A journey in decision making processes

Philippe Ciblat

Who I am ?

My very short resume : join S2A team in june 2023

- 1996 : Engineer degree from TP and M2 ATSI (UPSay)
- 2000 : PhD Thesis (*Blind frequency and channel estimation*)
- 2001 : Ass. Prof. at Digital Communications team in TP
- 2010 : Full Prof. and head of the team (until 2021)

Some editorial activities :

- BoG Gretsi, TPC Gretsi, Icassp, Eusipco, $\cdots$
- AE/SAE in IEEE TSP, IEEE TISPN, $\cdots$

Some teaching activities :

- COM105, Telecom track
- SI101, MD221, M2 MICAS
- TSE202, TSE101 (next year)

Other TSE-related activities :

- Member of GDR *Ecoinfo*, GDR *Internet et Société*

Outline

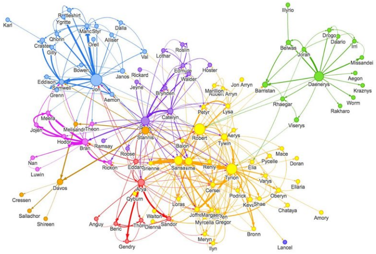
Two technical topics :

- Graph Node classification
  - No Graph Neural Network (GNN)
  - Interpretable algorithm
  - Less complex algorithm (with less hyperparameters)
    H. Hafidi et al., "Graph-assisted Bayesian node classifiers", IEEE Access, 2023
- Edge caching with popular time-sensitive contents
  - No neural network (while decision making agent)
  - Low-complex interpretable probabilistic approach
    H. Tang, et al., "Cache updating strategy minimizing AoI with time-varying popularity", IEEE Information Theory Workshop, 2021

Some perspectives :

- Wireless federative learning
- Graph based image classification
- Sustainable systems

Topic 1 : Graph Node classification



Idea : homophily principle

Predict class of each unlabeled node in the graph by relying

- on nodes' features and
- on nodes' graph connections

Examples :

- in social networks, people are more likely to connect with those who share the same areas of interest
- in research articles' database, more likely to have connections/citations between articles dealing with the same research topic

Problem statement

Classifier based on Bayesian decision theory : Maximum A Posteriori

- $\mathcal{V}_u$: set of nodes involved in the classification of node u.
- $\mathcal{X}_u = \{\boldsymbol{x}_u\} \cup \{\boldsymbol{x}_v, v \in \mathcal{V}_u\}$: set of features of node u and its "helping" nodes
- y_u : class of node u (what we are looking for !)
- D_k : probability density function of features belonging to class k. For any u,
$$D_k(\boldsymbol{x}_u) = p(\boldsymbol{x}_u | y_u = k).$$

Graph-Assisted Bayesian (GAB) Classifier

$$\hat{k}_u = \arg\max_k P_u(k)$$

with $P_u(k) = \Pr(y_u = k | \mathcal{X}_u, \mathcal{I}_{\mathcal{G}})$

Problem solution

Derivations of $P_u(k)$. Bayes' rule

$$P_u(k) = \frac{p(\mathcal{X}_u | y_u = k, \mathcal{I}_\mathcal{G}) \mathrm{Pr} P(y_u = k | \mathcal{I}_\mathcal{G})}{P(\mathcal{X}_u | \mathcal{I}_\mathcal{G})} \propto Q_u(k) \pi_k$$

with $\pi_k = \mathrm{Pr}(y_u = k | \mathcal{I}_\mathcal{G})$ a priori classes' probability
Let Δ_u be the diameter of the set $\mathcal{V}_u$.

$$Q_u(k) = D_k(\boldsymbol{x}_u) \prod_{d=1}^{\Delta_u} \prod_{v \in \mathcal{N}_u(d)} \left(\sum_{k'=1}^{K} r_{u,v}(k,k') D_{k'}(\boldsymbol{x}_v) \right)$$

with $r_{u,v}(k,k') = \mathrm{Pr}(y_v = k' | y_u = k, \mathcal{I}_\mathcal{G})$ the probability to be on class k' for node v given the fact that we are in class k for node u.

