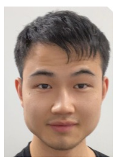
# A Large-scale Training Paradigm for Graph Generative Models



Yu Wang[1]

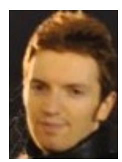

Ryan Rossi[4]


Namyong Park


Huiyuan Chen


Nesreen Ahmed[5]


Puja Trivedi[3]


Franck Dernoncourt[4]


Danai Koutra[3]


Tyler Derr[2]

KIND Lab[1], University of Oregon,
NDS Lab[2], Vanderbilt University
GEMS Lab[3], University of Michigan
Adobe Research[4], Cisco AI Lab[5]

# Background

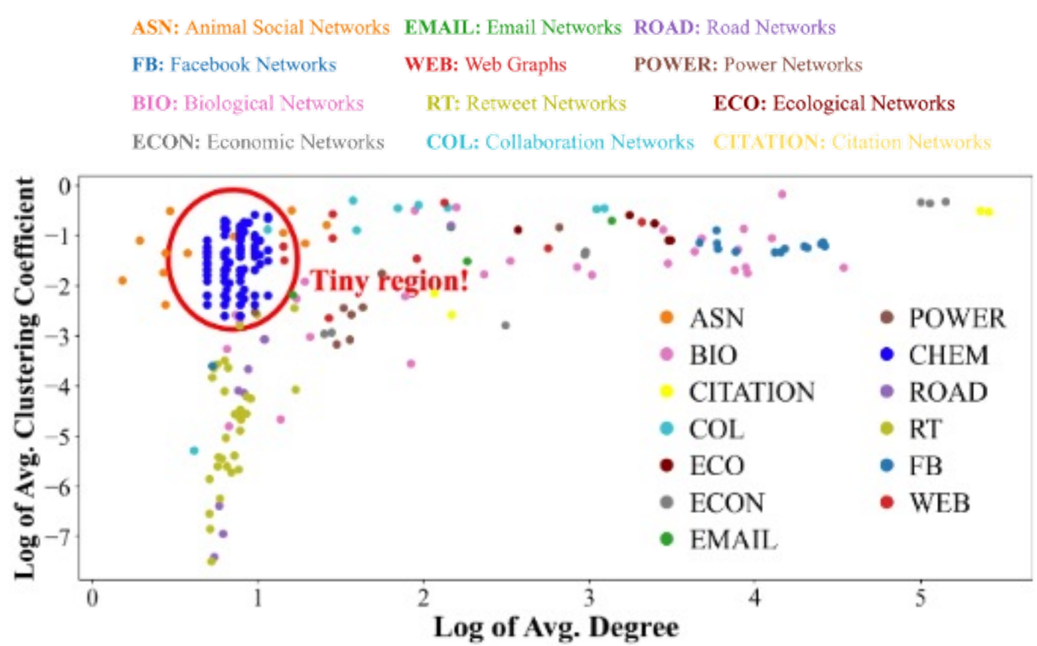
| Image | Language | Video | Audio |
|---|---|---|---|
|  |  |  |  |



