



**HAL**  
open science

## Multi-Player Bandits Revisited

Lilian Besson

► **To cite this version:**

Lilian Besson. Multi-Player Bandits Revisited. Séminaire “ IETR: Interagir Evaluer Transmettre Réunir ”, Jun 2018, Vannes, France. hal-02013847

**HAL Id: hal-02013847**

**<https://inria.hal.science/hal-02013847v1>**

Submitted on 11 Feb 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

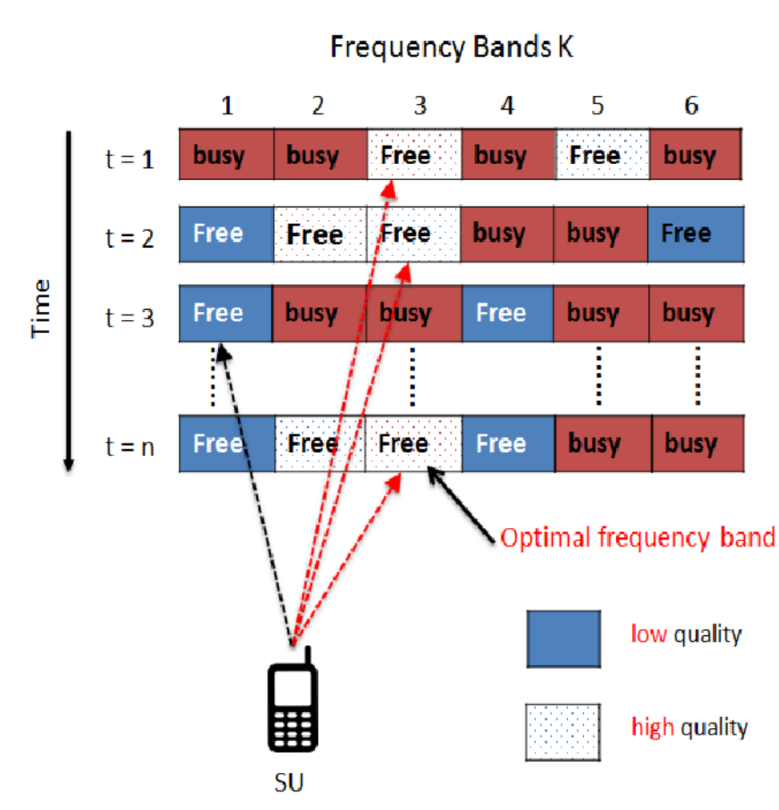
## 1. INTRODUCTION & GOAL

Goal: insert objects in a wireless network, keep a good *Quality of Service*.





- *Hypothesis*: object  $j$  choose channel  $A_j(t) \in \{1, \dots, K\}$ , to use to communicate at time  $t$ .
- *Idea*: use on-line **Machine Learning algorithms** ?
- *Not so easy*: each device takes its own decisions, without central control or communication, has light CPU/memory etc.
- $\implies$  **Solution: Decentralized MAB algorithms !**

## 2. MODEL: TIME/FREQUENCY PROTOCOL

## M PLAYERS IN THE NETWORK



Fix-duration communication. Channels can be free or busy.

- $K$  RF channels (of same bandwidth), e.g.,  $K = 9$ .
- Primary users  create a background traffic.
- Channel  $k$  is busy with mean  $\mu_1, \dots, \mu_K \in [0, 1]$ .
- Sensing gives binary feedback  $Y_{k,t} \sim \mathcal{B}(\mu_k)$ .
- $1 \leq M \leq K$  secondary users ,  $j \in \{1, \dots, M\}$ .
- Base station  replies if  packet is received.
- Collision indic.  $C^j(t) = \mathbb{1}(\exists j' \neq j, A^j(t) = A^{j'}(t))$ .
- Non-stochastic 0/1 reward  $r^j(t) := Y_{A^j(t),t} \times C^j(t)$ .

## 3. PREVIOUS WORKS (SINCE 2009)

Ideas: 1) use empirical means to learn the  $M$  best arms ( $\mu_1^*, \dots, \mu_M^*$ ), 2) and each user orthogonalize on this set.

