siam 2025

*Searchable Abstracts Document*

# SIAM International Conference on Data Mining (SDM25)

**May 1–3, 2025**

The Westin Alexandria Old Town, Alexandria Virginia, U.S.

*This document was current as of March 20, 2025. Abstracts appear as submitted.*

siam

3600 Market Street, 6th Floor
Philadelphia, PA 19104-2688 U.S.
Telephone: 800-447-7426 (U.S. & Canada) +1-215-382-9800 (Worldwide)
*meetings@siam.org*

**IP1**

**AI for America: NSF's Roadmap**

The term "Artificial Intelligence" (AI) has expanded to encompass machine learning, data mining, automation, and really most of computing and data science. The US has long been one of the worldwide leaders in research on AI, both in its original narrower and current more general senses, and much of the investment fueling this research has come from the US National Science Foundation (NSF). In this talk, I'll convey a bit about NSF's historical contributions and the organization's vision for future investments, focusing on the topics of Research, Translation, Infrastructure, and Workforce as well as a set of cross-cutting themes. The agency has research partnerships throughout the US Federal Government, the private sector, and internationally, so I'm hoping there will be elements of interest to a broad audience of researchers.

Michael Lederman Littman
National Science Foundation
mlittman@cs.brown.edu

**IP2**

**Thinking Outside the Ballot Box**

How should one design unprecedented democratic processes capable of handling enormous sets of alternatives like all possible policies, bills, or statements? I argue that this challenge can be addressed through a framework called generative social choice, which fuses the rigor of social choice theory with the flexibility and power of large language models. I then explore an application of generative social choice to the problem of identifying a proportionally representative slate of opinion statements. This includes a discussion of desired properties, an algorithm that provably achieves them, an implementation using GPT-4o, and insights from an end-to-end pilot. By providing guarantees, generative social choice could alleviate concerns about AI-driven democratic innovation and help unlock its potential.

Ariel Procaccia
Harvard University
arielpro@gmail.com

**IP3**

**Graph Reasoning in Large Language Models**

Large language models (LLMs) have demonstrated impressive capabilities in text generation, but their ability to reason over complex data remains an area of ongoing research. In this talk, we present three distinct approaches to improve LLM reasoning over complex structures. First, we leverage graph algorithms to analyze and understand the reasoning capabilities of transformer models. Our results establish a representational hierarchy, revealing the necessary Transformer capacity (number of layers, embedding dimension size) for solving different classes of reasoning tasks. Next, we exploit the topology of temporal reasoning to generate novel synthetic problem instances. This allows for a more robust evaluation of LLM reasoning capabilities. Finally, we introduce a method for improving in-context representations of structured data for pretrained LLMs, facilitating more effective reasoning over complex information.

Bryan Perozzi
Google Research
USA
bperozzi@cs.stonybrook.edu

**IP4**

**Interrupting Misinformation with Community Notes**

Social networks scaffold the diffusion of information on social media. Much attention has been given to the spread of true vs. false content on online social platforms, including the structural differences between their diffusion patterns. However, much less is known about how platform interventions on false content alter the engagement with and diffusion of such content. In this talk, I will survey work on the differences between true and false news diffusion, culminating in recent work on estimating the causal effects of Community Notes, the novel fact-checking feature adopted by X (formerly Twitter) to solicit and vet crowd-sourced notes for misleading content.

Johan Ugander
jugander@stanford.edu
Stanford University

**CP1**

**Acceleration in Low-Rank Tensor Completion**

This work studies the low-rank tensor completion problem based on partially observed entries. Inspired by the success of matrix completion based on low-rank property and its tightest convex relaxation – nuclear norm, we borrow the idea of low-rank scheme for tensor recovery. However, different from the matrix case, singular values of tensors are not straightforward. As a contribution, we reformulate tensor's nuclear norm into an equivalent form based on the unfoldings along each mode. We show that the new objective has nice properties by which we can make use of the alternating minimization method with Nesterov accelerated gradient descent. We prove the proposed algorithm has $O(\frac{1}{k^2})$ convergence rate. Numerical experiments demonstrate the superior performance of our proposed algorithm over its counterparts.