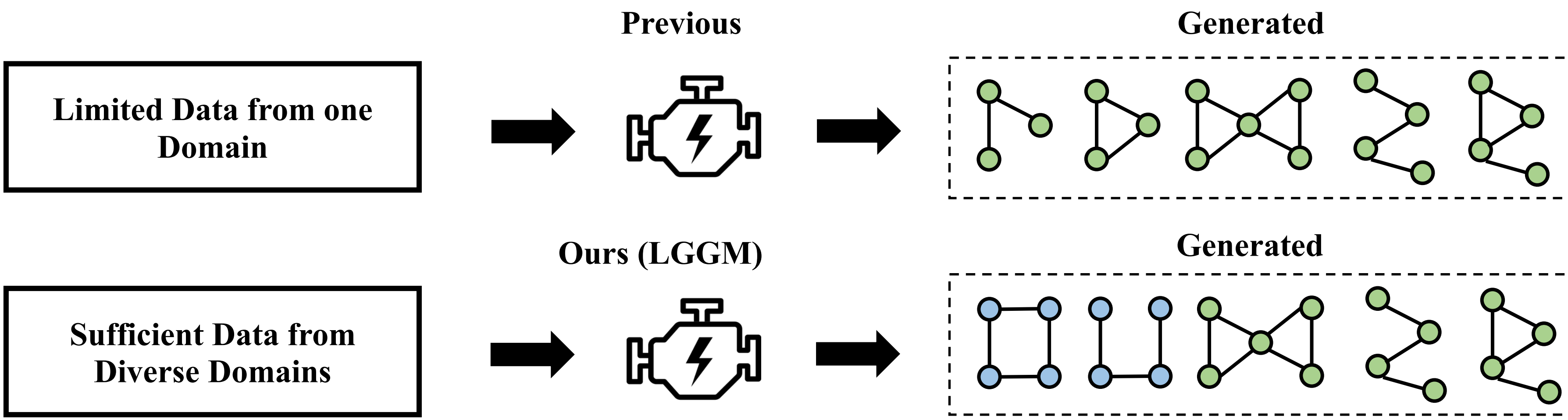

**Previous Small Model (Limited Data)** → **Recent Large Model (Sufficient well-curated Data)**

## How about graph generative models?

| Type | Model | # Domains | Multi-Domain Training |
|---|---|---|---|
| Auto-Regressive | GraphRNN [69] | 2 | ✘ |
| | EdgeRNN [3] | 3 | ✘ |
| | MolRNN [46] | 2 | ✘ |
| VAE | MDVAE [13] | 1 | ✘ |
| | PCVAE [14, 20] | 3 | ✘ |
| | (DE)CO-VAE [19] | 1 | ✘ |
| | GraphVAE [57] | 1 | ✘ |
| GAN | Mol-CycleGAN [40] | 1 | ✘ |
| | LGGAN [16] | 2 | ✘ |
| Flow | GraphNVP [38] | 1 | ✘ |
| | MoFlow [70] | 1 | ✘ |
| | GraphDF [36] | 2 | ✘ |
| Diffusion | GDSS [24] | 3 | ✘ |
| | DiGress [64] | 2 | ✘ |
| | GraphEBM [33] | 1 | ✘ |



