A Pseudocode

Algorithm 1: `Dec_UCB`

Input: $\mathbb{G}, T, C_{i,k}(t)$

Initialization Each agent samples each arm exactly once. Initialize $z_{i,k}(0) = \bar{x}_{i,k}(0) = X_{i,k}(0)$, $m_{i,k}(0) = n_{i,k}(0) = 1$, and $C_{i,k}(0) = 0$

for $t = 0, \ldots, T$ **do**
  $\mathcal{A}_i = \emptyset$
  if $n_{i,k}(t) \leq m_{i,k}(t) - M$ **then**
    Agent i puts k into a set $\mathcal{A}_i$ *# exploration consistency requirements*
  end
  if $\mathcal{A}_i = \emptyset$ **then**
    for $k = 1, \ldots, M$ **do**
      $Q_{i,k}(t+1) = z_{i,k}(t) + C_{i,k}(t)$ *# belief update*
    end
    $a_i(t+1) = \arg\max_k Q_{i,k}(t+1)$ *# optimal arm in belief*
  else
    $a_i(t+1)$ is randomly chosen from $\mathcal{A}_i$
  end
  Agent i sends $m_{i,k}(t)$ and $z_{i,k}(t)$ to each agent j satisfying $i \in \mathcal{N}_j$ *# information transmission*
  Agent i receives $m_{j,k}(t), z_{j,k}(t)$ from each neighbor $j \in \mathcal{N}_i$
  $n_{i,k}(t+1) = n_{i,k}(t), \ \forall k \in [M]$ *# information updating*
  $n_{i,a_i(t+1)}(t+1) = n_{i,a_i(t+1)}(t) + 1$
  $m_{i,k}(t+1) = \max\{n_{i,k}(t+1), m_{j,k}(t), j \in \mathcal{N}_i\}$
  $z_{i,k}(t+1) = \sum_{j=1}^{N} w_{ij} z_{j,k}(t) + \bar{x}_{i,k}(t+1) - \bar{x}_{i,k}(t)$
end

B Analysis and Proofs

In this appendix, we provide the analysis of `Dec_UCB` and proofs of Theorems 1 and 2.

We begin with some basic properties of sub-Gaussian random variables in B.1, and provide analysis on the exploration "consistency" of `Dec_UCB` in B.2, which theoretically validates Remark 2. Based on the properties in B.1 and results in B.2, we prove Theorems 1 and 2 in B.3 and B.4, respectively.

B.1 Sub-Gaussian Random Variables

A random variable X with $\mathbf{E}[X] = \mu$ is called σ^2 sub-Gaussian if there is a positive σ such that

$$\mathbf{E}(e^{\lambda(X-\mu)}) \leq e^{\frac{\sigma^2\lambda^2}{2}}, \quad \forall \lambda \in \mathbb{R}.$$

Such σ^2 is called a variance proxy, and the smallest variance proxy is called the optimal variance proxy. Sub-Gaussian random variables have the following three properties.

Lemma 1. *Let X be any σ^2 sub-Gaussian random variable $\mathbf{E}[X] = \mu$. Then, for any $a \geq 0$,*

$$\mathbf{P}(X - \mu \geq a) \leq e^{-\frac{a^2}{2\sigma^2}}, \qquad \mathbf{P}(\mu - X \geq a) \leq e^{-\frac{a^2}{2\sigma^2}}.$$

Proof: The proof can be found in Section 5.3 in [1]. ■

Lemma 2. *Let $X_1, \ldots, X_n$ be n independent random variables such that X_i is σ_i^2 sub-Gaussian random variable. Then, $X_1 + \cdots + X_n$ is $(\sigma_1^2 + \cdots + \sigma_n^2)$ sub-Gaussian.*

Proof: The proof can be found in Section 5.3 in [1]. ■

Lemma 3. *If a random variable X has a finite mean and $a \leq X \leq b$ almost surely, then X is $\frac{1}{4}(b-a)^2$ sub-Gaussian.*

Proof: The lemma is a direct consequence of Hoeffding's Lemma in [2]. ■

B.2 Exploration Consistency of `Dec_UCB`

In this subsection, we will show that after a finite number of pulls, for each agent i and each arm k, there holds $n_{i,k}(t) \leq 2\min_{j\in[N]} n_{j,k}(t)$; see Lemma 6. An immediate consequence of this property is that $\max_{j\in[N]} n_{j,k}(t) \leq 2\min_{j\in[N]} n_{j,k}(t)$; in other words, the global maximal number of pulls of each arm is no larger than twice the corresponding global minimum. This ensures that the explorations of each arm among all the agents are consistent in the sense that no agent is behind too much in arm exploration.

To proceed, we will make use of the concept of distance in a graph. For a strongly connected graph, the distance from vertex i to another vertex j is the length of the shortest directed path from i to j. The definition subsumes the special case of undirected graphs. For an undirected, connected graph, the distance between two different vertices is the length of the shortest path connecting them. We use the notation $d_{i,j}$ to denote the distance from vertex i to vertex j in a graph $\mathbb{G}$, regardless of it being directed or undirected. It is natural to define $d_{i,i} = 0$ for any vertex i, and easy to see that $d_{i,j} \leq d$, where d is the diameter of $\mathbb{G}$.

It is worth emphasizing that the following three lemmas hold for both a directed, strongly connected neighbor graph $\mathbb{G}$ and an unidirected, connected $\mathbb{G}$, using the same proofs.

For the purpose of analysis, we define $n_{i,k}(t) = m_{i,k}(t) = 0$ for all $i \in [N]$ and $k \in [M]$ whenever $t < 0$.

Lemma 4. *For any $i \in [N]$ and $k \in [M]$,*

$$m_{i,k}(t) = \max_{j\in[N]} \{n_{j,k}(t - d_{j,i})\}. \tag{9}$$

Proof: We will prove the lemma by induction on t. For the basis step, suppose that $t = 0$. In this case, $m_{i,k}(1) = \max\{n_{i,k}(0),\ m_{j,k}(0),\ j \in \mathcal{N}_i\} = 1$. Note that $\max_{j\in[N]}\{n_{j,k}(t - d_{j,i})\} = n_{i,k}(0) = 1$. Thus, (9) holds when $t = 0$.