Yifan Kang
Clemson University
yifank@clemson.edu

Mengyuan Zhang
Computer Science Division
Clemson University
mengyuz@clemson.edu

Kai Liu
Clemson University
kail@clemson.edu

**CP1**

**Task Aware Modulation Using Representation Learning: An Approach for Few Shot Learning in Environmental Systems**

We introduce TAM-RL (Task Aware Modulation using Representation Learning), a novel multimodal meta-learning framework for few-shot learning in heterogeneous systems, designed for science and engineering problems where entities share a common underlying forward model but exhibit heterogeneity due to entity-specific characteristics. TAM-RL leverages an amortized training process with

a modulation network and a base network to learn task-specific modulation parameters, enabling efficient adaptation to new tasks with limited data. We evaluate TAM-RL on two real-world environmental datasets: Gross Primary Product (GPP) prediction and streamflow forecasting, demonstrating significant improvements over existing meta-learning methods. On the FLUXNET dataset, TAM-RL improves RMSE by 18.9% over MMAML with just one month of few-shot data, while for streamflow prediction, it achieves an 8.21% improvement with one year of data. Synthetic data experiments further validate TAM-RL's superior performance in heterogeneous task distributions, outperforming the baselines in the most heterogeneous setting. Notably, TAM-RL offers substantial computational efficiency, with at least 3x faster training times compared to gradient-based meta-learning approaches while being much simpler to train due to reduced complexity. Ablation studies highlight the importance of pretraining and adaptation mechanisms in TAM-RL's performance.

Arvind Renganathan, Rahul Ghosh, Ankush Khandelwal, Vipin Kumar
University of Minnesota
renga016@umn.edu, ghosh128@umn.edu, khand035@umn.edu, kumar001@umn.edu

**CP1**

**OpenFE++: Efficient Automated Feature Generation via Feature Interaction**

Automated feature generation can greatly enhance the performance of machine learning models in many tabular and time-series prediction problems. The current state-of-the-art method, OpenFE, follows the "expand-and-reduce" framework, which generates a candidate feature set and subsequently extracts the most effective features. Nevertheless, with the increase in the number of features and the length of time series, feature generation algorithms based on expand-and-reduce often produce an overwhelmingly large pool of candidate features, rendering the identification of effective features quite time-consuming and more prone to overfitting. To resolve this issue, we propose OpenFE++, which leverages the feature interactions from both the feature and temporal dimensions to construct a substantially reduced candidate feature set, thereby enhancing efficiency and effectiveness. In the feature dimension, OpenFE++ utilizes locally interacted features to generate meaningful candidate features without exhaustively enumerating all possibilities. In the temporal dimension, it evaluates the lagged effects among different features and generates temporally meaningful features via the representative lagged periods. Thus, OpenFE++ can efficiently generate effective features to boost the model performance. We conduct extensive experiments on fourteen widely used benchmark datasets to demonstrate that OpenFE++ outperforms other baseline models in both efficiency and effectiveness.

Lei Wang, Yu Shi
Tsinghua University
wanglei20@mails.tsinghua.edu.cn, shi-y23@mails.tsinghua.edu.cn

Yifei Jin
X-technology (Beijing) Co. Ltd.
yfjin1990@gmail.com

Jian Li
Tsinghua University
lijian83@mail.tsinghua.edu.cn

**CP2**

**Trajectory Anomaly Detection with By-Design Complementary Detectors**

Trajectory anomaly detection is critical across a wide range of applications, from traffic control, and wildlife conservation, to public transportation optimization. However, detecting anomalies in trajectory data is challenging due to the diverse nature of anomalies. In this paper, we propose CETrajAD, an ensemble method for trajectory anomaly detection that integrates complementary detectors, each targeting different aspects of trajectory anomalies. Our approach leverages three types of trajectory embeddingsRoute, Speed, and Shapethat vary in their sensitivity to length, direction, shape, and speed, enabling the detection of diverse anomaly types. We combine detectors from both the embedding and input spaces and show how their complementary nature improves anomaly detection performance. Through theoretical analysis, we demonstrate the conditions when the proposed ensemble design outperforms traditional ensemble methods. Experiments on multiple real-world datasets, containing both simulated and ground-truth anomalies, show that the proposed model consistently outperforms existing baselines.

