On the stability of two-chunk file-sharing systems

Ilkka Norros, Hannu Reittu, Timo Eirola

On the stability of two-chunk file-sharing systems

Ilkka Norros and Hannu Reittu
VTT Technical Research Centre of Finland

Timo Eirola
Aalto University

Abstract

We consider five different peer-to-peer file sharing systems with two chunks, assuming non-altruistic peers who leave the system immediately after downloading the second chunk. Our aim is to find chunk selection algorithms that have provably stable performance with any input rate. We show that many algorithms that first looked promising lead to unstable or oscillating behavior. However, we end up with a system with desirable properties. Most of our rigorous results concern the corresponding deterministic large system limits, but in two simplest cases we provide proofs for the stochastic systems also.

1 Introduction

Consider the task of distributing a large file to a large number of recipients. Both the owner of the file, the 'seed', and the recipients will be called 'peers' from now on. Assume that each peer is able to open a transfer connection to any other peer, the transfer speeds being limited however. The most effective principle of organizing the task is that as soon as a peer has received (downloaded) the file, it takes the function of a seed itself and starts to upload it to new peers. The number of copies of the file then grows exponentially. The performance is still enhanced, if the file is divided into small pieces, 'chunks'. Then a peer can re-distribute parts of the file already before having the whole file in possession. This technique was introduced by B. Cohen with his BitTorrent [2] protocol. It became soon the dominant principle of sharing large files (e.g., movies) with peer-to-peer networking. While the real BitTorrent is highly sophisticated and evolving software, it seems to be customary to use the term 'BitTorrent-like systems' when referring to abstracted variations of its basic ideas.

This paper is motivated by the fascinating aspect of these systems that they can be arbitrarily scalable thanks to their distributed algorithms and resource sharing. Moreover, stable operation seems to be possible even if the peers are non-altruistic in the sense of leaving the system immediately after having downloaded all chunks (on the other hand, the peers of our models are altruistic in the sense that all download requests are fulfilled). We aim at mathematical understanding of the key issues behind such performance, and simplify the models as much as possible, focusing on the case of just two chunks. Two-chunk models are really toy models, since in real systems the number of chunks is large, e.g. of the order 10^3. However, it turns out that even the two-chunk

case is highly non-trivial as regards stability issues (in fact, the role of the important last chunk is maximally accentuated in the two-chunk case), and we hope that some of our solutions be useful in the analysis of more general models.

Although there is a wide research literature concerning the principles and algorithms of BitTorrent, the point of view mostly differs from ours. Qiu and Srikant [16] provided one of the first mathematical models of BitTorrent-like systems. They focus on the population dynamics of 'leechers' (downloaders) and seeds, assuming that as peers complete the download, they become seeds that stay in the system for a random time. The division of the file into chunks is hidden, assuming uniform availability of chunk replicas in the system. A simple deterministic fluid model is derived, and various aspects of such systems are studied using appropriate parametrizations of the fluid equation. Fan, Chiu and Lui [4] used stochastic differential equations to refine previous BitTorrent models. Chunks were now incorporated in the model, although assuming deliberately their distribution to be uniform. Liao, Papadopoulos and Psounis [10] added to modeling the aspect of peer heterogeneity in terms of their transfer capacities, analysing download delays and fairness issues. Yang and de Veciana [21] considered the service capacity of BitTorrent-like systems in transient regime to deal with so-called flash-crowd scenarios, where a large number of peers enter the system simultaneously. Using the tools of age-dependent branching processes, they show an exponential growth of system capacity during the activation of the crowd. More specifically, they give upper-bound type characteristics, without considering problems like bottlenecks caused by rare chunks however.

In this paper, we consider open systems with a constant rate of incoming peers. There is one permanent seed, but all other peers are non-altruistic. Our scenario is fully distributed and relies on randomness: each peer contacts another, uniformly randomly chosen peer, according to a standard Poisson process, and gets to know what chunks the latter possesses. What follows, depends on the particular algorithm. A possible chunk transfer is assumed to be instantaneous, which is a somewhat violent abstraction, since transfer times dominate search times in reality. We also remark that the model is download-restricted, whereas real peer-to-peer systems tend to be upload-restricted. Massoulie and Vojnovic [11, 12] were the first to propose this strongly abstracted model and to obtain rigorous mathematical results on it. They allowed an arbitrary number of chunks and analysed the corresponding deterministic large system limit (see Section 2) with the following remarkable result: if each new peer arriving in the system obtains a roughly uniformly random chunk, the system is stable even if all the remaining chunks are downloaded randomly (among those available by each contacted peer).

However, the results of [11] prompt for further research in at least two directions. First, the scheme where the seed gives a uniformly distributed chunk to every new peer makes the seed a potential bottleneck — thus, this algorithm is not fully distributed. In our systems, the seed has no special functions, its presence is needed only for guaranteeing that at least one copy of each chunk always remains in the system. The second problem is of theoretical nature: the stability of the large system limit does not automatically guarantee the stability of the original random system (see Section 2). Therefore, we try to establish the stability or non-stability of the stochastic systems as well as of their large system limits. In some cases, the question of stability remains open.

In our first paper [14] on this area, we considered the flash-crowd scenario with greedy copying

(a counterpart to the Plain Random Contact system of subsection 3.1). It was noticed that the first phase of the copying process is asymptotically (with increasing number of peers) equivalent to Pólya's urn model, where the vector of relative frequencies of different chunks converges by a well-known martingale argument to a random point of the simplex of distributions. In the second phase, where peers with complete chunk collections leave the system, this imbalance leads to the 'rare chunk phenomenon', whose avoidance is a central theme of this paper: one of the chunks is not able to become common and as a result forms a bottleneck of performance (see also [18]). In an open variant of this setup, with continuously incoming peers, this could lead to instability: the number of peers in the system could grow unboundedly, since more and more peers would be searching for the one rare chunk. BitTorrent counteracts to the rare chunk phenomenon by its 'Rarest First' principle, and our last two algorithms can be seen as distributed ways to implement this principle using very coarse rarity estimates. The central question of this paper is, how to avoid a severely imbalanced chunk distribution, implying instability, without using any centralized coordination of downloads. One source of ideas for this is provided by the wide literature on urn models (originally often related to physics; for a recent review, see [15]). For example, Ehrenfests' urn model gives an almost ideal balance in a closed system. Still more relevant is the so-called Friedman urn, with an analogous result for an open system, with a flow of incoming particles. We noticed in [14] that if an empty node first contacts a node having chunk 0 (resp. 1) but then downloads the opposite chunk 1 (resp. 0) first (neglecting the question how a peer with that chunk could be found), the distribution of chunks converges almost surely to $\frac{1}{2} - \frac{1}{2}$ as the number of peers goes to infinity.

We analyse five two-chunk models, progressing 'from problem to solution': (i) the Plain Random Contact system, which is found to be unstable (recently, Hajek and Zhu [7] gave a rigorous proof of this in the many-chunk scenario, generalizing Proposition 3.3 of this paper); (ii) the Deterministic First Chunk system, proposed in [14], but found very unsatisfactory in the present scenario; (iii) the ideal Friedman system (non-implementable in a distributed way in our scenario), which is proven to be stable; and two distributed algorithms that try to emulate the Friedman system: (iv) the Delayed Friedman system ([19]), which may be stable but oscillates heavily, and, finally, (v) the Enforced Friedman system which seems to provide the desired performance. Recently, we have proposed also a many-chunk variant [17] of the Enforced Friedman algorithm that seems to provide stable operation, although no proofs have been obtained so far. The paper is structured as follows: general definitions and some preliminaries are given in Section 2, and the five models are studied in Section 3, each in its own subsection.

2 Definitions and preliminaries

We study time-homogeneous continuous time Markov processes $S = (S_t)_{t\geq 0}$ with state space $\mathbb{N}^d$, where d is 2,3 or 4, depending on the particular model. Denoting the ith unit vector by e_i, $i = 1, \ldots, d$, the transitions are always of one of the three forms

$$s \to s + e_i, \quad s \to s - e_i, \quad s \to s + e_i - e_j.$$

The process S is thought as a model of a queueing network, where the i'th state component presents the number of customers in network node i, $i = 1, \ldots, d$. Denote the transition intensity from state $n \in \mathbb{N}^d$ to state $n + v$ by $q(n, n + v)$, and let

$$V = \left\{ v \in \mathbb{Z}^d : \ q(n, n + v) > 0 \text{ for some } n \in \mathbb{N}^d \right\}$$

denote the set of transition directions having positive probability. In our case, V is always finite. The Markov process S, started at time 0 from a fixed state n, is denoted as $S_t(n)$. We are interested in systems with high input rate whose other transition intensities scale linearly with respect to the state. Denote

$$q^{(N)}(n, n + v) = \begin{cases} Nq(n, n + v), & \text{if } v = e_i \text{ for some } i \in \{1, \ldots, d\} \\ q(n, n + v), & \text{otherwise.} \end{cases},$$

and let $S^{(N)}$ be the Markov process with intensities $q^{(N)}(n, n + v)$. Denote by $\lfloor y \rfloor$ the largest integer less than or equal to y and define it componentwise for vectors.

Definition 2.1 *We say that our Markov process S has the* large system limit

$$\dot{s} = F(s), \tag{1}$$

if the limits

$$\lim_{N \to \infty} \frac{1}{N} q^{(N)}(\lfloor Ns \rfloor, \lfloor Ns \rfloor + v) =: q_v(s) \tag{2}$$

exist for all $s \in (0, \infty)^d$, $v \in V$, and the function

$$F(s) = \sum_{v \in V} q_v(s) v \tag{3}$$

is locally Lipschitz continuous.

