SEDGG: Score Entropy Diffusion for Scalable and Controlled Graph Generation

Yangtian Zhang
Department of Computer Science
Yale University
yangtian.zhang@yale.edu

Abstract

Graph generation has become a pivotal task in a variety of domains, from molecular design to social network analysis, where generating high-quality and diverse graphs is crucial. Existing work either adopt autoregressive models, which neglect the inherent equivariance of graph-structured data, or utilize mean-parameterized diffusion models, which are computationally prohibitive for large graphs and suffer from high variance training objectives. In this work, we introduce SEDGG (Score Entropy Diffusion for Scalable and Controlled Graph Generation), a novel framework that leverages the principles of score entropy from diffusion models to efficiently generate graphs. Extending the Score Entropy Discrete Diffusion (SEDD) paradigm, SEDGG provides a scalable method for graph generation, balancing the trade-off between computational efficiency and output quality. Our model allows for fine-grained control over graph properties through its direct parameterization of probability ratios, enabling flexible sampling strategies that improve upon traditional graph generation methods. Future potential experimental results will demonstrate that SEDGG surpasses existing models in both scalability and control, generating high-quality graphs with reduced computational overhead. This approach opens new avenues for graph generation in various applications, offering enhanced flexibility and performance in a unified diffusion-based framework.

1 Introduction

Graph generation is a critical task in various fields such as molecular design, knowledge graph construction, and network analysis. Given a target graph structure $G = (V, E)$, where V represents the set of vertices and E the set of edges, the objective is to generate a new graph $\hat{G} = (\hat{V}, \hat{E})$ that adheres to specific distributional properties or constraints, such as node degrees, connectivity, or community structure.

Recent advancements in graph generation have employed diffusion models for discrete data [Vignac et al., 2022, Qin et al., 2023, Madeira et al., 2024], inspired by their success in continuous domains. The general approach in these models is to define a forward noising process, where a structured object like a graph is progressively corrupted by adding random noise at each time step t. A reverse process is then trained to recover the original data by denoising the corrupted samples.

Formally, the forward diffusion process is defined as a Markov chain:

$$p\left(x_t \mid x_{t-1}\right) = \mathcal{N}\left(x_t; \mu\left(x_{t-1}\right), \Sigma(t)\right)$$

where x_0 represents the original data (graph), x_t is the corrupted version at time step t, and $\Sigma(t)$ denotes the time-dependent noise variance. In previous works like DiGress [Vignac et al., 2022],

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

the reverse process is parameterized by mean prediction, focusing on estimating $p\left(x_0 \mid x_t\right)$, the probability of the original graph conditioned on the noisy sample at time t.

While effective in many cases, mean prediction-based diffusion models suffer from several limitations:

- **Inflexible Sampling:** Mean prediction-based models are inherently inflexible in sampling strategies, limiting the ability to explore diverse graph structures.
- **Variance in Training Objectives:** The variance in the training objectives can make the training process unstable and less sample efficient.
- **Limited Control:** The model's ability to control the generated graph properties is limited by the model's capacity to predict the mean accurately.

To address these challenges, we introduce SEDGG (Score Entropy Diffusion for Scalable and Controlled Graph Generation), a novel framework that leverages the principles of score entropy from diffusion models to efficiently generate graphs. Extending the Score Entropy Discrete Diffusion (SEDD) paradigm, SEDGG provides a scalable method for graph generation, balancing the trade-off between computational efficiency and output quality. Our model allows for fine-grained control over graph properties through its direct parameterization of probability ratios, enabling flexible sampling strategies that improve upon traditional graph generation methods.

By integrating this new approach, we aim to overcome the existing limitations and set a new standard for graph generation in various applications, offering enhanced flexibility and performance within a unified diffusion-based framework.

2 Related Works

Diffusion Models for Graph Generation: Denoising diffusion probabilistic models (DDPMs) were first introduced by Sohl-Dickstein et al. [2015] for continuous data, and later refined by Ho et al. [2020] to improve sampling efficiency and generation quality. Diffusion models have shown exceptional performance in domains like image and video generation [Dhariwal and Nichol, 2021], raising interest in adapting them to discrete data, including graphs.

Recent works have extended diffusion models to discrete domains. DiGress [Vignac et al., 2022] presents a discrete diffusion model specifically designed for generating graphs. DiGress operates by progressively adding noise through graph-editing operations (e.g., edge addition, removal, or node categorization changes) and training a graph transformer to reverse this process. The model achieved state-of-the-art performance on molecular and non-molecular datasets, including the Stochastic Block Model (SBM) and planar graph datasets, highlighting the power of diffusion in handling complex graph structures.

