Supplementary Information:

The origin of bursts and heavy tails in human dynamics

Albert-László Barabási

Department of Physics, University of Notre Dame, IN 46556, USA

(Dated: February 7, 2005)

Contents

I. Queuing Theory 2
A. Priority Queues with Discrete Priorities 2
B. Priority Queues With a Continuum Priority Distribution 3

II. Calculating $P(\tau)$ for the Priority List Model 4

III. Random Removal Limit of the Priority List Model 6

IV. Power Law Generating Processes 6

V. Mapping to evolutionary models 8

References 11

I. QUEUING THEORY

Since the pioneering work of the 19th century Swedish engineer Agner Krarup Erlang [1], queuing theory has been actively engaged in addressing queuing processes emerging in various industrial, communication and human environments. In this section we review the relevant models in queuing theory and discuss the degree they address the observed power law activity patterns.

A. Priority Queues with Discrete Priorities

The study of priority queues goes back to the seminal 1954 work of Alan Cobham, who introduced the M/G/1 queue with a priority selection rule [2]. In the model tasks arrive to the queue at rate λ, following an exponential arrival time distribution (M). The service time of each task follows a general (G) distribution. The most frequently studied service time is again the exponential, M, thus it is assumed that tasks are executed at rate μ. Each task is assigned a discrete priority parameter $p = 1, 2, ..., r$. Assuming that always the highest priority item is chosen for execution, Cobham derived the average waiting time for an item with priority p. While many variations of the model have been introduced, most work has concentrated on the case when there are two priorities in the system ($r = 2$) and within each priority class items are executed on a first-come-first-serve fashion. This model is motivated by processes taking place in computer and industrial environments, where tasks are typically assigned only into two priority classes, high or low.

In the current paper we are interested in the waiting time distribution $P(\tau_w)$ generated by a priority queue. Recently, Abate and Whitt obtained exact results for $r = 2$ [3], finding that the low priority cumulative waiting time distribution $\tilde{P}(\tau_w)$ has the asymptotic stationary form

$$1 - \tilde{P}(\tau_w) \sim \tau_w^{-\alpha} e^{-\beta\tau_w}, \tag{1}$$

where the exponent α has the value 3/2 (for some special parameter choices can also have the value 0 or 1/2).

B. Priority Queues With a Continuum Priority Distribution

While in industrial and computer environments discrete priorities are natural and common, in human environments priorities can take any value and do not need to be discrete. The question is, could Cobham's model with continuum priorities reproduce the observed power law waiting time distribution? In this case we are interested in a queue with random priorities in which the order of the service is sampled from a $\rho(x)$ distribution. A formal solution of the model for a uniform $\rho(x)$ distribution with $x \in [0, 1]$ is provided in chapter 3.III.11 of Cohen [4].

To show the form of the waiting time distribution I implemented a priority queue with priorities chosen from a uniform distribution $x_i \in [0, 1]$. As Fig. 1SM shows, the simulations indicate that the waiting time distribution is well approximated by the functional form $P(\tau) \sim \tau^{-3/2} e^{-\tau/\tau_0}$, functionally similar to the one derived for the derived for the $r = 2$ priority queue. Note an important difference, however: Eq. (1) was derived for the *low priority* items in the queue, while $P(\tau)$ shown in Fig. 1SM is for *all* items on the list, independent of their priority class.

The origin of the exponential cutoff is in the fluctuating queue length. Indeed, when the execution rate (μ) is larger than the arrival rate (λ), the queue will regularly run out of tasks, thus limiting the potential waiting time. In the unsaturated case $\lambda > \mu$ the queue length increases indefinitely, and the waiting time distribution does not reach a stationary regime. There is only one line in the (λ, μ) parameter space where a power law distribution dominates: $\mu = \lambda$, in which case $P(\tau)$ follows a power law with exponent $\alpha = 3/2$. In this case the queue length follows a random walk bounded at $L = 0$. Given this asymmetry of the fluctuations (i.e. we cannot have negative L values), for $\lambda = \mu$ the system is not stationary (see [5]). A stationary state develops only for $\lambda < \mu$.

In summary, a standard priority queue does not offer a satisfactory explanation for the observed email based communication patterns, for the following reasons:

(a) The exponent $\alpha = 3/2$ is 50% larger than the one observed in the email and other datasets, for which typically $\alpha \simeq 1$.