- RhoRand: use an index policy (UCB) and random *ranks*, [Anandkumar et al, 2011]
- MEGA: use a simple  $\varepsilon$ -greedy algorithm (hard to tune, and not so efficient), [Avner & Mannor, 2015]
- Musical Chair: pure exploration then random hopping until convergence on  $M$  best arms, [Shamir et al, 2016]
- Some others: TDFS uses a time sharing, RhoLearn uses BayesUCB to learn the ranks, etc.

## 4. OUR PROPOSAL

- 1) kl-UCB  $>$  UCB<sub>1</sub> for best arms identification.
- 2) New algorithm MCTopM for orthogonalization:
  - Estimate the set of  $M$  best arm  $\widehat{M}^j(t)$ ,
  - Randomly play in  $\widehat{M}^j(t)$  until fixed on a good one (no collision: fixed on a “chair”).

## 5. UCB<sub>1</sub> AND kl-UCB INDEXES

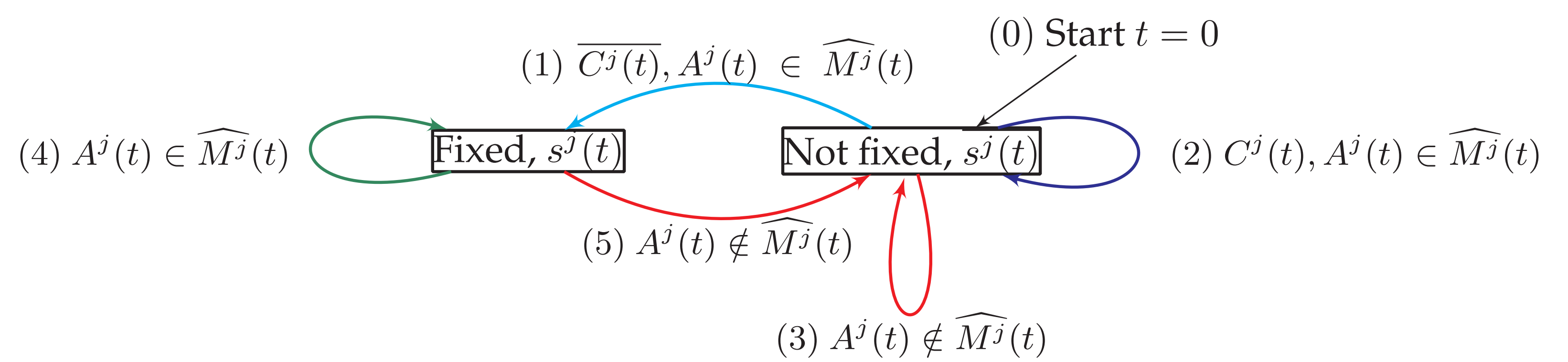
Let  $T_k^j(t)$  the selections of channel  $k$  for player  $j$ ,  $X_k^j(t)$  mean of free sensing.

$$\text{UCB}_1 : g_k^j(t) := \frac{X_k^j(t)}{T_k^j(t)} + \sqrt{\frac{\log(t)}{2T_k^j(t)}}$$

$$\text{kl-UCB} : g_k^j(t) := \sup_{q \in [0,1]} \left\{ q : \text{kl} \left( \frac{X_k^j(t)}{T_k^j(t)}, q \right) \leq \frac{\log(t)}{T_k^j(t)} \right\}$$

[Garivier & Cappé, 2011], [Cappé et al, 2013]

