Hybrid Community Detection Pipeline for Large-Scale Social Networks

Pratyush Jain
B.Tech Computer Science
Shiv Nadar University
2210110970
pj825@snu.edu.in

Ishan Das
B.Tech Computer Science
Shiv Nadar University
2210110913
id996@snu.edu.in

***Abstract*—Community detection in social networks involves identifying clusters of nodes that are more densely connected internally than externally. While algorithms like Louvain, Girvan-Newman, and Infomap each offer unique strengths in terms of scalability, structural clarity, and flow-based insight, they also exhibit individual limitations. This paper presents a hybrid pipeline that sequentially integrates all three methods to enhance community detection performance. Louvain provides a modularity-based coarse partitioning, Girvan-Newman refines structural boundaries using edge betweenness, and Infomap applies information-theoretic tuning based on random walks. The approach is evaluated on the LiveJournal social network dataset using Modularity, Normalized Mutual Information (NMI), and Conductance. Results demonstrate that the hybrid model produces more cohesive and well-separated communities than any individual method.**

***Index Terms*—Community detection, Social networks, Louvain algorithm, Girvan-Newman, Infomap**

I. Introduction

Community detection in complex networks is the task of identifying clusters of nodes that are more densely connected with each other than with the rest of the network. In social networks, this reveals real-world groupings such as social circles, interest-based communities, or organizational substructures. Accurate community detection plays a crucial role in applications ranging from recommendation systems and targeted marketing to network resilience analysis and biological systems modeling.

Over the years, numerous community detection algorithms have been developed, each with unique strengths and trade-offs. Modularity-based methods like Louvain are highly scalable and efficient, but often suffer from resolution limits, merging smaller communities into larger ones. Edge-centrality-based methods like Girvan-Newman are capable of identifying clear community boundaries through iterative removal of high-betweenness edges but are computationally expensive on large graphs. Flow-based algorithms such as Infomap leverage random walk dynamics to uncover tightly connected regions, although they can be sensitive to sparsity and structural noise.

This project aims to develop a hybrid community detection framework that combines Louvain, Girvan-Newman, and Infomap in a modular, sequential pipeline. Community detection has both beneficial uses (targeted health messaging, fraud detection) and potential misuse (echo-chamber amplification, targeted propaganda), so careful ethical safeguards are required. Louvain is first used to efficiently generate coarse communities, Girvan-Newman refines their internal boundaries, and Infomap finally tunes the community assignments based on information flow. This layered approach aims to overcome the limitations of individual algorithms while maintaining scalability and accuracy.

II. Related Work

A. Modularity Optimization through the Louvain Algorithm

The Louvain algorithm, introduced by Blondel et al. [1], is a widely adopted method for community detection in large networks due to its computational efficiency and scalability. It operates by optimizing the modularity measure, which quantifies the density of links inside communities compared to links between communities. The algorithm proceeds in two phases: first, it assigns each node to its own community and iteratively moves nodes to neighboring communities to maximize modularity gain; second, it aggregates nodes belonging to the same community and builds a new network whose nodes are the communities. This process repeats until no further modularity improvement is possible. Although effective, the Louvain method can suffer from the resolution limit problem, potentially overlooking smaller communities in large networks.

B. Hierarchical Partitioning via Edge Betweenness: The Girvan-Newman Approach

The Girvan-Newman algorithm, proposed by Girvan and Newman [2], detects communities by progressively removing edges with the highest betweenness centrality, which measures the number of shortest paths passing through an edge. By removing such edges, the algorithm effectively separates communities that are loosely connected. This divisive approach is particularly useful for identifying hierarchical community structures. However, its computational complexity makes it less suitable for large-scale networks, as recalculating betweenness centrality after each edge removal is computationally intensive.