# Our Large-scale Training Paradigm

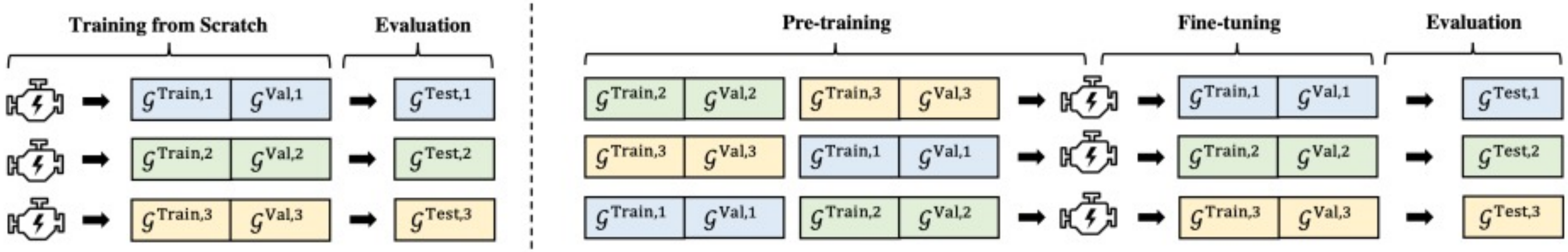



| | | | | | |
|---|---|---|---|---|---|
| ANIMAL SOCIAL NETWORKS | 816 | INTERACTION NETWORKS | 29 | SCIENTIFIC COMPUTING | 11 |
| BIOLOGICAL NETWORKS | 37 | INFRASTRUCTURE NETWORKS | 8 | SOCIAL NETWORKS | 77 |
| BRAIN NETWORKS | 116 | LABELED NETWORKS | 105 | FACEBOOK NETWORKS | 114 |
| COLLABORATION NETWORKS | 20 | MASSIVE NETWORK DATA | 21 | TECHNOLOGICAL NETWORKS | 12 |
| CHEMINFORMATICS | 646 | MISCELLANEOUS NETWORKS | 2669 | WEB GRAPHS | 36 |
| CITATION NETWORKS | 4 | POWER NETWORKS | 8 | DYNAMIC NETWORKS | 115 |
| ECOLOGY NETWORKS | 6 | PROXIMITY NETWORKS | 13 | TEMPORAL REACHABILITY | 38 |
| ECONOMIC NETWORKS | 16 | GENERATED GRAPHS | 221 | BHOSLIB | 36 |
| EMAIL NETWORKS | 6 | RECOMMENDATION NETWORKS | 36 | DIMACS | 78 |
| GRAPH 500 | 8 | ROAD NETWORKS | 15 | DIMACS10 | 84 |
| HETEROGENEOUS NETWORKS | 15 | RETWEET NETWORKS | 34 | NON-RELATIONAL ML DATA | 211 |

**Degree (DEG), Clustering Coefficient (CC)**

$$\mathrm{MMD}(\mathcal{G}_g, \mathcal{G}_r) = \frac{1}{m^2}\sum_{i,j=1}^{m} k(\mathbf{x}_i^r, \mathbf{x}_j^r) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(\mathbf{x}_i^g, \mathbf{x}_j^g) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(\mathbf{x}_i^g, \mathbf{x}_j^r) \qquad k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-d(\mathbf{x}_i, \mathbf{x}_j)/2\sigma^2)$$