## 7. ILLUSTRATION OF MCTOPM



## 6. MCTOPM ALGORITHM

- 1  $A^j(1) \sim \mathcal{U}(\{1, \dots, K\})$ ,  $C^j(1) = s^j(1) = \text{False}$
- 2 **for**  $t = 0, \dots, T - 1$  **do**
- 3   **if**  $A^j(t) \notin \widehat{M}^j(t)$  **then**   // (3) or (5)
- 4      $A^j(t+1) \sim \mathcal{U}(\widehat{M}^j(t) \cap \{k : g_k^j(t-1) \leq g_{A^j(t)}^j(t-1)\})$
- 5      $s^j(t+1) = \text{False}$    // arm with smaller UCB at  $t-1$
- 6   **else if**  $C^j(t)$  and  $\overline{s^j(t)}$  **then**   // (2)
- 7      $A^j(t+1) \sim \mathcal{U}(\widehat{M}^j(t))$
- 8      $s^j(t+1) = \text{False}$
- 9   **else**   // transition (1) or (4)
- 10     $A^j(t+1) = A^j(t)$    // same arm
- 11     $s^j(t+1) = \text{True}$    // on “chair”
- 12 **end**
- 13 Play arm  $A^j(t+1)$ , get new observations,
- 14 Compute the indexes  $g_k^j(t+1)$ , and set  $\widehat{M}^j(t+1)$  for next step.
- 15 **end**

## 8. THEOREMS

Let  $T_k(T) := \sum_{j=1}^M T_k^j(T)$  total selections of arm  $k$ , and  $\mathcal{C}_k(T)$  counts collisions on arm  $k$ .

$$1. \text{ Regret: } R_T(\mu, M) := \mathbb{E}_\mu \left[ \sum_{t=1}^T \sum_{j=1}^M \mu_j^* - r^j(t) \right] = \left( \sum_{k=1}^M \mu_k^* \right) T - \mathbb{E}_\mu \left[ \sum_{t=1}^T \sum_{j=1}^M r^j(t) \right]$$

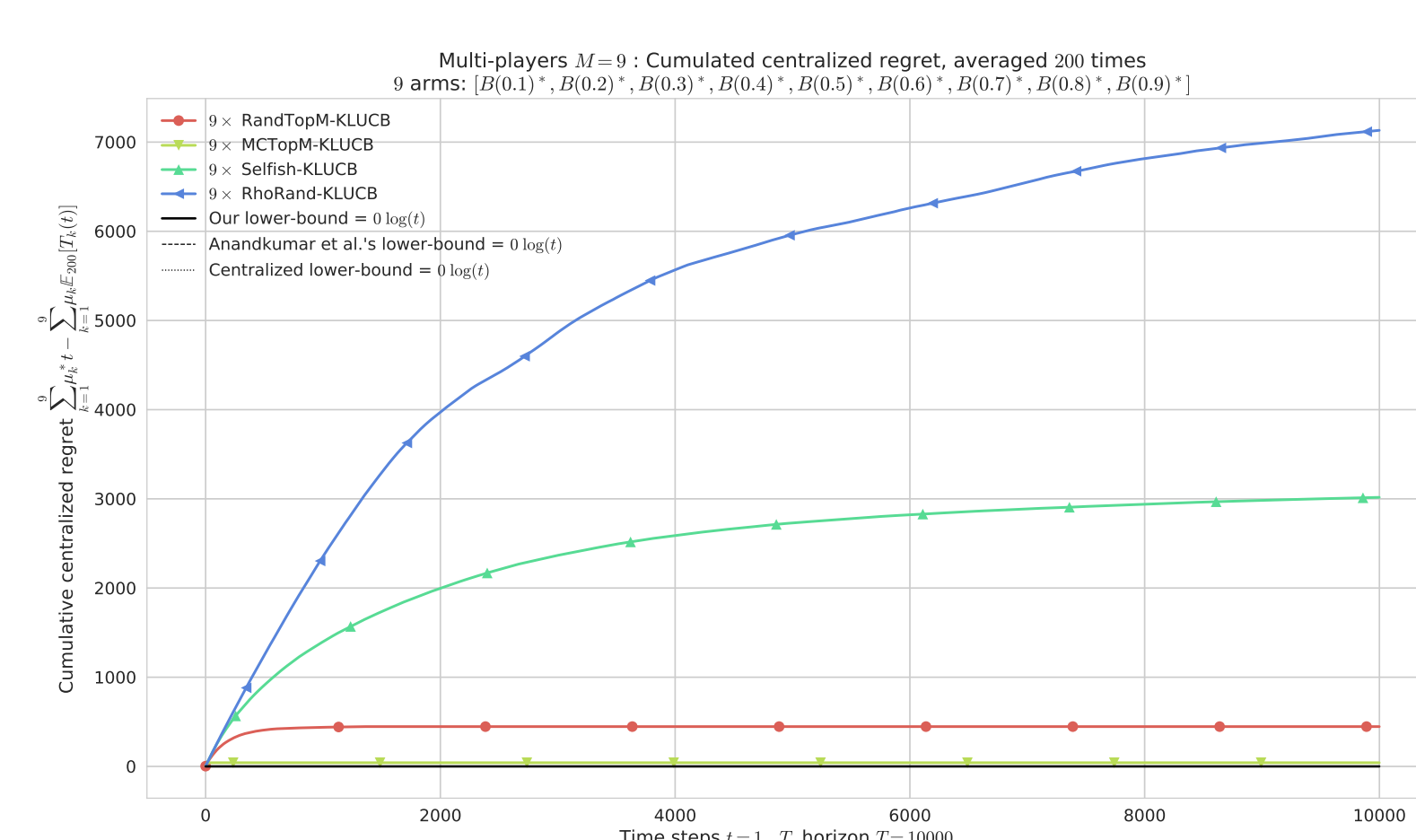
$$R_T = \sum_{k \in M\text{-worst}} (\mu_M^* - \mu_k) \mathbb{E}_\mu [T_k(T)] + \sum_{k \in M\text{-best}} (\mu_k - \mu_M^*) (T - \mathbb{E}_\mu [T_k(T)]) + \sum_{k=1}^K \mu_k \mathbb{E}_\mu [\mathcal{C}_k(T)].$$

