Work and information processing in a solvable model of Maxwell's demon

Dibyendu Mandal[a] and Christopher Jarzynski[b,1]

[a]Department of Physics, University of Maryland, College Park, MD 20742; and [b]Department of Chemistry and Biochemistry, and Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742

We describe a minimal model of an autonomous Maxwell demon, a device that delivers work by rectifying thermal fluctuations while simultaneously writing information to a memory register. We solve exactly for the steady-state behavior of our model, and we construct its phase diagram. We find that our device can also act as a "Landauer eraser", using externally supplied work to remove information from the memory register. By exposing an explicit, transparent mechanism of operation, our model offers a simple paradigm for investigating the thermodynamics of information processing by small systems.

Landauer's principle | nonequilibrium statistical mechanics

A system in thermal equilibrium undergoes random microscopic fluctuations, and it is tempting to speculate that an ingeniously designed device could deliver useful work by rectifying these fluctuations. The suspicion that this would violate the second law of thermodynamics has inspired nearly 150 years of provocative thought experiments (1–5), leading to discussions of the thermodynamic implications of information processing (6–12). Although both Maxwell (1) and Szilard (3) famously took the rectifying agent to be an intelligent being, later analyses have explored the feasibility of a fully mechanical "demon". There has emerged a kind of consensus, based largely on the works of Landauer (6) and Bennett (7, 8), and independently Penrose (13), according to which a mechanical demon can indeed deliver work by rectifying fluctuations, but in doing so it gathers information that must be written to physical memory. The eventual erasure of this information carries a thermodynamic cost, no less than $k_B T \ln 2$ per bit (*Landauer's principle*), which eliminates any gains obtained from the rectification of fluctuations.

The past few years have seen increased interest in the thermodynamics of information processing (14–19). Discussions of Maxwell's demon, Landauer's principle and related topics arise in contexts such as quantum information theory (20), the synthesis of artificial nanoscale machines (21), feedback control in microscopic systems (22–33), and single-photon cooling of atoms (34). Experiments have been performed with the explicit aim of testing theoretical predictions (27), including Landauer's principle (35). Moreover the consensus or "favored explanation" (36) described above is widely but not universally accepted, as suspicions persist that it assigns an unwarranted thermodynamic significance to random data (12, 36–39).

In spite of this attention, the field still lacks a tangible example or model of a device that converts heat into work at the expense of writing information. Discussions are often framed around general principles rather than a particular instance, and the demon is typically described in generic terms, as a system capable of performing microscopic feedback control, but otherwise unspecified. In this paper we propose an explicit, solvable model of a system that behaves as a Maxwell demon. Our device, which extracts energy from a single thermal reservoir and delivers it to raise a mass against gravity, is fully autonomous—it is neither manipulated by an external agent nor driven by an explicit thermodynamic force—but in order to lift the mass the device requires a memory register to which it can write information.

Briefly, in our model the device, or demon, is a three-state system that interacts with: a thermal reservoir, a mass that can be lifted or lowered, and a stream of *bits* that pass by the demon in sequence, as sketched in Fig. 1. The demon's dynamics consist of random transitions among its three states. These transitions are driven by thermal fluctuations from the reservoir, and are coupled to the bits and to the mass in a manner described in detail below. The model has three parameters: δ describes the initial statistical state of the bits, reflecting the initial ratio of 0's–1's; ϵ characterizes the weight of the mass; and τ is the duration of interaction with each bit in the stream. For any set of values (δ, ϵ, τ) the model reaches a unique periodic steady state, characterized by an average rate of work performed on the mass, and an average rate of information written to the bit stream.

We will first consider our model in the absence of an external load: the demon generates directed motion while writing information to the bits, but there is no provision for harnessing this motion to perform work. We then add a load by attaching a nonzero mass, as in Fig. 1, and we solve for the model's steady-state behavior. Our model exposes a specific mechanism for the operation of a mechanical Maxwell demon, allowing us to explore in detail the interplay between the gravitational pull on the mass and the changing information content of the stream of bits. Moreover, our demon is versatile: it is equally capable of acting as an eraser, using the energy of a falling mass to remove information from the memory register.

We now specify our model in detail, by introducing in turn its several elements (Fig. 2). The demon evolves by making thermally activated transitions among its three states, labelled A, B and C. We consider transitions in the direction $A \rightarrow B \rightarrow C \rightarrow A$ to be clockwise (CW), and those in the opposite direction to be counterclockwise (CCW), see Fig. 2A. The demon exhibits *directed rotation* if CW transitions occur with greater frequency than CCW transitions, or vice-versa. (Note that the demon must possess at least three states in order to exhibit directed rotation). To keep track of the net CW rotation, we introduce an integer variable χ, whose value increases by one unit whenever the demon makes a transition from C to A, and decreases by one unit with each transition from A to C.

We next describe the interaction between the demon and a single bit, with states labelled 0 and 1, Fig. 2B. The demon and bit together form a composite system with six states, $A0, \cdots C1$, depicted in Fig. 2C. The five lines, or edges, connecting pairs of states in this network specify the allowed transitions for the composite system. The demon can jump between states A and B, and between B and C, without involving the bit; these transitions are represented by the edges $A0$-$B0$, $B0$-$C0$, $A1$-$B1$ and $B1$-$C1$. Additionally, the demon can make a transition from C
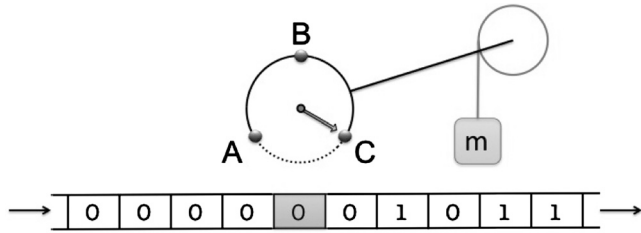
PHYSICS

Fig. 1. A sequence of bits moves at constant speed past the three-state demon, which interacts with the nearest bit (shaded) and a thermal reservoir (not shown). To model a positive external load $f > 0$, we imagine that a mass m is lifted by an amount Δh every time the demon makes a transition $C \rightarrow A$, and lowered with each transition $A \rightarrow C$. See text for details. For $f < 0$, the mass can be pictured as hanging off the rights side of the small circle, so that transitions $C \rightarrow A$ lower the mass and transitions $A \rightarrow C$ lift it.

to A if the bit simultaneously "flips" from 0 to 1, or from A to C if the bit flips from 1 to 0, as indicated by the line connecting $A1$ and $C0$. We model these transitions as a Poisson process, where R_{ij} is the probability per unit time to make a transition to state i, when the system is in state j, with $i, j \in \{A0, \cdots C1\}$. For the moment, we set $R_{ij} = 1$ for each of the ten allowed transitions (two per edge) depicted in Fig. 2C. These rates set the unit of time in our model. Because $R_{ij} = R_{ji}$ for every edge, we have implicitly assigned the same energy to all six states (40). Under these rates, the demon and bit relax toward equilibrium, in which all six states are equally likely. This relaxation occurs on a time scale of order unity: $\tau_r \sim 1$.