Shurui Cao, Leman Akoglu
Carnegie Mellon University
shuruic@andrew.cmu.edu, lakoglu@andrew.cmu.edu

**CP2**

**Convergence-Guaranteed Elastic Net Graphical Model Estimation with Applications to Anomaly Localization**

Estimating dependency structures from noisy multivariate variables is fundamentally important in many applications. Of particular importance in practice is anomaly localization, which is to compute a variable-wise anomaly score by comparing a target dependency structure to a reference structure. In this task, stably and accurately estimating the dependency structures is the key. First, we present an L0-elastic net model for estimating sparse inverse covariance matrices. Then we introduce a framework for anomaly localization that utilizes both the L0-elastic net model and a transfer learning model. Although L0-constrained optimization is known to be challenging, we introduce a hard thresholding line-search algorithm to efficiently solve these graphical models. Using synthetic and real-world data sets, we demonstrate that the proposed L0-based method systematically outperforms alternative methods in many use-cases.

Dzung Phan
IBM T.J. Watson Research Center
phandu@us.ibm.com

Matt Menickelly
Argonne National Laboratory
mmenickelly@anl.gov

Tsuyoshi Ide, Jayant Kalagnanam
IBM T.J. Watson Research Center
tide@us.ibm.com, jayant@us.ibm.com

**CP2**

**Anomaly Detection Via Graph Contrastive Learn-**

**ing**

Emre Sefer
Ozyegin University
emre.sefer@ozyegin.edu.tr

**CP2**

**Federated Koopman-Reservoir Learning for Large-Scale Multivariate Time-Series Anomaly Detection**

The proliferation of edge devices has dramatically increased the generation of multivariate time-series (MVTS) data, essential for applications from healthcare to smart cities. Such data streams, however, are vulnerable to anomalies that signal crucial problems like system failures or security incidents. Traditional MVTS anomaly detection methods, encompassing statistical and centralized machine learning approaches, struggle with the heterogeneity, variability, and privacy concerns of large-scale, distributed environments. In response, we introduce FedKO, a novel unsupervised Federated Learning framework that leverages the linear predictive capabilities of Koopman operator theory along with the dynamic adaptability of Reservoir Computing. This enables effective spatiotemporal processing and privacy-preserving for MVTS data. FedKO is formulated as a bi-level optimization problem, utilizing a specific federated algorithm to explore a shared Reservoir-Koopman model across diverse datasets. Such a model is then deployable on edge devices for efficient detection of anomalies in local MVTS streams. Experimental results across various datasets showcase FedKO's superior performance against state-of-the-art methods in MVTS anomaly detection. Moreover, FedKO reduces up to 8 times communication size and 2 times memory usage, making it highly suitable for large-scale systems.

Long T. Le, Tung-Anh Nguyen, Han Shu, Suranga Seneviratne
School of Computer Science
The University of Sydney
long.le@sydney.edu.au, tung6100@uni.sydney.edu.au, hshu3770@uni.sydney.edu.au, suranga.seneviratne@sydney.edu.au

Choong Seon Hong
School of Computing
Kyung Hee University
cshong@khu.ac.kr

Nguyen Tran
School of Computer Science, University of Sydney
nguyen.tran@sydney.edu.au

**CP3**

**$\ell_{1,\infty}$-Mixed Norm Promoted Row Sparsity for Fast Online Cur Decomposition Learning in Varying Feature Spaces**

