends on ε, but on no other parameters.

The analysis leading to Theorem 6.1 tracks the Metropolis process through three regimes. These are now briefly described in order to map out the general shape of the proof. The first two last for only $n^{1+\varepsilon}$ time, and hence, by Corollary 5.3, pass without the process visiting atypical states.

1. The first regime is in force until the imbalance k of the partition first reaches a value at least $\sqrt{n}$. (For a fortuitous initial partition, this condition may already hold at the outset.) The evolution of the imbalance k as a function of time is sufficiently close to that of an unbiased random walk on $\mathbf{N}$, that we are able to argue (in Subsection 6.1) that there is an appreciable probability that k reaches $\sqrt{n}$ in time $O(n)$. Thus with overwhelming probability we pass to the second regime in time $O(n^{1+\varepsilon})$.

2. The second regime lasts until $k = \Omega(n)$. In Subsection 6.2 we show that the process experiences a definite "drift" towards states (partitions) with higher values of k, with the result that $k = \Omega(n)$ holds with overwhelming probability after only $O(n^{1+\varepsilon})$ further steps.

3. When $k = \Omega(n)$, convergence begins to slow, because the probability of proposing a k-increasing move diminishes (though the probability of *accepting* a given k-increasing move is obviously high). As a result, we are no longer covered by Corollary 5.3. Fortunately, the progress towards the optimal partition is now inexorable, as we are able to demonstrate by rather crude arguments in Subsection 6.3.

As hinted above, we track the Metropolis process by observing the imbalance k as a function of time. The latter process is not Markovian, as the transition probabilities are a function of a "hidden" variable, namely the current partition π. However, we shall be able to reason about the the process $k(t)$ by comparing it with a simple random walk on $\mathbf{N}$. The latter will be chosen so that it approximates the former, and is dominated by it in an appropriate sense. The technology for making this comparison is encapsulated in the following theorem.

Theorem 6.2 *Let $X(t)$, $t = 1, 2, \ldots$, be a random process on $\mathbf{N}$, not necessarily Markovian, having $X(t+1) \in \{X(t)-1, X(t), X(t)+1\}$ as its only allowed transitions. (The transition probabilities may depend on some hidden state or even an external stimulus.) Let $Y(t)$, $t = 1, 2, \ldots$, be a random walk on $\mathbf{N}$ that is totally independent of X, but is* *dominated by X in the following sense: if $X(t) = Y(t) = k$, then*

$$\Pr[X(t+1) = k+1] \;\geq\; \Pr[Y(t+1) = k+1]$$

and

$$\Pr[X(t+1) = k-1] \;\leq\; \Pr[Y(t+1) = k-1].$$

Furthermore, assume that the loop probabilities $\Pr[Y(t+1) = k \mid Y(t) = k]$ *for the random walk are at least* $\frac{1}{2}$*, for all* k*. Then the process* $X(t)$ *dominates* $Y(t)$*: for every final state* j*, every choice of initial states satisfying* $Y(0) \leq X(0) \leq j$*, and every time bound* t*, it is the case that*

$$\Pr[X(t') = j, \text{for some } t' \leq t] \geq \Pr[Y(t') = j, \text{for some } t' \leq t].$$

Proof We shall base the processes $X(t)$ and $Y(t)$ on a common underlying sample space, so that $X(0) \geq Y(0)$ implies $X(t) \geq Y(t)$ for all t. The result follows immediately.

The sample space consists of a sequence $\alpha(0), \alpha(1), \ldots$ of independent "coin tosses", the result of the t'th toss being a real number $\alpha(t)$ uniformly distributed in $[0, 1)$. Focus attention on time step t, and let $p_X^+ = \Pr[X(t+1) = X(t)+1]$ and $p_X^- = \Pr[X(t+1) = X(t)-1]$, with analogous definitions for p_Y^+ and p_Y^-. (The probabilities p_X^+ and p_X^- may depend on a hidden state or external stimulus: the argument is not affected.) Then the next state of the process X is given by

$$X(t+1) = \begin{cases} X(t) - 1 & \text{if } \alpha(t) < p_X^-, \\ X(t) + 1 & \text{if } \alpha(t) \geq 1 - p_X^+, \\ X(t) & \text{otherwise}, \end{cases}$$

and similarly for the process Y.