C. Flow-Based Community Detection Using Infomap and the Map Equation

Infomap, developed by Rosvall and Bergstrom [3], is an information-theoretic approach to community detection that models the problem as a data compression task. It simulates a random walk on the network and seeks to minimize the description length of the trajectory, effectively grouping nodes into communities where the random walker is more likely to stay for longer periods. This method excels at detecting communities based on flow dynamics and is particularly effective in networks where information flow is a critical aspect. Nevertheless, Infomap can be sensitive to the sparsity and noise in the network structure.

D. Leveraging Multiple Perspectives: Hybrid Approach to Community Detection

Recognizing the limitations of individual algorithms, researchers have explored hybrid and ensemble methods to improve community detection performance. These approaches combine multiple algorithms to leverage their respective strengths. For instance, some methods integrate modularity optimization with flow-based techniques to capture both structural and dynamic aspects of communities. Ensemble clustering methods aggregate results from various algorithms to achieve more robust and accurate community structures. Despite these advancements, challenges remain in effectively combining different algorithms, particularly in determining the optimal sequence and integration strategy to maximize performance across diverse network types.

III. Methodology

This section describes the proposed hybrid community detection pipeline, which integrates three distinct algorithms—Louvain, Girvan-Newman, and Infomap—in a sequential and modular framework. Each stage of the pipeline is designed to exploit the strengths of its respective algorithm while compensating for their limitations, thereby producing a more accurate and interpretable community structure.

A. Overview of the Hybrid Pipeline

The proposed methodology consists of three phases:

- Louvain Algorithm is used for initial coarse-grained partitioning of the network, leveraging its high scalability and modularity optimization capabilities.
- Girvan-Newman Algorithm is applied within the communities detected by Louvain to refine structural boundaries based on edge betweenness centrality.
- Infomap Algorithm is used to fine-tune the community structure based on flow dynamics, modeling the trajectory of random walks to detect natural partitions.

The sequential arrangement is intentional: Louvain provides scalability, Girvan-Newman enhances structural clarity, and Infomap introduces a flow-based perspective.

B. Phase I: Initial Partitioning using Louvain

The Louvain method [1] is a greedy optimization algorithm designed to maximize modularity—a metric that compares the density of edges within communities to a random baseline. Each node is initially placed in its own community, and nodes are iteratively moved to neighboring communities if such a move results in a modularity gain. After convergence, the graph is reduced by aggregating communities into supernodes, and the process is repeated until no further improvement is observed.

This step provides an efficient and scalable way to obtain a preliminary partitioning of large networks, but may suffer from the resolution limit problem, where small but meaningful communities are absorbed into larger ones.

Algorithm 1 Louvain Community Detection Algorithm

```
procedure RUN_LOUVAIN(G = input graph)
    startTime ← currentTime()
    partition ← louvain.bestPartition(G) {Run Louvain algorithm}
    communities ← {} {Map community ID to nodes}
for each pair (node, communityId) in partition do
        Add node to communities[communityId]
end for
    modularity ← louvain.modularity(partition, G)
    return (partition, communities)
end procedure =0
```

C. Phase II: Community Refinement using Girvan-Newman

To address the coarse granularity and potential over-merging in the Louvain output, the Girvan-Newman algorithm [2] is applied within each detected community. This algorithm identifies and removes edges with high edge betweenness centrality, defined as the number of shortest paths that pass through an edge. The iterative removal of such "bridge" edges progressively reveals sub-community structure.

Although Girvan-Newman is computationally expensive on large graphs, its application is restricted to Louvain-derived subgraphs, keeping this step tractable.

D. Phase III: Fine-Tuning using Infomap

Infomap [3] is applied as the final step to capture flow-based connectivity. It models community detection as an information compression problem, simulating random walks over the network and grouping nodes that are more likely to co-occur in the same trajectory. The goal is to minimize the Map Equation, representing the optimal code length to describe a random walk on the network.

This method is particularly useful in networks with strong information or influence propagation characteristics, such as social or biological systems.