Online learning enables effective predictive modeling on complex data streams. To overcome the negative impact of possibly high-dimensional data, sparse online learning (SOL) has been proposed by imposing various sparse constraints to sheer the resultant model structure. However, most existing SOL studies focused on a fixed feature space, whereas in practice the steaming data observations may increase in both quantity and feature dimensions, leading to varying feature spaces. In this paper, we propose a novel $\ell_{1,\infty}$-mixed norm-based row sparsity SOL algorithm (SOOFS) to handle data streams in varying feature spaces. We empower SOOFS with a tailored online CUR matrix decomposition method based on the promoted row sparsity to actively and adaptively select informative instances in the sliding windows, facilitating stable online performance over time. Empirical results on ten benchmark datasets substantiate the superiority of SOOFS over three state-of-the-art competitors in terms of classification accuracy and model sparsity. **Keywords**: online learning, sparse learning, CUR decomposition, online active learning, $\ell_{1,\infty}$ mixed-norm.

Zhong Chen
Southern Illinois University
zhong.chen@cs.siu.edu

Yi He
College of William & Mary
yihe@wm.edu

Di Wu
Southwest University
wudi.cigit@gmail.com

Wenbin Zhang
Florida International University
wenbin.zhang@fiu.edu

Zhiqiang Deng
Louisiana State University
zdeng@lsu.edu

**CP3**

**TADAM: Learning Timed Automata from Noisy Observations**

Timed Automata (TA) are formal models capable of representing regular languages with timing constraints, making them well-suited for modeling systems where behavior is driven by events occurring over time. Most existing work on TA learning relies on active learning, where access to a teacher is assumed to answer membership queries and provide counterexamples. While this framework offers strong theoretical guarantees, it is impractical for many real-world applications where such a teacher is unavailable. In contrast, passive learning approaches aim to infer TA solely from sequences accepted by the target automaton. However, current methods struggle to handle noise in the data, such as symbol omissions, insertions, or permutations, often resulting in excessively large and inaccurate automata. In this paper, we introduce TADAM, a novel approach that leverages the Minimum Description Length (MDL) principle to balance model complexity and data fit, allowing it to distinguish between meaningful patterns and noise. We show that TADAM is significantly more robust to noisy data than existing techniques, less prone to overfitting, and produces concise models that can be manually audited. We further demonstrate its practical utility through experiments on real-world tasks, such as network flow classification and anomaly detection.

Lénaïg Cornanguer
CISPA
lenaig.cornanguer@cispa.de

Pierre-François Gimenez
CentraleSupélec, Inria, Univ. Rennes, IRISA

pierre-francois.gimenez@centralesupelec.fr

**CP3**

**Learning Confident Classifiers in the Presence of Label Noise**

The success of Deep Neural Network (DNN) models significantly depends on the quality of provided annotations. In medical image segmentation, for example, having multiple expert annotations for each data point is standard to minimize subjective annotation bias. Then, the goal of estimation is to filter out the label noise and recover the ground-truth masks, which are not explicitly given. This paper proposes a probabilistic model for noisy observations that allows us to build confident classification and segmentation models. We explicitly model label noise to accomplish this and introduce a new information-based regularization that pushes the network to recover the ground-truth labels. In addition, we adjust the loss function for the segmentation task by prioritizing learning in high-confidence regions where all the annotators agree on labeling. We evaluate the proposed method on a series of classification tasks such as noisy versions of MNIST, CIFAR-10, and Fashion-MNIST datasets, as well as CIFAR-10N, a real-world dataset with noisy human annotations. Additionally, for the segmentation task, we consider several medical imaging datasets, such as LIDC and RIGA, that reflect real-world inter-variability among multiple annotators. Our experiments show that our algorithm outperforms state-of-the-art solutions for the considered classification and segmentation problems.