The network of states in Fig. 2C forms a linear chain. Because this chain contains no closed loops, the model cannot yet exhibit directed rotation, only back-and-forth excursions along the chain. To introduce the possibility of rotation, let us now imagine a *bit stream* (see Fig. 1): a sequence of bits arranged at equally spaced intervals along a tape that is pulled through at constant speed, for instance by a frictionless flywheel. The demon remains at a fixed location, and interacts, in the manner discussed above, only with the bit that is currently closest to it. Let τ^{-1} denote the rate at which the bits pass by the demon, each interacting with the demon for a time interval of duration τ before the next bit in the stream takes its place. Thus τ determines the extent to which the composite system approaches equilibrium during one such *interaction interval*; for $\tau \ll \tau_r \sim 1$ the system hardly evolves during the interval, whereas for $\tau \gg 1$ the demon and the bit effectively reach equilibrium. Finally, let b_n and b'_n denote, respectively, the incoming and outgoing state of the n'th bit in the stream. The state of any bit can change only when it is interacting with the demon.

The incoming bits are statistically independent of one another, each with probability p_0 to be in state 0, and p_1 to be in state 1. The *excess parameter*
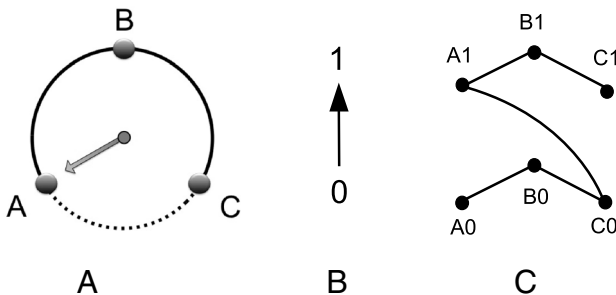

Fig. 2. Schematic depiction of the demon, the bit and their composite 6-state system. (A) The state of the demon is indicated by an arrow pointing in one of three directions (A, B, or C) on the face of a dial. (B) The bit is represented as an arrow pointing either up (1) or down (0). (C) Network depiction of the composite system, showing allowed transitions. The edge that connects $A1$ and $C0$ represents the coupling between the demon and bit.

$$\delta \equiv p_0 - p_1 \qquad [1]$$

quantifies the excess of 0's in the incoming bit stream.

To understand how the demon can exhibit directed rotation, in this paragraph let us consider the case in which every bit in the incoming stream is set to 0 (as in Fig. 1). The demon interacts with the n'th bit during the n'th interaction interval, $t_n \leq t < t_{n+1} \equiv t_n + \tau$. At the start of this interval, the composite system begins in state $A0$, $B0$ or $C0$, since $b_n = 0$. From $t = t_n$ to t_{n+1} the system evolves among the network of states depicted in Fig. 2C. It might repeatedly pass forward and back along the edge connecting $C0$ to $A1$, resulting in alternating increments and decrements of the counter $\chi(t)$. At the end of the interaction interval, if the system is found in state $A0$, $B0$ or $C0$ (i.e. if $b'_n = 0$) then we can infer that every transition $C0 \rightarrow A1$ was balanced by a transition $A1 \rightarrow C0$, hence $\Delta\chi_n \equiv \chi(t_{n+1}) - \chi(t_n) = 0$. If the system instead ends in $A1$, $B1$ or $C1$ ($b'_n = 1$), then the counter has advanced by one net unit: $\Delta\chi_n = +1$. At $t = t_{n+1}$, the n'th bit is replaced by the $(n + 1)$'th bit, and the next interval commences. Thus if the composite system is in state $B1$ at the end of one interval, then at the start of the next interval it is in state $B0$. This effective transition does not imply an actual change in the state of a given bit, but simply reflects the replacement of an outgoing bit in state 1 with an incoming bit in state 0. (Note that Fig. 2C depicts only the actual transitions that may occur during one interaction interval). Over time, the demon interacts with a sequence of bits, all initialized to 0, and the outgoing bit stream contains a record of the demon's rotary motion: each occurrence of an outgoing bit in state 1 indicates one full CW rotation, $\Delta\chi = +1$. Since the value of the counter can only increase or remain unchanged from one interval to the next, in the long run $\chi(t)$ grows with time and the demon undergoes directed CW rotation.

If the incoming stream were instead composed entirely of 1's, then full CW rotations would be prohibited, and full CCW rotations would be documented as outgoing 0's. For a more general distribution of incoming bits, the net change in the counter during the n'th interaction interval is

$$\Delta\chi_n = b'_n - b_n, \qquad [2]$$

and the outgoing stream provides partial information regarding the demon's gyrations.

The demon eventually reaches a periodic steady state in which its statistical behavior is the same from one interval to the next. If the outgoing bit stream is then characterized by values p'_0, p'_1 and $\delta' \equiv p'_0 - p'_1$, then the average number of full CW rotations per interaction interval is

$$\Phi \equiv \langle \Delta\chi_n \rangle = p'_1 - p_1 = \frac{1}{2}(\delta - \delta'). \qquad [3]$$

We will use Φ as our measure of directed rotation, and we will call it the *circulation*.

