Towards Addressing Frontiers in Graph Generation

Alex O. Davies

Centre for Doctoral Training in Interactive Artificial Intelligence
University of Bristol
alexander.davies@bristol.ac.uk

Abstract

This doctoral dissertation establishes and addresses frontiers in graph generation. I first apply a Graph Neural Network (GNN) model on social network data, a new domain, to establish what frontiers exist for graph generators. I establish that GNN models are currently limited in the diversity of feature sets that they can produce, the variety of graph structure types they can generate, and highly limited in the size of generated graphs. Further, I find that the quality metrics available for graph generation are aggregate-based and unexpressive. To address the issue of scale I propose Hierarchical Generation of Graphs (HiGGs), a framework for producing graphs orders of magnitude larger than is possible with a single model. As a step towards more expressive metrics I develop Topology only Pre-training (ToP), a pre-training framework for graph models that is capable of representing multiple domains of graphs simultaneously, without relying on tertiary models in downstream applications. The next stage of research will adapt ToP as a metric for graph generators.

Introduction

Graphs are unstructured data used to describe relational structures. Deep-learning methods for graph structures (Graph Neural Networks, GNNs) have typically lagged behind, and been adapted from, methods on more well behaved data. They initially relied on hard node orderings, describing a graph as a sequence, from which a model designed for sequential data could be applied. Gilmer et al. (2017) proposed Message Passing Neural Networks (MPNNs). These parameterise message passing and aggregation, allowing structural information to be strongly included in optimisation.

As with other data-types, generative models for graph data have direct applications, including drug design, synthetic datasets, data augmentation, and many others, but face complexities not present for other forms of data. Early graph generators were primarily rule-based, prior to the development of MPNNs, relying on set procedures. These were parameterised from the original graph datasets (Chakrabarti and Faloutsos 2006), often using a degree distribution or clustering distribution, and broadly meant as an exercise in mathematical modelling.

Related Work

Since MPNNs were developed, deep-learning graph generators have become possible. Early implementations were one-shot, producing whole graphs at once, and so were high order, sometimes up to $O(N^4)$. In the years since autoregressive models have become more efficient (Liao et al. 2019), although expressive denoising diffusion and similar models retain quadratic complexities (Vignac et al. 2023).

The two most researched applications for graph generators have so far been drug design and scene graph generation. Though both have obvious utility, these domains are fairly narrow in scope, with strict construction grammars and limited size. We pick social networks as a useful and challenging case study in graph generation. They are, or can be, richly and diversely attributed, directed, signed, multi-edged or temporal, not constructed according to a rigid grammar, and almost always large.

Case Study In the first part of this work, I apply GNN models in generating social networks as an investigative case study, and identify the current frontiers for graph generation (Davies, Ajmeri, and Silva Filho 2022). I benchmark an archetypal GNN model GRAN (Liao et al. 2019) against two rule-based models, the current status-quo in social network generation, and propose an extended set of aggregate metrics to assess their expressivity. I find that while GRAN out-performs the rule-based models, there are significant obstacles in graph generation:

- Current models are limited in the diversity of feature-sets and structure-types they can produce.
- In-memory costs for graph models are very high. For one-shot methods memory requirements scale at least $O(|V|^2)$ with $|V|$ the number nodes.
- The metrics currently employed in graph generation are simple aggregate measures. Without human qualitative understanding of graph data, model-based metrics are required, but are domain-locked.
- Current graph generators are locked to a single domain.

Very Large Graph Generation

To address the issue of scale I propose Hierarchical Generation of Graphs (HiGGs), a framework for using multiple generative models in hierarchies to produce very large

graphs (Davies, Ajmeri, and Filho 2023). I use graph partitioning to break a graph into hierarchies of resolution. The upper (less granular) hierarchies shows which partitions connect to one another. The lowest hierarchies (more granular) are then individual partitions.

Graphs are generated top-down, upper hierarchy to lower, with each stage conditioned on the previous. This divide-and-conquer approach allows far larger graphs to be generated without the necessity for them to be entirely in-memory, and individual generative stages trivially parallelised. For an n hierarchy HiGGs implementation with a generative model of maximum graph size $|V|$, the possible scale of graphs produced is $|V|^n$, ie a polynomial increase.

Using the graph diffusion model DiGress (Vignac et al. 2023) for each hierarchy, our demonstration implementation HiGGs-DGD produces graphs of over twenty thousand nodes and hundreds of thousands of edges. This is a quadratic increase in possible scale over DiGress alone. Although a few standalone models can reach this number of nodes (Dai et al. 2020), they have significant caveattes on density, and as such cannot produce near as many edges.

