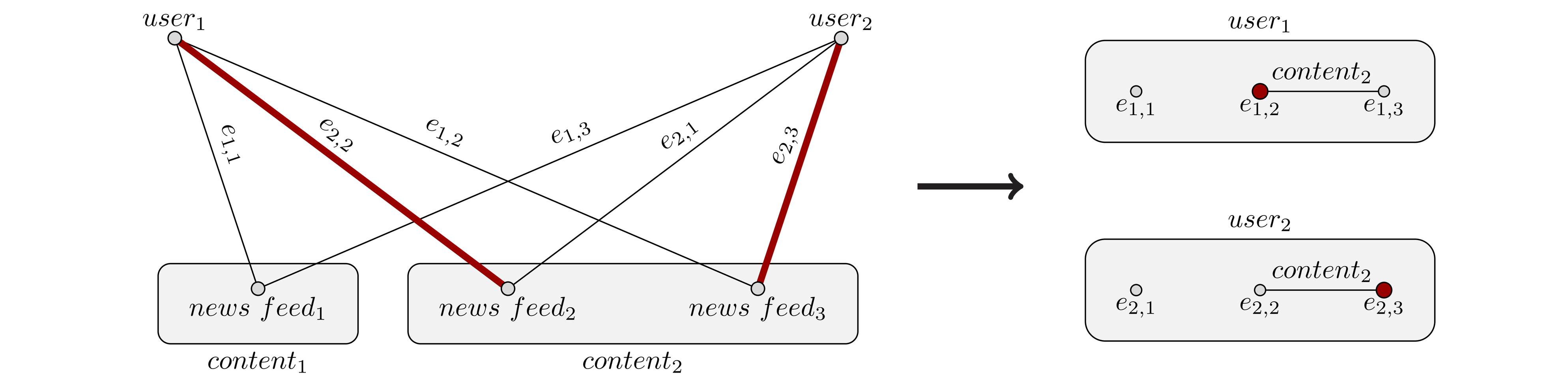
EFFICIENT LEARNING BY IMPLICIT EXPLORATION IN BANDIT PROBLEMS WITH SIDE OBSERVATIONS

TOMAS.KOCAK@INRIA.FR, GERGELY.NEU@INRIA.FR, MICHAL.VALKO@INRIA.FR, AND REMI.MUNOS@INRIA.FR

MOTIVATION - SEQUENTIAL NEWS RECOMMENDATION



- Select a matching covering users
- Obtain rewards of selected edges
- Observe additional rewards

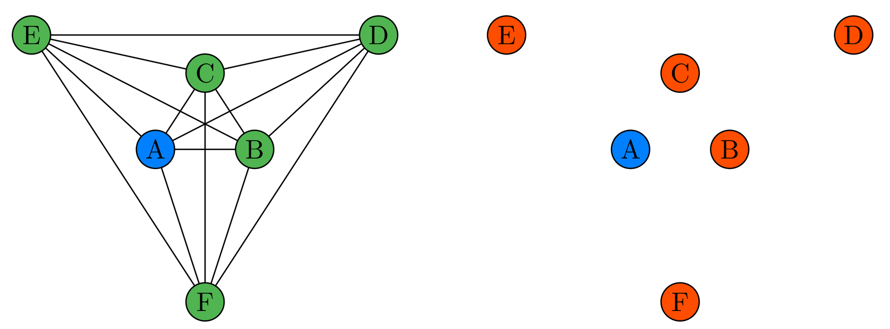
APPLICATIONS

- **Packet routing in computer networks**
  - **Typical feedback:** delays of our own packets
  - **Side observations:** other delays in network
- **Web advertising** – displaying one ad
  - Case 1: symmetric interrelations
    - **Example:** similar preferences for similar cars
    - **Typical feedback:** reward for displayed ad
    - **Side observations** for similar cars
    - **Model for observations:** undirected graph
  - Case 2: asymmetric interrelations
    - **Example:** electronics (interest in camera means interest in accessories, not vice versa)
    - **Typical feedback:** reward for displayed ad
    - **Side observations** for dependent products
    - **Model for observations:** directed graph

