Cooperative Dynamics of Loyal Customers
in Queueing Networks

Olivier Gallay*, Max-Olivier Hongler

Ecole Polytechnique Fédérale de Lausanne (EPFL)
STI-IPR-LPM
Station 17
CH-1015 Lausanne

Abstract

We consider queueing networks (QN's) with feedback loops roamed by "intelligent" agents, able to select their routing on the basis of their measured waiting times at the QN nodes. This is an idealized model to discuss the dynamics of customers who stay loyal to a service supplier, provided their service time remains below a critical threshold. For these QN's, we show that the traffic flows may exhibit collective patterns typically encountered in multi-agent systems. In simple network topologies, the emergent cooperative behaviors manifest themselves via stable macroscopic temporal oscillations, synchronization of the queue contents and stabilization by noise phenomena. For a wide range of control parameters, the underlying presence of the law of large numbers enables us to use deterministic evolution laws to analytically characterize the cooperative evolution of our multi-agent systems. In particular, we study the case where the servers are sporadically subject to failures altering their ordinary behavior.

Keywords: Queueing Networks with Feedback Loops, Loyal Customers, Cooperation, Stable Temporal Oscillations, Synchronization, Stabilization by Noise.

*Corresponding author. Tel: +41-21-6935817; fax: +41-21-6933891; e-mail address: olivier.gallay@epfl.ch

1 Introduction

The conception and the control of complex networks supporting random flows of items are key issues of several engineering domains ranging from car and cargo traffics, production systems, supply chains, water, electricity and information networks to only quote a few. For many actual situations, these random flows find a natural mathematical modeling framework in the Queueing Networks (QN's) formalism where the flows' randomness is introduced, via ad-hoc probability laws, in the dynamics of servers located at the vertices of the network. Under fairly general conditions and for arbitrary initial conditions, the transient regimes relax to stationary regimes. For a wide class of dynamics, the theory pioneered by J.R. Jackson and subsequently generalized by F.P. Kelly (see an up-to date account in (Chen and Yao, 2001)) offers powerful methods to calculate the time-invariant probability densities and therefore provides quantitative information for the stationary performance measures. While the basic QN's theory considers that the circulating items composing the flows are mainly passive entities (i.e. *tokens*), it is mandatory for numerous applications that each circulating item possesses his/her own identity[1], car traffic being a perfect illustration. When a collection of such intelligent items is considered, we shall speak of a *multi-agent* system. Based on his individual experience and information gathering, each agent is able to take a "private" decision affecting his subsequent dynamical behavior and/or routing in the network. It is widely recognized that the interactions between agents and their environment open possibilities for the emergence of *collective behaviors*, which result in macroscopic *spatio-temporal patterns*. The recent literature offers a wealth of illustrations where such auto-organizing features produce emergent structures (Schweitzer, 2003), (Mikhailov and Calenbuhr, 2002). In this general context, we will show below how QN's in which agents are allowed to modify their routing strategies according to measures of i) their waiting time and/or ii) the queue contents offer an ideal theoretical framework to investigate some aspects of multi-agent systems.

Generically, the emergence of macroscopic patterns in a society of interacting agents originates from the conjugate action of non-linearities in the dynamics coupled with the interactions between the "intelligent" members of the society. The interactions can either be direct (i.e. from agent to agent) or can be implemented via the reactions that each agent adopt when observing the macroscopic state of the society (i.e. the interaction between an agent and its social environment). This paper deals with this second type of interactions, operating in simple QN's. The interactions between the agents here depend on the waiting times they measure before being served at the vertices of the network. In addition, non-linearities into the dynamics are introduced via branching nodes where each agent has to decide either to engage himself in a feedback up-streaming route (and then to revisit certain nodes) or to travel downstream the network. The routing decision depends on the measured waiting time each agent suffers before a branching node. This is an idealization of the dynamics induced by a collection of customers who remain loyal to a server provided their service time stays below a critical threshold. By restricting our analysis to elementary topologies, we are here able to explicitly derive analytical results characterizing the emergence of *stable temporal oscillations* (i.e. temporal patterns). This cooperative evolution is entirely due to the interactions between the "intelligent" agents and their social environment (i.e. *stigmergic* interactions), the "intelligence" being the ability to monitor their waiting time. The agents' interactions generate, via their waiting time, an effective delay mechanism in the evolution of the queue contents. This time-delayed evolution can be represented by an hydrodynamic analogy (called the auto-*siphon* dynamics). Further, we increase the complexity of the network and consider a topology with two parallel feedback queues in which the agents, besides their ability to monitor waiting times, do also possess a vision capability (hence, the agents' "intelligence" is enhanced). In this situation, we are able to analytically characterize non-linear behaviors such as synchronization of oscillations and noise-induced stabilization. Besides the pure theoretical insights, a recent related contribution, (Zohar et al., 2002), testifies that our class of models offers a potential for relevant applications. Note in addition that siphon effects have been considered in the context of Petri nets (see (Chu and Xie, 1997) among others). However, contrary to the effect occurring in the Petri nets framework, which leads to infinite delays (*any empty siphon remains empty*), our dynamics deals with cyclo-stationary siphon effect.