For the inductive step, assume (9) holds at time t, and now consider time $t+1$. Note that

$$\begin{aligned} m_{i,k}(t+1) &= \max\{n_{i,k}(t+1),\ m_{j,k}(t),\ j \in \mathcal{N}_i\} \\ &= \max\{n_{i,k}(t+1),\ n_{h,k}(t - d_{j,h}),\ h \in [N],\ j \in \mathcal{N}_i\}. \end{aligned}$$

It is easy to see that $d_{h,i} \leq d_{j,i} + d_{h,j} = 1 + d_{h,j}$. Since $n_{i,k}(t)$ is a non-decreasing function of t by its definition,

$$m_{i,k}(t+1) \leq \max_{h\in[N]}\{n_{i,k}(t+1),\ n_{h,k}(t - d_{h,i} + 1)\} = \max_{j\in[N]}\{n_{j,k}(t - d_{j,i} + 1)\}. \tag{10}$$

Fix any vertex $j \in [N]$ and let $p = (j, v_{d_{j,i}}, \ldots, v_2, i)$ be a shortest path from j to i in $\mathbb{G}$. From (2),

$$\begin{aligned} m_{i,k}(t+1) &\geq m_{v_2,k}(t) \geq \cdots \geq m_{v_{d_{j,i}},k}(t - d_{j,i} + 2) \\ &\geq m_{j,k}(t - d_{j,i} + 1) \geq n_{j,k}(t - d_{j,i} + 1). \end{aligned} \tag{11}$$

Since j is arbitrarily chosen from $[N]$, we have $m_{i,k}(t+1) \geq \max_{j\in[N]}\{n_{j,k}(t - d_{j,i} + 1)\}$. Combining with (10), we have

$$m_{i,k}(t+1) = \max_{j\in[N]}\{n_{j,k}(t - d_{j,i} + 1)\}.$$

Thus, (9) also holds at $t+1$, which completes the induction. ∎

Lemma 5. *For any $i \in [N]$ and $k \in [M]$,*

$$n_{i,k}(t) > m_{i,k}(t) - M(M + 2d).$$

Proof: We will prove the lemma by contradiction. Suppose that, to the contrary, $\exists\ i, k_1$ such that $n_{i,k_1}(t) \leq m_{i,k_1}(t) - M(M + 2d)$. Let t' denote the first time at which the equality holds, i.e.,

$$n_{i,k_1}(t') = m_{i,k_1}(t') - M(M + 2d).$$

Here t' must exist, since when $t = 0$, we have $n_{i,k}(0) > m_{i,k}(0) - M(M + 2d)$, since both $n_{i,k}(t)$ and $m_{i,k}(t)$ increase by 0 and 1 at each time instance, if there exists some t such that

$n_{i,k_1}(t) < m_{i,k_1}(t) - M(M+2d)$, there must exist a t' between 0 and t, such that $n_{i,k_1}(t') = m_{i,k_1}(t') - M(M+2d)$. According to Lemma 4, $\exists\, j \in [N]$ such that

$$m_{i,k_1}(t') = n_{j,k_1}(t' - d_{j,i}). \tag{12}$$

Then,

$$n_{j,k_1}(t' - d_{j,i}) - n_{i,k_1}(t') = M(M+2d). \tag{13}$$

Since according to Lemma 4, $m_{i,k_1}(t) \geq n_{j,k_1}(t - d_{j,i})$ always holds, so for $t < t'$, we have

$$n_{j,k_1}(t - d_{i,j}) - n_{i,k_1}(t) \leq m_{i,k_1}(t) - n_{i,k_1}(t) < M(M+2d). \tag{14}$$

Since $n_{i,k}(t)$ is non-decreasing for all $i \in [N], k \in [M]$, (13) and (14) imply that $n_{j,k_1}(t' - d_{i,j}) > n_{j,k_1}(t' - d_{i,j} - 1)$. This further implies that at time $t' - d_{i,j}$, agent j pulls arm k_1.

Since each agent must pull an arm at each time, we have $\sum_k n_{i,k}(t) = t,\ \forall i \in [N]$. Then,

$$\sum_{k \in [M] \setminus k_1} n_{i,k}(t') - \sum_{k \in [M] \setminus k_1} n_{j,k}(t' - d_{j,i}) = M(M+2d) + d_{j,i}.$$

Applying the Pigeonhole principle, $\exists\, k_2 \in [M]$ such that

$$n_{i,k_2}(t') - n_{j,k_2}(t' - d_{j,i}) \geq \frac{M(M+2d)}{M-1} > M + 2d.$$

According to the definition of $n_{i,k}(t)$, it is non-decreasing and $n_{i,k}(t+1) \leq n_{i,k}(t) + 1$, we obtain

$$n_{i,k_2}(t') = n_{i,k_2}(t' - d_{j,i} - d_{i,j} + d_{j,i} + d_{i,j}) \leq n_{i,k_2}(t' - d_{j,i} - d_{i,j}) + d_{j,i} + d_{i,j}.$$

Thus,

$$\begin{aligned} n_{i,k_2}(t' - d_{j,i} - d_{i,j}) - n_{j,k_2}(t' - d_{j,i}) &> n_{i,k_2}(t') - n_{j,k_2}(t' - d_{j,i}) - d_{j,i} - d_{i,j} \\ &> M + 2d - d_{j,i} - d_{i,j} > M. \end{aligned}$$

Using (11), we have $m_{j,k_2}(t' - d_{j,i}) \geq n_{i,k_2}(t' - d_{j,i} - d_{i,j})$. Thus,

$$m_{j,k_2}(t' - d_{j,i}) - n_{j,k_2}(t' - d_{j,i}) > M.$$

From the above analysis, agent j must pull arm k_1 at time $t' - d_{j,i}$. According to the decision making step of the algorithm, there holds

$$m_{j,k_1}(t' - d_{j,i}) - n_{j,k_1}(t' - d_{j,i}) \geq M > 0. \tag{15}$$

Note that from (11),

$$m_{i,k_1}(t') \geq m_{j,k_1}(t' - d_{j,i}). \tag{16}$$

Combining (12) – (16) together, we have

$$n_{j,k_1}(t' - d_{j,i}) = m_{i,k_1}(t') \geq m_{j,k_1}(t' - d_{j,i}) > n_{j,k_1}(t' - d_{j,i}),$$

which is a contradiction. Therefore, the statement of the lemma is true. ∎

Lemma 6. *For any $i \in [N]$ and $k \in [M]$, if $n_{i,k}(t) \geq 2(M^2 + 2Md + d)$, then*

$$n_{i,k}(t) \leq 2 \min_{j \in [N]} n_{j,k}(t).$$

Proof: From (11), $m_{j,k}(t) \geq n_{i,k}(t - d_{i,j}),\ \forall i \in [N]$. Combining with $n_{j,k}(t+1) \leq n_{j,k}(t) + 1$, we have

$$m_{j,k}(t) \geq n_{i,k}(t) - d_{i,j} \geq n_{i,k}(t) - d.$$

From Lemma 5, we have

$$n_{j,k}(t) \geq n_{i,k}(t) - (M^2 + 2Md + d),\ \forall\, j, i \in [N].$$

Since j is arbitrarily chosen, when $n_{i,k}(t) \geq 2(M^2 + 2Md + d)$,

$$\min_{j \in [N]} n_{j,k}(t) \geq n_{i,k}(t) - (M^2 + 2Md + d) \geq n_{i,k}(t) - \frac{1}{2} n_{i,k}(t) = \frac{1}{2} n_{i,k}(t),$$

which completes the proof. ∎

B.3 Strongly Connected Graphs

To prove the performance of `Dec_UCB` on strongly connected graphs as stated in Theorem 1, we need to provide a tight bound of the variance proxy of $z_{i,k}(t)$ in this case.

Lemma 7. *For any $i \in [N]$, $k \in [M]$ and time $t \geq 0$, with W being defined in* (5)*, $z_{i,k}(t)$ is a sub-Gaussian random variable, and when*

$$n_{i,k}(t) \geq \max\left\{L,\ 2(M^2 + 2Md + d)\right\},$$

the optimal variance proxy of $z_{i,k}(t)$ is no larger than $\frac{1}{3n_{i,k}(t)}$.