# Experiment – Large-scale Training is Beneficial

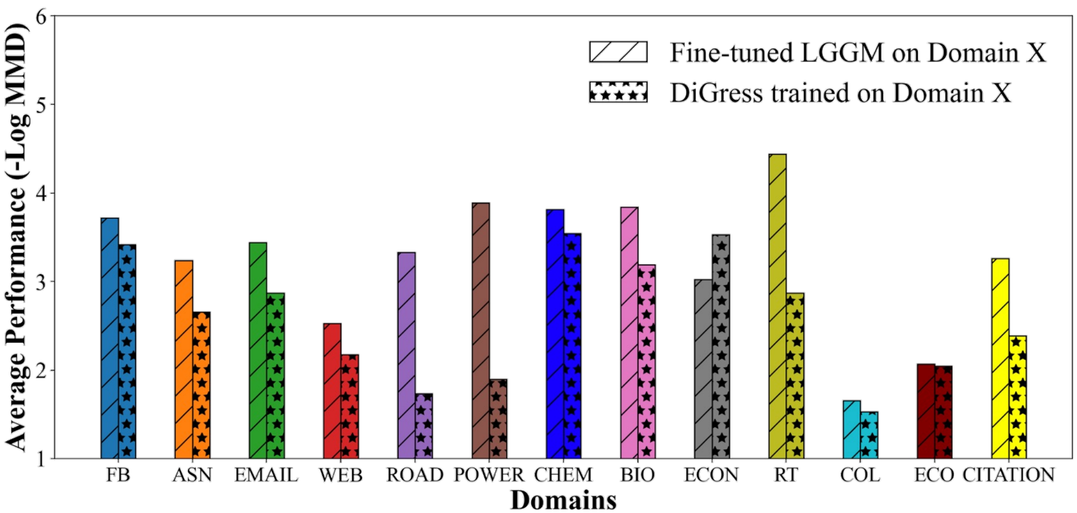


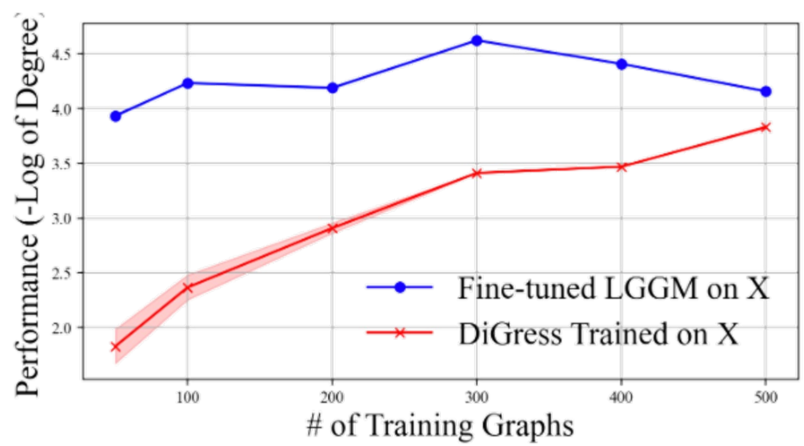

Digress trained from scratch on X

LGGM-X fine-tuned on X

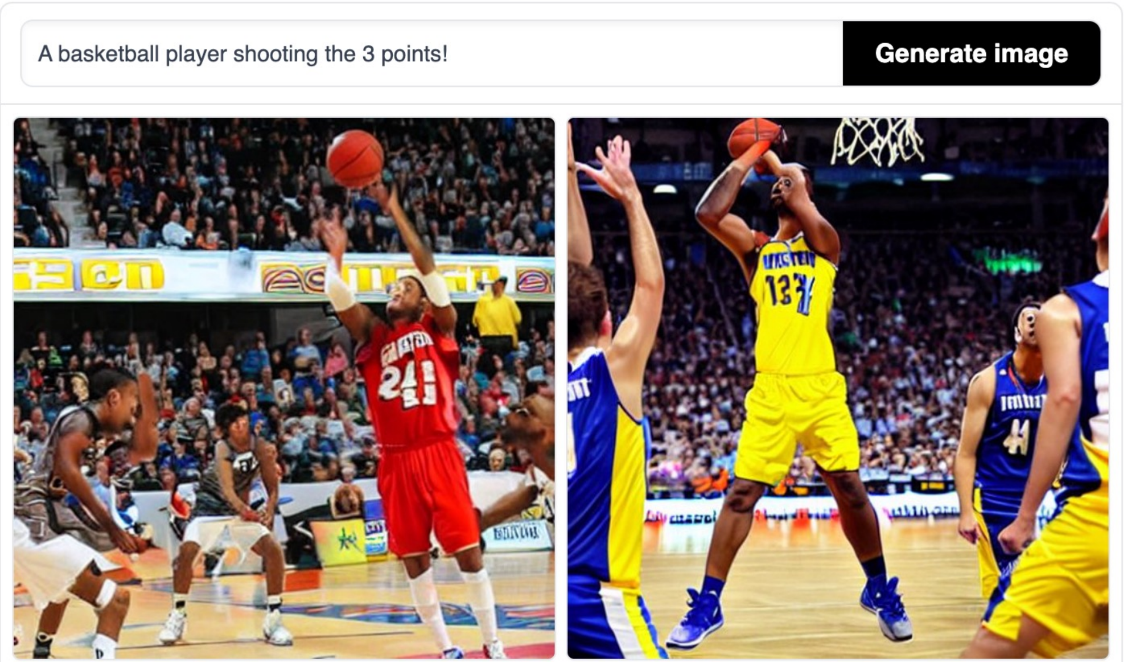
$$\Theta^{\star} = \underset{\Theta}{\operatorname{argmin}}\, \mathcal{L} = \mathbb{E}_{G\sim P(\mathbb{G})}\mathbb{E}_{t\sim\mathcal{T}}\mathbb{E}_{G^t\sim q(\mathbb{G}^t|\mathbb{G})}(-\log p_{\Theta}(G|G^t))$$

$$\Theta^{\star\star} = \underset{\Theta}{\operatorname{argmin}}\, \mathcal{L} = \mathbb{E}_{\widetilde{G}\sim P(\widetilde{\mathbb{G}})}\mathbb{E}_{t\sim\mathcal{T}}\mathbb{E}_{\widetilde{G}^t\sim q(\widetilde{\mathbb{G}}^t|\widetilde{\mathbb{G}})}(-\log p_{\Theta}(\widetilde{G}|\widetilde{G}^t))$$