Asma Ahmed Hashmi
ifi.lmu.edu
asmah17@gmail.com

Aigerim Zhumabayeva
MBZUAI, UAE
zhumabayeva.aigerim@gmail.com

Nikita Kotelevskii
Skoltech, Russia; MBZUAI, UAE
niket096@mail.ru

Artem Agafonov
Mohamed bin Zayed University of Artificial Intelligence
artem.agafonov@mbzuai.ac.ae

Mohammed Yaqub, Maxim Panov
MBZUAI, UAE
mohammad.yaqub@mbzuai.ac.ae,
panov.maxim@gmail.com

Martin Takac
Lehigh University
martin.taki@gmail.com

**CP3**

**Constraint-Focused Training for Multistate Survival Analysis with Neural Networks**

Deep learning has significantly improved performance in survival analysis (SA), which plays a crucial role in various fields, including medicine and engineering. A recent approach involves directly modeling the transition probability matrix using neural networks (NNs). However, modeling NNs as a transition probability matrix necessitates satisfying the constraints specific to transition probabilities. Existing methods often struggle to meet these constraints without compromising model performance (e.g., by limiting the model architectures). We propose a novel deep learning method for SA that models the transition probability matrix using NNs and trains them to meet the constraints through a specialized loss function. This loss function is designed to penalize the violation of the constraints using automatic differentiation (AD). Since any model capable of AD can compute this loss function, it enables the use of flexible and potentially higher-performing model architectures while forcing the model to satisfy the constraints of the transition probability. The experiments on real-world datasets demonstrate the superior performance of the proposed method compared to existing approaches. Additionally, ablation studies verify that all the components in the proposed method contribute to the performance.

Ryosuke Takayama, Masanao Natsumeda
NEC Corporation
ryosuke-takayama@nec.com, mnatsumeda@nec.com

**CP3**

**Label Shift Estimation With Incremental Prior Update**

An assumption often made in supervised learning is that the training and testing sets have the same label distribution. However, in real-life scenarios, this assumption rarely holds. For example, medical diagnosis result distributions change over time and across locations; fraud detection models must adapt as patterns of fraudulent activity shift; the category distribution of social media posts changes based on trending topics and user demographics. In label shift estimation, the goal is to estimate the changing label distribution $p_t(y)$ in the testing set, assuming the likelihood $p(x|y)$ does not change, implying no concept drift. In this paper, we propose a new post-hoc label shift estimation method, unlike prior methods that perform moment matching with confusion matrices or maximize the new data likelihood with an expectation-maximization algorithm. We incrementally update the prior for each sample, adjusting posteriors for more accurate label shift estimation. The proposed method is based on intuitive assumptions about modern probabilistic classifiers and relies on a weaker notion of calibration. As a post-hoc approach, it is versatile and can be applied to any black-box probabilistic classifier. Experiments on CIFAR-10 and MNIST show it consistently outperforms state-of-the-art maximum likelihood-based methods under different calibrations and label shift intensities.

Yunrui Zhang
University of New South Wales
yunrui.zhang@student.unsw.edu.au

Gustavo Batista
University of New South Wales (UNSW), Sydney
g.batista@unsw.edu.au

Salil Kanhere
University of New South Wales
salil.kanhere@unsw.edu.au

**CP4**

**Conformal Edge-Weight Prediction in Latent Space**

Predicting the edge weights of a graph is a critical task across many domains. Some examples include predict-

ing traffic flow in transportation networks, strength of interactions in protein-protein networks, and collaboration frequency in co-authorship networks. Graph Neural Networks have been very successful in edge-weight prediction tasks. However, these predictions lack rigorous statistical uncertainty quantification. Recent work has demonstrated the efficacy of conformal inference in quantifying the uncertainties of the predictions made by graph neural networks. However, there has been limited research in conformal inference for edge-weight prediction. A prior work has demonstrated that the powerful inductive bias of the representations deep neural networks can be leveraged for robust uncertainty quantification. In this paper, we extend the traditional conformal inference paradigm to compute uncertainty estimates in the latent space exploiting the deep representations learned by graph neural networks. Specifically, our method computes the non-conformity score in the latent feature space of the nodes and leverages the score for bandwidth estimation for weighted edge prediction. Experiments on a wide variety of edge-weighted networks show that Edge-CP always achieves the pre-defined target marginal coverage and obtains up to 38.16% shorter bands than the nearest baseline. Additionally, Edge-CP achieves the best conditional coverage among all methods.

