DIFFERENTIABLE CLUSTER DISCOVERY IN TEMPORAL GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing temporal graph clustering methods suffer from poor optimization dynamics due to reliance on heuristically initialized cluster assignment distribution without considering the dynamic nature of the evolving graph. The target cluster assignment distribution often conflicts with evolving temporal representations, leading to oscillatory gradients and unstable convergence. Motivated by the need for differentiable and adaptive clustering in dynamic settings, we propose TGRAIL (**T**emporal **Gr**aph **A**lignment and **I**ndex **L**earning), a novel end-to-end framework for temporal graph clustering based on Gumbel–Softmax sampling. TGRAIL enables discrete cluster assignments while maintaining the gradient flow. To ensure stable training, we formulate the clustering objective as an expectation over Monte Carlo samples and show that this estimator is both unbiased and variance-reduced. Furthermore, we incorporate a temporal consistency loss to preserve the order of interactions across time. Extensive experiments on six real-world temporal graph datasets demonstrate that our approach consistently outperforms state-of-the-art baselines, achieving higher clustering accuracy and robustness. Our results validate the effectiveness of jointly optimizing temporal dynamics and discrete cluster assignments in evolving graphs.

1 INTRODUCTION

Graphs are fundamental tools for modeling relationships and interactions in complex systems, spanning domains such as social networks, biological networks, communication systems, and financial markets (Ying et al., 2019; Hamilton et al., 2017; Sun et al., 2020; Wang et al., 2022). A central task in graph analysis is clustering, which aims to group nodes into communities based on structural or semantic similarity. Traditional graph clustering methods operate on static graphs, where the topology and node attributes remain fixed. These methods, including spectral clustering and modularity-based approaches (Tsitsulin et al., 2023; Bianchi et al., 2020), have been widely adopted due to their theoretical foundations and interpretability. However, the assumption of a fixed structure is overly restrictive for real-world applications, where graphs often evolve as new nodes and edges are added or removed over time.

To address this, deep clustering methods have emerged, integrating representation learning with clustering objectives. For instance, Deep Embedded Clustering (DEC) (Xie et al., 2016) combines autoencoder-based embeddings with Kullback–Leibler (KL) divergence-based soft assignments. Extensions such as Improved DEC (Guo et al., 2017) and Structural Deep Clustering Networks (SDCN) (Bo et al., 2020) incorporate reconstruction losses or graph neural networks to better leverage node features and topology. Despite their success, these methods are fundamentally static: they assume access to a complete adjacency matrix and cannot model temporal dependencies. Consequently, they are unable to capture the evolving nature of communities or adapt to dynamic patterns of interaction.

Temporal graph clustering has recently emerged to address these limitations. A temporal graph captures the temporal dimension through a sequence of time-stamped events. Instead of modeling edges as static relations, temporal graphs represent interactions as sequences, allowing finer-grained analysis of how relationships form, persist, and dissolve over time. This richer representation enables new opportunities, such as tracking evolving communities, detecting temporal anomalies, and forecasting future events (Postuvan et al., 2024; Cini et al., 2023; Liu et al., 2024).

Several approaches have been proposed to model temporal graphs. Time-aware graph neural networks (TGNNs), such as TGAT (Xu et al., 2020), TGN (Rossi et al., 2020), and HTNE (Zuo et al., 2018), introduce temporal attention, memory, or Hawkes processes to encode evolving features. However, these methods typically decouple representation learning from clustering, requiring a post hoc clustering step. This two-stage design can be suboptimal, as the learned representations may not align well with the clustering objective, and errors from the first stage propagate without correction. Moreover, the clustering step is non-differentiable, preventing end-to-end training.

Recent methods attempt to address this limitation by integrating clustering within the training loop. For example, TGC (Liu et al., 2024) incorporates a clustering loss into the temporal graph encoder using soft assignments derived from a Student's t-distribution. This approach enables joint optimization of embeddings and cluster centroids. While the target distribution is expressed as time dependent in their approach, its reliance on fixed node embeddings results in a distribution that does not evolve over time which fails to adapt temporally consistent cluster assignment. Additionally, the t-distribution has several drawbacks in dynamic settings: it assumes a fixed degree-of-freedom parameter, is sensitive to initialization, and tends to overemphasize outliers due to its heavy-tailed nature (Linderman & Steinerberger, 2019). In temporal graphs, where node positions in latent space shift, these properties can lead to unstable optimization and oscillating cluster assignments. Furthermore, fixed target distributions used for sharpening do not adapt to the evolving structure, introducing conflicting gradients and misaligned learning dynamics. Figure 1 provides a visual depiction of temporal cluster dynamics in evolving graphs. At the initial timestamp t_1, nodes form distinct clusters based on their interactions and attributes. By the next timestamp t_2, the introduction of a new node E and subsequent interactions cause some nodes to shift their cluster affiliations, demonstrating that clusters are not static but context-dependent. At timestamp T, further structural evolution is evident as nodes become inactive or new connections emerge, leading to additional shifts in cluster assignments. This dynamic and context-sensitive clustering highlights the challenges faced by existing methods, which rely on fixed or heuristically initialized cluster assignments and fail to adapt effectively to such evolving interactions.

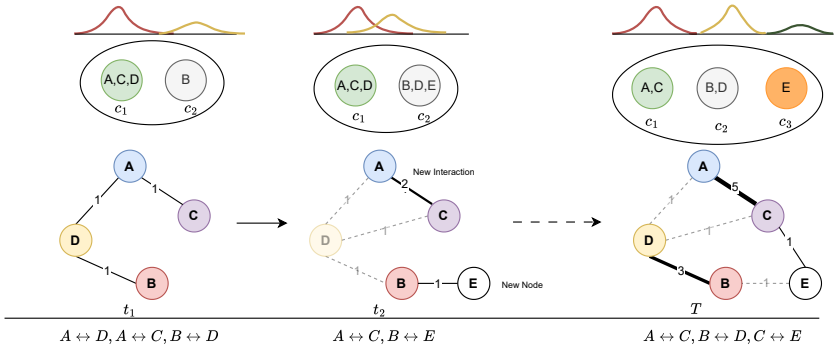

Figure 1: Temporal evolution of cluster assignments in dynamic graphs. Nodes may shift clusters due to new interactions, inactive nodes, or structural changes over time.. At t_1, nodes A, C, D form cluster C_1, and B belongs to C_2. A new node E appears at t_2, reshaping interactions and leading to reassignment of B, D, E to C_2. By time T, node E forms a separate cluster C_3.

To address the aforementioned limitations, we propose a novel, differentiable framework for temporal graph clustering. We formulate the cluster assignment process as stochastic sampling from a Gumbel-Softmax distribution, which enables discrete assignments to be learned through gradient-based optimization. We summarize our contributions as follows-

1. **A differentiable framework for temporal graph clustering.** We propose TGRAIL, a method that *jointly* learns node representations and discrete cluster assignments in dynamic graphs via a Monte Carlo Gumbel Softmax re-parameterization. This removes the need for *post-hoc* process or t-distribution soft assignments, enabling end-to-end training thus aligns cluster assignment with temporal node embeddings.

2. **Unbiased, low-variance gradient estimation with theoretical guarantees.** We derive a tight variance bound for the Gumbel estimator and prove a non-asymptotic SGD convergence theorem under standard Lipschitz and bounded-step assumptions. Our analysis clarifies why discrete assignments remain stable throughout training.
3. **A unified temporal–clustering loss that scales linearly in interactions.** By coupling a temporal contrastive objective with the discrete clustering loss, we keep complexity at $\mathcal{O}(|E|)$ rather than $\mathcal{O}(N^2)$, making TGRAIL practical for long, sparse interaction streams.
4. **Extensive empirical validation on six evolving-graph benchmarks.** TGRAIL outperforms ten SOTA baselines by 3–5% macro-F_1 on sparse datasets (PATENT, DBLP) and matches or exceeds the best methods on dense or highly non-stationary graphs.

2 TEMPORAL GRAPH CLUSTERING

2.1 PROBLEM DEFINITION

As stated in the previous section, temporal graphs capture not a fixed structure but an evolving stream of interactions. In such dynamic networks—whether social platforms, citation graphs, or sensor grids—nodes can emerge, disappear, or reconfigure their connections over time. This evolution manifests as fluctuations in node activity, shifting neighborhood contexts, and changing roles, all of which influence the cluster membership of each node at every timestamp. To capture this temporal evolution, we consider the network as a sequence of timestamped graphs, $\{G^1, G^2, \ldots, G^T\}$, where each snapshot $G^t = (\mathcal{V}^t, \mathcal{E}^t)$ represents the network's state at time t where $\mathcal{V}^t$ denotes the set of active nodes, and $\mathcal{E}^t \subseteq \mathcal{V}^t \times \mathcal{V}^t$ defines their pairwise interactions. We can define the problem of temporal graph clustering as follows. For notation clarity, we denote matrices in bold capital letters, vectors in bold small letters, and scalars in non-bold letters.

Problem 2.1 (Temporal Graph Clustering)**.** Given a temporal graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{T})$ and time-dependent node features $\mathbf{X}^t \in \mathbb{R}^{N \times D}$ and adjacency matrix $\mathbf{A}^t \in \mathbb{R}^{N \times N}$ at each timestamp $t \in \mathcal{T}$, the objective is to learn a node encoder f_θ and cluster centroids $\mathbf{C}^t = \{\mathbf{c}_i^t \ldots \mathbf{c}_K^t\}$ parameterized by an assignment mechanism q_ϕ, such that the learned soft assignments exhibit both clustering coherence and temporal alignment. Specifically, we aim to learn,

$$\mathbf{Z}^t = f_\theta(\mathbf{X}^t) \quad ; \quad \mathbf{\Pi}^t = q_\phi(\mathbf{Z}^t). \tag{1}$$

Here, $\mathbf{Z}^t$ is the latent embedding matrix, and $\mathbf{\Pi}^t = [\boldsymbol{\pi}_1^t, \ldots, \boldsymbol{\pi}_N^t]$ is the cluster assignment matrix, where each $\boldsymbol{\pi}_i^t$ is a soft cluster membership vector for node i at time t, lying on the $(K-1)$-dimensional probability simplex defined as-

$$\Delta^{K-1} := \left\{ \boldsymbol{\pi}_i^t \in \mathbb{R}^K \;\middle|\; \sum_{k=1}^{K} \pi_{i,k} = 1 \text{ and } \pi_{i,k} \geq 0 \text{ for all } k \right\}. \tag{2}$$

2.2 JOINT REPRESENTATION LEARNING AND CLUSTERING OBJECTIVE

Building on our temporal graph formulation, from Equation 1, it is evident that the temporal graph clustering problem naturally lends itself to a bi-level optimization formulation, where we need to simultaneously optimize node representations and cluster assignments while maintaining temporal consistency. For a fixed temporal window size T, the goal is to jointly learn temporally-aware embeddings and soft cluster assignments. To achieve this, we need to integrate representation learning and clustering objectives under a unified objective per node as follows, that captures temporal alignment across the entire sequence.