Example

$\mathcal{V}_u = \{v\}$, known $k_v = 1$, $\pi_1 = \pi_2 = 1/2$, and $\Delta_u = 1$:

$$Q_u(1) = D_1(\boldsymbol{x}_u) \frac{p}{p+q} \text{ and } Q_u(2) = D_2(\boldsymbol{x}_u) \frac{q}{p+q}$$

with p (resp. q) probability of intra (resp. inter)-class connection

Main result

Assumptions

- 2 equilikely classes
  - $p(k)$ probability that two nodes from class k are connected
    $\overline{p}_{\text{arithmetic}}$ arithmetic average of $\{p(k)\}_k$
  - q probability that two nodes from different classes are connected.
- Information on graph is 1-hop

We get

$$\begin{array}{c|c} r(1,2) = \frac{q}{p(1)+q} & r(2,2) = \frac{p(2)}{q+p(2)} \\ \hline r(1,1) = \frac{p(1)}{p(1)+q} & r(2,1) = \frac{q}{q+p(2)} \end{array}$$

Graph-agnostic iff $r(1,2) = r(2,2)$ and $r(1,1) = r(2,1)$

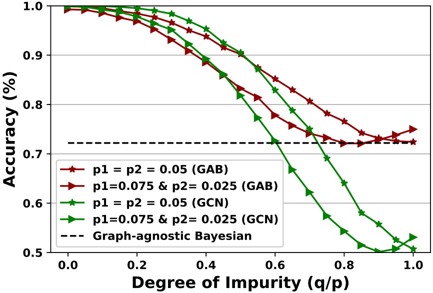
Main result

Graph-agnostic iff

- $q = \sqrt{p(1)p(2)} = \overline{p}_{\text{geometric}}$, or
- $\text{Degree of Impurity} = \frac{q}{\overline{p}_{\text{arithmetic}}} = \frac{\overline{p}_{\text{geometric}}}{\overline{p}_{\text{arithmetic}}} \leq 1$

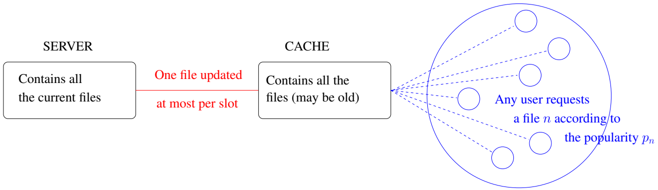
Numerical illustrations

- 2 classes
- Gaussian distributions with different means and covariance matrices
- Number of nodes $N = 5,000$ and number of features $F = 500$
- 500 (already-labeled) nodes



- GAB more robust to DoI than GCN
- GCN becomes worse than graph-agnostic (too confident)

Topic 2 : Edge caching



- Content n is time-sensitive ($X_n(t)$: age in caching)
- Content n has its own popularity (p_n : probability to be requested)
- Ex : newspaper website, web crawling, video last version, ...

Question

- Given a timeslot t, which item should be downloaded from the server to the cache to be as up-to-date as possible ?
- Scheduling problem

$$\{u_t\}_t = f(\text{information on the system})$$

Optimization problem

Optimization problem

$$\arg\min_{u_1,\cdots,u_T} \sum_{n=1}^{N} p_n \int_0^T X_n(t)\, dt$$

s.t. $u_t \in \{1,\cdots,N\}$ for all t, and $\sum_{t=1}^{T} \mathbf{1}\{u_t > 0\} = T$.

Approaches :

- As underlying Markov chain, constrained MDP well adapted
  - Optimal random policy exists
  - Suboptimal approach but simple : Whittle's index
- Probabilistic method by re-writing the problem

Extensions : i) when the transmission size depends on the age, ii) when the popularity is time-varying (satisfying Markov chain)

Approach 1 : concept of per-file update rate

- Consider λ_n the per-file update rate
- Actually, when T large enough,

$$\frac{1}{T}\int_0^T X_n(t)dt \approx \frac{1}{\lambda_n}$$

New optimization problem

$$\min_{\lambda_1,\dots,\lambda_N} \sum_{n=1}^{N} \frac{p_n}{\lambda_n}$$

s.t. $\lambda_n \geq 0$, and $\lambda_1 + \cdots + \lambda_N = 1$.

Main result

Problem is convex and leads to

$$\lambda_n^* = \frac{\sqrt{p_n}}{\sum_{m=1}^{N} \sqrt{p_m}}$$

Update rate of file n follows a square-root law wrt. its popularity

Practical protocol

Let $\tau_n^\star = 1/\lambda_n^\star$ be the optimal inter-update time for file n

$$u_t = \arg \max_{u\in\{1,\cdots,N\}} \underbrace{(X_u(t) - \tau_u^\star)}_{\text{Schedule-ordered by Age-based Priority (SOAP)}}$$