Our paper is organized as follows. In section 2, we introduce and study the flow dynamics of a single queueing node with feedback loop. The resultant flow dynamics can be viewed as formed by incoming potentially *loyal* customers[2], their loyalty being dependent on a patience parameter. Adopting this picture, the customers will visit the feedback loop (i.e. remain loyal) only when their experienced waiting time before the routing decision stays below a critical threshold value. In section 3, we study a network formed by two parallel feedback queues and a bifurcation point. Any

---

[1]From now on, we shall use "his" without sexist intention.

[2]From now on we shall speak of agents and customers interchangeably.

incoming agent has to choose between one of the two servers (*routing decision*), but once the choice made, can neither renege or nor jockey between the queues. The routing decision can be either deterministic, random, and/or guided by a partial or a full observation of the real time content of the queues. The simultaneous ability to observe queue contents and to monitor waiting times offers the possibility to generate new cooperative time evolutions. In section 3.2, we focus on situations where the agents can observe the queue content of a single server and the decision to engage into the observed queue is based on its content (the agent joins the queue if the population is below a critical population threshold). In this case, we show how the presence of random fluctuations in the service times can stabilize a flow dynamics which is otherwise unstable for purely deterministic service times. In particular, we provide an analytical study for the case where the servers are randomly subject to failures altering their ordinary behavior. Finally, in section 3.3, we allow the agents to observe both queue contents. In this situation, when a "shortest-queue-first" scheduling rule is adopted at the bifurcation node, a full synchronization of the queues' oscillations is observed.

2 Feedback queueing system - Siphon dynamics

Consider the single server queueing system sketched in Fig. 1. An incoming flow of customers,
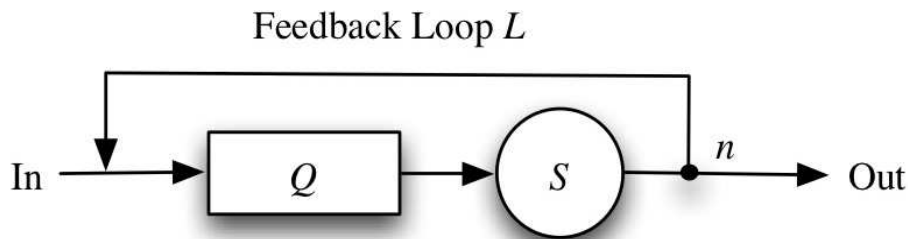

Figure 1: A single stage queueing system with feedback loop.

described by a renewal process with mean inter-arrival time $\frac{1}{\lambda}$ and probability distribution $A(x)$ with density $dA(x)$, is served by a processing unit which service times are i.i.d. random variables with mean $\frac{1}{\mu}$, probability distribution $B(x)$ and density $dB(x)$. Accordingly, the parameters λ and μ are respectively the incoming and service rates of the renewal processes. We assume that the distributions $A(x)$ and $B(x)$ have finite moments. Here, we suppose the traffic intensity $\rho = \frac{\lambda}{\mu} < 1 \Leftrightarrow \lambda < \mu$, which ensures the stability of the queueing system when there is no feedback loop. Assume also that the waiting room capacity is unlimited and that the service discipline is first-in-first-out (FIFO). After being served at the decision node n, each customer has to choose among two possibilities, namely:

a) either to quit the system definitively

b) or to follow the feedback loop and line up again for being served once more.

Several contributions (Takàcs, 1963), (D'avignon and Disney, 1976), (Peköz and Joglekar, 2002) consider the situation arising when the decision between the choices a) and b) is taken randomly. When this is the case, by imposing a stationary flow balance (i.e. incoming equals outgoing flow), we drive the system into a self-consistant *stationary* regime. As we shall now see, such purely stationary flows strongly differ from the queue dynamics that can be observed when "intelligent" agents circulate in the network. Specifically, assume now that each customer is able to record the total waiting time W he spent (i.e. W is the sum of the queueing and the processing times). Assume further that W controls the decision routing between the alternatives a) and b), namely: when W exceeds a critical value P (call it the *patience parameter*), the customer follows the alternative a), and the b) otherwise. When alternative b) is chosen, we shall speak of *loyal customers*, as the agents are pleased with the server and then return to it for another service. In the sequel, we focus on homogenous agents for which P is a common value. In this case, and for large enough P, quasi-deterministic cyclo-stationary regimes emerge, i.e. *stable temporal oscillations* of the queue level $Q(t)$ are observed and this independently of the detailed nature of the probability laws $A(x)$ and $B(x)$. Despite to the presence of the fluctuations, this robust and quasi-deterministic behavior is directly reminiscent from the *law of large numbers*. Indeed, the importance of the relative fluctuations around the average waiting time $\langle W \rangle$ (which is the sum of individual processing times) decreases for large queue content $Q(t)$ (a more formal characterization is given in (Filliger and Hongler, 2005)). Accordingly, for large P, the dynamics can approximatively be discussed via a purely deterministic approach (see (Hongler et al., 2004) and (Filliger and Hongler, 2005)) which follows when the service