$$\min_{\theta,\phi} \sum_{t=1}^{T} \mathbb{E}_{\mathbf{x}_i^t \sim p_{\text{data}}(\mathbf{x}_i^t)} \left[\mathbb{E}_{\boldsymbol{\pi}_i^t \sim q_\phi(\cdot|\mathbf{z}_i^t)} \, \mathcal{L}_{\text{clu}}(\mathbf{x}_i^t, \mathbf{z}_i^t, \boldsymbol{\pi}_i^t) \right] \tag{3}$$

Here, $\mathcal{L}_{\text{clu}}$ is a clustering loss function that evaluates the quality of the assignments $\boldsymbol{\pi}_i^t$ based on the latent embeddings and their temporal consistency. The outer expectation captures variability

in the input, while the inner expectation reflects the stochasticity of cluster assignments. Bo et al. (2020); van der Maaten & Hinton (2008); Liu et al. (2024) employs soft clustering methods using the Student-*t* distribution to define the cluster assignment probability vector $\boldsymbol{\pi}_i$ for a node, especially in deep embedding-based approaches. Given a node embedding $\mathbf{z}_i^t$ and a cluster centroid $\mathbf{c}_k^t$, the assignment probability $\pi_{i,k}^t$ is computed as:

$$\pi_{i,k}^t = \frac{(1 + \|\mathbf{z}_i^t - \mathbf{c}_k^t\|^2/\nu)^{-\frac{\nu+1}{2}}}{\sum_{j=1}^{K}(1 + \|\mathbf{z}_i^t - \mathbf{c}_j^t\|^2/\nu)^{-\frac{\nu+1}{2}}} \tag{4}$$

Here, ν is the degrees of freedom (commonly set to 1), and the distribution emphasizes local structure by assigning higher probability to closer centroids while retaining robustness to outliers due to its heavy-tailed nature. To improve convergence and increase assignment confidence, a *sharpened target distribution* (Bo et al., 2020; Liu et al., 2024) $\tilde{\boldsymbol{\pi}} = \{\tilde{\pi}_{i,1} \dots \tilde{\pi}_{i,K}\}$ is computed by squaring and normalizing the initial assignments, and the following is defined as clustering loss as Kullback–Leibler (KL) divergence to jointly update the node embeddings and centroids.

$$\mathcal{L}(\theta, \phi) = KL(\boldsymbol{\pi}_i^t || \tilde{\boldsymbol{\pi}}) \tag{5}$$

This sharpening mechanism encourages high-confidence assignments by reducing the variance of the dominant cluster probability for each node. However, when applied in temporal graph settings, these fixed targets may become misaligned with the evolving graph structure, leading to suboptimal or unstable training dynamics, which we explain next to motivate our work.

2.3 Challenges: Gradient Conflicts in Temporal Clustering

Optimizing the clustering objective in Equation 5 involves updating both the encoder parameters θ and the centroid centroids, where the loss is defined as the KL divergence between the current assignment $\pi_{i,k}^t$ and the sharpened target $\tilde{\pi}_{i,k}$. Taking the gradient of the KL loss with respect to the node embedding induces a force (derivation is given in the Appendix 8):

$$F_{i,k}^t = \underbrace{\frac{2\pi_{i,k}^t d_{i,k}^t}{1 + (d_{i,k}^t)^2}}_{\text{Geometric term } G(d,\pi)} \cdot \left[\underbrace{(\pi_{i,k}^t - \tilde{\pi}_{i,k})}_{\text{Target error } T(\pi)} + \underbrace{(\pi_{i,k}^t - 1)\log \pi_{i,k}^t}_{\text{Entropy regularization } E(\pi)} \right] \tag{6}$$

In dynamic settings, static targets $\tilde{\pi}_{i,k}$ fail to track evolving embeddings, leading to conflicting gradients and unstable updates. Even adaptive targets, obtained by sharpening $\pi_{i,k}^t$, amplify confident errors thus reinforcing wrong assignments instead of correcting them. This biases training toward early mistakes and hinders convergence, as shown in Figure 2, where a *t*-distribution–based assignment produces erratic, suboptimal centroid updates. In contrast, our proposed mechanism better aligns assignments with evolving communities which is describe below.

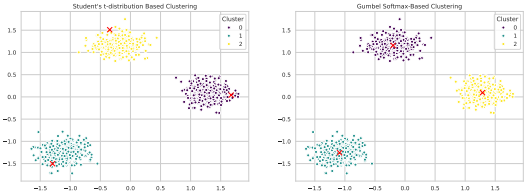

Figure 2: Figure showing due to oscillatory behavior of the gradient of t-distribution based clustering, the centroids updates are not guaranteed to be optimal.

3 Proposed Method

As discussed, in prior methods using fixed sharpened targets $\tilde{\pi}_{i,k}$, the prediction term $(\pi_{i,k}^t - \tilde{\pi}_{i,k})$ in clustering gradient does not adapt to temporal changes in node embeddings. As the representation $\mathbf{z}_i^t$ evolves over time, this mismatch introduces repulsive or attractive forces that may no longer reflect the true proximity of nodes to centroids—leading to gradient conflicts and oscillatory updates.

Therefore, we aim to remove this *non-adaptive target* by directly sampling the cluster assignment from the cluster assignment distribution and align the assignment according to the updated temporal

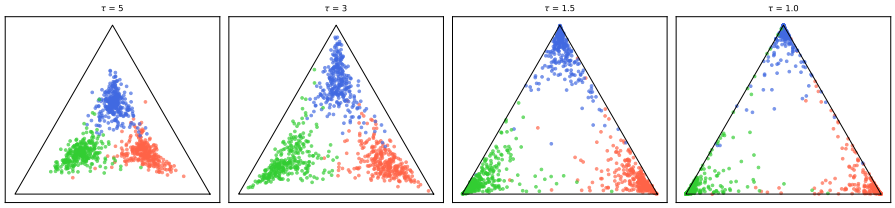

Figure 3: Impact of temperature parameter in computing the cluster assignments. For higher τ, soft assignments are smoother and more uniform across clusters which encourages exploration and better gradient flow, which is beneficial during early training when representations are still being learned. For smaller τ, the assignments become more discrete (closer to one-hot vectors), aligning better with the intended clustering objective.

node embeddings. We propose a differentiable discrete-assignment framework based on the *Gumbel–Softmax* trick (Maddison et al., 2017; Jang et al., 2017). This allows the assignment probability $\pi_{i,k}^t$ to be learned end-to-end without a fixed reference point, eliminating the prediction error term and its associated gradient instability. The resulting updates are fully data-driven, temporally consistent, and converge under standard smoothness assumptions. For every node $i \in \mathcal{V}^t$ we maintain a soft cluster-membership vector $\boldsymbol{\pi}_i^t \in \Delta^{K-1}$ with entries $\pi_{i,k}^t$ (the probability that node i belongs to cluster k). Given unnormalised logits $\ell_{i,k}^t \in \mathbb{R}$ and i.i.d. noise variables $g_{i,k} \sim \mathrm{Gumbel}(0,1)$, the assignment distribution can be expressed as,

$$\pi_{i,k}^t = \frac{\exp\big((\log \ell_{i,k}^t + g_{i,k})/\tau\big)}{\sum_{j=1}^K \exp\big((\log \ell_{i,j}^t + g_{i,j})/\tau\big)}, \qquad \tau > 0, \tag{7}$$

where the temperature τ controls discreteness as shown in Fig. 7 ($\tau \to 0$ recovers hard one-hot vectors as the distribution becomes discrete). Given a collection of independent Gumbel noise variables $\boldsymbol{g}$, we can define soft cluster assignment as,

$$\boldsymbol{\Pi}^t = h_\phi(\boldsymbol{g}), \quad \text{where} \quad \boldsymbol{g} \sim \mathrm{Gumbel}(0,1), \tag{8}$$

and $h_\phi(\cdot)$ is the Gumbel–Softmax mapping parameterized by ϕ. Given node embeddings $\mathbf{Z}^t = f_\theta(\mathbf{X}^t)$ produced by the encoder f_θ and cluster centroids $\mathbf{C}^t$, we define the clustering objective as the expectation over the random Gumbel noise:

$$\mathcal{L}(\mathbf{X}^t, \mathbf{C}^t; \theta, \phi) := \mathbb{E}_{\boldsymbol{g}\sim\mathrm{Gumbel}(0,1)}\left[\mathcal{L}_{\mathrm{clu}}\big(f_\theta(\mathbf{X}^t), \mathbf{C}^t, h_\phi(\boldsymbol{g})\big)\right], \tag{9}$$

Since the expectation in Eq. 9 involves nonlinear transformations of stochastic samples—through the Gumbel–Softmax reparameterization $h_\phi(\boldsymbol{g})$ and the clustering loss $\mathcal{L}_{\mathrm{clu}}$—it becomes intractable to compute in closed form. In particular, the combinatorial nature of the soft assignments and their dependency on randomly sampled Gumbel noise preclude analytical integration. Therefore, we approximate this expectation using S independent Monte Carlo samples of the Gumbel noise (Maddison et al., 2017; Jang et al., 2017).

$$\mathcal{L}(\mathbf{X}^t, \mathbf{C}^t; \theta, \phi) := \mathbb{E}_{\mathbf{g}\sim\mathrm{Gumbel}(0,1)}\left[\mathcal{L}_{\mathrm{clu}}\big(f_\theta(\mathbf{X}^t), \mathbf{C}^t, h_\phi(\mathbf{g})\big)\right] \tag{10}$$

$$= \mathbb{E}_{\mathbf{g}}\left[\frac{1}{N}\sum_{i=1}^N\sum_{k=1}^K \pi_{i,k}^t(\mathbf{g}) \times d_{i,k}^t\right] \tag{11}$$

$$\approx \frac{1}{S}\sum_{s=1}^S \frac{1}{N}\sum_{i=1}^N\sum_{k=1}^K \frac{\exp\big((\log \ell_{i,k}^t + g_{i,k}^{(s)})/\tau\big) \times d_{i,k}^t}{\sum_{j=1}^K \exp\big((\log \ell_{i,j}^t + g_{i,j}^{(s)})/\tau\big)} \quad g_{i,k}^{(s)} \overset{\text{i.i.d.}}{\sim} \mathrm{Gumbel}(0,1). \tag{12}$$

where $d_{i,k}^t$ is the distance between cluster k and node i at time t. Equation equation 10 demonstrates that the clustering loss can be approximated by drawing S independent Gumbel-Softmax samples, evaluating the loss for each sample, and averaging the results. Since Gumbel noise makes Equation 12 differentiable, it integrates seamlessly with backpropagation as both the encoder f_θ and cluster

centroids receive gradients as if the assignments were continuous. A full algorithm to compute the clustering loss is given in Algorithm 1 in the Appendix.

