Improvements on Community Membership Hiding via a Custom Reinforcement Learning Architecture*

Dante Culaciati[1,2], Esteban Feuerstein[1,2], and Gabriel Tolosa[3]

[1] Depto. de Computación, Fac. de Ciencias Exactas y Naturales, Universidad de Buenos Aires
[2] Instituto de Ciencias de la Computación, UBA-CONICET, Argentina
[3] Depto. de Ciencias Básicas, Universidad Nacional de Luján, Argentina

Abstract. As complex networks face increasingly advanced algorithmic analysis, the ability to conceal specific structural affiliations has emerged as a relevant graph algorithmic problem. Current online platforms rely on sophisticated techniques to characterize and cluster users in order to provide more contextualized content and services. One such approach is based on community detection algorithms applied to large underlying graphs. In this context, users are characterized by preferences or attributes, which can reveal information that may be considered sensitive. Community Membership Hiding is the problem of strategically rewiring network topology to camouflage a target node and evade detection by clustering algorithms. While recent Deep Reinforcement Learning (DRL) approaches have pioneered this line of work, they often face efficacy and scalability issues on large-scale graphs. By dissecting state-of-the-art models, we identified a critical architectural limitation: agents evaluate the network globally without explicitly conditioning on the target node, while relying on static embeddings that fail to capture real-time topological changes during the rewiring process. In this paper, we introduce a novel DRL architecture designed to overcome these fundamental limitations. We propose an Actor-Critic system powered by Graph Attention Networks v2 (GATv2). By leveraging a target-aware feature set that combines static original-graph indicators with dynamically updated structural signals, and by explicitly contextualizing the target node through a dynamic attention mechanism, our model accurately represents the evolving graph state. Extensive experimental evaluation confirms that our architecture significantly outperforms existing baselines, achieving high hiding success rates with minimal structural perturbation and notable gains in computational efficiency.

Keywords: community detection, community membership hiding, deep reinforcement learning, graph attention networks.

* Research funded by UBACYT Project 20020190100126BA Modelos y Herramientas Algorítmicas Avanzadas para Redes y Datos Masivos - Parte 2. The first author's research is supported by the "Becas de Iniciación a la Investigación en Cs. de la Computación (BIICC)" program, fully funded by alumni of Departamento de Computación, UBA. At present, national research funding has been drastically reduced, and no new competitive grants or operating budgets have been available for more than two years. As a result, the Argentine scientific system is undergoing an unprecedented contraction, leaving laboratories without the resources traditionally required for research continuity.

Mejoras en Ocultamiento de Pertenencia a Comunidades mediante una Arquitectura Especializada de Aprendizaje por Refuerzo

Resumen. A medida que las redes complejas enfrentan un análisis algorítmico cada vez más avanzado, la capacidad de ocultar afiliaciones estructurales específicas ha emergido como un problema algorítmico relevante de grafos. Las plataformas actuales dependen de técnicas sofisticadas para caracterizar y agrupar usuarios con el fin de ofrecer contenido y servicios más contextualizados. Uno de estos enfoques se basa en algoritmos de detección de comunidades aplicados a grafos de gran tamaño. En este contexto, los usuarios quedan caracterizados por preferencias o atributos que pueden revelar información sensible. El Ocultamiento de Pertenencia a Comunidades es el problema de reconfigurar estratégicamente la topología de la red para camuflar un nodo objetivo y evadir su detección por algoritmos de agrupamiento. Si bien los enfoques recientes de Aprendizaje por Refuerzo Profundo han sido pioneros en esta línea, suelen presentar problemas de eficacia y escalabilidad en grafos grandes. Al analizar modelos del estado del arte, identificamos una limitación arquitectónica crítica: los agentes evalúan la red de forma global sin condicionar explícitamente sobre el nodo objetivo, y dependen de embeddings estáticos que no capturan los cambios topológicos en tiempo real durante la reconfiguración. En este trabajo, introducimos una nueva arquitectura diseñada para superar estas limitaciones fundamentales. Proponemos un sistema Actor-Crítico impulsado por Redes de Atención de Grafos v2 (GATv2). Al aprovechar un conjunto de atributos orientado al nodo objetivo que combina indicadores estáticos del grafo original con señales estructurales actualizadas dinámicamente, y al contextualizar explícitamente el nodo objetivo mediante un mecanismo de atención dinámica, nuestro modelo representa con precisión el estado evolutivo del grafo. Una evaluación experimental extensa confirma que nuestra arquitectura supera significativamente a las líneas de base existentes, alcanzando altas tasas de ocultamiento con perturbación estructural mínima y mejoras notables en eficiencia computacional.