requires a fixed time interval $\frac{1}{\mu}$. Hence, for a given queue length N_c and a given corresponding patience parameter $P = \frac{N_c}{\mu}$, an incoming tagged customer (called $\mathcal{C}$ from now on) lining behind N_c other customers, will, when reaching node n, choose the alternative a) (i.e. leave the system), as for this deterministic regime, his measured waiting time $W = \frac{N_c}{\mu} + \frac{1}{\mu} > P$. However, before $\mathcal{C}$ reaches the node n, the queue content $Q(t)$ still increases at the rate λ (as nobody leaves the system during this time interval), implying a delay mechanism in the draining of the queue content. As soon as $\mathcal{C}$ reaches n, and thus leaves the system, a second dynamical phase is triggered. In this second phase, the customers arriving immediately after $\mathcal{C}$ do also experience a waiting time exceeding P and then will also leave the system. As $\lambda < \mu$, the queue population $Q(t)$ decreases in the second dynamical phase and the depletion lasts until a satisfied customer and his immediate successors reach the node n. When this happens, the first dynamical phase starts again and $Q(t)$ fills up at rate λ. The iteration of these two dynamical phases produces a cyclo-stationary behavior whose very existence is entirely due to the agents' ability to record W and take their decisions accordingly. It is instructive, in particular to understand the underlying delay mechanism, to visualize the dynamics of the queue content by using the hydrodynamic picture sketched in Fig. 2. The dynamics of the
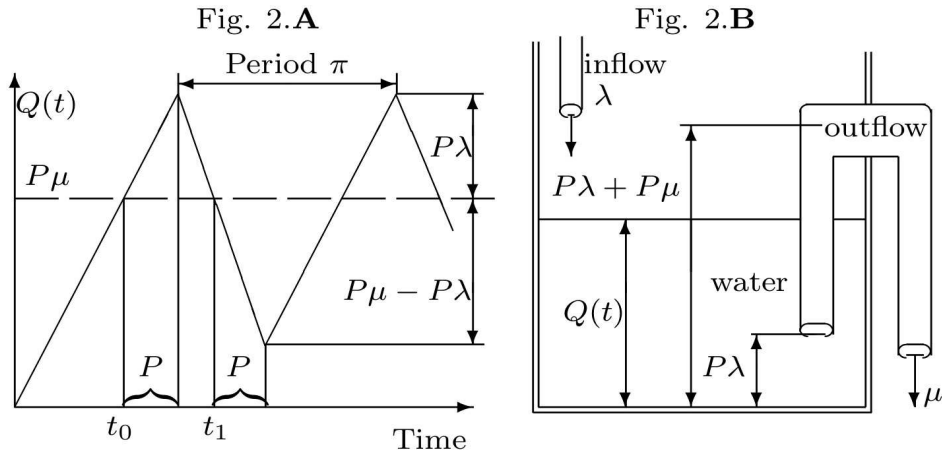

Figure 2: Following (Filliger and Hongler, 2005), we have **A.** The agent entering at t_0 is the first one of a whole cluster U of unsatisfied customers and triggers the alternation of $Q(t)$ from the increasing to the decreasing state at $t_0 + P$. The last agent belonging to the cluster of unsatisfied customers U is the one entering just before t_1 and triggers the switch of $Q(t)$ from the decreasing to the increasing state at $t_1 + P$. This simple delay dynamic repeats and creates stable oscillations. **B.** The siphon model. The queue length corresponds to the water level $Q(t)$. The inflow and outflow rates are λ respectively μ. The siphon leaves a water residue of hight $P\lambda$ due to the constant inflow during P. The effective siphon length is $P\mu$.

vessel content, in particular its oscillations is self-explanatory. In addition, the purely deterministic context enables an elementary derivation of both the amplitude Δ and the period Π of the queue population. Following (Filliger and Hongler, 2005), we obtain:

$$\Delta = P\mu, \tag{1}$$

$$\Pi = P\left[2 + \frac{\lambda}{\mu - \lambda} + \frac{\mu - \lambda}{\lambda}\right] \tag{2}$$

and both Eqs.(1) and (2) are in perfect agreement with simulation experiments, as discussed in (Hongler et al., 2004) and (Filliger and Hongler, 2005).