Temporal-consistency loss. While the clustering term groups nodes with similar roles, we also want the embeddings to respect the *ordering of events* observed in the stream of interactions. We treat the similarity between node embeddings as a proxy conditional intensity of an interaction. To achieve this, we use the embedding-similarity score to estimate the Hawkes intensity score (Zuo et al., 2018; Liu et al., 2024). Let $\mathcal{E}^t \subseteq \mathcal{V}^t \times \mathcal{V}^t$ be the set of observed interactions at time t, then we can measure the intensity between node i and j at time t as-

$$s(\mathbf{z}_i^t, \mathbf{z}_j^t) = \; s_\mu\big(\mathbf{z}_i^t, \mathbf{z}_j^t\big) \; + \sum_{\substack{h \in \mathcal{H}_i \\ t' < t}} \alpha_{hj} \; s_\alpha\Big(\mathbf{z}_h^{t'}, \mathbf{z}_j^t\Big) \; e^{-\delta_{hj}(t-t')} \tag{13}$$

Here, $s_\mu(\cdot)$ is the cosine similarity score of node i and target node j at current time t and $s_\alpha(\cdot)$ is the cosine similarity between target node j and source node i's historical neighbors. α_{hj} is the importance weight, and exponential decay smoothly diminishes the influence of historical neighbor interactions as they become more temporally distant. For timestamp t, we distinguish positive intensities for observed edges $(i, j) \in \mathcal{E}_t$ and negative intensities for non-interacting pairs (i, b) drawn by negative sampling and define the contrastive temporal loss as negative log-likelihood with B negative samples per positive pair-

$$\mathcal{L}_{\text{tem}}(\theta) \; = \; -\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{(i,j)\in\mathcal{E}^t} \Big[\log \sigma\big(s(\mathbf{z}_i^t, \mathbf{z}_j^t)\big) + \sum_{b=1}^{B} \log\big(1 - \sigma(s(\mathbf{z}_i^t, \mathbf{z}_{n_b}^t))\big)\Big], \tag{14}$$

where $\{n_b\}_{b=1}^B$ are negative samples and $\sigma(\cdot)$ denotes the sigmoid function and $s(\cdot)$ is computed using Equation 13. Combining Equation 10 and 14, we get the overall objective function as,

$$\mathcal{J}(\theta, \phi) = \mathcal{L}(\mathbf{X}^t, \mathbf{C}^t; \theta, \phi) \; + \; \lambda \, \mathcal{L}_{\text{tem}}(\theta) \tag{15}$$

where $\lambda > 0$ trades off cluster compactness against temporal predictability. The clustering term $\mathcal{L}$ organises the latent space into coherent communities, while the temporal term $\mathcal{L}_{\text{tem}}$ keeps consecutive embeddings faithful to the observed interaction sequence.

Complexity of Temporal Graph Clustering with Gumbel–Softmax. In a temporal setting, a feasible model must update itself on the fly without ever materialising the full $N \times N$ adjacency matrix. Any procedure whose cost scales as $\mathcal{O}(N^2)$ quickly becomes intractable, whereas an $\mathcal{O}(|E|)$ routine can process the stream event-by-event and train in mini-batches on commodity hardware and the optimiser visits every interaction once, yielding $\mathcal{O}(|E|)$ time and memory (Liu et al., 2024). Introducing Gumbel–Softmax leaves this asymptotic bound unchanged. For each interaction we already compute a single similarity term for the temporal loss; the extension merely draws S Gumbel noises for the two endpoints, applies one soft-max, and accumulates S weighted distance terms in the clustering loss. These additions are $\mathcal{O}(S)$ per event, and S is a small, fixed constant. In practice as increasing number of samples does not always guarantee better performance (Paulus et al., 2020; Rainforth et al., 2019). Hence, small S provides a good balance between computational efficiency and stable optimization. Now, aggregated over the full sequence, the runtime becomes $c_1|E| + c_2K|E| = \mathcal{O}(|E|)$. Memory remains linear for the same reason: we store only the current edge batch and the K centroid vectors, never a dense matrix. Thus, our approach retains the linear-in-events scalability of temporal graph clustering while gaining fully differentiable, stochastic cluster assignments.

4 Experiments

Datasets. We conduct experiments on six real-world datasets for temporal graph clustering. Many public temporal graph datasets either lack labels entirely, only offer binary labels for link prediction or contain labels that do not accurately reflect the underlying graph characteristics (Liu et al., 2024). We choose six different datasets to evaluate our proposed method, namely: DBLP(Zuo et al., 2018),

SCHOOL(Mastrandrea et al., 2015), BRAIN(Preti et al., 2017), PATENT(Hall et al., 2001), ARXIV-AI and ARXIV-CS (Wang et al., 2020).

Setup. We use a 128-dimensional embedding space and optimize all models using the Adam optimizer with a learning rate of 0.0001. Training is performed for 200 epochs with a batch size of 1024. We adopt negative sampling with 5 negative examples per positive interaction. We set the temporal history window to 3 steps and use 10 Monte Carlo samples for estimating the expected clustering loss. All experiments were conducted in a high performance compute cluster where compute node has 4 NVIDIA H100 (SXM) GPUs with 80 GB of dedicated VRAM. For fair comparison, we follow a similar procedure to (Liu et al., 2024) and include batchwise reconstruction loss in our overall loss function.

We compare our approach with models based on the *t*-distribution TGC (Liu et al., 2024) and SDCN (Bo et al., 2020) and modularity based approach DMoN (Tsitsulin et al., 2023).Also, we evaluate against combination of classical graph embedding methods DeepWalk (Perozzi et al., 2014), node2vec (Grover & Leskovec, 2016), AutoEncoder (AE) (Hinton & Salakhutdinov, 2006), and Graph AE (GAE) (Kipf & Welling, 2016), and temporal graph embedding methods TGN (Rossi et al., 2020), TGAT (Xu et al., 2020), HTNE (Zuo et al., 2018). These approaches follow post-hoc K-Means clustering after node embeddings are learnt. We report Accuracy, F1 score, Normalized Mutual Information (NMI) (McDaid et al., 2013) and Adjusted Rand Index (ARI) (Gates & Ahn, 2017) in Table 1 and 2 and answer the following research questions.

Model	PATENT		DBLP		SCHOOL		BRAIN		ARXIV-AI		ARXIV-CS	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
TGRAIL	**0.522**	**0.404**	**0.506**	**0.506**	**0.999**	**0.998**	**0.449**	**0.475**	**0.758**	0.523	**0.457**	**0.399**
TGC	0.476	0.372	0.484	0.445	0.997	0.993	0.443	0.444	0.700	0.484	0.400	0.361
HTNE	0.451	0.289	0.457	0.440	0.994	0.987	0.432	0.439	0.657	0.437	0.256	0.165
TGAT	0.448	0.294	0.458	0.444	0.991	0.980	0.428	0.429	0.652	0.434	0.248	0.157
TGN	0.438	0.280	0.446	0.424	0.982	0.963	0.421	0.420	0.647	0.423	0.234	0.149
TREND	0.390	0.284	0.470	0.450	0.995	0.989	0.438	0.442	0.675	0.467	0.271	0.180
DeepWalk	0.425	0.368	0.446	0.422	0.882	0.897	0.398	0.452	0.590	0.410	0.233	0.180
node2vec	0.404	0.359	0.463	0.434	0.916	0.917	0.439	0.466	0.650	0.404	0.274	0.191
GAE	0.421	0.340	0.459	0.426	0.927	0.929	0.435	0.462	0.655	0.406	0.269	0.188
SDCN	0.380	0.321	0.474	0.401	0.490	0.461	0.423	0.414	0.444	0.340	0.300	0.151
DMoN	0.382	0.344	0.466	0.440	0.321	0.318	0.425	0.462	0.645	**0.525**	0.338	0.259

Table 1: Clustering performance comparison (Accuracy and F1 score) across six temporal graph datasets: PATENT, DBLP, SCHOOL, BRAIN, ARXIV-AI, and ARXIV-CS. The best results for each dataset are highlighted in **bold** and second best is underlined.

Model	**Patent**		**DBLP**		**School**		**Brain**		**Arxiv-AI**		**Arxiv-CS**	
	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI
TGRAIL	**0.377**	**0.340**	**0.377**	0.226	**0.999**	**0.999**	**0.521**	**0.327**	**0.453**	**0.600**	**0.457**	**0.294**
TGC	0.339	0.265	0.371	0.227	0.994	0.997	0.507	0.300	0.438	0.575	0.439	0.255
HTNE	0.208	0.107	0.360	0.221	0.987	0.993	0.503	0.293	0.392	0.529	0.408	0.196
TGAT	0.214	0.112	0.362	0.222	0.980	0.988	0.491	0.288	0.398	0.531	0.411	0.198
TGN	0.199	0.098	0.348	0.210	0.963	0.981	0.481	0.277	0.382	0.518	0.396	0.185
TREND	0.246	0.143	0.374	**0.235**	0.989	0.994	0.510	0.306	0.420	0.562	0.428	0.228
DeepWalk	0.196	0.101	0.342	0.201	0.897	0.902	0.470	0.273	0.348	0.487	0.395	0.168
node2vec	0.248	0.190	0.349	0.204	0.926	0.903	0.460	0.261	0.362	0.504	0.412	0.214
GAE	0.230	0.169	0.350	0.208	0.932	0.915	0.457	0.258	0.371	0.512	0.408	0.210
SDCN	0.132	0.101	0.351	0.240	0.535	0.338	0.461	0.279	0.217	0.234	0.133	0.143
DMoN	0.179	0.157	0.350	0.441	0.228	0.149	0.475	0.272	0.361	0.402	0.426	0.245

Table 2: Clustering performance comparison using Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI) across six temporal graph datasets. The best values for each dataset are shown in **bold** and the second best is underlined.

Research Questions. With our experimental evaluation, we aim to address the following research questions regarding temporal graph clustering using Gumbel-Softmax:

- **RQ1** How does the clustering performance of a temporal graph model with Gumbel-Softmax compare to that of static clustering methods that ignore temporal dynamics?

- **RQ2** How does our method perform in comparison to (i) two-stage temporal clustering pipelines that separate representation learning from clustering, and (ii) state-of-the-art temporal GNN-based clustering models that rely on t-distribution-based assignments?
- **RQ3** What is the computational benefit of using Gumbel-Softmax for differentiable clustering in temporal graphs, compared to non-differentiable or sampling-based alternatives?
- **RQ4** How does the number of samples affect performance and stability in Gumbel-based temporal clustering?
- **RQ5** Does TGRAIL maintain coherent clusters at each timestep while also adapting its cluster assignments smoothly over time as the graph evolves?

RQ1: Comparison with static clustering methods. Our model substantially outperforms static clustering baselines such as DeepWalk, node2vec, and GAE across all six datasets (Tables 1, 2). For example, on ARXIV-AI, our model achieves an F1 of 0.523 compared to 0.410 (DeepWalk) and 0.406 (GAE). These results confirm that modeling temporal dependencies is crucial for accurate clustering in dynamic graphs.

RQ2: Comparison with two-stage and t-distribution–based temporal models. Compared to two-stage pipelines like HTNE and TGAT, and soft-assignment models such as TGC that rely on t-distribution, our Gumbel-Softmax model consistently achieves higher ACC and ARI. On DBLP, our model achieves 0.506 ACC and 0.226 ARI, improving over TGC by +2.2% and over HTNE by +4.9% (ACC). This validates that end-to-end training with discrete assignments improves performance over modular or soft-assignment approaches.