Proof. According to Lemma 3, $z_{i,k}(t)$ is a sub-Gaussian random variable as it is bounded. From (1) and (3), we know that $z_{i,k}(t)$ is a linear combination of $X_{j,k}(\tau)$, for all $j \in [N], \tau \in \{1, 2, \ldots, t\}$. According to Lemma 3 and Lemma 2, in order to find the variance proxy of $z_{i,k}(t)$, we need to estimate the coordinates of such $X_{j,k}(\tau)$ first.

Note that from (4),

$$\begin{aligned}
z_k(t) &= W z_k(t-1) + \bar{x}_k(t) - \bar{x}_k(t-1) \\
&= W^t z_k(0) + \sum_{\tau=0}^{t-1} W^\tau (\bar{x}_k(t-\tau) - \bar{x}_k(t-\tau-1)) \\
&= \sum_{\tau=0}^{t-1} (W^{t-\tau} - W^{t-\tau-1}) \bar{x}_k(\tau) + \bar{x}_k(t).
\end{aligned}$$

Thus,

$$z_{i,k}(t) = \sum_j \left\{ \sum_{\tau=0}^{t-1} [W^{t-\tau} - W^{t-\tau-1}]_{ij} \bar{x}_{j,k}(\tau) + [W^0]_{ij} \bar{x}_{j,k}(t) \right\}.$$

Denote $\tau_{i,1}, \tau_{i,2}, \ldots, \tau_{i,n_{i,k}(t)}$ as the ascending sequence of all time instances before time t at which agent i pulls arm k. From the initialization step of the algorithm, it is clear that $\tau_{i,1} = 0$. According to (1), if $\tau_{i,m} \leq \tau < \tau_{i,m+1}$, we have $\bar{x}_{i,k}(\tau) = \bar{x}_{i,k}(\tau_{i,m}),\ \forall i \in [N]$. Then,

$$z_{i,k}(t) = \sum_j \left\{ \sum_{h=1}^{n_{j,k}(t)-1} \left[W^{t-\tau_{j,h}} - W^{t-\tau_{j,h+1}}\right]_{ij} \bar{x}_{j,k}(\tau_{j,h}) + \left[W^{t-\tau_{j,n_{j,k}(t)}}\right]_{ij} \bar{x}_{j,k}(\tau_{j,n_{j,k}(t)}) \right\}, \tag{17}$$

Let $c_{i,k,j}^{(\tau)}(t)$ be the coefficient of $X_{j,k}(\tau)$ in $z_{i,k}(t)$, it is not hard to see from the above equation that when $\tau \neq \tau_{j,1}, \tau_{j,2}, \ldots, \tau_{j,n_{j,k}(t)}$, we have $c_{i,k,j}^{(\tau)}(t) = 0$. Also, when $\tau = \tau_{j,1}, \tau_{j,2}, \ldots, \tau_{j,n_{j,k}(t)}$, then $X_{j,k}(\tau) \cdot \mathbb{1}(a_j(\tau) = k) = X_{j,k}(\tau)$ are i.i.d. random variables. Thus, from Lemma 2 and Lemma 3,

$$\sigma_{i,k}^2 \triangleq \frac{1}{4} \sum_{j=1}^{N} \sum_{h=1}^{n_{j,k}(t)} \left| c_{i,k,j}^{(\tau_{j,h})}(t) \right|^2$$

is a variance proxy of $z_{i,k}(t)$. And we have

$$c_{i,k,j}^{(0)}(t) = \left[\sum_{h=1}^{n_{j,k}(t)-1} \frac{W^{t-\tau_{j,h}} - W^{t-\tau_{j,h+1}}}{h} + \frac{W^{t-\tau_{j,n_{j,k}(t)}}}{n_{j,k}(t)} \right]_{ij} \tag{18}$$

which holds for all $i \in [N], k \in [M]$. The equation also can be written as

$$c_{i,k,j}^{(0)}(t) = \left[W^t - \sum_{h=2}^{n_{j,k}(t)} \frac{W^{t-\tau_{j,h}}}{(h-1)h} \right]_{ij}. \tag{19}$$

From (6), $c^{(0)}_{i,k,j}(t)$ satisfies

$$
\begin{aligned}
|c^{(0)}_{i,k,j}(t)| \leq & [W_\infty]_{ij}\left(1 - \sum_{h=2}^{n_{j,k}(t)} \frac{1}{(h-1)h}\right) + c\rho_2^t + c\sum_{h=2}^{n_{j,k}(t)} \frac{\rho_2^{t-\tau_{j,h}}}{(h-1)h} \\
= & \frac{[W_\infty]_{ij}}{n_{j,k}(t)} + c\rho_2^t + \sum_{h=2}^{n_{j,k}(t)} \frac{c\rho_2^{t-\tau_{j,h}}}{(h-1)h}.
\end{aligned}
$$

Since $0 < \rho_2 < 1$, the smaller $t - \tau_{j,h}$ is, the larger the right side of the inequality would be, so $\rho_2^{t-\tau_{j,h}} \leq \rho_2^{n_{j,k}(t)-h}$. Since

$$
\begin{aligned}
& \sum_{h=2}^{n_{j,k}(t)} \frac{\rho_2^{n_{j,k}(t)-h}}{(h-1)h} \\
= & \sum_{h=2}^{\frac{12N\lceil c\rceil}{12N\lceil c\rceil+1} n_{j,k}(t)} \frac{\rho_2^{n_{j,k}(t)-h}}{(h-1)h} + \sum_{\frac{12N\lceil c\rceil}{12N\lceil c\rceil+1} n_{j,k}(t)+1}^{n_{j,k}(t)} \frac{\rho_2^{n_{j,k}(t)-h}}{(h-1)h} \\
= & \rho_2^{n_{j,k}(t)-2} + (1-\rho_2) \sum_{h=2}^{\frac{12N\lceil c\rceil}{12N\lceil c\rceil+1} n_{j,k}(t)} \frac{\rho_2^{n_{j,k}(t)-h-1}}{h} + \sum_{\frac{12N\lceil c\rceil}{12N\lceil c\rceil+1} n_{j,k}(t)+1}^{n_{j,k}(t)} \frac{\rho_2^{n_{j,k}(t)-h}}{(h-1)h} \\
\leq & \rho_2^{n_{j,k}(t)-2} + (1-\rho_2) \sum_{h=2}^{\frac{12N\lceil c\rceil}{12N\lceil c\rceil+1} n_{j,k}(t)} \rho_2^{n_{j,k}(t)-h-1} + \sum_{\frac{12N\lceil c\rceil}{12N\lceil c\rceil+1} n_{j,k}(t)+1}^{n_{j,k}(t)} \frac{1}{(h-1)h} \\
\leq & \rho_2^{n_{j,k}(t)-2} + \rho_2^{\frac{n_{j,k}(t)}{12N\lceil c\rceil+1}-1} + \frac{1}{12N\lceil c\rceil n_{j,k}(t)},
\end{aligned}
$$

we obtain that

$$
\begin{aligned}
|c^{(0)}_{i,k,j}(t)| \leq & \frac{[W_\infty]_{ij}}{n_{j,k}(t)} + c\left(\rho_2^t + \rho_2^{n_{j,k}(t)-2} + \rho_2^{\frac{n_{j,k}(t)}{12N\lceil c\rceil+1}-1} + \frac{1}{12N\lceil c\rceil n_{j,k}(t)}\right) \\
\leq & \frac{[W_\infty]_{ij}}{n_{j,k}(t)} + 3c\rho_2^{\frac{n_{j,k}(t)}{12N\lceil c\rceil+1}-1} + \frac{c}{12N\lceil c\rceil n_{j,k}(t)}.
\end{aligned}
$$