3   Bifurcation of feedback queues

Here, we consider the network $\mathcal{D}$, formed by a bifurcation of feedback queues, as sketched in Fig. 3. Two feedback queueing systems of the type introduced in section 2 are placed in parallel. The total incoming external customers feeding this system is a renewal process with rate Λ. At a first decision node (DN) n_e (where e stands for *entry*), the agents face two routing possibilities: to either join server S_u or to join S_d. In front of S_u and S_d, the agents wait in queues and the respective time dependent queue contents will be denoted by $Q_u(t)$ and $Q_d(t)$ (the indices u and d standing for *up* and *down* respectively). We shall write by μ_u and μ_d the respective service rates of S_u and S_d. The capacities of both queues $Q_u(t)$ and $Q_d(t)$ are assumed to be unlimited for both, the service policy is FIFO. The presence of the feedback loops introduces two DN's n_u and n_d. At n_u and
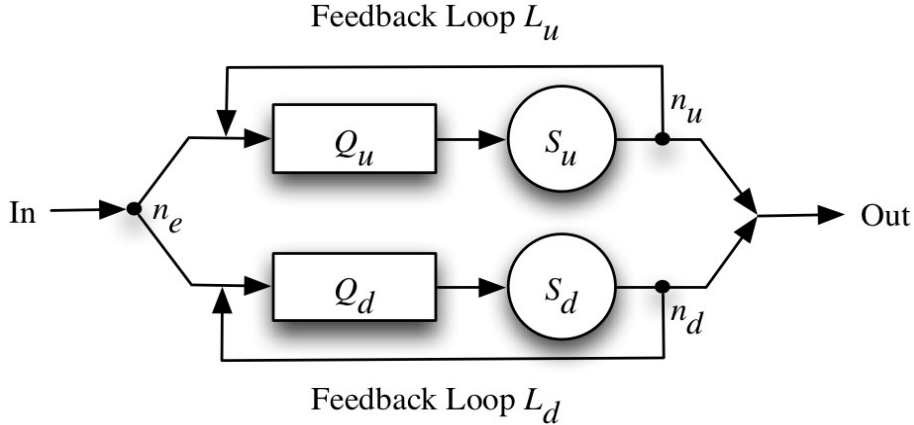
Figure 3: A bifurcation of queueing systems with feedback loop.

n_d, as in section 2, the decision to enter into the feedback loop depends on the waiting time W individually measured by each customer. In the sequel, we will separately consider three typical scenari depending on the agents' ability to gather information.

3.1 Fixed dispatching rule

Let us start with agents only being able to record the time spent in the system (i.e. queueing time + service time). In this case, the decision at node n_e does not depend on agents' "intelligence" and an incoming customer selects between S_u and S_d by using either a deterministic or a random rule, independently with regard to the content of the queues forming $\mathcal{D}$. Typical cases would be:

i) **Deterministic polling**. In this case, the time horizon is divided into deterministic intervals T_u and T_d during which S_u and S_d respectively are alternatively fed with the total incoming traffic Λ. The conditions $\rho_u = \frac{T_u}{T_u+T_d} \cdot \frac{\Lambda}{\mu_u} < 1$ and $\rho_d = \frac{T_d}{T_u+T_d} \cdot \frac{\Lambda}{\mu_d} < 1$ ensure the stability of the system. In view of section 2, it is not surprising that stable oscillations of the queue contents will, here again, be observed. However, the alternative feeding of the servers implies that the evolution of the queue contents exhibits indentations, instead of being smooth. The frequency of the alternations determines the indentation structure. Qualitatively, increasing the frequency of the alternations does decrease the roughness of the curve. For large P, the amplitudes and frequencies of the two decoupled oscillations are determined using Eqs.(1) and (2) with the parameters μ_u, $\lambda_u = \frac{T_u\Lambda}{T_u+T_d}$ on one hand and μ_d, $\lambda_d = \frac{T_d\Lambda}{T_u+T_d}$ on the other hand.

ii) **Random dispatching rule**. Here, we typically consider a Bernoulli sampling of the incoming flow, where the Bernoulli random variable is determined by a parameter r, $(0 \leq r \leq 1)$. A partial traffic with rate $r\Lambda$ enters into server S_u while a traffic with rate $(1-r)\Lambda$ enters into S_d. As the deterministic polling, the Bernoulli sampling implies that both systems S_u and S_d evolve independently. Their two individual dynamics follow the discussion given in section 2 and, for large P, two decoupled cyclo-stationary oscillations with amplitudes and frequencies determined using Eqs.(1) and (2) with the parameters μ_u, $r\Lambda$ on one hand and μ_d, $(1-r)\Lambda$ on the other hand.

3.2 Dispatching based on partial observation of the queues - Noise induced stabilization

Besides chronometers to record W, we equip now each customer with a vision system, thus enabling him to observe, in real time, the instantaneous queue content $Q_u(t)$ in front of S_u. In contrary, the real time content $Q_d(t)$ always remains hidden to the incoming agents, although they do know the average service rate μ_d. At time t, an incoming agent at node n_e first observes the queue content $Q_u(t)$ and, based on his observation, decides either to enter S_u or to join S_d. Once entered into a queue, neither reneging nor jockeying (i.e. jumping between S_u and S_d) is allowed. Note that except the presence of feedback loops, this network configuration is fully similar to the two gas stations network studied in (Hassin, 1996). In this contribution, two gas stations are located one after the other on a main road. A driver who needs to refuel is only able to observe the queue length $Q_u(t)$

5

at the first station (which would be here S_u). Then, he compares $Q_u(t)$ to the conditional expected queue content at the second station (here S_d) and decides either to enter into the first station or to wait and enter into the second one. Returning to our present model, we assume from now on that an incoming agent decides:

a) either to enter S_u whenever $Q_u(t)$ strictly stays below a threshold value N^* (i.e. when $Q_u(t) < N^*$)

b) or to enter into S_d otherwise.