Akash Choudhuri, Yongjian Zhong, Mehrdad moharrami
University of Iowa
Department of Computer Science
akash-choudhuri@uiowa.edu, yongjian-zhong@uiowa.edu, mehrdad-moharrami@uiowa.edu

Christine Klymko, Mark Heimann
Lawrence Livermore National Laboratory
klymko1@llnl.gov, heimann2@llnl.gov

Jayaraman thiagarajan
Apple Inc.
jjthiagarajan@gmail.com

Bijaya Adhikari
Department of Computer Science,
University of Iowa
bijaya-adhikari@uiowa.edu

**CP4**

**Defense Against Shortest Path Attacks**

Identifying shortest paths between nodes in a network is an important task in many applications. Recent work has shown that a malicious actor can manipulate a graph to make traffic between two nodes of interest follow their target path. In this paper, we develop a defense against such attacks by modifying the edge weights that users observe. The defender must balance inhibiting the attacker against any negative effects on benign users. Specifically, the defenders goals are: (a) recommend the shortest paths to users, (b) make the lengths of the shortest paths in the published graph close to those of the same paths in the true graph, and (c) minimize the probability of an attack. We formulate the defense as a Stackelberg game in which the defender is the leader and the attacker is the follower. We also consider a zero-sum version of the game in which the defenders goal is to minimize cost while achieving the minimum possible attack probability. We show that the defense problem is NP-hard and propose heuristic solutions for both the zero-sum and non-zero-sum settings. By relaxing some constraints of the original problem, we formulate a linear program for local optimization around a feasible point. We present defense results with both synthetic and real networks and show that our methods often reach the lower bound of the defenders cost.

Benjamin A. Miller
MIT Lincoln Laboratory
miller.be@northeastern.edu

Zohair Shafi
Northeastern University
shafi.z@northeastern.edu

Wheeler Ruml
University of New Hampshire
ruml@cs.unh.edu

Yevgeniy Vorobeychik
Washington University in St. Louis
yvorobeychik@wustl.edu

Tina Eliassi-Rad
Northeastern University
t.eliassirad@northeastern.edu

Scott Alfeld
Amherst College
alfeld@amherst.edu

**CP4**

**Efficient Sampling of Temporal Networks with Preserved Causality Structure**

In this paper, we extend the classical Color Refinement algorithm for static networks to temporal (undirected and directed) networks. This enables us to design an algorithm to sample synthetic networks that preserves the $d$-hop neighborhood structure of a given temporal network. The higher $d$ is chosen, the better the temporal neighborhood structure of the original network is preserved. Specifically, we provide efficient algorithms that preserve time-respecting ("causal") paths in the networks up to length $d$, and scale to real-world network sizes. We validate our approach theoretically (for Degree and Katz centrality) and experimentally (for edge persistence, causal triangles, and burstiness). An experimental comparison shows that our method retains these key temporal characteristics more effectively than existing randomization methods.

Felix I. Stamm
RWTH Aachen University
felix.stamm@cssh.rwth-aachen.de

Mehdi Naima
Sorbonne Université, CNRS, LIP6, F-75005
mehdi.naima@lip6.fr

Michael T. Schaub
RWTH Aachen University
schaub@cs.rwth-aachen.de

**CP4**

**StarRec: A Hypergraph-Based Framework with Star-Expansion for Multi-Behavior Recommendation**

In modern recommendation systems, leveraging multiple types of user-item interaction behaviors (e.g., click, add-to-cart, and purchase) presents both advantages and

challenges. Recent studies organize multi-behavior data into heterogeneous bipartite graphs and used graph neural networks to learn latent representations. However, these methods struggle to model higher-order interactions and capture complex dependencies across various behaviors. In this paper, we propose a novel graph construction method that converts multi-behavior interactions into dual star-expansion hypergraphs by introducing a new type of node called hypernode. Subsequently, we develop StarRec, a hypergraph-based framework for multi-behavior recommendation. StarRec utilizes a spatial-based two-stage intra-behavior message passing strategy and a cross-behavior propagation layer to accurately and efficiently propagate information, modeling both high-order and cross-behavior relationships through the hypergraphs. This approach yields comprehensive representations that enhance recommendation performance. Experimental results on two real-world datasets demonstrate the superiority of StarRec. Our extensive experiments show that StarRec significantly outperforms state-of-the-art methods while maintaining competitive model scalability.

