MGTCOM: Community Detection in Temporal Multimodal Graphs

EGOR DMITRIEV, Utrecht University, The Netherlands
MEL CHEKOL, Utrecht University, The Netherlands
SHIHAN WANG, Utrecht University, The Netherlands

Community detection is the task of discovering groups of nodes sharing similar patterns within a network. With recent advancements in deep learning, methods utilizing graph representation learning and deep clustering have shown great results in community detection. However, these methods often rely on the topology of networks (i) ignoring important features such as network heterogeneity, temporality, multimodality and other possibly relevant features. Besides, (ii) the number of communities is not known a priori and is often left to model selection. In addition, (iii) in multimodal networks all nodes are assumed to be symmetrical in their features; while true for homogeneous networks, most of the real-world networks are heterogeneous where feature availability varies. In this paper, we propose a novel framework (named MGTCOM) that overcomes the above challenges (i)–(iii). MGTCOM allows to discover dynamic communities through multimodal feature learning by leveraging a new sampling technique for unsupervised learning of temporal embeddings. Importantly, MGTCOM is an end-to-end framework optimizing network embeddings, communities, and number of communities in tandem. In order to assess its performance, we carried out extensive evaluation on a number of multimodal networks. We found out that our method is competitive against state-of-the-art and performs well under inductive setting.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**; • **Information systems** → **Document representation**; • **Mathematics of computing** → *Probabilistic algorithms.*

Additional Key Words and Phrases: community detection, representation learning, dynamic networks

ACM Reference Format:
Egor Dmitriev, Mel Chekol, and Shihan Wang. 2022. MGTCOM: Community Detection in Temporal Multimodal Graphs. In . ACM, New York, NY, USA, 41 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

CIKM '22, Oct 17-22, 2022, Virtual Event, Atlanta, Georgia, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Contents

1 INTRODUCTION

Various systems can be modelled as complex networks such as social [28], citation [44, 59], biological [24] and transaction [55] networks. The task of identifying patterns of nodes with common properties, in such networks, is referred to as community detection. There is an abundant number of community detection methods in literature that approach this problem through modularity optimization [4, 49, 58], clique identification [19, 39], and spectral optimization [30, 68]. With recent advancements in graph representation learning a new type of methods have emerged which utilize context-based learning techniques (e.g., DeepWalk [54], LINE [62] or Node2Vec [27]) to obtain topology-aware node embeddings. These embeddings are either combined with existing clustering methods [7, 80] or are jointly optimized with found clusters [9, 33, 57] to obtain communities.

In the above studies, the dynamic and multimodal characteristics of real-world networks are overlooked. These characteristics can manifest as meta-topological features (node and relation types) [8], temporal features, and contentual features (e.g., text and image attributes). Introduction of multimodality contrasts *homophily* assumed by previous methods as *heterophily* and can play an essential role in detecting communities in multimodal networks, as connected nodes may belong to different communities when multiple feature types are considered [93]. While it is common for causal links to be present between these features, it cannot be assumed without extensive domain knowledge. Various algorithms have been devised to address the issue of temporality and multimodality [18, 26, 40, 42], though as far as we are aware none of the methods are able to address the lossless setting where all the features are incorporated.

Another challenge is information variance present in heterogeneous real-world networks. Different node or relation types may have different feature subsets and/or dimensionality. Let us consider the Twitter dataset (SDS) we use as a case study in Section 7. This network consists of users, tweets, hashtags, and various relations in-between. Here tweets have content as textual features and post dates as temporal features, while users only have biography as textual features, and hashtags have neither. Similarly, users form a directed follower relation link, while multiple relations may be present between tweets such as retweet, mention or quote. The meta-topological information describes important semantics of this network, while varying features and topology can be used to identify individual nodes. If these characteristics are ignored by a model, then the quality of the communities discovered can be affected.

With the emergence of web-scale network datasets (often exceeding billions of nodes), recent advancements have pushed for scalability in graph representation learning [29, 87]. To this end, graph convolution methods have allowed for inductive inference on unseen nodes no longer requiring storing full graph Laplacian or node embeddings in memory. Utilizing this representation function learning helps solve scaling issues faced by many auto-encoder-based and shallow embedding community detection methods [46, 51, 70].

In this paper, we propose a novel community detection framework (MGTCOM) that is able to address the aforementioned challenges. MGTCOM discovers dynamic communities through multimodal feature learning and unsupervised learning with a new sampling technique. In particular, our key contributions include:

(i) A robust method for unsupervised representation learning on multimodal networks
(ii) A new sampling technique for unsupervised learning of temporal embeddings
(iii) An end-to-end framework optimizing network embeddings, communities, and number of communities in tandem
(iv) Extensive evaluation on the quality of various features in multimodal networks
(v) Implementation of various graph sampling algorithms found in the literature (See repository).

We compare MGTCOM with state-of-the-art methods and demonstrate its robustness on inference tasks.

The rest of the paper is organized as follows. Related works and relevant material is discussed in Section 2 and Section 3 respectively. Section 4 covers the details of our frameworks. In Section 5 we present extensive experimental results including comparison with baseline methods. Section 6 provides ablation studies to support

our design decisions. Finally, in Section 7 we provide a deep dive into results produced on the Social Distancing Students dataset as a case study.

Table 1. A comparison of MGTCOM with state-of-the-art on embedding (node, meta-topology, content and temporal information) and ability to infer the number of communities k. (top: graph embedding methods; bottom: community detection methods).

	topology	meta-topology	content	temporal	infers k
GraphSAGE [77]	•		•		
SageDy [77]	•		•	•	
CTDNE [50]	•			•	
HGT [31]	•	•	•		
ComE [9]	•				
GEMSEC [57]	•				
GRACE [81]	•		•		
Fani et al. [18]	•		•	•	
CP-GNN [42]	•	•			
MGTCOM	•	•	•	•	•

2 RELATED WORK

In this section, we provide an in-depth overview of related work and highlight important differences with our work by mainly focusing on graph embedding and community detection methods. A comparison of MGTCOM with the state-of-the-art is given in Table 1. As can be seen, MGTCOM is able to generate: (i) node, (ii) meta-topology, (iii) content, and (iv) temporal embeddings as well as (v) is able to infer the number of communities (K). By contrast, state-of-the-art methods (such as GraphSAGE and ComE) are able to produce either two or three of the above. A commonality of all the methods is that they are all utilize topological features. Similarly, we focus on representation based community detection methods in contrast to traditional link-based methods. We explain in detail these methods below. Throughout the paper we will use the terms network and graph interchangeably.

2.1 Graph Embedding

With the growing amount of rich graph data, efficient representation is highly demanded for retrieval and analytical purposes. Graph embedding focuses on the representation of nodes into low-dimensional vectors. The graph representation field stems from computational linguistics, which relies heavily on the notion of *distributional semantics*, stating that words occurring in the same context are semantically similar. By creating a parallel between words and nodes the linguistic approaches can be generalized to work in the context of graphs and vice versa.

Approaches such as DeepWalk [54], LINE [62], SDNE [67] and Node2Vec [27] utilize random walks as a means to generate context and adopt the Skip-gram [47] model to directly learn the node embeddings (*shallow embedding* methods). By defining a trade-off between first- and higher-order proximity they provide a way to fine-tune the learned topological representations for the task at hand. Grover and Leskovec [27] observe in their work that depth-first search sampling strategies (higher-order proximity) encourage network communities while breadth-first search (first-order proximity) encourages structural similarity as the local neighborhood is more thoroughly explored.

On the other hand, matrix factorization-based approaches represent first-order proximity using an adjacency or Laplacian matrix. Consequently, they decompose the matrix in order to obtain node-based representation matrix [6]. As this process is quite expensive $O(n^{2.372}$, graph autoencoders (GAE) [36, 64] and graph convolutional networks (GCN) [37] are used instead.

Newer methods aim to solve various issues with current approaches involving scalability [29, 87], incorporation of node/edge features [29, 78], application to heterogeneous [5, 16, 31], attributed [12, 76] and temporal [14, 50, 77] networks (see Table 1 for comparison). In line with this, our proposed method (MGTCOM) is able to address all of the above issues.

2.1.1 Scalability. The authors of GraphSAGE (Hamilton et al. [29]) argue that many graph embedding methods are *transductive* and therefore have to be retrained upon the introduction of new or unseen nodes (nodes that are not part of the training data). Additionally, with the emergence of web-scale graphs containing billions of nodes, it is not possible to keep all node embeddings in memory [87]. Hence, in GraphSAGE, they introduce a local k-hop neighborhood sampling strategy and a GCN architecture that is able to infer node representations based on the sampled subgraph. A caveat of this approach is that the GCN architecture requires the presence of node features. While various workarounds exist to use zero or random vectors for missing features, this limits its application for various graph datasets. In MGTCOM we overcome this issue by introducing auxiliary embeddings for nodes with missing features. By keeping auxiliary embeddings of most important nodes at hand, a primary embedding can be computed for each node within the graph.

2.1.2 Heterogeneous networks. The above methods mainly work on homogeneous networks in which all nodes and edges belong to the same types. Often real-world data cannot be efficiently represented using homogeneous networks. Hence, to accurately represent real-world information heterogeneous networks are used. These networks involve *meta-topological* information that characterizes various relationships between different types of nodes/entities [82]. Since most graph embedding methods are designed for homogeneous networks, extending them to incorporate heterogeneous networks is not trivial.

One way to address meta-topological features is by using meta-path constrained random walks to capture semantic and structural relations between different node/entity types [16, 21]. Meta-path describes a sequence of entity and relation types. For example, an "APA" meta-path would define a path between Author-Paper-Author node types. Derivative works introduce attention-based mechanisms to learn the importance of the meta-types [74]. While highly successful, the technique faces certain limitations, mainly that the construction of meta-paths requires extensive domain knowledge and that in highly heterogeneous networks such as knowledge graphs, the amount of meta-paths becomes unmanageable.

Other methods utilize a representation-based approach [5, 75] to explicitly capture meta-topological features by defining relations as translations between different node types. This approach is further utilized in GCN-based methods [31, 90] to apply them in an inductive setting. Furthermore, in Hu et al. [31] the authors improve the neighborhood sampling algorithm by introducing a type-based budget for unbiased sampling.

2.1.3 Temporal networks. Multimodal networks are dynamic and may evolve over time. Temporal networks are a specialization of multimodal networks as they attach a start and end timestamp to each node and edge. Accordingly, graph representation methods should have the ability to capture this evolution.

Temporal graph embedding approaches are mainly split into two categories. Snapshot-based approaches operate by temporally splitting the graph into multiple snapshots or subgraphs and applying (modifying) existing graph embedding methods by temporally smoothing between the snapshots [25, 43, 52, 92]. The second category are the continuous temporal representation approaches which attempt to capture temporal information within the learned embeddings. Generally, these methods look at the temporal progression of individual nodes rather than utilizing predefined snapshots. The techniques vary; CTDNE introduces biased temporal random walks [50]; SageDy introduces a neighborhood sampling technique to filter for temporal neighborhood [77]; BurstGraph captures node representation changes using a RNN [91]; HyTe [14] modifies representation-based techniques to explicitly learn temporal information.

2.2 Community Detection

A community reveals patterns within its members that are different from those in other communities in a network. There is an abundance of work concerning the finding of community structures by relying mainly on topological features [20, 79]. Despite this, the term *community* does not have a universally accepted definition. In their work Peel et al. [53] argue that community detection does not have a one size fits all solution and that definition and quality highly depend on the task at hand. Similarly, they observe that the task of community detection is analogous to finding clusters in document vectors. Nevertheless a few common characteristics distinct community detection from tasks such as topic analysis and clustering including the involvement of topological information and the fact that the number of communities is not known a priori.

Recent community detection methods focus on exploiting feature-rich information found in multimodal networks [60]. The focus has shifted from link-based methods towards deep learning methods which combine graph embedding methods with clustering algorithms (such as k-means or spectral clustering) [38, 64]. Similar methods are employed to learn find communities that take into account global context by utilizing graph autoencoders [7, 73] or graph affiliation networks [83]. More advanced methods utilize multi-objective optimization by combining topological accuracy and cluster quality metrics during graph representation learning [9, 57, 69, 89]. Other methods focus on modifying [33] or augmenting [34] graph context sampling algorithms to reinforce communities within learned representations.

2.2.1 Multimodal Methods. Many methods rely on *homophily* which refers to the assumption that "individuals" sharing similar patterns are more likely to be connected [45]. With the emergence of multimodal community detection methods *heterophily* becomes equally important as similarity may not always be correlated with topological features [93]. Some argue that consideration of link or content information alone is sufficient for identifying communities [18]. Sparsity, noise, and irrelevant information may mislead traditional community detection or topic modeling algorithms. However, both types of information may be of interest for analysis and may be valuable in overcoming noise in multimodal networks.

In line with this, various methods [41, 61] modify Latent Dirichlet Allocation algorithm to incorporate attribute, topological and meta-topological information. Cao et al. [7] and Yang et al. [81] utilize autoencoders to jointly optimize graph embeddings on content and topological information. Fani et al. [18] use topic models to construct a user interest histogram over a time axis, which in turn is used to learn temporal content-based node representations. These representations are interpolated with topological representations, the similarity along edges is computed, and fed as edge weight to the existing link-based community detection algorithm (Louvain [4]).

2.2.2 Heterogeneous Networks. Meta-topological information is a valuable asset for the analysis of found communities. This information can be used in various ways to assist in community detection, for instance, by the representation of small node-specific ego-networks [32] and learning the importance of network relations [61]. Luo et al. [42] propose CP-GNN which combines a heterogeneous graph transformer architecture with k-means clustering to find communities in content-rich heterogeneous graphs. Moreover, they devise a context-path-based k-hop neighborhood sampler to reinforce the discovery of community structures in topological data.

2.3 Clustering

The task of clustering is to find groups of documents in a d-dimensional vector space based on a predefined similarity metric in an unsupervised manner [48]. Many community detection algorithms rely on existing clustering algorithms such as k-means [7, 38, 42, 64, 85] and Gaussian Mixture Models (GMM) [9, 13]. Others employ end-to-end clustering techniques such as deep embedding clustering [81] and clustering loss based parameter optimization [57, 89].

As it is uncommon to know the number of clusters in community detection tasks [20], determining the optimal cluster count is often left to model selection which might become computationally expensive. While various non-parametric clustering algorithms exist such as DBSCAN [17], OPTICS [2] and BIRCH [88] they are not straightforward to incorporate into end-to-end applications.

Bayesian non-parametric methods such as Dirichlet Process Mixture (DPM) have had great results in clustering and community detection tasks where the number of communities is unknown [65, 94, 95]. As these models can evaluate the likelihood of a set of cluster parameters being drawn from a prior distribution, the task is transformed into a Markov Chain Monte Carlo (MCMC) sampling problem. Since there are prohibitively many possible parameter states, various hierarchical algorithms are proposed to explore the most promising states efficiently [11, 63]. Because these methods can be estimated using Expectation-Maximization (EM) algorithms, the previously introduced embedding methods can be utilized to learn representations and clusters in an end-to-end manner [9, 56].

Table 2. Notation used in this paper.

Notation	Description
$\mathcal{V}$	The set of nodes in a graph
$\mathcal{A}$	The set of node types
$\mathcal{E}$	The set of edges in a graph
$\mathcal{R}$	The set of edge types/relations
$\mathcal{T}$	The set of timestamps in a graph
$\mathcal{X}_{\phi(\cdot)}$	Feature matrix for node type $\phi(\cdot)$
$\phi(v) \in \mathcal{A}$	Type of node v
$\psi(e) \in \mathcal{R}$	Type of edge e
G_v	k-hop neighborhood subgraph for node v
$d \in \mathbb{N}$	Size of node embedding vector
$E_v \in \mathbb{R}^d$	Auxiliary embedding vector for node v
$Z_v \in \mathbb{R}^d$	Primary embedding vector for node v
$Z_v^{\mathcal{E}}, Z_v^{\mathcal{T}} \in \mathbb{R}^d$	Task specific embedding vectors for node v
$K \in \mathbb{N}$	Number of communities
$\mathcal{N}(\mu_k, \Sigma_k), \theta_k$	Parameters for the k'th cluster/community
$\mu_k \in \mathbb{R}^d$	k'th cluster mean vector
$\Sigma_k \in \mathbb{R}^{d \times d}$	k'th cluster covariance vector
$\mathbf{z} \in \{0, .., K\}^{\lvert\mathcal{V}\rvert}$	Community membership assignment vector
ω	Interval window for temporal context sampling
$P_l \in \mathcal{V}^l$	Sampled context window of length l

3 PRELIMINARIES

We present a brief overview of the key concepts and notations used in community detection and graph representation learning. The notations can be found in Table 2.