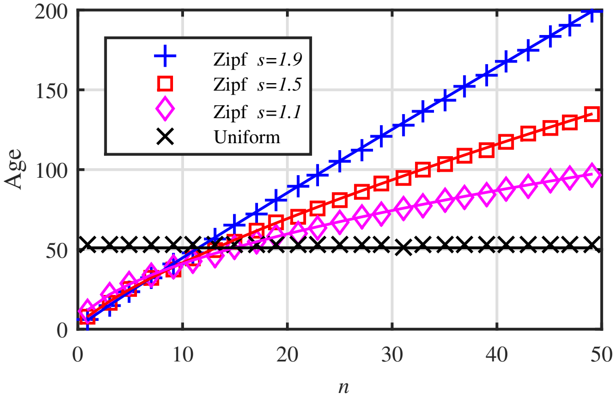
General context :

- $r(D, X)$: rank function with descriptor D and age X
- Scheduled user

$$u_t = \arg \max_{u\in\{1,\cdots,N\}} r(D_u(t), X_u(t))$$

- Many policies follow this shape
  - Round-Robin (RR), $r(\emptyset, X_u) = X_u$
  - "Weighted Round-Robin", $r(d_u, X_u) = d_u.X_u$
    Example : does it make sense to choose $d_u = \sqrt{p_u}$?

Numerical illustrations



Average AoI vs file index n (with $N = 50$)

Proposed policy reduces the average age of more popular items at the expense of less popular items

Approach 2 : index based policy

Find a suboptimal policy based on an index :

$$u_t = \arg \max_{u \in \{1, \cdots, N\}} \mathcal{I}_u(S_u)$$

- $\mathcal{I}$: it is an heuristic
- Whittle's index : methodology for exhibiting a reasonnable index in Restless Multi-Arm Bandit problem
  - N bandits/players/agents
  - At each timeslot, select one bandit (let's say u_t)
  - Its state s_{u_t} is modified according to its action, and is rewarded

  but states of other bandits also modified and rewarded in different ways (restless)
- When non-playing bandits are frozen (no state evolution) and not rewarded : Gittins' index is optimal

$$\mathcal{IG}_u(D) = \sup_{\tau > 0} \frac{\mathbb{E}[\sum_{t=0}^{\tau-1} \gamma^t r_u(S_t) | s_0 = S]}{\sum_{t=0}^{\tau-1} \gamma^t}$$

Whittle index (1/2)

$$\arg \max_{\{a_n(t)\}_{n,t}} \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t \sum_{n=1}^{N} r_n(s_n(t), a_n(t))\right]$$

s.t. $\sum_{n=1}^{N} a_n(t) = 1$ (C1) and $a_n(t) \in \{0, 1\}$ (C2)

Two modifications :

- Relaxation : C1 replaced with $\sum_{t=0}^{\infty} \gamma^t \sum_{n=1}^{N} a_n(t) = 1/(1-\gamma)$
- Lagrangian penalty

$$\arg \max_{\{a_n(t)\}_{n,t}} \mathcal{L}(\lambda)$$

s.t. C2 and with

$$\begin{aligned}
\mathcal{L}(\lambda) &= \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t \sum_{n=1}^{N} r_n(s_n(t), a_n(t))\right] - \lambda \left(\sum_{n=1}^{N} a_n(t) - 1/(1-\gamma)\right) \\
&= \lim_{T\to\infty} \mathbb{E}\left[\sum_{n=1}^{N} \left(\sum_{t=0}^{T-1} \gamma^t r_n(s_n(t), a_n(t)) - \lambda a_n(t)\right)\right]
\end{aligned}$$

Whittle index (2/2)

- The problem is now decoupled
- For each bandit and fixed λ, we maximize

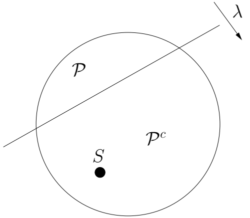
$$\mathcal{L}_n(\lambda) = \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t r_n(s_n(t), a_n(t)) - \lambda a_n(t)\right]$$

- $\mathcal{S}(\lambda)$: set of states to be idle obtained via $\mathcal{L}_n(\lambda)$
- **Optimal policy :** play ($a = 1$) if $s \in \mathcal{S}^c(\lambda)$ else idle ($a = 0$)
- **Indexability :** if idle for λ, then still idle for $\lambda' > \lambda$ (if higher penalty for being active, stay idle)

Definition

$$\mathcal{IW}(S) = \lambda^\star$$

s.t. $S \in \partial\mathcal{P}(\lambda^\star)$
Remark : if $\lambda < \lambda^\star$, $S \in \mathcal{P}(\lambda)$, else $S \in \mathcal{P}^c(\lambda)$



