# GTLM: Graph Transformer Language model for conditional subgraph generation

## Abstract

## 1 Introduction

Graph node classification and linkage/edge prediction are two important tasks with key applications on graph-structured data. In particular, a linkage prediction graph neural network has important application of gene regulatory networks (GRN) inference. Recently, there has been work which cast both problems into a single generative framework. The current approach, however, is limited as the linearizing algorithm for autoregressive generation is stringent and does not allow for conditioning of partial information of a vertex or an edge (e.g., predicting the type or existence of an edge from its incidence of vertices).

We plan to explore a generative training approach with more flexible linear order of decoding from a partial graph, along with a special masking scheme during training to maximize representation learning and generative capacity. Specifically, given a graph, we represent "true" vertices and edges as unmasked tokens in some random order with some constraints (e.g, no edges are placed before their incident vertices), with a triangular attention mask such that tokens can only attend to itself and the previous ones. The fully or partially masked tokens, or "target" tokens, of the same ordering are concatenated at the end of this "true" sequence. Importantly, the attention mask for the "target" tokens is again triangular with respect to either the "true" tokens or the "target" tokens. The attention masking and training scheme ensures the model can learn to generate under diverse scenarios and conditioning. We plan to evaluate the model on standard graph generation benchmark, as well as the more niche task of GRN inference.

## References

## A Technical Appendices and Supplementary Material

Preprint. Under review.