By a classical result, the local Lipschitz continuity of F implies that the autonomous ordinary differential equation (1) has a unique solution $s(t) = s(t, w)$, $t \in [0, T_w)$, $T_w \in (0, \infty]$, for every starting point $s(0) = w \in (0, \infty)^d$. The domain of F can often be extended to vectors s that have one or several components equal to zero (we prefer the open state space since our models include the asymptotically negligible but functionally crucial seed). Kurtz' little monograph [8] studies various types of population processes in the large system scaling regime and gives limit theorems that can be used for the justification of approximations. The basic limit relation in our case is the following ([8], Theorem 8.1):

Theorem 2.2 *With notation as above, assume that S possesses a large system limit and*

$$\sum_{v \in V} |v| \sup_{w \in K} q_v(w) < \infty$$

for all compact subsets K of $(0, \infty)^d$. Then for all $w \in (0, \infty)^d$ and all $t > 0$

$$\lim_{N \to \infty} \sup_{u \in [0,t]} |\frac{1}{N} S_t^{(N)}(\lfloor Nw \rfloor) - s(u, w)| = 0 \quad \textit{a.s.}.$$

Let us now turn to the question of stability. It is important to understand the differences of the notions of stability in stochastic and deterministic systems. The Markov process S is called *stable*, if it is irreducible and positively recurrent. This is equivalent to the existence of a unique stationary probability measure. Assuming irreducibility and finiteness of transition graph neighborhoods, stability is equivalent to the existence of a finite set of states $C \subset \mathbb{N}^d$ such that with any starting point S_0, the process reaches C in a time with finite expectation.

On the other hand, the dynamical system (1) is called *locally asymptotically stable* around an equilibrium state s^* (that is, a state with $F(s^*) = 0$), if there exists an open set U containing s^* such that $\lim_{t\to\infty} s(t) = s^*$ for any initial state $s(0) \in U$. The system is called *globally asymptotically stable*, if it has a unique equilibrium state s^*, such that $\lim_{t\to\infty} s(t) = s^*$ for any initial state $s(0) \in (0,\infty)^d$.

It is interesting to note that there seem to be no general results concerning the relation between the stability of a Markovian population model and that of its large system limit. Results of the type of Theorem 2.2 tell that, with large populations and high input rates, the stochastic system follows for some time closely the corresponding large system limit path. However, they tell nothing about the stability of the stochastic system $S^{(N)}$ with any finite N. However, these circumstances often coincide, and, as we shall see here also, proving the stability of the dynamical system is usually much easier than proving the stability of the Markov process. Therefore it is interesting to consider the large system limits together with the original random systems.

Another type of scaling to be mentioned is the *fluid scaling*

$$\frac{1}{N} S_{Nt}(\lfloor Nw \rfloor), \quad N \to \infty,$$

where w is a point of the simplex $\left\{ w \in [0,1]^d : \sum_1^d w_i = 1 \right\}$. This scaling plays a central role in J. Dai's seminal paper [3] on the stability of stochastic processing networks and subsequent work by many authors. Dai's theory is able to provide necessary and sufficient conditions for stability: the typical form of results is that a stochastic processing system is stable if and only if all its fluid scaling limits reach the origin a.s. in a bounded time and stay there. Although we did not find ways to apply this approach to our systems, the possibility is worth of more consideration.

All the systems studied in this paper possess differentiable large system limits, and the existence of a unique solution from any starting point is thus always granted. They are, however, non-linear, and proving their stability seems to be very hard in some cases. There are no black-box tools applicable in general. The following elementary lemma is sometimes useful when considering the asymptotic behaviour of a dynamical system. For completeness, a proof is given in the Appendix.

Lemma 2.3 *Let a and b be locally Lipschitz continuous functions $[0,\infty) \to (0,\infty)$. The unique solution u of the differential equation*

$$\dot{u}_t = b_t - a_t u_t, \quad t \geq 0,$$

with initial condition $u_0 \geq 0$ is positive for every $t > 0$ and satisfies

$$\frac{\liminf_{t\to\infty} b_t}{\limsup_{t\to\infty} a_t} \leq \liminf_{t\to\infty} u_t \leq \limsup_{t\to\infty} u_t \leq \frac{\limsup_{t\to\infty} b_t}{\liminf_{t\to\infty} a_t}$$

whenever the fractions are well-defined.

3 Models and results

3.1 Plain Random Contact system

Our first and simplest model is defined in Figure 1. The number of non-seed peers with chunk 0 (resp. 1) is denoted by X (resp. Y). Peers arrive according to a Poisson process with parameter λ, make a random contact, download whatever chunk the contacted peer has (if the seed was contacted, the downloaded chunk is chosen randomly), then make repeated random contacts at Poisson rate 1 until the remaining chunk is found, and leave the system. The system relies entirely on randomness, with fatal consequences.

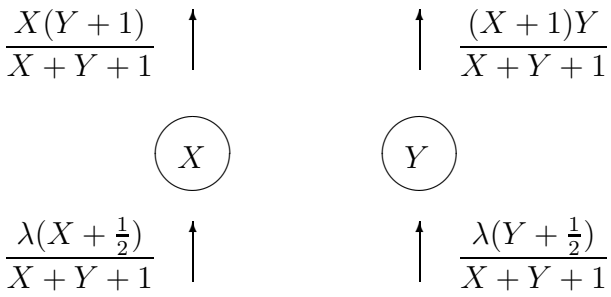

Figure 1: Plain Random Contact system

In the general notation of Section 2, we have $V = \{e_1, -e_1, e_2, -e_2\} = \{e_x, -e_x, e_y, -e_y\}$, the intensities $q(n, n+v)$ are those depicted in Figure 1, and the limits (2) are $q_{e_x} = \lambda x/(x+y)$, $q_{e_y} = \lambda y/(x+y)$, $q_{-e_x} = q_{-e_y} = xy/(x+y)$. Thus, the ordinary differential equation (3) appearing as the large system limit of the Plain Random Contact system is

$$\dot{x} = \frac{(\lambda - y)x}{x+y}, \quad \dot{y} = \frac{(\lambda - x)y}{x+y}. \tag{4}$$

In this paper, we are not interested in the possible stability of the system when the input rate λ is sufficiently low. In all our models, λ appears in the large scale limits as a pure scaling parameter that can be as well chosen to be one. The stability of the stochastic system may, however, depend on λ. Susitaival and Aalto [20] study by simulations several two-chunk systems also from the point of view of stability regions in terms of λ.

The dynamical system (4) is easily seen to be unstable even close to its equilibirium $x = y = \lambda$:

Proposition 3.1 *In $(0, \infty)^2$, the system (4) has a single equilibrium $(x^*, y^*) = (\lambda, \lambda)$. The system is not stable in any environment of (x^*, y^*). Starting with $x_0 > \lambda > y_0$ we have $x_t \to \infty$ and $y_t \to 0$, and vice versa.*

Proof This is immediate from the equations (4). □

As long as both x and y are large and roughly of the same size, the system empties rapidly. However, it is rather straightforward to prove that the stochastic system is unstable when λ is larger than one. We use a coupling argument and state first the behavior of an auxiliary process in a separate lemma:

Lemma 3.2 *Let U and V be mutually dependent, time-inhomogeneous birth and death processes with stochastic up-jump ('α') and down-jump ('β') intensities defined as follows, respectively:*

$$\alpha_t^U = \frac{a}{t \vee 1}, \quad \beta_t^U = bU_t, \qquad \alpha_t^V = c, \quad \beta_t^V = (U_t + d)1_{\{V_t \geq 1\}},$$

where the parameters a, b, c, d are positive numbers and $c > d \geq 1$. Then

$$\mathbb{P}(U_t \in \{0,1\} \text{ eventually}) = \mathbb{P}\left(\lim_{t\to\infty} V_t = \infty\right) = 1.$$

Moreover, with initial state $(U_0, V_0) = (0, M)$, where $M \in \mathbb{N} \setminus \{0\}$, the event

$$A_M = \left\{\forall t \geq 0 : \ U_t \leq 1 \text{ and } V_t \geq M \vee \left(\frac{c-d}{2}t\right)\right\}$$

has positive probability.