At the DN's n_u and n_d, the routing depends, as in section 2, on a patience parameter P which is again assumed to be common to all agents. The patience P and the threshold control parameter N^* will be related by assuming that:

$$P \geq \frac{N^* + 1 + \delta}{\mu_u}, \qquad \delta \in \mathbb{N}^+. \qquad (3)$$

We can interpret the decision at n_u whether to engage or not into the feedback loop as a formal illustration of the H. Maister's first principle of the psychology of waiting lines (Maister, 1985), namely: *"Satisfaction equals perception minus expectation"*. Indeed, at the DN n_e, the level N^* defines via P as given by Eq.(3) an expected admissible waiting time. Later, when reaching n_u, each agent compares his actually measured waiting time (playing the role of the *perceived* waiting time) with P (playing the role of the *expected* waiting time) and then take his routing decision.

Consider first the deterministic dynamics where S_u operates with a fixed service time $\frac{1}{\mu_u}$. When, at a given time t, $Q_u(t) = N^*$ agents are waiting in front of S_u, they will remain loyal to S_u forever (i.e. these agents will loop forever and ever). Indeed, their measured waiting time W never exceeds P and, the dynamics being deterministic, no perturbation will alter this dynamically "frozen" situation. In particular, once $Q_u(t) \equiv N^*$, the server S_u is definitively unavailable for any external incomer and the global incoming traffic with rate Λ is entirely dispatched to S_d. Whenever $\frac{\Lambda}{\mu_d} > 1$, the queueing system will thus be unstable (i.e. $\lim_{t \to \infty} Q_d(t) = \infty$). Assume now that random fluctuations affect the service times of S_u. While Eq.(3) is still satisfied on average, service time noise triggers, at node n_u, a random flow of unsatisfied customers, who will definitively leave the system. Hence, the very presence of noise (in the service time) does effectively increase the availability of S_u. Consequently, a part of the global incoming traffic will now be processed by S_u. For a selected range of control parameters, we may simultaneously have:

$$\frac{\alpha \Lambda}{\mu_u} < 1 \quad \text{and} \quad \frac{(1 - \alpha) \Lambda}{\mu_d} < 1, \quad 0 \leq \alpha \leq 1, \qquad (4)$$

where $\alpha \Lambda$ and $(1 - \alpha)\Lambda$ stand for the rates of the stationary average partial traffic flows feeding S_u and S_d respectively. When no fluctuations affect the service time of S_u, then $\alpha \equiv 0$. Whenever Eq.(4) holds, both queueing branches are stable. The previous qualitative reasoning suggests that it exists a critical variance $\sigma_{u,c}^2$ of the service time of S_u (and hence a critical value α_c) such that:

a) for $\sigma_u^2 \geq \sigma_{u,c}^2$, the queueing system is stable.

b) for $\sigma_u^2 < \sigma_{u,c}^2$, the queueing system is unstable.

3.2.1 Experimental observations.

The above dynamical behavior is easily observed in simulation experiments where the incoming flow of customers is an exponential process with parameter Λ and the S_u service times are drawn from a probability density $dB_u(x)$ being:

a) uniform with support $\left[\frac{1}{\mu_u} - \xi, \frac{1}{\mu_u} + \xi\right]$ with $\xi \geq 0$ (thus $\sigma_u^2 = \frac{\xi^2}{3}$). The following numerical values were used: $\Lambda = 1.11$, $\frac{1}{\mu_u} = \frac{1}{\mu_d} = 1$, $N^* = 28$ and $P = 30$ (i.e. $\delta = 1$ in Eq.(3)). We observe that for $\xi \geq 0.118 \Rightarrow \sigma_{u,c}^2 \geq 0.0046$, the queueing system remains stable, while it becomes unstable (i.e. $\lim_{t \to \infty} Q_d(t) = \infty$) for smaller values of ξ.

b) a Normal law $\mathcal{N}(\frac{1}{\mu_u}, \sigma_u^2)$. For the same numerical values as above, we observe that for $\sigma_u^2 \geq \sigma_{u,c}^2 = 0.0046$, the queueing system remains stable, while it becomes unstable for $\sigma_u^2 < \sigma_{u,c}^2$.

3.2.2 Analytical approach.

To analytically discuss the stability issue reported above, let us consider the situation where the service times of S_u are independent Bernoulli random variables with values $\left\{\frac{1}{\mu_u}, \frac{1}{\mu^+}\right\}$ and corresponding probabilities $(1-q)$ and q respectively, $0 \leq q << 1$. We assume that $\mu^+ < \mu_u$ and interpret

$\frac{1}{\mu^+}$ (with $\frac{1}{\mu^+} > \frac{1}{\mu_u}$) as the effective service time occurring when a failure alters the ordinary behavior of the server S_u. Remember that the agents follow the FIFO rule and are homogenous in their patience parameter P, chosen here to fulfill:

$$P < \frac{N^*}{\mu_u} + \frac{1}{\mu^+} \quad \text{and} \quad P > \frac{N^* + 1}{\mu_u}. \tag{5}$$

When, at a given time t, $Q_u(t) \equiv N^* - 1$, an incoming tagged customer $\mathcal{C}$ at DN n_e will decide to enter S_u. Later on, when $\mathcal{C}$ reaches n_u, he will, according to Eq.(5), choose:

    a) either to follow the feedback loop, whenever no failure occurred during the service of the N^* customers who were directly in front of him (including the customer who was served when $\mathcal{C}$ joined $Q_u(t)$) and during his own service

    b) or to leave the system, whenever one or more failures occurred during the service of the N^* customers who were directly in front of him and during his own service.