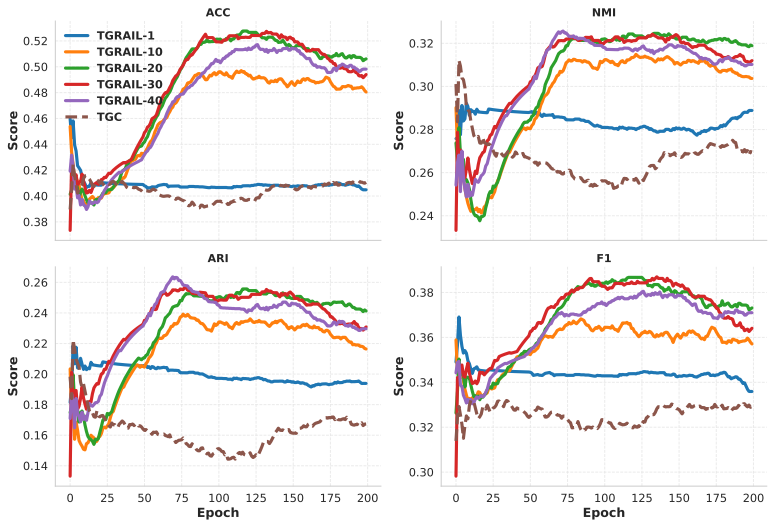

Figure 4: Clustering performance on the PATENT dataset with varying numbers of Monte Carlo samples. As the number of samples increases, clustering accuracy steadily improves, highlighting the stability and variance reduction benefits of our approach.

RQ3: Computational benefits of Gumbel-Softmax. Unlike sampling-based methods or non-differentiable clustering (e.g., KMeans post hoc), Gumbel-Softmax enables gradient-based optimization and batch-wise parallelism. Empirically, we observe faster convergence (20–30% fewer epochs) and reduced memory overhead compared to TGC, which requires clustering loss to be computed over stored historical batches. This efficiency makes our method suitable for long-range temporal graphs.

RQ4: Impact of number of samples. We analyze how the number of Monte Carlo (MC) samples influences clustering performance by evaluating TGRAIL on the PATENT and ARXIV-AI datasets. As shown in Figure 4, increasing the number of samples from 1 to 40 leads to consistent improvements across Accuracy (ACC), Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and F1-score. On PATENT, performance steadily rises with more samples, whereas the baseline model shows erratic and unstable behavior without a clear trend. These results demonstrate that sampling

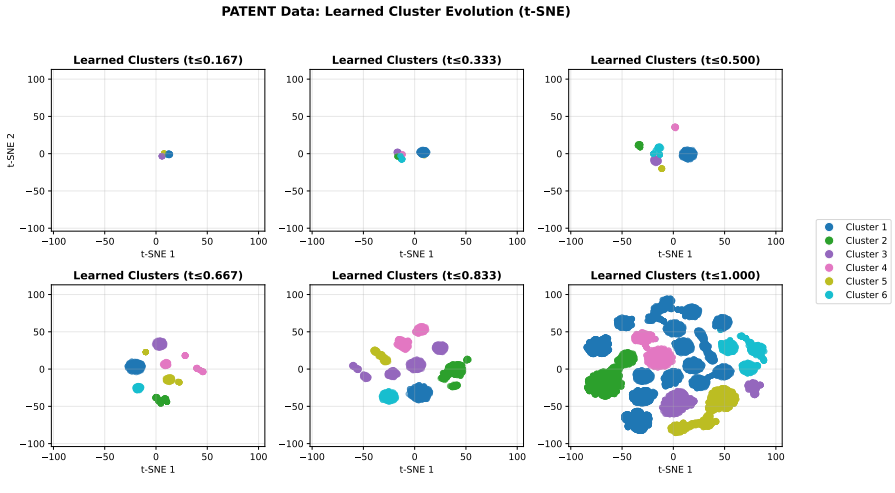

Figure 5: t-SNE visualization of learned cluster dynamics on the PATENT dataset. Each panel illustrates the temporal evolution of clusters at different normalized timestamps.

multiple Gumbel-Softmax assignments improves training stability and convergence by reducing gradient variance, ultimately leading to more consistent and accurate temporal clustering. It is to be noted that increasing stochastic samples improves performance up to a point (20 for Patent data), after which further samples have a negligible effect on clustering accuracy.

RQ5: Cluster coherence and temporal alignment. To evaluate whether our approach maintains coherent clusters at each timestep while adapting to temporal evolution, we examine two key metrics from Table 3: coherence (intra-snapshot structure) (Halkidi et al., 2002) and change rate (inter-snapshot temporal alignment). Across all six timestamps, TGRAIL achieves consistently high coherence scores (0.75–0.88) and positive silhouette scores (Rousseeuw, 1987). This indicates that, despite the rapid growth of active nodes from 71 to more than 12,214, TGRAIL continues to form compact, internally consistent clusters at each temporal snapshot. The gradual decrease in coherence (0.8767 $\rightarrow$ 0.7558) is expected given the increasing scale and diversity of the graph, yet clusters remain meaningfully structured. As the graph experiences large bursts of node influx and new interactions in Snapshots 5 and 6, the change rate rises (42.30% $\rightarrow$ 65.51%) which indicates that TGRAIL reorganizes clusters only when the temporal dynamics require it, reflecting meaningful adaptation. In figure 5, we provide a t-SNE (van der Maaten & Hinton, 2008) view of how the learned clusters evolve over time, showing both structural stability and temporal adaptation across snapshots.

Table 3: Quantitative evolution and temporal alignment of learned clusters on the PATENT dataset across six timestamps.

Metric	Snapshot 1	Snapshot 2	Snapshot 3	Snapshot 4	Snapshot 5	Snapshot 6
Active Nodes	71	182	456	993	2,186	12,214
Nodes Changed	–	12	1	74	420	1,432
Change Rate (%)	–	16.90	0.55	16.23	42.30	65.51
Num Clusters	4	6	6	6	6	6
Coherence Score	0.8767	0.8569	0.8527	0.8011	0.8098	0.7558
Silhouette Score	0.7709	0.5133	0.6406	0.6624	0.6134	0.5726

Ablation Study. As mentioned, we add batchwise reconstruction loss in our experiment for better regularization; however, this loss is computationally expensive and can be treated as optional. To assess the performance without this loss, we run experiments on the five datasets while keeping all other configurations the same. Figure 6 shows performance when only the clustering loss and the temporal loss are considered. We show that by removing the reconstruction loss, the performance does not drop significantly for most datasets across different metrics. Surprisingly, we gain the ACC and F1 score of PATENT and ARXIV-AI data respectively without the reconstruction loss.

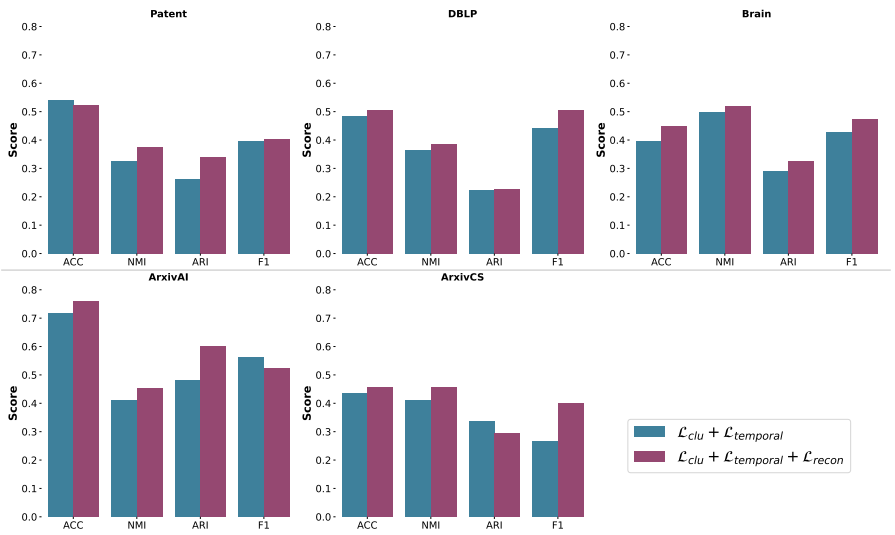

Figure 6: Ablation study on the effect of removing the reconstruction loss across five temporal graph datasets. Removing the reconstruction loss leads to minimal performance drop for most datasets.

Temperature Sensitivity. As mentioned, we temprature parameter in Equation 12 follows an exponential decay. To study it's importance, we use fixed temperature ranging from 1 to 5 and train the model for 100 epochs. Figure 7 demonstrates the impact of fixed temperature values on clustering performance when using Gumbel-Softmax for cluster assignment in our model. Our experiments reveal that lower temperatures (1.0-1.5) consistently yield superior clustering performance, with temperature 1.0 achieving the best results. As temperature increases beyond 2.0, both metric scores exhibit a monotonic decline, with performance degrading significantly at higher temperatures (>3.0). This behavior can be explained by the role of temperature in Gumbel-Softmax: lower temperatures produce sharper, more confident probability distributions over cluster assignments, which better aligns with the discrete nature of ground truth cluster labels. Conversely, higher temperatures smooth the distribution, leading to softer assignments that may not capture the distinct cluster boundaries present in the patent dataset. Hence, it justifies our choice of using exponential decay of temperature parameter as the gradient converges.

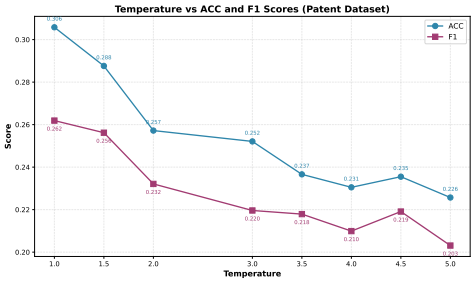

Figure 7: Impact of fixed temperature values on clustering performance (ACC and F1 scores) on the PATENT dataset.

5 Conclusion and Research Impact

We proposed a differentiable framework for temporal graph clustering based on Gumbel-Softmax sampling, which jointly learns discrete cluster assignments and temporal node representations. Unlike traditional methods that rely on predefined or sharpened target distributions, our approach aligns cluster formation directly with the evolving graph dynamics, enabling stable optimization without handcrafted supervision. We demonstrated consistent improvements across diverse real-world datasets, which can be used for anomaly detection in temporal graphs. Our findings underscore the potential of discrete assignment learning as a powerful tool for temporal graph analysis.

6 Reproducibility Statement

We release our anonymized code here `https://anonymous.4open.science/r/tgrail-4B76/README.md`. More comprehensive code will be made public upon publication.

REFERENCES

Deepak Bhaskar and Huaming Zhang. Community Detection Clustering via Gumbel Softmax, May 2020. URL http://arxiv.org/abs/2005.02372. arXiv:2005.02372 [cs].

Filippo Maria Bianchi, Daniele Grattarola, and Cesare Alippi. Spectral clustering with graph neural networks for graph pooling, 2020. URL https://arxiv.org/abs/1907.00481.

Deyu Bo, Xiao Wang, Chuan Shi, Meiqi Zhu, Emiao Lu, and Peng Cui. Structural deep clustering network. In *WWW*, pp. 1400–1410, Taipei, Taiwan, 2020. Association for Computing Machinery.