Palabras clave: detección de comunidades, ocultamiento de pertenencia a comunidades, aprendizaje por refuerzo profundo, redes de atención de grafos.

1 Introduction

The modern science of networks has brought significant advances to our understanding of complex systems. One of the most relevant features of graphs representing real-world structures, ranging from social media platforms to biological networks, is their community structure, i.e. the identification of tightly connected groups of nodes, a task for which there exist highly effective algorithms (Fortunato, 2010). However, in some contexts, this functionality inherently risks exposing individuals to privacy breaches by inadvertently revealing sensitive affiliations, such as political leanings or religious beliefs, based solely on their structural position.

To counter this, the concept of Community Membership Hiding (CMH) has been proposed (Bernini et al., 2024). The objective of CMH is to strategically alter the structural properties of a network to prevent a specific target node from being grouped into its original community by a detection algorithm, while maintaining overall graph integrity and ideally treating the detection algorithm as a "black box" with respect to its internal mechanics.

Recent studies have formulated this evasion tactic as a Markov Decision Process (MDP) solved via Deep Reinforcement Learning (DRL) (Bernini et al., 2024). While that work demonstrated the viability of the method, it exhibits notable performance degradation when scaled to larger graphs, and has fundamental limitations that hinder its ability to achieve optimal performance. This work directly develops from (Bernini et al., 2024) and treats its DRL architecture as the primary baseline reference. Building on that foundation, we dissect its architectural shortcomings and propose a novel, highly efficient RL architecture that leverages modern Graph Neural Networks to preserve privacy while maintaining overall graph characteristics. We compare our results (GATv2 Actor-Critic) against three strong baselines: the original agent introduced in (Bernini et al., 2024), and two gradient-based methods (∇-CMH and ∇-CMH (projected)) introduced in (Silvestri et al., 2025), which improve over the original one while achieving better computational efficiency. We show in Section 4 that our method consistently outperforms the original method with less computation time. Against the $\nabla-$ variants, we obtain better effectiveness at a cost of more computation time, so there is a trade-off to consider.

2 Community Evasion

2.1 The Problem

The following definitions are borrowed from (Bernini et al., 2024). Let $\mathcal{G} = (V, E)$ be a graph where V is the set of nodes and E is the set of edges. Given a community detection algorithm $f(\cdot)$ and a target node $u \in V$, the algorithm assigns u to an original community C_i.

The goal of CMH is to find a modified graph $\mathcal{G}' = (V, E')$ such that upon re-evaluating $f(\mathcal{G}')$, the new community assigned to the target node, denoted as C_i', is sufficiently dissimilar to the original community. Formally, the node is considered hidden when:

$$\mathrm{sim}(C_i - \{u\}, C_i' - \{u\}) \leq \tau$$

where $\tau \in [0, 1]$ is a predefined hiding threshold that determines the required divergence, and $\mathrm{sim}(\cdot, \cdot)$ is a similarity function, such as the Sørensen-Dice coefficient.

To aid in successfully hiding the target node, modifications to the rewiring are restricted. The agent may only remove outgoing edges from u to nodes $v \in C_i$ and add outgoing edges from u to nodes $v \notin C_i$; that is, it can only remove edges inside its original community, and add edges to nodes not in its original community.

In addition to achieving hiding success, a desired trait is to maximize Normalized Mutual Information (NMI, a measure of similarity between two clustering assignments) by introducing the smallest possible structural perturbation, which often also involves hiding the node in a minimal number of steps. The maximum number of modifications (edge budget) is defined through a parameter β as: $\max\left(1, \left\lfloor \frac{|E|}{|V|} \cdot \beta \right\rfloor\right)$.

2.2 Original Agent Architecture

To process the graph state within the established MDP framework, the baseline architecture introduced by (Bernini et al., 2024) utilizes a Graph Convolutional Network (GCN) (Kipf & Welling, 2017). Crucially, the initial node features fed into the GCN are generated using *node2vec* (Grover & Leskovec, 2016), which maps each node's structural role into a latent vector space based on the initial, unaltered graph. Furthermore, the agent is constrained to act on the target node not through an internal target-aware mechanism, but via an external environment wrapper that masks invalid actions, artificially restricting all rewiring to the target's immediate neighborhood.