Definition 3.1 (Heterogeneous graph). A heterogeneous graph, denoted as $G = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{R})$ consists of a set of nodes $\mathcal{V}$, a set of edges $\mathcal{E}$, and their associated type mapping functions $\phi : \mathcal{V} \rightarrow \mathcal{A}$ and $\psi : \mathcal{E} \rightarrow \mathcal{R}$. ϕ (resp. ψ) maps a node (resp. edge) to its type. $\mathcal{A}$ and $\mathcal{R}$ denote predefined sets of node and edge types, respectively, where $|\mathcal{A}| \geq 1$ and $|\mathcal{R}| \geq 1$. G is a homogeneous graph if $|\mathcal{A}| = 1$ and $|\mathcal{R}| = 1$.

We define a multimodal information network by combining the notion of heterogeneous, continuous-time, and contentual networks.

Definition 3.2 (Multimodal graph). A multimodal graph is defined as $G = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \mathcal{X})$ where $\mathcal{T}$ is a set of timestamps t and $\mathcal{X}$ is a set of type specific feature matrices $\mathcal{X}_{\phi(\cdot)}$. Each node $v \in \mathcal{V}$ (resp. edge $e \in \mathcal{E}$) has a timerange $\tau(v) = [t_s, t_e]$ (resp. $\tau(e) = [t_s, t_e]$), indicating the time period on which it is considered valid, where $t_s, t_e \in \mathcal{T}$. In addition, each node v has an attribute vector $\mathbf{x} \in \mathcal{X}_{\phi(v)}$.

Definition 3.3 (Incompleteness constraints). Real-world multimodal networks can be noisy, incomplete, and may change over time. In order to represent this information we introduce additional indicator functions to denote whether a node has a timerange $\mathbf{1}_{\mathcal{T}} : \mathcal{V} \rightarrow \{0, 1\}$, has a feature vector $\mathbf{1}_{\mathcal{X}} : \mathcal{V} \rightarrow \{0, 1\}$, or is unseen during training $\mathbf{1}_{\mathcal{V}} : \mathcal{V} \rightarrow \{0, 1\}$. We refer to the noisiness, incompleteness, and temporality as incompleteness constraints.

Definition 3.4 (Context window). A context window connects nodes based on some predefined criteria. Two nodes are *context neighbors* if they occur in the same context window. In our work, we use two different kinds

of context windows. The first is the *topological context window*. It connects two nodes v_i and v_j if there exists a k-hop path $p_k^{\mathcal{E}}$ in graph G through which they are connected. The second is *temporal context window* $p_\omega^{\mathcal{T}}$. It connects v_i and v_j if they occur within a given time window $\omega = [t_s, t_e]$. Going forward we use $P_k^{\mathcal{E}}$ and $P_\omega^{\mathcal{T}}$ to denote a fixed size sample of all possible context windows.

Definition 3.5 (Gaussian Mixture Model). Gaussian mixture models (GMM) is a clustering algorithm that assumes the data points are generated by K d-dimensional multivariate Gaussian distributions Eq. (1). Here cluster parameters θ_k for $k \in \{1, ..., K\}$ consist of the mean vector $\mu_k \in \mathbb{R}^d$ and the covariance matrix $\Sigma_k \in \mathbb{R}^{d \times d}$. A K-dimensional binary variable $\mathbf{z}$ is used to denote membership of a particular point n where $\sum_k z_{nk} = 1$. Mixing coefficients π_k specify a marginal distribution over $\mathbf{z}$, such that $\sum_{n \in N} p(z_{nk} = 1) = \pi_k$ where $\pi_k \in [0, 1]$ and $\sum_{k=1}^K \pi_k = 1$. Consequently r_k represents conditional probability of $\mathbf{z}$ given a data point $\mathbf{x}$ Eq. (2).

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \Sigma_k) \tag{1}$$

$$r_k = p(z_k = 1|\mathbf{x}) = \frac{\pi_k \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}_k, \Sigma_k\right)}{\sum_{j=1}^K \pi_j \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}_j, \Sigma_j\right)} \tag{2}$$

Assuming that the points $\mathbf{X} \in \mathbb{R}^{N \times d}$ are drawn independently from the distribution, the log-likelihood function is given by Eq. (3). The value of μ_k, Σ_k, π_k can be found by setting derivative of $\ln p(\mathbf{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma)$ to zero with respect to their values yielding closed form equations Eqs. (4) to (6). N_k represents the number of points assigned to cluster k. While model parameters can be computed given values of $\mathbf{X}$ and $\mathbf{r}$ are known, it is important to note that $\mathbf{r}$ is dependent on the model parameters Eq. (2). Expectation-Maximization (EM) is an elegant iterative technique devised to find such clustering parameters. Given an initial cluster assignment that may be obtained using k-means or a similar technique, *expectation* (E) and *maximization* (M) steps are applied alternatively Fig. 1. The E step uses current cluster parameters to evaluate posterior probabilities Eq. (2), while the M step uses these probabilities to compute new model parameters using Eqs. (4) to (6). The model is deemed converged once the change in parameters or assignment falls below a certain threshold.

$$\ln p(\mathbf{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k\right) \right\} \tag{3}$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \mathbf{x}_n \tag{4}$$

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \left(\mathbf{x}_n - \boldsymbol{\mu}_k\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^{\mathrm{T}} \tag{5}$$

$$\pi_k = \frac{N_k}{N} \tag{6}$$

Definition 3.6 (Dirichlet Process Mixture Model). Gaussian Mixture Models suffer from severe overfitting problems in form of single-point collapse and the fact that the number of clusters needs to be known a priori. Bayesian parametric (BP) and non-parametric (BNP) mixture models aim to solve these issues by introducing prior distributions governing the model parameters (π, μ, Σ) and using maximum a priori (MAP) instead of maximum likelihood estimation.

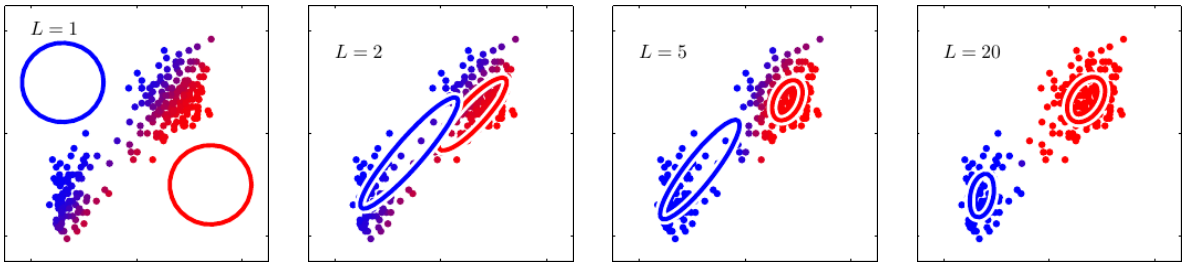


Fig. 1. Visualization Expectation-Maximization algorithm with random initial clusters assignment (a). [3]

Dirichlet process mixture model (DPMM) is a part of BNP mixture models which finds a clustering solution when K is unknown. DPMM extends GMM as it is an infinite mixture model Eq. (7) with the Dirichlet process as prior distribution on the number of clusters Eq. (8). Here hyperparameter α_0 is the concentration parameter referring to the prior amount of observations associated with each component. The cluster parameters θ are assumed to be i.i.d. distributed and are drawn from a prior distribution. In our case, Normal Wishart Distribution (NW) Eq. (9) where hyperparameters κ and ν represent the concentration parameter and degrees of freedom of the Wishart distribution respectively. The data is parameterized by the data mean μ and $\mathbf{\Lambda}$ which is the precision matrix (inverse of the covariance matrix Σ).

$$p(\boldsymbol{x}) = \sum_{k=1}^{\infty} \pi_k \mathcal{N}\left(\boldsymbol{x}, \boldsymbol{\mu}_k, \mathbf{\Lambda}^{-1}\right) \tag{7}$$

$$p(\boldsymbol{\pi}) = \mathrm{Dir}\left(\boldsymbol{\pi}; \boldsymbol{\alpha}_0\right) = \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K} \Gamma(\alpha_0)} \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1} \tag{8}$$

$$\begin{aligned} p(\boldsymbol{\mu}, \mathbf{\Lambda}) &= \mathrm{NW}(\boldsymbol{\mu}, \mathbf{\Lambda}; \kappa_0, \mu_0, \nu_0, \boldsymbol{W}_0) \\ &= \prod_{k=1}^{K} \underbrace{\mathcal{N}\left(\boldsymbol{\mu}_k \mid \mu_0, (\kappa_0 \mathbf{\Lambda}_k)^{-1}\right)}_{p(\mu_k \mid \mathbf{\Lambda}, \kappa_0, \mu_0)} \underbrace{\mathcal{W}\left(\mathbf{\Lambda}_k \mid \mathbf{W}_0, \nu_0\right)}_{p(\mathbf{\Lambda} \mid \mathbf{W}_0, \nu_0)} \end{aligned} \tag{9}$$

The prior parameters α_0, κ_0, and ν_0 are set to a predetermined value, whereas prior parameters μ_0 and W_0 are calculated on a sample of the full dataset using Eqs. (14) and (15). Here $\alpha_0, \nu_0, \kappa_0 \in \mathbb{R}^+$ and $\nu_0 > d + 1$.

$$\overline{\mathbf{x}}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \mathbf{x}_n \tag{10}$$

$$\mathbf{S}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \left(\mathbf{x}_n - \overline{\mathbf{x}}_k\right) \left(\mathbf{x}_n - \overline{\mathbf{x}}_k\right)^{\mathrm{T}} \tag{11}$$

EM can similarly be used to approximate solutions for DPM models. During the E step Eq. (2) is once again used to estimate the assignments. While during the M step Eqs. (10) and (11) equations analogous to Eqs. (4) and (5) are used to estimate the data covariance and data mean. Subsequently the following closed form equations are used to compute posterior parameters for the given prior Eqs. (12) to (16). Given the posterior parameters, the

new cluster parameters are inferred using Eqs. (14) and (17). When α_0, κ_0, and ν_0 are much smaller than N, the posterior distribution will be influenced primarily by the data rather than the prior. We use λ to denote computed posterior parameters.

$$\pi_k = \frac{N_k}{\sum_{k=1}^{k} N_k + \alpha_0} \tag{12}$$

$$\kappa_k = \kappa_0 + N_k \tag{13}$$

$$\mu_k = \frac{1}{\kappa_k} \left(\kappa_0 \mu_0 + N_k \overline{\mathbf{x}}_k\right) \tag{14}$$

$$\mathbf{W}_k^{-1} = \mathbf{W}_0^{-1} + N_k \mathbf{S}_k + \frac{\kappa_0 N_k}{\kappa_0 + N_k} \left(\overline{\mathbf{x}}_k - \mu_0\right) \left(\overline{\mathbf{x}}_k - \mu_0\right)^{\mathrm{T}} \tag{15}$$

$$\nu_k = \nu_0 + N_k \tag{16}$$

$$\Sigma_k = \frac{\nu \mathbf{W}_k^{-1}}{\nu - d + 1} \tag{17}$$

The described implementation solves overfitting and cluster count, though it is an incomplete one since in practice the cluster count has a defined upperbound K (computationally and storage-wise). While the clusters can get pushed out of existence, no additional clusters can be created. To solve this issue many variants of DPMM have been proposed utilizing the Chinese Restaurant process, Collapsed Weight sampling, etc. We focus on the split/merge sampling algorithm introduced by Chang and Fisher III (DPMMSC) [11]. For a exhaustive discussion, we refer interested readers to [3, 10].

DPMMSC exploits an alternate perspective in which DPMM is defined as a Monte Carlo Markov Chain if all the chosen priors are conjugate (i.e. prior distribution is in the same form as the posterior distribution). The stationary distribution is defined by the probability of cluster parameters given the data observations Eq. (19). Intuitively in this approach sampling methods are used to approximate the E step of EM by sampling from the current estimate of posterior distribution $p(\mathbf{z}|\mathbf{X}, \theta^{\text{old}})$ (proposal distribution), where during M step the new state θ is found. A similar methodology is employed to transition between different values of K by proposing θ directly. As proposal space is unmanageably large, a greedy strategy is employed to propose the most promising states.

$$H_s = \frac{\alpha \Gamma(N_{k_1}) p(X_{k_1}; \lambda_{k_1}) \Gamma(N_{k_2}) p(X_{k_2}; \lambda_{k_2})}{\Gamma(N_k) p(X_k; \lambda_k)} \tag{18}$$

$$p(\mu_k, \Sigma_k|\mathbf{X}_k) = \mathrm{NW}(\boldsymbol{\mu}, \boldsymbol{\Lambda}; \kappa_0, \mu_0, \nu_0, \boldsymbol{W}_0) \tag{19}$$

$$\begin{aligned} p(\mathbf{X}; \lambda) &= \int p\left(\mathbf{X} \mid \boldsymbol{\mu}_k, \Sigma_k\right) p\left(\boldsymbol{\mu}_k, \Sigma_k; \lambda\right) d\left(\boldsymbol{\mu}_k, \Sigma_k\right) \\ &= \frac{1}{\pi^{\frac{Nd}{2}}} \frac{\Gamma_d\left(\nu_0/2\right)}{\Gamma_d(\nu_k/2)} \frac{\left|\nu_0 \boldsymbol{\Lambda}_0\right|^{\nu_0/2}}{\left|\nu_k \boldsymbol{\Lambda}_k\right|^{\nu_k/2}} \left(\frac{\kappa_k}{\kappa_0}\right)^{d/2} \end{aligned} \tag{20}$$

For each supercluster k, two auxiliary subclusters are defined with parameters θ_{k_1} and θ_{k_2} forming a two-component GMM. Once subclusters are in a converged state, the split proposals are made given the supercluster and its two subcomponents. Similarly, supercluster merges are proposed by picking k nearest candidates for each supercluster.

The proposed candidates are either accepted or rejected by the Metropolis-Hastings (MH) algorithm moving the model to the next state. As the The split acceptance ratio is defined by the probability of data being sampled from the split state in contrast to the current state Eq. (18). Analogously, the merge ratio is its inverse, namely $\frac{1}{H_s}$.

Eqs. (3), (9), (12) and (19) are used to derive the marginal probability of data being generated by parameter set λ given prior parameters Eq. (20) (note that π refers to the mathematical constant).

The proposals are considered once the supercluster model has converged. If no proposal is accepted, then DPMMSC is considered as converged.

Definition 3.7 (Graph Convolutional Neural Networks). Graph Convolutional Neural Networks (GCN) [29, 35, 37] generate node embeddings given a spatial filter which is applied as a convolution for each node their neighborhood. The convolution operation enables GCNs to propagate structural information of graphs throughout the network (referred to as message-passing). By layering this process, the receptive field of each node expands to its k-hop neighborhood.

Suppose H_t^l is the representation of node t at layer l, a forward step of the message-passing procedure is defined as Eq. (21) where $N(t)$ is t's neighboring node set and $\mathcal{E}(s,t)$ is the set of edges between nodes t and its neighbor s. Here the operator **Message**$(\cdot)$ extracts useful information from the neighboring source nodes s, while the **Aggregate**$(\cdot)$ operator gathers the neighborhood information via some aggregation operator such as *mean*, *sum* or *max* to get contextualized representation of t.

$$H_t^{(l)} = \underset{\forall s \in N(t), \forall e \in \mathcal{E}(s,t)}{\textbf{Aggregate}} \left[\textbf{Message}\left(H_s^{(l-1)}, e, H_t^{(l-1)} \right) \right] \tag{21}$$

The time complexity to run a forward step over the entire training set is $O(|\mathcal{V}| \cdot deg \cdot d^2)$ where deg refers to the average node degree. While $deg \ll |\mathcal{V}|$ is true for most graphs, a vital optimization step is to sample a fixed size N_v ensuring that deg is bounded by a constant.