Observe that the high loop probabilities of process Y, combined with dominance of the transition probabilities of Y by those of X, ensure that $p_Y^+ \leq \frac{1}{2}$ and $p_X^- \leq \frac{1}{2}$. Thus it cannot happen that $X(t+1) = X(t) - 1$ and $Y(t+1) = Y(t) + 1$, for the former implies $\alpha(t) < p_X^- \leq \frac{1}{2}$ and the latter $\alpha(t) \geq 1 - p_Y^+ \geq \frac{1}{2}$.

The proof of $X(t) \geq Y(t)$ for all t is by induction on t, the case $t = 0$ being satisfied by hypothesis. If $X(t) > Y(t)$ then it is immediate that $X(t+1) \geq Y(t+1)$ by the previous observation. If $X(t) = Y(t)$ then there are two possible bad cases: either Y increases and X does not; or Y is unchanged and X decreases. But Y increases only if $\alpha(t) \geq 1 - p_Y^+ \geq 1 - p_X^+$, in which case X also increases; similarly a decrease in X requires $\alpha(t) < p_X^- \leq p_Y^-$, which in turn requires Y to decrease. ∎

The machinery required for analyzing the three regimes described earlier is now in place.

6.1 Small imbalance

Here we consider the Metropolis process on partitions with imbalance at most $\sqrt{n}$. We show that the process exhibits the following favorable behavior.

Lemma 6.3 *Let $p-r$ and T be as in Theorem 6.1, and select $G \in \mathcal{G}_{4n,p,r}$. At constant temperature T and starting at any initial state, the Metropolis process on bisections of G reaches a bisection with imbalance at least $\sqrt{n}$ in time $O(n^{1+\varepsilon})$, with overwhelming probability.*

Proof Let $\lambda = n^{-1/2}$. By Corollary 5.3, with overwhelming probability no λ-atypical states are encountered in the first $O(n^{1+\varepsilon})$ steps of the Metropolis process. It is therefore enough to prove the result under the assumption that no λ-atypical states are encountered.

Set $b = \lceil \sqrt{n} \rceil$, and consider the random walk $Y(t)$ on $\mathbf{N}$ defined by transition probabilities

$$\Pr[Y(t+1) = k+1 \mid Y(t) = k] = \begin{cases} u & \text{if } 0 \le k < b, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\Pr[Y(t+1) = k-1 \mid Y(t) = k] = \begin{cases} d & \text{if } 0 < k \le b, \\ 1 & \text{if } k > b, \end{cases}$$

where $u = \frac{1}{8} - \delta$, $d = \frac{1}{8} + \delta$, and $\delta > 0$ is a constant to be chosen shortly. (Loop probabilities are implicitly defined by complementation.) Let $k(t)$, $t = 0, 1, \ldots$, be the sequence of imbalances recorded as the Metropolis process evolves from a random initial state. Under the assumption that the current state is not λ-atypical, Corollary 3.3 assures us that k-increasing and k-decreasing transitions both occur with probability $\frac{1}{8} \pm O(n^{-1/2})$ when $k \le b = \lceil \sqrt{n} \rceil$. Thus, for some choice of $\delta = O(n^{-1/2})$ the random walk $Y(t)$ stands in relation to $k(t)$ as its namesake does to the process $X(t)$ in Theorem 6.2.[1] We are thus reduced to showing that the random walk Y reaches b in time $O(n^{1+\varepsilon})$ with overwhelming probability when started in state 0.

For $0 \le k \le b$, let μ_i be the expected time for the random walk Y to hit state b, given that the initial state is i. According to Feller [7, eq. XIV.3.3], the general solution for μ_i is[2]

$$\mu_i = \frac{i}{d-u} + A + B\left(\frac{d}{u}\right)^i, \tag{10}$$

where A and B are constants that must be fitted to the boundary conditions; in our case, these are $\mu_b = 0$ and $\mu_0 = 1 + u\mu_1 + (1-u)\mu_0$, the latter simplifying to $\mu_0 - \mu_1 = u^{-1}$. Substituting the general solution from equation (10) into each of these in turn yields

$$A + B\left(\frac{d}{u}\right)^b + \frac{b}{d-u} = 0$$

and

$$B\left(1 - \frac{d}{u}\right) - \frac{d}{u(d-u)} = 0.$$

[1]The transition probabilities for Y when $Y(t)$ exceeds b are artificially defined to ensure a formal correspondence with the condition of Theorem 6.2; in fact we are only interested in the evolution of $k(t)$ until $k(t) \ge b$.