Zijian Song, Wenhan Zhang, Yihuan Wu
Peking University
2201111590@stu.pku.edu.cn, pku_zwh@pku.edu.cn,
2200013181@stu.pku.edu.cn

Lifang Deng, Jiandong Zhang
Lazada Group
wanmei.dlf@alibaba-inc.com,
chensong.zjd@alibaba-inc.com

Kaigui Bian, Bin Cui
Peking University
bkg@pku.edu.cn, bin.cui@pku.edu.cn

**CP5**

**Beyond Models! Explainable Data Valuation and Metric Adaption for Recommendation**

User behavior records serve as the foundation for recommender systems. While the behavior data exhibits ease of acquisition, it often suffers from varying quality. Current methods employ data valuation to discern high-quality data from low-quality data. However, they tend to employ black-box design, lacking transparency and interpretability. Besides, they are typically tailored to specific evaluation metrics, leading to limited generality across various tasks. To overcome these issues, we propose an explainable and versatile framework DVR which can enhance the efficiency of data utilization tailored to any requirements of the model architectures and evaluation metrics. For explainable data valuation, a data valuator is presented to evaluate the data quality via calculating its Shapley value from the game-theoretic perspective, ensuring robust mathematical properties and reliability. In order to accommodate various evaluation metrics, including differentiable and non-differentiable ones, a metric adapter is devised based on reinforcement learning, where a metric is treated as the reinforcement reward that guides model optimization. Extensive experiments conducted on various benchmarks verify that our framework can improve the performance of current recommendation algorithms on various metrics including ranking accuracy, diversity, and fairness. Specifically, our framework achieves up to 34.7% improvements over existing methods in terms of representative NDCG metric.

Renqi Jia
City University of Hong Kong
renqijia2-c@my.cityu.edu.hk

Xiaokun Zhang, Bowei He
City University of Hong Kong
Hong Kong SAR
dawnkun1993@gmail.com, boweihe2-c@my.cityu.edu.hk

Qiannan Zhu
Beijing Normal University
zhuqiannan@bnu.edu.cn

Weitao Xu
City University of Hong Kong
Hong Kong SAR
weitaoxu@cityu.edu.hk

Jiehao Chen
China Academy of Industrial Internet
chenjiehao@china-aii.com

Chen Ma
City University of Hong Kong
chenma@cityu.edu.hk

**CP5**

**Cdsrnp: Cross-Domain Sequential Recommendation Via Neural Process**

Cross-Domain Sequential Recommendation is a hot-topic in user interest modeling, which aims at utilizing a single model to predict the next items for different domains. Many methods are focused on domain overlapped users behaviors fitting, which heavily relies on the same user s different-domain item sequences collaborating signals to capture the synergy of cross-domain item-item correlation. Indeed, these overlapped users occupy a small fraction of the entire user set only, which introduces a strong assumption that the small group of domain overlapped users is enough to represent all domain user behavior characteristics. However, intuitively, such a suggestion is biased, and it will inevitably limit model performance. Further, it is not trivial to model non-overlapped user behaviors in CDSR because there are no other domain behaviors to collaborate with, which causes the observed single-domain users behavior sequences to be hard to contribute to cross-domain knowledge mining. We raise a challenging and unexplored question: How to unleash the potential of non-overlapped users behaviors to empower CDSR? To this end, we propose a novel CDSR framework with neural processes (NP), briefly termed CDSRNP, where NP combines the advantages of meta-learning and stochastic processes. Experimental results illustrate that CDSRNP outperforms state-of-the-art methods in real-world datasets.