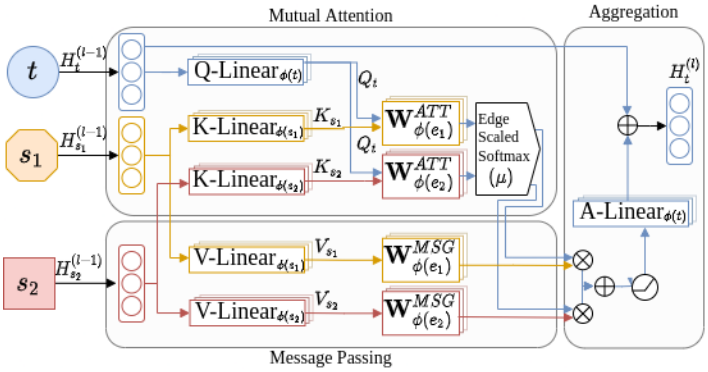


Fig. 2. Visualization of a Heterogeneous Graph Transformer Layer. Given target node t and neighboring source nodes s_1 and s_2 by edges e_1 and e_2, mutual attention and messages are computed. Within aggregation step the messages are attended and combined with previous target node embedding $H_t^{(l-1)}$ resulting in the new embedding vector $H_t^{(l)}$

Definition 3.8 (Heterogeneous Graph Transformer). Classical GCNs focus mainly on homogeneous graphs. A fair amount of works describe ways to adapt existing algorithms by introducing a **Message** step parameterized by meta-topological types. Based on the observation that the value of different connections varies given a node type, attention-based mechanisms are introduced into the aggregation process. Inspired by success in NLP Heterogeneous Graph Transformer (HGT) [31] adopts the transformer architecture [66] to by calculating mutual attention based on representation and meta-types of source, target and relation information.

$$H_t^l = \underset{\forall s \in N(t), \forall e \in \mathcal{E}(s,t)}{\textbf{Aggregate}} \left[\textbf{Attention}\,(s,e,t) \cdot \textbf{Message}\,(s,e,t) \right] \tag{22}$$

HGT consists mainly of three components, mutual attention possession performance of each source node, message passing extracts information from source nodes and target-specific aggregation which combines the neighborhood messages. A general form for a forward pass is defined as Eq. (22).

The attention vector is calculated by mapping source node s into Key K and target node t into a Query Q vectors Eqs. (25) and (26). A single head attention vector is calculated an inner product similarity vector between Key K and Query Q vectors Eq. (24) given a relation specific interaction matrix, where prior tensor $\mu \in \mathbb{R}|\mathcal{A}| \times |\mathcal{R}| \times |\mathscr{A}|$ denotes significance of each relation triplet. $K(s)$ and $Q(t)$ are computed as projections of source s and target t nodes respectively Eqs. (25) and (26). The final attention vector results from a concatenation of h attention heads per source node Eq. (23).

$$\textbf{Attention}_{HGT}(s,e,t) = \underset{\forall s \in N(t)}{\text{Softmax}} \left(\underset{i \in [1,h]}{\Big\|} \mathit{ATT\text{-}Head}^i(s,e,t) \right) \tag{23}$$

$$\mathit{ATT\text{-}Head}^i(s,e,t) = \left(K^i(s) W^{ATT}_{\psi(e)} Q^i(t)^T \right) \cdot \frac{\mu_{\langle \phi(s), \psi(e), \phi(t) \rangle}}{\sqrt{d}} \tag{24}$$

$$K^i(s) = \text{K-Linear}^i_{\phi(s)} \left(H_s^{(l-1)} \right) \tag{25}$$

$$Q^i(t) = \text{Q-Linear}^i_{\phi(t)} \left(H_t^{(l-1)} \right) \tag{26}$$

Similarly, the multi-head message is computed by applying type-dependent projection (M-Linear) to the input source node representation and transforming it using the edge type matrix $W^{MSG}_{\psi(e)} \in \mathbb{R}^{\frac{d}{h} \times \frac{d}{h}}$ to incorporate the relation dependency into the result Eq. (28). In both operations, edge interaction matrices and the head-specific type projection matrices are shared to minimize the number of used parameters.

$$\textbf{Message}_{HGT}(s,e,t) = \underset{i \in [1,h]}{\Big\|} \mathit{MSG\text{-}Head}^i(s,e,t) \tag{27}$$

$$\mathit{MSG\text{-}Head}^i(s,e,t) = \text{M-Linear}^{\,i}_{\phi(s)} \left(H_s^{(l-1)} \right) W^{MSG}_{\psi(e)} \tag{28}$$

Finally, during the aggregation step, the calculated attention is applied to neighborhood messages and summed into the neighborhood representation vector Eq. (29). The final node representation vector $H_t^{(l)}$ results from the summation of the projected neighborhood vector into the target node space and the previous representation of the target vector Eq. (30).

$$\widetilde{H}_t^{(l)} = \underset{\forall s \in N(t)}{\bigoplus} \left(\textbf{Attention}_{HGT}(s,e,t) \cdot \textbf{Message}_{HGT}(s,e,t) \right) \tag{29}$$

$$H_t^{(l)} = \text{A-Linear}_{\phi(t)} \left[\sigma \left(\widetilde{H}_t^{(l)} \right) \right] + H_t^{(l-1)} \tag{30}$$

See Fig. 2 for a visualization of a forward pass of single layer HGT.

Problem formulation. Given a multimodal graph G, our goal is to learn a node embedding function $\zeta : G_v \rightarrow \mathbb{R}^d$ which given a k-hop neighborhood subgraph G_v of node v produces a d-dimensional embedding vector Z_v. The objective is to minimize the distance between embedding Z_v to other node embeddings, given that they are

topological and/or temporal context neighbors of node v. Taking into account incompleteness constraints, ζ should work under any valuation of $(\mathbf{1}_{\mathcal{X}(v)}, \mathbf{1}_{\mathcal{T}(v)}, \mathbf{1}_{\mathcal{V}(v)})$. We also aim to find community parameters $\theta = \{\mathcal{N}(\mu_1, \Sigma_1), ..., \mathcal{N}(\mu_k, \Sigma_k)\}$ and node-to-community assignment $\mathbf{z} \in \{0, .., K\}^{|\mathcal{V}|}$ such that their members have a low inter-proximity in contrast to other nodes. Finally, the found community count K should approximate the ground truth number of communities.

4 THE PROPOSED APPROACH

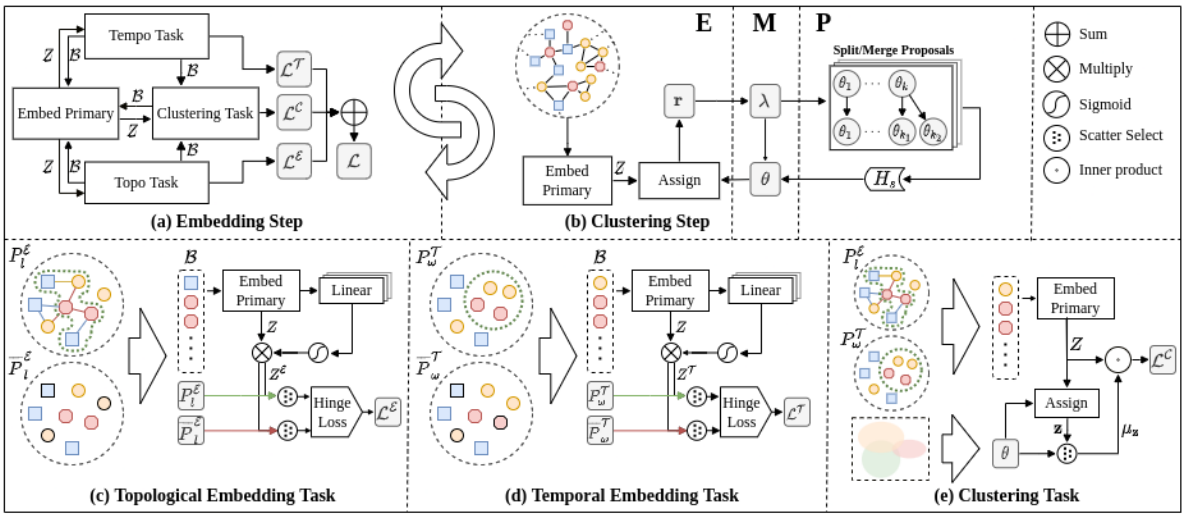


Fig. 3. Overview of the MGTCOM framework. (a) In the embedding step primary embeddings are used in auxiliary tasks to construct the multi-objective loss. (b) Clustering step updates clustering by alternating between Expectation, Maximization, and Proposal steps. (c) In the topological embedding task, random walk sampling and feature-wise attention minimize inter-node proximity. (d) In the temporal embedding task, ballroom walk sampling and feature-wise attention minimize proximity between temporally related nodes. (e) Clustering task adds community awareness to the embeddings by minimizing proximity between nodes within the same cluster.

We present our framework for Community Detection in Temporal Multimodal Graphs (MGTCOM) that learns multimodal representation vectors for graph nodes and detects communities in tandem. We achieve this by leveraging heterogeneous graph transformers [31] to learn primary node embedding function ζ. In order to handle the incompleteness constraints, we introduce an auxiliary embedding vector E for known (or seen) nodes with missing features. Next, we learn task-specific node representation for topological and temporal information by combining primary embeddings with task-specific transformation/attention and context sampling. As we utilize random walks for topological context sampling, we introduce its analogue as an unbiased temporal window sampling algorithm for temporal context collection. Finally, we adopt DPMM for community detection and close the loop by introducing cluster-based loss to ensure the graph embeddings are *community-aware*. MGTCOM consists of three major components (as can be seen in Fig. 3): primary embedding module, task-specific learning, and community detection/clustering module. MGTCOM also has a graph sampling component. In the following, we describe the components in detail.

4.1 Primary embedding module

The central component of our framework is responsible for inferring the primary representation vector Z_v given a node $v \in \mathcal{V}$ in a graph G. Motivated by the success of inductive GCN-based methods [29, 31, 87], we build our architecture by combining L graph convolution layers (*HeteroConv* or HGCN) and a graph subsampler (*HeteroSample*).

Algorithm 1: Batchwise primary node embedding

1 **Procedure** `EmbedPrimary()`

Input: multimodal graph $G = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{R}, \mathcal{X})$, mini-batch $\mathcal{B} \subseteq \mathcal{V}$, auxiliary node embedding $E \in \mathbb{R}^{N_{\bar{\mathcal{X}}} \times d}$ where $N_{\bar{\mathcal{X}}} = |\{v | v \in \mathcal{V}, \mathbf{1}_{\mathcal{X}}(v) = 0\}|$, number of convolutional layers L

Output: The primary embedding $Z_{\mathcal{B}}$ for nodes in batch $\mathcal{B}$

2 $G_{\mathcal{B}}(\mathcal{V}_{\mathcal{B}}, \mathcal{E}_{\mathcal{B}}, \mathcal{A}_{\mathcal{B}}, \mathcal{R}_{\mathcal{B}}, \mathcal{X}_{\mathcal{B}}) \leftarrow \text{HeteroSample}(G, \mathcal{B}, L)$;

3 **for** $s \in \mathcal{V}_{\mathcal{B}}$ **do**

4 $$H_s^{(0)} = \begin{cases} \text{Linear}(\mathbf{x_s}) & \mathbf{1}_{\mathcal{X}}(s) = 1 \\ \text{Dropout}(\mathbf{E}_s) & \mathbf{1}_{\mathcal{V}}(s) = 1 \\ 0_d & \text{otherwise} \end{cases};$$

5 **end**

6 **for** $l = 1$ **to** L **do**

7 $H^{(l)} = \text{GeLU}(\text{HeteroConv}(G_{\mathcal{B}}, H^{(l-1)}))$;

8 **end**

9 $Z_{\mathcal{B}} = \{H_t^{(L)} | t \in \mathcal{B}\}$;

10 **return** $Z_{\mathcal{B}}$

Specifically, we use the budget-based subgraph sampling algorithm and the heterogeneous graph transformer (HGT) proposed by Hu et al. [32]. HGT captures topological, meta-topological, and contentual aspects by combining off-the-shelf graph convolution with node type-specific projection and edge type-based attention.

Algorithm 1 provides a full overview of the primary embedding algorithm. The basic idea is to infer node representation from its k-hop heterogeneous neighborhood subgraph G_v while handling edge cases introduced by the incompleteness constraints (Definition 3.3) in order to handle web-scale multimodal graphs. The inference starts by sampling a subgraph $G_{\mathcal{B}}$ given a batch of central nodes using the *budget sampling* algorithm on line 2. We use node type bound budget required by *budget sampling* algorithm as a multiple of $|\mathcal{B}|$ for each of the layers to hyperparameter retuning for different datasets.

Once the graph is sampled we split the task of initial feature inference into three cases to handle the incompleteness constraints. (i) If a feature vector is present, then it is simply projected into the representation space. (ii) If the node is in the training set while no feature vector is present, then its representation is drawn from the *auxiliary embedding* matrix $\mathbf{E}$. To avoid overreliance on the embeddings in preference for feature vectors we apply dropout on the resulting representation. (iii) Finally, if an unseen node without a feature vector is encountered, the zero vector (denoted as 0_d) is used, indicating that its feature vector has zero weight during the aggregation step of the graph convolution. Note, that as the graph may become too big, it may not be feasible to keep a full auxiliary embedding matrix in memory. In Section 6.3 we explore a setting where auxiliary embeddings are limited to a subset of important nodes.

Given the subgraph and the initial representation vector $H^{(0)}$ L layers of HGT graph convolutions are applied on line 7. Each layer uses the representation vector of the previous layer and feeds its output through a GeLU activation function (See Section 6.5 for performance comparison). Finally, the output vectors at L'th layer are used as primary representation vectors and output for each query node in the batch on line 9.

4.2 Multi-task representation learning

During task-specific learning we focus on two main tasks capturing which capture the intricacies of multimodal networks. The topological task identified by $\mathcal{E}$ focuses on minimizing the representation distance between nodes that are proximate within the network. Analogously, the temporal task $\mathcal{T}$ focuses on minimizing the distance

between nodes that co-occur at the same timeframes. While fundamentally different since the tasks are trained in parallel, they benefit from weight sharing and from node sharing during primary embedding as the subgraph batches are centered around the same nodes.

4.2.1 Task-based attention. An important observation is that while temporal and topological communities are both important during analysis, they are not always correlated. In fact, in most of the benchmarking datasets such as Cora and DBLP temporal features and graph structure show low correlation. While it is very rare that contentual features are completely independent of topology and temporality, we describe a general implementation that can be applied to such a case.

Given the above observation, we admit that it may not be possible to train an embedding that excels at both tasks. To work around this issue while still capturing both tasks in a single embedding vector we introduce *task-based attention*. The basic idea is that while primary embedding extracts suitable features from the multimodal network, task-specific attention selects the most relevant of these features for the task at hand.

Inspired by transformers [66] we define multi-head attention to capture various feature patterns Eq. (32). The task-based transformation function is defined as Eq. (31) where the primary representation vector is attended to using a simple matrix multiplication operation. We specialize this function for topological task as $f^{\mathcal{E}}(\mathbf{Z})$ producing $\mathbf{Z}^{\mathcal{E}}$ and temporal task $f^{\mathcal{T}}(\mathbf{Z})$ producing $\mathbf{Z}^{\mathcal{T}}$.

$$f^{task}(\mathbf{Z}) = \mathbf{Z} \cdot ATT^{task}(\mathbf{Z}) \tag{31}$$

$$ATT^{task}(\mathbf{Z}) = \Big\|_{i \in [1..h]} \sigma\left[\text{Linear}^{i}_{task}(\mathbf{Z})\right] \tag{32}$$

4.2.2 Objective function. In order to learn model parameters in an unsupervised way, we define contrastive loss. Task-specific positive context sample P and a negative context sample $\bar{P}$, both sharing a central query node q are used to construct positive (q, p) and negative (q, n) node pairs respectively. Define a max-margin-based loss function (Eq. (33)) which aims to maximize the inner product similarity between the query and positive examples. On the other hand, the inner product of query and negative samples is minimized to be smaller than that of the positive samples by some predefined value of Δ (See Section 6.5 hyperparameter experiments). In our tests, we found that averaging similarity over positive samples within the max loop helps to smoothen out the noise caused by context sampling Eq. (34).

$$\text{MM-Loss}(\mathbf{Z}, P, \bar{P}, q) = \max_{n \in \bar{P}} \left\{0, \mathbf{Z}_q\mathbf{Z}_n - \widetilde{\mathbf{Z}_q\mathbf{Z}_p} + \Delta\right\} \tag{33}$$

$$\widetilde{\mathbf{Z}_q\mathbf{Z}_p} = \frac{1}{|P|} \sum_{p \in P} \left[\mathbf{Z}_q\mathbf{Z}_p\right] \tag{34}$$

4.2.3 Temporal context sampling. Temporal features are often not correlated with network topology. We propose a separate context sampling function that, given a query node and an interval window ω, returns other nodes occurring within the same time window. The interval window ω is determined by using the dataset statistics as a fraction of the complete time range $\mathcal{T}$. By picking a small enough interval window, a fine-grained continuous-time representation vector can be learned as during sampling it is shifted to be centered around the query node.