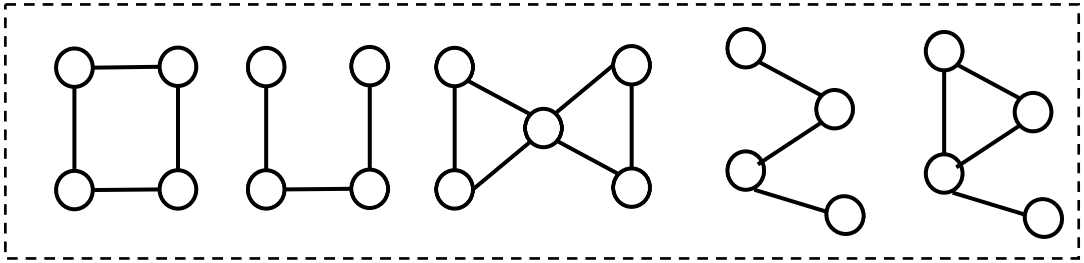



# Experiment – Text2Graph Generation

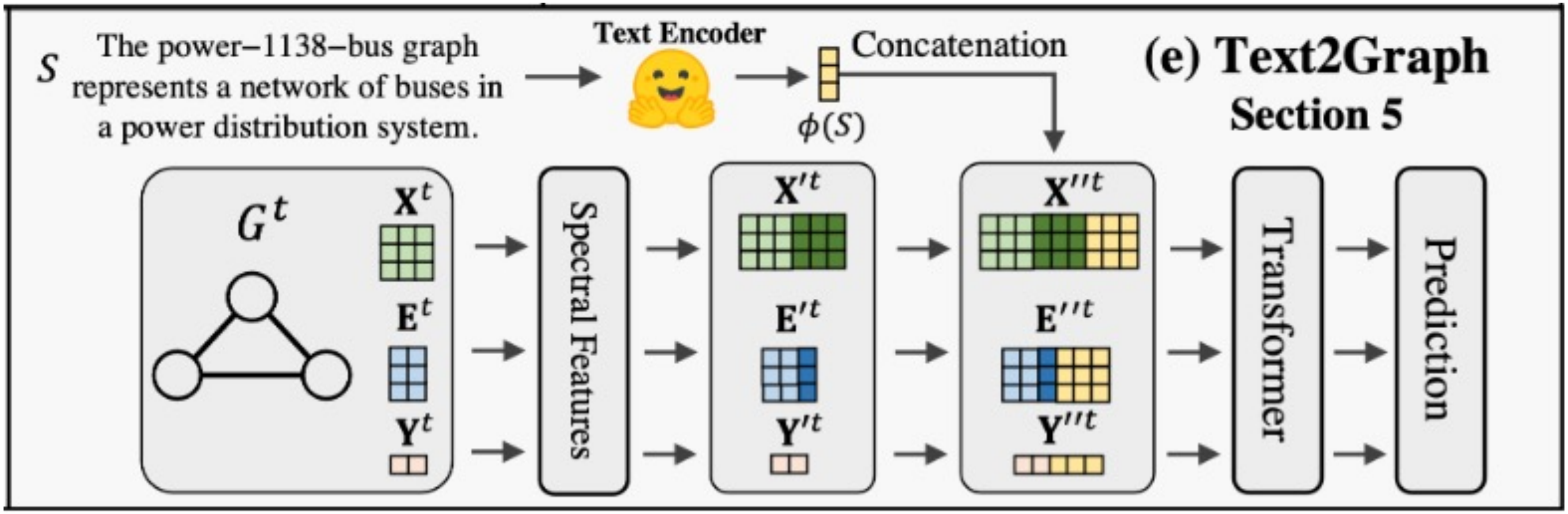



Generated Images are of Low-quality due to Online Free Version

**Images have their semantic-meaning for us to specify**



**Graphs have their own meta-data: Degree/Density/Clustering Coefficient/Domain…**

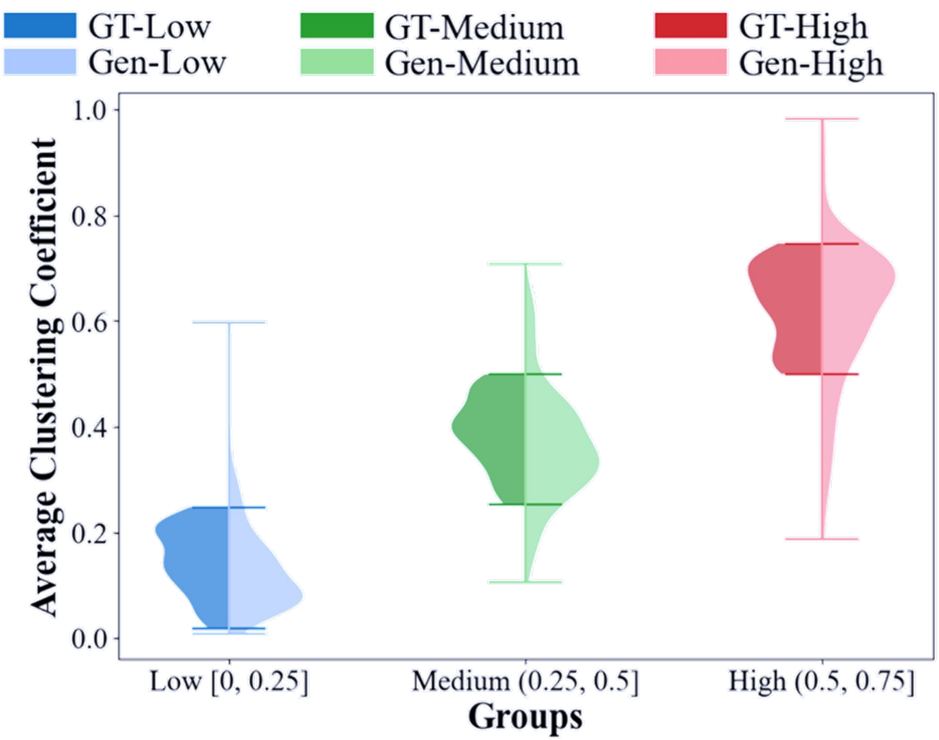