[2]Feller has the convention $u + d = 1$, but his derivation is valid for the general case.

Solving for A and B, and substituting in $\mu_0 = A + B$, we obtain

$$\mu_0 = -\frac{b}{d-u} + \frac{d}{(d-u)^2}\left[\left(\frac{d}{u}\right)^b - 1\right]$$

Noting that $(d-u)^2 = 4\delta^2 = O(n^{-1})$ and $d/u = (1+8\delta)/(1-8\delta) = 1 + O(n^{-1/2})$, we see immediately that $\mu_0 = O(n)$, and hence $\mu_i = O(n)$ for all $i \le b$. By Markov's inequality, there is a non-zero constant probability that the process $Y(t)$ will hit state b in $O(n)$ steps starting at any initial state $i \le b$. Thus, starting in state 0, the process will hit state b in $O(n^{1+\varepsilon})$ steps, with overwhelming probability. ∎

6.2 Moderate imbalance

Up to $\sqrt{n}$, the imbalance k evolves roughly as an unbiased random walk on $\mathbf{N}$. Beyond this threshold a positive bias takes hold, and k increases at faster and faster rates. The effect can be quantified, and leads to the following lemma.

Lemma 6.4 *Let $c > 0$ be a sufficiently small constant, $p-r$ and T be as in Theorem 6.1, and select $G \in \mathcal{G}_{4n,p,r}$. Suppose that the Metropolis process on bisections is allowed to evolve at constant temperature T, starting at any initial state. Then the transit time from the first partition with imbalance $k \ge \sqrt{n}$ to the first partition with imbalance $k \ge cn$ is $O(n^{1+\varepsilon})$, with overwhelming probability.*

Proof Let $\lambda = n^{-1/2}$, and consider the evolution of the Metropolis process starting at the first state with imbalance $k \ge \sqrt{n}$. By Corollary 5.3 and Lemma 6.3, with overwhelming probability no λ-atypical states are encountered in the following $O(n^{1+\varepsilon})$ steps of the Metropolis process. It is therefore enough to prove the result under the assumption that no λ-atypical states are encountered.

We show that in a sequence of "phases" the imbalance will rise to $k \ge cn$ with overwhelming probability. In a typical phase, k will increase from some initial value i to a final value at least $b = \min\{\lceil in^{\varepsilon}\rceil, cn\}$ in time $O(n^{1+\varepsilon})$. Since the number of phases is constant, the overall transit time is also $O(n^{1+\varepsilon})$.

We proceed to analyze a single phase with initial state $i \ge \sqrt{n}$. Set $a = \lfloor i/2 \rfloor$, and consider the random walk $Y(t)$ on $\mathbf{N}$ defined by transition probabilities

$$\Pr[Y(t+1) = k+1 \mid Y(t) = k] \quad = \quad \begin{cases} u & \text{if } a \le k < b, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\Pr[Y(t+1) = k-1 \mid Y(t) = k] \quad = \quad \begin{cases} d & \text{if } a < k \le b, \\ 1 & \text{if } 0 < k \le a \text{ or } k > b, \end{cases}$$

where $u = \frac{1}{8} + \delta$, $d = \frac{1}{8} - \delta$, and $\delta > 0$ will be chosen presently. Let $k(t)$, $t = 0, 1, \ldots,$ be the sequence of imbalances recorded as the Metropolis process evolves from an initial state with imbalance $k = i$. Under the assumption that the current state is not λ-atypical, Corollary 3.3 assures us that k-increasing transitions occur with probability $\frac{1}{8}+$

$\Omega(\min\{in^{-1+\varepsilon}, 1\})$, and k-decreasing transitions with probability $\frac{1}{8} - \Omega(\min\{in^{-1+\varepsilon}, 1\})$, provided $a \le k \le b$. Thus, for an appropriate choice of $\delta = \Theta(\min\{in^{-1+\varepsilon}, 1\})$, the random walk $Y(t)$ and process $X(t) = k(t)$ satisfy the conditions of Theorem 6.2. For future reference, note that $\delta = \Omega(n^{-1/2+\varepsilon})$. We are thus reduced to showing that the random walk Y reaches b in time $O(n^{1+\varepsilon})$ with overwhelming probability when started in state i.