Recall that L is the smallest value such that when $t \geq L$, we have $3\rho_2^{t/12N\lceil c\rceil} \leq \frac{\rho_2}{24N\lceil c\rceil t}$. Since $c \leq \lceil c\rceil$, when $n_{j,k}(t) \geq L$, we have

$$
|c^{(0)}_{i,k,j}(t)| \leq \frac{[W_\infty]_{ij}}{n_{j,k}(t)} + \frac{1}{8Nn_{j,k}(t)}.
$$

Note that the expression in $[\cdot]$ of (18) is a summation of $n_{j,k}(t)$ terms. From the derivation of (18) and the definition of $c^{(\tau)}_{i,k,j}(t)$, it is straightforward to verify that for each $l \in \{2, \ldots, n_{j,k}(t)\}$,

$$
c^{(\tau_{j,l})}_{i,k,j} = \left[\sum_{h=l}^{n_{j,k}(t)-1} \frac{W^{t-\tau_{j,h}} - W^{t-\tau_{j,h+1}}}{h} + \frac{W^{t-\tau_{j,n_{j,k}(t)}}}{n_{j,k}(t)}\right]_{ij},
$$

in which $[\cdot]$ is the summation of the last $n_{j,k}(t) - l + 1$ terms in $[\cdot]$ of (18). Then, following the same steps as above, we can conclude that for all $l \in \{2, \ldots, n_{j,k}(t)\}$, there holds

$$
\left|c_{i,k,j}(t)^{(\tau_{j,l})}\right| \leq \frac{[W_\infty]_{ij}}{n_{j,k}(t)} + \frac{1}{8Nn_{j,k}(t)}
$$

when $n_{j,k}(t) \geq L$. Using Lemma 6, when $n_{j,k}(t) \geq \max\{L, 2(M^2 + 2Md + d)\}$, we have

$$\begin{aligned}
\sigma_{i,k}^2(t) &= \frac{1}{4} \sum_{j=1}^{N} \sum_{h=1}^{n_{j,k}(t)} \left| c_{i,k,j}^{(\tau_{j,h})} \right|^2 \\
&\leq \sum_{j=1}^{N} \frac{[W_\infty]_{ij}^2 + \frac{1}{4N} \cdot [W_\infty]_{ij} + \frac{1}{64N^2}}{4n_{j,k}(t)} \\
&\leq \frac{1}{2n_{i,k}(t)} \left(\sum_{j=1}^{N} [W_\infty]_{ij}^2 + \frac{1}{4N} \sum_{j=1}^{N} [W_\infty]_{ij} + \frac{1}{64N} \right).
\end{aligned} \tag{20}$$

Since W is stochastic, so is W_∞. Then,

$$\sum_{j=1}^{N} [W_\infty]_{ij} = 1, \tag{21}$$

and $W_\infty \cdot \mathbf{1} = \mathbf{1}$. Let $a^\top$ be the "normalized" dominant left eigenvector of W, i.e., $a^\top \cdot \mathbf{1} = 1$ and $a^\top \cdot W = a^\top$. Then, $a^\top \cdot W_\infty = a^\top$ and moreover $W_\infty = \mathbf{1} \cdot a^\top$ [3]. Since W is an irreducible nonnegative matrix, by the Perron-Frobenius Theorem, each entry of a is positive. With these facts, we have

$$\begin{aligned}
\sum_{j=1}^{N} [W_\infty]_{ij}^2 &= a^\top a = a^\top W a = \mathrm{tr}(a^\top W a) = \mathrm{tr}(W \cdot a a^\top) \qquad (\mathrm{tr}(AB) = \mathrm{tr}(BA)) \\
&\leq \sum_{i=1}^{N} \frac{\sum_{j=1}^{N} a_i a_j}{|\mathcal{N}_i|} \qquad (a \textit{ is positive}) \\
&= \sum_{i=1}^{N} \frac{a_i}{|\mathcal{N}_i|} \qquad (\textit{the sum of all entries of } a \textit{ equals 1}) \\
&\leq \sum_{i=1}^{N} \frac{a_i}{2} \qquad (|\mathcal{N}_i| \geq 2) \\
&= \frac{1}{2},
\end{aligned} \tag{22}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a square matrix. Substituting (21) and (22) in (20), we have

$$\sigma_{i,k}^2(t) \leq \frac{1}{2n_{i,k}(t)} \cdot \left(\frac{1}{2} + \frac{1}{4N} + \frac{1}{64N} \right) < \frac{1}{3n_{i,k}(t)},$$

which completes the proof. ∎

We are now in a position to prove Theorem 1.

Proof of Theorem 1: According to Lemma 1 and Lemma 7, when

$$n_{i,k}(t) \geq \max\left\{L, 2(M^2 + 2Md + d)\right\},$$

we have

$$\mathbf{P}\left(z_{i,k}(t) - \mu_k \geq \sqrt{\frac{4 \log t}{3n_{i,k}(t)}} \right) \leq \exp\left(-\frac{2 \log t}{3n_{i,k}(t) \sigma_{i,k}^2(t)} \right) \leq \frac{1}{t^2}.$$

Similarly,

$$\mathbf{P}\left(\mu_k - z_{i,k}(t) \geq \sqrt{\frac{4 \log t}{3n_{i,k}(t)}} \right) \leq \frac{1}{t^2}.$$

Now let us go back to the algorithm and set

$$C_{i,k}(t) = \sqrt{\frac{4 \log t}{3 n_{i,k}(t)}}.$$

The Decision Making step of `Dec_UCB` ensures that agent i chooses an arm $k \neq 1$, instead of the optimal arm 1, at time t only if one of the following four cases occurs:

Case 1: $n_{i,k}(t) \leq m_{i,k}(t) - M$;

Case 2: $z_{i,k}(t) - \mu_k \geq C_{i,k}(t)$;

Case 3: $\mu_1 - z_{i,1}(t) \geq C_{i,1}(t)$;

Case 4: $\mu_1 - \mu_k < 2C_{i,k}(t)$.