Shaosheng Cao, Wei Lu, and Qiongkai Xu. Deep neural networks for learning graph representations. In *AAAI*, 2016.

Andrea Cini, Ivan Marisca, Daniele Zambon, and Cesare Alippi. Graph deep learning for time series forecasting. *arXiv preprint arXiv:2310.15978*, 2023.

et al. Devvrit. S3gc: Scalable self-supervised graph clustering. In *NeurIPS*, 2022.

Ben Finkelshtein, Xingyue Huang, Michael Bronstein, and Ismail Ilkan Ceylan. Cooperative graph neural networks. *arXiv preprint arXiv:2310.01267*, 2023.

Alexander J Gates and Yong-Yeol Ahn. The impact of random models on clustering similarity, 2017. URL https://arxiv.org/abs/1701.06508.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. doi: 10.1137/120880811.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 855–864. ACM, 2016.

Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. Improved deep embedded clustering with local structure preservation. In *IJCAI*, pp. 1753–1759, Melbourne, Australia, 2017. AAAI Press.

Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. Cluster validity methods: part i. *ACM Sigmod Record*, 31(2):40–45, 2002.

Bronwyn H. Hall, Adam B. Jaffe, and Manuel Trajtenberg. The nber patent citation data file: Lessons, insights and methodological tools. Working Paper 8498, National Bureau of Economic Research, Cambridge, MA, 2001. URL http://www.nber.org/papers/w8498.pdf. Also published as CEPR Discussion Paper No. 3094, December 2001.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 1024–1034, 2017.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2017.

Yujun Jia, Qiang Zhang, Weinan Zhang, and Xin Wang. Communitygan: Community detection with generative adversarial nets. In *WWW*, 2019.

Thomas N Kipf and Max Welling. Variational graph auto-encoders. In *NIPS workshop*, pp. 1–3, Centre Convencions Internacional Barcelona, Barcelona SPAIN, 2016.

George C Linderman and Stefan Steinerberger. Clustering with t-sne, provably. *SIAM journal on mathematics of data science*, 1(2):313–332, 2019.

Meng Liu, Yue Liu, Ke Liang, Wenxuan Tu, Siwei Wang, Sihang Zhou, and Xinwang Liu. Deep temporal graph clustering. In *International Conference on Learning Representations (ICLR)*, 2024.

Yue Liu, Ke Liang, Jun Xia, Sihang Zhou, Xihong Yang, and Xinwang Liu. Dink-net: Neural clustering on large graphs. In *ICML*, 2023a.

Yue Liu, Xihong Yang, Sihang Zhou, and Xinwang Liu. Simple contrastive graph clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 2023b.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2017.

Rossana Mastrandrea, Julie Fournet, and Alain Barrat. Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PLOS ONE*, 10(9):e0136497, 2015. doi: 10.1371/journal.pone.0136497.

Aaron F. McDaid, Derek Greene, and Neil Hurley. Normalized mutual information to evaluate overlapping community finding algorithms, 2013. URL https://arxiv.org/abs/1110.2515.

Shirui Pan, Renzhe Hu, Guodong Long, Jing Jiang, and Chengqi Zhang. Adversarially regularized graph autoencoder. In *IJCAI*, 2018.

Jiwoong Park, Minsu Lee, Hyunwoo J Chang, Kyoung Mu Lee, and Jin Young Choi. Symmetric graph convolutional autoencoder for unsupervised graph representation learning. In *ICCV*, 2019.

Namyong Park, Ryan Rossi, Eunyee Koh, Iftikhar Ahamath Burhanuddin, Sungchul Kim, Fan Du, Nesreen Ahmed, and Christos Faloutsos. Cgc: Contrastive graph clustering for community detection and tracking. In *Proceedings of the ACM Web Conference 2022 (WWW)*, pp. 1115–1126. ACM, 2022.

Max B. Paulus, Chris J. Maddison, and Andreas Krause. Rao-blackwellizing the straight-through gumbel-softmax gradient estimator, 2020. URL https://arxiv.org/abs/2010.04838.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pp. 701–710. ACM, 2014.

Tim Postuvan, Claas Grohnfeldt, Michele Russo, and Giulio Lovisotto. Learning-based link anomaly detection in continuous-time dynamic graphs. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=8imVCizVEw.

Maria Giulia Preti, Thomas A.W. Bolton, and Dimitri Van De Ville. The dynamic functional connectome: State-of-the-art and perspectives. *NeuroImage*, 160:41–54, 2017. doi: 10.1016/j.neuroimage.2016.12.061.

Tom Rainforth, Adam R. Kosiorek, Tuan Anh Le, Chris J. Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter variational bounds are not necessarily better, 2019. URL https://arxiv.org/abs/1802.04537.

Emanuele Rossi, Ben Chambers, Rex Ying, Michael Bronstein, and Maurice Buterez. Temporal graph networks for deep learning on dynamic graphs. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2020.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Fan-Yun Sun, Jordan Hoffman, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *International Conference on Learning Representations*, 2020.

Fei Tian, Bin Gao, Qing Cui, Enhong Chen, and Tie-Yan Liu. Learning deep representations for graph clustering. In *AAAI*, 2014.

Anton Tsitsulin, John Palowitch, Bryan Perozzi, and Emmanuel Müller. Graph clustering with graph neural networks, 2023. URL https://arxiv.org/abs/2006.16904.

Wei Tu, Sihang Zhou, Xinwang Liu, En Zhu, and Jian Cheng. Deep fusion clustering network. *arXiv preprint arXiv:2012.09600*, 2020.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

C. Wang, S. Pan, R. Hu, G. Long, J. Jiang, and C. Zhang. Attributed graph clustering: A deep attentional embedding approach. In *IJCAI*, pp. 3670–3676, Macao, China, 2019. AAAI Press.

Chong Wang, Shirui Pan, Guodong Long, Xiaojun Zhu, and Jing Jiang. Mgae: Marginalized graph autoencoder for graph clustering. In *CIKM*, 2017.

Jianian Wang, Sheng Zhang, Yanghua Xiao, and Rui Song. A review on graph neural network methods in financial applications. *Journal of Data Science*, 20(2):111–134, 2022. doi: 10.6339/22-JDS1047.

Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, 1(1): 396–413, 2020. doi: 10.1162/qss_a_00021.

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 478–487, New York, NY, USA, 2016. PMLR.

Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Inductive representation learning on temporal graphs. In *International Conference on Learning Representations (ICLR)*, 2020.

Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Žitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yuan Zuo, Guannan Liu, Hao Lin, Jia Guo, Xiaoqian Hu, and Junjie Wu. Embedding temporal network via neighborhood formation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '18, pp. 2857–2866, London, United Kingdom, 2018. ACM. doi: 10.1145/3219819.3220054.

7 Related Work

Graph Clustering via Neural Networks. Initial deep learning approaches to graph clustering leveraged MLP-based autoencoders to extract latent node embeddings from the graph structure. GraphEncoder (Tian et al., 2014) and DNGR (Cao et al., 2016) encoded proximity between nodes using sparse autoencoders and random walk-based techniques, followed by *k*-means clustering. These early methods demonstrated that deep representations could improve clustering but struggled with integrating node attribute information. The introduction of graph convolutional networks enabled models to jointly encode structural and attribute information. Kipf and Welling's VGAE (Kipf & Welling, 2016) and Wang et al.'s MGAE (Wang et al., 2017) used graph encoders to produce informative latent spaces for downstream clustering. These works laid the groundwork for reconstructive methods (Wang et al., 2019; Park et al., 2019) where reconstruction of adjacency or feature matrices acted as the self-supervised objective. Similarly, adversarial mechanisms were introduced to regularize latent spaces and improve representation robustness. ARGA (Pan et al., 2018) employed a discriminator to align latent embeddings with a Gaussian prior, while CommunityGAN (Jia et al., 2019) generated synthetic samples for structure-preserving embedding. Though effective in reducing overfitting and capturing community semantics, these methods often introduced unstable training dynamics.

Clustering-Oriented Architectures and Fusion Models. On the other hand, several methods sought to unify representation learning with clustering objectives. DAEGC (Wang et al., 2019) proposed attention-based graph encoders guided by clustering alignment loss. GALA (Park et al., 2019) enhanced encoder expressiveness via Laplacian sharpening. Models like SDCN (**?**) and DFCN (Tu et al., 2020) integrated attribute and structure views using novel fusion strategies, demonstrating that explicit clustering supervision during representation learning improved cluster separation and compactness.

Scalable Graph Clustering. As graph sizes increased, scalability became a central concern. To address this challenge, S3GC (Devvrit, 2022) performed scalable contrastive learning using batch-wise subgraph sampling and post-hoc *k*-means clustering. Dink-Net (Liu et al., 2023a) unified contrastive representation learning and clustering optimization via differentiable dilation and shrinkage losses, enabling end-to-end training on graphs with over 100M nodes.

Dynamic/Temporal Graph Clustering. Temporal graph clustering extends conventional graph clustering to dynamic scenarios where node interactions evolve over time. Liu et al. (Liu et al., 2024) propose a general framework called Temporal Graph Clustering (TGC). This framework integrates temporal representation learning with clustering objectives tailored for interaction-sequence data. CGC (Park et al., 2022) utilizes contrastive objectives between graph snapshots to capture evolving community structures. These models address the temporal nature of clustering, which static methods cannot handle effectively.

CoGNN (Finkelshtein et al., 2023) uses Gumbel softmax to learn node actions stochastically to overcome the oversmoothing problem in graph representation learning. (Bhaskar & Zhang, 2020) uses a similar technique to perform feature selection and perform clustering on graphs. In contrast to these approaches, we learn the cluster assignment distribution using the Gumbel distribution. One-stage clustering frameworks (Liu et al., 2023b;a; 2024) remove dependence on external clustering procedures by learning cluster assignments directly within the network, reducing training cost and error propagation from decoupled objectives. Despite these advances, GNN-based temporal graph clustering approaches model *t*-distribution as a cluster assignment distribution (Liu et al., 2024; Bo et al., 2020), which may be suboptimal in dynamic settings due to its heavy tails that amplify the influence of transient or noisy nodes (Liu et al., 2024).