However, DiGress and similar approaches typically rely on mean prediction-based diffusion where the reverse process estimates $p(x_0 \mid x_t)$, focusing on generating the mean of the original distribution. This can lead to constrained or rigid sampling, limiting the flexibility needed to adapt to diverse graph structures and specific properties during generation.

Score-Based Generative Models Score-based generative models, as introduced by Song and Ermon [2019], propose an alternative approach by estimating the gradient of the data distribution, or the score function, instead of directly modeling the distribution itself. This approach has shown promise in continuous domains, particularly for high-dimensional data, and provides an efficient sampling process. Lou et al. has further extended this idea to discrete data generation. In graph generation, the ability to model data via score functions could enable more flexible sampling mechanisms compared to traditional mean prediction models.

Building on these ideas, SEDGG extends score-based models by introducing score entropy diffusion for graph generation, parameterizing the reverse process through probability ratios rather than mean predictions. This enables flexible sampling, offering greater control over the generated graph's properties, including topological features and node or edge attributes.

3 Method

In this section, we propose the Score Entropy Diffusion for Graph Generation (SEDGG), adapting the discrete diffusion framework from the Score Entropy Discrete Diffusion (SEDD) to the domain of graph-structured data $G = (V, E)$. First, we introduce the discrete diffusion process and then describe how we apply this approach to graph data.

3.1 Discrete Diffusion Process with Score Entropy

We model probability distributions over a discrete finite support $\mathcal{X} = \{1, \ldots, N\}$, where our distributions are represented by probability mass vectors $p \in \mathbb{R}^N$ that are positive and sum to 1 . The discrete diffusion process evolves a family of distributions $p_t \in \mathbb{R}^N$ according to a continuous-time Markov process, described by the following linear ordinary differential equation:

$$\frac{dp_t}{dt} = Q_t p_t \quad p_0 \approx p_{\text{data}}$$

Here, $Q_t \in \mathbb{R}^{N \times N}$ represents the diffusion matrices. These matrices have nonnegative nondiagonal entries, and their columns sum to zero, ensuring that the total mass of the distribution remains constant over time. The matrix Q_t is typically designed such that p_t approaches a limiting distribution p_{base} as $t \to \infty$, which acts as a base distribution for the diffusion process.

The reverse process, which recovers the original distribution from the noisy data, is given by another diffusion matrix $\bar{Q}_t$. This reverse process evolves the noisy data back towards the original data distribution using the following differential equation:

$$\frac{dp_{T-t}}{dt} = \bar{Q}_{T-t} p_{T-t}$$

The reverse diffusion matrix $\bar{Q}_t$ is defined in terms of the forward diffusion matrix Q_t and the ratio of probabilities between the two time steps:

$$\bar{Q}_t(y, x) = \frac{p_t(y)}{p_t(x)} Q_t(x, y)$$

The matrix $\bar{Q}_t$ allows us to reverse the diffusion process, gradually denoising the data back to its original form. The term $\frac{p_t(y)}{p_t(x)}$, known as the concrete score, plays a role analogous to the typical score function $\nabla_x \log p_t$ in continuous diffusion processes.

To train the SEDD model, we minimize the denoising score entropy loss as described in Lou et al..

$$\underset{\substack{x_0 \\ x \sim p(\cdot|x_0)}}{\mathbb{E}} \left[\sum_{y \neq x}^{y \neq x} \left[w_{xy} \left(s_\theta(x)_y - \frac{p\,(y \mid x_0)}{p\,(x \mid x_0)} \log s_\theta(x)_y \right) \right] \right]$$

3.2 Discrete Diffusion to Graph-Structured Data

Since graph is also a kind of discrete data, we can use SEDD to generate graph by defining diffusion process directly on the node and edge.

- **Node Diffusion:** The node set V is corrupted by applying the discrete diffusion process to node attributes or by modifying the connectivity of the nodes.
- **Edge Diffusion:** Similarly, the edge set E is corrupted by applying the discrete diffusion process to the edge features or by modifying the connectivity of the edges.

And by reversing the process, we can generate graph with desired properties.