Proof The process $(U_t)_{t\geq 0}$ can be interpreted as an infinite-server queue with initial state $U_0 = u_0$, arrival rate α_t^U and independent $Exp(b)$-distributed service times $J_1, J_2, \ldots$. Denote by ν_t the counting process of its arrivals after time 0. The process ν_t is an inhomogeneous Poisson process, and $\nu_t \to \infty$ and $\nu_t/t \to 0$ hold almost surely. It follows that

$$\limsup_{t\to\infty} \frac{1}{t}\int_0^t U_s \,\mathrm{d}s \leq \limsup_{t\to\infty} \frac{1}{t}\left(\sum_{n=1}^{u_0} J_n + \sum_{n=u_0+1}^{u_0+\nu_t} J_n\right) = \frac{1}{b}\limsup_{t\to\infty}\frac{\nu_t}{t} = 0 \quad \text{a.s.}, \tag{5}$$

where the first equality follows from the ergodic theorem. On the other hand, we have

$$\mathbb{P}(U_t \geq 1) \leq \mathbb{E}\{U_t\} = u_0 e^{-bt} + \int_0^t \alpha_s^U e^{-b(t-s)}\,\mathrm{d}s \sim \frac{a}{t}$$

for large t, where the asymptotics is easily obtained after a division of the integration interval into sub-intervals $(0,1]$, $(1, t - ((2/b)\log t]$ and $(t - (2/b)\log t, t]$. Denote then by ν_t^* the counting process of upward jumps of U after time 0 to any state strictly larger than 1. To establish the claim concerning U, it suffices to show that ν_∞^* is almost surely finite. Now,

$$\mathbb{E}\{\nu_\infty^*\} = \mathbb{E}\left\{\int_0^\infty \alpha_t^U 1_{\{U_t\geq 1\}}\,\mathrm{d}t\right\} = \int_0^\infty \frac{a}{t\vee 1}\mathbb{P}(U_t \geq 1)\ \mathrm{d}t < \infty,$$

since the rightmost integrand is $O(t^{-2})$. As regards V, note that it is a pure jump process with compensator

$$ct - \int_0^t (U_s + d)1_{\{V_s\geq 1\}}\,\mathrm{d}s.$$

Since $c > d$, (5) implies that $V_t \to \infty$ a.s. as $t \to \infty$.

Denote $\tau_1 = \sup\{t \geq 0 :\ U_t \geq 2\}$ and $\tau_0 = \inf\{t > \tau_1 :\ U_t = 0\}$. We just showed that $\tau_0 < \infty$ a.s.. Since U is an (inhomogeneous) Markov process, there is some $t_0 > 0$ such that

$$\mathbb{P}[U_{t_0+s} \in \{0,1\}\ \forall s \geq 0 \,|\, U_{t_0} = 0] > 0. \tag{6}$$

Since the probability that U makes no jumps on $[0, t_0]$ is positive, (6) holds also with $t_0 = 0$. Finally, conditioned on a realization $U = u$, the process V can be decomposed as

$$V_t = V_0 + N_t^+ - \int_0^t 1_{\{V_{s-} \geq 1\}}\, \mathrm{d}N_s^-,$$

where N^+ is a homogeneous Poisson process with rate c and N^- is an independent, inhomogeneous Poisson process with rate $u_t + d$. Since $\lim_{t\to\infty} V_t/t = c - d$ a.s., we have, almost surely, eventually $V_t > \frac{c-d}{2}t$. Let $V_0 \geq M$. To prove the last claim, it suffices to note that, conditionally on any realization of N^-, there is a positive probability that the points of N^+ are arranged in such a way that the resulting realization of (U, V) lies in A_M. □

Proposition 3.3 *With $\lambda > 1$, the (stochastic) Plain Random Contact system is unstable. More exactly, almost surely either X or Y escapes to infinity whereas the other obtains eventually only the values 0 and 1.*

Proof We couple (X, Y) with a copy of the process pair (U, V) of Lemma 3.2. Choose the parameters of (U, V) as

$$a = \frac{3\lambda}{c-1},\ b = \frac{1}{2},\ c \in (1, \lambda),\ d = 1,$$

and take

$$M = \left\lceil \frac{2c}{\lambda - c} \vee \frac{c-1}{2} \right\rceil.$$

With $X_t \in \{0, 1\}$ and $Y_t \geq ((c-1)/2)t \vee M$, the up- and down-intensities of X and Y satisfy

$$\begin{aligned}
\alpha_t^X &= \frac{\lambda(X_t + \frac{1}{2})}{X_t + Y_t + 1} \leq \frac{\frac{3}{2}\lambda}{Y_t} \leq \frac{a}{t \vee \frac{2M}{c-1}} \leq \frac{a}{t \vee 1} \\
\beta_t^X &= \frac{X_t(Y_t + 1)}{X_t + Y_t + 1} \geq \frac{Y_t + 1}{Y_t + 2} X_t \geq \frac{1}{2} X_t \\
\alpha_t^Y &= \frac{\lambda(Y_t + \frac{1}{2})}{X_t + Y_t + 1} \geq \frac{\lambda Y_t}{Y_t + 2} \geq \frac{\lambda}{1 + 2/M} \geq c \\
\beta_t^Y &= \frac{(X_t + 1)Y_t}{X_t + Y_t + 1} \leq X_t + 1.
\end{aligned}$$

If

$$X_t \leq U_t \leq 1 \text{ and } Y_t \geq V_t \geq \frac{c-1}{2}t \vee M, \tag{7}$$

then $\alpha_t^X \le \alpha_t^U$, $\beta_t^X \ge \beta_t^U 1_{\{X_t>0\}}$, $\alpha_t^Y \ge \alpha_t^V$, and $\beta_t^Y \le \beta_t^V$. It follows by a standard argument that, with any initial state satisfying $X_0 = U_0 = 0$, $Y_0 \ge V_0 = M$, the processes (X, Y) and (U, V) can be coupled in such a way that (7) holds for all $t \ge 0$ on the event A_M of Lemma 3.2. Hence, there exists a positive number p such that

$$\mathbb{P}\Big[\lim_{t\to\infty} X_t = \infty,\ Y_t \in \{0,1\} \text{ eventually} \Big| X_0 = x_0, Y_0 = 0\Big] > p > 0 \tag{8}$$

when $x_0 \ge M$. By symmetry, the corresponding relation holds if X and Y are interchanged.

Next, let $R > (3/2)\lambda$ and note that

$$\mathbb{P}((X_t, Y_t) \in [R \vee M, \infty) \times [R \vee M, \infty) \text{ eventually}) = 0.$$

Indeed, when $X_t \wedge Y_t \ge R$, the total output rate of the system is larger than the total input rate λ:

$$\frac{2(X_t \wedge Y_t)}{1 + \dfrac{X_t \wedge Y_t + 1}{X_t \vee Y_t}} > \frac{2}{3}R > \lambda.$$

Thus, $\{X_t \wedge Y_t \le R \vee M\}$ is a recurrent event. Since the intensities depicted in Figure 1 are bounded on $\{X_t \wedge Y_t \le R \vee M\}$, the probability of moving from a state with $X \wedge Y \le R \vee M$ to a state with $X \wedge Y = 0$ before a change in $X \vee Y$ is bounded from below by a positive constant not depending on the value of $X \vee Y$. It follows that

$$\{X_t \vee Y_t \ge R \vee M,\ X_t \wedge Y_t = 0\}$$

is a recurrent event as well. Now, (8) yields the proposition. □

Our Plain Random Contact system can be considered as a two-chunk version of the system studied by Massoulie and Vojnovic [11, 12]. In their model, the nodes of a feed-forward queueing network correspond to subsets B of the set of chunks C satisfying $0 < |B| < |C|$. When a peer founds another having one or more new chunks, it downloads one of them randomly and moves forward to the corresponding node. The authors group nodes into 'layers' according to the cardinalities of the chunk sets and obtain most results in the layered version of the model, where a peer restricts its contacts to its own layer. Their main stability result is the following (note that the condition of the theorem is impossible in the two-chunk case).

Theorem 3.4 *([12], Theorem 3) Assume that $|C| > 2$ and that the input rates λ_c, $c \in C$, to nodes of the first (one-chunk) layer satisfy the condition*

$$\forall c \in C \quad \lambda_c < \sum_{c' \in C \setminus \{c\}} \lambda_{c'}.$$

Then the large system limit of the layered model is asymptotically globally stable.

It is quite remarkable that when the input to the first layer is sufficiently balanced, a stochastic system with high input rate can obviously stay in a well-balanced regime long times (or maybe forever) with just random chunk selection. The balancing of the input is however critical — as shown by Hajek and Zhu [7], choosing the first chunk from the entire population leads to instability also in a many-chunk system (on the other hand, they also show that the system is stable for $\lambda < 1$). The rest of this paper is devoted to augmenting the Plain Random Contact system with various distributed algorithms aimed at balancing the chunk populations X and Y.

3.2 Deterministic First Chunk system

Our first attempt to overcome the spontaneous imbalance tendency of the Plain Random Contact system was the Deterministic Last Chunk mechanism introduced in [14], where each peer decides in advance (randomly) which chunk it will download as the last one. The idea was to prevent peers from downloading systematically the rarest chunk as the last one before leaving the system. In the two-chunk case, defined by Figure 2, it is more natural to speak about a Deterministic First Chunk system. The number of empty peers determined to download chunk 0 (resp. 1) first is denoted by A (resp. B), while X and Y have their previous meaning. (The denominators $X + Y + 1$ refer to an overlay network structure where the empty-handed peers in the 'waiting rooms' A and B can contact random peers in the union of X and Y, while being invisible for the latter. For example, in our experimental Chord-based file-sharing implementation (see [13]), a new peer remains invisible for Chord searches (a 'parasite') by delaying its first `stabilize` command until it has downloaded one chunk.)