We have solved for the periodic steady state, obtaining (as described in the *Methods*)

$$\Phi(\delta; \tau) = \frac{\delta}{2}\left[1 - \frac{1}{3}K(\tau)\right], \qquad [4a]$$

where

$$K(\tau) = e^{-2\tau}\frac{(1 + 8\alpha + 4\sqrt{3}\beta) - (2 + 7\alpha + 4\sqrt{3}\beta)e^{-2\tau}}{3 - (2 + \alpha)e^{-2\tau}} \qquad [4b]$$

and $\alpha = \cosh(\sqrt{3}\tau)$, $\beta = \sinh(\sqrt{3}\tau)$. The function $K(\tau)$ decreases monotonically from $K(0^+) = 3$ to $K(\infty) = 0$, hence the magnitude of the circulation increases with τ, from $\Phi(\delta; 0^+) = 0$ to

$\Phi(\delta; \infty) = \delta/2$. These values are easily understood. When $\tau \to 0$, the probability to observe any transition during a given interaction interval vanishes, and therefore so does Φ. When $\tau \gg 1$, during each interaction interval the composite system has sufficient time to relax to equilibrium, with all six states in Fig. 2C equally likely, hence $\delta' = 0$ and $\Phi = \delta/2$ [Eq. **3**]. Interestingly, the average *rate* of rotations, $\omega = \Phi/\tau$, achieves its maximal value $\omega = \delta/6$ at $\tau = 0^+$, and then decreases with τ.

Since the quantity in square brackets in Eq. **4a** is non-negative, the demon effectively converts an excess of 0's or 1's into directed CW or CCW rotation. We now harness this rotation to an external load, by attaching a mass m to the demon in such a way that the mass is lifted a distance Δh whenever the demon makes a transition from C to A, and is lowered by Δh for the reverse transition (Fig. 1). The corresponding energy, $\pm mg\Delta h$, is exchanged with the thermal reservoir: with every transition $C0 \to A1$, heat is withdrawn from the bath to lift the mass, and with every transition $A1 \to C0$ that energy is released to the bath. (There is no exchange of energy with the flywheel that pulls the bits past the demon, in particular the flywheel does not contribute energy to lift the mass). To incorporate these considerations into our dynamics, we modify the transition rates in accordance with detailed balance (40):

$$\frac{R_{A1,C0}}{R_{C0,A1}} = e^{-f}, \qquad f \equiv \frac{mg\Delta h}{k_B T} > 0, \qquad \textbf{[5]}$$

where T is the temperature of the thermal reservoir and k_B is Boltzmann's constant, with all other rates R_{ij} unchanged. The parameter f quantifies a thermodynamic force that favors CCW rotations ($R_{C0,A1} > R_{A1,C0}$) as gravity tugs the mass downward. In terms of Fig. 2C, Eq. **5** effectively increases the energies of states $A1$, $B1$ and $C1$ by an amount $mg\Delta h$, relative to states $A0$, $B0$ and $C0$, reflecting the energy that is withdrawn from the reservoir during the transition $C0 \to A1$.

If the demon interacts with a single bit for a sufficiently long time then the two will reach equilibrium, with

$$p_{A0}^{eq} = p_{B0}^{eq} = p_{C0}^{eq} = \frac{e^f}{Z}, \qquad p_{A1}^{eq} = p_{B1}^{eq} = p_{C1}^{eq} = \frac{1}{Z}, \qquad \textbf{[6a]}$$

where $Z = 3(1 + e^f)$. After summing over the states of the demon, the equilibrium probabilities for the bit itself are found to satisfy

$$p_0^{eq} - p_1^{eq} = \tanh\left(\frac{f}{2}\right) \equiv \epsilon, \qquad \textbf{[6b]}$$

where the *weight parameter*, ϵ, is a rescaled version of the thermodynamic force, f.

Eq. **5** leaves us with some freedom in assigning the rates $R_{C0,A1}$ and $R_{A1,C0}$. We have chosen

$$R_{A1,C0} = 1 - \epsilon, \qquad R_{C0,A1} = 1 + \epsilon. \qquad \textbf{[7]}$$

With this choice, we are again able to solve analytically for the periodic steady state, obtaining (see *Methods*)

$$\Phi(\delta, \epsilon; \tau) = \frac{\delta - \epsilon}{2}\left[1 - \frac{1}{3}K(\tau) + \frac{\epsilon\delta}{6}J(\tau, \epsilon\delta)\right] \qquad \textbf{[8a]}$$

with

$$J(\tau, \epsilon\delta) = \frac{(1 - e^{-\tau})[2e^{-2\tau}(\alpha + \sqrt{3}\beta - 1)]^2}{[3(1 - \epsilon\delta e^{-\tau}) - (1 - \epsilon\delta)(2 + \alpha)e^{-2\tau}][3 - (2 + \alpha)e^{-2\tau}]} \qquad \textbf{[8b]}$$

and K, α and β as in Eq. **4**. These results extend to negative values of f, if we interpret these as indicating that gravity exerts a *clockwise* torque (see caption of Fig. 1). Eq. **8**, our central result, is then valid for $|\epsilon| < 1$, $|\delta| \leq 1$, and $0 < \tau < \infty$.

In the limits $\tau = 0^+$ and $\tau \to \infty$ Eq. **8** gives $\Phi = 0$ and $\Phi \to (\delta - \epsilon)/2$, respectively. The latter reflects equilibration between the demon and each bit: $p_{0,1}' = p_{0,1}^{eq}$, hence $\delta' \equiv p_0' - p_1' = \epsilon$ (see Eqs. **3** and **6b**). On the right side of Eq. **8a**, the quantity in square brackets is non-negative, as determined by numerical inspection, and the prefactor indicates a competition between the parameters δ and ϵ. When $\delta > \epsilon$, the incoming bits contain a surplus of 0's, relative to the equilibrium proportions [Eq. **6a**]; during each interaction interval the composite system relaxes toward equilibrium, generating CW rotation as 0's are converted to 1's, on average. Similarly when $\delta < \epsilon$ the relative surplus of incoming 1's generates CCW rotation. When $\delta = \epsilon$ there is no directed rotation, as the bits arrive distributed in the equilibrium ratio.

The new term appearing in Eq. **8a**, $\epsilon\delta J/6$, does not affect the sign of the quantity in square brackets. However, there is a succinct way to describe its action: it goes against the loser if there is a competition between δ and ϵ, and against both if there is cooperation. This follows from the inequality $J(\tau, \epsilon\delta) \geq 0$ (determined by numerical inspection). E.g. if $\delta > \epsilon > 0$, so that in competition δ wins giving rise to $\Phi > 0$, then the term $\epsilon\delta J/6$ makes a positive contribution to Φ. If $\delta > 0 > \epsilon$, so that both parameters favor CW rotation, the contribution due to this term is negative.