Edge cases arising from the incompleteness constraints need to be handled where the nodes are missing timestamps $\mathbf{1}^{\mathcal{T}} = 1$. While the usual semantic approach is to consider these nodes omnipresent (static), the naive window sampling methods quickly get congested with static to static context pairs. Our aim to alleviate this issue using biased sampling in favor of non-static pairs.

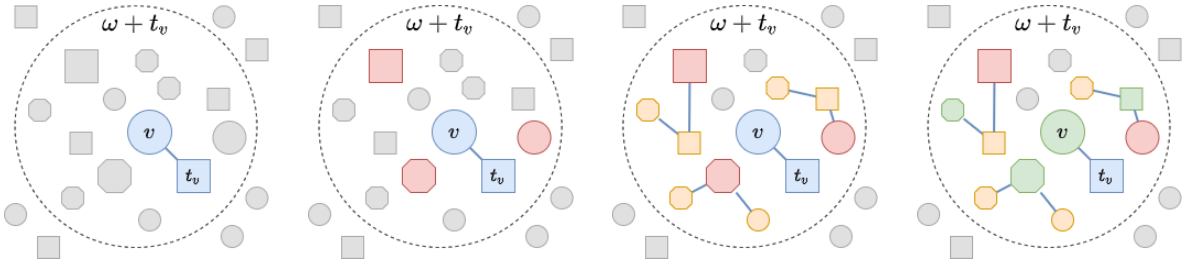


Fig. 4. Visual overview of Ballroom Walk temporal sampling algorithm. (a) The sampling timestamp t_v for query node v is inferred given the nearest neighbor if the node is static (blue). The relative time window is determined as $\omega + t_v$. (b) The root context nodes are sampled from the relative time window (red). (c) Context is extended with temporal random walks from the root nodes (yellow). (d) The context path is sampled from the collected context (green).

Algorithm 2: Temporal Random Walk

1 **Procedure** `TemporalRW()`
  Input: center node v, temporal window ω^*
  Output: Temporal random walk P_l
2 Initialize P_l;
3 $(u, t_u) = (v, \emptyset)$;
4 **for** $i = 1$ **to** l **do**
5 $N(u) = \{w | w \in \mathcal{V}, (u, w) \in \mathcal{E}, \tau(w) \cap \omega^* \neq \emptyset\}$;
6 **if** $N(u) = \emptyset$ **then** `/* Restart on dead end */`
7 $(u, t_u) = (v, t_v)$;
8 **go to** 5;
9 **end**
10 $w \sim N(u)$;
11 $t_u = \max\{t_u, \min \tau(w), \tau((u, w))\}$;
12 $u = w$;
13 $P_l.push((u, u_t))$;
14 **end**
15 **return** P_l

We start by introducing the temporal random walk procedure shown in Algorithm 2 which enforces standard random walks over the network to stay within a predetermined temporal window ω^*. Here random walk of size l is constructed by picking a randomly connected node to the current head node (line 5) within a time window. If no such node is present, then the random walk is restarted from any already picked node line 7. The walk is extended with a new head node until it reaches the desired length.

An outline of our proposed sampling method "Ballroom Walk" is shown in Algorithm 3. It starts by inferring the sampling timestamp t_v by picking a random timestamp the query node v occurs in. If the node is static, the timestamp of its nearest neighbor reachable through temporal random walk is selected (line 3). To reliably sample the temporal neighborhood, n root nodes w are picked occurring in the time-window relative to the sampling timestamp on line 4. Temporal context C is constructed by collecting temporal random walks starting from

Algorithm 3: Ballroom walk sampling

1 **Algorithm** `BallroomWalk()`
  Input: multimodal graph $G = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{R}, \mathcal{X})$, relative temporal window ω, walks per node n, walk length l, center node v,
  Output: Temporal n random walks P_l
2 Initialize C;
3 $t_v = \begin{cases} t_v \sim \tau(v) & \mathbf{1}_{\mathcal{T}}(v) = 1 \\ \texttt{TemporalRW}(v, (-\infty, \infty)).first() & \text{otherwise} \end{cases}$;
4 $N(v) = \{w | w \in \mathcal{V}, \tau(w) \cap \omega + t_v \neq \emptyset\}$;
5 **for** $i = 1$ **to** n **do**
6 $w \sim \mathcal{N}$;
7 $C = C \cup \texttt{TemporalRW}(w, \omega + v_t)$;
8 **end**
9 `RandomPermute`(C);
10 **for** $i = 1$ **to** n **do**
11 $P_l = \{C_j | i \cdot l \leq j < i \cdot l + l\}$;
12 **return** P_l;
13 **end**

root nodes w given a relative time window $\omega + t_v$ on line 7. Finally, l long context paths are created as random subsets of C. Note that because a sampled context is valid for all member nodes, random walk-like throughput optimization can be used by setting a larger window length than context size [54].

Due to timestamp inference, the first- and second-order proximity static to static pairs are ignored. By only passively sampling omnipresent nodes we mitigate the over-saturation issue while still being fair. Most importantly the neighborhood of central nodes is being sampled independently of their topology. By sampling within a temporal window, we avoid not relying on the correlation of temporality with topology.

4.2.4 Graph sampling. The objective of task-specific learning is mainly defined by the context sampling method. As our method allows for inference of primary and task-specific representations for unseen nodes, we assume that topological, meta-topological and contentual features contain enough information / are correlated with the objective of the tasks.

To gather the topological context $P^{\mathcal{E}}$ Node2Vec biased random-walk algorithm is utilized [27]. By choosing a low value for its control parameter q we discourage structural/topological equivalence representation in favor of larger neighborhood exploration (depth-first strategy) which is useful for community representation. Similarly, we use ballroom sampling to collect temporal context $P^{\mathcal{T}}$ of size l as introduced in the previous section. The negative nodes are collected by sampling random nodes from the graph. In our framework, the query nodes and negative contexts **??** are shared across both tasks.

4.3 Community detection

For community detection, we adopt the DPMM split/merge algorithm proposed by Chang and Fisher III [11] as discussed in Definition 3.6. In our implementation, we use Normal Wishart (NW) as a conjugate prior and use variational lower bound in our convergence criteria.

Specifically, we monitor the log sum of the variational lower bound Eq. (35) for the supercluster and subcluster models. The variational lower bound is computed as the product of variational distribution $q(\mathbf{z})$, the normalizing constant of the Dirichlet distribution $B(\alpha_0)$, and the normalizing constant of the Normal Wishart distribution

$C(W, \nu)$. Once its monitored value starts oscillating, we deem the models converged and move into the proposal state. If the model parameters remain unchanged during the proposal stage (no split or merge is accepted), then the clustering is complete.

$$\text{Lower-Bound}(r) = \underbrace{\left[\prod_{n=1}^{N}\prod_{k=1}^{K} r_{nk} e^{r_{nk}}\right]}_{q(\mathbf{z})} \underbrace{\frac{\prod_{i=1}^{K}\Gamma(\alpha_0)}{\Gamma\left(\sum_{k=1}^{K}\alpha_0\right)}}_{B(\alpha_0)} \underbrace{2^{\frac{\nu d}{2}}|W|^{\frac{\nu}{2}}\Gamma_d\left(\frac{\nu}{2}\right)}_{C(W,\nu)} \tag{35}$$

The only parameters relevant for our clustering method are the prior hyperparameters (See Definition 3.6). Most of the parameters (i.e. α, κ, and ν) are not very relevant if they are much smaller than the sample count. We use the Σ_{scale} parameter to scale the dataset covariance for more effective control over the strength of the data-bound prior parameters μ_0, Σ_0. In Section 6.5 we provide a more detailed analysis of result sensitivity to prior parameters.

To fit the clustering model we use primary embedding to calculate assignment and posterior parameters as it contains features relevant for both temporal and topological tasks. While not explored in this thesis, it is worth noting that it is not necessary to have all the embeddings in memory as exact posterior parameters depend on data μ and Σ which can be calculated over multiple batches.

4.4 End-to-end approach

Given a graph embedding, it is straightforward to find communities by performing the embedding and clustering tasks sequentially. This approach lacks a unified objective, thus, the node embeddings may not be optimized for community detection. We extend the objective with cluster-based loss calculated as the distance between node embedding and its assigned cluster z_v Eq. (36). This introduces a feedback loop that encourages the model to reinforce community structures while optimizing the topological and temporal objectives Eq. (39).

The influence of three objectives can be controlled using hyperparameters $\beta^{\mathcal{E}}$, $\beta^{\mathcal{T}}$, and β^{C}.

$$\mathcal{L}^{C} = \left\|Z_v - \mu_{z_v}\right\|_{\ell 2}^{2} \tag{36}$$

$$\mathcal{L}^{\mathcal{E}} = \text{MM-Loss}(\mathbf{Z}, P^{\mathcal{E}}, \bar{P}, v) \tag{37}$$

$$\mathcal{L}^{\mathcal{T}} = \text{MM-Loss}(\mathbf{Z}, P^{\mathcal{T}}, \bar{P}, v) \tag{38}$$

$$\mathcal{L} = \beta^{\mathcal{E}}\mathcal{L}^{\mathcal{E}} + \beta^{\mathcal{T}}\mathcal{L}^{\mathcal{T}} + \beta^{C}\mathcal{L}^{C} \tag{39}$$

With this closed feedback loop, the training procedure consists of two alternating stages (See Fig. 3 and Algorithm 4). The *embedding optimization* stage (line 2), is responsible for optimizing the graph embedding function parameters while keeping cluster parameters θ fixed. Once the graph embeddings are updated we run I_c clustering/EM steps to optimize cluster parameters θ while keeping node representations fixed line 11. Note that the *representation optimization* stage is run until convergence as part of pretraining beforehand to ensure the clusters are initialized properly.

Algorithm 4: MGTCOM learning pipeline

for *subiter* = 1 **to** I **do**
  for $v \in \mathcal{V}$ **do**
    /* Gather context samples */
    $P^{\mathcal{E}}$ = `Node2VecRandomWalk`(G, l, v);
    $P^{\mathcal{T}}$ = `BallroomWalk`(G, ω, l, v);
    $\overline{P} \overset{l}{\sim} \mathcal{V}$ Negative sampling;
    $\mathcal{B} = P_l^{\mathcal{E}} \cup P_l^{\mathcal{T}} \cup \bar{P}_l$;
    $\mathbf{Z}$ = `EmbedPrimary`$(G, \mathcal{B})$;
    Compute task embeddings $Z^{\mathcal{E}}$, $Z^{\mathcal{T}}$ using Eq. (31);
    Compute loss $\mathcal{L}^{\mathcal{E}}$, $\mathcal{L}^{\mathcal{T}}$, $\mathcal{L}^{C}$, $\mathcal{L}$ using Eqs. (36) to (39) given respective context $P^{\mathcal{E}}$, $P^{\mathcal{T}}$;
  end
  for *iter* = 1 **to** I_c **do**
    if $i = 1$ **then**
      Initialize θ using K-means
    end
    Update θ using EM given Z
  end
end

5 EXPERIMENTS

In this section, we investigate the effectiveness of the proposed framework *MGTCOM* (in Section 4) by evaluating its performance on auxiliary tasks related to multimodal networks. We start by describing our experimental setup, whereafter we compare the performance of our model against baseline methods.

5.1 Evaluation metrics

There are no measures that can assess the quality of communities in multimodal networks. Therefore, we evaluate our model component-wise by defining related auxiliary tasks. On a high level, these tasks evaluate the efficiency of topological and temporal node embeddings and found communities. The found communities shall capture important patterns in the data which are useful for further analysis. In order to measure predictive performance over distinct aspects of our data, we first define the following labels for calculating performance metrics, then describe the auxiliary tasks.

- **Ground truth labels** L_y. Various datasets include manually selected ground truth labels which capture valuable higher-order relations within data. By measuring prediction performance on this label we gauge the quality of found communities.
- **Node timestamps** $L_{\mathcal{T}}$. We split the nodes evenly into snapshot labels given the timestamp of their first occurrences. This allows measuring the quality of node embeddings on temporal prediction.
- **Link-based communities** L_G. While other measures such as modularity and link prediction are well-suited for measuring the quality of node embeddings in capturing the structure of a given network, they either require community assignment or measure low-proximity similarity. In order to overcome this, we first identify community labels using the Louvain method [4]. Then we use those labels to assess the quality of individual node embeddings for community detection. As the Louvain method greedily approximates optimal communities, we don't use this label for formal comparison.

5.1.1 Classification (CF). In the classification experiment, we evaluate predictive performance given task-related labels. To elaborate, given a set of node embeddings and their respective ground truth labels, we train a logistic regression model to predict node labels. For the predicted node labels, we calculate average accuracy, F_1-micro, and F_1-macro classification measures.

5.1.2 Link prediction (LP). In this set of experiments, we evaluate link prediction performance. Given a set of positive and negative node pairs, binary classification is used to predict whether an edge exists within the graph. We use a held-out positive and randomly sampled negative sets of edges to train a logistic regression model. The inner-product similarity between a pair of node embeddings is used as input for the model. By repeating this process three times, the average accuracy, ROC AUC, and F_1 score are calculated.

5.1.3 Cluster quality. Given node embeddings and their respective labeling, we calculate the silhouette coefficient and Davies-Bouldin index which are helpful to estimate how coherent a clustering is. In this case, a coherent clustering indicates how well defined the correlated patterns are within the embeddings.

Definition 5.1 (**Davies-Bouldin Index**). Davies-Bouldin Index (DBI) is the ratio of the sum of the average distance to the distance between the centers of mass of the two clusters. In other words, it is defined as a ratio of within-cluster, to the between cluster separation. This measure is defined as an average over all the found clusters and is therefore also a good measure to deciding how many clusters should be used (See Eqs. (40) and (41)). The s_i refers to the average distance between each point in cluster i to its cluster center μ_i, and d_{ij} refers to the distance between cluster centers μ_i and μ_j

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{40}$$

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij} \tag{41}$$

$$\tag{42}$$

5.1.4 Link-based Community quality. In this experiment, we measure link-based community quality. Girvan and Newman [23] defined community structure as a group of nodes where inter-community connectivity is higher than intra-connectivity. Following this definition, they introduce a modularity measure to evaluate the quality of found communities in a given network. We make use of this measure in our empirical evaluation. Note that we use modularity (Definition 5.2) to measure the quality of topological communities.

Definition 5.2 (**Modularity**). Modularity directly measures the density of links inside a graph and is therefore computed on communities (sets of nodes) individually by weighing edges using community similarity (or exact matching). Calculation of modularity is done by aggregating per community r for each pair of nodes vw the difference between the expected connectivity $\frac{k_v k_w}{2m}$ (expected amount of edges between the nodes) and the actual connectivity A_{vw} (existence of an edge) given their degrees (k_v and k_w). The final result represents the connectivity difference between the current and a random graph, as expected connectivity is determined by random rewirings. Because intracommunity pairs are weighted less than intercommunity pairs, the score can vary. See Eq. (43), where S_{vr} indicates membership of node v for community r, and m represents the total edge count.

$$Q = \frac{1}{2m} \sum_{vw} \sum_{r} \left[\overbrace{A_{vw}}^{\text{Connectivity}} - \underbrace{\frac{k_v k_w}{2m}}_{\text{Expected Connectivity}} \right] \overbrace{S_{vr} S_{wr}}^{\text{Community Similarity}} \tag{43}$$

5.1.5 Ground-truth Community quality (COM). To measure the predictive quality of the embeddings, we defined task-based labels. Similarly, to measure the quality of detected communities for specific tasks, we measure the Normalized Mutual Information Score (NMI) score given a task-based label (Definition 5.3).

Definition 5.3 (**Normalized Mutual Information Score (NMI)).** Normalized Mutual Information is a popular measure used to evaluate network partitioning. It is a variant of a common measure in information theory called Mutual Information defined by $I(X;Y) = H(X) - H(X|Y)$ and represents a reduction in entropy of variable X by observing the random variable Y or vice versa. In the context of ground-truth community evaluation setting this measure is used to quantify the overlap between two sets of partitions. The Mutual Information score for two sets of partitions X and Y is computed as Eq. (44), where $|X|$ is the size of set X, $X)_i$ refers to i'th partition of set X, and N is the total number of data points. Finally, the NMI score is computed by normalizing the MI score using the arithmetic mean of entropy of respective partitions Eq. (45).

$$MI(X;Y) = \sum_{i=1}^{|X|} \sum_{j=1}^{|Y|} \frac{\left|X_i \cap Y_j\right|}{N} \log \frac{N \left|X_i \cap Y_j\right|}{|X_i| \left|Y_j\right|} \tag{44}$$

$$NMI(X;Y) = \frac{MI(X;Y)}{(H(X) + H(Y))/2} \tag{45}$$

5.2 Experimental setup

As shown in Section 2 there are no directly comparable methods to ours in terms of features. For a fair and coherent comparison define three variants of the MGTCOM model for evaluation. In addition to the complete end-to-end model *MGTCOM*, we split our framework into a temporal model $MGTCOM^{\mathcal{T}}$ and topological model $MGTCOM^{\mathcal{E}}$, by removing $\mathcal{L}_{\mathcal{T}}$ and $\mathcal{L}_{\mathcal{E}}$ from the objective respectively.