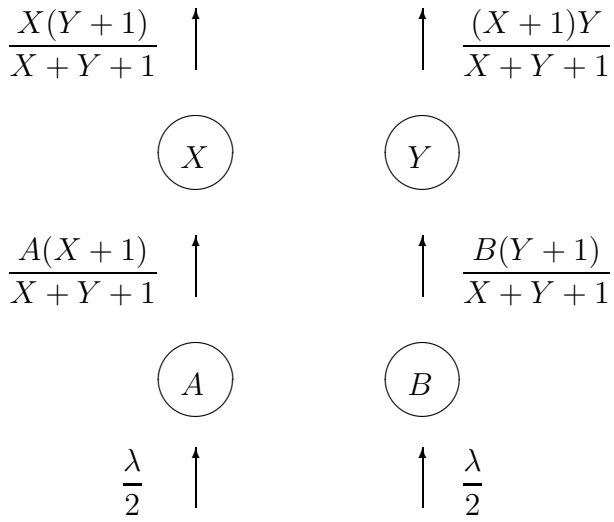

Figure 2: Deterministic First Chunk system

This balancing rule worked promisingly well in our flash-crowd setup with many chunks in [14]. However, its large system limit

$$\dot{a} = \frac{\lambda}{2} - \frac{ax}{x+y} \qquad \dot{b} = \frac{\lambda}{2} - \frac{by}{x+y} \tag{9}$$
$$\dot{x} = \frac{(a-y)x}{x+y} \qquad \dot{y} = \frac{(b-x)y}{x+y}$$

turns out to be unstable:

Proposition 3.5 *The system (9) has no unique equilibrium — its equilibria form the unbounded curve*

$$(a(\theta), b(\theta), x(\theta), y(\theta)) = (\theta, \frac{\lambda\theta}{2\theta-\lambda}, \frac{\lambda\theta}{2\theta-\lambda}, \theta), \quad \theta \in (\frac{\lambda}{2}, \infty).$$

Moreover, it has an open set of trajectories where two components (either x and b or y and a) grow to infinity while the other two remain bounded.

Proof Choose the initial values so that

$$y(0) > a(0) > \lambda, \quad x(0) < \frac{1}{2}\lambda, \quad b(0) > \frac{(x(0)+y(0))\lambda}{2y(0)}.$$

Then the same relations hold for the whole paths, as seen, with help of Lemma 2.3, by writing

$$\begin{aligned}
\dot{a} &= \frac{x}{x+y}\left(\frac{\lambda}{2} + \frac{\lambda}{2x}\cdot y - a\right), \\
\dot{b} &= \frac{y}{x+y}\left(\frac{(x+y)\lambda}{2y} - b\right), \\
\dot{y} - \dot{a} &= \frac{x}{x+y}\left(\frac{y}{x}\left(b - \frac{(x+y)\lambda}{2y}\right) - (y-a)\right).
\end{aligned}$$

Moreover, y and a grow toward infinity, x decreases and b approaches the value $\lambda/2$. □

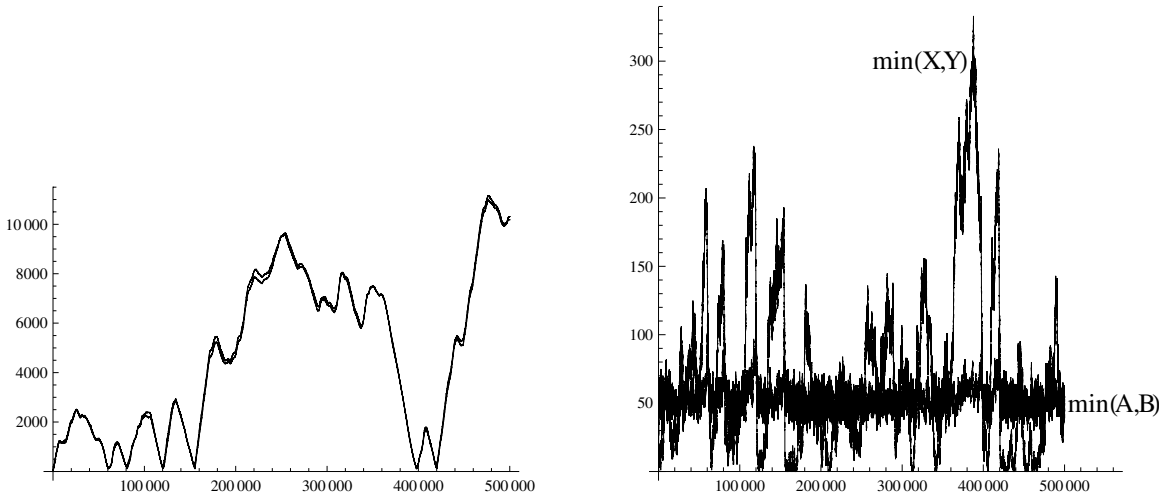

Figure 3: A simulation of the evolution of the Deterministic First Chunk system with $\lambda = 100$. *Left:* The curves $X_t \vee Y_t$ and $A_t \vee B_t$, which almost overlap. *Right:* The curves $X_t \wedge Y_t$ and $A_t \wedge B_t$.

What about the stochastic system? Closer examination of the rates in Figure 2 reveals that the behavior in an unbalanced state differs remarkably from the transient behavior of the Plain Random Contact system (Proposition 3.3). If, say, A and Y are large and about the same size, and B and X are small, then both the input and output rates of X are roughly equal to X itself. Thus, X fluctuates like a time-scaled symmetric random walk reflected at the origin. Meanwhile, Y receives from B rather stable input at rate $\lambda/2$ (B is approximately an $M/M/\infty$ queue), whereas its output

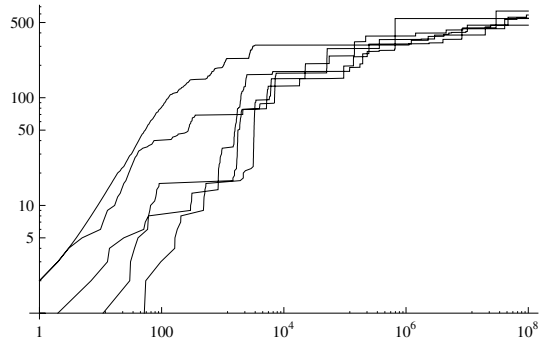

Figure 4: Five long simulations of the Deterministic First Chunk system with $\lambda = 10$ in a log-log plot. Only the record value of total system volume in each run is shown.

rate is roughly equal to X. In such a regime, X exceeds sooner or later the level $\lambda/2$, causing for Y a drift downwards. Although this situation may remain temporary, simulations[1] indicate that Y cannot keep its majority for a very long time, and once X and Y have become equal, it usually happens that X keeps growing while Y shrinks, and their roles change. This oscillatory behavior is rather irregular. Figure 3 confirms the above intuitive analysis. The maxima $X_t \vee Y_t$ and $A_t \vee B_t$ turn out to be very close to each other. On the other hand, $A_t \wedge B_t$ is very stable as expected, whereas $X_t \wedge Y_t$ varies wildly. To get an idea of the kind of recurrence, we made five very long simulations, this time with $\lambda = 10$. Figure 4 shows the growth of the record value of total system population in each run. The peak heights exceed the high but not huge level 500 seldom and little, and we don't observe as strong irregularity as could be expected from a null-recurrent system. Thus, despite Proposition 3.5, we make the following

Conjecture 3.6 *The Deterministic First Chunk system is stable.*

Note that if Conjecture 3.6 is true, we have also found an example of a stable stochastic system whose large system limit has an open set of diverging trajectories.

3.3 Friedman system

Consider an urn containing balls with two colors. The simplest version of Friedman's urn [6] works so that one repeatedly picks a random ball from the urn and returns it together with a ball of the opposite color. The proportions of the two colors approach $\frac{1}{2} - \frac{1}{2}$ [6, 5].

We now modify the Plain Random Contact system by assuming deliberately (this cannot be realised in the kind of distributed systems considered in this paper) that arriving peers make a random contact and then enter the system with a copy of the chunk that the contacted peer did *not*

[1]The simulations of this paper were made with discrete time versions of the models, where times between state changes are constant instead of being exponential with state-dependent parameters. This makes no qualitative difference.

have (in the case that the seed was contacted, the downloaded chunk is chosen randomly); we call this complementary random input 'Friedman input'. This system is defined by Figure 5.

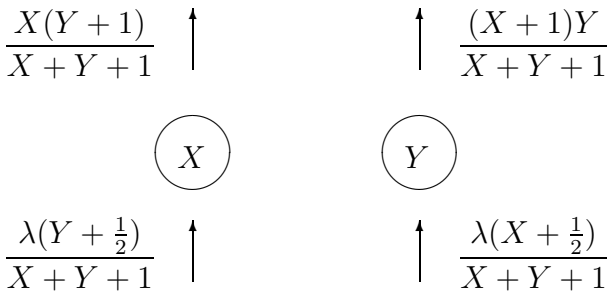

Figure 5: Friedman system

The corresponding large system limit is

$$\dot{x} = \frac{(\lambda - x)y}{x+y} \quad \dot{y} = \frac{(\lambda - y)x}{x+y} \tag{10}$$

Proposition 3.7 *The system (10) has a single equilibrium $(x^*, y^*) = (\lambda, \lambda)$ which is globally asymptotically stable.*

Proof The equations tell immediately that with any initial state in $(0, \infty)^2$, both x and y converge monotonically to (λ, λ). □

Moreover, we note that $(x-\lambda)^2 + (y-\lambda)^2$ is monotonically decreasing:

$$\frac{\mathrm{d}}{\mathrm{d}t}((x-\lambda)^2 + (y-\lambda)^2) = -\frac{2}{x+y}\Big((x-\lambda)^2 y + (y-\lambda)^2 x\Big) < 0.$$

Thus, the Euclidean distance from equilibrium is a Lyapunov function for the dynamical system (10). This observation suggests to study whether and where the squared distance $X_t^2 + Y_t^2$ in the stochastic system is diminishing in some stochastic sense. Indeed, the stochastic Friedman system can be proven to be stable in this way.