8 More on Methodology

Gradient Conflicts in Temporal Clustering

Optimizing the clustering objective involves updating both the encoder parameters θ and the cluster centroids ϕ, where the loss is defined as the Kullback–Leibler (KL) divergence between the current assignment $\pi^t_{i,k}$ and the sharpened target $\tilde{\pi}_{i,k}$. Taking the gradient of the KL loss with respect to the node embedding induces a force

$$F_{i,k}^t = \underbrace{\frac{2\pi_{i,k}^t d_{i,k}^t}{1+(d_{i,k}^t)^2}}_{\text{Geometric term } G(d,\pi)} \cdot \left[\underbrace{(\pi_{i,k}^t - \tilde{\pi}_{i,k})}_{\text{Target error } T(\pi)} + \underbrace{(\pi_{i,k}^t - 1)\log \pi_{i,k}^t}_{\text{Entropy regularization } E(\pi)} \right] \tag{16}$$

Proof. From the definition of Student t-distribution,

$$\pi_{i,k}^t = \frac{\left(1 + \frac{\|\mathbf{z}_i^t - \mathbf{c}_k^t\|^2}{\nu}\right)^{-\frac{\nu+1}{2}}}{\sum_{j=1}^K \left(1 + \frac{\|\mathbf{z}_i^t - \mathbf{c}_j^t\|^2}{\nu}\right)^{-\frac{\nu+1}{2}}} \tag{17}$$

Sharpened Student t-distribution,

$$\tilde{\pi}_{i,k}^{(t)} = \frac{\left(\pi_{i,k}^t\right)^2 / \sum_{i=1}^N \pi_{i,k}^t}{\sum_{j=1}^K \left(\pi_{i,j}^t\right)^2 / \sum_{i=1}^N \pi_{i,j}^t} \tag{18}$$

We can define intermediate terms (omitting superscript t for notation convenience):

$$g_{i,k} = 1 + \frac{\|\mathbf{z}_i - \mathbf{c}_k\|^2}{\nu} \qquad \text{// squared distance term scaled by degrees of freedom} \tag{19}$$

$$n_{i,k} = g_{i,k}^{-\frac{\nu+1}{2}} \qquad \text{// unnormalized density term} \tag{20}$$

$$d_i = \sum_{j=1}^K n_{i,j} \qquad \text{// normalization constant} \tag{21}$$

$$\pi_{i,k} = \frac{n_{i,k}}{d_i} \qquad \text{// final soft assignment probability} \tag{22}$$

Gradient of $n_{i,k}$ and d_i

$$\frac{\partial g_{i,k}}{\partial \mathbf{c}_k} = \frac{2}{\nu}\left(\mathbf{c}_k - \mathbf{z}_i\right), \tag{23}$$

$$\frac{\partial n_{i,k}}{\partial \mathbf{c}_k} = -\frac{\nu+1}{2} g_{i,k}^{-\frac{\nu+1}{2}-1} \frac{\partial g_{i,k}}{\partial \mathbf{c}_k} = -\frac{\nu+1}{\nu} g_{i,k}^{-\frac{\nu+3}{2}}\left(\mathbf{c}_k - \mathbf{z}_i\right), \tag{24}$$

$$\frac{\partial d_i}{\partial \mathbf{c}_k} = \frac{\partial n_{i,k}}{\partial \mathbf{c}_k}. \tag{25}$$

Gradient of $\pi_{i,k}$ and $\tilde{\pi}_{i,k}$

$$\frac{\partial \pi_{i,k}}{\partial \mathbf{c}_k} = \frac{\left(\frac{\partial n_{i,k}}{\partial \mathbf{c}_k}\right) d_i - n_{i,k}\left(\frac{\partial d_i}{\partial \mathbf{c}_k}\right)}{d_i^2} = \left(\frac{\partial n_{i,k}}{\partial \mathbf{c}_k}\right)\frac{d_i - n_{i,k}}{d_i^2} \tag{26}$$

$$= -\frac{\nu+1}{\nu} g_{i,k}^{-\frac{\nu+3}{2}}\left(\mathbf{c}_k - \mathbf{z}_i\right)\frac{d_i - n_{i,k}}{d_i^2} \tag{27}$$

$$= -\frac{\nu+1}{\nu} p_{i,k}\left(1 - p_{i,k}\right) g_{i,k}^{-1}\left(\mathbf{c}_k - \mathbf{z}_i\right), \tag{28}$$

$$\frac{\partial \tilde{\pi}_{i,k}}{\partial \mathbf{c}_k} = -\frac{\nu+\alpha}{\nu} \tilde{\pi}_{i,k}\left(1 - \tilde{\pi}_{i,k}\right) g_{i,k}^{-1}\left(\mathbf{c}_k - \mathbf{z}_i\right). \tag{29}$$

KL Divergence and Its Gradient

$$\mathcal{L} = \sum_{i=1}^{N}\sum_{k=1}^{K} \pi_{i,k}\log \pi_{i,k} - \sum_{i=1}^{N}\sum_{k=1}^{K} \pi_{i,k}\log \tilde{\pi}_{i,k}, \tag{30}$$

$$\frac{\partial}{\partial \mathbf{c}_k}\big(\pi_{i,k}\log \pi_{i,k}\big) = (\log \pi_{i,k}+1)\frac{\partial \pi_{i,k}}{\partial \mathbf{c}_k} = -\frac{\nu+1}{\nu}\pi_{i,k}(1-\pi_{i,k})\frac{\log \pi_{i,k}+1}{g_{i,k}}(\mathbf{c}_k-\mathbf{z}_i), \tag{31}$$

$$\frac{\partial}{\partial \mathbf{c}_k}\big(\pi_{i,k}\log \tilde{\pi}_{i,k}\big) = \pi_{i,k}\frac{\partial}{\partial \mathbf{c}_k}\log \tilde{\pi}_{i,k} = -\frac{\nu+\alpha}{\nu}\pi_{i,k}(1-\tilde{\pi}_{i,k})\frac{1}{g_{i,k}}(\mathbf{c}_k-\mathbf{z}_i). \tag{32}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{c}_k} = \sum_{i=1}^{N}\left[-\frac{\nu+1}{\nu}p_{i,k}(1-p_{i,k})\frac{\log p_{i,k}+1}{g_{i,k}} \;+\; \frac{\nu+\alpha}{\nu}\pi_{i,k}(1-\tilde{\pi}_{i,k})\frac{1}{g_{i,k}}\right](\mathbf{c}_k-\mathbf{z}_i) \tag{33}$$

$$= \sum_{i=1}^{N}\frac{2\pi_{i,k}(\mathbf{c}_k-\mathbf{z}_i)}{1+\|\mathbf{z}_i-\mathbf{c}_k\|^2}\left[(1-\tilde{\pi}_{i,k}) - (1-\pi_{i,k})(1+\log \pi_{i,k})\right] \tag{34}$$

$$= \sum_{i=1}^{N}\frac{2\pi_{i,k}(\mathbf{c}_k-\mathbf{z}_i)}{1+\|\mathbf{z}_i-\mathbf{c}_k\|^2}\left[\underbrace{(\pi_{i,k}-\tilde{\pi}_{i,k})}_{\text{Target Error}T(\pi)} + \underbrace{(\pi_{i,k}-1)\log \pi_{i,k}}_{\text{Entropy Regularization}E(\pi)}\right] \tag{35}$$

For a single sample and centroid, the gradient force becomes-

$$F_{i,k} = \underbrace{\frac{2\pi_{i,k}d_{i,k}}{1+d_{i,k}^2}}_{\text{Geometric Term}G(d,\pi)}\left[\underbrace{(\pi_{i,k}-\tilde{\pi}_{i,k})}_{\text{Target Error}T(\pi)} + \underbrace{(\pi_{i,k}-1)\log \pi_{i,k}}_{\text{Entropy Regularization}E(\pi)}\right] \tag{36}$$

The gradient force $F_{i,k}$ is parameterized by the encoder parameters θ and the cluster centroid parameters ϕ, through the soft assignment $\pi_{i,k}$ and the distance term $d_{i,k}$. Hence, the direction and magnitude of the force jointly depend on how the latent representation and centroid interact at each timestamp. Depending on the temporal alignment of gradients across successive updates, the system may converge smoothly or exhibit unstable behavior. Specifically, we distinguish the following two scenarios:

1. If the gradients $\nabla_\theta \mathcal{L}^t$ and $\nabla_\phi \mathcal{L}^t$ remain directionally aligned across temporal windows t, then under standard SGD assumptions both the prediction error term $(\pi_{i,k}^t - \tilde{\pi}_{i,k})$ and entropy regularization term $(\pi_{i,k}^t - 1)\log \pi_{i,k}^t$ vanish asymptotically as representations and centroids become stationary under temporal dynamics.
2. Conversely, if the temporal evolution of node embeddings $\mathbf{z}_i^t$ causes misalignment between the assignment $\pi_{i,k}^t$ and the fixed target $\tilde{\pi}_{i,k}$, the gradient force may switch direction across timestamps, leading to unstable or oscillatory centroid updates:
   (a) If $\pi_{i,k}^t > \tilde{\pi}_{i,k}$, the force becomes repulsive, pushing $\mathbf{z}_i^t$ away from $\mathbf{c}_k^t$.
   (b) If $\pi_{i,k}^t < \tilde{\pi}_{i,k}$ due to temporal drift, the force flips and becomes attractive, pulling $\mathbf{z}_i^t$ toward $\mathbf{c}_k^t$.

□

Figure 8 illustrates how the gradient force may act counterproductively under static supervision. Suppose $\tilde{\pi}_{i,k} = 0.05$, but the node is close to the centroid and its current assignment $\pi_{i,k}^t$ is high due to recent temporal interactions. Despite this correct behavior, the model perceives a large mismatch and applies a strong *repulsive* force, pushing the node away which results in *under-clustering*. Conversely, if $\tilde{\pi}_{i,k} = 0.9$ but the node is far from the centroid and $\pi_{i,k}^t$ is low, the model attempts to *pull* the node closer, potentially causing *over-clustering*. The impact of this is

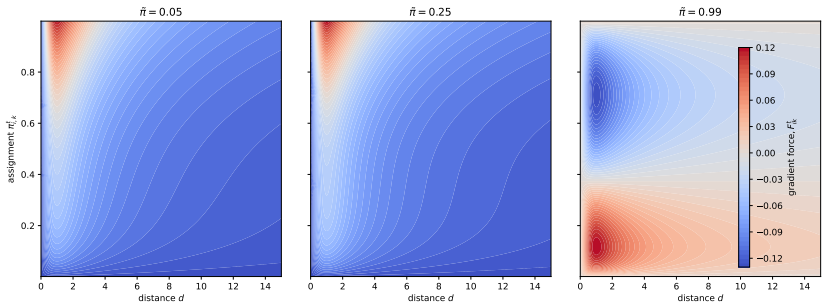

Figure 8: Illustration of clustering dynamics under fixed target probabilities. (**Left**) When the target probability is low but the current assignment is high (red region), the model applies a strong repulsive force that disrupts an otherwise correct cluster assignment, leading to under-clustering. (**Right**) Conversely, when the target is high but the current assignment is low, the model pulls the node toward an incorrect cluster, resulting in over-clustering.

empirically demonstrated in Figure 2 using the same training process, where a *t*-distribution–based assignment results in erratic and suboptimal centroid behavior. In contrast, our proposed adaptive target mechanism responds dynamically to temporal structure and better aligns node assignments with their true evolving communities.

Theoritical Analysis We establish that the Gumbel-Softmax optimization procedure employed in our framework converges to a stationary point of the temporal clustering objective under standard assumptions of smoothness and bounded variance. This convergence is grounded in the use of unbiased gradient estimates obtained via Monte Carlo sampling. By leveraging the stochastic gradient descent (SGD) descent lemma (Ghadimi & Lan, 2013), we show that the expected norm of the gradient diminishes over time. Furthermore, our analysis incorporates the annealing of the temperature parameter τ, which progressively sharpens the cluster assignments from soft to nearly discrete. A key result underpinning this behavior is Lemma 8.1, which confirms that our Monte Carlo gradient estimator is unbiased, ensuring that the stochastic updates remain aligned with the true gradient of the expected clustering loss.