For evaluation, we split the network edges into disjoint training (80%), validation (10%), and testing (10%) sets. During link prediction, we exclusively use links in the respective set as positive pairs. Negative pairs are sampled given the full set of edges. Similarly, the clustering is computed on the training embeddings while cluster-based metrics are calculated using test and validation sets. During the calculation of predictive metrics such as link prediction and classification, we run logistic regression three times and use the average to get an accurate measurement.

5.2.1 Hyperparameters. The hyperparameters for *MGTCOM* model can be attributed to either network architecture, topological random walk, temporal random walk or clustering. In Section 6.5 we explore the sensitivity of our model to these hyperparameters. In Appendix A we display a complete overview of all the hyperparameter values used for evaluation. The most important hyperparameters are specified below.

For primary embedding, we use two HGT layers with neighborhood sampling sizes of 8 and 4. All the hidden dimensions are equal to the representation dimension, which is 64 ($d = 64$). For temporal and topological context sampling we use walk length $l = 10$ with 10 walks per node. Node2Vec is configured to use $q = 0.5$ to favor neighborhood exploration. The temporal sampling window ω for ballroom walk is determined for each dataset by

Table 3. Dataset statistics. *Temporal* indicates if a dataset is temporal and *labelled* refers to the availability of ground truth labels.

Dataset	Node type	# Nodes	Edge type	# Edges	Temporal	Labelled
DBLP	Author (A)	5,162	A - Authored - P	11,022	•	•
	Paper (P)	5,511	P - Published In - V	5,511		
	Venue (V)	14				
IMDB	Person (P)	8,491	P - Directed - M	4,939	•	
	Movie (M)	5,043	P - Acted In - M	15,086		
	Genre (G)	26	M - Tagged - G	14,504		
SDS	User (U)	34,919	U - Tweeted - T	56,173	•	
	Hashtag (H)	2,341	T - Reply To - U	21,769		
	Tweet (T)	56,173	T - Reply To - T	4,296		
			T - Quote - T	882		
			T - Mention - U	70,367		
			T - Mention - H	12,313		
			U - Follows - U	5,649,098		
ICEWS	Entity (E)	10,463	123 different types	915,028	•	
Cora	Paper (P)	2,708	P - Cites - P	10556		•

splitting $\mathcal{T}$ into 20 even partitions. For the clustering module we define prior parameters as $\nu = d+1, \kappa = 1, \alpha = 10$ and $\Sigma_{scale} = 0.05$. We set trade-off parameters as $\beta_{\mathcal{E}} = 1, \beta_{\mathcal{T}} = 1, \beta_C = 0.01$. For max-margin loss we set Δ to 0.1.

5.2.2 Baselines. We use various graph embeddings and community detection algorithms as baselines, covering state-of-art developments in related fields. For the baselines, we use the hyperparameters reported in their respective papers. To keep the results comparable, we use representation dimension $d = 64$ throughout.

- **ComE [9]** uses Gaussian mixture model to learn homogeneous graph embeddings and cluster parameters jointly while utilizing random walk based context sampling.
- **GEMSEC [57]** uses random walks to learn community structure and embeddings simultaneously on homogeneous graphs.
- **CP-GNN [42]** learns node embeddings from a heterogeneous graph by utilizing transformers and k-hop context sampling.
- **CTDNE [50]** utilizes time-based biased random walks to learn spatio-temporal node representations from dynamic networks.
- **GraphSAGE [29]** uses k-hop neighborhood sampling to learn node embeddings from homogeneous graphs. Its unsupervised variant combines contrastive link sampling with hinge loss.
- **Node2Vec [27]** adopts biased random walk and Skip-Gram to learn node embeddings from homogeneous graphs.

5.2.3 Datasets. We use four widely used real-world (temporal) datasets for evaluation. These graphs are of different types and contain information on different modalities. We applied additional preprocessing on the IMDB, DBLP-HCN, and ICEWS datasets to include the multimodal features present in the datasets but often not included in the graph due to sparsity of temporal or content-based features. See Table 3 for a detailed comparison of node features.

- **DBLP** [84] is a citation network consisting of Authors, Papers and Venues. Aside from being heterogeneous, the dataset also contains timestamps representing paper publication dates and abstracts. There are thirteen ground-truth communities representing publication venues. The used graph contains 10687 nodes and 33066 edges. This dataset includes ground truth labels.
- **ICEWS** [22] is a temporal knowledge graph in which nodes represent entities and timestamped edges the relationship between them. We model this data as a highly heterogeneous network consisting of different types of nodes (10463 in total) connected by 915028 timestamped edges. Edges are labeled with relations.
- **IMDB5000** [1] network consists of Actor, Director, Movie, and Genre nodes where each Movie node type has a timestamp denoting the release date. Additionally, each actor node has a set of attributes characterizing information unique to the actor such as age and popularity, while movies have box-office data and keywords encoded as feature vectors. This network includes 13560 nodes and 69058 edges.
- **SocialDistancingStudents (SDS)** [71] represents a small part of the Twitter network around a set of hashtags related to the COVID pandemic. This heterogeneous network models connections between Users, Tweets, and Hashtags where parallel edges are possible due to relations such as tweeted, retweeted, quoted, etc. The tweet nodes contain post timestamps and content encoded as feature vectors. 93433 nodes and 7420366 edges are included.
- **Cora** [86] is a homogeneous citation network. Nodes represent published papers and contain feature vectors representing specific term occurrences in the abstract. Each node is associated with one of the seven ground-truth labels.

5.3 Performance Comparison

In this experiment, we evaluate the performance of learned node embeddings and detected clusters. In particular, we evaluate the predictive quality of embeddings using classification and link prediction, i.e., link prediction accuracy (LP_{ACC}), temporal $L_{\mathcal{T}}$ and ground truth L_y label classification accuracy CF_{ACC}. We evaluate the quality of detected clusters by calculating their NMI score based on predefined ground-truth communities L_y, $L_{\mathcal{T}}$, L_G. This tells us how whether detected clusters approximate user-defined communities L_y, temporal partitioning $L_{\mathcal{T}}$ or the topology L_G. Additionally, we calculate cluster and community quality scores for the learned community assignments, specifically Davies Bouldin score and modularity.

The embeddings obtained from non-community detection methods were clustered using k-means clustering with $K = 20$. Similarly, we use $K = 20$ for community detection methods (ComE, GEMSEC, CP-GNN) that assume a predefined cluster count.

The results are reported in Table 4. It can be seen that while MGTCOM is competitive on task-specific measures such as link prediction and timestamp prediction, the community detection methods still have an edge on link-based modularity measures. A possible explanation for this would be the fact that the DPMM process is more prone to getting stuck in local minima as the clusters split and merge. Another possibility is that node features do not contain enough information to model very specific network features such as modularity. In Section 6.3 we further explore this issue by varying the auxiliary embedding ratio.

While CTDNE performs comparatively well in capturing the temporal aspect of the network, we see that it still yields inferior results on datasets where temporal features are weakly correlated with topology.

It is an interesting thing to note that algorithms that rely on pairwise loss measures such as GraphSAGE and CP-GNN perform relatively well on classification-based measures while performing very poorly on cluster quality measures such as DBI and modularity. A possible explanation for such observation is that the combination of neighborhood sampling and pairwise loss reinforces structural similarity despite having a large receptive field. Our method successfully overcomes this issue by modifying Hinge loss to work in a context path setting (See Section 4.2.2.

Table 4. Comparison of performance of baselines on multimodal graph learning tasks. ("-" means no data available, for example for temporal methods on static datasets such as Cora). The calculated metrics are the link prediction accuracy (LP_{ACC}), predictive accuracy on ground truth communities CF_{ACC} L_y, timestamp predictive accuracy CF_{ACC} $L_{\mathcal{T}}$, NMI score of detected communities (COM_{NMI}) given predefined communities (L_y, $L_{\mathcal{T}}$, L_G), Davies-Bouldin Index (DBI) and Modularity.

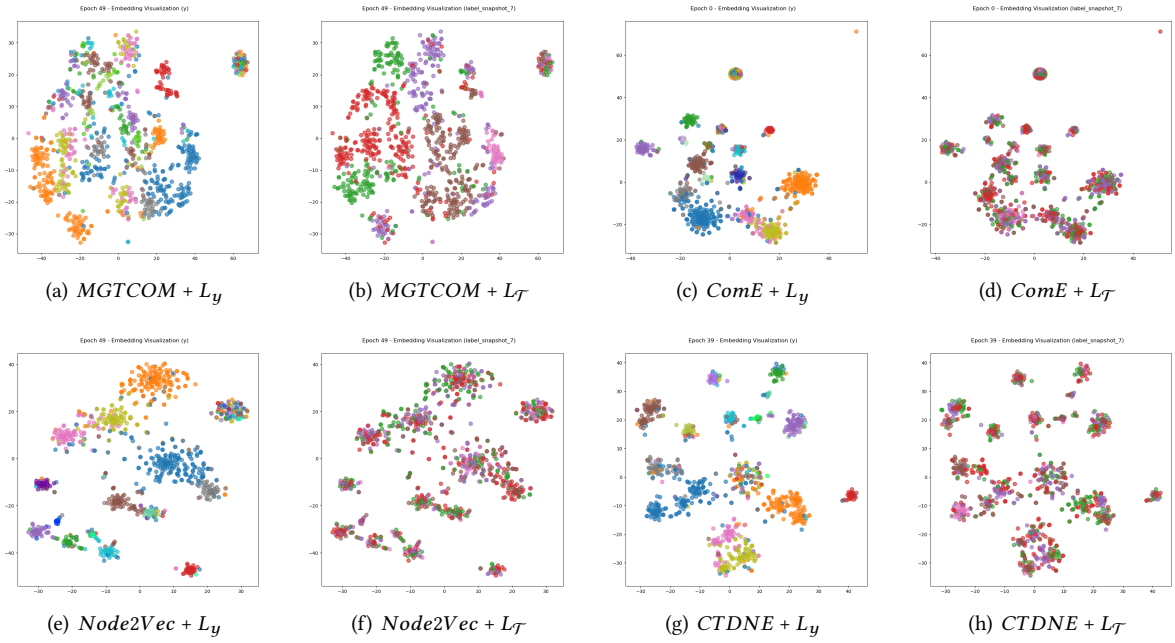
Dataset		GraphSAGE	Node2Vec	ComE	GEMSEC	CTDNE	CP-GNN	*MGTCOM*	$MGTCOM^{\mathcal{T}}$	$MGTCOM^{\mathcal{E}}$
DBLP	LP_{ACC}	0.624	0.710	0.735	0.544	0.701	0.522	0.743	0.634	**0.794**
	CF_{ACC} L_y	0.315	0.832	0.842	0.831	0.809	0.506	**0.896**	0.330	0.884
	CF_{ACC} $L_{\mathcal{T}}$	0.309	0.308	0.328	0.324	0.488	0.313	**0.758**	0.508	0.320
	COM_{NMI} L_y	0.051	**0.549**	0.463	0.385	0.537	0.209	0.465	0.059	0.492
	COM_{NMI} $L_{\mathcal{T}}$	0.006	0.033	0.025	0.022	0.059	0.022	**0.209**	0.168	0.026
	COM_{NMI} L_G	0.040	0.425	**0.470**	0.314	0.401	0.107	0.336	0.039	0.371
	DBI	0.472	2.305	2.205	4.056	1.206	4.780	2.039	4.205	**5.188**
	Modularity	0.028	**0.662**	0.636	0.492	0.642	-0.035	0.427	0.137	0.514
ICEWS	LP_{ACC}	0.525	0.936	0.880	0.768	0.921	0.709	0.903	0.896	**0.945**
	CF_{ACC} $L_{\mathcal{T}}$	0.294	0.301	0.264	0.310	0.285	0.273	0.316	**0.318**	0.313
	COM_{NMI} $L_{\mathcal{T}}$	0.018	0.040	0.015	0.022	0.022	0.013	**0.057**	0.002	0.011
	COM_{NMI} L_G	0.227	0.354	**0.548**	0.309	0.347	0.204	0.119	0.001	0.447
	DBI	1.027	1.697	2.559	3.867	1.533	**4.737**	3.883	3.598	3.182
	Modularity	0.218	0.215	**0.483**	0.311	0.239	0.199	0.007	0.001	0.390
IMDB	LP_{ACC}	0.714	0.757	0.666	0.637	0.728	0.598	0.721	0.724	**0.773**
	CF_{ACC} $L_{\mathcal{T}}$	0.346	0.373	0.394	0.380	0.488	0.316	**0.659**	0.556	0.377
	COM_{NMI} $L_{\mathcal{T}}$	0.022	0.025	0.031	0.013	0.065	0.004	**0.239**	0.231	0.026
	COM_{NMI} L_G	0.039	0.181	**0.197**	0.094	0.160	0.033	0.107	0.031	0.158
	DBI	0.301	1.803	3.840	**4.951**	1.749	4.806	2.257	1.285	4.013
	Modularity	-0.172	0.190	**0**.395	0.073	0.196	0.053	0.119	0.114	0.286
SDS	LP_{ACC}	0.922	0.953	0.758	0.878	0.955	-	0.934	0.616	**0.956**
	CF_{ACC} $L_{\mathcal{T}}$	0.521	0.445	0.386	0.384	0.447	-	0.523	**0.887**	0.492
	COM_{NMI} $L_{\mathcal{T}}$	0.250	0.149	0.117	0.015	0.161	-	0.204	**0.536**	0.044
	COM_{NMI} L_G	0.186	0.277	0.346	0.117	0.233	-	0.120	0.043	**0.389**
	DBI	1.108	2.355	**3.986**	3.410	2.890	-	2.474	1.519	2.559
	Modularity	0.088	0.163	0.301	0.037	0.016	-	0.015	0.005	**0.374**
Cora	LP_{ACC}	0.505	0.939	**0.962**	0.923	-	0.829	-	-	0.958
	CF_{ACC} L_y	0.659	0.798	**0.864**	0.845	-	0.780	-	-	0.854
	COM_{NMI} L_y	0.376	0.345	0.434	0.437	-	0.370	-	-	**0.439**
	COM_{NMI} L_G	0.507	0.543	0.635	0.632	-	0.501	-	-	**0.643**
	DBI	1.526	1.250	2.021	1.500	-	2.634	-	-	**2.647**
	Modularity	0.636	0.691	**0.785**	0.780	-	0.677	-	-	0.754

We also observe that the MGTCOM model performs well on both topology and temporal prediction tasks in comparison to its task-specific counterparts.

5.4 Qualitative Results

We further compare *MGTCOM* and the baseline models on the DBLP-HCN network. We apply the T-SNE dimensionality reduction technique to visualize the trained node embeddings in 2d space colored by the ground truth label and the node timestamp (See Fig. 5).

Since in the DBLP-HCN dataset the timestamps are weakly correlated with its topology, we can see that topology-focused embedding (and community detection) methods such as ComE and Node2Vec do not capture temporal relations of nodes. On contrary, we observe distinct patterns emerge when looking at *MGTCOM*



(a) $MGTCOM + L_y$ (b) $MGTCOM + L_{\mathcal{T}}$ (c) $ComE + L_y$ (d) $ComE + L_{\mathcal{T}}$

(e) $Node2Vec + L_y$ (f) $Node2Vec + L_{\mathcal{T}}$ (g) $CTDNE + L_y$ (h) $CTDNE + L_{\mathcal{T}}$

Fig. 5. Visualization of trained embedding against ground truth labels (L_y, left) and timestamp labels ($L_{\mathcal{T}}$, right) for DBLP-HCN dataset. (Note: The embeddings are calculated on the training dataset. Each of the plots contains a blob of nodes that have no edges in the training set due to the validation split. None of the methods is equipped to handle disconnected nodes.)

generated embeddings for both of the labels. Similar to that of *ComE* the community structures are visible in the node embeddings though they are not as distinct.

5.5 Inference Results

Because the *MGTCOM* model operates on sampled neighborhood subgraphs, in contrast to other methods it can operate in an inductive setting. Meaning that it is not necessary to retrain the model to infer representation vectors for previously unseen nodes.