We are then prepared to find a bound for $\mathbf{E}(n_{i,k}(T))$. First, it is easy to verify that when

$$n_{i,k}(t) \geq \frac{16}{3\Delta_k^2} \log T,$$

Case 4 does not hold. To proceed, we define t' as the first time instance, if any, that satisfies

$$n_{i,k}(t') = \max\left\{ \frac{16}{3\Delta_k^2} \log T,\ L,\ 2(M^2 + 2Md + d) \right\}.$$

In the case when there does not exist such a $t' \leq T$, it immediately follows that

$$n_{i,k}(T) < n_{i,k}(t') = \max\left\{ \frac{16}{3\Delta_k^2} \log T,\ L,\ 2(M^2 + 2Md + d) \right\}.$$

Next consider the case when t' exists and $t' \leq T$. Then,

$$\sum_{t > t'} \left[\mathbf{P}\Big(z_{i,k}(t) - \mu_k \geq C_{i,k}(t) \Big) + \mathbf{P}\Big(\mu_1 - z_{i,1}(t) \geq C_{i,1}(t) \Big) \right] \leq \sum_{t > t'} \frac{2}{t^2} = \frac{\pi^2}{3},$$

which implies that after t', the expected number of pulls of agent i on arm k due to Case 2 and Case 3 is no more than $\frac{\pi^2}{3}$. Since the difference between $n_{i,k}(t)$ and $m_{i,k}(t)$ is at most $M^2 + 2Md$ by Lemma 5 and $n_{i,k}(t)$ keeps increasing until the difference is less than M according to the Decision Making step of `Dec_UCB`, the expected number of pulls due to Case 1 must be no larger than $\frac{\pi^2}{3} + M^2 + (2d-1)M$. Thus,

$$\begin{aligned}
\mathbf{E}(n_{i,k}(T)) &\leq \mathbf{E}\left(n_{i,k}(T) \mid T \geq t' \right) \\
&= n_{i,k}(t') + \frac{2\pi^2}{3} + M^2 + (2d-1)M \\
&= \max\left\{ \frac{16}{3\Delta_k^2} \log T,\ L,\ 2(M^2 + 2Md + d) \right\} + \frac{2\pi^2}{3} + M^2 + (2d-1)M.
\end{aligned}$$

Now we can get an upper bound of agent i's regret as follows:

$$\begin{aligned}
R_i(T) &= T\mu_1 - \sum_{t=1}^{T} \mathbf{E}(X_{a_i(t)}) \\
&= \sum_{k:\Delta_k > 0} \mathbf{E}(n_{i,k}(T)) \cdot \Delta_k \\
&\leq \sum_{k:\Delta_k > 0} \left(\max\left\{ \frac{16}{3\Delta_k^2} \log T,\ L,\ 2(M^2 + 2Md + d) \right\} + \frac{2\pi^2}{3} + M^2 + (2d-1)M \right) \Delta_k.
\end{aligned}$$

This completes the proof. ■

B.4 Undirected and Connected Graphs

In this subsection, we prove the performance of `Dec_UCB` on undirected and connected graphs as stated in Theorem 2. The procedure is the same as that in B.3. We first provide a tight bound of the variance proxy of $z_{i,k}(t)$.

Lemma 8. *For any $i \in [N]$, $k \in [M]$, and time t, with W being defined in* (8)*, $z_{i,k}(t)$ is a sub-Gaussian random variable, and when*

$$n_{i,k}(t) \geq \max\left\{L,\ 2(M^2 + 2Md + d)\right\},$$

the optimal variance proxy of $z_{i,k}(t)$ is no larger than $\frac{3}{4|\mathcal{N}_i|n_{i,k}(t)}$.

Proof. The arguments in the proof of Lemma 7 until (20) still hold here, that is, when $n_{j,k}(t) \geq \max\{L, 2(M^2 + 2Md + d)\}$, we still have

$$\sigma_{i,k}^2(t) \leq \frac{1}{2n_{i,k}(t)}\left(\sum_{j=1}^{N}[W_\infty]_{ij}^2 + \frac{1}{4N}\sum_{j=1}^{N}[W_\infty]_{ij} + \frac{1}{64N}\right),$$

where now W is a doubly stochastic matrix defined by (8). Since W is irreducible and doubly stochastic, its dominant left and right eigenvectors of eigenvalue 1 are $\mathbf{1}^\top$ and $\mathbf{1}$, respectively, and thus $W_\infty = \frac{1}{N}\mathbf{1}\mathbf{1}^\top$ [3]. Thus,

$$\sigma_{i,k}^2(t) \leq \frac{1}{2n_{i,k}(t)}\left(\frac{1}{N} + \frac{1}{4N} + \frac{1}{64N}\right) \leq \frac{3}{4Nn_{i,k}(t)} \leq \frac{3}{4|\mathcal{N}_i|n_{i,k}(t)},$$

which proves the lemma. ∎

Proof of Theorem 2: According to Lemma 1 and Lemma 8, when

$$n_{i,k}(t) \geq \max\left\{L,\ 2(M^2 + 2Md + d)\right\},$$

we have

$$\mathbf{P}\left(z_{i,k}(t) - \mu_k \geq \sqrt{\frac{3\log t}{|\mathcal{N}_i|n_{i,k}(t)}}\right) \leq \exp\left(-\frac{3\log t}{2|\mathcal{N}_i|n_{i,k}(t)\sigma_{i,k}^2(t)}\right) \leq \frac{1}{t^2}.$$

Similarly,

$$\mathbf{P}\left(\mu_k - z_{i,k}(t) \geq \sqrt{\frac{3\log t}{|\mathcal{N}_i|n_{i,k}(t)}}\right) \leq \frac{1}{t^2}.$$

Again, the Decision Making step of `Dec_UCB` guarantees that agent i chooses an arm $k \neq 1$ instead of the optimal arm 1 at time t only if one of the four cases listed in the proof of Theorem 1 occurs. First, it is easy to verify that when

$$n_{i,k}(t) \geq \frac{12}{|\mathcal{N}_i|\Delta_k^2}\log T,$$

Case 4 does not hold. Then, let t' be the first time instance, if any, that satisfies

$$n_{i,k}(t') = \max\left\{\frac{12}{|\mathcal{N}_i|\Delta_k^2}\log T,\ L,\ 2(M^2 + 2Md + d)\right\}.$$

In the case when there does not exist such a $t' \leq T$, it immediately follows that

$$n_{i,k}(T) < n_{i,k}(t') = \max\left\{\frac{12}{|\mathcal{N}_i|\Delta_k^2}\log T,\ L,\ 2(M^2 + 2Md + d)\right\}.$$

In the other case when t' exists and $t' \leq T$, using the same arguments as in the proof of Theorem 1, the expected number of pulls due to Case 1 is no larger than $\frac{\pi^2}{3} + M^2 + (2d-1)M$. Then,

$$\begin{aligned}
\mathbf{E}(n_{i,k}(T)) &\leq \mathbf{E}\left(n_{i,k}(T) \mid T \geq t'\right) \\
&= n_{i,k}(t') + \frac{2\pi^2}{3} + M^2 + (2d-1)M \\
&= \max\left\{\frac{12}{|\mathcal{N}_i|\Delta_k^2}\log T, L, 2(M^2 + 2Md + d)\right\} + \frac{2\pi^2}{3} + M^2 + (2d-1)M.
\end{aligned}$$

We thus get the following upper bound of agent i's regret:

$$\begin{aligned}R_i(T) &= T\mu_1 - \sum_{t=1}^{T}\mathbf{E}(X_{a_i(t)})\\&= \sum_{k:\Delta_k>0}\mathbf{E}(n_{i,k}(T))\cdot\Delta_k\\&\le \sum_{k:\Delta_k>0}\left(\max\left\{\frac{12}{|\mathcal{N}_i|\Delta_k^2}\log T,\ L,\ 2(M^2+2Md+d)\right\}+\frac{2\pi^2}{3}+M^2+(2d-1)M\right)\Delta_k,\end{aligned}$$

which completes the proof. ■

C Additional Simulations

In this appendix, we provide a set of additional simulations to complement those in Section 4. We compare the homogeneous and heterogeneous settings, and include simulations for special cases of interest.

