Summary of Graph Clustering using Deep Modularity Networks (DMoN)

November 17, 2024

Introduction

Graph clustering is a key unsupervised learning task focused on identifying clusters within graph-structured data. This summary reviews the approach by Tsitsulin et al. (2023), which presents the Deep Modularity Networks (DMoN) model, designed to enhance clustering via Graph Neural Networks (GNNs) by directly optimizing modularity in a differentiable framework.

1. Key Scientific Question

The research addresses a fundamental question in machine learning and graph theory: how to improve graph clustering using GNNs, particularly in handling complex, sparse graphs. Traditional clustering methods like spectral clustering and k-means are less effective with high-dimensional or noisy graph data. DMoN leverages GNNs to improve clustering by optimizing modularity—a metric assessing the strength of network divisions directly, aiming to capture the community structure within graphs.

2. Contributions and Novelty

DMoN's main contributions include:

- An unsupervised, end-to-end GNN clustering approach through modularity optimization.
- A unique objective integrating modularity directly within the GNN framework, improving clustering quality.
- The introduction of "collapse regularization" to avoid trivial clustering solutions (i.e., all nodes in a single cluster) and maintain balanced clusters, even in sparse and noisy data contexts.

DMoN is designed to improve on limitations found in other GNN-based pooling techniques, notably DiffPool and MinCutPool, which have high computational costs and may struggle on sparse graphs.

3. Existing Clustering Methods and Limitations

Traditional clustering (spectral clustering, k-means) and GNN-based methods (DiffPool, MinCutPool) face challenges with scalability, stability, and accurately capturing community structures:

- **Spectral Clustering:** Projects graphs to a lower-dimensional space, suitable for smaller graphs but computationally expensive for larger, sparse ones.
- **Graph Embedding with k-means:** Uses embeddings for clustering, but may ignore graph topology and struggle with sparse, complex graphs.
- **GNN-based Pooling (DiffPool, MinCutPool):** Integrates node features and graph topology but has limitations in unsupervised settings, with DiffPool requiring complex multi-step optimization and MinCutPool facing convergence issues on irregular graphs.

DMoN directly addresses these limitations by optimizing modularity within a GNN, providing a scalable, robust framework suited for diverse graph types.

4. Technical Approach

DMoN incorporates modularity as a differentiable objective within the GNN, aiming to enhance the clustering quality by capturing community structures. Its modularity optimization uses an end-to-end differentiable framework, setting it apart from multi-step models. Additionally, DMoN introduces a collapse regularization term that balances cluster sizes by penalizing unbalanced cluster assignments. This term is calculated via the Frobenius norm on the sum of the cluster assignment matrix, ensuring stability across varied datasets.

5. Experimental Results

DMoN demonstrates superior performance in modularity and clustering accuracy, outperforming DiffPool and MinCutPool on synthetic and real-world datasets. Experimental results show that:

- DMoN achieves up to 40% improvement in modularity, indicating better community structure detection.
- Its robustness and generalizability across diverse datasets, including sparse and noisy networks, make it a promising method for real-world clustering.

- Performance remains stable even under challenging conditions such as sparse connectivity and high noise levels in node attributes.

6. Strengths and Weaknesses

Strengths: DMoN's strengths lie in its robustness, modularity-based optimization that aligns with clustering goals, and collapse regularization that ensures meaningful cluster sizes. These features make it effective for complex graph structures.

Weaknesses: The modularity optimization approach, although efficient, may encounter scalability challenges on extremely large graphs. Additionally, skewed degree distributions can affect the clustering balance.

7. Implementation and Experimental Verification

To further evaluate the effectiveness of DMoN, we implemented the model following the approach outlined by Tsitsulin et al. This allowed us to conduct our own experiments on synthetic and real-world datasets to verify the results reported in the paper.

Synthetic Data Experiments

We first tested DMoN on synthetic data generated using the Stochastic Block Model. The parameters for this random graph were as follows:

- Number of nodes: 300
- Number of clusters: 4
- Probability of intra-cluster links: `[0.2, 0.3, 0.1, 0.4]`
- Probability of inter-cluster links (q): 0.001

After generating the graph, we applied the DMoN model to clusterize it. The results are shown in the following plots:

- **Convergence Plot**: The evolution of the loss and evaluation metrics over epochs indicates that the model takes approximately 5000 epochs to converge on this synthetic data.