Hence, in absence of failures and when $Q_u(t) \equiv N^*$, the agents will remain in the feedback loop forever and, at DN's n_e and n_u, neither an externally new incomer nor a leaving customer will be observed. However, as soon as failures occur in S_u, Eq.(5) implies that one or more customers will definitively leave the system after the decision at n_u. Hence, this implies that the global incoming traffic will now be shared between S_u and S_d. Assume that:

$$\mu_d < \Lambda \qquad \Longleftrightarrow \qquad \rho_d = \frac{\Lambda}{\mu_d} > 1. \tag{6}$$

Thus, S_d cannot sustain alone the full traffic load without being in an unstable regime ($\rho_d > 1 \Rightarrow \lim_{t \to \infty} Q_d(t) = \infty$). Remember that $\alpha\Lambda$ and $(1-\alpha)\Lambda$ denote the rates of the average partial traffics processed by S_u and S_d respectively. It exists a critical incoming flow, defined by $(1 - \alpha_c)\Lambda$, above which the queue $Q_d(t)$ becomes unstable. For the associated traffic intensities, this implies that:

$$\rho_u = \frac{\alpha\Lambda}{\mu_u} < 1 \quad \text{and} \quad \rho_d = \frac{(1-\alpha)\Lambda}{\mu_d} < 1, \ \forall \alpha > \alpha_c, \tag{7}$$

$$\rho_{d,c} = \frac{(1-\alpha_c)\Lambda}{\mu_d} = 1, \tag{8}$$

where $\rho_{d,c}$ is the critical traffic load driving the queue $Q_d(t)$ to its marginal stability regime.

To proceed further with analytical considerations, let us now focus on rare events regimes (RER), for which more than a single failure during $N^* + 1$ consecutive ordinary services is a highly improbable event. As N^* is the threshold value governing the decision at node n_e and P fulfills Eq.(5), the RER is expected when $N^* + 1 << \frac{1}{q}$. Under the RER, each failure triggers the drainage of the queue $Q_u(t)$. Indeed, due to the FIFO scheduling rule, when a failure occurs at time t, the last agent in $Q_u(t)$ will experience a waiting time larger than P when arriving at n_u. So will also do the $N^* - 1$ agents directly lining behind him (i.e. these are the loyal customers traveling in the loop and feeding $Q_u(t')$ for $t' > t$). As it has been discussed in section 2, this produces a *siphon avalanche*, here of size N^*. In the RER, the succession of these siphon events will be approximately uncorrelated. Hence, in the stationary regime, we can simply estimate the outgoing flow rate λ_u at DN n_u as being given by:

$$\lambda_u = \text{Prob}\{a \text{ single failure occurs}\} N^* \mu_u = q N^* \mu_u. \tag{9}$$

When Eq.(9) holds, the partial traffic on S_d is given by:

$$\rho_d = \frac{\lambda_d}{\mu_d} = \frac{\Lambda - \lambda_u}{\mu_d} = \frac{\Lambda - q N^* \mu_u}{\mu_d}. \tag{10}$$

The marginal stability of queue $Q_d(t)$ is attained at the critical traffic $\rho_d = \rho_{d,c} = 1$, which implies:

$$q \geq q_c := \frac{\Lambda - \mu_d}{N^* \mu_u}. \tag{11}$$

In terms of α_c, we can write:

$$\alpha_c = 1 - \frac{\mu_d}{\Lambda}. \tag{12}$$

Finally, we can also express the stability condition given by Eq.(11) in terms of the critical variance $\sigma_{u,c}^2$ of the underlying Bernoulli random variable. We obtain:

$$\sigma_u^2 \geq \sigma_{u,c}^2 = q_c(1 - q_c)\left(\frac{1}{\mu^+} - \frac{1}{\mu_u}\right)^2. \tag{13}$$

The numerical experiments reported in Tab. 1 are in perfect agreement with Eqs.(11) to (13).

Global incoming traffic Λ	Simulated stability condition on q	Simulated stability condition on σ_u^2
1.05	0.0017	0.00075
1.1	0.0034	0.0015

Table 1: Stability conditions obtained when using a discrete events simulator with the following parameters: $N^* = 28$, $\frac{1}{\mu_d} = \frac{1}{\mu_u} = 1$, $\frac{1}{\mu_+} = 3$ and $P = 30$. No discrepancy between simulated and theoretical results have been observed up to the shown precision.