(b) In the absence of a criteria that would allow us to guarantee $\mu = \lambda$, $P(\tau)$ is dominated by prominent exponential cutoff (see Fig. 1SM).

(c) As we show in Fig. 2c in the manuscript, the $\lambda = \mu$ condition is not valid in email communications: for most users the number of incoming and outgoing messages is very different.

Note, however, that in some tasks heavy tailed distributions with $\alpha > 1$ have been observed, and for some of these a queuing mechanism similar to the one incorporated in the Cobham model may be appropriate. In this case, however, one needs to uncover the mechanism that maintains $\lambda = \mu$, as only when this is satisfied can a power law distribution emerge.

II. CALCULATING $P(\tau)$ FOR THE PRIORITY LIST MODEL

To determine the behavior of $P(\tau)$ analytically we consider a stochastic version of the priority list model (stochastic priority list, or SPL) in which the probability that a task with priority x is chosen for execution in a unit time is

$$\Pi(x) = \frac{x^\gamma}{\sum_{i=1}^{L} x_i^\gamma}, \qquad (2)$$

where γ is a parameter that allows us to interpolate between the random choice limit ($\gamma = 0$, or $p = 0$ in the discrete model) and the case when always the highest priority item is chosen for execution ($\gamma = \infty$ or $p = 1$). Note that this parametrization captures the scaling of the model only in the $p \to 0$ and $p \to 1$ limits, but not for intermediate p values, thus it is chosen only for mathematical convenience. The probability that a task with priority x waits a time interval t before execution is

$$f(x,t) = (1-\Pi(x))^{t-1}\Pi(x). \tag{3}$$

The average waiting time of a task with priority x is obtained by averaging over t weighted with $f(x,t)$, obtaining

$$\tau(x) = \Pi(x)\sum_{t=1}^{\infty} t(1-\Pi(x))^{t-1}. \tag{4}$$

To calculate this sum we use the equality

$$\sum_{t=0}^{\infty} t(1-\Pi(x))^{t-1} = \frac{1}{[\Pi(x)]^2}, \tag{5}$$

which gives

$$\tau(x) = \frac{1}{\Pi(x)} \sim \frac{1}{x^\gamma}. \tag{6}$$

Note that $\tau(x)$ has the unit of time, as $\Pi(x)$ is the probability to choose a task in a *unit time*, thus with unit $[s^{-1}]$.

The calculation leading to (6) represents a mean-field approximation, as we assumed that $\sum_{i=1}^{L} x_i^\gamma$ is constant in time. In reality, even in the stationary case the sum displays fluctuations. Yet, the fact that the predictions of this mean-field approach agree with the results of the numerical simulations indicate that these fluctuations do not affect the SPL model's scaling properties.

To calculate the probability distribution $P(\tau)$ of the time τ spent by a task on the priority list we use the fact that the priorities are chosen from the $\rho(x)$ distribution. That is, a task chosen from $\rho(x)$ distribution enters the queue, and it is delayed a time interval $\tau(x)$ given by (6), which allows us to determine $P(\tau)$ using $\rho(x)dx = P(\tau)d\tau$, providing

$$P(\tau) = \frac{1}{\gamma}\frac{\rho(\tau^{-1/\gamma})}{\tau^{1+1/\gamma}}. \tag{7}$$

Evidence for the validity of (7) is provided by Fig. 2SM, where we show the numerically determined $P(\tau)$ for various values of the γ exponent in Eq. (2), in agreement with the analytical predictions.

In the $\gamma \to \infty$ limit, which corresponds to the strictly priority based choice ($p = 1$) in the model, we find that

$$P(\tau) \sim \tau^{-1}, \tag{8}$$

in agreement with the numerical results (see Fig 3a in the manuscript), as well as the empirical data on the email interarrival times (see Fig 2a in the manuscript). In the $\gamma = 0$ ($p = 0$) limit $\tau(x)$ is independent of x, thus $P(\tau)$ converges to an exponential distribution, as shown in Fig. 3b in the manuscript, and as calculated next.