2.3 Architectural Limitations in the Baseline

Upon analyzing the baseline implementation, we identified two critical bottlenecks that limit its scalability and effectiveness on larger, more complex graphs:

Global Evaluation without Target Context: The GCN receives the entire graph as input but lacks explicit, localized knowledge of the target node u. The environment implicitly enforces valid actions, meaning the agent's policy evaluates the global graph structure rather than explicitly contextualizing the specific neighborhood it must camouflage.

Static Embeddings: The *node2vec* embeddings are computed only once on the initial graph $\mathcal{G}$. As the agent actively rewires the network, these static embeddings cease to reflect the current, evolving topological reality. This creates a severe information lag, effectively blinding the agent to the consequences of its own structural modifications and degrading learning efficiency.

3 Proposed Architecture: GATv2 Actor-Critic

Our approach consists of an **Actor-Critic** model with a **shared graph encoder** that uses a special version of Graph Attention Networks (Veličković et al., 2018), designed to keep the state representation synchronized with the evolving topology after each rewiring action. Instead of relying on static embeddings, we build a *target-aware* feature space that combines signals from the original graph with signals recomputed after each modification. Essentially, we are reinterpreting the attention concept by feeding back the updated context into the neural network on each step. The feature design is organized into two complementary groups:

Static target-context indicators (computed on the original graph): Identify whether a node is the *target node*, whether it belongs to the *target community*, and whether it is an *original neighbor* of the target.

Dynamic structural indicators (updated after rewiring): Include *normalized degree*, *current neighbor status* with respect to the target, *normalized co-citation/common-neighbor score*, and *Jaccard similarity* to the target.

At the core of the model, a **shared GNN backbone** based on stacked GATv2 (Brody et al., 2022) layers encodes node features and edge information into node embeddings. These shared embeddings are consumed by both **actor** and **critic** heads. The actor computes node-level logits conditioned on the target embedding, while the critic combines an attention-pooled graph representation with the target embedding to estimate the state value. This shared-encoder design reduces redundant computation and enforces a consistent latent representation for policy and value estimation.

The training signal is aligned with the CMH objective and action costs. A successful hiding outcome receives a **positive terminal reward**, while exhausting the edge budget without hiding yields a **negative terminal reward**. For non-terminal steps, edge addition and edge removal receive different step costs, encouraging *efficient rewiring* under the budget constraint. Hidden status is determined by the community-similarity criterion relative to the threshold controlled by τ. For reproducibility, the implementation is publicly available[1].

4 Results

We report results using the F1 score, defined as the harmonic mean of the hiding success rate and Normalized Mutual Information (NMI), following the standard CMH evaluation protocol (Bernini et al., 2024). We executed thorough experiments over five datasets of different sizes and densities, on an AWS EC2 g4dn.xlarge instance, with 4 Intel Xeon Scalable vCPUs, 16GiB of RAM, and an NVIDIA T4 Tensor Core GPU. To evaluate the robustness of our approach, specifically its sensitivity to the hiding threshold (τ) and the edge budget (β), we performed a parameter sweep with $\tau \in \{0.3, 0.5, 0.8\}$ and $\beta \in \{0.5, 1.0, 2.0\}$ and compared the success rates against the original agent. Table 1 shows the overall consistency of the results. Figure 1 summarizes our results for specific values of these parameters.

