Bandit networks

Charles Auguste & Louis Trezzini

January 24, 2019

Goal of the project

Study the reinforcement learning framework of 2 articles :

- **DBLP:journals/corr/DaiKZDS17**
- **DBLP:journals/corr/BelloPLNB16**

Multi-agent stochastic multi-armed bandit (MAB) problem

Conclusion

- The reinforcement learning methods presented here are not problem-specific

Thank you

Any Questions ?

Additional slide : Pointer Networks

Issue with the FYL policy

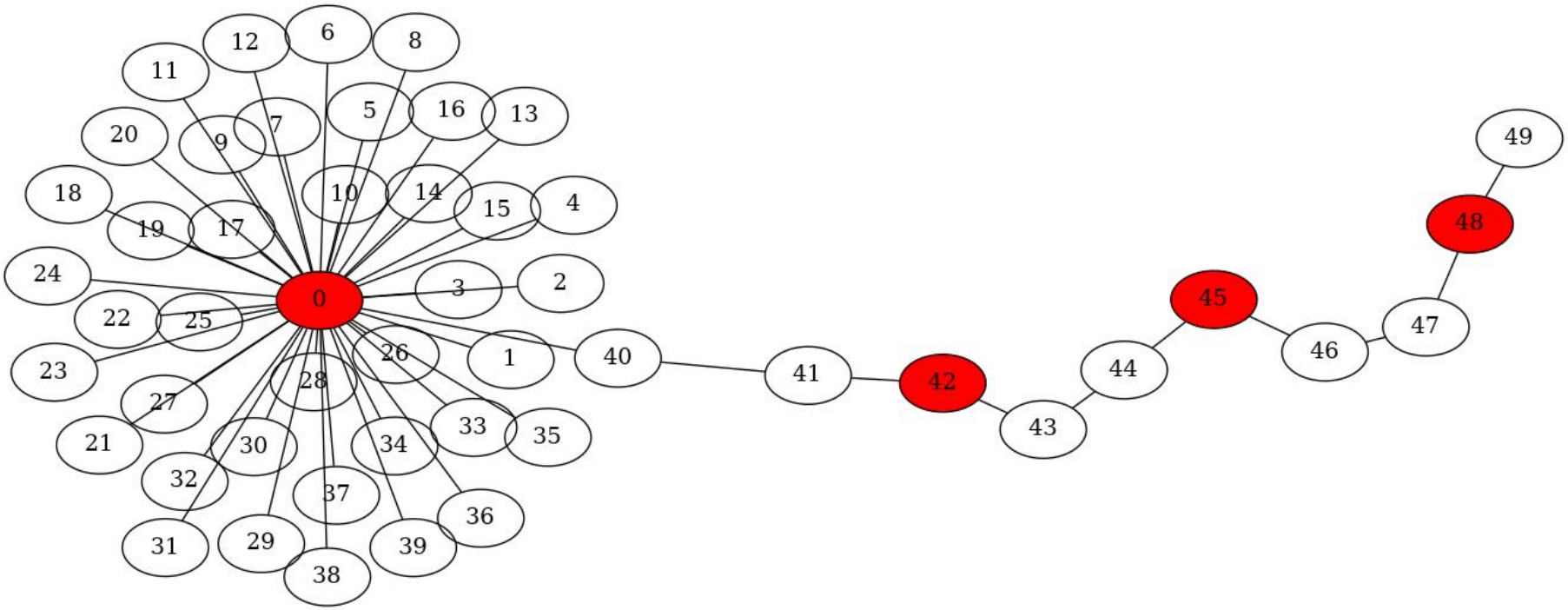

Figure 1: Star-chain graph, with optimal dominating set in red

Nodes 41-49 are missing on a lot of information !

Follow Best Informed (FBI) policy

- FYL policy is myopic
- In addition to their previous action, nodes can output the number of samples (information) they used to compute it
- Nodes can follow their best informed neighbor and use UCB-policy if they are better informed
- Actually, the structure of a graph fully determines the behavior of the nodes (but not their precise actions obviously)

