CS60078 Complex Network Spring 2025-26 Projects

Instructors: Niloy Ganguly, Somak Aditya

TAs: Sachin Vashistha, Mainak Chaudhury

There are ten projects:

- Project 1: Graph Meets ASR
- Project 2: LLM Meets Diffusion
- Project 3: Graph Anomalies
- Project 4: Multivariate Time Series Analysis
- Project 5: Seizure Detection
- Project 6: Drug Target Binding Affinity
- Project 7: Learning Persistent Community Structures in Dynamic Networks via Topological Data Analysis
- Project 8: Neural Common Neighbor with Completion for Link Prediction

The following table shows the assignment of the aforementioned projects to the students.

Project No.	Assigned students (Group)	Concerned TA
6	21CS30030, 21CS30054, 21ME3FP55	Mainak
4	22CS10012, 22CS10072, 22CS30016	
2	22CS30038, 22CS30045, 22CS30046	
5	25CS60R22, 25CS60R47, 25CS60R60, 25CS60R79	
1	25BM6JP24, 25BM6JP41, 25BM6JP50	Sachin
7	25BM6JP57, 25CS60R01, 25CS60R02, 25CS60R77	
8	25BM6JP06, 25BM6JP17, 25BM6JP21	
3	22CS30072, 22IM30034, 23EX10026	

Midterm Task (Total marks: 100):

1. Task T1: Read the research paper associated with the assigned project, and prepare a report summarizing it. The report should contain the following: (Marks: 20)
   - Problem statement that the authors worked on.
   - Brief description of the previous solutions and their limitations.

- The novel solution that the authors provided, and how they cover the limitations of the previous approaches/ solutions.
- Brief description of the datasets with an example.
- Summary of the results and conclusion.

2. Task T2: Download and setup the Github repo of the assigned project. (Marks: 20)
3. Task T3: Reproduce the paper results (should include the novel solution proposed by authors). (Marks: 30)
4. Task T4: Carry out qualitative and quantitative analysis of the experiments conducted. (Marks: 30)

Midterm Deliverables:

1. Prepare a three page report in the PDF format for Task T1.
2. For Task T2, host the setup code on your github repo, make it public, and share with the assigned TAs.
3. For Task T3, and T4, prepare a report in PDF format that contains all the results (tables, graphs, etc.), and analysis (qualitative as well as quantitative).
4. NOTE: For each project, students have to submit a .zip file in the following format: **`<Group_number>_<Project_number>_MidTerm.zip`**
   Example: Prepare the .zip file "1_1_MidTerm.zip", if your group no is 1, and project no is 1.
   This zip file will contain all reports, codes, etc.

Endterm Task (Total marks: 100):

1. Task 1: Based on the qualitative and quantitative analysis carried out in the "Midterm task", prepare a report containing the issues and limitations of the "novel approach" introduced by the authors of the paper. (Marks: 30)
2. Task 2: This task involves students proposing their approach to improve upon the current approach of the authors. This can be novel - i.e., you can propose your own novelty on top of existing algorithms. OR, it can be a re-implementation of a new technique. (Marks: 70)

Endterm Deliverables:

1. Prepare a two page report in the PDF format for Task T1.
2. Prepare a detailed report (no page limit) in the PDF format for Task T2. Students should mathematically introduce your proposed approach. Hence, it would be good to prepare the report in latex format. Also, host the setup code on your github repo, make it public, and share with the assigned TAs.
3. NOTE: For each project, students have to submit a .zip file in the following format: **`<Group_number>_<Project_number>_EndTerm.zip`**
   Example: Prepare the .zip file "1_1_EndTerm.zip", if your group no is 1, and project no is 1.
   This zip file will contain all reports, codes, etc.
4. *NOTE: For each project, you may receive some specific feedback or suggestions to do specific experiments during midterm. That will be included for endterm and marks will be adjusted accordingly.*

General Instructions

1. All groups are final. For project assignment, if anyone has difficulty with the topic, first discuss with the concerned TA and clarify. Please do it within the next two days (**within Jan 24th**). After that, no change will be accepted.
2. **A Google Form will be circulated to collect the submissions. Deadline Feb 12 11:59 PM and can not be extended.**
3. **A Pre-mid term evaluation session will be conducted on Feb 13, 11AM - 1PM**.
4. Your codes should print out exactly what is asked, and in the specified format.
5. In the code hosted on the github, along with the code, also submit an additional text file called “instructions.md” where you should state how to run your codes as well as any additional information you want to convey, such as the libraries of Python used.
6. *You are requested not to take help of any AI tools for preparing your report. If your report matches with AI tool generated outputs (which we will perform at our end), you may lose all marks for the entire Project.*

Project 1: Graphs meet ASR

This project studies Automatic Speech Recognition (ASR) performance through the lens of Complex Networks by modeling speakers as nodes in a graph. The central idea is to construct speaker similarity graphs where edges encode acoustic or error-based similarity between speakers, and then analyze how graph structure relates to ASR error patterns.