Closed-form expression for $\mathcal{IW}$

Main result `[unpublished]`

We have $S_u = X_u$, and

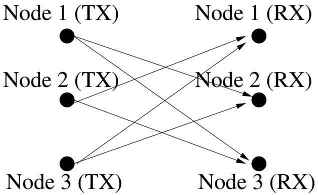
$$\mathcal{IW}_u(X_u) = \sqrt{p_u} X_u$$

- Is it close to square-root law ?
- Extension to 2-D state (age, popularity) when time-varying popularity
  - to be complete
  - condition for indexability
  - never done for any 2-D state contex

Perspective 1 : Wireless federative learning

- Learn a w-NN
- but database is split over K agents : $(x_k, y_k)_{k=1,\cdots,K}$

$$w^\star = \arg\min_w \sum_{k=1}^{K} f_k(w)$$



- Agents are wirelessly connected

Algorithm :

- Local gradient computation : $\nabla f_k(w_t)$
- Sharing gradient and update : $w_{t+1} \leftarrow w_t - \mu \sum_{k=1}^{K} \nabla f_k(w_t)$

Sharing step is a bottleneck !

- If no interference (baseline) : user rate $= \log_2(\mathrm{SNR})$
- If time-sharing : user rate $= \frac{1}{K}\log_2(\mathrm{SNR})$
- If our scheme : user rate $= \frac{K(K-1)-1}{K(2K-3)}\log_2(\mathrm{SNR}) \sim \frac{1}{2}\log_2(\mathrm{SNR})$
  More than **half the cake** for each agent ! (if $K = 3$, then $5/6$)

Perspective 1 : why does it work ?

When you precode well, you have more than the cake :

- Example : interference channel

$$y_n = x_n + i\left(\sum_{m\neq n} x_m\right) + z_n, \ \forall n$$

- If $x_n \in \mathbb{R}$, then $\Re\{y_n\} = x_n + \Re\{z_n\}, \ \forall n$
- User rate is then

$$\frac{1}{2}\log_2(1+\mathrm{SNR}) \sim \frac{1}{2}\log_2(\mathrm{SNR})$$

- Interference Alignement (IA) : put interference in a common subspace at RX side. The orthogonal is interference-free

An other application : distributed estimation in complete graph (rather than star graph and 1-bit time-sharing communication)

Collaboration with Shanghai Jiao Tong University (co-PhD degree)

Perspective 2 : Graph image classification

Representing an image (or set of patches) as a graph

Some examples :

- node $=$ pixel ; feature $=$ color level ; edge $=$ two pixels belonging to similar patches ; weight $=$ location distance within the image
- node $=$ patch ; feature $=$ patch learnt (strongly compressed) representation ; edge+weight $=$ related to original patch feature

Application : lung cancer

Collaboration with Universidad Nacional de Colombia (UNAL)

Perspective 3 : Efficiency

$$\textbf{Efficiency} = \frac{\textbf{metric of performance}}{\textbf{consumed energy}}$$

- OPEX-like energy : operational one
- CAPEX-like energy : embodied one
  - mining, manufacturing, recycling, $\cdots$: Life-Cycle Assessment

Example : Machine Learning

- Operational energy : computation energy during usage phase
- Embodied energy : Training, Computer's manufacturing, Cooling
- Metric of performance : Customer Satisfaction Rate

Main concerns :

- Open-data are missing for this kind of evaluation
- Depreciation duration, energy assignement (training/test)

Perspective 3 : or Sustainability ?

Sustainable (or resilient) system "*meets the needs of present generations without compromizing the ability of future generations to meet their own needs*" [Brundtland1987]

Implementation : given an application/usage, level of power is fixed

Why is it different from energy efficient system ?

- rebound effect has to be taken into account.
- if gain in energy consumption comes from enablement effect, customer behavior has to be predicted

Main concerns :

- Does not depend only on engineers' answers
- Required Science and Technology Studies (STS)

Perspective 3 : sustainable systems

Cars' traffic management :

- Given an area, amount of energy is limited per prefixed duration
- Speed limited to satisfy the energy constraint
- Avoid Stop-and-Go policy (consuming the whole budget once)
- Long-term policy is required to be smoother
  - at which spatial scale : road, county (but long-haul traffic ?), country
  - at which time scale : day, week, year
  - traffic prediction or adaptation ?
- Machine Learning is a relevant tool since highly-complex problem

Back to communication network :

- Given an area, amount of energy is limited per pre-fixed duration
- Packet traffic has to adapt
  - Quality of Service is moving
  - Outage is possible
- Here : available traffic model via stochastic geometry (ANR and PEPR grants in collaboration with *Infres*)