2. Lemma for upper-bound: only two terms to focus on (bad selections and collisions).

$$\sum_{k \in M\text{-best}} (\mu_k - \mu_M^*) (T - \mathbb{E}_\mu [T_k(T)]) \leq (\mu_1^* - \mu_M^*) \left( \sum_{k \in M\text{-worst}} \mathbb{E}_\mu [T_k(T)] + \sum_{k \in M\text{-best}} \mathbb{E}_\mu [\mathcal{C}_k(T)] \right).$$

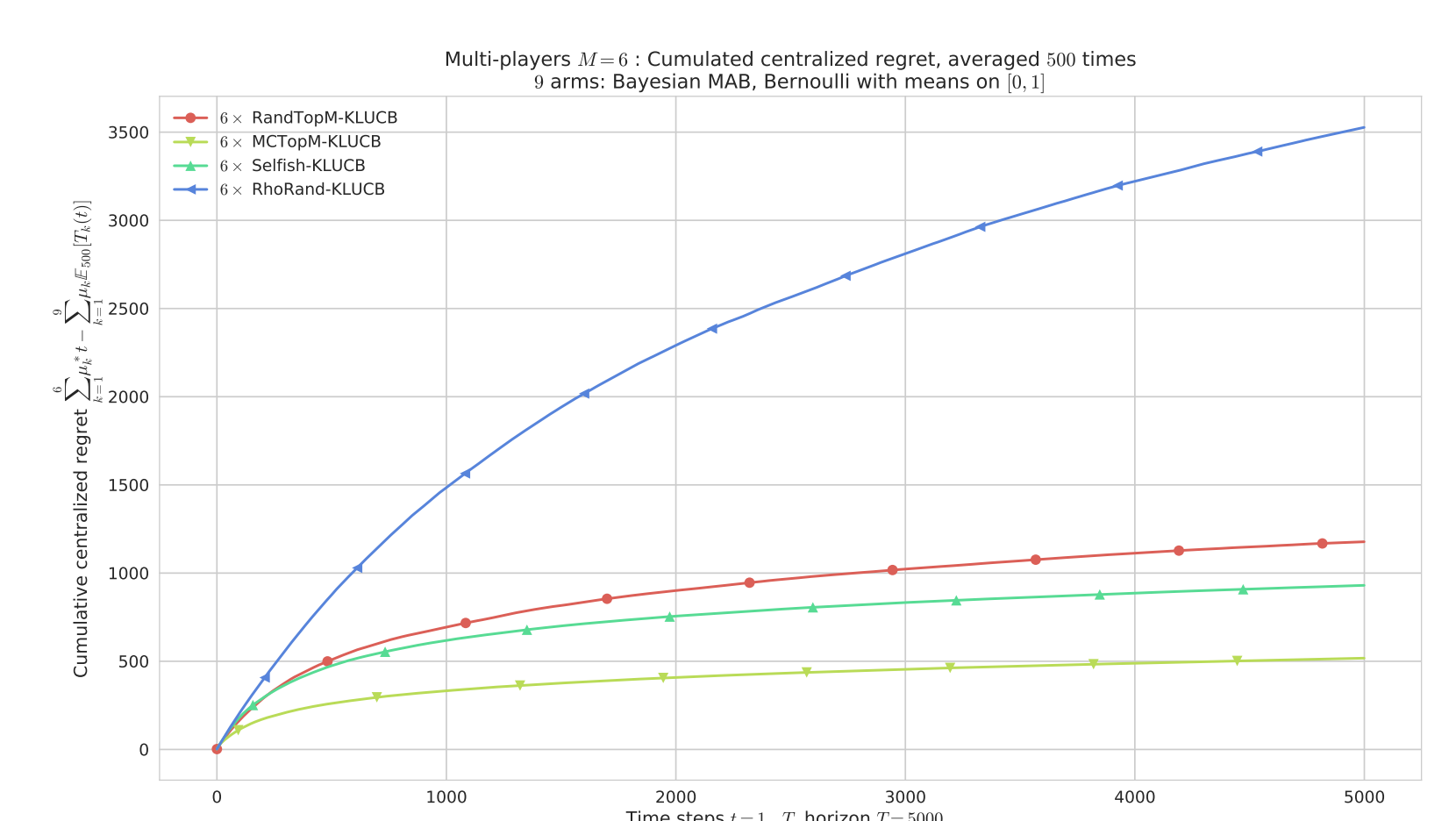
3. If all  $M$  players use MCTopM with kl-UCB,  $\forall \mu, \exists G_{M,\mu}, R_T(\mu, M) \leq G_{M,\mu} \times \log(T) + o(\log T)$ .

### 9.1. ILLUSTRATION FOR $M = K$



Only **RandTopM** and **MCTopM** achieve constant regret in this saturated case (proven).

### 9.2. ILLUSTRATION FOR $M < K$



**RhoRand** < **RandTopM** < **Selfish** < **MCTopM**.

## 10. CONCLUSIONS

- Our algorithm MCTopM is uniformly better than all previous proposals.
- *Good bounds*:  $\log(T)$  collisions, arm switches, bad selections and regret.
- Real-world implementation?  $\leftrightarrow$  Yes, presented at ICT 2018!