While the concept of stabilization by noise is already abundantly discussed in the context of stochastic differential equations (Has'minskiĭ, 1980), (Arnold et al., 1983) and (Ruszczynski and Kish, 2000), our present class of models exemplifies clearly that such a random stabilization can be encountered in multi-agent systems where a non-linearity (in our case, the feedback loop) is present.

Remark 3.1. When the assumptions for RER are not satisfied (i.e. more than a single failure during $N^* + 1$ consecutive ordinary services is not a highly improbable event), then Eqs.(11) to (13) are not valid anymore. Indeed, it is now possible (and not improbable) that the siphon avalanches due to two successive single failures overlap. Hence, in this case, the number of agents that leave the system after a failure is not always equal to N^* (but is above-bounded by N^*) and thus the number of single failure events needed to reach the critical partial traffic $\rho_{d,c}$ is underestimated in Eqs.(11) to (13), as it is testified by numerical experiments (see Tab. 2).

Global incoming traffic Λ	Theoretical stability condition on q	Simulated stability condition on q
1.2	0.0069	0.0074
1.3	0.0103	0.124

Table 2: Theoretical vs. numerical stability conditions obtained when using a discrete events simulator with the following parameters: $N^* = 28$, $\frac{1}{\mu_d} = \frac{1}{\mu_u} = 1$, $\frac{1}{\mu_+} = 3$ and $P = 30$.

3.3 Flow dispatching based on fully observable queues - Synchronization of oscillations

In this section, we assume that both queues $Q_u(t)$ and $Q_d(t)$ can be observed simultaneously by the incoming agents. Thus, compared with section 3.2, the information gathering process has been further increased. Based on the queue contents, several dispatching policies at the DN n_e can be constructed. Among the simplest and natural rules, let us here focus on the policy sending a new externally incoming customer to the shortest observed queue. As we shall now see, this shortest queue first (SQF) rule implies the natural emergence, for large common patience parameter P, of *synchronized stable temporal oscillations* of the queue contents $Q_u(t)$ and $Q_d(t)$. This happens for any initial conditions of the queue populations. As before, when P is large and common to all agents, a purely deterministic approach is perfectly suitable. We assume that $\frac{\Lambda}{\mu_u + \mu_d} < 1$, which ensures the stability of the system. Let us consider the two selected configurations characterized by:

a) two identical servers (i.e. $\frac{1}{\mu_u} = \frac{1}{\mu_d} = \frac{1}{\mu}$).

b) two servers with service rate ratio $\mathcal{S} = \frac{\mu_u}{\mu_d} \neq 1$ (i.e. one of the two servers is faster than the other).

In configuration a), the total incoming traffic is evenly divided between the two servers, both receiving a partial traffic with rate $\frac{\Lambda}{2}$. The amplitude and period of the common synchronized stable temporal oscillations of the queue contents $Q_u(t)$ and $Q_d(t)$ are given by Eqs.(1) and (2) with parameters $\frac{\Lambda}{2}$ and μ.

Consider now configuration b) and suppose, without loss of generality, that $\frac{1}{\mu_u} > \frac{1}{\mu_d}$. We observe the following dynamics: even though the servers do not work at the same speed, the queue contents $Q_u(t)$ and $Q_d(t)$ are equal at any time, provided $\frac{\Lambda}{\mu_d} > 1$ (i.e. provided S_d is not able to handle alone the total incoming flow). The greater speed of S_d implies that the customers joining

this server will remain satisfied for a longer queue length than with S_u. As a consequence of the SQF rule, there will be more unsatisfied customers with server S_u and this server will thus process a greater part of the global incoming traffic than S_d (i.e. S_u will absorb more fresh customers, but these customers will stay less time in the system than those joining S_d). As shown in Fig. 4, two cases may emerge. For both, it is possible to fully characterize the emergent common synchronized stable temporal oscillations.
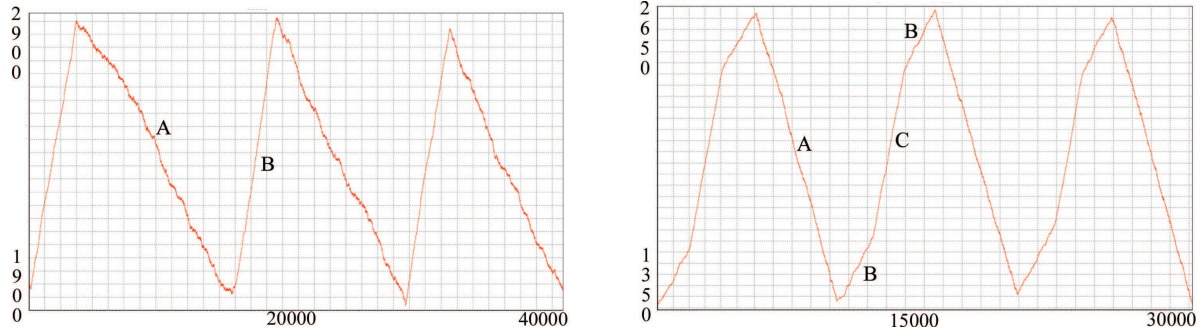