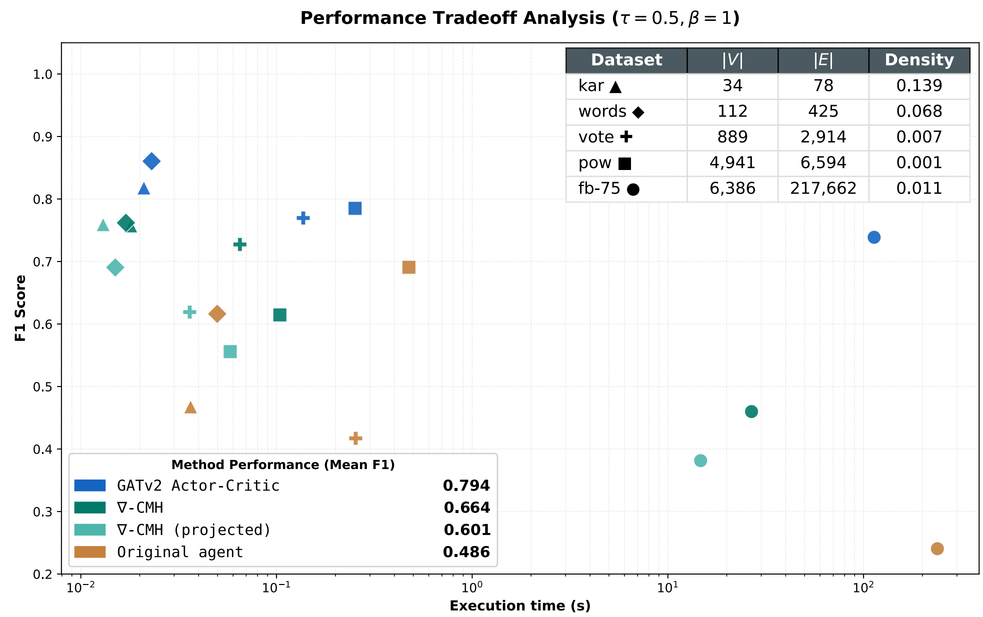

Fig. 1. Tradeoff between F1 score and execution time ($\tau = 0.5,\ \beta = 1$). Time axis in logarithmic scale.

As shown in Figure 1, our GATv2 Actor-Critic method consistently outperforms the baseline models across the evaluated datasets. Notably, on the largest graph, our approach achieves up to a 40% improvement in F1 score compared to ∇-CMH, the second most effective method. While the gradient-based ∇-CMH methods achieve the best computational efficiency metrics, they fail to reach the high F1 effectiveness demonstrated by our DRL architecture, showing a trade-off between F1 score and execution time.

[1] https://github.com/Dante-010/codigo_tesis.

As shown in Table 1, our proposed agent demonstrates a substantial and consistent improvement in success rate over the original agent across all tested configurations.

These trends indicate that while stricter hiding constraints (higher τ) naturally increase the difficulty of the task, our dynamic, target-aware architecture maintains a robust performance gap over the baseline. Furthermore, varying the edge budget (β) confirms that moderate to high budget allowances allow our model to capitalize on the extra flexibility to achieve near-perfect success rates, whereas the original agent's performance prematurely plateaus. Importantly, the proposed architecture maintains consistent behavior across the parameter grid, proving it does not overfit to a single scenario.

β	τ	**GATv2**	**Original**
	0.3	61.67	41.67
0.5	0.5	70.21	38.22
	0.8	79.38	62.67
	0.3	79.17	61.67
1.0	0.5	91.78	58.44
	0.8	96.25	90.00
	0.3	82.08	77.67
2.0	0.5	98.75	74.67
	0.8	99.17	96.33

Table 1. Success rates.

5 Conclusions and Further Work

We show that architectural alignment between representation, decision, and reward is central to scalable CMH. By combining target-aware feature construction, a shared GATv2 encoder, and coordinated actor-critic heads, the proposed method improves the tradeoff between hiding success and structural preservation while also reducing convergence time relative to reference baselines. Beyond the specific benchmark gains, the results suggest that RL-based graph evasion benefits substantially from feature designs that remain synchronized with graph evolution during rewiring. This perspective opens a practical path toward extending the framework to broader hiding scenarios and more complex community-detection settings. In the near future, we will extend this framework to overlapping community detection scenarios using proxy node injection and to benchmark against other formulations of CMH.

Acknowledgments. The authors would like to thank the original authors of the Community Membership Hiding baseline for their prompt clarifications regarding their codebase.

References

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3–5), 75–174. https://doi.org/10.1016/j.physrep.2009.11.002

Grover, A., & Leskovec, J. (2016). Node2vec: Scalable feature learning for networks. https://arxiv.org/abs/1607.00653

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. https://arxiv.org/abs/1609.02907

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. https://arxiv.org/abs/1710.10903

Brody, S., Alon, U., & Yahav, E. (2022). How attentive are graph attention networks? https://arxiv.org/abs/2105.14491

Bernini, A., Silvestri, F., & Tolomei, G. (2024). Evading community detection via counterfactual neighborhood search. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 131–140. https://doi.org/10.1145/3637528.3671896

Silvestri, M., Gabrielli, E., Silvestri, F., & Tolomei, G. (2025). The right to hide: Masking community affiliation via minimal graph rewiring. https://arxiv.org/abs/2502.00432