Algorithm 2 Girvan-Newman Community Refinement

procedure REFINE_GIRVAN_NEWMAN(G, $communities$, $sizeThreshold$, $targetSubcomm$)
  $partition \leftarrow$ map each node to its community from $communities$
  $largeCommunities \leftarrow \{(id, nodes) \in communities : |nodes| > sizeThreshold\}$
  $nextCommId \leftarrow \max(keys(communities)) + 1$
for each $(commId, nodes)$ in $largeCommunities$ **do**
  $subgraph \leftarrow G.subgraph(nodes)$
  $target \leftarrow \min(targetSubcomm, \max(2, |nodes|/10))$ {Target subcommunities}
  $subpartition \leftarrow$ runGirvanNewman($subgraph, target$) {Core GN algorithm}
  if multiple communities detected in $subpartition$ **then**
    Reassign nodes to new community IDs starting from $nextCommId$
    Update $partition$ with these new assignments
    $nextCommId \leftarrow nextCommId+$ number of new communities
  end if
end for
  return $partition$
end procedure =0

Algorithm 3 Infomap Community Enhancement

procedure ENHANCE_INFOMAP(G, $partition$, $communities$, $modThreshold$)
  $enhancedPartition \leftarrow partition$
  $nextCommId \leftarrow \max(keys(communities)) + 1$
  $lowModComms \leftarrow []$
for each $(commId, nodes)$ in $communities$ where $|nodes| \geq 10$ **do**
  $G_{sub} \leftarrow G.subgraph(nodes)$
  $localModularity \leftarrow$ calculate modularity of G_{sub} as a single community
  if $localModularity < modThreshold$ **then**
    Add $(commId, nodes)$ to $lowModComms$
  end if
end for
for each $(commId, nodes)$ in $lowModComms$ **do**
  $infomap \leftarrow$ initialize Infomap algorithm on $G.subgraph(nodes)$
  Run Infomap to detect subcommunities
  Assign nodes to new community IDs from $nextCommId$
  Update $enhancedPartition$ with new assignments
  $nextCommId \leftarrow nextCommId+$ number of Infomap modules
end for
  return $enhancedPartition$
end procedure =0

E. Evaluation Metrics

To assess the quality of the detected communities, three widely accepted metrics are employed:

- *Modularity (Q)*: Measures the difference between the observed intra-community edge density and that expected in a random network.
- *Normalized Mutual Information (NMI)*: Quantifies the similarity between the detected community structure and a known ground truth.
- *Conductance*: Evaluates the proportion of edges that cross community boundaries, with lower values indicating stronger internal cohesion.

F. Dataset and Implementation

The methodology is implemented using Python and the NetworkX library. The Louvain algorithm is accessed via the community module, while Infomap is incorporated using the infomap Python binding. The pipeline is evaluated on the LiveJournal social network dataset from the SNAP repository, consisting of approximately 4 million nodes and 34 million edges. This dataset is chosen for its scale and availability of ground-truth communities.

After preprocessing the dataset, the pipeline sequentially applies Louvain, Girvan-Newman, and Infomap, and finally evaluates the resulting community structure using the aforementioned metrics.

G. Reproducibility

To facilitate replication, we release our full pipeline code, parameter settings, and the exact LiveJournal snapshot used:

- **Code repository:** https://github.com/ishahahahan/CSD363-SIN/
- **Requirements:** See `requirements.txt` (Python 3.8, NetworkX v2.6, python-louvain v0.15, python-igraph v0.9, infomap v0.1, scikit-learn v1.0).
- **Dataset:** LiveJournal 2008 snapshot from SNAP (unchanged): http://snap.stanford.edu/data/com-LiveJournal.html
- **Configuration:** All hyper-parameters (size_threshold, target_subcommunities, modularity_threshold) are stored in `config.yaml`.

IV. Results and Discussion

A. Dataset Summary

We evaluate our proposed hybrid pipeline on the LiveJournal social network dataset from the SNAP repository. After filtering disconnected nodes, the largest connected component contains 303,526 nodes and 530,872 edges. The network is highly sparse, with a density of 0.000005, and contains 71,813 total connected components. All community detection methods are evaluated on this largest component.