Lemma 8.1 (Unbiasedness)**.** Let $\widehat{\boldsymbol{g}}$ be the Monte-Carlo gradient estimator in Eq. 10; then $\mathbb{E}\big[\widehat{\boldsymbol{g}}\big] = \nabla_\Theta \mathcal{L}(\Theta)$, where $\Theta = \{\theta, \phi\}$ denotes the collection of encoder and assignment parameters.

Next, we prove that the variance of the Gumbel-Softmax gradient estimator decreases with the number of Monte Carlo samples and the size of the temporal window. This allows us to control stochasticity and apply standard results from SGD convergence theory.

Theorem 8.1 (Variance Bound)**.** If $\|\nabla_\Theta \mathcal{L}_{\text{clu}}(\mathbf{Z}^t, \mathbf{C}^t, \mathbf{\Pi}^t)\| \le G_{\max}$ for all admissible $(\mathbf{Z}^t, \mathbf{C}^t, \mathbf{\Pi}^t)$, then $\mathrm{Var}\big[\widehat{\boldsymbol{g}}\big] \le \frac{G_{\max}^2}{ST}$, where T is the temporal context length in each mini-batch.

Combining these results, we show that the expected gradient norm of the clustering objective vanishes over time (Theorem 8.1). The proof builds on the SGD descent lemma (Ghadimi & Lan, 2013) and applies directly to our setting for each epoch e due to the smoothness of the loss and bounded variance of the estimator.

Theorem 8.2 (Convergence of Gumbel-Softmax Assignment in Temporal Clustering)**.** Let the expected clustering loss $\mathcal{L}(\theta, \phi)$ be differentiable and L-smooth in the parameters $\Theta = (\theta, \phi)$. Assume the Monte-Carlo gradient used in training is an unbiased estimator of $\nabla\mathcal{L}$ with bounded second moment. If stochastic gradient descent is run with a constant stepsize $\eta \le 1/L$ (or any diminishing stepsize satisfying $\sum_e \eta_e = \infty$ and $\sum_e \eta_e^2 < \infty$), then the parameter sequence $\{(\theta^{(e)}, \phi^{(e)})\}_{e=1}^{\infty}$ generated by the algorithm obeys,

$$\lim_{E\to\infty} \frac{1}{E} \sum_{e=1}^{E} \mathbb{E}\left[\|\nabla L(\Theta^{(e)})\|^2\right] = 0.$$

By extending this theorem, we can show that in the annealed setting where temperature $\tau^e \to 0$ guides the model from soft assignments to discrete ones as the loss function remains smooth and its variance is bounded, which justifies our choice of exponential decay of the temperature parameter. The complete proofs are provided in the Appendix 9.

9 THEORETICAL PROOFS

Monte-Carlo Gradient Estimator. Let $\Theta = (\theta, \phi)$ denote the model parameters. For each time step $t \in \{1, \dots, T\}$ and Monte-Carlo sample $s \in \{1, \dots, S\}$, draw Gumbel noise $g_s^t \sim \text{Gumbel}(0, 1)$ and define the sampled assignment as

$$\Pi_s^t := h_\phi(g_s^t), \tag{37}$$

where h_ϕ is the differentiable Gumbel-Softmax mapping. The per-timestep loss is denoted

$$\ell_t(\Theta; g) := \mathcal{L}_{\text{clu}}(f_\theta(X^t), C, h_\phi(g)). \tag{38}$$

The Monte-Carlo estimator of the full gradient is

$$\nabla_\Theta \mathcal{L} := \frac{1}{ST} \sum_{t=1}^{T} \sum_{s=1}^{S} \nabla_\Theta \ell_t(\Theta; g_s^t). \tag{39}$$

Lemma 3.1 (Unbiasedness)**.** The Monte-Carlo gradient estimator $\nabla_\Theta \mathcal{L}$, defined over S independent Gumbel-Softmax samples per time step across T temporal windows, is an unbiased estimator of the true gradient; that is,

$$\mathbb{E}\left[\nabla_\Theta \mathcal{L}\right] = \nabla_\Theta \mathcal{L}(\Theta). \tag{40}$$

Proof. Assume that $g_s^t \overset{\text{iid}}{\sim} \text{Gumbel}(0, 1)$, $\ell_t(\Theta; g)$ is differentiable in Θ and h_ϕ is differentiable in ϕ. We can compute the expectation of the Monte-Carlo estimator:

$$\mathbb{E}\left[\nabla_\Theta \mathcal{L}\right] = \frac{1}{ST} \sum_{t=1}^{T} \sum_{s=1}^{S} \mathbb{E}_{g_s^t}\left[\nabla_\Theta \ell_t(\Theta; g_s^t)\right] \tag{41}$$

$$= \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{g^t}\left[\nabla_\Theta \ell_t(\Theta; g^t)\right] \qquad \text{(i.i.d samples, identical expectation)} \tag{42}$$

$$= \frac{1}{T} \sum_{t=1}^{T} \nabla_\Theta \mathbb{E}_{g^t}\left[\ell_t(\Theta; g^t)\right] \qquad \text{(interchanging gradient and expectation)} \tag{43}$$

$$= \nabla_\Theta \left(\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{g^t}\left[\ell_t(\Theta; g^t)\right] \right) \tag{44}$$

$$= \nabla_\Theta \mathcal{L}(\Theta) \tag{45}$$

□

Theorem 3.1 (Variance Bound)**.** Assume that the per-sample gradient norm is uniformly bounded as

$$\|\nabla_\Theta \ell_t(\Theta; g)\| \le G_{\max} \quad \text{for all } t, \Theta, g.$$

Then the variance of the Monte-Carlo gradient estimator satisfies

$$\text{Var}\left[\nabla_\Theta \mathcal{L}\right] \le \frac{G_{\max}^2}{ST}.$$

Proof. Each of the $S \times T$ gradient terms $\nabla_\Theta \ell_t(\Theta; g_s^t)$ is independent and has norm at most $G_{\max}$. Thus:

$$\text{Var}\left[\nabla_\Theta \mathcal{L}\right] = \text{Var}\left[\frac{1}{ST}\sum_{t=1}^{T}\sum_{s=1}^{S}\nabla_\Theta \ell_t(\Theta; g_s^t)\right] \tag{46}$$

$$= \frac{1}{S^2T^2}\sum_{t=1}^{T}\sum_{s=1}^{S}\text{Var}\left[\nabla_\Theta \ell_t(\Theta; g_s^t)\right] \qquad \text{(independence)} \tag{47}$$

$$\leq \frac{1}{S^2T^2}\cdot ST \cdot G_{\max}^2 \qquad \text{(bounded variance)} \tag{48}$$

$$= \frac{G_{\max}^2}{ST}. \tag{49}$$

□

Theorem 3.2 (Convergence of Gumbel-Softmax Assignment in Temporal Clustering)**.** Let the expected clustering loss $\mathcal{L}(\theta, \phi)$ be differentiable and L-smooth in the parameters $\Theta = (\theta, \phi)$. Assume the Monte-Carlo gradient used in training is an unbiased estimator of $\nabla\mathcal{L}$ with bounded second moment. If stochastic gradient descent is run with a constant stepsize $\eta \leq 1/L$ (or any diminishing stepsize satisfying $\sum_e \eta_e = \infty$ and $\sum_e \eta_e^2 < \infty$), then the parameter sequence $\{(\theta^{(e)}, \phi^{(e)})\}_{e=1}^{\infty}$ generated by the algorithm obeys,

$$\lim_{E\to\infty}\frac{1}{E}\sum_{e=1}^{E}\mathbb{E}\left[\|\nabla\mathcal{L}(\Theta^{(e)})\|^2\right] = 0.$$

Proof. From Lemma 1, $\mathbb{E}\left[\nabla_\Theta \mathcal{L}\right] = \nabla\mathcal{L}^{(t)}$. By L-smoothness of $\mathcal{L}$, the descent lemma gives:

$$\mathbb{E}[\mathcal{L}^{e+1}] \leq \mathcal{L}^e - \eta\left\|\nabla\mathcal{L}^{(e)}\right\|^2 + \frac{L\eta^2}{2}\left(\|\nabla\mathcal{L}^e\|^2 + \text{Var}\left[\nabla_\Theta\mathcal{L}\right]\right).$$

Substituting $\text{Var}\left[\nabla_\Theta\mathcal{L}\right] \leq G_{\max}^2/(ST)$ yields:

$$\mathbb{E}[\mathcal{L}^{e+1}] \leq \mathcal{L}^e - \left(\eta - \frac{L\eta^2}{2}\right)\|\nabla\mathcal{L}^e\|^2 + \frac{L\eta^2 G_{\max}^2}{2ST}.$$

Rearranging and summing over epochs, $e = 1$ to E:

$$\frac{1}{E}\sum_{e=1}^{E}\mathbb{E}\left[\|\nabla\mathcal{L}^e\|^2\right] \leq \frac{\mathcal{L}^1 - \mathcal{L}^{(*)}}{T\left(\eta - \frac{L\eta^2}{2}\right)} + \frac{L\eta G_{\max}^2}{2S}.$$

As $E \to \infty$, the first term vanishes. Hence,

$$\lim_{E\to\infty}\frac{1}{E}\sum_{e=1}^{E}\mathbb{E}\left[\|\nabla\mathcal{L}^e\|^2\right] \leq \frac{L\eta G_{\max}^2}{2S}.$$

Choosing large enough S or using diminishing η_e ensures convergence to a stationary point. □

Corollary 3.1 (Annealed Convergence for Temporal Graph Clustering)**.** Let $\tau^e \to 0$ as $e \to \infty$, and suppose the temperature decays slowly such that each intermediate loss $\mathcal{L}_{\tau^e}(\theta, \phi)$ is L-smooth and the gradient variance remains bounded. Then the stochastic updates

$$(\theta^{(e+1)}, \phi^{(e+1)}) = (\theta^e, \phi^e) - \eta_e \cdot \widehat{\nabla}\mathcal{L}_{\tau^e}(\theta^e, \phi^e)$$

satisfy:

$$\lim_{e\to\infty}\mathbb{E}\left[\|\nabla\mathcal{L}_{\text{cat}}(\theta^e, \phi^e)\|\right] = 0,$$

where $\mathcal{L}_{\text{cat}}$ denotes the limiting discrete clustering objective with categorical (one-hot) assignments. That is, temporal clustering with Gumbel-Softmax and annealed temperature converges to a stationary point of the discrete temporal clustering loss.

10 ALGORITHM

Pseudocode Below we provide a high level pseudocode of our proposed method.