Proposition 3.8 *The Friedman system defined in Figure 5 is stable for any input rate λ.*

Proof Denote by N^{x+} and N^{x-} the counting processes of the up- and down-jumps of the process X, respectively, so that $X_t = X_0 + \int_0^t (\,\mathrm{d}N_s^{x+} - \,\mathrm{d}N_s^{x-})$. We can then write

$$\begin{aligned} X_t^2 - X_0^2 &= \int_0^t ((X_{s-}+1)^2 - X_{s-}^2)\,\mathrm{d}N_s^{x+} - \int_0^t ((X_{s-}-1)^2 - X_{s-}^2)\,\mathrm{d}N_s^{x-} \\ &= \int_0^t (2X_{s-}+1)\,\mathrm{d}N_s^{x+} - \int_0^t (2X_{s-}-1)\,\mathrm{d}N_s^{x-}. \end{aligned}$$

(Note that N^{x-} does not jump when $X_{s-} = 0$.) The compensators of N^{x+} and N^{x-} are, respectively,

$$A_t^{x+} = \int_0^t \frac{\lambda(Y_s + \frac{1}{2})}{X_s + Y_s + 1}\,\mathrm{d}t, \quad A_t^{x-} = \int_0^t \frac{X_s(Y_s + 1)}{X_s + Y_s + 1}\,\mathrm{d}t.$$

Using similar notation for the process Y, we see that the process $X_t^2 + Y_t^2$ is compensated to a martingale by subtracting from it the process $A_t = \int_0^t a_s\,\mathrm{d}t$, where

$$a = \left(4\lambda(X + \frac{1}{2})(Y + \frac{1}{2}) - 2X(X - \frac{1}{2})(Y + 1) - 2(X + 1)(Y - \frac{1}{2})Y\right)/(X + Y + 1).$$

Denoting $m = X \wedge Y$, $M = X \vee Y$, we have the estimate

$$\begin{aligned} a &= \frac{1}{M + m + 1}\left(4\lambda(M + \frac{1}{2})(m + \frac{1}{2}) - 2M(M - \frac{1}{2})(m + 1) - 2(M + 1)m(m - \frac{1}{2})\right) \\ &\leq \frac{2M(m + 1)}{M + m + 1}\left(2\lambda(1 + \frac{1}{2M}) - M + \frac{1}{2}\right) \\ &\leq \frac{1}{2}(3\lambda + \frac{1}{2} - M) \\ &\leq -\frac{1}{4} \end{aligned}$$

on the set $\{M > 3\lambda + 1\}$. Denote $\tau = \inf\{t :\ M_t \leq 3\lambda + 1\}$. Assume now that $M_0 = X_0 \vee Y_0 > 3\lambda + 1$. Then, the process $B_t = X_{t\wedge\tau}^2 + Y_{t\wedge\tau}^2$ is a non-negative supermartingale and thus converges to an integrable limit satisfying $\mathbb{E}\{B_\infty\} \leq \mathbb{E}\{B_0\}$, and $B_t - A_{t\wedge\tau}$ is a martingale (for integrability, note that B_t is dominated by $(X_0 + Y_0 + N_t^{x+} + N_t^{y+})^2$, where $N_t^{x+} + N^{y+}$ is a Poisson process). Since $A_{t\wedge\tau}$ is non-increasing, we have

$$B_0 \geq B_0 - \lim_{t\to\infty}\mathbb{E}\{B_t\} = -\lim_{t\to\infty}\mathbb{E}\{A_{t\wedge\tau}\} \geq \mathbb{E}\left\{\int_0^\tau \frac{1}{4}\,\mathrm{d}t\right\} = \frac{1}{4}\mathbb{E}\{\tau\}.$$

Thus, the finite set $\{(x, y) \in \mathbb{N}^2 :\ x \vee y \leq 3\lambda + 1\}$ is reached from any fixed initial state outside of it in a time having finite expectation. □

3.4 Delayed Friedman system

Our first distributed implementation of the idea of the Friedman system (see [19]) was the system defined by Figure 6. An arriving peer first makes a random contact and then *decides* to download first the chunk that the contacted peer does *not* have (in the case that the seed was contacted, the peer decides randomly). As with the Deterministic First Chunk system, the number of empty peers determined to download chunk 0 (resp. 1) first is denoted by A (resp. B), while X and Y have the same meaning as before. We call this the Delayed Friedman system, because the subsystem (X, Y) obtains Friedman input with stochastic delay.

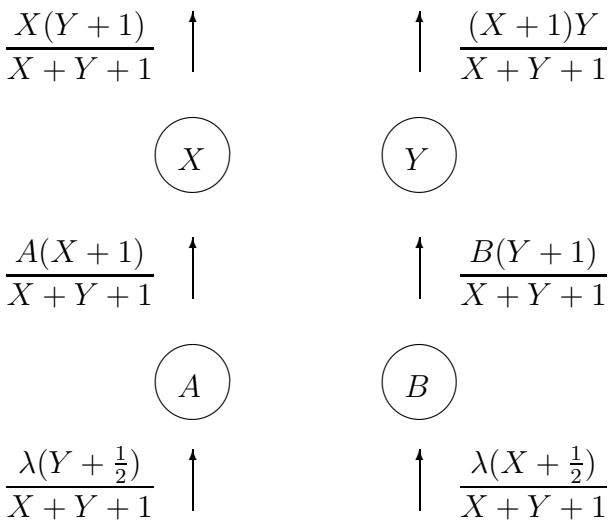

Figure 6: Delayed Friedman system.

The corresponding large system limit is the dynamical system

$$\begin{aligned} \dot{a} &= \frac{\lambda y - ax}{x+y} \qquad \dot{b} = \frac{\lambda x - by}{x+y} \\ \dot{x} &= \frac{(a-y)x}{x+y} \qquad \dot{y} = \frac{(b-x)y}{x+y}. \end{aligned} \tag{11}$$

This system is difficult to analyse, because it oscillates heavily around its unique equilibrium $(a^*, b^*, x^*, y^*) = (\lambda, \lambda, \lambda, \lambda)$. The 'logic' of the oscillating system evolution from an imbalanced state, depicted in the left plot of Figure 7, is the following:

- there is hardly any input to nor output from y
- the 'Friedman rule' directs input to a, which accumulates almost all of it, since x is negligible
- when enough mass has accumulated to a, the balance starts to improve
- once x has become macroscopic, a and y empty rapidly
- b has not had time to grow, so we get a situation close to the mirror image of the original.

Numerical experiments like that shown in Figure 7 (right) hint to global stability, but we have not found an explicit Lyapunov function or other means to prove this. As regards local behavior near equilibrium, the linearised system at equilibrium is essentially a two-dimensional harmonic oscillator — the two remaining dimensions correspond to negative eigenvalues. If we continue the numerical computation of a trajectory of (11), we see slow convergence toward equilibrium (by the way, it is interesting to observe that $x+y$ seems to never descend below 2). Indeed, the system

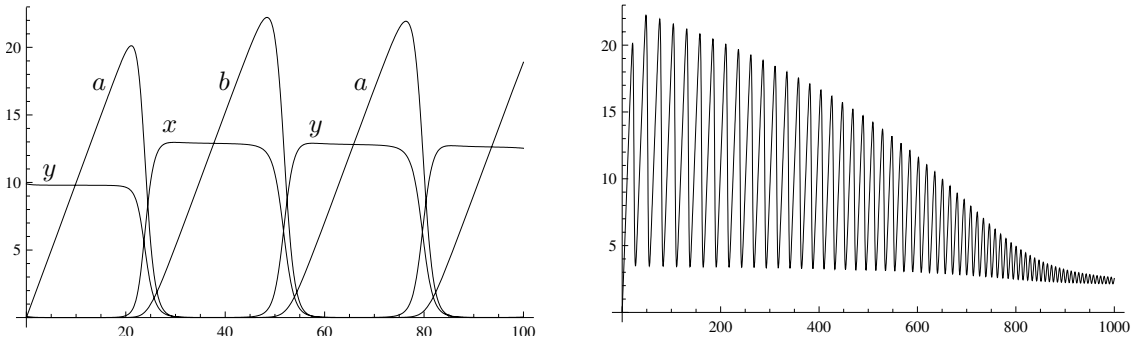

Figure 7: *Left:* A trajectory of the large system limit (11) of the Delayed Friedman system with initial values $a = b = 0$, $x = 0.1$, $y = 9.9$. The curves of x_t and y_t have flat tops while those of a_t and b_t have peaked tops. *Right:* The behaviour of $x + y$, initial state as before.

(11) turns out to be locally asymptotically stable, and this can be shown in a non-elementary but basically straightforward way through a center manifold analysis (see [1, 9]).