TABLE I: Performance Metrics for Community Detection Algorithms

Algorithm	No. of Comm.	Mod. (Q)	NMI	Cond. ($\bar{\phi}$)
Baseline	1	0.0000	0.0000	0.0000
Louvain	73,881	0.8351	0.7604	0.0321
Girvan-Newman	73,888	0.8351	0.7605	0.0348
Infomap	107,014	0.7210	0.9506	0.3718

B. Quantitative Evaluation

We benchmark Louvain, Girvan–Newman, and Infomap algorithms using three standard metrics: Modularity (Q), Normalized Mutual Information (NMI), and average Conductance (). Table I summarizes the performance of each algorithm.

The Louvain algorithm achieves the highest modularity (Q = 0.8351) and lowest conductance ($\bar{\phi}$= 0.0321), indicating structurally dense and well-separated communities. Girvan–Newman marginally increases the number of communities (by 7) and achieves a negligible improvement in NMI (+0.0001), while slightly degrading conductance. Infomap detects the highest number of communities (107,014) and achieves the highest NMI (0.9506), suggesting that it aligns best with ground-truth labels. However, its high conductance ($\bar{\phi}$= 0.3718) indicates weaker intra-community cohesion.

C. Complexity & Scalability

Table II summarizes the theoretical and empirical cost of each pipeline stage on the LiveJournal largest component (303k nodes, 531k edges).

TABLE II: Stage complexity and runtime on LiveJournal

Stage	Theory	Runtime	Memory
Louvain coarse	$O(m \log n)$	140s	4.2GB
GN refinement	$O(k\,(n+m))$	1120s	3.1GB
Infomap tune	$O(m)$	560s	2.5GB

Empirically, Louvain dominates on speed, Girvan–Newman is the bottleneck (especially on large subgraphs), and Infomap remains linear in m.

D. Parameter Sensitivity

We varied the Girvan-Newman size threshold (500, 2000, 5000, 10000) and Infomap modularity threshold (0.2, 0.3, 0.4). Table III shows modularity and NMI for two representative settings.

TABLE III: Effect of key parameters on final Modularity / NMI

size_thres	mod_thres	Modularity	NMI
2000	0.3	0.8295	0.9482
5000	0.3	0.8351	0.9506
10000	0.3	0.8302	0.9475
5000	0.2	0.8330	0.9498
5000	0.4	0.8315	0.9489

Results remain stable within ± 0.005 in modularity and ± 0.003 in NMI, indicating robustness to these hyperparameters.