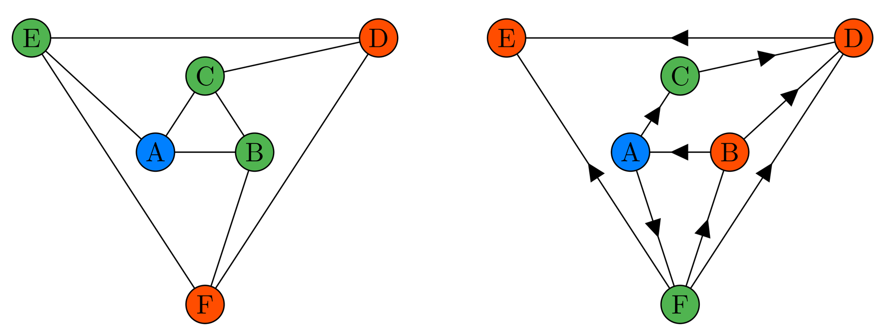
EXAMPLES OF GRAPH STRUCTURES

Side observations can be modeled as a graph

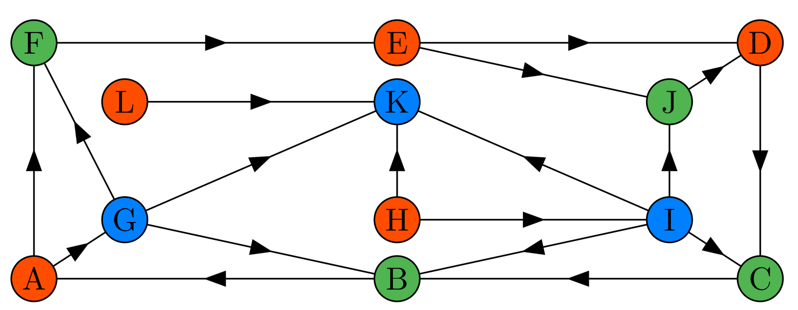
Full information and bandit setting – simple action



Undirected and directed case – simple action



Directed combinatorial case – complex action



LEARNING SETTING

In every round $t = 1, 2, \dots, T$:

- **Environment**
  - Privately assigns vector $\boldsymbol{\ell}_t$ of losses to actions
  - Generates an observation graph
    - Undirected / Directed
    - Disclosed / Not disclosed
- **Learner**
  - Plays action $\boldsymbol{V}_t \in \mathcal{S} \subseteq \{0,1\}^N$
    - Each action $\boldsymbol{v} \in \mathcal{S}$ satisfies $\|\boldsymbol{v}\|_1 \le m$
    - I.e. action consists of playing at most m nodes
    - Case $m = 1$: we denote $I_t \in [N]$ a node played
  - Obtain loss $\boldsymbol{V}_t^\intercal \boldsymbol{\ell}_t$ corresponding to nodes played
  - Observe losses of neighbors of played nodes
    - Graph disclosed

Performance measure: total expected regret

$$R_T = \max_{\boldsymbol{v} \in \mathcal{S}} \mathbb{E}\left[\sum_{t=1}^T (\boldsymbol{V}_t - \boldsymbol{v})^\intercal \boldsymbol{\ell}_t\right]$$

IMPLICIT EXPLORATION

Usual approach to exploration:

- Bias sampling distribution as $\tilde{\boldsymbol{p}}_t = (1-\gamma)\boldsymbol{p}_t + \gamma\boldsymbol{\mu}$
  - Needs to know graph structure
  - Constructing a good $\boldsymbol{\mu}$ is expensive
- Construct unbiased loss estimates

$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{o_{t,i}} \mathbb{1}\{\ell_{t,i} \text{ is observed}\}$$

- $o_{t,i}$ – probability of observing $\ell_{t,i}$

Our new approach:

- Do not touch sampling distribution
- Construct optimistically biased loss estimates

$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{o_{t,i} + \gamma} \mathbb{1}\{\ell_{t,i} \text{ is observed}\} \quad \text{s.t.} \quad \mathbb{E}\left[\hat{\ell}_{t,i}\right] \le \ell_{t,i}$$

- Encourages exploration by optimism
- Does not require knowledge of observation graph
- Cheaper than computing $\boldsymbol{\mu}$

EXP3-IX ALGORITHM

- **Compute weights** using loss estimates $\hat{\ell}_{t,i}$.

$$w_{t,i} = \exp\left(-\eta \sum_{s=1}^{t-1} \hat{\ell}_{s,i}\right)$$

- **Play action** I_t according to probability distribution

$$\mathbb{P}(I_t = i) = p_{t,i} = \frac{w_{t,i}}{W_t} = \frac{w_{t,i}}{\sum_{j=1}^N w_{t,j}}$$

- **Compute loss estimates** (using observability graph)

$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{o_{t,i} + \gamma} \mathbb{1}\{\ell_{t,i} \text{ is observed}\}$$