Using standard ASR systems and public datasets, we compute speaker-level error statistics such as Word Error Rate (WER), substitutions, insertions, and deletions. Speakers are connected based on similarity of learned speech embeddings (e.g., x-vectors, Wav2Vec embeddings) or similarity in ASR error profiles.

Reference: https://ieeexplore.ieee.org/document/10448308, https://arxiv.org/pdf/2305.18824
GitHub Repo : https://github.com/BriansIDP/espnet/tree/GNN

Project 2: LLM meets Diffusion for Crystal Material Generation

Recent advances in generative modeling have demonstrated substantial potential for the discovery and design of novel periodic crystal structures. Existing approaches typically rely on either large language models (LLMs) or equivariant denoising-based generative models, each offering complementary strengths and limitations. LLMs are particularly effective at modeling discrete information, such as atomic species and stoichiometry, but often struggle with continuous variables, including atomic coordinates and lattice parameters. In contrast, equivariant denoising models excel at capturing continuous geometric features but face challenges in accurately generating discrete atomic compositions.

To leverage the strengths of both paradigms, recent works have proposed hybrid approaches in which LLMs—fine-tuned on materials science datasets—jointly predict discrete (atomic types) and continuous (atomic positions and lattice parameters) information. In these frameworks, discrete predictions are typically more accurate, while the continuous outputs are comparatively noisy.

Building on this insight, we treat the continuous predictions as an intermediate noisy representation corresponding to a specific diffusion time step τ\tauτ. Conditioned on the highly reliable discrete information produced by the LLM, a diffusion model is then employed to denoise this intermediate representation from time step τ\tauτ to the final clean structure, yielding physically plausible crystal geometries.

The objectives of this work are threefold:

1. **LLM exploration:** Investigate variants of LLM architectures that achieve superior performance on discrete atomic predictions compared to current state-of-the-art models.

2. **Diffusion modeling:** Evaluate and develop graph-based diffusion models that effectively denoise continuous structural information while respecting physical and symmetry constraints.

3. **Intermediate transition modeling:** Identify and characterize accurate intermediate diffusion states of continuous variables that serve as optimal starting points for conditional denoising

Together, these efforts aim to create a controlled and modular generative framework that combines the discrete reasoning capabilities of LLMs with the geometric precision of diffusion models for crystal structure generation.

References :
https://arxiv.org/pdf/2510.23040
https://github.com/kdmsit/crysllmgen

Project 3: Controlled and Interpretable Graph Anomaly Detection via Diffusion Models

Graph anomaly detection is critical for identifying abnormal nodes or substructures within networks and has broad applications across domains such as fraud detection, fault diagnosis in sensor networks, and cybersecurity. The objective is to detect entities whose structural patterns, attributes, or temporal behaviors deviate significantly from the norm.

Most existing unsupervised graph anomaly detection methods rely on reconstructing unlabeled data from compressed latent representations. While effective in capturing global patterns, these reconstruction-based approaches often fail to preserve anomaly-specific discriminative information, resulting in diminished detection performance—particularly for subtle or context-dependent anomalies.

Recent work addresses this limitation by leveraging diffusion models to learn a more discriminative latent space. By progressively perturbing and denoising graph representations, diffusion-based methods capture anomaly-relevant information across multiple scales, enabling more accurate and scalable detection of abnormal nodes and subgraphs.

However, the diffusion-enhanced latent space intertwines multiple sources of information, making it difficult to interpret why a particular node or subgraph is classified as anomalous. The lack of interpretability limits practical adoption, especially in high-stakes domains.

This work aims to address this challenge by disentangling latent representations during the diffusion process and introducing controllable mechanisms for anomaly detection. The proposed approach seeks to balance detection performance with interpretability, enabling not only accurate identification of anomalies in graphs but also meaningful explanations of the underlying causes.

Reference : https://arxiv.org/abs/2410.06549
GitHub Repo Link : https://github.com/fortunato-all/DiffGAD

Project 4: Multivariate Time Series Analysis using Graphs

Multivariate time series (MTS) forecasting is critical across many real-world applications, yet existing GNN-based approaches typically rely on separate spatial and temporal modules, such as graph convolutions and recurrent networks. This separation not only increases architectural complexity and handcrafted design choices, but also fails to model spatiotemporal dependencies in a unified manner. To address this, recent work reformulates MTS forecasting from a pure graph perspective by introducing the hypervariate graph, where each variable at each timestamp is treated as a node and sliding windows are modeled as space–time fully connected graphs. By operating entirely in the Fourier domain through the proposed Fourier Graph Operator, this work unifies spatial and temporal modeling within a single graph framework, achieving high efficiency, strong expressiveness, and reduced model complexity.