Proposition 3.9 *The system defined by (11) is locally asymptotically stable. If the starting point* (a_0, b_0, x_0, y_0) *is close enough to the equilibrium* $(\lambda, \lambda, \lambda, \lambda)$*, the distance to it decreases proportionally to* $1/\sqrt{t}$*.*

Proof Since the common denominator of the right hand sides of (11) does not affect the trajectories, and since λ is a pure scaling factor, it is sufficient to consider the system

$$\boldsymbol{x}' = \begin{bmatrix} x_4 - x_1\, x_3 \\ x_3 - x_2\, x_4 \\ (x_1 - x_4)\, x_3 \\ (x_2 - x_3)\, x_4 \end{bmatrix} \qquad \text{with equilibrium} \qquad \boldsymbol{p} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} .$$

At $\boldsymbol{p}$, the linearized system is given by the matrix

$$\boldsymbol{A} = \begin{bmatrix} -1 & 0 & -1 & 1 \\ 0 & -1 & 1 & -1 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \end{bmatrix} .$$

This has eigenvalues $\lambda_{1,2} = \pm i$ and $\lambda_{3,4} = -1$. The latter has just one linearly independent eigenvector. Let $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ be the real and imaginary parts of an eigenvector corresponding to $\lambda_1 = i$ and let $\boldsymbol{v}_3$ be an eigenvector for $\lambda_3 = -1$ and $\boldsymbol{v}_4$ such that $\boldsymbol{A}\,\boldsymbol{v}_4 + \boldsymbol{v}_4 = \boldsymbol{v}_3$. Putting these into matrix $\boldsymbol{V}$ we obtain the similarity transformation

$$\boldsymbol{V}^{-1}\,\boldsymbol{A}\,\boldsymbol{V} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & -1 \end{bmatrix} \qquad \text{with} \qquad \boldsymbol{V} = \begin{bmatrix} 1 & 1 & 0 & 1 \\ -1 & -1 & 0 & 1 \\ -1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix} .$$

Thus, the system has a two-dimensional stable manifold W^s and a two-dimensional center manifold W^c. The corresponding columns of $\boldsymbol{V}$ are tangents to these at $\boldsymbol{p}$ (see e.g. [1]), and the manifolds are invariant under the flow of the system. Trajectories close to $\boldsymbol{p}$ approach the center manifold like $t\,e^{-t}$. By the reduction principle (see [9]), in order to study the stability of the system, it suffices to know how it behaves on the local center manifold. For this we will reduce the system to a normal form (see [9]).

Take first a linear change of coordinates:

$$\boldsymbol{x} = \boldsymbol{p} + \boldsymbol{V}\,\left[\begin{smallmatrix}\boldsymbol{u}\\ \boldsymbol{v}\end{smallmatrix}\right]\,, \qquad \text{where} \qquad \boldsymbol{u},\boldsymbol{v} \in \mathbb{R}^2\,.$$

Then the local center manifold can be expressed as

$$W^c = \{\boldsymbol{p} + \boldsymbol{V}\,\left[\begin{smallmatrix}\boldsymbol{u}\\ \boldsymbol{h}(\boldsymbol{u})\end{smallmatrix}\right] :\ \|\boldsymbol{u}\| < d\}\ ,$$

where $\boldsymbol{h}(\boldsymbol{u})$ is a two-vector with $\boldsymbol{h}(0,0) = 0$ and $D\boldsymbol{h}(0,0) = 0$ (the latter because of tangency). Write $\left[\begin{smallmatrix}\boldsymbol{f}^c\\ \boldsymbol{f}^s\end{smallmatrix}\right] = \boldsymbol{V}^{-1}\,\boldsymbol{f}$. Then the requirement of invariance of W^c amounts to

$$\boldsymbol{f}^s(\boldsymbol{p} + \boldsymbol{V}\,\left[\begin{smallmatrix}\boldsymbol{u}\\ \boldsymbol{h}(\boldsymbol{u})\end{smallmatrix}\right]) = D\boldsymbol{h}(\boldsymbol{u})\,\boldsymbol{f}^c(\boldsymbol{p} + \boldsymbol{V}\,\left[\begin{smallmatrix}\boldsymbol{u}\\ \boldsymbol{h}(\boldsymbol{u})\end{smallmatrix}\right])\,.$$

This is the equation for $\boldsymbol{h}$. We can recursively solve the coefficients of the Taylor expansion of $\boldsymbol{h}$. Note that the constant and linear terms are zero. This way we get:

$$\boldsymbol{h}(\boldsymbol{u}) = \tfrac{1}{9}\,\begin{bmatrix} 5\,u_1^2 + u_1u_2 - u_2^2 \\ -5\,u_1^2 - 14\,u_1u_2 - 2\,u_2^2 \end{bmatrix} + O(\|\boldsymbol{u}\|^3)$$

and

$$\boldsymbol{g}(\boldsymbol{u}) := \boldsymbol{f}^c(\boldsymbol{p} + \boldsymbol{V}\,\left[\begin{smallmatrix}\boldsymbol{u}\\ \boldsymbol{h}(\boldsymbol{u})\end{smallmatrix}\right]) = \begin{bmatrix} -u_2 + (-10\,u_1^3 + 8\,u_1^2u_2 - 2\,u_1u_2^2 + u_2^3)/9 \\ uu_1 \end{bmatrix} + O(\|\boldsymbol{u}\|^4)\,.$$

The behaviour of the system on the center manifold is given by equation $\boldsymbol{u}' = \boldsymbol{g}(\boldsymbol{u})$. Following [9] we change to complex variables. An eigenvector of the linear part of $\boldsymbol{g}$ corresponding to the eigenvalue i is $\boldsymbol{q} = \left[\begin{smallmatrix} i \\ 1 \end{smallmatrix}\right]$. Setting $\boldsymbol{u} = z\,\boldsymbol{q} + \bar{z}\,\bar{\boldsymbol{q}}$ we get $z = (u_2 - i\,u_1)/2$ and

$$z' = i\,z + \tfrac{8+7i}{18}\,z^3 + \tfrac{32+11i}{18}\,z^2\,\bar{z} + \tfrac{32-11i}{18}\,z\,\bar{z}^2 + \tfrac{-8+7i}{18}\,\bar{z}^3 + O(|z|^4)\,.$$

Taking substitution $z = w + \beta_3\,w^3 + \beta_2\,w^2\,\bar{w} + \beta_1\,w\,\bar{w}^2 + \beta_0\,\bar{w}^3$ we obtain

$$w' = i\,w + (\tfrac{8+7i}{18} - 2\,\beta_3)\,w^3 - \tfrac{32+11i}{18}\,w^2\,\bar{w} + (\tfrac{32-11i}{18} + 2i\,\beta_1)\,w\,\bar{w}^2 + (\tfrac{-8+7i}{18} + 4i\,\beta_0)\,\bar{w}^3 + O(|w|^4)\,.$$

We see that we can kill other third order terms except the $w^2\,\bar{w}$-term. Hence choosing

$$\beta_0 = \tfrac{7-8i}{72}\,,\ \beta_1 = \tfrac{11+32i}{36}\,,\ \beta_2 = 0\,,\ \beta_3 = \tfrac{7-8i}{36}\,,$$

we obtain the equation

$$w' = i\,w - \tfrac{32+11i}{18}\,|w|^2\,w + O(|w|^4)\,.$$

Now, since

$$\tfrac{d}{dt}\,|w|^2 = (i\,w - \tfrac{32+11i}{18}\,|w|^2\,w)\,\bar{w} + w\,(-i\,\bar{w} - \tfrac{32-11i}{18}\,|w|^2\,\bar{w}) + O(|w|^5) = -\tfrac{32}{9}\,|w|^4 + O(|w|^5)\,,$$

we see that $w \to 0$ like $1/\sqrt{t}$. Hence the system is locally asymptotically stable. □

If the third order terms of the equation of w would not have determined the stability, we should have gone to higher order expansions.

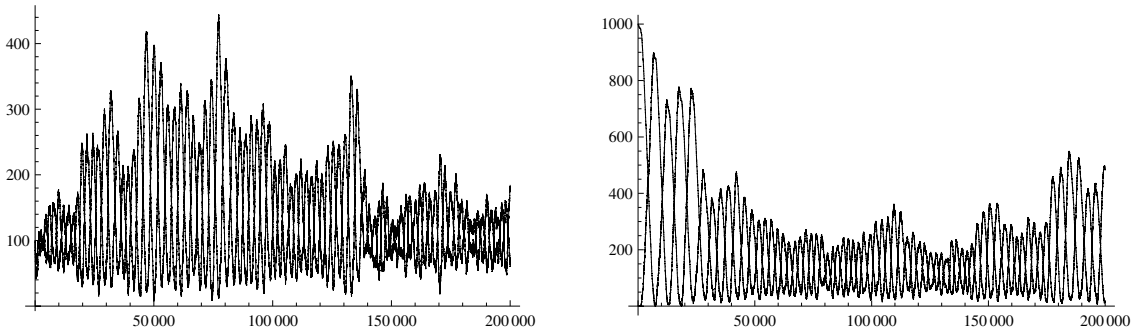

Figure 8: Simulations of the Delayed Friedman system with $\lambda = 100$, trajectories of X_t and Y_t shown. *Left:* Starting from the dynamic system equilibrium $X_0 = Y_0 = A_0 = B_0 = \lambda = 100$. *Right:* Starting with $X_0 = 1000$, $Y_0 = A_0 = B_0 = 0$.

Simulations show that also the original stochastic system oscillates strongly, even when it is started from a balanced state — see Figure 8, left plot. The frequency of the oscillations is very regular compared with that of the Deterministic First Chunk system of Section 3.2. The amplitude is clearly a stochastic process with memory. When started from heavy imbalance, the process returns to more typical amplitudes in a similar fashion as the dynamical system — compare the right plots of Figures 7 and 8. Thus, the simulations suggest stability, however we have not found a way to prove (or disprove) it. From a practical viewpoint, the oscillations indicate that the system is not in good balance, which could be fatal in an application with rapidly changing arrival rates and flash-crowd scenarios.