While the SPL has been introduced for its mathematical convenience, it may also play a role in explaining some human behavior. For example, in some cases individuals may find difficult to adhere to a strict priority based execution, as they may not always have the resources (time and material) to execute the highest priority task. For example, if the top priority item requires the cooperation of some other agent (like a bank) when it is not available (bank closed), some lower priority items will be executed, until the conditions are met to tackle the highest priority item. Therefore, a priority based execution can be often described as a stochastic process, in which a task with priority x_i is executed with a probability $\Pi(x_i)$ that increases with x_i. Such process favors the highest priority items, but allows for the occasional execution of some lower priority items as well.

III. RANDOM REMOVAL LIMIT OF THE PRIORITY LIST MODEL

In the random removal limit of the model ($p = 0, \gamma = 0$) the probability to choose a task for execution is $\Pi = 1/L$, where L is the length of the priority queue. The probability that a task waits a time interval τ before being chosen for execution is

$$P(\tau) = (1 - \Pi)^{\tau-1}\Pi = \frac{1}{L}(1 - \frac{1}{L})^{\tau-1}, \tag{9}$$

i.e. the waiting time distribution follows a simple exponential, in agreement with the numerical simulation shown in Fig. 3b in the manuscript.

IV. POWER LAW GENERATING PROCESSES

In the past few decades several mechanism have been proposed to explain the widespread occurrence of heavy tailed distributions observed in various systems (for recent reviews see [6, 7]). Some of the most studied proposals include:

1. *Preferential attachment based mechanism* (also called the Yule, Simon or Price process, as they were independently introduced by several authors [8–11].) The common element of

this process is that the relevant quantities whose distribution we aim to calculate change in time proportional to their size (which is at the basis of the preferential attachment name [11], often refereed to as the Matthew effect as well). Variations of this process have been invoked to explain a wide range of observed heavy tailed distributions, ranging from the citation distributions of papers [10] to the degree distribution of various complex networks [11], or file size distributions on the Internet [12, 13].

2. *Optimization based processes:* The earliest proposal that optimization could lead to power laws was put forward by Mandelbrot [14], who used it to explain the rank frequency of words. Similar arguments have been suggested to apply to processes ranging from the structure of the Internet graph to file sizes and the degree distribution of complex networks [15].

3. *Critical phenomena:* Power laws naturally emerge in critical phenomena, characterizing the distribution of a number of measurable quantities near a phase transition point. The mechanism giving rise to power laws is rooted in the high fluctuations that a system displays during a phase transition, and have been well understood thanks to intensive work between 1960s and the 1980s, culminating in the scaling theory of phase transitions and the renormalization group approach.

4. *Self-organized criticality (SOC):* Aiming to explain the prevalence of power laws in nature, SOC offered a wide range of models that lead to power law distributions in various quantities, applied from granular media to earthquakes [16, 17]. In these models power laws emerge thanks to the nonlinear interaction between a large number of discrete components.

The queuing model discussed in this paper belongs to a class of stochastic processes that require neither growth (as models invoking preferential attachment do), nor interactions between spatially extended elements, and finally, it is not based on an optimization principle either. The heterogeneous, heavy tailed distribution generated by the model is rooted in a queuing process, that filters a high priority task rapidly through the priority list, while forcing a long waiting time on the low priority tasks.

As we show next, while the model has close links to punctuated equilibrium models [18], the mechanism responsible for the power law distributed waiting times is different from the mechanism responsible for the power distributed avalanches in the punctuated equilibrium models.

V. MAPPING TO EVOLUTIONARY MODELS

Next we show that the priority list model introduced in the paper can be mapped into the random neighbor annealed (RNA) [20, 21] version of the Bak-Sneppen (BS) [18] punctuated equilibrium model.

The RNA model is defined as follows: the state of the system is uniquely characterized by L real numbers x_i, $i = 1, ..., L$. At each time step the smallest x_i, as well as $K-1$ randomly chosen x_j elements are replaced with random numbers between [0,1], chosen from a uniform distribution. In the BS model the $K-1$ sites are not selected randomly, but the x_i numbers are placed on a lattice (most often one dimensional, $d = 1$), and after locating the site i with the smallest x_i, it and its immediate neighbors (x_{i-1} and x_{i+1}) are all replaced by random numbers (i.e. for $d = 1$ we have $K = 3$).