4 Experiment

4.1 Dataset

To evaluate the performance of the proposed SEDGG model, we can leverage a variety of well-established graph generation datasets, similar to those used in prior works like DiGress. These datasets include both molecular and non-molecular benchmarks, offering a broad spectrum of graph generation tasks that challenge the scalability, flexibility, and accuracy of the model. Below are the key datasets that can be used for comprehensive evaluation:

Stochastic Block Model (SBM) Dataset: This dataset, introduced by Martinkus et al. [2022], consists of synthetic graphs generated using a stochastic block model, which is commonly used to model community structures in graphs. It contains graphs with varying node and edge properties, making it suitable for assessing the model's ability to handle diverse topologies and sizes. Each graph contains up to 200 nodes, and the dataset includes different metrics for evaluating the validity, uniqueness, and novelty (V.U.N.) of the generated graphs.

Planar Graph Dataset: The planar graph dataset, described by Martinkus et al. [2022], consists of planar graphs, which are graphs that can be embedded in a plane without any edges crossing. The task involves generating graphs with up to 64 nodes, testing the model's ability to maintain geometric constraints such as planarity and connectivity. The dataset is useful for evaluating the structural integrity of generated graphs, particularly in terms of ensuring that generated samples conform to graph-theoretic properties like planarity and orbit counts.

QM9 Molecular Dataset: The QM9 dataset, introduced by Wu et al. [2018], is a widely-used molecular dataset containing 134,000 small organic molecules with up to 9 heavy atoms. It includes molecular structures along with several associated properties. QM9 serves as an excellent benchmark for graph generation models in molecular chemistry, assessing the models' ability to generate chemically valid molecules while adhering to structural and property-based constraints.

4.2 Evaluation Metrics and Expected Results

To evaluate the performance of SEDGG, we will use a combination of quantitative and qualitative metrics. For quantitative evaluation, we will consider the following metrics for different graph generation tasks:

- **V.U.N. Score:** Measures the validity, uniqueness, and novelty of generated graphs.
- **Planarity Score:** Assesses the proportion of generated planar graphs in the dataset.
- **Chemical Validity:** Evaluates the chemical validity of generated molecules using established benchmarks like Fréchet ChemNet Distance (FCD) and Scaffold Similarity (SNN).
- **Graph Similarity:** Compares the structural similarity between generated and real graphs using metrics such as Jaccard Similarity and Graph Edit Distance.

4.3 Training Details

For the discrete diffusion process, we employ an absorbing transition matrix, ensuring stability in the noising process. We adopt a log-linear diffusion noise schedule as described in [Lou et al.], with noise parameters $\sigma_{\min} = 10^{-4}$ and $\sigma_{\max} = 20$. The diffusion process is configured to run for 500 steps by default, with an ablation study conducted to evaluate the impact of varying this parameter on model performance.

The backbone model is a GraphTransformer [Yun et al., 2019] with 5 layers. The hidden dimension of the node embeddings is set to 256, while the hidden dimension of the edge embeddings is set to 128, balancing model expressiveness and computational efficiency.

We optimize the model using the AdamW optimizer, with a weight decay of 10^{-12} to prevent overfitting. The learning rate is set to 2×10^{-4}. To ensure efficient training on large datasets, we use a batch size of 512.

4.4 Synthetic Graph Generation

We evaluate the performance of synthetic graph generation on the SBM dataset. Results for DiGress and our method are averaged over three independent runs, while results for other models are sourced from Martinkus et al. [2022]. As shown in Table 1, our model demonstrates superior novelty and achieves better in orbit metrics. However, it performs slightly worse than DiGress on the Degree and Clustering metrics, suggesting that the explicit x_0 parameterization used by DiGress may better capture these specific graph properties.

Table 1: Comparison of synthetic graph generation models on key metrics. Lower values for Degree (Deg), Clustering (Clus), and Orbit (Orb) indicate better alignment with ground truth. Higher values for Validity, Uniqueness, and Novelty (V.U.N.) are desirable.

Model	**Deg** ↓	**Clus** ↓	**Orb** ↓	**V.U.N.** ↑
GraphRNN	6.9	1.7	3.1	5%
GRAN	14.1	1.7	2.1	25%
GG-GAN	4.4	2.1	2.3	25%
SPECTRE	1.9	1.6	1.6	53%
ConGress	34.1	3.1	4.5	0%
DiGress	**1.6** ± 0.1	**1.5** ± 0.1	1.8 ± 0.2	76%
SEDGG (Ours)	2.1 ± 0.1	1.6 ± 0.1	**1.6** ± 0.1	**81**%

4.5 Molecule Graph Generation

We evaluate molecular graph generation performance using key metrics, including Validity, Uniqueness, Atom Stability, and Molecule Stability, as shown in Table 2. Our method, SEDGG, achieves competitive results, particularly excelling in Uniqueness with a score of 98.9%, surpassing both ConGress and DiGress. Additionally, SEDGG performs on par with DiGress in Atom Stability, achieving 98.0%, and outperforms ConGress across all metrics.