Conjecture 3.10 *The Delayed Friedman system is stable.*

3.5 Enforced Friedman system

The reason of the oscillatory behaviour of the Delayed Friedman system was that when a peer finally succeeds in downloading its first chunk according to the 'Friedman rule', applied possibly long ago, the choice may already be outdated and thus counterproductive. Our last model avoids this problem by realising choices only immediately or not at all. An empty peer makes three contacts simultaneously (sampling with replacement; the seed shows a random chunk). If two of the chunks of the contacted peers differ from the third one, the latter is downloaded. If all three chunks are similar, nothing is done, but the empty peer stays in a waiting room and repeats the triple contact operation after Exp(1)-distributed waiting time. The number of peers in the waiting room is denoted by Z. Note that if the experiment was successful (that is, a 2-1 situation was obtained), the probability of downloading chunk 0 is $(Y + \frac{1}{2})/(X + Y + 1)$ — exactly the same as in the Friedman system (this circumstance was indeed the design goal of the algorithm). Therefore we call this system, defined by Figure 9, the Enforced Friedman system.

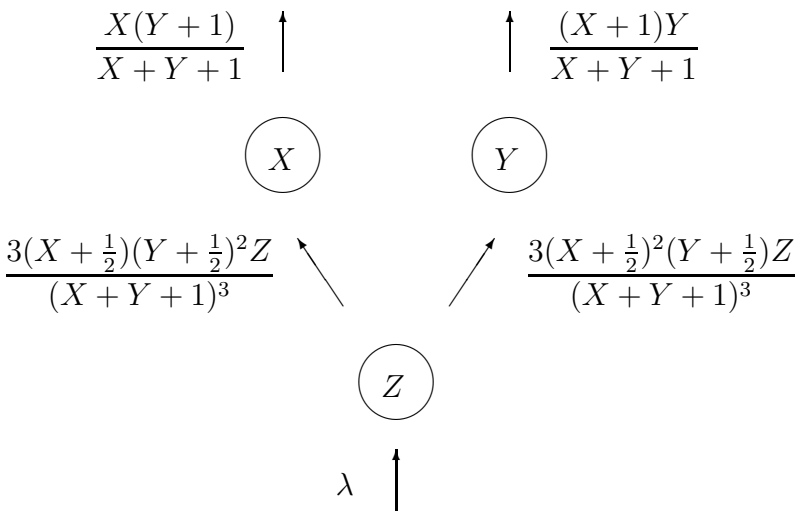

Figure 9: Enforced Friedman system.

The large system limit of the Enforced Friedman system is

$$\dot{z} = \lambda - \frac{3xyz}{(x+y)^2} \tag{12}$$

$$\dot{x} = \frac{3xy^2z}{(x+y)^3} - \frac{xy}{x+y} \tag{13}$$

$$\dot{y} = \frac{3x^2yz}{(x+y)^3} - \frac{xy}{x+y}.$$

This is trickier than the plain Friedman system, but nevertheless tractable:

Proposition 3.11 *The system (12) has in* $(0,\infty)\times(0,\infty)\times(0,\infty)$ *a single equilibrium point* $(z^*,x^*,y^*) = (\frac{4}{3}\lambda,\lambda,\lambda)$. *Locally, the equilibrium is a sink, i.e. all eigenvalues of the limiting linear system are real and negative. This equilibrium is globally asymptotically stable.*

Proof The uniqueness and local character of the equilibrium are found by easy computations. It remains to prove the global stability. Since λ is a pure scaling parameter, we can choose $\lambda = 1$. Let us change to the variables

$$z, \quad \rho = x+y, \quad \beta = \left(\frac{x-y}{x+y}\right)^2,$$

which lose the differentiation between x and y in favour of a single imbalance characteristic β.

The new system satisfies the equations

$$\dot{z} = 1 - \frac{3}{4}(1-\beta)z \tag{14}$$

$$\dot{\rho} = \frac{1}{4}(1-\beta)(3z-2\rho) \tag{15}$$

$$\dot{\beta} = -\beta(1-\beta)\left(\frac{3z}{\rho}-1\right). \tag{16}$$

Fix arbitrary initial values $z_0 > 0$, $\rho_0 > 0$, $\beta_0 \in (0,1)$ ($\beta_0 = 0$ gives a simple linear system).

$1^{\rm o}$. Note first that the system (z_t, ρ_t, β_t) cannot escape from the set $(0,\infty)\times(0,\infty)\times[0,1)$ in finite time. Denote $\tau = \inf\{t \geq 0 : \beta_t = 1\}$. For $t \in [0,\tau)$, $\dot{z}_t > 0$ whenever $z_t < 4/3$, and $\dot{\rho}_t > 0$ whenever $\rho_t < (3/2)z_t$. Since the system freezes at τ, it follows that $\inf_{t\geq 0} z_t \geq \min(z_0, 4/3)$ and $\inf_{t\geq 0}\rho_t > 0$. Equation (16) then yields that β cannot reach 1 at any finite time, so $\tau = \infty$.

$2^{\rm o}$. A basic observation from the equations (13) is that the absolute value of $x-y$ is non-increasing. In terms of the variables z, ρ, β this means that $\beta\rho^2$ is non-increasing, and the corresponding equation from which this is seen reads

$$\frac{\mathrm{d}}{\mathrm{d}t}(\beta_t\rho_t^2) = -\frac{3}{2}\beta_t(1-\beta_t)z_t. \tag{17}$$

By point $1^{\rm o}$, we get

$$\lim_{t\to\infty}\int_t^\infty \beta_s(1-\beta_s)z_s\,\mathrm{d}s = 0. \tag{18}$$

Assume that $\beta_t \to 1$. Since $\beta_t\rho_t^2$ is decreasing by (17), ρ_t cannot be eventually increasing, and it approaches some finite limit value. On the other hand, Lemma 2.3, applied to (14), yields with $\beta_t \to 1$ that $z_t \to \infty$, and (15) makes ρ_t eventually increasing. This contradiction shows that $w := 1 - \liminf_{t\to\infty}\beta_t > 0$.

Since we already saw that eventually $z_t \geq 1$, we obtain by (18) and (16) (neglecting the term $3z/\rho$) that for any $\epsilon > 0$ there exists a number t_0 such that

$$\epsilon > \int_t^\infty \beta_s(1-\beta_s)z_s\,\mathrm{d}s \geq \int_t^T \dot{\beta}_s\,\mathrm{d}s = \beta_T - \beta_t \tag{19}$$

for any t and T such that $T > t > t_0$. Choosing $\epsilon = w/2$ we deduce that $\limsup_{t\to\infty}\beta_t < 1$. But then, (18) and (19) imply $\lim_{t\to\infty}\beta_t = 0$.

$3^{\rm o}$. Now, Lemma 2.3 yields that z_t and ρ_t converge to their stable values. □

In simulations, the Enforced Friedman scheme seems to work perfectly, and to be capable to deal with flash-crowds and rapid input rate variations as well (say, when λ is large for a while and then suddenly drops). Figure 10 shows by a simulation how a heavy initial imbalance is rapidly improved, after which a stable regime persists with fairly small fluctuations. To understand what happens, note that the waiting room population Z must first grow sufficiently large to succeed in finding the rare copies of chunk 2. Once such a state is reached, the system emulates the ideal Friedman system well.

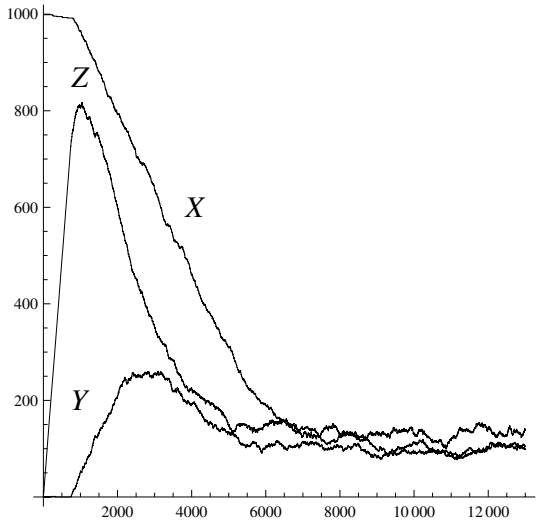

Figure 10: A simulation of the evolution of the Enforced Friedman system with $\lambda = 100$ and initial state $X_0 = 1000, Y_0 = Z_0 = 0$.

As noted earlier, the output from the waiting room is pure Friedman input to the subsystem (X, Y), just with a randomly varying rate. Although it was easy to believe that the stochastic Enforced Friedman system be stable, we did not find a proof for

Conjecture 3.12 *The (stochastic) Enforced Friedman system is stable with any input rate λ.*

However, shortly before the submission of the final version of this paper, Conjecture 3.12 was proven by Barlas Oğuz (University of Berkeley). The details will be published later.

The idea of the Enforced Friedman system can be generalized to the many-chunk case as follows ([17]):

Algorithm 3.13 Enforced Friedman algorithm for many chunks:

(i) a peer makes always three contacts simultaneously (with replacement);

(ii) if there are 'minority chunks' possessed by exactly one of the three peers, download one of them at random;

(iii) if there are no minority chunks, wait for the next triple contact.

