Heterogeneous Graph Generation: A Hierarchical Approach using Node Feature Pooling

Hritaban Ghosh[*] Changyu Chen[†] Arunesh Sinha[‡]
Shamik Sural[*]

[*]Indian Institute of Technology Kharagpur, India

[†]Singapore Management University, Singapore

[‡]Rutgers University, USA

July 02, 2025

1. Introduction
2. Motivation
3. Dataset
4. Methodology
5. Results

Introduction

- Heterogeneous graphs can be used to model systems in various domains such as social networks, recommendation systems, and biological networks.
- Generating realistic heterogeneous graphs that capture complex interactions among diverse entities is a difficult task.
- The generator has to capture both the node-type distribution and the feature distribution for each node type.
- In this paper, we address the challenges in the generation of heterogeneous graphs employing a two-phase hierarchical approach called HG2NP (Heterogeneous Graph Generation using Node Feature Pooling).

Heterogeneous Graphs

- Heterogeneous graphs are also known as multi-type graphs.
- Characterized by the presence of multiple types of nodes and edges representing diverse entities and relationships.
- We denote a heterogeneous graph by $G = (V, E)$, where V is the set of nodes/vertices and E is the set of edges.
- Given a set of M observed heterogeneous graphs $\{G_i\}_{i=1}^{M}$, the problem of generating heterogeneous graphs is to learn the probability distribution $\mathbb{P}$ of these graphs from which new graphs can be sampled $G \sim \mathbb{P}$.

Motivation

- In order to highlight the importance of a heterogeneous graph, let us take a simplified example of a social network.
- In any social network, an individual can have any relationship (including none) with another individual. Individuals can be part of a group. Groups can be part of a community.

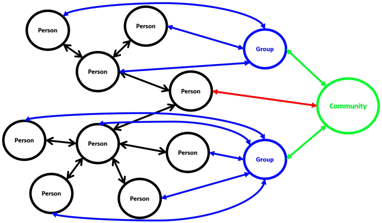

Figure 1: Example Heterogeneous Graph

- This level of detail and meaningful relationships cannot be achieved by homogeneous graphs without several assumptions and constraints.
- The limited availability of real-world data can hinder the development and performance of downstream analytic models, leading to suboptimal results.
- Need for methods to generate synthetic heterogeneous graphs that closely resemble their real-world counterparts.
- Scarcity of research work on heterogeneous graph generation.

Datasets

- Digital Bibliography & Library Project (DBLP) and Internet Movie Database (IMDB).
- The original DBLP dataset is a heterogeneous graph consisting of authors (4,057 nodes), papers (14,328 nodes), terms (7,723 nodes), and conferences (20 nodes). $\sim$ [26,128 nodes]
- The original IMDB dataset is a heterogeneous graph consisting of three types of entities - movies (4,278 nodes), actors (5,257 nodes), and directors (2,081 nodes). $\sim$ [11,616 nodes]

Node category	Representation	Size of feature vector
author	Bag-Of-Words of paper keywords	334
paper	Bag-Of-Words of paper titles	4231
term	GloVe vector	50

Table 1: DBLP graph description

Node category	Representation	Size of feature vector
movie	Bag-Of-Words of plot keywords	3066
actor	Mean of associated movies' features	3066
director	Mean of associated movies' features	3066

Processing the Datasets - DBLP

- Modified the DBLP dataset to include another categorical feature - the type of the publication.
- Used categories and their combinations to segregate the large graph into smaller components.
  - author research area (indicated as author)
  - conference
  - type of paper publication
- Restricted ourselves to work with graphs whose number of nodes ≤ 200.

Split criteria	Number of graphs with nodes $\leq$ 200
conference	1
type	4
author and conference	25
author and type	8
conference and type	29
author, conference and type	77

Table 3: DBLP graph sets

Processing the Datasets - IMDB

- Modified the IMDB dataset to include three categorical features - year, language and country.
- Used categories and their combinations to segregate the large graph into smaller components.
  - movie classes (indicated as movie)
  - year
  - language
  - country
- Restricted ourselves to work with graphs whose number of nodes ≤ 200.

Split criteria	Number of graphs with nodes ≤ 200
year	64
language	43
country	56
year and language	246
year and country	441
language and country	111
movie, language and country	179
year, language and country	523
movie, year, language and country	802

Table 4: IMDB graph sets

Methodology

HG2NPis a two-phase scheme, in which the first phase leverages an existing state-of-the-art homogeneous graph generation framework to produce a *skeleton graph* with nodes and edges and node type. Then, in the second phase, we assign feature vectors to the nodes. The second phase is setup as a generative adversarial network.