Real-World Data Experiments

We also evaluated DMoN on a real-world dataset, using the widely studied Cora dataset. The following results were obtained:

- **Convergence Plot**: Similar to the synthetic data, we recorded the evolution of loss and metrics with epochs on the Cora dataset, observing the convergence behavior.

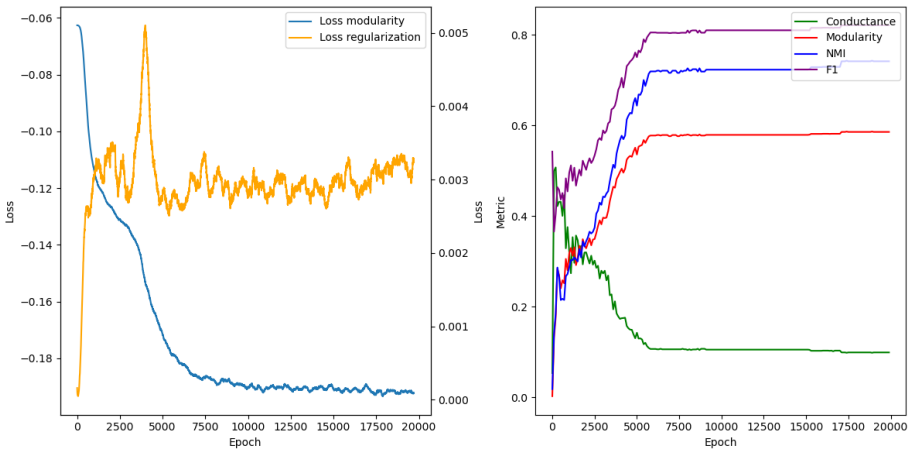

Figure 1: Loss and metrics evolution over epochs on synthetic data

- **Replication of Results**: We ran the model multiple times to evaluate the consistency and reliability of the clustering performance. Our experiments revealed that, while DMoN performs effectively on the Cora dataset, the original results reported by Tsitsulin et al. seem somewhat overestimated. Our repeated trials yielded slightly lower metrics, indicating a potential variance in performance.

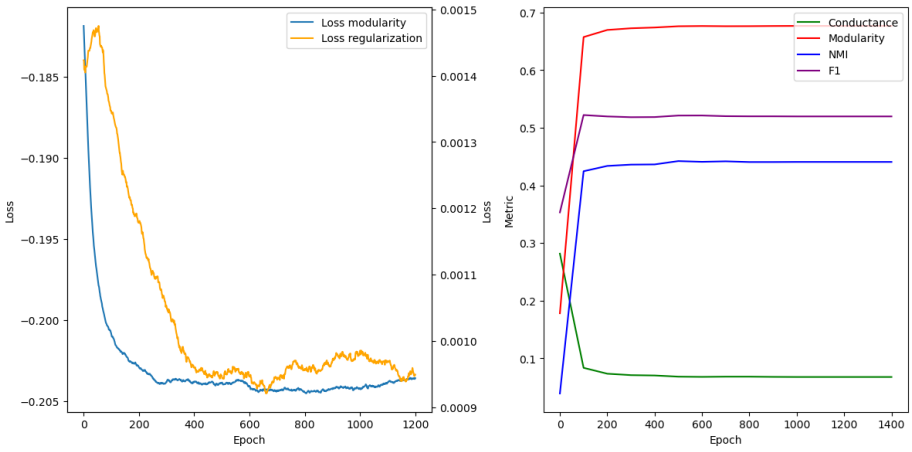

Figure 2: Loss and metrics evolution over epochs on real world data

These experiments confirm the overall effectiveness of DMoN in clustering but suggest that the performance may be more variable across different datasets and trials than initially reported.

8. Conclusion

DMoN provides a notable advancement in unsupervised graph clustering by directly optimizing modularity within a GNN. Its design overcomes challenges of traditional clustering methods and GNN-based pooling techniques, making it a valuable tool for diverse clustering applications. Further work may focus on optimizing computational efficiency and addressing limitations in scalability to broaden DMoN's applicability in large-scale graph analysis.