We now explore the thermodynamic and information-processing behavior of our device, and we ask when it might perform a "useful" service. Consider the square region representing allowable values of the excess parameter δ and the weight parameter ϵ, depicted in Fig. 3 for $\tau = 1$ and 10. The line $\epsilon = \delta$ is the contour of zero steady-state rotation: to the left of this line the rotation is CCW ($\Phi < 0$) and to the right it is CW ($\Phi > 0$). The product $W \equiv k_B T f \Phi$ represents the average work that the device delivers to the mass, per interaction interval. Since sign(ϵ) = sign(f), the two lightly shaded triangles in Fig. 3 ($\epsilon\Phi > 0$) are the regions in which the device acts as an *engine*, converting heat from the thermal bath into work to lift the mass. For example, when $\delta > \epsilon > 0$ gravity exerts a CCW torque, but the excess of incoming 0's generates a greater CW torque.

Now consider the quantities

$$S_b = -\sum_{i=0,1} p_i \ln p_i \quad \text{and} \quad S_b' = -\sum_{i=0,1} p_i' \ln p_i'. \qquad \textbf{[9]}$$

For convenience we will call these the *disorder* (per bit), although this terminology ignores correlations between successive bits in the outgoing stream. S_b quantifies the information content of the incoming stream, and is related to its capacity to record new information, in the following sense. When $S_b = 0$ the incoming stream is a blank slate composed entirely of 0's (or entirely of 1's), and the outgoing stream contains a faithful record of CW (or CCW) rotations, as discussed earlier. When $S_b = \ln 2$ (its maximum possible value) the incoming stream is saturated with an equal mixture of 0's and 1's, and in this case the outgoing stream does not chronicle the demon's rotations. We will interpret the difference $\Delta S \equiv S_b' - S_b$ as a measure of the degree to which new information is written to the bits, as they interact with the demon.

Since the rotation of the demon couples tightly to the flipping of bits [Eq. **2**], the line $\epsilon = \delta$ (where $\Phi = 0$) is a contour along which $\Delta S(\delta, \epsilon; \tau) = 0$; here, there is no net rotation and no net change in the bit statistics: $p_0' = p_0$ and $p_1' = p_1$. The other solid line depicted in Fig. 3, running from the upper left to the lower right, is also a contour along which $\Delta S = 0$, representing the *inversion* of bit statistics: $p_0' = p_1$ and $p_1' = p_0$. The two lines divide
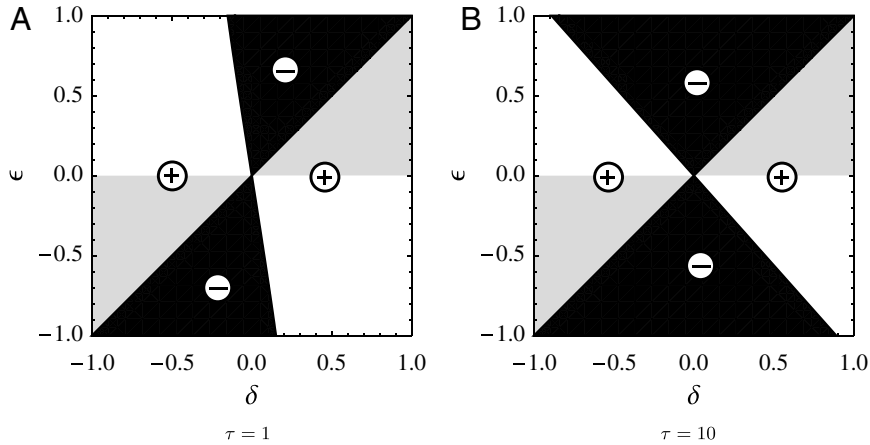
Fig. 3. Behavior of our model as a function of δ and ε, for τ = 1 and 10. The demon can act as an engine (lightly shaded region), an eraser (darkly shaded) or a dud (unshaded). These regions are delineated by the lines ε = 0 and ε = δ, together with a third line (see text), shown passing through the second and fourth quadrants, which depends on τ and is nearly but not exactly straight. The symbols + and − indicate the sign of ΔS, the average change in disorder per bit. The circulation Φ is positive (CW) in the *Lower Right* half of the figure, and negative (CCW) in the *Upper Left*.

the (δ, ε)-square into four regions, with the +'s and −'s in Fig. 3 denoting the sign of ΔS in these regions.

We see in Fig. 3 that ΔS > 0 whenever our device acts as an engine. This is consistent with the proposition that a mechanical demon, in order to convert heat to work, must write information to a memory register. Indeed, Fig. 3 shows that the greater the storage capacity of the incoming bit stream, the larger the mass the demon can hoist against gravity: when presented with a blank slate (δ = ±1) the demon can lift any mass; but when the incoming bit stream is saturated (δ = 0) the demon is incapable of delivering work. Thus a blank or partially blank memory register acts as a thermodynamic resource that gets consumed when the demon acts as an engine.

In the above description, the demon is an active rectifying agent and the bit stream merely a passive receptacle for information. From another perspective, however, the interaction with the demon presents an opportunity for the bits to evolve to a more disordered sequence of 0's and 1's. The bits' role then appears more assertive: their evolution toward greater randomness is what drives the engine, and the demon simply facilitates the process.

In the darkly shaded regions in Fig. 3, the demon acts as an *eraser*, removing information from the memory register: ΔS < 0. For example, if δ = 0, f ≫ 1 (i.e. ε ≈ 1) and τ ≫ 1, then the bits arrive in an equal mixture of 0's and 1's, but each bit has sufficient time to equilibrate with the demon, hence at the end of each interaction interval the composite system is almost certainly in state *A*0, *B*0 or *C*0 [Eq. **6**]. As a result, the outgoing bits are nearly all 0's, and the memory is effectively wiped clean as the mass drops by a distance Δh/2 (on average) per interaction interval.

Our model thus reflects the interplay between two effective forces, one associated with the randomization of the bits and the other with the pull of gravity. When our model acts as an engine, it consumes one resource—a blank or partially blank memory register—to build up another: the gravitational potential of the mass. When it acts as an eraser the roles are reversed. In the unshaded regions in Fig. 3, both resources are squandered (the mass falls and the bits' disorder increases) and our model is a dud, accomplishing nothing useful.