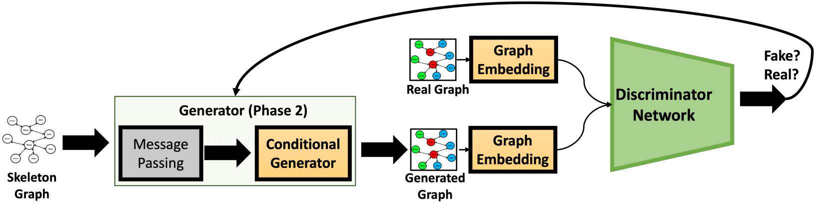

Figure 2: Overall generation process for HG2NP. The skeleton graph is the output of Phase 1. Phase 2 uses a GAN architecture for training.

Methodology

Phase 1: Skeleton Graph Generation

- We leverage DiGress, which is designed for generating homogeneous graphs.
- It is a diffusion based generative model which is utilized to learn the distribution patterns of node type underlying the heterogeneous graphs in our datasets.

Phase 2: Heterogeneous Feature Vector Assignment

- Given a skeleton graph from Phase 1, in the second phase, we first utilize a message passing process for updating each node's type vector to obtain a graph with updated node type vectors

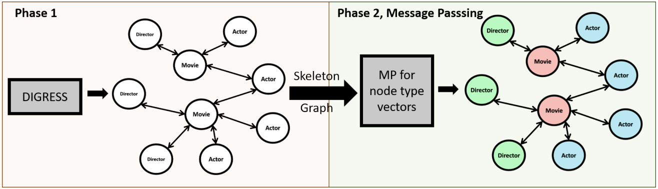

Figure 3: In Phase 1, we use DiGress to output a graph with node types only. Subsequently, the first part in Phase 2 is to perform message passing to embed neighbor node type information in each node type vector.

Methodology

Phase 2: Heterogeneous Feature Vector Assignment

- Conditioned on the updated node type vector, we sample a feature vector from a node type specific pool of feature vectors.
- In this way, all the nodes of the graph are assigned feature vectors. This yields a generated heterogeneous graph.

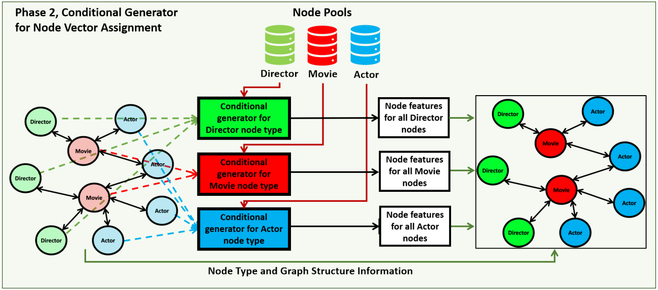

Figure 4: The second part is to perform node vector assignment using conditional generators. The node vectors are picked from the feature vector pool based on the probability distribution generated by the conditional generator of that node type.

Multiple Graphs as Train Data

IMDB

	Year			Country, Language and Movie		
Metrics	VGAE	VGAE-H	HG2NP	VGAE	VGAE-H	HG2NP
Degree Dist.	0.35 ± 0.004	0.34 ± 0.008	**0.18 ± 0.001**	0.30 ± 0.002	0.30 ± 0.003	**0.15 ± 0.001**
Clust. Coeff.	0.92 ± 0.024	0.88 ± 0.015	**5.2e-5 $\pm$ 1.4e-5**	$— \pm —$	0.75 ± 0.012	**7.8e-6 $\pm$ 4.6e-6**
Spect. Dens.	0.38 ± 0.013	0.36 ± 0.011	**0.06 ± 0.003**	0.23 ± 0.003	0.24 ± 0.004	**0.04 ± 0.002**
Type Degree	0.28 ± 0.007	0.26 ± 0.007	**0.05 $\pm$ 3.5e-4**	0.18 ± 0.002	0.17 ± 0.001	**0.03 $\pm$ 2.3e-4**
W-Dist.	43.85e-6 $\pm$ 5.47e-6	49.64e-6 $\pm$ 7.95e-6	**2.93e-6 $\pm$ 2.01e-6**	39.07e-6 $\pm$ 3.20e-6	35.27e-6 $\pm$ 3.41e-6	**8.63e-6 $\pm$ 2.35e-6**
1-NN Accuracy	0.85 ± 0.112	0.79 ± 0.019	**0.38 ± 0.196**	0.47 ± 0.011	0.48 ± 0.009	**0.48 ± 0.010**
FID	2.53e-3 $\pm$ 5.42e-5	2.61e-3 $\pm$ 9.19e-5	**0.99e-3 $\pm$ 1.56e-5**	1.18e-3 $\pm$ 1.20e-5	1.19e-3 $\pm$ 2.76e-5	**1.07e-3 $\pm$ 3.63e-5**