After N time steps in the RNA model N sites will be replaced because they represent the lowest numbers on the list, and $(K-1)N$ sites will be chosen randomly. The ratio between the sites selected by the two protocols is $1/(K-1)$. In the priority list model after N time steps pN sites are chosen with maximum priority and $(1-p)N$ sites are chosen randomly, the ratio between the deterministically and randomly chosen sites being $p/(1-p)$. Thus the priority list and the RNA models can be mapped into each other by assuming that

$$\frac{1}{K-1} = \frac{p}{1-p}, \tag{10}$$

providing

$$K = \frac{1}{p}. \tag{11}$$

Note that this mapping is not *exact*, as in the RNA model at each time step $K-1$ sites are updated randomly, while in the priority list model the number of randomly updated sites changes from iteration to iteration (*i.e.* it is only probabilistically defined, and it fluctuates, its average being determined by p). Yet, the overall dynamics of the two models should nevertheless map into each other.

This mapping allows us to highlight the difference between the nature of the power law distributions discussed in this paper and the one studied extensively in the context of the RNA and BS models. First we note that the quantity of interest in the two models is different. The BS or the RNA models generate power law distributed avalanches, where an avalanche is defined as the number of consecutive mutations (or iterations) during which all

x_i elements chosen for removal are below a preset threshold. This avalanche size distribution follows $P(s) \sim s^{-3/2}$. The concept of the avalanche, however, at this point does not appear to have a tangible interpretation in the priority list model. Instead, we are interested in the waiting time, *i.e.* the time it takes for the individual x_i tasks to be executed (replaced with a different random number).

To demonstrate the equivalence of the *mechanisms* generating the power laws in the two models we would need to show that the avalanche size distribution of the RNA model is related to the waiting time distribution of the priority list model. In the following we show, however, that this is *not* the case, because the avalanches follow a power law for any K (or p) value, while the waiting times are power law distributed only in the $p \to 1$ (or $K \to 0$) limit. Thus, using the language of SOC, the priority list model is tuned, which is not a problem given that we are interested in the impact of human decision making and prioritizing, which corresponds to the $p \to 1$ limit of the priority list model.

In the BS model K represents the number of species whose fitness is changed during an iteration. Given the lattice based nature of the model, only integer K values are typically explored. In the most studied $d = 1$ case we have $K = 3$, which, according to (11) corresponds to $p = 1/3$ in the priority list model. In general for the d dimensional BS model we have $K = 2d + 1$, providing $p = 1/(2d + 1)$.

In summary, the largest p value of interest in the BS and RNA models is $p = 1/3$, all higher dimensional cases corresponding to even smaller p. Power law distributed avalanches emerge for any integer K value in both the RNA and the BS model.

In contrast, the waiting times follow a power law only in the $p \to 1$ limit in the priority list model (which corresponds to the $K \to 0$ limit of the RNA model). As Fig 3SM shows, for $p = 1/3$, which is the largest p explored in the BS context, the priority list model does *not* generate power law distributed waiting times. Thus, despite the mapping, the mechanism responsible for the power laws in the BS context and the one of interest for the priority list are fundamentally different. The power law avalanche sizes are rooted in the critical threshold dynamics of the BS and RNA models, and they emerge for any K value. In contrast, the power law waiting times are driven by a queuing mechanism, effective only in the $p \to 1$ ($K \to 0$) limit of the model. To be sure, one can define avalanches in the priority list model, and they would be power law distributed, but they do not appear to have a meaningful interpretation in the context of human behavior, thus they just represent a technical feature of the model.

It is easy to see that further generalizations of the priority list model, including more realistic human driven processes, would drive the models further away from the punctuated equilibrium models. Thus the class of priority based models that describe human behavior and the class of punctuated equilibrium models will likely meet only in the case of the minimalist priority list model discussed in the paper, and the minimalist RNA version of the punctuated equilibrium models.

Despite the difference between the mechanisms explored in the two modeling frameworks, the mapping between the priority list and the RNA models has potential benefits, as it allows the experience and the tools developed to understand punctuated equilibrium [16–21] to be used to address human decision driven queuing processes, and inversely, to use the rich mathematical framework of queuing theory to have a fresh look at punctuated equilibrium models.

[1] A.K. Erlang, Solution of Some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges, Elektrotkeknikeren **13** (1917).

