CCN Interest Forwarding Strategy as Multi-Armed Bandit Model with Delays

Konstantin Avrachenkov
INRIA Sophia Antipolis
France
Email: k.avrachenkov@sophia.inria.fr

Peter Jacko
BCAM – Basque Center for Applied Mathematics
Bilbao, Spain
Email: jacko@bcamath.org

***Abstract*—We consider Content Centric Network (CCN) interest forwarding problem as a Multi-Armed Bandit (MAB) problem with delays. We investigate the transient behaviour of the ε-greedy, tuned ε-greedy and Upper Confidence Bound (UCB) interest forwarding policies. Surprisingly, for all the three policies very short initial exploratory phase is needed. We demonstrate that the tuned ε-greedy algorithm is nearly as good as the UCB algorithm, commonly reported as the best currently available algorithm. We prove the uniform logarithmic bound for the tuned ε-greedy algorithm in the presence of delays. In addition to its immediate application to CCN interest forwarding, the new theoretical results for MAB problem with delays represent significant theoretical advances in machine learning discipline.**

I. Introduction

There is a conceptual clash between rapidly expanding digital information dissemination and the host-based network architecture of the current Internet. To facilitate the dissemination of digital information, several Information-Centric Network (ICN) architectures have been proposed: TRIAD [9], DONA [13], CCN/NDN [11]. Since the CCN/NDN (Content-Centric Networking / Named Data Networking) proposal appears to be the most elaborate, we develop our contribution in the framework and within the terminology of CCN/NDN. For the sake of brevity, we shall refer to CCN/NDN as CCN. The main features of the ICN paradigm, and the CCN architecture in particular, are that the content is addressed by a unique name and can have many identical cached copies in the CCN routers across the Internet. Any of such copies can be retrieved independently of its location. The content is typically divided into several small chunks. A chunk is also uniquely identified. A chunk of content is located and requested by forwarding so-called interests. A user or a CCN router can forward interests to one or more neighbour CCN routers. Clearly, if there is no bandwidth limitation the most efficient way is to forward interests to all available neighbour routers. However, if there is a bandwidth limitation or the interest sender has to pay for the interest or/and delivered content, there can be better interest forwarding strategies than simple flooding.

In the present work we suggest to view the problem of optimal interest forwarding strategy as a variant of the Multi-Armed Bandit (MAB) problem. The MAB problem is a classical problem in probability theory in which a decision maker finds an optimal balance between exploration and exploitation efforts. Its name originates in the example of a gambler facing the problem of playing at every slot one of multiple one-armed bandits (slot machines) that yield random rewards. The optimal solution is known [8], but depends heavily on the statistical assumptions of the rewards, and may be computationally intractable in many cases. Here we adopt three well known algorithms from the machine learning literature: ε-greedy [17], tuned ε-greedy and UCB [1], which are considerably simpler and more robust.

Our study brings advances to both networking and machine learning disciplines. We show that the MAB algorithms allow to detect the optimal router (i.e., any router of those with smallest mean delay) with very small number of interests sent to sub-optimal routers. The novelty from machine learning perspective is that we analyze the transient period of the MAB algorithms with delays. This is a very challenging topic with hardly any results available in the literature. In fact, we are only aware of the work [7], [16] and [4] on MAB with delay. However, the approach in these papers is to partially characterize the optimal or asymptotical optimal strategy, which was performed under very restrictive assumptions that are not realistic for the Internet.

We expect that our MAB-based mechanisms can be integrated in the Interest Control Protocol (ICP) which regulates the pacing of interests in CCN [5]. Several studies in that respect have appeared recently, realizing the importance of this problem by its analogy to the congestion control and routing in the present Internet. For instance, [15] introduced a hop-by-hop rate-based mechanism to control the transmission buffer occupancy. Further, [6] argued that when no (or not useful) information is available in the forwarding table, then the neighbourhood need to be explored, but multi-path flooding might lead to inefficient use of resources such as link capacity and cache space.

The paper is organized as follows. In Section II we present a formal model of the problem and describe three algorithms that we propose for CCN interest forwarding. We analyze the initial exploratory phase of these algorithms in Section III, both numerically and mathematically, providing a bound and an approximation of its duration. In Section IV we study the exploitation phase of the tuned ε-greedy algorithm and prove a logarithmic bound on the probability of choosing a suboptimal router. Section V concludes. Because of the lack of space, we omit all the proofs which can be found in the accompanying

research report available online [2].