Figure 4: The SQF policy implies that $0 \leq |Q_u(t) - Q_d(t)| \leq 1$, $\forall t$. Thus, we only show the state of $Q_u(t)$ in the above figures. *Left*: Queue content $Q_u(t)$ when $P = 2500$ and processes are exponential with parameters $\Lambda = 1.25$, $\frac{1}{\mu_u} = 1.6$ and $\frac{1}{\mu_d} = 1.2$. The amplitude and period of the common synchronized stable temporal oscillations are given by $\Delta = \frac{P\mu_d}{2}$ and $\Pi = P \left(\frac{\mu_d}{\mu_d + \mu_u - \Lambda} + \frac{\mu_d}{\Lambda - \mu_u} \right)$ respectively. The two different slopes are given by $\mathbf{A} = \frac{\Lambda - \mu_u - \mu_d}{2}$ and $\mathbf{B} = \frac{\Lambda - \mu_u}{2}$. *Right*: Queue content $Q_u(t)$ when $P = 2500$ and processes are exponential with parameters $\Lambda = 0.9$, $\frac{1}{\mu_u} = 1.6$ and $\frac{1}{\mu_d} = 1.2$. The dynamics differs from the *Left* behavior by the presence of a time interval with slope $\mathbf{C} = \frac{\Lambda}{2}$. During this interval, customers in S_u and S_d are all satisfied. On the other hand, during the time intervals with slope $\mathbf{A}$ and $\mathbf{B}$, the customers in S_u are unsatisfied (the customers in S_d being unsatisfied only during the interval with slope $\mathbf{A}$). For instance, in the *Left* configuration, all the customers joining S_u are unsatisfied, because $Q_u(t)$ always remains above the critical threshold. The complexity of the dynamics in the *Right* case requires more involved computations, which precludes to give simple and compact expressions for the amplitude and the period of the synchronized oscillations. However, due to the deterministic nature of the dynamics (when P is large), an analytical characterization is still feasible.

4 Conclusion

Networks where circulating items are endowed with elementary forms of "intelligence" which affects their routing decisions clearly offer a high relevance for applications. As so far relatively little analytical work has being devoted to such systems, we address here this issue. By adopting a multi-agent point of view, our note explores, mostly analytically, the behavior of simple queueing networks where agents decide, based on individually measured waiting times, whether to visit or not feedback loops present in the network topology. The underlying idea to study such systems originates from questions related to the loyalty of customers to a particular service provider. It is remarkable that this simple class of models is already rich enough to exhibit temporal patterns of the queue contents (i.e. stable temporal oscillations), synchronization and stabilization by noise phenomena, which are typical for multi-agent systems.

Acknowledgements

References

[1] Arnold, L., Crauel, H.,Wihstutz, V. (1983), Stabilization of Linear Systems by Noise. SIAM Journal on Control and Optimization, 21(3):451-461.

[2] Chen, H., Yao, D.D. (2001). Fundamental of Queueing Networks. Springer, Inc., New York.

[3] Chu, F., Xie, X.L. (1997). Deadlock Analysis of Petri Nets Using Siphons and Mathematical Programming. IEEE Transactions on Robotics and Automation, 13(6):793-804.

[4] D'Avignon, G.R., Disney, R.L. (1976). Single-Server Queues with State-Dependent Feedback. INFOR Journal, 14(1):71-85.

[5] Filliger, R., Hongler, M.-O. (2005). Syphon Dynamics - A Soluble Model of Multi-Agents Cooperative Behavior. Europhysics Letters, 70:285-291.

[6] Has'minskiĭ, R.Z. (1980). Stochastic Stability of Differential Equations, Sijthoff & Noordhoff, Alphen aan den Rijn.

[7] Hassin, R. (1996). On the Advantage of Being the First Server. Management Science, 42(4):618-623.

[8] Hongler, M.-O., Cheihkrouhou, N., Glardon, R. (2004). An Elementary Model for Customer Fidelity. Proceedings of MOSIM-04, 2(899-906), Lavoisier Editions, Nantes (France).

[9] Maister, D.H. (1985). The Psychology of Waiting Lines. The Service Encounter, Chap. 8, Czepiel, J.A., Solomon, M.R., Surprenant D.F. (eds.), D.C. Heath, Lexington, Mass.

[10] Mikhailov, A.S., Calenbuhr, V. (2002). From Cells to Societies. Springer, Inc., New York.

[11] Peköz, E.A., Joglekar, N. (2002). Poisson Traffic Flow in a General Feedback Queue. Journal of Applied Probability, 39(3):630-636.

[12] Ruszczynski, P.S., Kish, L.B. (2000). Noise Enhanced Efficiency of Ordered Traffic. Physics Letters A, 276:187-190.

[13] Schweitzer, F. (2003). Brownian Agents and Active Particles. Springer, Inc., New York.

[14] Takàcs, L. (1963). A Single-Server Queue with Feedback. The Bell System Technical Journal, 42:505-519.

[15] Zohar, E., Mandelbaum, A., Shimkin, N. (2002). Adaptive Behavior of Impatient Customers in Tele-Queues: Theory and Empirical Support. Management Science, 48(4):566-583.