Example of the usefulness of the FBI policy

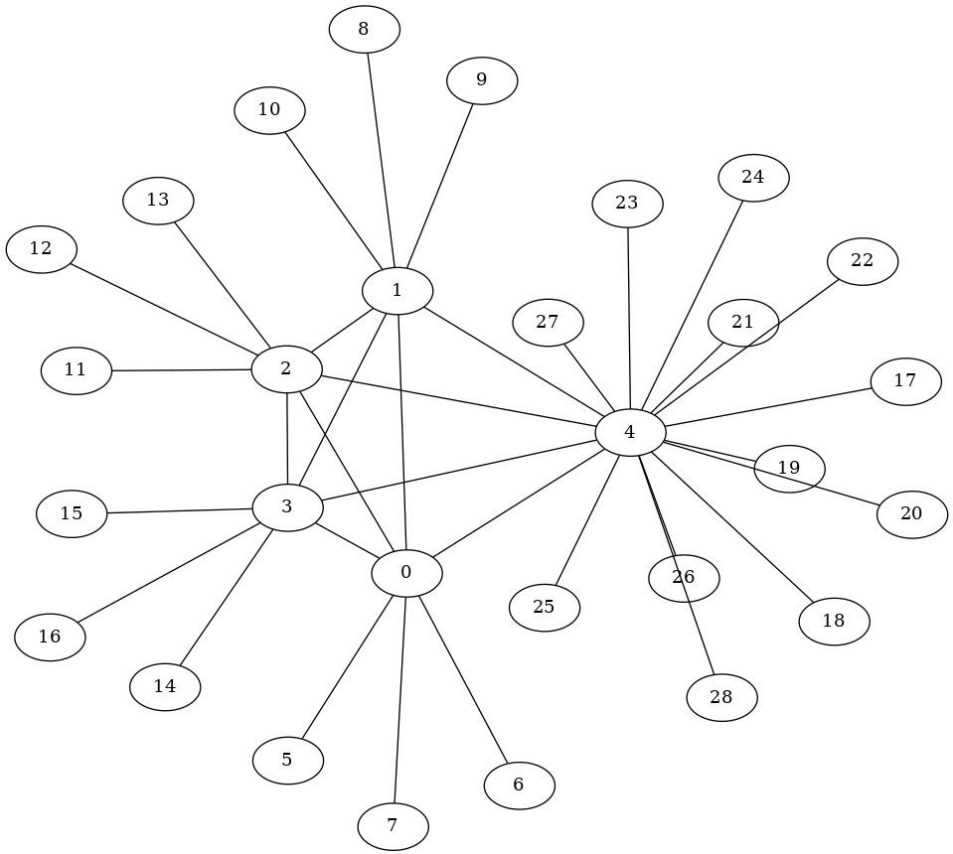

Figure 2: Fully connected stars graph

After first iteration, node 4 has the most information. It can pass it to nodes 0-3, who will then pass it to their children.