II. Model and interest forwarding strategies

We suppose that a CCN router or a user can forward interests to K CCN neighbour routers. We consider a discrete time model. The slot duration can be chosen equal to the minimal duration of packet generation at the MAC layer. Therefore, we assume that at each time slot $t \in \mathcal{T} := \{0, 1, 2, \dots\}$ the user can send only one interest to one of K CCN neighbour routers.

CCN routers reply with delays distributed according to discrete distribution functions $F_k(x)$, $k = 1, ..., K$, $x = 1, 2, ...$ with mean denoted by μ_k. Specifically, we assume that a chunk corresponding to the interest generated at the present slot and forwarded to the neighbour router k is delivered by router k after a random number of slots distributed according to the distribution function $F_k(x)$. Thus, we shall know the effect of the action taken at the time slot t only at the future time slot $t + X_k(t)$, where $X_k(t)$ is an i.i.d. random variable generated according to $F_k(x)$. This delay may be a consequence of several aspects, including the number of hops necessary to pass in order to find the chunk, caching algorithms, network conditions, etc.

We are interested in minimizing the expected number of interests sent to sub-optimal routers, or to sub-optimal arms in terminology of the multi-armed bandit framework [17]. The challenging novelty of our setting with respect to the classical multi-armed bandit problem formulation is that the cost becomes known to the decision maker with delays. In fact, the costs are the delays.

The optimal policy in the classical setting without delay is obtained by the Gittins index rule [8], which breaks the combinatorial complexity of the problem by computing the Gittins index (a history-dependent function) for each router in isolation and then simply sending the interest at every slot to the router whose current Gittins index value is lowest. This result significantly reduces the dimensionality of the problem, but the evaluation of the Gittins index may still be computationally tedious, especially if the index depends on the whole history, not only on the last observed state. Moreover, the Gittins optimality result requires that the evolution of costs from routers be mutually independent, while the algorithms described below are efficient even for dependent arms [1].

Since strictly speaking optimal policy is very likely to be very complex even in the classical setting without delay, many researchers have proposed sensible policies and shown desirable properties of such policies [12], [1]. One desirable property of the multi-armed bandit problem policy is the uniform logarithmic bound on the number of sub-optimal arms chosen by the decision maker. We shall establish the uniform logarithmic bound for the tuned ε-greedy policy in the case of delayed information in Section IV.

In the present work we consider the following three algorithms: ε-greedy algorithm, tuned ε-greedy algorithm, and UCB (Upper Confidence Bound) algorithm. These are the most popular multi-armed bandit algorithms, and in this paper

1) **Initialization:** Choose $t_0 \in \mathcal{T}$ and $\varepsilon \in (0, 1)$. During the first t_0 slots keep sending interests to routers in round robin fashion or randomly to routers chosen according to the uniform distribution.
2) **at each time slot** $t \geq t_0$ **do**
3) For each router k, compute the average delay:
$$\overline{X}_{k,T_k(t)} = \frac{1}{T_k(t)} \sum_{\tau=0}^{t-1} A_k(\tau, t) X_k(\tau)$$
4) For each router k, set the index:
$$\nu_k(t) = \overline{X}_{k,T_k(t)}.$$
5) With probability $1 - \varepsilon$ send new interest to the router with the smallest index or with probability ε send new interest to a uniformly randomly chosen router.
6) **end for**

Fig. 1. Algorithm ε-greedy

1) **Initialization:** Choose $t_0 \in \mathcal{T}$ and $\varepsilon_0 \in (0, t_0)$. During the first t_0 slots keep sending interests to routers in round robin fashion or randomly to routers chosen according to the uniform distribution.
2) **at each time slot** $t \geq t_0$ **do**
3) For each router k, compute the average delay:
$$\overline{X}_{k,T_k(t)} = \frac{1}{T_k(t)} \sum_{\tau=0}^{t-1} A_k(\tau, t) X_k(\tau)$$
4) For each router k, set the index:
$$\nu_k(t) = \overline{X}_{k,T_k(t)}.$$
5) With probability $1 - \varepsilon_0/t$ send new interest to the router with the smallest index and with probability ε_0/t send new interest to a uniformly randomly chosen router.
6) **end for**