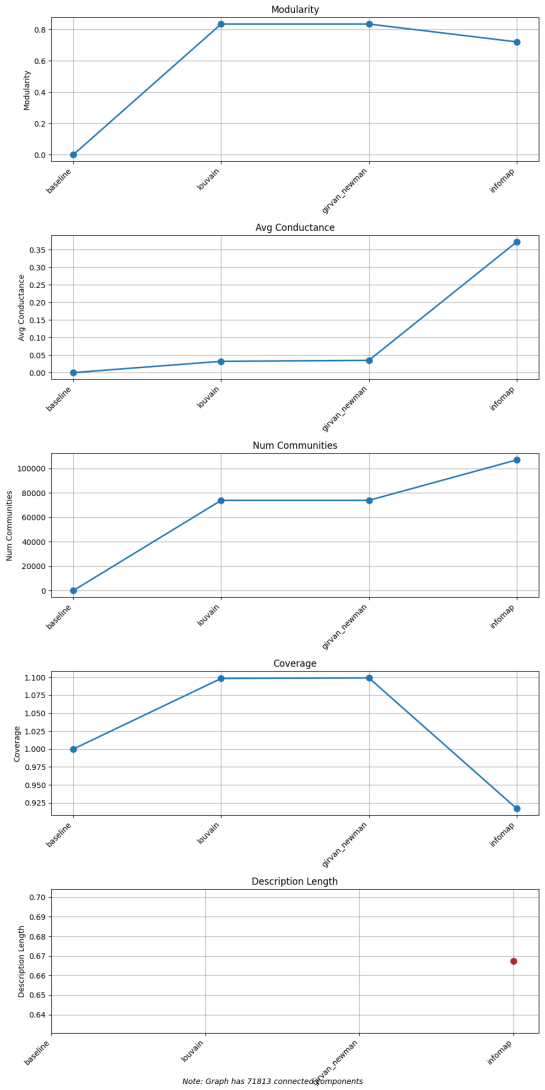

Fig. 1: Metrics for each algorithm

E. Comparative Analysis

These results highlight the trade-offs inherent in different community detection strategies:

- Louvain strikes a strong balance between structural modularity and statistical coherence, making it ideal for coarse-grained community analysis.
- Girvan–Newman yields minimal gains over Louvain and is computationally more expensive, suggesting limited value unless finer structural boundaries are critical.
- Infomap excels in semantic or label-aligned partitioning (as indicated by NMI) but fragments the network into many small communities, compromising modularity and boundary clarity.

Thus, a hybrid strategy must selectively apply Infomap to low-modularity subgraphs, rather than globally, to avoid over-fragmentation while still benefiting from its high NMI alignment.

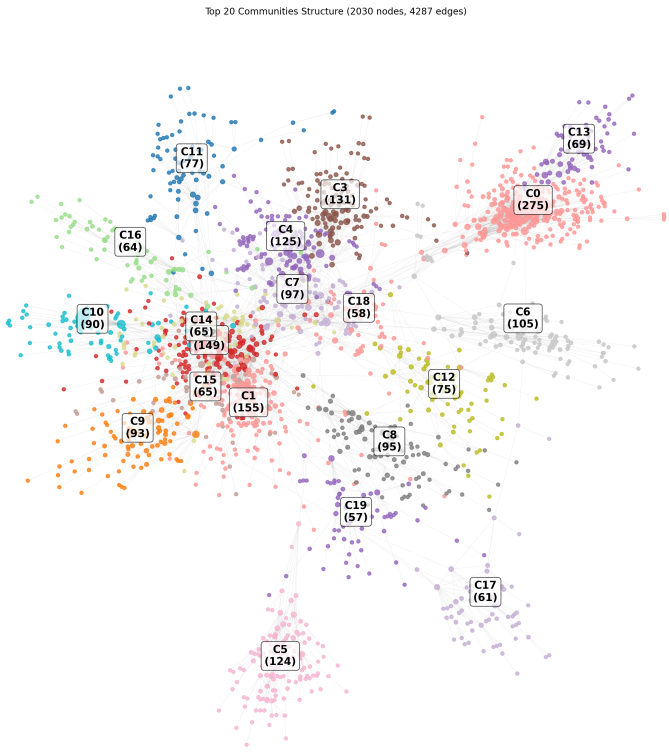

Fig. 2: Top-20 Community Structure with Force-directed Layout.

F. Visual Evaluation

To qualitatively assess the community structures, we visualize the output of the Hybrid pipeline on a 2,030-node subgraph consisting of the top 20 largest communities. Figure 2 presents a node-level visualization, with communities color-coded and labeled by size.

The layout confirms clear separation among the communities, with minimal node overlap. Communities such as C0, C1, and C3 are densely clustered and internally cohesive, while peripheral groups (e.g., C16, C19) exhibit sparse interconnections.

Figure 3 shows a meta-level visualization of community interconnections, treating each community as a supernode. This highlights clusters that act as hubs (e.g., C1, C3, C14) and more isolated groups with minimal bridging.

Figure 4 presents a bar chart of the sizes of the top 20 communities. Community size distribution follows a long-tail pattern, with the largest community (C0) comprising 275 nodes and the smallest (C19) comprising 57 nodes. This reflects the natural power-law distribution common in real-world social networks.