We evaluate the performance of *MGTCOM* and its task-specific variants in inductive settings by controlling the ratio of nodes in the training set to the validation set. The training set remains constant throughout the experiment to accurately assess performance on inferred nodes. The relevant quality measures are computed exclusively on the test set and can be found in Table 5.

In figure Fig. 6 we see the same measures plotted with the training ratio on the x-axis. From figure Fig. 6 (a) we observe that varying training set size does not affect link-prediction tasks as much as node classification tasks (b, c, d). Throughout the measures, we can see that using only 75% of the data does not substantially affect the results. Finally, it is an interesting observation can be made that the variance on the temporal prediction task increases when more data is provided.

5.6 Learnable parameter reduction

An important goal of our work is to prove that inductive-based community detection is feasible. We address the structural similarity bias found in many unsupervised inductive algorithms by introducing a custom loss and

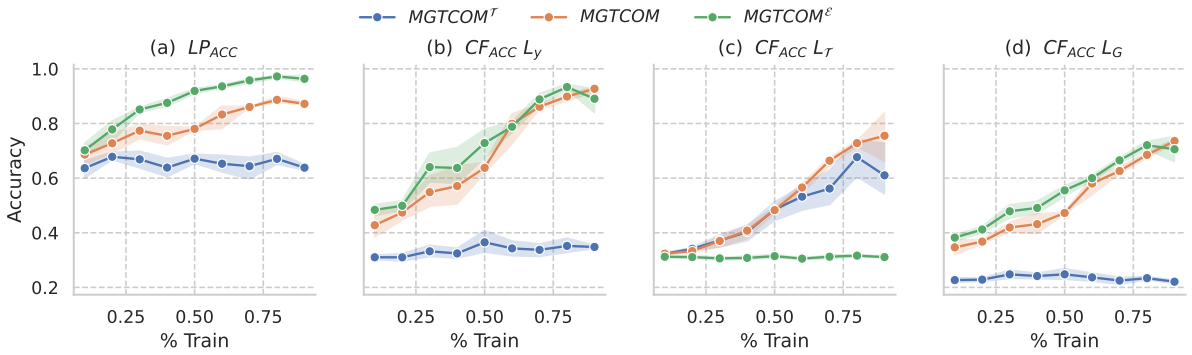


Fig. 6. Visual comparison of different model variants in the inference-based setting. Graph nodes are split into three disjointed sets (train, validation, and test). The metrics are measured while the training to validation ratio is varied. The test set is set to 10% of the nodes and is kept constant. The average metrics per data value are plotted along with their standard deviation.

Table 5. Comparison of different model variants in the inference-based setting. Graph nodes are split into three disjointed sets (train, validation, and test). The metrics are measured while the training to validation ratio is varied. The test set is set to 10% of the nodes and is kept constant.

Model	**% Train**	10%	20%	30%	40%	50%	60%	70%	80%	90%
MGTCOM	LP_{ACC}	0.686	0.728	0.774	0.755	0.780	0.833	0.861	0.887	0.872
	CF_{ACC} L_y	0.428	0.474	0.548	0.571	0.638	0.799	0.861	0.899	0.927
	CF_{ACC} $L_{\mathcal{T}}$	0.323	0.333	0.370	0.408	0.483	0.566	0.664	0.728	0.755
	CF_{ACC} L_G	0.347	0.368	0.419	0.432	0.473	0.580	0.626	0.685	0.736
MGTCOM$^{\mathcal{E}}$	LP_{ACC}	0.702	0.778	0.851	0.876	0.919	0.936	0.958	0.972	0.963
	CF_{ACC} L_y	0.484	0.499	0.640	0.637	0.729	0.788	0.888	0.933	0.891
	CF_{ACC} $L_{\mathcal{T}}$	0.312	0.311	0.306	0.308	0.314	0.305	0.312	0.316	0.311
	CF_{ACC} L_G	0.382	0.412	0.479	0.491	0.555	0.600	0.665	0.721	0.706
MGTCOM$^{\mathcal{T}}$	LP_{ACC}	0.636	0.678	0.669	0.639	0.671	0.653	0.644	0.671	0.638
	CF_{ACC} L_y	0.310	0.310	0.332	0.324	0.365	0.343	0.337	0.352	0.348
	CF_{ACC} $L_{\mathcal{T}}$	0.323	0.341	0.372	0.403	0.485	0.532	0.562	0.677	0.610
	CF_{ACC} L_G	0.227	0.228	0.248	0.242	0.248	0.237	0.225	0.234	0.221

sampling methodology in Section 4.2.2. While our model still utilizes embeddings to address the incompleteness constraints, we show in Section 6.3 that importance-based pruning is an effective optimization to keep the model scalable.

As result, our model reaps the benefits of scalability characteristic of inductive community detection methods. In Table 6 we compare the parameter count of the *MGTCOM* model to the node2vec model which directly learns node embeddings. Overall *MGTCOM* has a lower number of parameters since the model size is bound by meta-topology. In highly heterogeneous graphs such as the ICEWS dataset, the number of parameters may become larger than expected.

Table 6. Parameter count comparison between node2vec and the *MGTCOM* model.

Dataset	node2vec	*MGTCOM*
DBLP	683,968	173,910
ICEWS	669,632	1,072,302
IMDB	867,840	170,846
SDS	5,979,712	231,282
Cora	173,312	136,390

6 ABLATION STUDIES

In this section, we investigate the sensitivity of our model to the described design choices and hyperparameter values. Throughout the experiments keep the same base parameters as described in the experimental setup. Similarly, the DBLP dataset is used throughout as it provides a wide range of features suitable for the evaluation of all supported tasks.

6.1 Auxiliary Embedding Ratio

To address the incompleteness constraints, *MGTCOM* introduces auxiliary embeddings for nodes without features. Zero-vector features are used for nodes that are unseen during training and don't have their own feature vector to encourage its inference from neighboring nodes. While doing this introduces performance benefits, for large datasets it may not be possible to store the auxiliary embeddings in memory.

We define a procedure to work around this scaling issue by noting that embeddings only need to be constructed for a fraction of the most important nodes. This is due to scaling laws applicable to most real-world networks. Specifically, in this experiment, we sort all the nodes without features by their degree and use a fraction of the highest degree nodes for auxiliary embeddings. Other nodes are given a zero-vector upon inference.

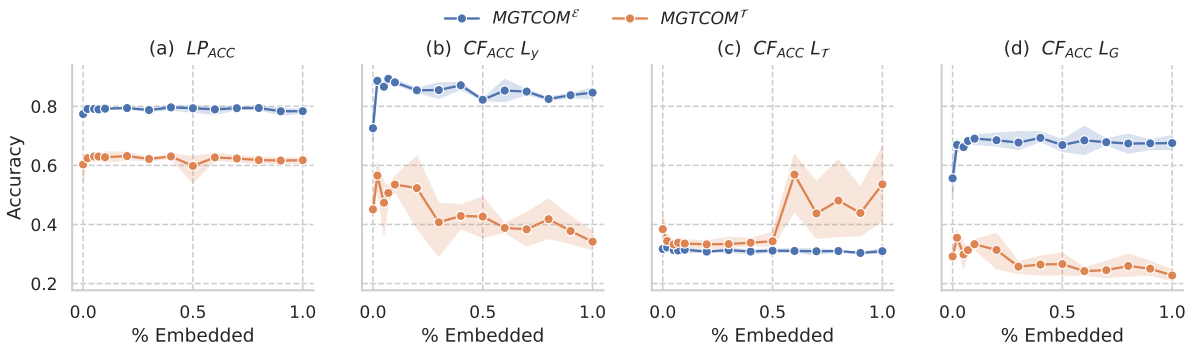


Fig. 7. Performance results for topological $MGTCOM^{\mathcal{E}}$ and temporal $MGTCOM^{\mathcal{T}}$ models on various tasks where the ratio of auxiliary embedded nodes varies.

In Fig. 7 we see the results of the tasks specific models when the auxiliary ratio is varied. From figure (a) we can observe that while auxiliary embeddings don't have a large influence during link prediction, they are in fact necessary on prediction tasks as figures (b), (c), and (d) indicate. It can be rightfully deduced that embeddings are necessary for temporal tasks (figure (c)) since topology and content-based features are weakly correlated with temporal features.

6.2 Meta-topological features

Meta-topological features are an important part of multimodal graphs. In this experiment, we aim to determine the importance of meta-topology in our evaluation setting. We measure performance measures on heterogeneous and homogeneous variants of the DBLP dataset. By varying convolutional layers between Heterogeneous Graph Transformer and GraphSAGE [15] (each edge type has a separate set of weights), we additionally aim to determine the importance of meta-topology-based attention used during the aggregation step.

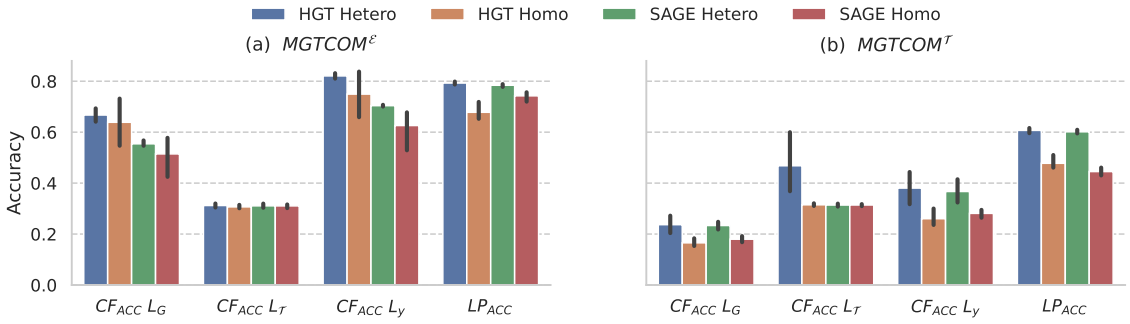


Fig. 8. Performance results for (a) topological $MGTCOM^{\mathcal{E}}$ and (b) temporal $MGTCOM^{\mathcal{T}}$ models, on prediction tasks for heterogeneous and homogeneous variants of the DBLP dataset. To determine the importance of meta-topological attention we vary the convolutional layers HGT and GraphSAGE (which is adopted for heterogeneous graphs).

Overall we see that the addition of meta-topological features has a positive effect on the classification performance of both topological as well as temporal models. This effect is especially pronounced on link prediction and topology-based classification tasks for the temporal model. The cause for this may be that while topological features are not provided during training, meta-topology still conveys enough information about the topology.

From the results, we see that meta-topology-based attention yields benefits in classification performance in contrast to naive aggregation techniques.

6.3 Trade-off Parameter

During analysis the trade-off parameters ($\beta^{\mathcal{E}}$, $\beta^{\mathcal{T}}$, $\beta^{\mathcal{C}}$) are used to guide the trained embeddings to favor specific tasks. In this experiment, we explore the trade-off between temporal and topological tasks by varying value of $\beta^{\mathcal{E}}$, $\beta^{\mathcal{T}}$ while setting constraint $1 = \beta^{\mathcal{E}} + \beta^{\mathcal{T}}$.

In figure Fig. 9 (a) we can see an almost linear correlation between link prediction accuracy and the topological weight parameter $\beta^{\mathcal{E}}$. On the other hand, in figures (b) and (d) we see a more logarithmic curve for topology correlated classification measures. The most interesting takeaway is that while variance is quite high on the temporal classification task, its curve peaks at a value of 0.5. In further work, it may be worth exploring this phenomenon in more detail. The most probable assumption would be that the temporal model still benefits from the fact that temporal features are weakly correlated with the topology.

6.4 Initial K sensitivity

We evaluate the sensitivity of the clustering results to the initial K value selection. While our method does not require setting the cluster to count K, it can still be set to find more accurate initial clustering, and help DPMM avoid local minima. In this experiment, we have varied the initial cluster count while keeping all other parameters fixed.

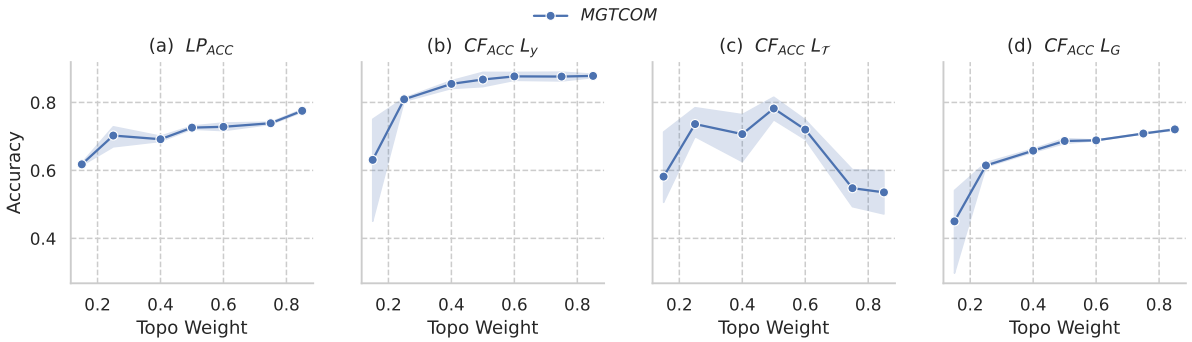


Fig. 9. Performance of *MGTCOM* model while varying topological loss weight parameter $\beta^{\mathcal{T}}$ under $1 = \beta^{\mathcal{E}} + \beta^{\mathcal{T}}$ contraint.

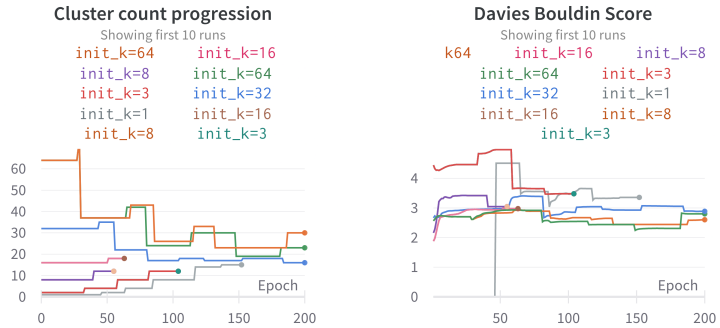


Fig. 10. (a) and (b): Cluster count progression during DPMM clustering given an initial cluster count. The clustering is done on pretrained *MGTCOM* embeddings for the DBLP dataset.

In Fig. 10 (a) we see that despite varying starting values, all the runs converge at 12-18 cluster range. Having a value that strongly deviates from the "optimal" cluster count causes a slower convergence since more split/merge operations are required. We can see a similar pattern in the measured Davies-Bouldin index in Fig. 8 (b).

6.5 Hyperparameter sensitivity

In this part, the sensitivity of other hyperparameters on the model performance is discussed.

The node2vec random walk algorithm used for the topological task relies on parameters such as walk length l, the number of random walks started for each node n, and the exploration trade-off parameter q. In Fig. 11 we see that while the choice of random walk length has a significant impact on link-prediction and classification performance (a), the model is not as sensitive to the other parameters. A surprising observation is that the trade-off parameter does not significantly affect the productivity accuracy of ground communities (CF_{ACC} L_y). A possible explanation for this may be the fact that we stack random walk and neighborhood sampling algorithms making the trade-off ineffective.

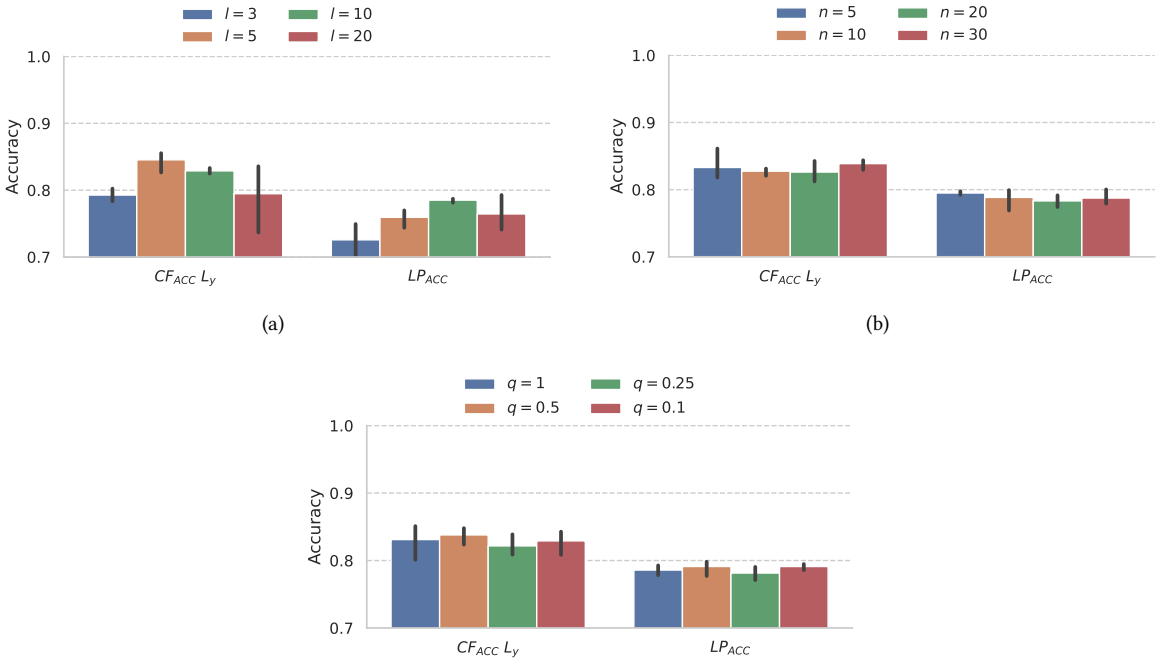


Fig. 11. Performance of $MGTCOM^{\mathcal{E}}$ model with varying (a) random walks length l, (b) number of random walks per node n, and (c) the exploration trade-off parameter q.