Fig. 2. Algorithm tuned ε-greedy

1) **Initialization:** Choose $t_0 \in \mathcal{T}$ and $L > 0$. During the first t_0 slots keep sending interests to routers in round robin fashion or randomly to routers chosen according to the uniform distribution.
2) **at each time slot** $t \geq t_0$ **do**
3) For each router k, compute the average delay:
$$\overline{X}_{k,T_k(t)} = \frac{1}{T_k(t)} \sum_{\tau=0}^{t-1} A_k(\tau, t) X_k(\tau)$$
4) For each router k, set the index:
$$\nu_k(t) = \overline{X}_{k,T_k(t)} - \sqrt{\frac{L \ln(t)}{T_k(t)}}$$
where L is so-called exploration parameter.
5) Send new interest to the CCN router with the smallest index.
6) **end for**

Fig. 3. Algorithm Upper Confidence Bound (UCB)

Parameters	Router 1	Router 2	Router 3
propagation delay	2	2	2
p parameter	0.8	0.7	0.6
r parameter	10	10	10
mean delay	4.5	6.29	8.67
std	1.77	2.47	3.33

TABLE I
THE VALUES OF PARAMETERS IN THE NUMERICAL EXAMPLE.

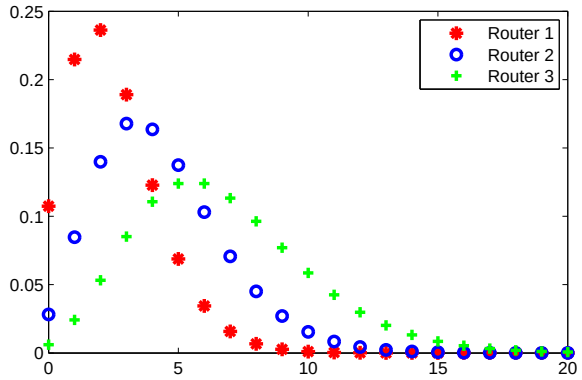

Fig. 4. Negative binomial distributions in example.

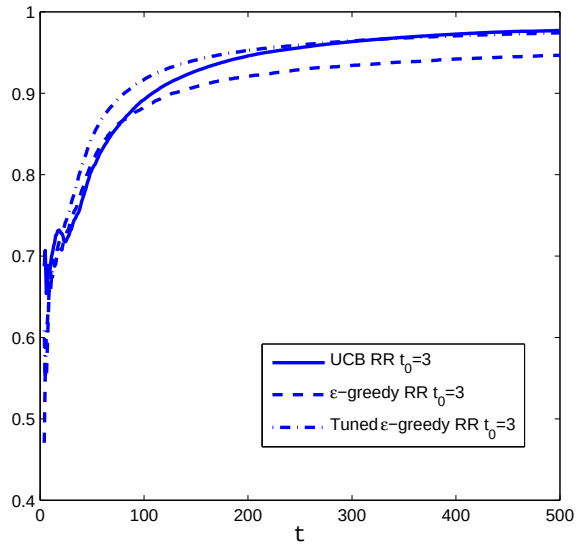

Fig. 5. The time evolution of the probability of sending the interest to an optimal router by the three MAB algorithms.

we propose their generalizations to the setting with delayed information.

Denote by $T_k(t)$ the total number of interests sent to router k and answered up to the end of slot $t-1$, and

$$A_k(\tau,t) := 1\{\text{interest sent to } k \text{ at } \tau \text{ and answered up to the end of slot } t-1\}.$$

Let us formally describe each algorithm. The ε-greedy algorithm presented in Figure 1 is the simplest algorithm. Its main drawback is that the expected number of sub-optimal arms grows linearly in time. A variant of ε-greedy algorithm was proposed in [17] for Markov Decision Process models without delay.

The tuned ε-greedy algorithm and UCB algorithm for models without delays have been proposed and analysed in [1]. Both the tuned ε-greedy and UCB algorithms have logarithmic bounds on the number of sub-optimal arms in the case of no delays [1]. The respective variants of these algorithms are presented in Figure 2 and Figure 3.

In our case, since we minimize the cost, we should more appropriately call this algorithm the lower confidence bound algorithm. However, to make an explicit connection with [1] we shall continue to call it the UCB algorithm. In the previous works the UCB algorithm has shown slightly better performance than the tuned ε-greedy algorithm.