Results for a fully connected stars graph

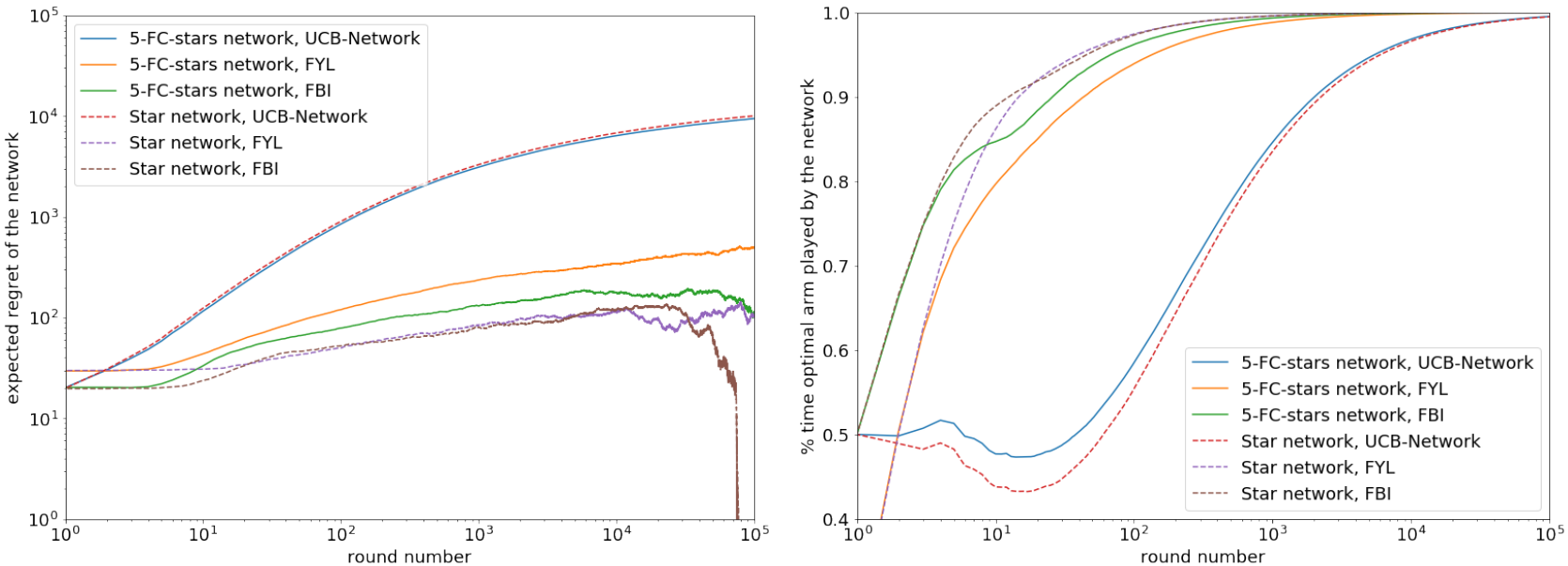

Figure 3: Performance comparison of UCB-Network, FYL, and FBI policies on a 100-nodes star network and on the 100-nodes 5-FC-stars network: 2 arms, Bernoulli rewards with means 0.5 and 0.7 (1000 sample paths).

Results star-chain graph

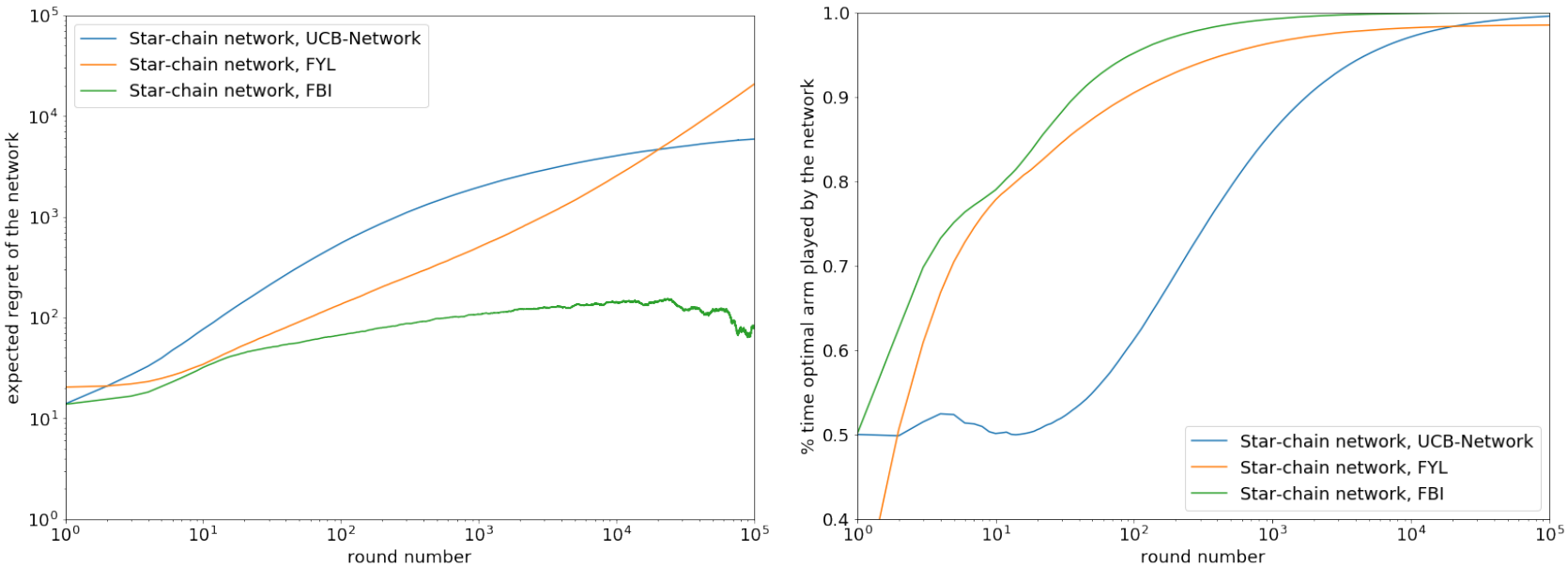

Figure 4: Performance comparison of FYL and FBI policies on the pathological graph structure (star graph with 70 nodes, among which a 20-nodes long chain): 2 arms, Bernoulli rewards with means 0.5 and 0.7 (1000 sample paths).

A deeper look at the FBI policy

- **Downside** : If one node has more information than the rest, every node it going to follow it (at a delayed rate) $\Rightarrow$ Strong correlation in the nodes actions

- **Further improvements** : When a node has multiple neighbors informed about in the same way, it may be smart to randomly follow one with a probability depending on its amount of information. But then the behavior of the nodes is not determined by the structure of the graph...