FPL-IX ALGORITHM

- **Draw perturbation** $Z_{t,i} \sim \text{Exp}(1)$ for all $i \in [N]$
- **Play "the best" action** $\boldsymbol{V}_t$ according to total loss estimate $\widehat{\boldsymbol{L}}_{t-1}$ and perturbation $\boldsymbol{Z}_t$

$$\boldsymbol{V}_t = \arg\min_{\boldsymbol{v} \in \mathcal{S}} \boldsymbol{v}^\intercal \left(\eta_t \widehat{\boldsymbol{L}}_{t-1} - \boldsymbol{Z}_t\right)$$

- **Compute loss estimates**

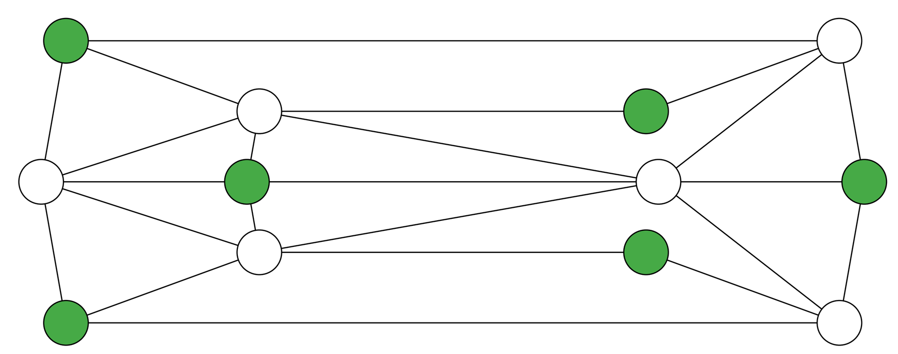
$$\hat{\ell}_{t,i} = \ell_{t,i} K_{t,i} \mathbb{1}\{\ell_{t,i} \text{ is observed}\}$$

- $K_{t,i}$: geometric random variable with

$$\mathbb{E}[K_{t,i}] = \frac{1}{o_{t,i} + (1 - o_{t,i})\gamma}$$

INDEPENDENCE SET

- Nodes of independence set are not connected
- α - size of the largest independence set



Independence set of size 6

MAIN RESULTS

Regret bound of Exp3-IX

$$R_T = \widetilde{\mathcal{O}}\left(\sqrt{\sum_{t=1}^T \alpha_t}\right) = \widetilde{\mathcal{O}}\left(\sqrt{\overline{\alpha} T}\right)$$

$\overline{\alpha}$ - average independence number of observation graph

Regret bound of FPL-IX

$$R_T = \widetilde{\mathcal{O}}\left(m^{3/2}\sqrt{\sum_{t=1}^T \alpha_t}\right) = \widetilde{\mathcal{O}}\left(m^{3/2}\sqrt{\overline{\alpha} T}\right)$$

RELATED WORK

- **Undirected case** – simple action ($m = 1$)
  - **ELP** (Mannor, Shamir)
    - Graph **disclosed before** action
    - Need to compute linear program for mixing
    - Regret bound of order $\widetilde{\mathcal{O}}(\sqrt{cT})$
  - **Exp3-Set** (Alon, Cesa-Bianchi, Gentile, Mansour)
    - Graph **disclosed after** action
    - Regret bound of order $\widetilde{\mathcal{O}}(\sqrt{\alpha T})$
- **Directed case** - simple action ($m = 1$)
  - **Exp3-Dom** (Alon, Cesa-Bianchi, Gentile, Mansour)
    - Graph **disclosed before** action
    - Need to find dominating set of graph
    - Regret bound of order $\widetilde{\mathcal{O}}(\sqrt{\alpha T})$
  - Exp3-IX
    - Graph **disclosed after** action
    - Computationally efficient
    - Regret bound of order $\widetilde{\mathcal{O}}(\sqrt{\alpha T})$
- **Directed case** - complex action ($m > 1$)
  - FPL-IX
    - Graph **disclosed after** action
    - Computationally efficient
    - Regret bound of order $\widetilde{\mathcal{O}}(m^{3/2}\sqrt{\alpha T})$

ANALYSIS

Analysis of Exp3 algorithms in general - tracking evolution of $\log(W_{t+1}/W_t)$

$$\underbrace{\mathbb{E}\left[\sum_{t=1}^T\sum_{i=1}^N p_{t,i}\hat{\ell}_{t,i}\right]}_{A} - \underbrace{\mathbb{E}\left[\sum_{t=1}^T \hat{\ell}_{t,k}\right]}_{B} \le \mathbb{E}\left[\frac{\log N}{\eta}\right] + \underbrace{\mathbb{E}\left[\frac{\eta}{2}\sum_{t=1}^T\sum_{i=1}^N p_{t,i}(\hat{\ell}_{t,i})^2\right]}_{C}$$