To get an idea of the performance of the above algorithms in the presence of delay, we provide a numerical example. We have taken the negative binomial distribution with deterministic shift as the distribution of delay $F_k(x)$ in our numerical examples. There are several reasons for this choice. The negative binomial distribution is quite versatile. With two parameters, we can easily choose any mean and variance, which have simple explicit expressions. The distribution shape can take diverse forms such as the shape of geometric distribution and the shape close to that of the normal distribution. The negative binomial distribution represents the distribution of a sum of geometrically distributed random variables. Since the waiting time distribution in many queueing systems is exponential or close to exponential, the negative binomial distribution represents well the response time of queueing systems in cascade. We introduce the deterministic shift to model the propagation delay. In Table I we present the parameters of our numerical example and in Figure 4 we plot the negative binomial distributions with the chosen parameters.

In Figure 5 we plot the fraction of interests sent to the optimal arm as a function of time for the three algorithms with Round Robin strategy employed in the initial phase. This numerical example demonstrates that despite the presence of delays, the three algorithms perform well. In particular, as in the case of no delay, the performances of the UCB and tuned ε-greedy algorithms are comparable and the ε-greedy algorithm performs not too badly. In the following sections we will provide a detailed analysis of these three algorithms.

III. ANALYSIS OF INITIAL EXPLORATORY PHASE

Let us now investigate the effect of the duration of the initial, purely exploratory, phase on the algorithm performance. We shall consider two possible initial strategies: the Round Robin (RR) strategy and the strategy when the arm chosen randomly with uniform probability (Uni). Note that in the Round Robin strategy the initial arm and the order are chosen randomly with uniform distribution.

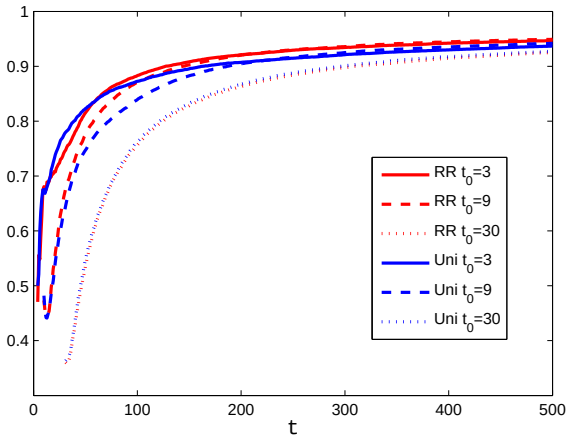

Fig. 6. The effect of the initial phase duration and initial strategy of the ε-greedy algorithm on the probability of sending the interest to an optimal router.

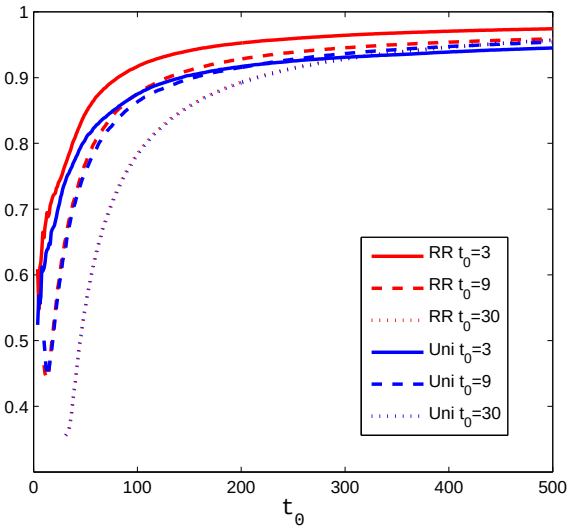

Fig. 7. The effect of the initial phase duration and initial strategy of the tuned ε-greedy algorithm on the probability of sending the interest to an optimal router.

In Figures 6-8 for our numerical example we plot the fraction of interests sent to the optimal arm for different durations ($t_0 = 3, 9, 30$) of the initial phase for different algorithms with different initial phase strategies.

A bit surprisingly, it turns out that it is better to set up very short duration of the initial phase. Another important observation is that it is better to use the Round Robin initial strategy rather than the uniformly random strategy. This is intuitively expected as by using the Round Robin strategy we reduce the randomness. Below we provide theoretical explanation of these phenomena.