Feller shows [7, eq. XIV.2.4] that the probability, given $Y(0) = i$, that the event[3] $Y(t) = a$ occurs before the event $Y(t) = b$ is equal to $(r^{i-a} - r^{b-a})/(1 - r^{b-a})$, where $r = d/u$. Now, $r = 1 - \Omega(\delta) = 1 - \Omega(n^{-1/2+\varepsilon})$, and $b - a \ge i - a = \lceil i/2 \rceil = \Omega(n^{1/2})$. So the probability of "ruin" is $\exp(-\Omega(n^{\varepsilon}))$, and the event $Y(t) = b$ occurs first with overwhelming probability.

Again from Feller [7, eq. XIV.3.4], the expected time to hit either a or b is

$$-\frac{i-a}{2\delta} + \frac{b-a}{2\delta} \cdot \frac{1 - r^{i-a}}{1 - r^{b-a}},$$

which is $O(n)$, since $b - a = O(\min\{in^{\varepsilon}, cn\})$, and $\delta = \Omega(\min\{in^{-1+\varepsilon}, 1\})$. Thus the hitting time for state b is $O(n^{1+\varepsilon})$ with overwhelming probability. ■

6.3 Large imbalance

When k is large, progress towards partitions with higher imbalance is slowed, since the probability of *selecting* a k-increasing move declines. Ultimately, when $k = n - 1$, the probability of selecting the unique k-increasing move is just $1/4n^2$. As a result, we can no longer remain within the time bound of Corollary 5.3. Fortunately, the Metropolis process is now safely locked on to the optimal solution, and rather crude arguments suffice to prove the lemma relating to the third and final regime.

Lemma 6.5 *Let $c > 0$, $p - r$ and T be as in Lemma 6.1, and select $G \in \mathcal{G}_{4n,p,r}$. Suppose that the Metropolis process on bisections is allowed to evolve at constant temperature T, starting at any bisection with imbalance $k \ge cn$. Then the hitting time for the unique minimal bisection with imbalance $k = n$ is $O(n^{2+\varepsilon})$, with overwhelming probability.*

Proof Denote the set of neighbors of vertex v in G by $G(v)$. Recall the terminology and notation of Section 2, in particular, that block R (block L) of the partition is the one with a majority of white (black) vertices. Call a white vertex v *reliable* if

$$|G(v) \cap R| - |G(v) \cap L| \ge n^{5/6+3\varepsilon/2};$$

symmetrically, a black vertex will be reliable if

$$|G(v) \cap L| - |G(v) \cap R| \ge n^{5/6+3\varepsilon/2}.$$

[3] so-called "ruin"

Intuitively, a reliable vertex is one that is will readily move to the "correct" side of the (L, R) partition if it is not already there, but is almost impossible dislodge once it has arrived at the correct side. More precisely: a proposed k-increasing (decreasing) move that nominates two reliable vertices will be accepted (rejected) with overwhelming probability. (The change in cost is in absolute value at least $2(n^{5/6+3\varepsilon/2}-1)$ and hence the acceptance (rejection) probability is at least $1-\exp(-\Omega(n^{\varepsilon/2}))$.

For a given partition π with imbalance $k \geq \lfloor cn/2 \rfloor$, given vertex v, and random $G \in \mathcal{G}_{4n,p,r}$, the probability that v is unreliable is very low: for a white vertex $v \in L$ the quantity $|G(v) \cap R| - |G(v) \cap L|$ is distributed as[4]

$$(n-k)\cdot B_r + (n+k)\cdot B_p - (n+k)\cdot B_r - (n-k-1)\cdot B_p$$

which has expectation $(2k+1)p - 2kr = \Omega(n(p-r)) = \Omega(n^{5/6+2\varepsilon})$. Similar calculations can be undertaken for white vertices in R, and black vertices in L or R, with the same result. By the bounded difference inequality (Theorem 3.4) or that of Chernoff, the probability that v is unreliable is $\exp(-\Omega(n^{2/3+4\varepsilon}))$.