[2] A. Cobham, Priority Assignment in Waiting Line Problems, J. Oper. Res. Soc. Amer. **2**, 70–76 (1954).

[3] J. Abate and W. Whitt, Assymptotics for M/G/1 low-priority waiting time tail probabilities, Queuing Systems **25**, 173-233 (1997).

[4] J. W. Cohen, *The Single Server Queue*, (North Holland, Amsterdam, 1969).

[5] D. Gross and C.M. Harris, *Fundamentals of Queuing Theory* (Wiley, New York, 1998).

[6] M.E.J. Newman, Power Laws, Pareto Distributions and Zipf's Law, cond-mat/0412004

[7] M. Mitzenmacher, A brief history of generative models for power law and lognormal distributions, Internet Mathematics **1**, 226–251 (2004).

[8] H. A. Simon, On a class of skew distribution functions. Biometrika **42** 425–440 (1955).

[9] G.V. Yule, A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis, Philos. Trans. R. Soc., London B. **213**, 21–87 (1925).

[10] D.J. Price, A general theory of bibliometric and other cumulative advantage processes. J. Amer. Soc. Inform. Sci. **27** 292–306 (1976).

[11] A.-L. Barabási, and R. Albert, Emergence of scaling in random networks, Science **286**, 509–512 (1999).

[12] M. Mitzenmacher, Dynamic Models for File Sizes and Double Pareto Distributions. Internet Mathematics (in press).

[13] A.B. Downey, The structural causes of file size distributions, *Proceedings of MASCOTS2001* (August 15 - 18, 2001 Cincinnati, Ohio).

[14] B. Mandelbrot, An informal theory of the statistical structure of languages, in *Communications Theory* ed. W. Jackson, pp. 486–502 (Butterworth, Woburn, MA, 1953).

[15] S. Valverde, R. Ferrer Cancho and R. V. Solé, Scale-free networks from optimal design, Europhys. Lett. **60**, 512-517 (2002).

[16] H.J. Jensen, *Self-Organised Criticality* (Cambridge Univerity Press, Cambridge, 1998).

[17] P. Bak, C. Tang and K. Wiesenfeld, Self-organized criticality–an explanation of 1/f noise, Phys. Rev. Lett. **59**, 381-384 (1987)

[18] P. Bak and K. Sneppen, Punctuated equilibrium and criticality in a simple model of evolution, Phys. Rev. Lett. **71** 4083–4086 (1993).

[19] M. Paczuski, S. Maslov, and P. Bak, Avalanche dynamics in evolution, growth, and depinning models, Phys. Rev. E **53**, 414443 (1996).

[20] H. Flyvbjerg, K. Sneppen, and P. Bak, Mean field theory of a simple model of evolution, Phys. Rev. Lett. **71**, 4087–4090 (1993).

[21] J. de Boer, B. Derrida, H. Flyvbjerg, A.D. Jackson, and T. Wettig, Simple model of self-organiszed biological evolution, Phys. Rev. Lett. **73** 906–909 (1994).