C.1 Homogeneous vs. Heterogeneous Settings

Recall that a homogeneous setting requires that all agents observe the same reward distribution for any given arm, while in a heterogeneous setting, agents may observe different reward distributions for a given arm, as long as the mean reward for the arm is consistent across all agents. Below we present side by side homogeneous and heterogeneous simulations for otherwise identical settings, first for a case where only one class of reward distribution is present, and next in a case where three different classes of reward distributions are present.

C.1.1 One Distribution

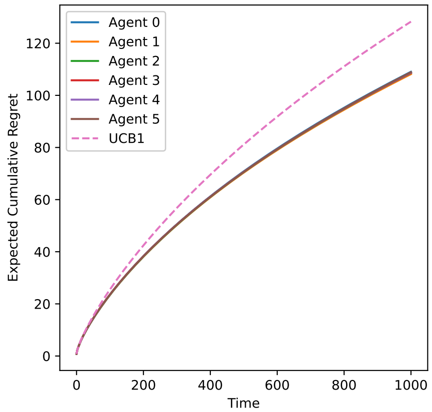
In this subsection we consider side by side homogeneous and heterogeneous simulations for strongly connected, undirected connected, and weakly connected graphs with 6 agents able to choose from a set of 6 arms for $T = 1000$ time steps, and average the results over 100 trials. Arm means are chosen uniformly at random from [0.05, 0.95]. Reward distributions follow a Beta distribution with a standard deviation of either 0.01, 0.05, or 0.1. In the homogeneous case, each arm is assigned a Beta distribution with a certain standard deviation uniformly at random. In the heterogeneous case, each agent/arm pair is assigned a Beta distribution with a certain standard deviation uniformly at random. See Figures 5, 6, and 7.

C.1.2 Three Distributions

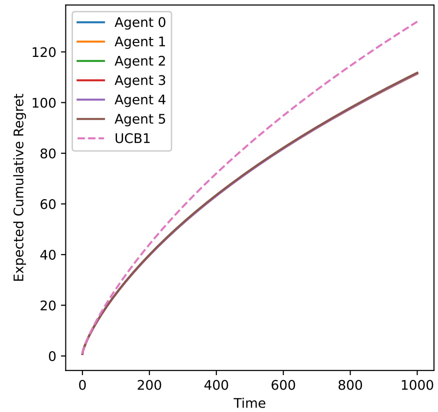
In this subsection we consider side by side homogeneous and heterogeneous simulations for strongly connected, undirected connected, and weakly connected graphs with 15 agents able to choose from a set of 10 arms for $T = 1000$ time steps, and average the results over 100 trials. Arm means are chosen uniformly at random from [0.05, 0.95]. Reward distributions follow either a truncated normal distribution with standard deviation 0.2, a Bernoulli distribution, or a uniform distribution with the greatest width possible in [0,1] given a particular mean. In the homogeneous case, each arm is assigned a distribution uniformly at random. In the heterogeneous case, each arm/agent pair is assigned a distribution uniformly at random. See Figures 8, 9, and 10.

C.1.3 Discussion

Our simulations on homogeneous and heterogeneous settings reveal no significant differences between the two in terms of regret performance. To complement Section 4 with some variety we opted to use new graph sizes, new quantities of arms, and new distributions, namely, Beta distributions with standard deviations other than 0.05, a truncated normal distribution with a standard deviation four times greater than that of Section 4, and a uniform distribution. Momentarily digressing from the homogeneous vs. heterogeneous comparison, we notice that these different settings still agree with our results in Section 3.2. If any difference between the homogeneous and heterogeneous settings

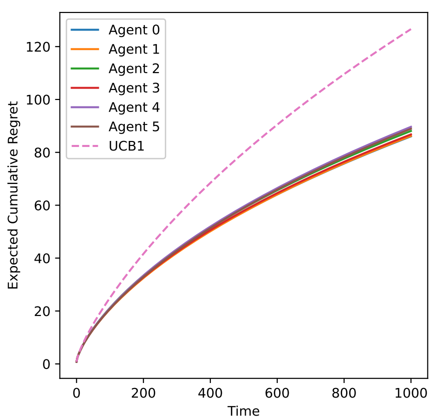


(a) Results for the homogeneous strongly connected setting. Each arm follows a Beta distribution with standard deviation chosen randomly from 0.01, 0.05, 0.1

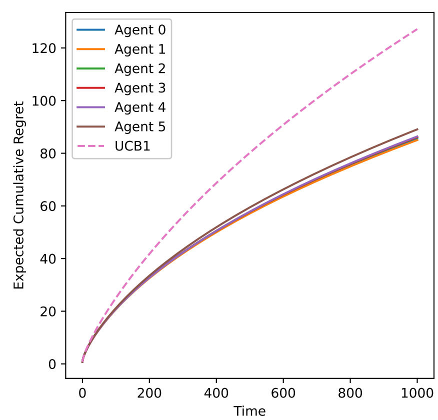


(b) Results for the heterogeneous strongly connected setting. Each agent/arm pair follows a Beta distribution with standard deviation chosen randomly from 0.01, 0.05, 0.1.

Figure 5: Homogeneous vs. heterogeneous regret plots for every agent running `Dec_UCB` and the best agent running UCB1. Results for each plot are averaged over 100 different randomly generated (Erdős–Rényi) strongly connected graphs.



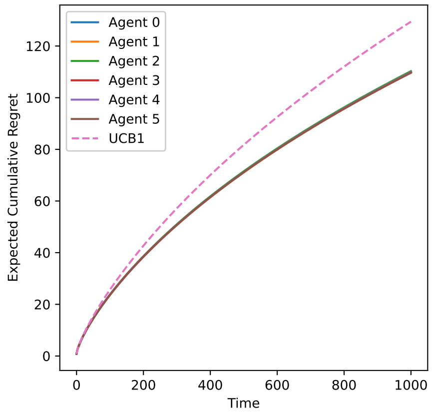
(a) Results for the homogeneous undirected connected setting. Each arm follows a Beta distribution with standard deviation chosen randomly from 0.01, 0.05, 0.1



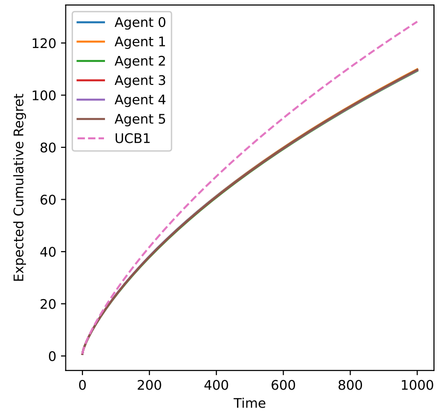
(b) Results for the heterogeneous undirected connected setting. Each agent/arm pair follows a Beta distribution with standard deviation chosen randomly from 0.01, 0.05, 0.1.