The initial phase $[0, t_0 - 1]$ is characterized by large exploration effort. Here we would like to provide an estimate for the period after which we can with high certainty rely on the

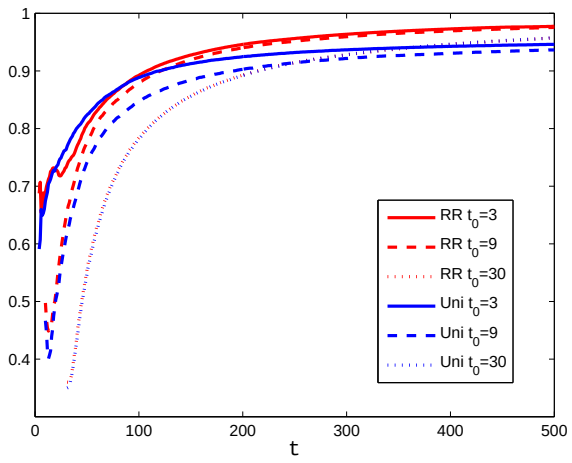

Fig. 8. The effect of the initial phase duration and initial strategy of the UCB algorithm on the probability of sending the interest to an optimal router.

choice of the best performing arm based on evaluated averages. Specifically, let us estimate the probability of choosing the best arm (denoted by $*$) given the arms are chosen independently before the end of the initialization phase.

Denote by I_t the arm chosen at time slot t. Assume first that arms are chosen randomly and independently during the initial phase with probability $p_j := \mathbb{E}[1\{I_t = j\}]$, $j = 1, ..., K$. In the case of uniformly random strategy we have $p_j = 1/K$. Let further D be the maximum possible delay between choosing the arm and observing the realization ($D = 1$ corresponds to no delay, i.e., receiving the chunk always before the end of the slot when an interest was sent) and

$$c_j := D^2 + \frac{\Delta_j}{2} D + \frac{\Delta_j}{2} p_* D,$$

where $\Delta_j = \mu_j - \mu_*$. Then, we have the following result.

Theorem 1: If during the exploration phase we choose the arms randomly and independently with uniform distribution ($p_j = 1/K$), and at the end of the exploration period, at slot t_0, we choose the arm according to the estimated average, the probability of choosing the best arm is lower bounded as follows:

$$\mathbb{P}[\overline{X}_{*,T_*(t_0)} < \min_{j \neq *} \overline{X}_{j,T_j(t_0)}]$$

$$\geq \prod_{j \neq *} \left(1 - \exp\left(-\frac{\Delta_j^2 (t_0 - D)^2}{8K^2 c_j^2 t_0}\right)\right)^2 \qquad (1)$$

A strong point of the above result is that the derived lower bound is given in terms of exponential function, which means that starting from some value of t_0 the probability of success will be very high. However, the bound (1) can be loose. Therefore, next we suggest an approximation of the success probability based on the central limit theorem.

Also, it turns out that if the maximal delay is not too large, we do not introduce a large error by considering only interests sent by the time $t_0 - D$. Then, by the time t_0 we observe reply from all sent interests.

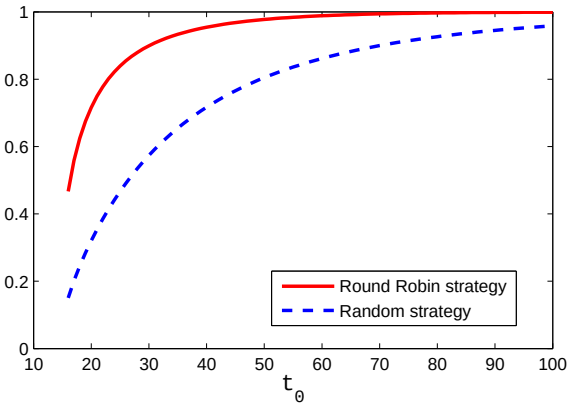

Fig. 9. Approximations for the probability of sending the interest to an optimal router in the first slot after the the end of the initial phase.