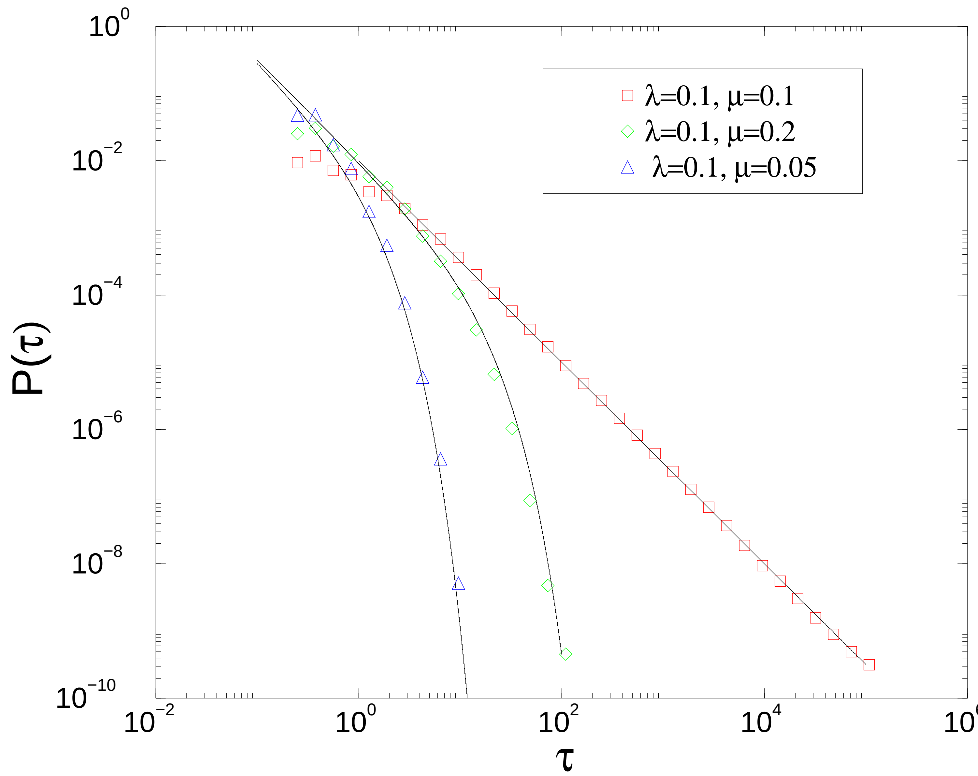

FIG. 1: The waiting time distribution generated by a priority model for which the priorities are chosen from a uniform distribution $\rho(x)$ with $x \in [0, 1]$. Tasks arrive at a constant rate λ and are executed at a rate μ, whose value for each curve being shown in the figure legend. The continuous lines corresponds to $P(\tau) \sim \tau^{-3/2}$ for $\lambda = \mu = 0.1$, and to $P(\tau) \sim \tau^{-3/2} e^{\tau/\tau_0}$ for the other two curves. Note that in general a power law emerges only when $\lambda = \mu$.