However, SEDGG falls slightly short of DiGress in Validity (93.3% vs. 95.4%) and Molecule Stability (73.4% vs. 79.8%), suggesting that DiGress's explicit x_0 parameterization may provide an advantage in preserving chemical validity and overall molecular integrity. Despite this, SEDGG demonstrates comparable performance.

Table 2: Comparison of models on molecular graph generation metrics. Higher values indicate better performance.

Model	**Valid** ↑	**Unique** ↑	**Atom Stable** ↑	**Mol Stable** ↑
Dataset (Ground Truth)	97.8	100.0	98.5	87.0
ConGress	86.7 ± 1.8	98.4 ± 0.1	97.2 ± 0.2	69.5 ± 1.6
DiGress	**95.4** ± 1.1	97.6 ± 0.4	**98.1** ± 0.3	**79.8** ± 5.6
SEDGG (Ours)	93.3 ± 2.5	**98.9** ± 0.3	98.0 ± 0.2	73.4 ± 3.9

4.6 Ablation Study

To assess the impact of the number of diffusion steps on SEDGG's performance, we conducted an ablation study on the QM9 dataset, varying the steps from 100 to 500. As shown in Figure 1, reducing the steps leads to a gradual decline in performance across all metrics. Validity decreases from 93.3% to 82.7%, while Molecule Stability shows the sharpest drop, from 73.4% to 55.9%. This highlights the importance of sufficient diffusion steps for capturing structural and property-based constraints. Based on these findings, we use 500 steps as the default, balancing performance and computational efficiency.

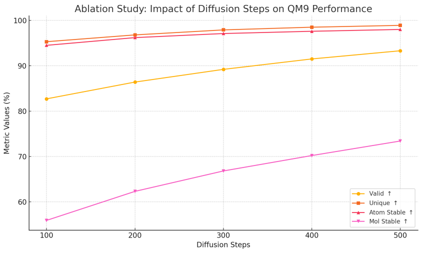

Figure 1: Ablation study: Impact of diffusion steps on QM9 performance.

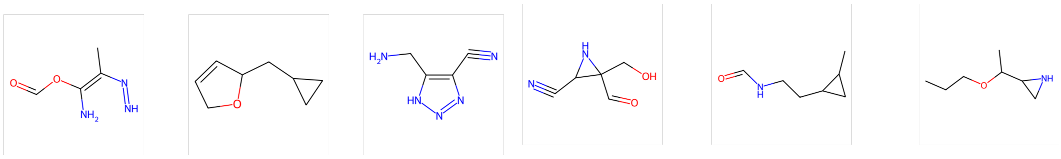

Figure 2: Sampling results with 500 diffusion steps.

5 Conclusion

In this paper, we introduced SEDGG, a novel framework for scalable and controlled graph generation based on score entropy diffusion. By leveraging direct probability ratio parameterization, SEDGG achieves competitive performance across synthetic and molecular graph generation tasks, offering enhanced flexibility and control. Our experimental results demonstrate that SEDGG excels in generating unique and high-quality graphs while balancing computational efficiency.

Despite its strengths, SEDGG shows room for improvement in capturing certain structural properties, as observed in specific metrics like Degree and Molecule Stability. Future work could explore adaptive diffusion schedules and enhanced parameterizations to further refine its performance. Overall, SEDGG sets a strong foundation for advancing diffusion-based approaches in graph generation, paving the way for broader applications in areas such as molecular design, network analysis, and beyond.

Reproducibility Statement

The repository is available at: https://github.com/zytzrh/SEDGG.

References

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Forty-first International Conference on Machine Learning*.

Manuel Madeira, Clement Vignac, Dorina Thanou, and Pascal Frossard. Generative modelling of structurally constrained graphs. *arXiv preprint arXiv:2406.17341*, 2024.

Karolis Martinkus, Andreas Loukas, Nathanaël Perraudin, and Roger Wattenhofer. Spectre: Spectral conditioning helps to overcome the expressivity limits of one-shot graph generators. In *International Conference on Machine Learning*, pages 15159–15179. PMLR, 2022.

Yiming Qin, Clement Vignac, and Pascal Frossard. Sparse training of discrete diffusion models for graph generation. *arXiv preprint arXiv:2311.02142*, 2023.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation. *arXiv preprint arXiv:2209.14734*, 2022.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. Graph transformer networks. *Advances in neural information processing systems*, 32, 2019.