Theorem 2: If during the exploration phase we choose the arms randomly and independently with uniform distribution ($p_j = p_* = 1/K$), and if at the end of the exploration period, at slot t_0, we choose the arm according to the estimated average, the probability of choosing the best arm can be approximated as follows:

$$\mathbb{P}[\overline{X}_{*,T_*(t_0-D)} < \min_{j\neq *} \overline{X}_{j,T_j(t_0-D)}]$$

$$\approx \prod_{j\neq *} \Phi\left(\Delta_j \sqrt{\frac{p_j(t_0-D)}{4Var(X_j)+\Delta_j^2(1-p_j)}}\right)$$

$$\Phi\left(\Delta_j \sqrt{\frac{p_*(t_0-D)}{4Var(X_*)+\Delta_j^2(1-p_*)}}\right), \qquad (2)$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal random variable.

In the case when the Round Robin strategy is used in the initial phase, we can provide even sharper approximation.

Theorem 3: If during the exploration phase we choose the arms according to the Round Robin strategy with the first arm and the order chosen randomly with the uniform distribution, and if at the end of the exploration period, at slot t_0, we choose the arm according to the estimated average, the probability of choosing the best arm can be approximated as follows:

$$\mathbb{P}[\overline{X}_{*,T_*(t_0-D)} < \min_{j\neq *} \overline{X}_{j,T_j(t_0-D)}]$$

$$\approx \prod_{j\neq *} \Phi\left(\Delta_j \sqrt{\frac{t_0-D}{3(Var(X_*)+Var(X_j))}}\right). \qquad (3)$$

We consider now our numerical example with truncated negative binomial distributions with $D = 15$. In Figure 9 we plot the approximations (2) and (3) as functions of the duration of the initial exploratory phase t_0, which firstly support our numerical finding that it is enough to have a very short initial phase and secondly confirm our intuition that the Round Robin strategy is better than the random strategy.

One may be interested in rough estimation of the number of time slots after which using estimated averages the optimal arm will be selected with high probability. We can provide recommendation for such value based on (3) and 2-sigma rule. If the arguments of the standard normal distribution function are equal to two, then respective probabilities are greater than 0.977. Thus, we conclude that after the time

$$T \geq D + 12\frac{Var(X_*)+\max_j Var(X_j)}{\min_j \Delta_j^2}, \qquad (4)$$

using the estimated averages and the RR strategy, we select the optimal arm with probability at least 0.977^{K-1}. In our numerical example, after 68 time slots the probability of choosing correctly the optimal arm is estimated to be more than 0.95. This is a conservative estimation and in reality we need even shorter exploratory period.

IV. Logarithmic bound for the tuned ε-greedy algorithm

In this section we finally prove that the regret (cumulative suboptimality) of employing the tuned ε-greedy algorithm is bounded logarithmically in t, which is the same result as for the case without delay (and known to be the best achievable) [1].

Theorem 4: Let $a > 0$ and $0 < d \leq \min_{k:\mu_k>\mu_*} \Delta_k$, and let initial phase be run with the uniformly random strategy. For all $K > 1$ and for all delay distributions $F_1, \ldots, F_K$ with support in $[1, D]$, if algorithm tuned ε-greedy is run with input parameters $t_0 > \varepsilon_0 := aK/d^2$, then the probability that the algorithm chooses in slot $t \geq t_0$ a suboptimal arm j is at most

$$2D\frac{a}{d^2}\left(\ln\frac{td^2e^{1/2}}{aK}\right)\left(\frac{aK}{td^2e^{1/2}}\right)^{\frac{3a}{14d^2}}$$
$$+\frac{16D^3}{d^2}\exp\left\{\frac{D+1}{8}\right\}\left(\frac{aK}{td^2e^{1/2}}\right)^{\frac{a}{8D^2}}+\frac{a}{d^2t}.$$

This bound says that the cumulative probability of suboptimal decisions is logarithmic for a large enough (surely if $a > \max\{14d^2/3, 8D^2\}$), because the instantaneous suboptimality at any slot $t \geq t_0$ is of the order $(K-1)a/d^2t+o(1/t)$ for $t \to \infty$. We conclude that the smaller the number of arms (CCN neighbour routers) and the larger d, the difference between the mean delays of the best and the strictly second-best arm, the better the performance of the tuned ε-greedy algorithm.

V. Conclusion

The contribution of this paper is twofold. First, we have proposed tractable and well-performing interest forwarding algorithms for CCN networks. We have demonstrated that the algorithms work fast and logarithmically few interests are send suboptimally, which means that the resources of the user and CCN routers are efficiently managed. Theoretical bounds show that the learning process is best achievable.