It is very interesting that as weak favoring of rare chunks as that implemented in Algorithm 3.13 turns out to be sufficient for stable operation, at least by our simulation evidence. As an example, Figure 11 shows a simulation where all non-seed peers except one are initially missing only one and the same chunk. It is crucial that new peers keep arriving, and most of them download the rare chunk in an early phase.

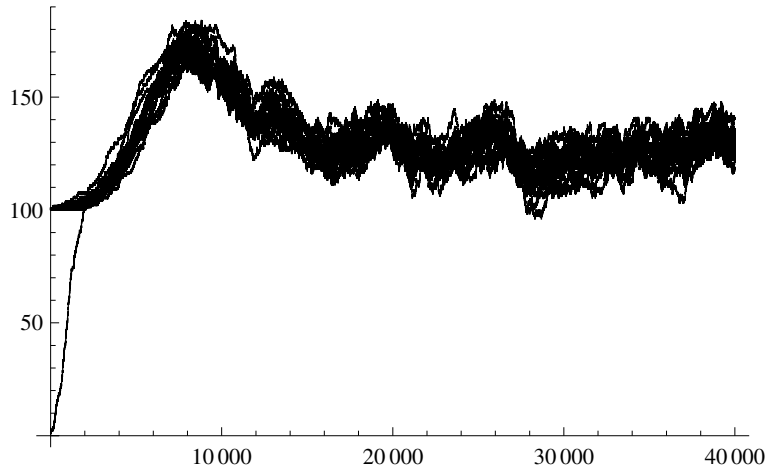

Figure 11: A simulation of Algorithm 3.13 with 20 chunks, $\lambda = 10$, starting from a 'rare chunk' configuration with 100 peers that are missing only chunk 1, and one peer that has only that rare chunk. The evolution of the number of copies of each chunk is shown by a curve.

As a conclusion of our study, we have seen that it is possible to create effectively stabilizing 'Friedman input' using a simple fully distributed algorithm based on random contacts. We also observed that one should use a 'waiting room' concept with memoryless re-trials rather than launch an active hunt for a rare chunk — the rarity may dissolve during the hunt, leading to oscillations. The extra delay caused by the waiting room stage is bearable and worth its price. As regards mathematical theory, we have no proofs yet even for the large system limit in the many-chunk case, and also the less successful algorithms left many interesting problems open. Probably one needs some new insights and methods, because with m chunks the Enforced Friedman system has $d = 2^m - 1$ nodes which is a large number even for a modest m.

Acknowledgement. We thank Balakrishna Prabhu, Rudesindo Nunez Queija, Philippe Robert, Florian Simatos (other members of the Euro-FGI SCALP project 2007-08), as well as Lasse Leskelä and Sergey Foss for fruitful discussions and insights. We also thank two anonymous referees for valuable feedback.

References

[1] S-N. Chow and J. Hale. *Methods of Bifurcation Theory*. Springer-Verlag, New York, 1982.

[2] B. Cohen. BitTorrent specification, 2006. `http://www.bittorrent.org`.

[3] J.G. Dai. On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Ann. Appl. Probability*, 5:49–77, 1995.

[4] B. Fan, D.-M. Chiu, and J.C.S. Lui. Stochastic differential equation approach to model BitTorrent-like P2P systems. In *ICC'06*, Istanbul, Turkey, 2006. DOI: 10.1109/ICC.2006.254824.

[5] D. Freedman. Bernard Friedman's urn. *Ann. Math. Statist.*, 36:956–970, 1965.

[6] B. Friedman. A simple urn model. *Comm. Pure Appl. Math.*, 2:59–70, 1949.

[7] B. Hajek and J. Zhu. The missing piece syndrome in peer-to-peer communication. 2010. Submitted for publication. Preprint: http://arxiv.org/abs/1002.3493.

[8] T. Kurtz. Approximation of population processes. In *CBMS-NSF Regional Conference Series in Applied Mathematics*, volume 36, 1981.

[9] Y.A. Kuznetsov. *Elements of Applied Bifurcation Theory*. Springer-Verlag, New York, 1995.

[10] W.-C. Liao, F. Papadopoulos, and K. Psounis. Performance analysis of BitTorrent-like systems with heterogeneous users. *Performance Evaluation*, 64:876–891, 2007.

[11] L. Massoulie and M. Vojnovic. Coupon replication systems. *ACM SIGMETRICS Performance Evaluation Review*, 33(1):2–13, 2005.

[12] L. Massoulie and M. Vojnovic. Coupon replication systems. *IEEE/ACM Trans. Networking*, 16(3):603–616, 2005.

[13] I. Norros, V. Pehkonen, H. Reittu, A. Binzenhöfer, and K. Tutschku. Relying on randomness-PlanetLab experiments with distributed file-sharing protocols. In *Proceedings of the 3 rd EURO-NGI conference on Next Generation Internet Networks (NGI 2007)*, Trondheim, Norway, 2007. Downloadable from IEEE Xplore. DOI: 10.1109/NGI.2007.371191.

[14] I. Norros, B. Prabhu, and H. Reittu. Flash crowd in a file sharing system based on random encounters. In *Inter-Perf*, Pisa, Italy, 2006. Downloadable from ACM Digital Library.

[15] R. Pemantle. A survey of random processes with reinforcement. *Probability Surveys*, 4:1–79, 2007.

[16] D. Qiu and R. Srikant. Modeling and performance analysis of BitTorrent-like peer-to-peer networks. *ACM SIGCOMM Computer Communication Review*, 34(4):367–378, 2004.

[17] H. Reittu. A stable random-contact algorithm for peer-to-peer file sharing. In *Self-Organizing Systems. Proceedings of 4th IFIP TC 6 International Workshop, IWSOS 2009, Zurich, Switzerland*, volume 5918 of *Lecture Notes in Computer Science*. Springer, 2009. DOI: 10.1007/978-3-642-10865-5_16.

[18] H. Reittu and I. Norros. Toward modeling of a single file broadcasting in a closed network. In *Proceedings of IEEE SPASWIN2007*, Limassol, Cyprus, 2007. DOI: 10.1109/WIOPT.2007.4480079.

[19] H. Reittu and I Norros. Urn models and peer-to-peer file sharing. In *Proc. IEEE PHYSCOMNET'08*, Berlin, 2008. DOI: 10.1109/WIOPT.2008.4586131.

[20] R. Susitaival and S. Aalto. Analyzing the file availability and download time in a P2P file sharing system. In *Proceedings of NGI 2007*, 2007. DOI: 10.1109/NGI.2007.371202.

[21] X. Yang and G. de Veciana. Service capacity of peer to peer networks. In *Proceedings of INFOCOM 2004*, pages 2242 – 2252, 2004.

A Appendix

Proof of Lemma 2.3.

Denote $\tau = \inf\{t > 0 : u_t = 0\}$. Since $\dot{u}_0 = b_0 > 0$ in the case that $u_0 = 0$, we always have $\tau > 0$. Assume for the contrary that $\tau < \infty$. For $t \in [0, \tau)$ we have

$$\dot{u}_t \geq \inf_{s\in[0,\tau)} b_s - (\sup_{s\in[0,\tau)} a_s)u_t.$$

It follows that

$$u_t \geq \frac{\inf_{s\in[0,\tau)} b_s}{\sup_{s\in[0,\tau)} a_s} + (u_0 - \frac{\inf_{s\in[0,\tau)} b_s}{\sup_{s\in[0,\tau)} a_s}) \exp(-(\sup_{s\in[0,\tau)} a_s)t), \quad t \in [0, \tau).$$

Since the right hand side is positive also for $t = \tau$, we deduce that $\tau = \infty$.

Assume first that the functions a and b are bounded away from zero and infinity. Let a_*, a^*, b_*, b^* be any positive numbers such that

$$a_* < \liminf_{t\to\infty} a_t \leq \limsup_{t\to\infty} a_t < a^*, \quad b_* < \liminf_{t\to\infty} b_t \leq \limsup_{t\to\infty} b_t < b^*.$$

Denote

$$\sigma = \inf\left\{t \geq 0 : \inf_{s\in[t,\infty)} a_s \geq a_*, \sup_{s\in[t,\infty)} a_s \leq a^*, \inf_{s\in[t,\infty)} b_s \geq b_*, \sup_{s\in[t,\infty)} b_s \leq b^*\right\}.$$

Since σ is finite and

$$\inf_{s\in[\sigma,\infty)} b_s - (\sup_{s\in[\sigma,\infty)} a_s)u_{\sigma+t} \leq \dot{u}_{\sigma+t} \leq \sup_{s\in[\sigma,\infty)} b_s - (\inf_{s\in[\sigma,\infty)} a_s)u_{\sigma+t}, \quad t \geq 0,$$

we obtain that

$$\frac{b_*}{a^*} + (u_\sigma - \frac{b_*}{a^*})e^{-a^*t} \leq u_{\sigma+t} \leq \frac{b^*}{a_*} + (u_\sigma - \frac{b^*}{a_*})e^{-a_*t}, \quad t \geq 0,$$

which implies

$$\frac{b_*}{a^*} \leq \liminf_{t\to\infty} u_t \leq \limsup_{t\to\infty} u_t \leq \frac{b^*}{a_*}.$$

Since a_*, a^*, b_*, b^* were arbitrary, the assertion follows.

Finally, when a and b are not both bounded away from zero and infinity and the fractions in the assertion are well-defined, the non-trivial cases are obtained as above, simply relaxing some conditions in the definition of σ. □