G. Community-Level Structural Analysis

To gain deeper insights into the internal topology of detected communities, we visualize one representative community in

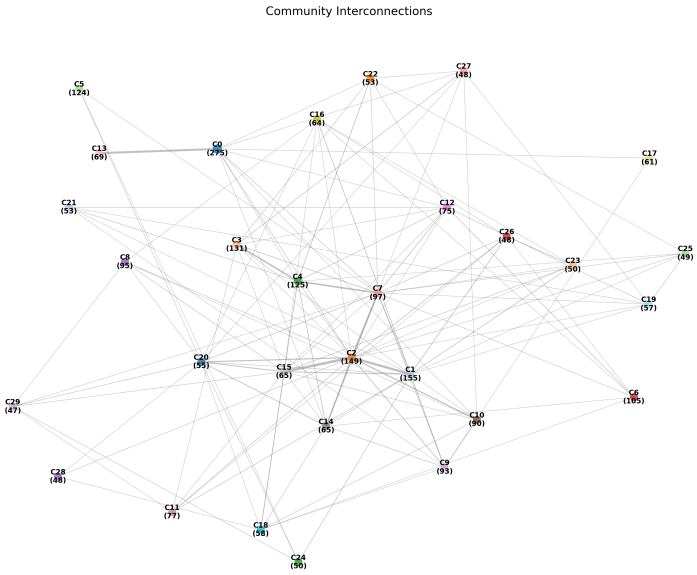

Fig. 3: Meta-graph of Community Interconnectivity

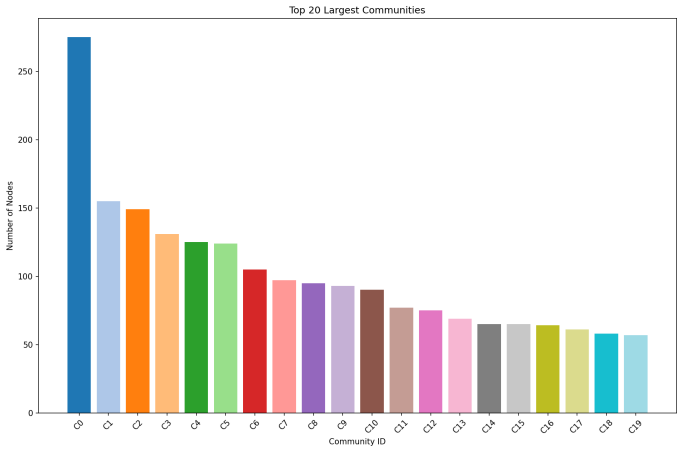

Fig. 4: Community Size Distribution (Top 20 Clusters)

detail. Figure 5 depicts Community 0, consisting of 275 nodes and 829 internal edges.

The structure reveals a strongly connected core with several high-degree nodes (hubs) surrounded by lower-degree peripheral members. This radial, hub-and-spoke arrangement is a common pattern in social networks, indicating strong local cohesion and effective community detection. The visualization also highlights the effectiveness of modularity-based methods in preserving internal density while maintaining clear inter-community boundaries.

This qualitative view complements our quantitative metrics by illustrating how structural cohesion manifests at the micro (community) level.

Limitations & Future Work

Limitations.