Second, we have also contributed to the theory of the multi-armed bandit problem with delayed information. This is an important and challenging topic with few existing results. We

have provided finite-time analysis of algorithms extended to this setting and showed that the deterioration of their performance due to delays is not significant. Perhaps surprisingly, there is no need to include a long exploratory phase, just a single datum from each arm is sufficient for an efficient performance of the algorithms.

The CCN interest forwarding model presented here is very simple in order to be able to perform its mathematical analysis. From the practical point of view, it would be desirable to study the performance of the proposed algorithms in more realistic systems, for instance taking into account interest forwarding strategies at subsequent routers, caching parameters, correlation in the stream of interests (it is likely that the optimal CCN router changes over time), packet losses and timeouts, etc.

ACKNOWLEDGMENT

We would like to thank Bruno Kauffmann, Luca Muscariello and Alain Simonian for stimulating discussions. This work is supported by Orange France Telecom Grant on Content-Centric Networking. The work of P. Jacko is partially supported by grant MTM2010-17405 (Ministerio de Ciencia e Innovación, Spain) and grant PI2010-2 (Department of Education and Research, Basque Government).

REFERENCES

[1] P. Auer, N. Cesa-Bianchi and P. Fischer, "Finite-time analysis of the multiarmed bandit problem", *Machine Learning*, v.47, pp.235-256, 2002.

[2] K. Avrachenkov and P. Jacko, "CCN Interest Forwarding Strategy as Multi-Armed Bandit Model with Delays", *INRIA Research Report no.7917*, March 2012, available at `http://hal.inria.fr/hal-00683827`.

[3] G. Bennett, "Probability inequalities for the sum of independent random variables", *Journal of the American Statistical Association* 57, pp. 33-45, 1962.

[4] F. Caro and O. S. Yoo, "Indexability of bandit problems with response delays," *Probability in the Engineering and Informational Sciences*, vol. 24, no. 3, pp. 349–374, 2010.

[5] G. Carofiglio, M. Gallo, and L. Muscariello, "ICP: Design and evaluation of an interest control protocol for content-centric networking", in Proceedings of IEEE INFOCOM Workshop on emerging design choices in name oriented networking, Orlando, USA, March 2012.

[6] R. Chiocchetti, D. Rossi, G. Rossini, G. Carofiglio, and D. Perino, "Exploit the known or explore the unknown? Hamlet-like doubts in ICN," in *Proceedings of the second edition of the ICN workshop on Information-centric networking*. ACM, 2012, pp. 7–12.

[7] S.G. Eick, "Gittins procedures for bandits with delayed responses", *Journal of the Royal Statistical Society. Series B (Methodological)*, v. 50(1), pp.125-132, 1988.

[8] J.C. Gittins, "Bandit processes and dynamic allocation indices", *Journal of the Royal Statistical Society, Series B*, v. 41(2), pp.148-177, 1979.

[9] M. Gritter and D.R. Cheriton, "An architecture for content routing support in the internet", in Proceedings of the USENIX Symposium on Internet Technologies and Systems, March 2001.

[10] W. Hoeffding, "Probability inequalities for sums of bounded random variables", *Journal of the American Statistical Association* 58, pp. 13-30, 1963.

[11] V. Jacobson, D. Smetters, J. Thornton, M. Plass, N. Briggs and R. Braynard, "Networking named content", in Proceedings of ACM CoNEXT 2009.

[12] T.L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules", *Advances in Applied Mathematics*, v.6(1), pp.4-22, 1985.

[13] T. Koponen, M. Chawla, B. Chun, A. Ermolinskiy, K. Kim, S. Shenker and I. Stoica, "A data-oriented (and beyond) network architecture", in Proceedings of ACM SIGCOMM 2007.

[14] D. Pollard, *Convergence of Stochastic Processes*, Springer-Verlag, 1984.

[15] N. Rozhnova and S. Fdida, "An effective hop-by-hop Interest shaping mechanism for CCN communications," in *in Proceedings of IEEE Conference on Computer Communications (INFOCOM) Workshops,*. IEEE, 2012, pp. 322–327.

[16] X. Wang and M. G. Bickis, "One-armed bandit models with continuous and delayed responses," *Mathematical Methods of Operations Research*, vol. 58, no. 2, pp. 209–219, 2003.

[17] R. Sutton and A. Barto, *Reinforcement learning: An Introduction*, MIT Press, 1998.