ABSTRAK

Masalah *Stochastic Multi-Armed Bandit* (MAB) merupakan kerangka kerja fundamental dalam pemodelan pengambilan keputusan sekuensial di bawah ketidakpastian. Secara tradisional, model ini diselesaikan oleh agen yang beroperasi secara mandiri dan terisolasi. Namun, sifat kesendirian ini menjadi kelemahan fatal ketika agen dihadapkan pada ruang keputusan yang sangat masif, seperti pada sistem rekomendasi skala industri, yang mana memicu *exploration bottleneck* dan menghasilkan *cumulative regret* yang tinggi. Untuk mengatasi keterbatasan tersebut, penelitian ini mendemonstrasikan bahwa sekumpulan agen dapat berkooperasi melalui komunikasi *peer-to-peer* terbatas guna menurunkan akumulasi *regret* secara drastis. Efisiensi sistem kooperatif ini dievaluasi secara komparatif melintasi tiga kebijakan (*policy*) fondasional: *Upper Confidence Bound* (UCB), *Epsilon-Greedy*, dan *Thompson Sampling*. Selanjutnya, kami mengungkap kerentanan arsitektur ini dengan menginjeksi *malicious agents* yang menyebarkan rekomendasi acak (*data poisoning*). Hasil simulasi membuktikan bahwa tanpa mekanisme pertahanan aktif, jaringan kooperatif mengalami degradasi performa yang parah. Bahkan sebagian kecil penyerang mampu mengeksploitasi celah komunikasi untuk melumpuhkan keuntungan kerja sama, menyeret efisiensi jaringan kembali mendekati tingkat agen yang terisolasi. Kerentanan kritis ini secara tegas menyoroti urgensi pengembangan protokol pertahanan aktif pada riset mendatang.

Kata Kunci: *Multi-Armed Bandit*, *Cooperative Multi-Agent*, *Data Poisoning*, *Malicious Agents*, *Thompson Sampling*, Sistem Rekomendasi.

ABSTRACT

Stochastic Multi-Armed Bandit (MAB) problems serve as fundamental frameworks for modeling sequential decision-making under uncertainty. Traditionally, these models are solved by agents operating in strict isolation. However, this solitary nature becomes a critical limitation when facing massive decision spaces, such as in industrial-scale recommendation systems, triggering severe exploration bottlenecks and resulting in unacceptably high cumulative regret. To overcome this limitation, this study demonstrates that a network of agents can cooperate via limited peer-to-peer communication to drastically reduce overall regret. The efficiency of this cooperative system is comparatively evaluated across three foundational policies: Upper Confidence Bound (UCB), Epsilon-Greedy, and Thompson Sampling. Furthermore, we expose the inherent vulnerability of this architecture by introducing malicious agents that disseminate toxic recommendations (data poisoning). Simulation results prove that without active defense mechanisms, cooperative networks suffer catastrophic performance degradation. Even a small fraction of attackers can exploit the communication protocol to negate the benefits of cooperation, dragging network efficiency back toward isolated agent levels. This systemic fragility underscores the absolute necessity of developing active defense protocols in future works.

Keywords: Stochastic Multi-Armed Bandit, Cooperative Multi-Agent, Data Poisoning, Malicious Agents, Thompson Sampling, Recommendation Systems.