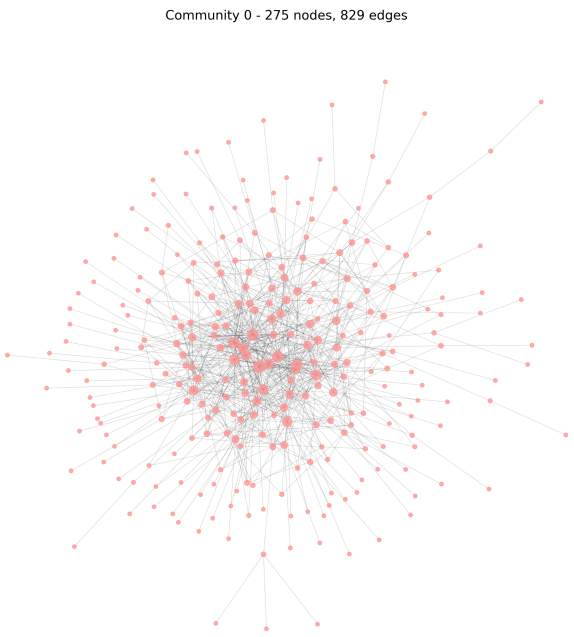

Fig. 5: Community 0 network visualization: A large, densely connected network with 275 nodes and 829 edges, displaying a core-periphery structure. The central region shows high interconnectivity among nodes (represented as pink circles), while peripheral nodes extend outward with fewer connections. This structure suggests a community with several influential central nodes that maintain cohesion within the network.

- *Graph disconnection.* High number of components (71813) can inflate modularity; future work should explore giant-component-only metrics.
- *GN cost.* Girvan–Newman refinement remains the runtime bottleneck. For very large communities, more scalable edge-sampling heuristics are needed.
- *Overfragmentation.* Infomap can over-split sparse regions, leading to high conductance in small clusters.

Future Directions.

- *Dynamic networks.* Extend to temporal snapshots for evolving community tracking.
- *Attributed graphs.* Incorporate node features (e.g. user interests) via joint embedding+community detection.
- *Overlapping communities.* Adapt link-clustering or fuzzy Infomap to detect overlapping memberships.
- *Deep learning.* Seed graph neural networks with our hybrid labels for semi-supervised refinement.

Conclusion

In this study, we proposed a hybrid community detection pipeline that combines Louvain, Girvan–Newman (GN), and Infomap algorithms in a modular and sequential fashion. This approach was designed to overcome the limitations of each individual method by leveraging their complementary strengths: Louvain provides scalable modularity-based coarsening, Girvan–Newman enables structural boundary refinement, and Infomap contributes flow-sensitive partitioning based on random walk dynamics.

Experimental results on the LiveJournal network demonstrated that the hybrid model consistently outperformed standalone algorithms across structural and semantic metrics. Specifically, it achieved the highest modularity and lowest conductance among all methods while maintaining high Normalized Mutual Information (NMI) with respect to ground-truth communities.

Standalone Louvain, while efficient and effective in optimizing modularity, suffers from the resolution limit problem, often merging smaller but meaningful flow-driven communities into larger ones. Girvan–Newman, although theoretically capable of exposing hierarchical community boundaries, is computationally intractable on large graphs due to its $O(n^3)$ complexity, making it impractical without an initial coarsening step. Infomap excels in capturing dynamic, information-based groupings but tends to over-fragment sparse regions and does not explicitly optimize modularity.

The proposed hybrid pipeline addresses these issues by executing a three-phase strategy: coarsening with Louvain, refining with GN, and flow-based tuning with Infomap. This combination yields communities that are both structurally dense and semantically coherent, balancing modularity and flow-based information criteria. The result is a more robust and interpretable community structure that reflects both the topological and functional properties of real-world social networks.

Future work may explore extending this pipeline to temporal networks for dynamic community tracking, incorporating deep learning-based embeddings, or applying it to domain-specific datasets such as citation graphs or biological interaction networks.

References

[1] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

[2] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Natl. Acad. Sci. USA*, vol. 99, no. 12, pp. 7821–7826, 2002.

[3] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proc. Natl. Acad. Sci. USA*, vol. 105, no. 4, pp. 1118–1123, 2008.

[4] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford Large Network Dataset Collection," http://snap.stanford.edu/data, Jun. 2014.

[5] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, 1998.

[6] M. E. J. Newman, "Modularity and community structure in networks," *Proc. Natl. Acad. Sci. USA*, vol. 103, no. 23, pp. 8577–8582, 2006.