Our model satisfies the inequality

$$W \leq k_B T \Delta S, \qquad [10]$$

for any ε, δ and τ, with the equality holding only when ε = δ. (See *SI Text*). Thus, the increase in the information content of the bit stream places an upper limit on the work that can be delivered,

when the model is an engine. Analogous inequalities arise in the context of feedback control, where an external agent manipulates the system on the basis of outcomes of explicit measurements (22–33). When our model acts an eraser (ΔS < 0), Eq. **10** reveals the minimum amount of work that must be *supplied*, by the falling mass, in order to reduce the information content by a given amount. In the case of full erasure ($S_b' = 0$) this becomes Landauer's principle, $|W| > k_B T S_b$. Note that if we are willing to assign thermodynamic meaning to the randomness in a string of data, Eq. **10** can be interpreted as the second law of thermodynamics (or rather as a weak statement of it, since S_b' ignores correlations between outgoing bits): the decrease in the entropy of the reservoir, $-\Delta S_r = W/k_B T$, must not exceed the increase in the entropy of the bit stream:

$$\Delta S_r + \Delta S \geq 0. \qquad [11]$$

While both sides of Eq. **10** approach zero as ε → δ, their ratio approaches unity in that limit (see *SI Text*). Thus in the immediate vicinity of the line ε = δ, the bound represented by Eq. **10** becomes saturated, and our model behaves with maximal efficiency, acting as a thermodynamically *reversible* engine or eraser. Note however that the *rate* at which the demon either delivers work or erases information approaches zero in this reversible limit.

We conclude by mentioning two extensions of the present work. First, we can reformulate our model as one in which the bits arrive as a sequence of matched pairs, or *dimers*, $b_{2k-1} = b_{2k}$ (e.g. 0011110011…), and the demon interacts with the bit stream, one dimer at a time. Even if the incoming stream contains an equal proportion of 0's and 1's, the demon is able to lift the mass, effectively by "digesting" the pairwise correlations between the bits, which depart in a less ordered sequence (e.g. 0111101011…). This suggests the possibility of a more complex information-processing engine, driven by the recognition of specific patterns in the bit stream.

We have also sketched a mechanistic version of our model, composed of frictionless paddles, pulleys and axles immersed in a dilute gas. While highly idealized, this model is more easily visualized as a material physical system than the discrete-state model described in the present paper. See ref. (8) for an analogous model of a Turing machine.

Methods

To obtain Φ(δ, ε; τ) we solve for the periodic steady state of the demon, then we use that solution to determine the distribution of outgoing bits (p_0', p_1'). The value of Φ then follows from Eq. **3**.

Mandal and Jarzynski

Let $\mathcal{T}_{3\times 3}$ denote the transition matrix whose component $T_{\mu\nu}$ gives the probability to find the demon in state $\mu \in \{A, B, C\}$ at the end of one interaction interval, given that it began in state ν at the start of the interval. Explicitly,

$$\mathcal{T} = \mathcal{P}_D e^{\mathcal{R}\tau} \mathcal{M}. \qquad [12]$$

Here, $\mathcal{M}_{6\times 3} = \begin{pmatrix} p_0 \mathbb{1} \\ p_1 \mathbb{1} \end{pmatrix}$, with $\mathbb{1}$ the 3×3 identity matrix; $\mathcal{R}_{6\times 6}$ is the transition rate matrix whose elements $\{R_{ij}\}$ are discussed in the main text; and $(\mathcal{P}_D)_{3\times 6} = (\mathbb{1}, \mathbb{1})$. Specifically, if $\mathbf{q}_0 \equiv (q_0^A, q_0^B, q_0^C)^T$ gives the probability distribution that describes the demon at the start of one interval, then the six-component vector $\mathcal{M}\mathbf{q}_0$ gives the combined state of the (initially uncorrelated) demon and bit. The factor $e^{\mathcal{R}\tau}$ propagates this distribution for the duration of the interval, and $\mathcal{P}_D$ projects out the state of the bit, so that $\mathbf{q}_\tau = \mathcal{T}\mathbf{q}_0$ is the statistical state of the demon at the end of the interval.

In the periodic steady state, the initial probability distribution of the demon is given by the vector satisfying $\mathcal{T}\mathbf{q}^{\mathrm{pss}} = \mathbf{q}^{\mathrm{pss}}$, whose uniqueness is guaranteed by the Perron-Frobenius Theorem (41). At the end of the interaction interval the state of the (now correlated) demon and bit is $e^{\mathcal{R}\tau}\mathcal{M}\mathbf{q}^{\mathrm{pss}}$.

The statistics of the outgoing bit are obtained by projecting this correlated state to that of the bit:

$$\begin{pmatrix} p_0' \\ p_1' \end{pmatrix} = \mathcal{P}_B e^{\mathcal{R}\tau} \mathcal{M}\mathbf{q}^{\mathrm{pss}}, \qquad \mathcal{P}_B \equiv \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}. \qquad [13]$$

Thus to obtain (p_0', p_1') we must determine $\mathcal{T}$, solve for its eigenstate $\mathbf{q}^{\mathrm{pss}}$, and apply Eq. **13**. This calculation involves a straightforward if tedious exercise in the spectral decomposition of $\mathcal{R}$, which we detail in the *SI Text*.