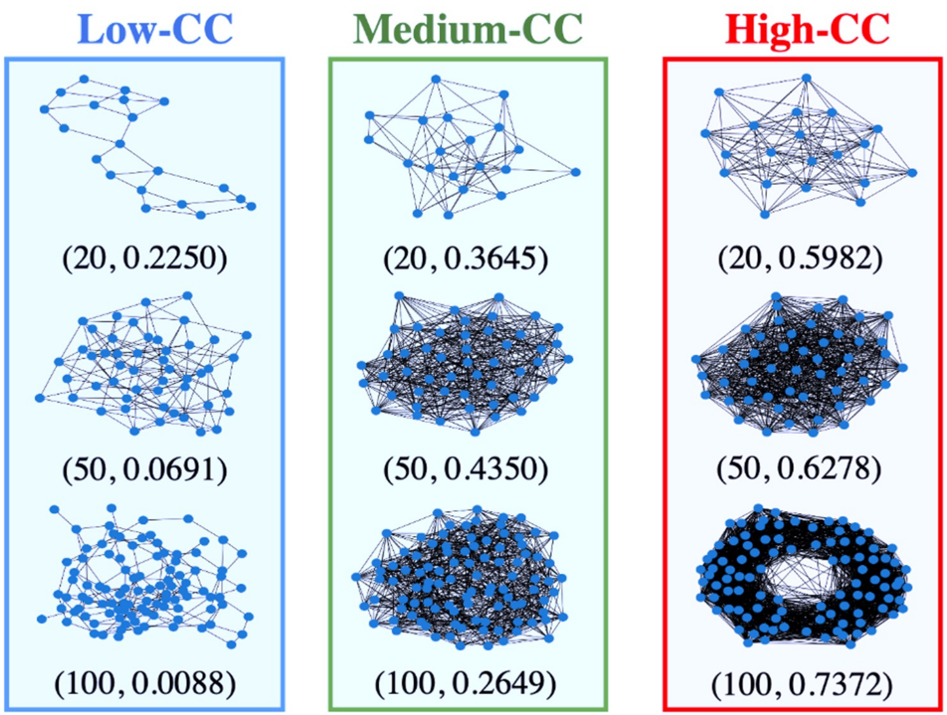

**Pretrained encoder to obtain text embeddings**

**Fuse the text embeddings into the latent diffusion**

# Experiment – Text2Graph Generation

## Control the Clustering Coefficient (somewhat # of triangles) in the graph



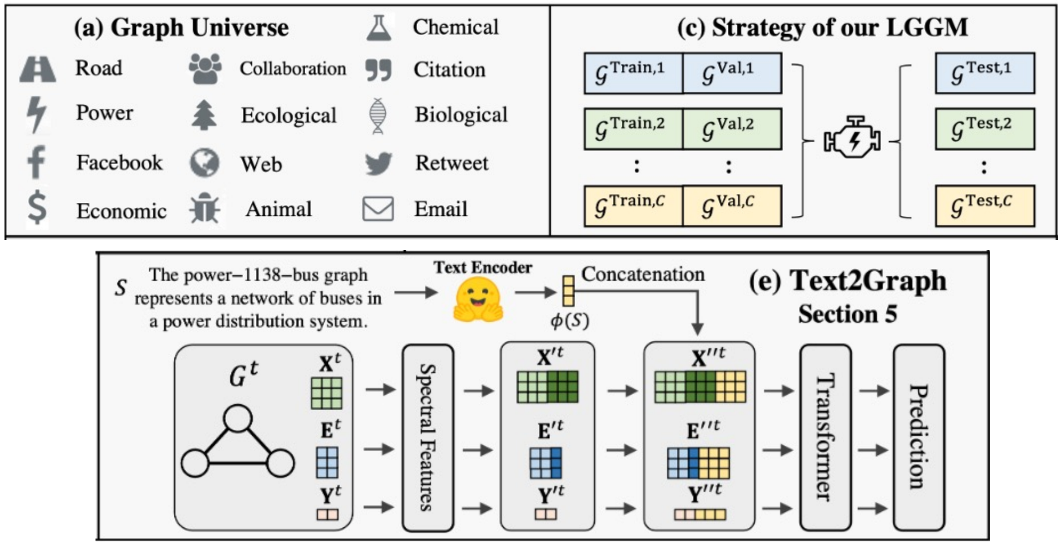