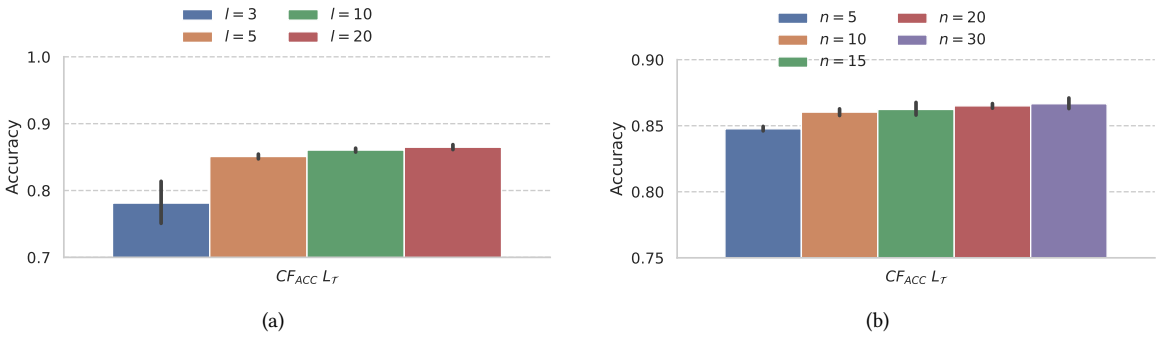


Fig. 12. Performance of $MGTCOM^{\mathcal{T}}$ model with varying (a) random walks length l and (b) number of random walks per node n.

The ballroom walk algorithm introduced in Section 4.2.3 similarly relies on the walk length l and the number of random walks started for each node n hyperparameters, though they serve a different purpose. Increasing either the l or the n parameter only marginally increases the models performance at timestamp prediction (See

Fig. 12). For both parameters, there is a positive correlation between performance and an increase in the receptive field.

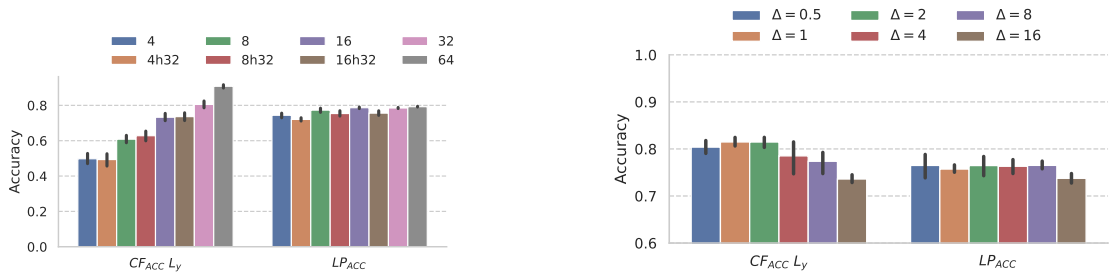


Fig. 13. Performance of $MGTCOM^{\mathcal{E}}$ model with varying (a) representation dimension d, and (b) the margin (Δ) parameter for the hinge loss.

The most sensitive/important parameter for our model is the representation dimension size d. In Fig. 13 (a) we plot the predictive performance of the topological model while varying the model representation dimension d and the hidden representation dimension used in in-between layers of graph convolution. The link-prediction performance seems to benefit the most from a larger d, while classification only sees a marginal improvement. Moreover having hidden dimension size deviate from the representation dimension only seems to degrade the model performance.

In Fig. 13 (b) we vary the margin parameter of hinge loss. It is conventional to use $\Delta = 1$ if the similarity is bounded (as is in our case), therefore we can see the model performance degrade as the margin exceeds this threshold. Increasing loss beyond 1 amplifies the relative relevance of small loss samples, which in turn makes the model more prone to noise.

In Fig. 14 we vary the convolution architecture of the topological model and measure the resulting test performance. As observed earlier HGT convolutional layers perform better since they introduce meta-topology-based attention. Varying the layer neighborhood size does not seem to affect the performance substantially, except for the fact that computed performance measures during training are a lot smoother throughout.

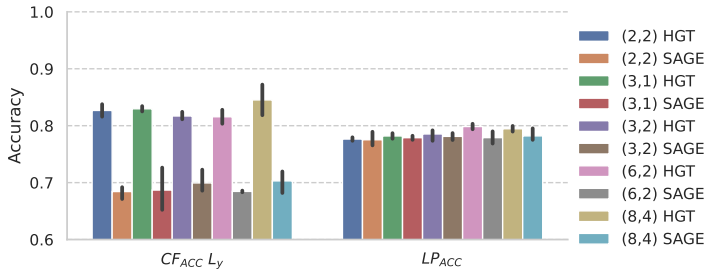


Fig. 14. Performance of $MGTCOM^{\mathcal{E}}$ model with varying convolution layers. We vary the architecture by switching between Heterogeneous Graph Transformer (HGT) and Heterogenous GraphSAGE convolutional layers. Similarly, we also modify the neighborhood size of each node within the two-layer convolution setup. Format (x, y) represents the number of neighbors per node in the first (x) and second (y) layer respectively.

7 CASE STUDY: SOCIAL DISTANCING STUDENTS DATASET

Work in progress

8 FUTURE WORK

While we have explored a multitude of topics, the is still a lot of room for further improvements and exploration. Below we list a multitude of possible further exploration scenarios.

Our experiments have shown that temporal representation learning benefits greatly from auxiliary embeddings as node features may often be too weakly correlated with temporality. In contrast to topological tasks, auxiliary embeddings have been shown to be effective only for the most important nodes. In future work, it may be valuable to explore more flexible settings where representations are augmented with embeddings only for temporal tasks, therefore reducing parameters and inference latency.

The scale of our model is meta-topology bound, meaning that amount of learnable parameters increases if there are more node or edge types. This reduces the effectiveness of our framework on highly heterogeneous networks such as knowledge graphs. Future works may explore improvements to our embedding method by utilizing techniques used knowledge graph embedding field.

While detected communities excel in topological and temporal predictive capabilities, the detected communities still under-perform on the modularity measure. Further work may explore swapping node2vec random walk algorithm by motif-sampling [33] to encourage strong link-based proximity.

The presented framework uses DPMMSC algorithm as introduced in the original paper [11]. Meanwhile, a multitude of works has been published that extend the algorithm to a deep learning setting [56] or that address local minima issues faced by the algorithm. Hierarchical DPMM algorithms have been studied [10, 63] and may be invaluable for community detection in analytical settings. Our clustering implementation can be further improved by exploring the effectiveness of different priors and introducing new split/merge proposal methods. Finally, we note that the detected communities are mainly dictated by the structure of node embeddings. Introducing a control parameter to bias communities towards temporal and topological communities would improve ergonomics of community detection when reusing the learned embeddings.

9 CONCLUSION

In this paper, we introduce the MGTCOM framework for community detection in multimodal graphs. It utilizes meta-topological, topological, content features, and temporal information to detect communities. Moreover, we address common issues in multimodal graphs such as information incompleteness, and inference on unseen data by adopting a graph convolutional network architecture that combines k-hop neighborhood sampling and random walk context sampling. We devise a unified objective and an efficient temporal sampling method to learn multimodal community-aware node embeddings in an unsupervised manner. Consequently, we leverage a split/merge-based Dirichlet process mixture model for community detection where the number of communities are not known a priori. Our empirical evaluation shows that MGTCOM is quite competitive with the state-of-the-art.

REFERENCES

[1] [n.d.]. IMDB 5000 Movie Dataset. https://kaggle.com/carolzhangdc/imdb-5000-movie-dataset.
[2] Mihael Ankerst, Markus M. Breunig, Hans-peter Kriegel, and Jörg Sander. 1999. OPTICS: Ordering Points To Identify the Clustering Structure. ACM Press, 49–60.
[3] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer, New York.
[4] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (Oct. 2008), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008
[5] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems*, Vol. 26. Curran Associates, Inc.
[6] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. 2014. Spectral Networks and Locally Connected Networks on Graphs. https://doi.org/10.48550/arXiv.1312.6203 arXiv:1312.6203 [cs]

[7] Jinxin Cao, Di Jin, Liang Yang, and Jianwu Dang. 2018. Incorporating Network Structure with Node Contents for Community Detection on Large Networks Using Deep Learning. *Neurocomputing* 297 (July 2018), 71–81. https://doi.org/10.1016/j.neucom.2018.01.065

[8] Yuwei Cao, Hao Peng, Jia Wu, Yingtong Dou, Jianxin Li, and Philip Yu. 2021. Knowledge-Preserving Incremental Social Event Detection via Heterogeneous GNNs. 3383–3395. https://doi.org/10.1145/3442381.3449834

[9] Sandro Cavallari, Vincent W. Zheng, Hongyun Cai, Kevin Chen-Chuan Chang, and Erik Cambria. 2017. Learning Community Embedding with Community Detection and Node Embedding on Graphs. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, Singapore Singapore, 377–386. https://doi.org/10.1145/3132847.3132925

[10] Jason Chang. [n.d.]. Sampling in Computer Vision and Bayesian Nonparametric Mixtures. ([n. d.]), 241.

[11] Jason Chang and John W Fisher III. 2013. Parallel Sampling of DP Mixture Models Using Sub-Cluster Splits. In *Advances in Neural Information Processing Systems*, Vol. 26. Curran Associates, Inc.

[12] Weijian Chen, Fuli Feng, Qifan Wang, Xiangnan He, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. CatGCN: Graph Convolutional Networks with Categorical Node Features. *IEEE Transactions on Knowledge and Data Engineering* (2021), 1–1. https://doi.org/10.1109/TKDE.2021.3133013

[13] Jun Jin Choong, Xin Liu, and Tsuyoshi Murata. 2018. Learning Community Structure with Variational Autoencoder. In *2018 IEEE International Conference on Data Mining (ICDM)*. 69–78. https://doi.org/10.1109/ICDM.2018.00022

[14] Shib Sankar Dasgupta, Swayambhu Nath Ray, and Partha Talukdar. 2018. HyTE: Hyperplane-based Temporally Aware Knowledge Graph Embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2001–2011. https://doi.org/10.18653/v1/D18-1225

[15] CSIRO's Data61. 2018. Stargazers · Stellargraph/Stellargraph. https://github.com/stellargraph/stellargraph.

[16] Yuxiao Dong, Nitesh V. Chawla, and Ananthram Swami. 2017. Metapath2vec: Scalable Representation Learning for Heterogeneous Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. Association for Computing Machinery, New York, NY, USA, 135–144. https://doi.org/10.1145/3097983.3098036

[17] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*. AAAI Press, Portland, Oregon, 226–231.

[18] Hossein Fani, Eric Jiang, Ebrahim Bagheri, Feras Al-Obeidat, Weichang Du, and Mehdi Kargar. 2020. User Community Detection via Embedding of Social Network Structure and Temporal Content. *Information Processing & Management* 57, 2 (March 2020), 102056. https://doi.org/10.1016/j.ipm.2019.102056

[19] Illés Farkas, Dániel Ábel, Gergely Palla, and Tamás Vicsek. 2007. Weighted Network Modules. *New Journal of Physics* 9, 6 (June 2007), 180–180. https://doi.org/10.1088/1367-2630/9/6/180

[20] Santo Fortunato. 2010. Community Detection in Graphs. *Physics Reports* 486, 3 (Feb. 2010), 75–174. https://doi.org/10.1016/j.physrep.2009.11.002

[21] Tao-yang Fu, Wang-Chien Lee, and Zhen Lei. 2017. HIN2Vec: Explore Meta-paths in Heterogeneous Information Networks for Representation Learning. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17)*. Association for Computing Machinery, New York, NY, USA, 1797–1806. https://doi.org/10.1145/3132847.3132953

[22] Alberto García-Durán, Sebastijan Dumancic, and Mathias Niepert. 2018. Learning Sequence Encoders for Temporal Knowledge Graph Completion. 4816–4821. https://doi.org/10.18653/v1/D18-1516

[23] M. Girvan and M. E. J. Newman. 2002. Community Structure in Social and Biological Networks. *Proceedings of the National Academy of Sciences* 99, 12 (June 2002), 7821–7826. https://doi.org/10.1073/pnas.122653799

[24] Marko Gosak, Rene Markovič, Jurij Dolenšek, Marjan Slak Rupnik, Marko Marhl, Andraž Stožer, and Matjaž Perc. 2018. Network Science of Biological Systems at Different Scales: A Review. *Physics of Life Reviews* 24 (March 2018), 118–135. https://doi.org/10.1016/j.plrev.2017.11.003

[25] Palash Goyal, Sujit Rokka Chhetri, and Arquimedes Canedo. 2020. Dyngraph2vec: Capturing Network Dynamics Using Dynamic Graph Representation Learning. *Knowledge-Based Systems* 187 (Jan. 2020), 104816. https://doi.org/10.1016/j.knosys.2019.06.024

[26] Derek Greene, Dónal Doyle, and Pádraig Cunningham. 2010. Tracking the Evolution of Communities in Dynamic Social Networks. In *2010 International Conference on Advances in Social Networks Analysis and Mining*. 176–183. https://doi.org/10.1109/ASONAM.2010.17

[27] Aditya Grover and Jure Leskovec. 2016. Node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 855–864. https://doi.org/10.1145/2939672.2939754

[28] Loni Hagen, Thomas Keller, Stephen Neely, Nic DePaula, and Claudia Robert-Cooperman. 2018. Crisis Communications in the Age of Social Media: A Network Analysis of Zika-Related Tweets. *Social Science Computer Review* 36, 5 (Oct. 2018), 523–541. https://doi.org/10.1177/0894439317721985

[29] William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 1025–1035.

[30] Xiaofei He, Deng Cai, Yuanlong Shao, Hujun Bao, and Jiawei Han. 2011. Laplacian Regularized Gaussian Mixture Model for Data Clustering. *IEEE Trans. Knowl. Data Eng.* 23 (Sept. 2011), 1406–1418. https://doi.org/10.1109/TKDE.2010.259

[31] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous Graph Transformer. In *Proceedings of The Web Conference 2020*. ACM, Taipei Taiwan, 2704–2710. https://doi.org/10.1145/3366423.3380027

[32] Mingqing Huang, Qingshan Jiang, Qiang Qu, Lifei Chen, and Hui Chen. 2022. Information Fusion Oriented Heterogeneous Social Network for Friend Recommendation via Community Detection. *Applied Soft Computing* 114 (Jan. 2022), 108103. https://doi.org/10.1016/j.asoc.2021.108103

[33] Yuting Jia, Qinqin Zhang, Weinan Zhang, and Xinbing Wang. 2019. CommunityGAN: Community Detection with Generative Adversarial Nets. In *The World Wide Web Conference*. ACM, San Francisco CA USA, 784–794. https://doi.org/10.1145/3308558.3313564

[34] Yoonsuk Kang, Jun-Seok Lee, Won-Yong Shin, and Sang-Wook Kim. 2021. Community Reinforcement: An Effective and Efficient Preprocessing Method for Accurate Community Detection. *Knowledge-Based Systems* (Nov. 2021), 107741. https://doi.org/10.1016/j.knosys.2021.107741

[35] Seyed Mehran Kazemi, Rishab Goel, Kshitij Jain, Ivan Kobyzev, Akshay Sethi, Peter Forsyth, and Pascal Poupart. 2020. Representation Learning for Dynamic Graphs: A Survey. *Journal of Machine Learning Research* 21, 70 (2020), 1–73.