1. Maxwell JC (1871) *Theory of Heat* (Longmans, London).
2. Smoluchowski M (1912) Experimentell nachweisbare, der üblichen Thermodynamik widersprechende Molekularphänomene [Experimentally verifiable phenomena of molecules contradicting usual thermodynamics]. *Physikalische Zeitschrift* 13:1069–1080 German.
3. Szilard L (1929) Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen [On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings]. *Zeitschrift für Physik* 53:840–856 German.
4. Brillouin M (1951) Maxwell's demon cannot operate: Information and entropy. I. *J Appl Phys* 22:334–337.
5. Feynman RP, Leighton RB, Sands M (1966) *The Feynman Lectures on Physics* (Addison-Wesley, Reading, MA), Chap 46.
6. Landauer R (1961) Irreversibility and heat generation in the computing process. *IBM J Res Develop* 5:183–191.
7. Bennett CH (1982) The thermodynamics of computation—a review. *Int J Theor Phys* 21:905–940.
8. Bennett CH, Landauer R (1985) The fundamental physical limits of computation. *Sci Am* 253:48–56.
9. Zurek WH (1989) Thermodynamic cost of computation, algorithmic complexity and the information metric. *Nature* 341:119–124.
10. Bub J (2001) Maxwell's demon and the thermodynamics of computation. *Stud Hist Phil Mod Phys* 32:569–579.
11. Leff HS, Rex AF, eds. (2003) *Maxwell's Demon 2: Entropy, Classical and Quantum Information, Computing* (Institute of Physics Publishing, Bristol).
12. Maroney O (2009) Information processing and thermodynamic entropy. *The Stanford Encyclopedia of Philosophy*, ed EN Zelta.
13. Penrose O (1970) *Foundations of Statistical Mechanics: A Deductive Treatment* (Pergamon Press, Oxford).
14. Quan HT, Wang YD, Liu YX, Sun CP, Nori F (2006) Maxwell's demon assisted thermodynamic cycle in superconducting quantum circuits. *Phys Rev Lett* 97:180402.
15. Andrieux D, Gaspard P (2008) Nonequilibrium generation of information in copolymerization processes. *Proc Natl Acad Sci USA* 105:9516–9521.
16. Maruyama K, Nori F, Vedral V (2009) Colloquium: The physics of Maxwell's demon and information. *Rev Mod Phys* 81:1–23.
17. Lambson B, Carlton D, Bokor J (2011) Exploring the thermodynamic limits of computation in integrated systems: Magnetic memory, nanomagnetic logic, and the landauer limit. *Phys Rev Lett* 107:010604.
18. Hosoya A, Maruyama K, Shikano Y (2011) Maxwell's demon and data compression. *Phys Rev E* 84:061117.
19. Granger L, Kantz H (2011) Thermodynamic cost of measurements. *Phys Rev E* 84:061110.
20. del Rio L, Aberg J, Renner R, Dahlsten O, Vedral V (2011) The thermodynamic meaning of negative entropy. *Nature* 474:61–63.
21. Kay ER, Leigh DA, Zerbetto F (2007) Synthetic molecular motors and mechanical machines. *Angew Chem (Int Ed)* 46:72–191 and references therein.
22. Kim KH, Qian H (2007) Fluctuation theorems for a molecular refrigerator. *Phys Rev E* 75:022102.
23. Sagawa T, Ueda M (2008) Second law of thermodynamics with discrete quantum feedback control. *Phys Rev Lett* 100:080403.
24. Sagawa T, Ueda M (2009) Minimal energy cost for thermodynamic information processing: Measurement and information erasure. *Phys Rev Lett* 102:250602.
25. Cao FJ, Feito M, Touchette H (2009) Information and Flux in a feedback controlled Brownian ratchet. *Physica A* 388:113–119.
26. Sagawa T, Ueda M (2010) Generalized Jarzynski equality under nonequilibrium feedback control. *Phys Rev Lett* 104:090602.
27. Toyabe S, Sagawa T, Ueda M, Muneyuki E, Sano M (2010) Experimental demonstration of information-to-energy conversion and validation of the generalized Jarzynski equality. *Nat Phys* 6:988–992.
28. Ponmurugan M (2010) Generalized detailed fluctuation theorem under nonequilibrium feedback control. *Phys Rev E* 82:031129.
29. Horowitz JM, Vaikuntanathan S (2010) Nonequilibrium detailed fluctuation theorem for repeated discrete feedback. *Phys Rev E* 82:061120.
30. Horowitz JM, Parrondo JMR (2011) Thermodynamic reversibility in feedback processes. *Eur Phys Lett* 95:10005.
31. Abreu D, Seifert U (2011) Extracting work from a single heat bath through feedback. *Eur Phys Lett* 94:10001.
32. Vaikuntanathan S, Jarzynski C (2011) Modeling Maxwell's demon with a microcanonical Szilard engine. *Phys Rev E* 83:061120.
33. Sagawa T (2012) Thermodynamics of information processing in small systems. *Prog Theor Phys* 127:1–56.
34. Raizen MG (2011) Demons, entropy and the quest for absolute zero. *Sci Am* 304:55–59.
35. Bérut A, et al. (2012) Experimental verification of Landauers principle linking information and thermodynamics. *Nature* 483:187–189.
36. Norton JD (2011) Waiting for Landauer. *Stud Hist Philos Mod Phys* 42:184–198.
37. Earman J, Norton JD (1998) Exorcist XIV: The wrath of Maxwell's demon. Part I from Maxwell to Szilard. *Stud Hist Philos Mod Phys* 29:435–471.
38. Earman J, Norton JD (1999) Exorcist XIV: The wrath of Maxwell's demon. Part II. From Szilard to Landauer and beyond. *Stud Hist Philos Mod Phys* 30:1–40.
39. Hemmo M, Shenker OR (2010) Maxwell's demon. *J Philos* 107:389–411.
40. van Kampen NG (2007) *Stochastic Processes in Physics and Chemistry* (Elsevier, Amsterdam), Chap V.4, 3rd Ed.
41. Meyer CD (2000) *Matrix Analysis and Applied Linear Algebra* (SIAM, Philadelphia, PA), Chap 8.

PHYSICS

Supporting Information

Mandal and Jarzynski 10.1073/pnas.1204263109

SI Text

Here we detail key steps in the derivation of our expression for $\Phi(\delta, \epsilon; \tau)$, and we discuss the inequality $W \leq k_\text{B} T \Delta S$, Eq. **10** of the main text.