Now, keeping π as before, consider the probability that some fixed set U of vertices are all unreliable, where $|U| = O(n^{5/6})$. A necessary condition for a white vertex $u \in U$ to be unreliable is

$$|G(v) \cap (R-U)| - |G(v) \cap (L-U)| < n^{5/6+3\varepsilon/2} + |U| \tag{11}$$

an event that, as before, occurs with probability at most $\exp(-\Omega(n^{2/3+4\varepsilon}))$. (Here we use the fact that $|U| = O(n^{5/6})$.) Observe that the events (11), where u ranges over U, are probabilistically independent, so the probability that *all* vertices in U are unreliable is at most $\exp\{-|U|\,\Omega(n^{2/3+4\varepsilon})\}$.

Still keeping π as before, consider the probability that G contains at least $\lceil n^{-2/3}(n-k)\rceil$ unreliable vertices. This probability is bounded above by the expected number of subsets $U \subset V$ of size $|U| = \lceil n^{-2/3}(n-k)\rceil = O(n^{5/6})$ that contain only unreliable vertices. The expected number of such sets U is

$$\begin{aligned}\binom{4n}{|U|}\exp\{-|U|\,\Omega(n^{2/3+4\varepsilon})\} &\leq \exp\{|U|\ln 4n - |U|\,\Omega(n^{2/3+4\varepsilon})\}\\ &\leq \exp\{-\Omega((n-k)n^{4\varepsilon})\}.\end{aligned}$$

Observe that this bound is exponentially smaller than the reciprocal of the number of partitions with imbalance k; thus, with overwhelming probability, if $G \in \mathcal{G}_{4n,p,r}$ is chosen at random, *every* k-partition with $k \geq cn$ has fewer than $\lceil n^{-2/3}(n-k)\rceil$ unreliable vertices.

There are two cases to analyze, depending on the size of $n-k$. If $n-k < n^{2/3}$, then *all* vertices are reliable, and the probability that a k-increasing move will be proposed and accepted is $((n-k)/2n)^2$ — to within an exponentially small additive term — and

[4] Here we re-use the notation from the proof of Claim 3.2.

certainly greater than $\frac{2}{3}((n-k)/2n)^2$; similarly, the probability that a k-decreasing move will be proposed and accepted is 0 — again to within an exponentially small additive term — and certainly less than $\frac{1}{3}((n-k)/2n)^2$. If $n-k \geq n^{2/3}$ then the total number of possible moves involving unreliable vertices is at most $2n^{1/3}(n-k) \leq 2n^{-1/3}(n-k)^2$, i.e., at most a factor $2n$ greater than the number of unreliable vertices. These potentially bad moves form only a small fraction of the total number of k-increasing moves that could be proposed, and it is easy to check that the bounds on transition probabilities computed in the case of small $n-k$ hold here also.

Finally, in the style of Lemmas 6.3 and 6.4, we introduce a random walk $Y(t)$ on $\mathbf{N}$ that is dominated by the Metropolis process in the sense of Theorem 6.2. Taking our cue from the earlier analysis, we define the transition probabilities for $Y(t)$ as follows:

$$\Pr[Y(t+1) = k+1 \mid Y(t) = k] \;=\; \begin{cases} u(k) & \text{if } a \leq k < n, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\Pr[Y(t+1) = k-1 \mid Y(t) = k] \;=\; \begin{cases} d(k) & \text{if } a < k \leq n, \\ 1 & \text{if } 0 < k \leq a \text{ or } k > n, \end{cases}$$

where $a = \lfloor cn/2 \rfloor$, $u(k) = \frac{2}{3}((n-k)/2n)^2$, and $d(k) = \frac{1}{3}((n-k)/2n)^2$. As before, by Theorem 6.2, we are reduced to showing that the random walk Y reaches state n in time $O(n^{2+\varepsilon})$ with overwhelming probability when started in state $i \geq cn$.