Multi-Domain Graph Models

A generalised representation learner does not exist for graph data, and the development of such a model would open avenues in developing both multi-domain graph generators and quality metrics for those graph generators. The main hurdle is that there are hugely varied feature sets on nodes and edges between domains. A social network user might be described by a text embedding, and an atom in a molecule by chirality and 3D position.

LLM approaches to allow multi-domain graph models with features included exist (Liu et al. 2023), but these works are still in their infancy, and requires that features are easily expressed textually. I show that, with features excluded, representation learning techniques can be used to produce multi-domain graph models. I call this pre-training method Topology only Pre-training (ToP) (Davies et al. 2024). I also demonstrate how domain features can be re-included downstream. Transfer is significantly positive on 75% of experiments, and where it is not significantly positive, it is never significantly negative.

A secondary finding from this work is that in the absence of node and edge features, diversity in pre-training data is more beneficial than being in-domain. A model pre-trained only on molecules, compared to a model pre-trained only on non-molecules, performs significantly worse on the majority of datasets. Through application of noise on features and structure, I show that ToP pre-training leads models to rely more strongly on structural information during downstream fine-tuning.

Planned Works

With HiGGs, I demonstrated how simple frameworks can be used to produce very large graphs. ToP in turn demonstrated that a single GNN model can represent multiple domains of graphs. For the rest of my PhD, I plan to both extend the results of ToP, and employ ToP as the basis for expressive graph generation metrics.

Quality metrics in other generative fields are often model-based, primarily as distances between sets of embeddings. In this section, at the time of writing on-going research, I experiment with different normalisations and distance measures between embeddings from ToP models. By using controlled perturbations of graph datasets, I benchmark the response of these constructed metrics, and compare ToP-based metrics to the current aggregates. I hope to show that ToP models respond more expressively than un-pre-trained models, and much more expressively than aggregate metrics, but expect that models pre-trained only on each dataset prove more expressive. The advantage of ToP metrics over these domain-specific models is in their consistency between uses - they do not require users to train their own models, and are not sensitive to differences between metric deployment.

I also plan, once I've established these ToP-based metrics, to significantly expand the scope of my case study for inclusion in my PhD. Here I'll include a wider, and more recent, range of graph generators and datasets. The aim is to represent the broad state of graph generation, with a selection from each specific school, for example auto-regressive and diffusion models.

References

Chakrabarti, D.; and Faloutsos, C. 2006. Graph mining. *ACM Computing Surveys (CSUR)*, 38(1): 2.

Dai, H.; Nazi, A.; Li, Y.; Dai, B.; and Schuurmans, D. 2020. Scalable Deep Generative Modeling for Sparse Graphs. In *Proceedings of the 37th International Conference on Machine Learning*, 2302–2312. PMLR.

Davies, A.; Ajmeri, N.; and Silva Filho, T. 2022. Realistic Synthetic Social Networks with Graph Neural Networks. *arXiv Pre-Print*, 1(1): 1–12.

Davies, A. O.; Ajmeri, N. S.; and Filho, T. M. S. 2023. Size Matters: Large Graph Generation with HiGGs. In *Proceedings of the NeurIPS 2023 Synthetic Data Generation with Generative AI workshop*.

Davies, A. O.; Green, R. W.; Ajmeri, N. S.; and Filho, T. M. S. 2024. Towards Generalised Pre-Training of Graph Models.

Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural Message Passing for Quantum Chemistry. *In Proceedings of the 34th International Conference on Machine Learning, (ICML) 2017*, 3: 2053–2070.

Liao, R.; Li, Y.; Song, Y.; Wang, S.; Hamilton, W.; Duvenaud, D. K.; Urtasun, R.; and Zemel, R. 2019. Efficient Graph Generation with Graph Recurrent Attention Networks. In *In Proceedings of Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, volume 32, 1–11. Vancouver: Curran Associates, Inc.

Liu, H.; Feng, J.; Kong, L.; Liang, N.; Tao, D.; Chen, Y.; and Zhang, M. 2023. One for All: Towards Training One Graph Model for All Classification Tasks. *arXiv Pre-Print*.

Vignac, C.; Krawczuk, I.; Siraudin, A.; Wang, B.; Cevher, V.; and Frossard, P. 2023. DiGress: Discrete Denoising diffusion for graph generation. In *The Eleventh International Conference on Learning Representations*.