Solving for Φ. As explained in the "Methods" section, to solve for Φ we first obtain the stationary probability distribution $\mathbf{q}^{pss}$ of the transition matrix $\mathcal{T}_{3\times3}$. This matrix describes the evolution of the demon over one interaction interval, and is the product of three matrices, $\mathcal{T} = \mathcal{P}_D e^{\mathcal{R}\tau} \mathcal{M}$ (Eq. **12**). Expressions for $\mathcal{P}_D$ and $\mathcal{M}$ were provided explicitly in the "Methods" section. For $\epsilon \in (-1, 1)$, the transition rate matrix for the composite demon-and-bit is

$$\mathcal{R} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -2+\epsilon & 1+\epsilon & 0 & 0 \\ 0 & 0 & 1-\epsilon & -2-\epsilon & 1 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}. \quad \text{[S1]}$$

This matrix has six real, non-degenerate eigenvalues that are (surprisingly) independent of ϵ:

$$\{\lambda_i\} = \{0, -c, -1, -2, -3, -d\}, \quad \text{[S2]}$$

with

$$\begin{aligned} a &= 1 - \sqrt{3}, c = 2 - \sqrt{3}, x = 1 + \epsilon \\ b &= 1 + \sqrt{3}, d = 2 + \sqrt{3}, y = 1 - \epsilon. \end{aligned} \quad \text{[S3]}$$

The quantities a, b, x and y will be used momentarily.

We have found the following spectral decomposition of $\mathcal{R}$ to be convenient:

$$\mathcal{R} = \sum_{i=1}^{6} \frac{|i\rangle \lambda_i \langle i|}{\langle i|i\rangle} = U N^{-1} \Lambda V$$

$$= \begin{pmatrix} \uparrow & & \uparrow \\ \mathbf{u}_1 & \cdots & \mathbf{u}_6 \\ \downarrow & & \downarrow \end{pmatrix} \begin{pmatrix} n_1^{-1} & & \\ & \ddots & \\ & & n_6^{-1} \end{pmatrix} \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_6 \end{pmatrix} \begin{pmatrix} \leftarrow & \mathbf{v}_1 & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{v}_6 & \rightarrow \end{pmatrix}. \quad \text{[S4]}$$

Here, the columns of U are right eigenvectors of $\mathcal{R}$, and the rows of V are its left eigenvectors. We denote the right eigenvectors by $\mathbf{u}_i$ or $|i\rangle$, and the left eigenvectors by $\mathbf{v}_i^T$ or $\langle i|$. These form a biorthogonal pair of basis sets: $\mathbf{v}_i^T \cdot \mathbf{u}_j = \langle i|j\rangle = n_i \delta_{ij}$, i.e. $VU = N$. Explicitly,

$$U = \begin{pmatrix} x & 1 & x & 1 & x & 1 \\ x & -a & 0 & -1 & -2x & -b \\ x & c & -x & -1 & x & d \\ y & -c & -y & 1 & y & -d \\ y & a & 0 & 1 & -2y & b \\ y & -1 & y & -1 & y & -1 \end{pmatrix},$$

$$V = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ y & -ay & cy & -cx & ax & -x \\ 1 & 0 & -1 & -1 & 0 & 1 \\ y & -y & -y & x & x & -x \\ 1 & -2 & 1 & 1 & -2 & 1 \\ y & -by & dy & -dx & bx & -x \end{pmatrix}, \quad \text{[S5]}$$

and $\{n_i\} = \{6, 12c, 4, 6, 12, 12d\}$. Note that since $\mathcal{R}$ is not symmetric, its left and right eigenvectors differ. The matrices N and Λ are diagonal. While it is usual to normalize the left and right eigenvectors so that they are biorthonormal ($n_i = 1$), we have found that the choice of normalization given above leads to less cumbersome expressions in the subsequent analysis.

In terms of this decomposition, we have

$$\mathcal{T} = \mathcal{P}_D e^{\mathcal{R}\tau} \mathcal{M} = (\mathbb{1} \quad \mathbb{1}) U N^{-1} e^{\Lambda\tau} V \begin{pmatrix} p_0 \mathbb{1} \\ p_1 \mathbb{1} \end{pmatrix} \quad \text{[S6]}$$

where $\mathbb{1}$ is the 3×3 identity matrix (see "*Methods*"). An explicit evaluation yields

$$\mathcal{T} = \frac{1}{12} \begin{pmatrix} F+G+\delta H & M-2\delta L & F-G+\delta H \\ M & M+12\sigma^3 & M \\ F-G-\delta H & M+2\delta L & F+G-\delta H \end{pmatrix}$$

$$- \frac{\epsilon}{12} \begin{pmatrix} H+\delta(G-6\sigma) & -2L & H-\delta(G-6\sigma) \\ 0 & 0 & 0 \\ -H-\delta(G-6\sigma) & 2L & -H+\delta(G-6\sigma) \end{pmatrix} \quad \text{[S7]}$$

where $\sigma = e^{-\tau}$ and

$$\begin{aligned} F &= 4 + 2\sigma^3, \, G = 4\sigma^2 + \sigma^c + \sigma^d, \, H = \sqrt{3}(\sigma^c - \sigma^d) \\ L &= 2\sigma^2 - \sigma^c - \sigma^d, \, M = 4 - 4\sigma^3. \end{aligned} \quad \text{[S8]}$$

Solving the equation $\mathcal{T}\mathbf{q}^{\text{pss}} = \mathbf{q}^{\text{pss}}$ (see "*Methods*") we obtain

$$\mathbf{q}^{\text{pss}} = \frac{1}{3} \begin{pmatrix} 1+N \\ 1 \\ 1-N \end{pmatrix}, \qquad N(\delta, \epsilon) = \frac{(\delta-\epsilon)(H-L)}{6 - G + \epsilon\delta(G-6\sigma)}. \quad \text{[S9]}$$

Combining this result with Eq. **13** of the text yields the statistics of the outgoing bits, (p_0', p_1'), from which we then obtain the circulation using the relation $\Phi = p_1' - p_1$.

Relationship between *W* and Δ*S*. We now obtain Eq. **10** of the main text: $W \leq k_\text{B} T \Delta S$. Since

$$W = k_\text{B} T \Phi f = k_\text{B} T \Phi \ln \frac{1+\epsilon}{1-\epsilon}, \quad \text{[S10]}$$

we must establish the non-negativity of the *dissipation* function:

$$\Omega \equiv \Delta S - \Phi \ln \frac{1+\epsilon}{1-\epsilon} \geq 0. \qquad \textbf{[S11]}$$

We will first prove this for the quasistatic case $\tau \to \infty$, and then extend it to finite τ. Let

$$S(X) = -\frac{1-X}{2}\ln\frac{1-X}{2} - \frac{1+X}{2}\ln\frac{1+X}{2} \qquad \textbf{[S12]}$$

denote the entropy of a bit as a function of an excess parameter X.