Figure 6: Homogeneous vs. heterogeneous regret plots for every agent running `Dec_UCB` and the best agent running UCB1. Results for each plot are averaged over 100 different randomly generated (Erdős–Rényi) undirected connected graphs.

must be pointed out, we could observe from the figures in C.1.1 and C.1.2 that the homogeneous regrets appear to be somewhat smaller than the heterogeneous regrets. However, the difference is quite small, and without any theoretical backing we abstain from making any claims about differences in the two settings.

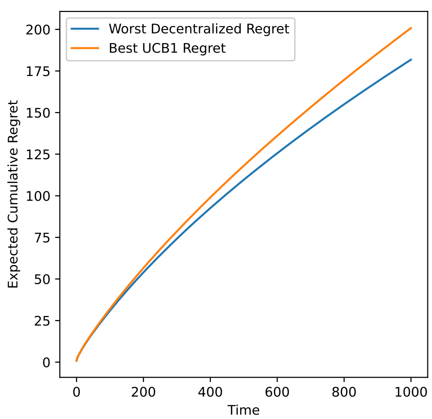


(a) Results for the homogeneous weakly connected setting. Each arm follows a Beta distribution with standard deviation chosen randomly from 0.01, 0.05, 0.1

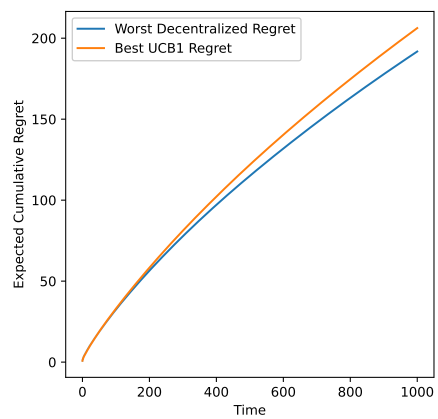


(b) Results for the heterogeneous weakly connected setting. Each agent/arm pair follows a Beta distribution with standard deviation chosen randomly from 0.01, 0.05, 0.1.

Figure 7: Homogeneous vs. heterogeneous regret plots for every agent running `Dec_UCB` and the best agent running UCB1. Results for each plot are averaged over 100 different randomly generated (Erdős–Rényi) weakly connected graphs.



(a) Results for the homogeneous strongly connected setting. Each arm follows either a Bernoulli, uniform, or truncated normal ($\sigma = 0.2$) distribution chosen at random.



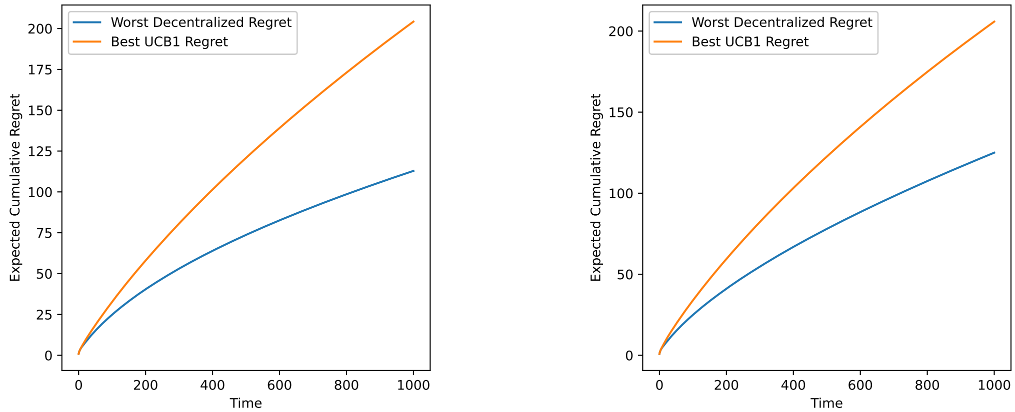
(b) Results for the heterogeneous strongly connected setting. Each agent/arm pair follows either a Bernoulli, uniform, or truncated normal ($\sigma = 0.2$) distribution chosen at random.

Figure 8: Homogeneous vs. heterogeneous regret plots for both the worst performing agent of `Dec_UCB` and best performing agent of UCB1 in otherwise identical settings. Results for each plot are averaged over 100 different randomly generated (Erdős–Rényi) strongly connected graphs.

C.2 Selected Graphs

C.2.1 Corollary 2: Undirected vs. Strongly Connected

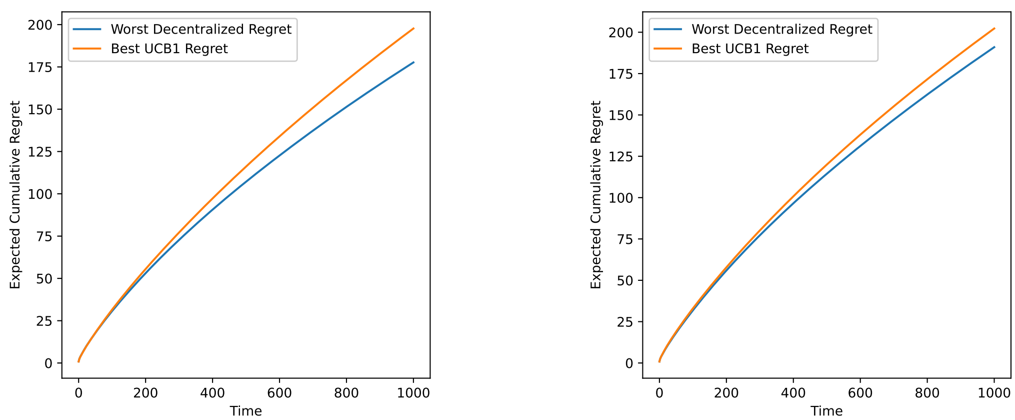
We first demonstrate the validity of Corollary 2. This is accomplished by first comparing the performance of `Dec_UCB` on an undirected, connected graph where all agents have at least two neighbors with the performance of `Dec_UCB` on a directed, strongly connected graph, keeping all other parameters the same. Six agents and six arms were used, with set means of [0.10, 0.25, 0.45,



(a) Results for the homogeneous undirected connected setting. Each arm follows either a Bernoulli, uniform, or truncated normal ($\sigma = 0.2$) distribution chosen at random.

(b) Results for the heterogeneous undirected connected setting. Each agent/arm pair follows either a Bernoulli, uniform, or truncated normal ($\sigma = 0.2$) distribution chosen at random.

Figure 9: Homogeneous vs. heterogeneous regret plots for both the worst performing agent of `Dec_UCB` and best performing agent of UCB1 in otherwise identical settings. Results for each plot are averaged over 100 different randomly generated (Erdős–Rényi) undirected connected graphs.