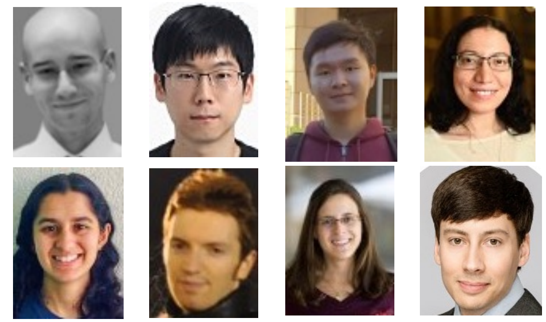

**GT – Ground-truth ones**

**Gen – Generated ones**

# Summary of LGGM

## Method - LGGM



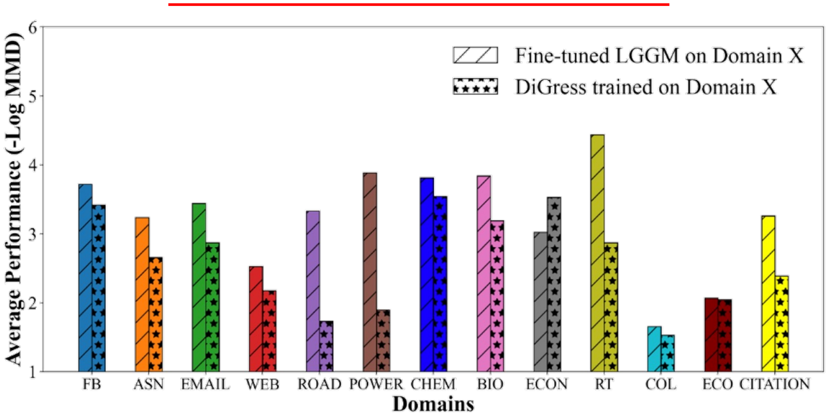
Adobe

NSF

Knowledge Intelligence for Discovery and Decision-making (KIND) Lab

https://kindlab-fly.github.io/

## Better Generative Performance




## Text-to-Graph Generation Control