DBLP

	Author and Conference			Author, Conference, and Pub. Type		
Metrics	VGAE	VGAE-H	HG2NP	VGAE	VGAE-H	HG2NP
Degree Dist.	0.41 ± 0.002	0.43 ± 0.007	**0.16 ± 0.004**	0.36 ± 0.003	0.35 ± 0.006	**0.01 ± 0.001**
Clust. Coeff.	$— \pm —$	1.08 ± 0.007	**3.6e-4 $\pm$ 4.7e-5**	$— \pm —$	0.92 ± 0.018	**6.0e-5 $\pm$ 1.3e-5**
Spect. Dens.	0.48 ± 0.023	0.52 ± 0.020	**0.07 ± 0.006**	0.38 ± 0.015	0.36 ± 0.006	**1.2e-3 $\pm$ 3.2e-4**
Type Degree	0.30 ± 0.004	0.29 ± 0.016	**0.04 ± 0.001**	0.22 ± 0.002	0.19 ± 0.003	**1.5e-3 $\pm$ 9.3e-5**
W-Dist.	28.07e-6 $\pm$ 3.50e-6	42.38e-6 $\pm$ 3.80e-6	**0.00 ± 0.000**	58.91e-6 $\pm$ 4.65e-6	63.32e-6 $\pm$ 6.31e-6	**0.00 ± 0.000**
1-NN Accuracy	0.92 ± 0.112	0.98 ± 0.011	**0.61 ± 0.024**	1.00 ± 0.002	1.00 ± 0.000	**0.47 ± 0.053**
FID	3.01e-3 $\pm$ 1.54e-5	3.23e-3 $\pm$ 2.44e-5	**0.87e-3 $\pm$ 6.41e-5**	3.45e-3 $\pm$ 3.72e-5	3.53e-3 $\pm$ 2.07e-5	**0.14e-3 $\pm$ 6.11e-5**

Single Graph as Train Data

IMDB

	Year		
Metrics	VGAE	NetGAN	HG2NP
Degree Dist.	0.37 ± 0.004	0.37 ± 0.004	**0.23 ± 0.002**
Clust. Coeff.	0.95 ± 0.016	0.73 ± 0.008	**6.8e-5 $\pm$ 7.2e-6**
Spect. Dens.	0.61 ± 0.178	0.64 ± 0.168	**0.22 ± 0.131**
Type Degree	0.30 ± 0.006	0.24 $\pm$ 4.6e-4	**0.07 $\pm$ 4.9e-4**
W-Dist.	5.55e-5 $\pm$ 5.01e-6	4.52e-5 $\pm$ 5.06e-6	**1.90e-5 $\pm$ 3.83e-6**
1-NN Accuracy	0.71 ± 0.026	0.78 ± 0.063	**0.55 ± 0.151**
FID	2.61e-3 $\pm$ 7.44e-5	2.32e-3 $\pm$ 1.07e-5	**1.21e-3 $\pm$ 5.85e-5**

DBLP

	Author and Conference		
Metrics	VGAE	NetGAN	HG2NP
Degree Dist.	0.41 ± 0.003	0.73 ± 0.000	**0.17 ± 0.001**
Clust. Coeff.	— $\pm$ —	2.00 ± 0.000	**0.01 $\pm$ 3.3e-4**
Spect. Dens.	0.68 ± 0.080	0.75 ± 0.000	**0.29 ± 0.205**
Type Degree	0.30 ± 0.004	0.36 $\pm$ 4.1e-4	**0.040 $\pm$ 3.1e-4**]
W-Dist.	2.19e-5 $\pm$ 3.38e-6	3.05e-5 $\pm$ 3.61e-6	**1.28e-5 $\pm$ 3.58e-6**
1-NN Accuracy	0.98 ± 0.011	0.99 ± 0.005	**0.63 ± 0.033**
FID	3.04e-3 $\pm$ 2.44e-5	3.11e-3 $\pm$ 1.20e-5	**1.76e-3 $\pm$ 15.95e-5**

Concluding Remarks

- We see that for majority of the cases, our hierarchical model outperforms the baselines.
- The idea of node pooling is used to sample node feature vectors and assign them to the nodes, thereby reconstructing the heterogeneous graphs.
- An important limitation, and potential future work, is to design our approach to explicitly work with edge types.
- Another potential improvement is to have iterative refinement of feature vector assignments. Vectors once assigned are not revisited in HG2NP. Assignment is one shot in the sense that all of them are done in parallel in one go.
- Overall, our work HG2NPprovides a hierarchical approach that successfully generates heterogeneous graphs with promising results for the domains that we experimented with.

Thank you