(a) Results for the homogeneous weakly connected setting. Each arm follows either a Bernoulli, uniform, or truncated normal ($\sigma = 0.2$) distribution chosen at random.

(b) Results for the heterogeneous weakly connected setting. Each agent/arm pair follows either a Bernoulli, uniform, or truncated normal ($\sigma = 0.2$) distribution chosen at random.

Figure 10: Homogeneous vs. heterogeneous regret plots for both the worst performing agent of `Dec_UCB` and best performing agent of UCB1 in otherwise identical settings. Results for each plot averaged over 100 different randomly generated (Erdős–Rényi) weakly connected graphs.

0.65, 0.75, 0.90] for the arms. The reward means μ_k were the same for all agents on a given arm, with each μ_k randomly chosen from a uniform distribution on $[0.05, 0.95]$. Possible reward distributions again included the Beta, Bernoulli, and truncated normal distributions, following the three distribution heterogeneous setting. Experiments were ran for $T = 1000$ time steps for a total of 100 experiments. The used graphs and their performances are shown in Figures 11 and 12 respectively. As shown, every agent in the undirected graph performs better than the agents in the strongly connected graph, with the performance of an undirected agent being directly related to its number of neighbors.

Next we compare the performance of `Dec_UCB` on an undirected, connected graph where all agents are not guaranteed to have at least two neighbors with the performance of `Dec_UCB` on the same directed, strongly connected graph. This undirected graph is shown in Figure 13. As illustrated, the undirected agents with only one neighbor (Agents 0 and 5) perform approximately equivalent to the strongly connected agents. The undirected agents with two neighbors perform better than the strongly connected agents.

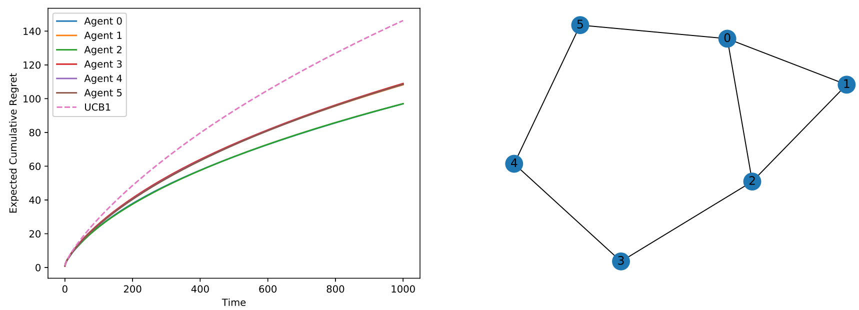


Figure 11: The used undirected graph, ensuring that all agents have at least two neighbors. Reward distributions used vary between agents for a given arm.

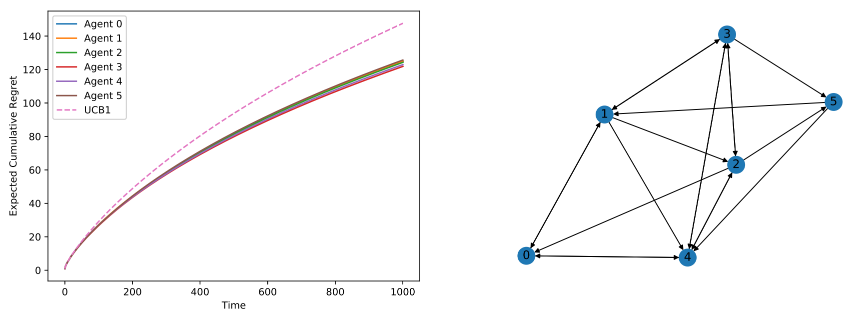


Figure 12: The used strongly connected graph for comparison. Reward distributions used vary between agents for a given arm.

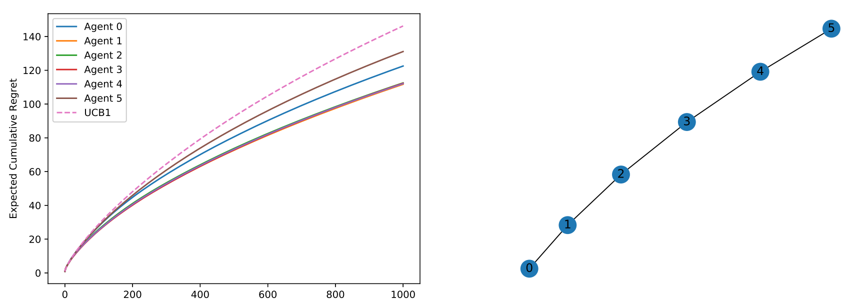


Figure 13: The used undirected graph, allowing for some agents to have only one neighbor. Reward distributions used vary between agents for a given arm.

C.2.2 (Weakly Connected) Directed Path Graph

We next illustrate results for a special weakly connected graph, directed path, for `Dec_UCB`, compared with UCB1. Six agents and six arms were used, with results averaged over 100 experiments, generating new random means μ_k each time. The reward means μ_k were the same for all agents on a given arm, with each μ_k randomly chosen from a uniform distribution on $[0.05, 0.95]$. Possible reward distributions again included the Beta, Bernoulli, and truncated normal distributions, following the three distribution heterogeneous setting. Each experiment was ran for $T = 1000$ time steps. The results are illustrated in Figure 14. These results further suggest that `Dec_UCB` can perform better than UCB1 for this special type of weakly connected graphs, though we are unable to definitively confirm or deny this until theoretical backing is developed or a counterexample is discovered.

It is worth emphasizing that in such a directed path, the "root" agent (e.g., agent 0 in Figure 14) does not have any incoming neighbor and thus receives no external information. The root agent therefore solves the bandit problem as in the conventional single-agent case. Since the upper confidence bound function of `Dec_UCB` is smaller than that of UCB1, it is likely that there exists counterexamples, with certain "worst-case" reward distributions, in which such a "root" agent cannot solve the bandit problem using `Dec_UCB`.

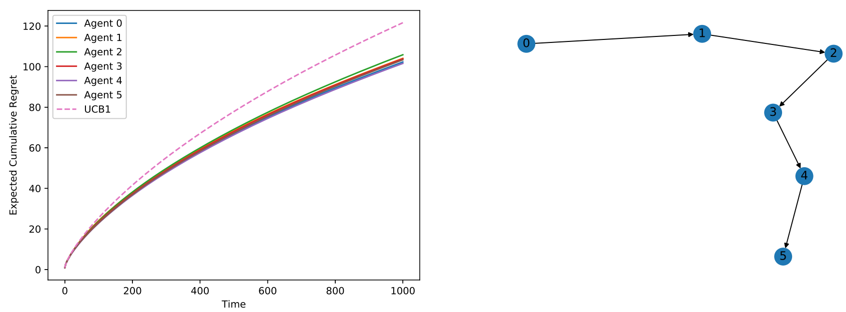


Figure 14: A plot of the regret for both `Dec_UCB` and UCB1 of the weakly connected directed path graph, averaged over 100 experiments. Reward distributions used vary between agents for a given arm.

References

[1] T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2018.

[2] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[3] L. Xiao and S. Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78, 2004.