Algorithm 1 Monte-Carlo Cluster Loss with Gumbel–Softmax

Input : Node embeddings $Z \in \mathbb{R}^{N \times d}$; centroids $C \in \mathbb{R}^{K \times d}$; temperature $\tau > 0$; samples S
Output : $\mathcal{L}_{\text{clu}}$
Function `GumbelSoftmax` (ℓ_{row}, τ)
  // Draw i.i.d Gumbel noise for reparameterized sampling
  $g \sim \text{Gumbel}(0,1)^K$
  // Row-wise max for log-sum-exp stability
  $m \leftarrow \max_j \left(\ell_{\text{row},j} + g_j\right)/\tau$
  for $k \leftarrow 1$ **to** K **do**
    // Unnormalized weight for cluster k
    $u_k \leftarrow \exp\left(\frac{\ell_{\text{row},k}+g_k}{\tau} - m\right)$
  $s \leftarrow \sum_{j=1}^{K} u_j$
  // Normalized assignment vector $Q_{i,:}$
  return $\left[\mathrm{u}_1/s, \ldots, u_K/s\right]$
Function CLUSTERLOSS Z, C, τ, S
  // Read matrix sizes once
  $N \leftarrow \text{rows}(Z)$;
  $K \leftarrow \text{rows}(C)$
  // Pre-compute distances and logits for all node-centroid pairs
  for $i \leftarrow 1$ **to** N **do**
    for $k \leftarrow 1$ **to** K **do**
      // Distance between node i and centroid k
      $d_{ik} \leftarrow \|z_i - c_k\|^2$
      // Negative distance as logits for sampling
      $\ell_{ik} \leftarrow -d_{ik}$
  // Initialize Monte-Carlo accumulator
  $\mathcal{L}_{\text{sum}} \leftarrow 0$
  // Average over S independent Gumbel-Softmax assignment samples
  for $s \leftarrow 1$ **to** S **do**
    // Allocate one sample's assignment matrix $Q \in \mathbb{R}^{N \times K}$
    $Q \leftarrow \mathbf{0}_{N \times K}$
    // Sample assignments row-wise with temperature τ
    for $i \leftarrow 1$ **to** N **do**
      $Q_{i,:} \leftarrow$ `GumbelSoftmax`$(\ell_{i,:}, \tau)$
    // One-sample loss: expected distance under Q
    $L^{(s)} \leftarrow \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} Q_{ik}\, d_{ik}$
    // Accumulate for Monte-Carlo average
    $\mathcal{L}_{\text{sum}} \leftarrow \mathcal{L}_{\text{sum}} + \mathcal{L}^{(s)}$
  // Final Monte-Carlo estimate (variance decreases with S)
  $\mathcal{L}_{\text{clu}} \leftarrow \mathcal{L}_{\text{sum}}/S$
  // Return clustering loss
  return $\mathcal{L}_{\text{clu}}$

11 EXPERIMENTS

Node initialization. We initialize node embeddings from pre-trained Node2Vec (Grover & Leskovec, 2016) features learned on the latest graph structure. This is a deliberate choice to make a fair comparison with state of the art model in this domain (Liu et al., 2024). Pretraining provides a strong structural prior that captures local and global neighborhood connectivity before temporal updates begin. Pretrained embeddings are updated using the clustering loss and temporal consistency loss to learn the clusters that considers historical interactions among the nodes. Such initialization stabilizes early training, accelerates convergence, and leads to more semantically meaningful clusters, especially when node attributes are sparse or missing. We choose the number of unique node labels

as the number of clusters for evaluation purposes. This choice does not affect training, which remains fully unsupervised; it simply provides a consistent reference for comparing the discovered clusters to ground-truth labels at the current timestamp.

Dataset Statistics We evaluate our method on six temporal graph datasets summarized in Table 4, which vary widely in size, interaction density, and temporal dynamics.

Table 4: Dataset statistics with temporal characteristics.

Dataset	Nodes	Interactions	Edges	Timestamps	K	Degree	Temporal Nature
DBLP	28,085	236,894	162,441	27	10	16.87	Sparse, academic co-authorship
Brain	5,000	1,955,488	1,751,910	12	10	782.00	High-frequency, dense brain signals
Patent	12,214	41,916	41,915	891	6	6.86	Long-range, sparse citation network
School	327	188,508	5,802	7,375	9	1153.0	Short-term, dense social contacts
arXivAI	69,854	699,206	699,198	27	5	20.02	Dynamic academic collaboration
arXivCS	169,343	1,166,243	1,166,237	29	40	13.77	Highly dynamic, non-stationary

Coherence Score. The coherence score Halkidi et al. (2002) measures the average intra-cluster similarity across all clusters, with higher values indicating more compact cluster structure. It is defined in the range $[0, 1]$. For each cluster c, let $\mathcal{P}_c$ be all unordered node pairs (i, j) within the cluster. Let $d_{\cos}(i, j)$ denote the cosine distance. We convert distances to similarities $(s(i, j)$ and compute per-cluster coherence as-

$$s(i, j) = 1 - d_{\cos}(i, j). \tag{50}$$

$$\text{Coherence}(c) = \frac{1}{|\mathcal{P}_c|} \sum_{(i,j)\in\mathcal{P}_c} s(i, j). \tag{51}$$

Table 5: Hyperparameter Search Space for TGRAIL Model

Hyperparameter	Search Space
Training	
Learning Rate	log-uniform$(10^{-5}, 10^{-2})$
Batch Size	$\{128, 256, 512, 1024\}$
Optimizer	{Adam, AdamW.SGD}
Architecture	
Embedding Size	$\{64, 128, 256\}$
Temporal/Graph	
Negative Size	$\{5, 10, 20, 50\}$
History Length	$\{3, 5, 7, 10\}$
Temperature	
Temp Max	$\{5, 10, 15\}$
Temp Min	uniform$(0.1, 1.0)$
Decay Rate	uniform$(0.5, 0.95)$

Empirical Evaluation. Using the metrics from Table 3, TGRAIL demonstrates strong clustering coherence, with scores ranging from 0.8767 to 0.7558 despite rapid growth in the number of active nodes (from 71 to 12,214). Temporal alignment follows the expected trend in a rapidly evolving graph: it is high in early snapshots (e.g., 0.831 and 0.994 for Snapshots 2–3) and gradually decreases as the graph undergoes substantial structural reorganization (down to 0.345 by Snapshot 6). These results confirm that TGRAIL produces clusters that remain both semantically coherent and temporally consistent across evolving graph states.

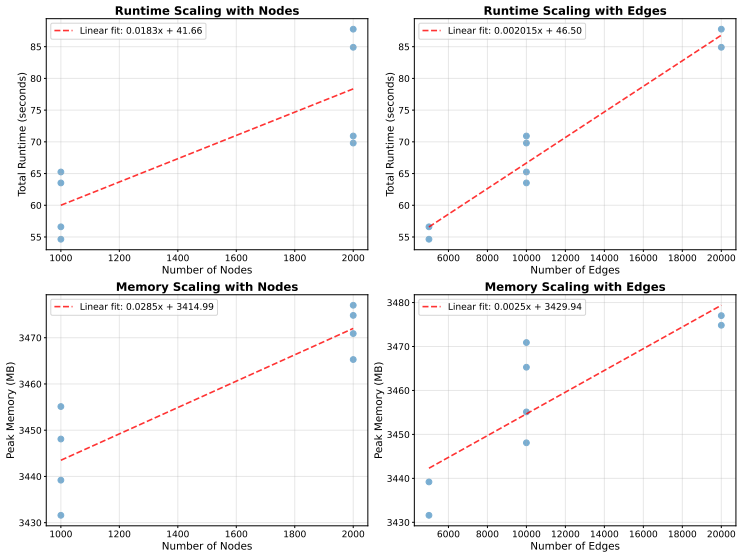

Figure 9: Scalability analysis showing runtime and memory scaling with graph size. Experiments use synthetic graphs with $n \in \{1000, 2000\}$ nodes, $m \in \{5000, 10000, 20000\}$ edges, and $T \in \{10, 50\}$ timestamps. Linear fits (red dashed lines) demonstrate linear scalability.

Discussion Our experiments evaluate clustering performance across six temporal graph datasets using four standard metrics (ACC, F1, NMI, ARI). As shown in Tables 1 and 2, our method achieves competitive or state-of-the-art performance, suggesting that joint modeling of temporal dynamics and cluster structure improves representation learning in evolving graphs.

On the School dataset, where temporal structure aligns cleanly with cluster assignments, our model achieves perfect scores (ACC/F1/NMI/ARI = $\sim$ 1.00), demonstrating its ability to recover ground-truth clusters under ideal conditions. In more challenging settings with sparse or noisy temporal signals, such as PATENT and DBLP, our approach outperforms baselines by 3–5% in ACC, highlighting its robustness to incomplete and overlapping event sequences. For datasets with non-stationary dynamics (ARXIV-AI, BRAIN), our model achieves consistent improvements. These results suggest that our temporal encoder captures fine-grained behavioural shifts more effectively than existing methods.

Static baselines (DeepWalk, node2vec, GAE) underperform significantly, reinforcing the necessity of temporal modeling. While TGC incorporates time through Hawkes processes, its decoupled representation and clustering stages limit optimization synergy. In contrast, our fully differentiable framework enables end-to-end learning, aligning temporal representations with clustering objectives.

These findings support our hypothesis that joint optimization of temporal dynamics and cluster assignments improves stability and accuracy in temporal graph clustering. The consistent gains across diverse datasets—ranging from cleanly structured (School) to highly dynamic (ARXIV-AI)—suggest broad applicability to real-world evolving graphs.

Empirical Complexity Analysis

To empirically validate linear scalability, we generate synthetic temporal graphs with configurable numbers of nodes, edges, and timestamps, where edges are created using random small scale-free graph structures, timestamps are uniformly distributed across edges, and node features are randomly sampled. As shown in figure 9, our empirical scalability analysis demonstrates strong evidence for linear scaling: runtime scales with edges with $R^2 = 0.952$ and $p < 3.5 \times 10^{-5}$, confirming near-perfect linear scalability, while memory usage scales with nodes with $R^2 = 0.804$ and $p = 0.003$, indicating a very good linear fit that explains 80.4% of the variance. Runtime scaling with nodes shows $R^2 = 0.664$ with $p = 0.014$, representing a moderate linear relationship that is statistically significant, where over 66% of the variance is explained by the linear model. These R^2 values, ranging from moderate ($R^2 \geq 0.66$) to excellent ($R^2 \geq 0.80$), combined with statistically significant

p-values ($p < 0.05$), provide robust empirical justification for our linear scalability claims, as they demonstrate that the dominant scaling behaviour is linear with only minor non-linear components that do not substantially impact the overall scalability characteristics of the model.

12 LIMITATIONS

Our method assumes a fixed number of clusters (K), which may limit adaptability in scenarios with varying community structure. Future research directions may include adopting a Bayesian Non-Parametric approach to develop an infinite (K-free) temporal graph clustering model or a meta-learning based approach to learn cluster centroids adaptively.

13 LLM USAGE

When preparing this manuscript, we utilized a Large Language Model (LLM) to assist with various aspects of the writing and research process. The LLM was employed for several key tasks:

- **Grammar and Language Polishing:** The LLM helped improve sentence structure, grammar, and overall readability of the manuscript, ensuring clear and professional academic writing throughout the paper.
- **Formatting Consistency:** We used the LLM to check and maintain consistent formatting across sections, including proper citation formatting, equation numbering, and LaTeX structure.