Observe that if the loop probabilities are factored out, so that only transitions to distinct states are recorded, the resulting random walk $Y'(t)$ has constant transition probabilities in the range $a < k < n$, namely, “up” with probability $\frac{2}{3}$ and “down” with probability $\frac{1}{3}$. Using the line of argument in the proof of Theorem 6.4, the expected time for $Y'(t)$ to hit the state n (starting at any state $i \geq cn$) is $O(n)$. In fact, it is not difficult to show that the expected number of visits to any given state is $O(1)$.

Now consider the actual random walk $Y(t)$. The expected sojourn time in a state k (corresponding to a single visit by $Y'(t)$) is $(2n/(n-k))^2$. Thus the expected total time spent by $Y(t)$ in state k is $O(n^2/(n-k)^2)$; summing over k, the expected hitting time for state n is $O(n^2)$, since $\sum_{i=1}^{\infty} i^{-2}$ is bounded. The state n will be reached in time $O(n^{2+\epsilon})$ with overwhelming probability. ■

References

[1] Y. Azar, A. Z. Broder, A. R. Karlin, N. Linial, and S. Phillips. Biased random walks. In *Proceedings of the 24th ACM Symposium on Theory of Computer Science*, pages 1–9, 1992.

[2] B. Bollobás. Martingales, isoperimetric inequalities and random graphs. In A. Hajnal, L. Lovász, and V. T. Sós, editors, *Combinatorics*, number 52 in Colloq. Math. Soc. János Bolyai, pages 113–139. North Holland, 1988.

[3] R. B. Boppana. Eigenvalues and graph bisection: An average-case analysis. Technical report, Rutgers University. Extended abstract appears in *Proceedings of the 28th IEEE Symposium on Foundations of Computer Science*, 1987, pp 280–285.

[4] T. Bui, S. Chaudhuri, T. Leighton, and M. Sipser. Graph bisection algorithms with good average case behavior. In *Proceedings of the 25th IEEE Symposium on Foundations of Computer Science*, pages 181–192, 1984.

[5] T. N. Bui. *Graph Bisection Algorithms.* PhD thesis, M. I. T., 1986.

[6] M. E. Dyer and A. M. Frieze. Fast solution of some random NP-hard problems. In *Proceedings of the 27th IEEE Symposium on Foundations of Computer Science*, pages 331–336, 1986.

[7] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, New York, NY, 3d edition, 1968.

[8] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. Wiley, New York, NY, 2nd edition, 1971.

[9] A. M. Frieze. A note on computing random permanents. Unpublished manuscript, 1989.

[10] M. R. Garey and D. S. Johnson. *Computers and Intractability.* Freeman, San Francisco, CA, 1979.

[11] M. R. Garey, D. S. Johnson, and L. Stockmeyer. Some simplified NP-complete graph problems. *Theoretical Computer Science*, 1:237–267, 1976.

[12] M. Jerrum. Large cliques elude the Metropolis process. *Random Structures and Algorithms*, 3(4):347–359, 1992.

[13] M. Jerrum and A. Sinclair. Approximating the permanent. *SIAM J. Comput.*, 18(6):1149–1178, Dec. 1989.

[14] M. R. Jerrum and A. Sinclair. Conductance and the rapid mixing property for Markov chains: The approximation of the permanent resolved. In *Proceedings of the 20th Annual ACM Symposium on Theory of Computing*, pages 235–244, 1988.

[15] S. Kirkpatrick, C. D. Gelatt, Jr., and M. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, May 1983.

[16] C. McDiarmid. On the method of bounded differences. In *London Mathematical Society Lecture Note Series*, volume 141, pages 148–188. Cambridge University Press, 1989.

[17] R. Motwani. Expanding graphs and average-case analysis of algorithms for matchings and related problems. In *Proceedings of the 21st Annual ACM Symposium on Theory of Computing*, pages 550–561, 1989.

[18] G. Sasaki. The effect of the density of states on the Metropolis algorithm. *Information Processing Letters*, 37:159–163, 1991.

[19] G. H. Sasaki and B. Hajek. The time complexity of maximum matching by simulated annealing. *Journal of the Association for Computing Machinery*, 35(2):387–403, Apr. 1988.

[20] G. B. Sorkin. *Theory and Practice of Simulated Annealing on Special Energy Landscapes*. PhD thesis, University of California at Berkeley, May 1991.