- *Future?* learn  $M$ , arrival/departures of users, dynamic problems, jammers etc.

## 11. MAIN REFERENCES

MORE ON-LINE  $\rightarrow$  <http://lbo.k.vu/JdD2018>

- [BBM<sup>+</sup>17] R. Bonnefoi, L. Besson, C. Moy, E. Kaufmann, and J. Palicot (2017). *Multi-Armed Bandit Learning in IoT Networks: Learning helps even in non-stationary settings*. In *12th EAI Conference on Cognitive Radio Oriented Wireless Network and Communication*.
- [BK18] L. Besson and E. Kaufmann (April 2018). *Multi-Player Bandits Revisited*. In *Algorithmic Learning Theory*. Lanzarote, Spain. URL <https://hal.inria.fr/hal-01629733>.
- [B18] Simulation code on [GitHub.com/SMPyBandits/SMPyBandits](https://github.com/SMPyBandits/SMPyBandits), open source (MIT license)!

## 12. THANKS TO ...

- Organizers of the *Workshop on MAB and Learning Algorithms!*
- ADDI association for the *PhD Students Day 2018!*
- SCEE team at IETR, CentraleSupélec (Rennes).
- SequeL team at Inria (Lille), and CNRS.
- My PhD advisors: Émilie Kaufmann, Christophe Moy.