In the quasistatic limit (specified below by the subscript "∞") the outgoing bits reflect full equilibration between the demon and the bit; see Eq. **6** of the main text. In that limit we have

$$\Phi \to \frac{\delta - \epsilon}{2} \equiv \Phi_\infty, \qquad \delta' \to \epsilon, \qquad \Delta S_b \to S(\epsilon) - S(\delta), \qquad \textbf{[S13]}$$

hence

$$\Omega \to S(\epsilon) - S(\delta) - \frac{\delta - \epsilon}{2}\ln\frac{1+\epsilon}{1-\epsilon} \equiv \Omega_\infty. \qquad \textbf{[S14]}$$

Now note that

$$\Omega_\infty = 0 \quad \text{for } \epsilon = \delta \quad \text{and} \quad \frac{\partial}{\partial \epsilon}\Omega_\infty = \frac{\epsilon - \delta}{1 - \epsilon^2}\begin{cases} >0 & \text{if } \epsilon > \delta \\ <0 & \text{if } \epsilon < \delta \end{cases}. \qquad \textbf{[S15]}$$

Thus for any fixed value of δ, the function $\Omega_\infty(\delta, \epsilon)$ is zero at the point $\epsilon = \delta$, and as a function of ϵ it decreases when $\epsilon < \delta$ and increases when $\epsilon > \delta$. This establishes that $\Omega_\infty \geq 0$.

We have verified by explicit numerical investigation that

$$\Phi(\delta, \epsilon; \tau) = \eta \Phi_\infty \quad \text{where } 0 \leq \eta \leq 1. \qquad \textbf{[S16]}$$

That is, the quantity in square brackets in Eq. **8a** of the main text (which we here label η) falls in the range $[0, 1]$. While we have not been able to establish this analytically, we believe it is related to the fact that all eigenvalues of the transition rate matrix $\mathcal{R}$ are real and non-positive [Eq. **S2**], with the consequence that the composite demon-and-bit system relaxes monotonically toward equilibrium during each interaction interval. This suggests that $\text{sign}(\Phi) = \text{sign}(\Phi_\infty)$, and that the maximum circulation is obtained by allowing the composite system to relax fully to equilibrium.

For finite τ, the excess parameter δ' for the outgoing stream is a linear average of δ and ϵ:

$$\delta' = \delta - 2\Phi = (1 - \eta)\delta + \eta\epsilon, \qquad \textbf{[S17]}$$

using Eq. **3** of the main text, Eq. **S16** and $\Phi_\infty = (\delta - \epsilon)/2$. Since $S(X)$ is concave ($d^2 S/dX^2 < 0$),

$$S_b' = S(\delta') \geq (1 - \eta)S(\delta) + \eta S(\epsilon) = S(\delta) + \eta[S(\epsilon) - S(\delta)]. \qquad \textbf{[S18]}$$

From Eqs. **S13**, **S14** and the non-negativity of Ω_∞, we have

1. Cover TM, Thomas JA (2006) *Elements of Information Theory* (Wiley-Interscience, Hoboken, NJ).

$$S(\epsilon) - S(\delta) \geq \Phi_\infty \ln\frac{1+\epsilon}{1-\epsilon}. \qquad \textbf{[S19]}$$

Combining Eqs. **S18** and **S19** we get

$$S_b' \geq S(\delta) + \eta\Phi_\infty \ln\frac{1+\epsilon}{1-\epsilon} = S_b + \Phi \ln\frac{1+\epsilon}{1-\epsilon}, \qquad \textbf{[S20]}$$

which is the result we set out to establish [Eq. **S11**].

As mentioned in the main text, the inequality $W \leq k_B T\Delta S$ can be viewed as a weak statement of the second law of thermodynamics, where the weakness is due to the neglect of correlations in the outgoing bit stream.* If we accept an equivalence between thermodynamic entropy and the information content of a random data set, then we might expect the second law to be represented more accurately by the inequality

$$W \leq k_B T\Delta\mathcal{H}, \qquad \textbf{[S21]}$$

where $\Delta\mathcal{H} = \mathcal{H}_b' - \mathcal{H}_b$ is the change in the information entropy per bit, *including* correlations among the outgoing bits. For our model Eq. **S21** can be derived directly from general properties of Markov processes, without involving our solution for $\Phi(\delta, \epsilon; \tau)$. We omit this derivation, which makes use of relative entropy (1) as a Lyapunov function (2) characterizing the relaxation of the demon and bits. Since $\mathcal{H}_b = S_b$ (the bits arrive uncorrelated) and $\mathcal{H}_b' \leq S_b'$ (1), we get

$$\Delta\mathcal{H} \leq \Delta S, \qquad \textbf{[S22]}$$

which provides an alternative derivation of the inequality $W \leq k_B T\Delta S$.

Finally, setting $k_B T = 1$ for convenience, we establish the result

$$\lim_{\epsilon \to \delta}\frac{W}{\Delta S} = 1 \qquad \textbf{[S23]}$$

mentioned near the end of the main text. Taking the partial derivatives of the quantities

$$W = \Phi \ln\frac{1+\epsilon}{1-\epsilon} \quad \text{and} \quad \Delta S = S(\delta') - S(\delta) \qquad \textbf{[S24]}$$

with respect to ϵ, at fixed δ and τ, we get

$$\frac{\partial W}{\partial \epsilon} = \frac{\partial\Phi}{\partial\epsilon}\ln\frac{1+\epsilon}{1-\epsilon} + \frac{2\Phi}{1-\epsilon^2} \qquad \frac{\partial\Delta S}{\partial\epsilon} = \frac{\partial\Phi}{\partial\epsilon}\ln\frac{1+\delta'}{1-\delta'} \qquad \textbf{[S25]}$$

(using $\delta' = \delta - 2\Phi$). Along the line $\epsilon = \delta$ we have $W = \Delta S = 0$ as well as

$$\frac{\partial W}{\partial\epsilon} = \frac{\partial\Delta S}{\partial\epsilon}. \qquad \textbf{[S26]}$$

Eq. **S23** then follows by l'Hôpital's rule.

2. Schnakenberg J (1976) Network theory of microscopic and macroscopic behavior of master equation systems. *Rev Mod Phys* 48:571–585.

*Empirically, these correlations are quite small, though non-zero.