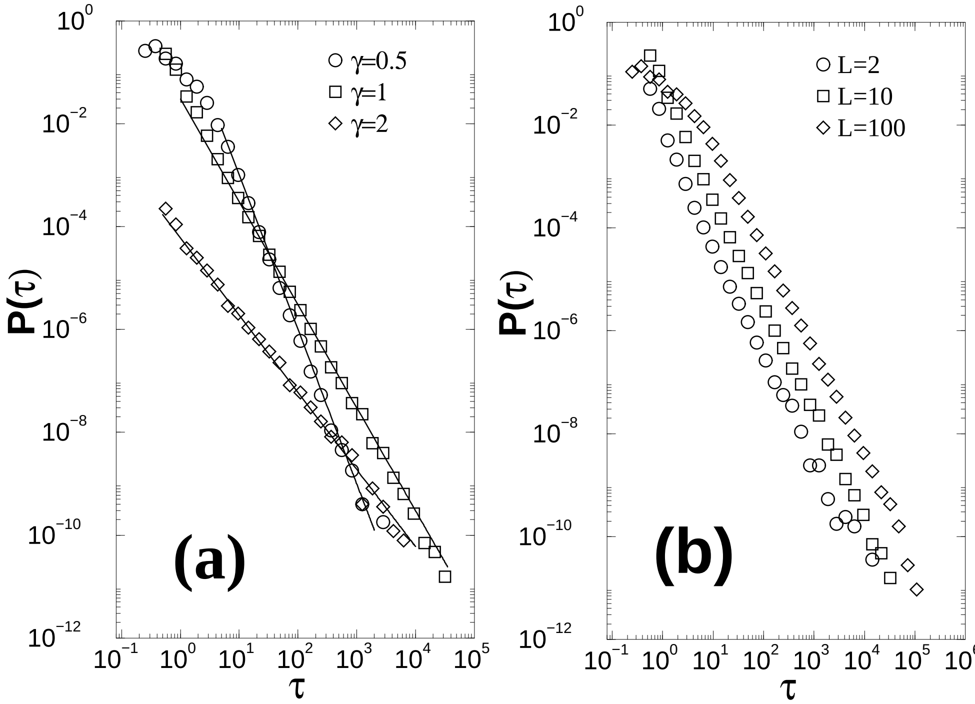

FIG. 2: The waiting or interevent time distribution for the model in which the tasks are chosen stochastically, using (2). The priorities were assigned from a uniform distribution $x_i \in [0, 1]$, and we monitored a priority list of length L over $T = 10^6$ time steps. **(a)** The probability $P(\tau)$ that a task spends τ time on the list, shown for three different γ values in (2), obtained for a priority list of length $L = 10$. The continuous lines correspond to the scaling predicted by (7) for the corresponding γ value, indicating an excellent agreement between the numerical results and the analytical predictions. Note that the $\gamma \to \infty$ and $\gamma \to 0$ cases are discussed separately in Fig. 3 in the manuscript. **(b)** The waiting time distribution obtained for $\gamma = 1$ and varying priority list lengths L, demonstrating that the tail of the distribution is independent of L. The data shown in (a) and (b) were log-binned, to reduce the uneven statistical fluctuations common in heavy tailed distributions, a procedure that does not alter the slope of the tail.

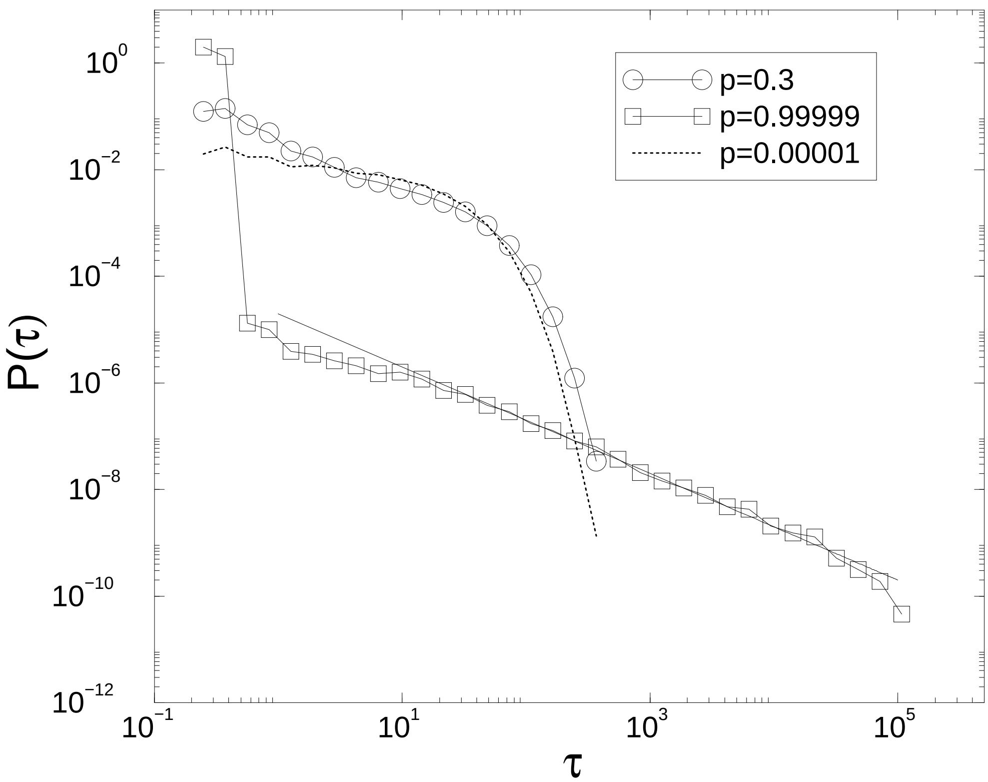

FIG. 3: The waiting or interevent time distribution in the priority list model, demonstrating that for $p = 0.3$, which corresponds to the case $K = 3$ in the RNA model, $P(\tau)$ does not follow a power law. Indeed, for comparison we also show the $p = 0.00001$ curve, which is identical to the one shown in Fig. 3b in the paper, which follows an exponential distribution. As one can see, the difference between the $p = 0.3$ and the $p = 0.00001$ curves is minimal, and they both deviate significantly from the $p = 0.99999$ curve shown as reference. For the $p = 0.99999$ curve here we show the full distribution (in contrast with the tail shown in Fig 3 in the paper), indicating the high peak emerging at $P(\tau = 1)$ due to the fact that for large p values the priority list is dominated by low priority items, thus the newly arriving items will be executed immediately. The closer p is to 1, the more prominent the peak is at $\tau = 1$.