Despite these advantages, Fourier based Graph Models remains a deterministic forecasting model that primarily captures global spectral patterns. As a result, it is limited in handling non-stationary dynamics, regime shifts, stochastic variations, and noise, and cannot provide uncertainty-aware predictions. These limitations motivate a natural extension of Fourier based Graph Models toward diffusion-based modeling. By transforming FourierGNN into a conditional diffusion framework—where diffusion is guided by Fourier based Graph Models latent representations—forecasting can be formulated as an iterative denoising process. This enables probabilistic forecasting, improves robustness under distribution shift, and better captures irregular and uncertain temporal dynamics. Consequently, diffusion serves as a complementary and principled extension to Fourier based Graph Models, enhancing its expressiveness and uncertainty modeling while preserving its unified and efficient spatiotemporal encoding.

Reference : https://arxiv.org/pdf/2311.06190

GitHub Repo : https://github.com/aikunyi/FourierGNN

Project 5: Seizure Detection using EEG Data

Automated seizure detection and classification from electroencephalography (EEG) has the potential to significantly improve epilepsy diagnosis and treatment. Early graph-based approaches addressed key challenges by modeling the non-Euclidean spatiotemporal structure of EEGs using graph neural networks, improving rare seizure classification through self-supervised pre-training, and introducing quantitative interpretability methods for seizure localization. These methods demonstrated strong detection and classification performance on large public datasets, with notable gains for rare seizure types and improved localization of focal seizures compared to CNN baselines.

However, this line of work is now relatively dated. With recent advances in dynamic and attention-based graph neural networks, graph transformers, and generative models, there is substantial scope to further improve seizure detection accuracy, robustness to class imbalance, and clinically meaningful interpretability. Moreover, modern generative and self-supervised frameworks open new possibilities beyond classification, such as synthetic EEG generation, rare seizure augmentation, uncertainty-aware prediction, and counterfactual seizure localization, which were not explored in earlier studies.

Reference : https://arxiv.org/pdf/2104.08336

GitHub code : https://github.com/tsy935/eeg-gnn-ssl

Project 6: Graph Based Methods for predicting drug-target binding affinity

The development of new drugs is a costly and time-intensive process. Drug repurposing offers an effective alternative by leveraging existing FDA-approved drugs and their known properties to identify new therapeutic applications. Prior work employs a hybrid Graph Convolutional Network (GCN)–LSTM architecture to predict drug–protein binding affinity, where molecular graphs are used to represent drug structures and sequential models capture protein characteristics. Trained on multiple benchmark bioactivity datasets, this approach ranks approved drugs against target proteins using a combined affinity score, enabling efficient prioritization of promising candidates for further investigation.

Recent advances have surpassed GCN–LSTM-based models by addressing their limitations in capturing long-range dependencies and complex interaction patterns. Graph Attention Networks (GATs) and Graph Transformers enhance expressiveness by adaptively weighting atom–atom interactions, improving molecular representation learning. Cross-attention–based drug–target interaction models directly model interactions between drug atoms and protein residues, leading to more accurate affinity estimation. Additionally, pretrained molecular foundation models and protein language models leverage large-scale unlabeled data to provide richer representations, significantly outperforming task-specific architectures. Diffusion-based generative models and equivariant graph neural networks further advance binding prediction by incorporating uncertainty estimation and 3D structural inductive biases, establishing new state-of-the-art performance in drug–target interaction tasks.

Reference : https://epubs.siam.org/doi/10.1137/1.9781611977172.82
GitHub Repo: https://github.com/shrimonmuke0202/DeepGLSTM

Project 7: Learning Persistent Community Structures in Dynamic Networks via Topological Data Analysis

This work studies **community detection in dynamic networks**, where the network changes over time. The task is to group nodes into communities at each time step, while ensuring that meaningful communities remain **stable across time**. The input is a sequence of graph snapshots (e.g., monthly email networks or yearly collaboration graphs), and the output is a community assignment for each node at every time step.

The key challenge is that small changes in the network can cause large, unstable changes in detected communities. To solve this, the paper introduces a method that combines deep clustering with **topological data analysis (TDA)**. Instead of only looking at node-level changes, the method analyzes how **communities relate to each other** over time and penalizes unnecessary changes. This encourages communities to persist unless there is a real structural shift in the network.

Paper Link: https://arxiv.org/pdf/2401.03194
Github Link: https://github.com/kundtx/MFC-TopoReg

Project 8: Neural Common Neighbor with Completion for Link Prediction

This paper focuses on **link prediction**, where the goal is to predict whether an edge exists (or will exist) between two nodes. The input is a graph with some edges removed for training, and the output is a score for each candidate node pair indicating the likelihood of a link. A classic idea in link prediction is that two nodes are more likely to connect if they share many **common neighbors**.

The paper proposes **Neural Common Neighbor (NCN)**, which learns how to use common-neighbor information instead of relying on fixed formulas. However, real graphs are often incomplete, which weakens common-neighbor signals. To address this, the authors introduce **completion**, where the model first predicts missing edges and then recomputes common neighbors using the completed graph. This leads to more accurate and robust link predictions than both traditional heuristics and standard GNNs.

Paper Link: https://arxiv.org/pdf/2302.00890
Github Link: https://github.com/GraphPKU/NeuralCommonNeighbor