Lower bound of A (using definition of loss estimates)

$$\mathbb{E}\left[\sum_{t=1}^T\sum_{i=1}^N p_{t,i}\hat{\ell}_{t,i}\right] \ge \mathbb{E}\left[\sum_{t=1}^T\sum_{i=1}^N p_{t,i}\ell_{t,i}\right] - \mathbb{E}\left[\gamma\sum_{t=1}^T Q_t\right]$$

Lower bound of B (optimistic loss estimates: $\mathbb{E}[\hat{\ell}] < \mathbb{E}[\ell]$)

$$-\mathbb{E}\left[\sum_{t=1}^T \hat{\ell}_{t,k}\right] \ge -\mathbb{E}\left[\sum_{t=1}^T \ell_{t,k}\right]$$

Upper bound of C (using definition of loss estimates)

$$\mathbb{E}\left[\frac{\eta}{2}\sum_{t=1}^T\sum_{i=1}^N p_{t,i}(\hat{\ell}_{t,i})^2\right] \le \mathbb{E}\left[\frac{\eta}{2}\sum_{t=1}^T Q_t\right]$$

Together we have

$$R_T \le \frac{\log N}{\eta} + \left(\frac{\eta}{2} + \gamma\right)\sum_{t=1}^T \mathbb{E}[Q_t]$$

$$Q_t = \sum_{i=1}^N \frac{p_{t,i}}{o_{t,i} + \gamma}$$

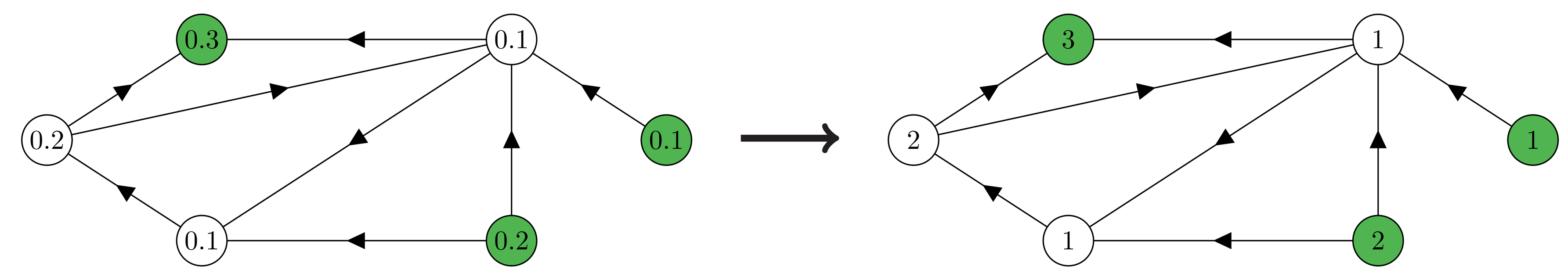
Lemma 1. *Let G be a directed graph, with $V = \{1, \dots, N\}$. Let d_i^- be the indegree of the node i and $\alpha = \alpha(G)$ be the independence number of G. Then*

$$\sum_{i=1}^N \frac{1}{1 + d_i^-} \le 2\alpha \log\left(1 + \frac{N}{\alpha}\right).$$

Step 1 of applying Lemma 1 to upper bound Q_t - **Discretization**

$$Q_t = \sum_{i=1}^N \frac{p_{t,i}}{o_{t,i} + \gamma} = \sum_{i=1}^N \frac{p_{t,i}}{p_{t,i} + \sum_{j \in N_i^-} p_{t,j} + \gamma} \le \sum_{i=1}^N \frac{\hat{p}_{t,i}}{\hat{p}_{t,i} + \sum_{j \in N_i^-} \hat{p}_{t,j}} + 2 \quad \text{for } M = \lceil N^2/\gamma \rceil$$

Step 2 of applying Lemma 1 to upper bound Q_t - **Construction of a "clique graph"**



$$\sum_{i=1}^N \frac{\hat{p}_{t,i}}{\hat{p}_{t,i} + \sum_{j \in N_i^-} \hat{p}_{t,j}} = \sum_{i=1}^N \frac{M\hat{p}_{t,i}}{M\hat{p}_{t,i} + \sum_{j \in N_i^-} M\hat{p}_{t,j}} = \sum_{i=1}^N \sum_{k \in C_i} \frac{1}{1 + d_k^-} \le 2\alpha \log\left(1 + \frac{M + N}{\alpha}\right)$$