[36] Thomas N. Kipf and Max Welling. 2016. Variational Graph Auto-Encoders. *arXiv:1611.07308 [cs, stat]* (Nov. 2016). arXiv:1611.07308 [cs, stat]

[37] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv:1609.02907 [cs, stat]* (Feb. 2017). arXiv:1609.02907 [cs, stat]

[38] Mark Kozdoba and Shie Mannor. 2015. Community Detection via Measure Space Embedding. In *Advances in Neural Information Processing Systems*, Vol. 28. Curran Associates, Inc.

[39] Jussi M. Kumpula, Mikko Kivelä, Kimmo Kaski, and Jari Saramäki. 2008. Sequential Algorithm for Fast Clique Percolation. *Physical Review E* 78, 2 (Aug. 2008), 026109. https://doi.org/10.1103/PhysRevE.78.026109

[40] Ye Li, Chaofeng Sha, Xin Huang, and Yanchun Zhang. 2018. Community Detection in Attributed Graphs: An Embedding Approach. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (April 2018).

[41] Hongtao Liu, Hui Chen, Mao Lin, and Yu Wu. 2014. Community Detection Based on Topic Distance in Social Tagging Networks. *Indonesian Journal of Electrical Engineering and Computer Science* 12, 5 (May 2014), 4038–4049.

[42] Linhao Luo, Yixiang Fang, Xin Cao, Xiaofeng Zhang, and Wenjie Zhang. 2021. Detecting Communities from Heterogeneous Graphs: A Context Path-based Graph Neural Network Model. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. Association for Computing Machinery, New York, NY, USA, 1170–1180.

[43] Sedigheh Mahdavi, Shima Khoshraftar, and Aijun An. 2020. Dynamic Joint Variational Graph Autoencoders. In *Machine Learning and Knowledge Discovery in Databases (Communications in Computer and Information Science)*, Peggy Cellier and Kurt Driessens (Eds.). Springer International Publishing, Cham, 385–401. https://doi.org/10.1007/978-3-030-43823-4_32

[44] Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. 2000. Automating the Construction of Internet Portals with Machine Learning. *Information Retrieval* 3, 2 (July 2000), 127–163. https://doi.org/10.1023/A:1009953814988

[45] Miller Mcpherson, Lynn Smith-Lovin, and James Cook. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 27 (Jan. 2001), 415. https://doi.org/10.3410/f.725356294.793504070

[46] Nikhil Mehta, Lawrence Carin Duke, and Piyush Rai. 2019. Stochastic Blockmodels Meet Graph Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 4466–4474.

[47] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. *arXiv:1310.4546 [cs, stat]* (Oct. 2013). arXiv:1310.4546 [cs, stat]

[48] Erxue Min, Xifeng Guo, Qiang Liu, Gen Zhang, Jianjing Cui, and Jun Long. 2018. A Survey of Clustering With Deep Learning: From the Perspective of Network Architecture. *IEEE Access* 6 (2018), 39501–39514. https://doi.org/10.1109/ACCESS.2018.2855437

[49] M. E. J. Newman. 2004. Fast Algorithm for Detecting Community Structure in Networks. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* 69, 6 Pt 2 (June 2004), 066133. https://doi.org/10.1103/PhysRevE.69.066133

[50] Giang Hoang Nguyen, John Boaz Lee, Ryan A. Rossi, Nesreen K. Ahmed, Eunyee Koh, and Sungchul Kim. 2018. Continuous-Time Dynamic Network Embeddings. In *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*. ACM Press, Lyon, France, 969–976. https://doi.org/10.1145/3184558.3191526

[51] Shirui Pan, Ruiqi Hu, Guodong Long, Jing Jiang, Lina Yao, and Chengqi Zhang. 2018. Adversarially Regularized Graph Autoencoder for Graph Embedding. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18)*. AAAI Press, Stockholm, Sweden, 2609–2615.

[52] Aldo Pareja, Giacomo Domeniconi, Jie Chen, Tengfei Ma, Toyotaro Suzumura, Hiroki Kanezashi, Tim Kaler, Tao Schardl, and Charles Leiserson. 2020. EvolveGCN: Evolving Graph Convolutional Networks for Dynamic Graphs. *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (April 2020), 5363–5370. https://doi.org/10.1609/aaai.v34i04.5984

[53] Leto Peel, Daniel B. Larremore, and Aaron Clauset. 2017. The Ground Truth about Metadata and Community Detection in Networks. *Science Advances* 3, 5 (2017), e1602548. https://doi.org/10.1126/sciadv.1602548

[54] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. DeepWalk: Online Learning of Social Representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. Association for Computing Machinery, New York, NY, USA, 701–710. https://doi.org/10.1145/2623330.2623732

[55] Stephen D. Pryke. 2004. Analysing Construction Project Coalitions: Exploring the Application of Social Network Analysis. *Construction Management and Economics* 22, 8 (Oct. 2004), 787–797. https://doi.org/10.1080/0144619042000206533

[56] Meitar Ronen, Shahaf E. Finder, and Oren Freifeld. 2022. DeepDPM: Deep Clustering With an Unknown Number of Clusters. *arXiv:2203.14309 [cs, stat]* (March 2022). arXiv:2203.14309 [cs, stat]

[57] Benedek Rozemberczki, Ryan Davies, Rik Sarkar, and Charles Sutton. 2019. GEMSEC: Graph Embedding with Self Clustering. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '19)*. Association for Computing Machinery, New York, NY, USA, 65–72. https://doi.org/10.1145/3341161.3342890

[58] Philipp Schuetz and Amedeo Caflisch. 2008. Multistep Greedy Algorithm Identifies Community Structure in Real-World and Computer-Generated Networks. *Physical Review E* 78, 2 (Aug. 2008), 026112. https://doi.org/10.1103/PhysRevE.78.026112

[59] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. 2008. Collective Classification in Network Data. *AI Magazine* 29, 3 (Sept. 2008), 93–93. https://doi.org/10.1609/aimag.v29i3.2157

[60] Xing Su, Shan Xue, Fanzhen Liu, Jia Wu, Jian Yang, Chuan Zhou, Wenbin Hu, Cecile Paris, Surya Nepal, Di Jin, Quan Z. Sheng, and Philip S. Yu. 2022. A Comprehensive Survey on Community Detection With Deep Learning. *IEEE Transactions on Neural Networks and Learning Systems* (2022), 1–21. https://doi.org/10.1109/TNNLS.2021.3137396

[61] Yizhou Sun, Charu Aggarwal, and Jiawei Han. 2012. Relation Strength-Aware Clustering of Heterogeneous Information Networks with Incomplete Attributes. *Proceedings of the VLDB Endowment* 5 (Jan. 2012). https://doi.org/10.14778/2140436.2140437

[62] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: Large-scale Information Network Embedding. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1067–1077. https://doi.org/10.1145/2736277.2741093

[63] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical Dirichlet Processes. *J. Amer. Statist. Assoc.* 101, 476 (Dec. 2006), 1566–1581. https://doi.org/10.1198/016214506000000302

[64] Fei Tian, Bin Gao, Qing Cui, Enhong Chen, and Tie-Yan Liu. 2014. Learning Deep Representations for Graph Clustering. *Proceedings of the AAAI Conference on Artificial Intelligence* 28, 1 (June 2014).

[65] Stefano F. Tonellato. 2020. Bayesian Nonparametric Clustering as a Community Detection Problem. *Computational Statistics & Data Analysis* 152 (Dec. 2020), 107044. https://doi.org/10.1016/j.csda.2020.107044

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.

[67] Daixin Wang, Peng Cui, and Wenwu Zhu. 2016. Structural Deep Network Embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, San Francisco California USA, 1225–1234. https://doi.org/10.1145/2939672.2939753

[68] Fei Wang, Tao Li, Xin Wang, Shenghuo Zhu, and Chris Ding. 2011. Community Discovery Using Nonnegative Matrix Factorization. *Data Min. Knowl. Discov.* 22 (May 2011), 493–521. https://doi.org/10.1007/s10618-010-0181-y

[69] Hongwei Wang and Jure Leskovec. 2021. Combining Graph Convolutional Neural Networks and Label Propagation. *ACM Transactions on Information Systems* 40, 4 (Nov. 2021), 73:1–73:27. https://doi.org/10.1145/3490478

[70] Peizhuo Wang, Lin Gao, and Xiaoke Ma. 2017. Dynamic Community Detection Based on Network Structural Perturbation and Topological Similarity. *Journal of Statistical Mechanics: Theory and Experiment* 2017, 1 (Jan. 2017), 013401. https://doi.org/10.1088/1742-5468/2017/1/013401

[71] Shihan Wang, Marijn Schraagen, Erik Tjong Kim Sang, and Mehdi Dastani. 2020. Public Sentiment on Governmental COVID-19 Measures in Dutch Social Media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.nlpcovid19-2.17

[72] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. https://doi.org/10.48550/arXiv.2002.10957 arXiv:2002.10957 [cs]

[73] Xiao Wang, Peng Cui, Jing Wang, Jian Pei, Wenwu Zhu, and Shiqiang Yang. [n.d.]. Community Preserving Network Embedding. ([n. d.]), 7.

[74] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. 2019. Heterogeneous Graph Attention Network. In *The World Wide Web Conference (WWW '19)*. Association for Computing Machinery, New York, NY, USA, 2022–2032. https://doi.org/10.1145/3308558.3313562

[75] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge Graph Embedding by Translating on Hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI'14)*. AAAI Press, Québec City, Québec, Canada, 1112–1119.

[76] Chengyuan Wu and Carol Hargreaves. 2020. *Topological Machine Learning for Mixed Numeric and Categorical Data.*

[77] Jiaming Wu, Meng Liu, Jiangting Fan, Yong Liu, and Meng Han. 2021. SageDy: A Novel Sampling and Aggregating Based Representation Learning Approach for Dynamic Networks. In *Artificial Neural Networks and Machine Learning – ICANN 2021 (Lecture Notes in Computer Science)*, Igor Farkaš, Paolo Masulli, Sebastian Otte, and Stefan Wermter (Eds.). Springer International Publishing, Cham, 3–15. https://doi.org/10.1007/978-3-030-86383-8_1

[78] Xiaodong Wu, Weizhe Lin, Zhilin Wang, and Elena Rastorgueva. 2020. Author2Vec: A Framework for Generating User Embedding. *arXiv:2003.11627 [cs, stat]* (March 2020). arXiv:2003.11627 [cs, stat]

[79] Jierui Xie, Stephen Kelley, and Boleslaw K. Szymanski. 2013. Overlapping Community Detection in Networks: The State-of-the-Art and Comparative Study. *Comput. Surveys* 45, 4 (Aug. 2013), 43:1–43:35. https://doi.org/10.1145/2501654.2501657

[80] Shan Xue, Jie Lu, and Guangquan Zhang. 2019. Cross-Domain Network Representations. *Pattern Recognition* 94 (Oct. 2019), 135–148. https://doi.org/10.1016/j.patcog.2019.05.009

[81] Carl Yang, Mengxiong Liu, Zongyi Wang, Liyuan Liu, and Jiawei Han. 2017. Graph Clustering with Dynamic Embedding. *arXiv:1712.08249 [physics]* (Dec. 2017). arXiv:1712.08249 [physics]

[82] Carl Yang, Yuxin Xiao, Yu Zhang, Yizhou Sun, and Jiawei Han. 2020. Heterogeneous Network Representation Learning: A Unified Framework with Survey and Benchmark. *IEEE Transactions on Knowledge and Data Engineering* PP (Dec. 2020), 1–1. https://doi.org/10.1109/TKDE.2020.3045924

[83] Jaewon Yang and Jure Leskovec. 2012. Community-Affiliation Graph Model for Overlapping Network Community Detection. In *2012 IEEE 12th International Conference on Data Mining*. 1170–1175. https://doi.org/10.1109/ICDM.2012.139

[84] Jaewon Yang and Jure Leskovec. 2012. Defining and Evaluating Network Communities Based on Ground-Truth. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics (MDS '12)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/2350190.2350193

[85] Liang Yang, Xiaochun Cao, Dongxiao He, Chuan Wang, Xiao Wang, and Weixiong Zhang. 2016. Modularity Based Community Detection with Deep Learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*. AAAI Press, New York, New York, USA, 2252–2258.

[86] Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. 2016. Revisiting Semi-Supervised Learning with Graph Embeddings. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48 (ICML'16)*. JMLR.org, New York, NY, USA, 40–48.

[87] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, and Jure Leskovec. 2018. Graph Convolutional Neural Networks for Web-Scale Recommender Systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. Association for Computing Machinery, New York, NY, USA, 974–983. https://doi.org/10.1145/3219819.3219890

[88] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. 1996. BIRCH: An Efficient Data Clustering Method for Very Large Databases. *ACM SIGMOD Record* 25, 2 (June 1996), 103–114. https://doi.org/10.1145/235968.233324

[89] Tianqi Zhang, Yun Xiong, Jiawei Zhang, Yao Zhang, Yizhu Jiao, and Yangyong Zhu. 2020. CommDGI: Community Detection Oriented Deep Graph Infomax. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 1843–1852. https://doi.org/10.1145/3340531.3412042

[90] Kai Zhao, Ting Bai, Bin Wu, Bai Wang, Youjie Zhang, Yuanyu Yang, and Jian-Yun Nie. 2020. Deep Adversarial Completion for Sparse Heterogeneous Information Network Embedding. In *Proceedings of The Web Conference 2020*. Association for Computing Machinery, New York, NY, USA, 508–518.

[91] Yifeng Zhao, Xiangwei Wang, Hongxia Yang, Le Song, and Jie Tang. 2019. Large Scale Evolving Graphs with Burst Detection. (2019), 4412–4418.

[92] Yujing Zhou, Weile Liu, Yang Pei, Lei Wang, Daren Zha, and Tianshu Fu. 2019. Dynamic Network Embedding by Semantic Evolution. In *2019 International Joint Conference on Neural Networks (IJCNN)*. 1–8. https://doi.org/10.1109/IJCNN.2019.8852247

[93] Jiong Zhu, Mark Heimann, Yujun Yan, Lingxiao Zhao, Leman Akoglu, and Danai Koutra. [n.d.]. Beyond Homophily in Graph Neural Networks: Current Limitations and Effective Designs. ([n. d.]), 12.

[94] Ruimin Zhu and Wenxin Jiang. 2016. Combining Random Walks and Nonparametric Bayesian Topic Model for Community Detection. https://doi.org/10.48550/arXiv.1607.05573 arXiv:1607.05573 [stat]

[95] Ruimin Zhu and Wenxin Jiang. 2019. Bayesian Complex Network Community Detection Using Nonparametric Topic Model. In *Complex Networks and Their Applications VII (Studies in Computational Intelligence)*, Luca Maria Aiello, Chantal Cherifi, Hocine Cherifi, Renaud Lambiotte, Pietro Lió, and Luis M. Rocha (Eds.). Springer International Publishing, Cham, 280–291. https://doi.org/10.1007/978-3-030-05411-3_23

A SUPPLEMENTAL MATERIAL

A.1 Dataset construction

The datasets IMDB, DBLP and ICEWS were preprocessed into multimodel networks for evaluation of our model. In the following sections we discuss the construction of these multimodal networks.

A.1.1 IMDB. IMDB dataset is originally made up of rows detailing movies and their information. To construct the multimodal we normalize this dataset by splitting listed actors, directors and genres per movie as a separate entity. Actors and directors are merged as person entity since both sets overlap. Each movie is associated with a set of keywords. By collecting these words into a vocabulary of 80 most frequent keywords, we construct 80-dimensional one-hot feature vector for each movie. The Genre and Actor entities have no feature representations. Each movie is given $[t_v, \infty]$ time range, where t_v is the release data of the respective movie. Edges are constructed as person directed movie, person acted in movie, and movie has genre pairs.

A.1.2 DBLP. The DBLP dataset is constructed in a similar way as IMDB since the dataset consists of Papers with their respective citations, authors and venues. To construct the representation vector each paper, we use pretrained sentence transformers [72] to embed the abstract text. Authors and venues have no features. Each paper is given $[t_v, \infty]$ time range, where t_v is the publication date of the respective movie. Edges are constructed as paper cites paper, author wrote paper and paper was published in venue entity pairs.

A.1.3 ICEWS. The ICEWS datasets consists of triplets between subject predicate and object entities. As ICEWS is a temporal knowledge graph, each triplet is associated with a timestamp. We model this dataset by combining subject and object into one single entity type. Between these entities we create typed edges corresponding to the respective predicate. The name of each entity is embedded into a feature vector using a pretrained language transformer [72]. Each edge is given $[t_v, t_v]$ time range, where t_v represents timestamp when the corresponding triplet was valid.

A.2 Exact model parameters