Haipeng Li
Nanjing University of Science and Technology
lihaipeng@njust.edu.cn

Jiangxia Cao
Kuaishou Technology
caojiangxia@kuaishou.com

Yiwen Gao
Nanjing University of Science and Technology
gaoyiwen@njust.edu.cn

Yunhuai Liu
Peking University

yunhuai.liu@pku.edu.cn

Shuchao Pang
Nanjing University of Science and Technology
pangshuchao@njust.edu.cn

**CP5**

**A Look into News Avoidance through AWRS: An Avoidance-Aware Recommender System**

In recent years, journalists have expressed concerns about the increasing trend of news article avoidance, especially within specific domains. This issue has been exacerbated by the rise of recommender systems. Our research indicates that recommender systems should consider avoidance as a fundamental factor. We argue that news articles can be characterized by three principal elements: exposure, relevance, and avoidance, all of which are closely interconnected. To address these challenges, we introduce AWRS, an Avoidance-Aware Recommender System. This framework incorporates avoidance awareness when recommending news, based on the premise that news article avoidance conveys significant information about user preferences. Evaluation results on three news datasets in different languages (English, Norwegian, and Japanese) demonstrate that our method outperforms existing approaches.

Igor Lima Rocha Azevedo, Toyotaro Suzumura
The University of Tokyo
igorlima1740@gmail.com, suzumura@g.ecc.u-tokyo.ac.jp

Yuichiro Yasui
Nikkei Inc.
yuichiro.yasui@nex.nikkei.com

**CP5**

**Ranking with Confidence for Large Scale Comparison Data**

In this work, we leverage a generative data model considering comparison noise to develop a fast, precise, and informative ranking algorithm from pairwise comparisons that produces a measure of confidence on each comparison. The problem of ranking a large number of items from noisy and sparse pairwise comparison data arises in diverse applications, like ranking players in online games, document retrieval or ranking human perceptions. Although different algorithms are available, we need fast, large-scale algorithms whose accuracy degrades gracefully when the number of comparisons is too small. Fitting our proposed model entails solving a non-convex optimization problem, which we tightly approximate by a sum of quasi-convex functions and a regularization term. Resorting to an iterative reweighted minimization and the Primal-Dual Hybrid Gradient method, we obtain PD-Rank, achieving a better Kendall tau than comparing methods, even for 10% of wrong comparisons in simulated data matching our data model and without the assumption of strong connectivity. In real data, PD-Rank requires less computational time to achieve the same Kendall tau than active learning methods.

Filipa Valdeira
Nova School of Science and Technology, NOVA FCT
f.valdeira@fct.unl.pt

Cláudia Soares
NOVA School of Science and Technology, NOVA FCT
claudia.soares@fct.unl.pt

**CP6**

**DMDHC: Discovery of Multi-Density Hierarchical Cluster Structures**

Hierarchical clustering techniques can reveal nested structures within data by representing patterns in a tree-like form. However, when dealing with complex data, many traditional hierarchical methods produce cluttered and hard-to-interpret trees. To address this, we propose a novel hierarchical clustering method called Discovery of Multi-Density Hierarchical Cluster structures (DMDHC), which introduces a new type of cluster tree to represent hierarchical information more effectively. Our approach automatically generates hierarchical local cuts along the tree structure. In contrast to state-of-the-art methods like PEARCH, which typically apply only a single cut across the hierarchy, DMDHC takes advantage of density-based insights to perform multiple cuts at different levels. This results in a more compact and comprehensible representation of intricate hierarchical structures. Extensive experiments on real-world datasets demonstrate that DMDHC, along with its newly introduced tree structure, outperforms existing methods.

Walid Durani
Ludwig-Maximilians-Universität
durani@dbs.ifi.lmu.de

Dominik Mautz
LMU Munich
mautz@dbs.ifi.lmu.de

Claudia Plant, Christian Böhm
University of Vienna
claudia.plant@univie.ac.at, christian.boehm@univie.ac.at

**CP6**

**Hierarchical Superpixel Segmentation Via Structural Information Theory**

Superpixel segmentation is a foundation for many higher-level computer vision tasks, such as image segmentation, ob- ject recognition, and scene understanding. Existing graph- based superpixel segmentation methods typically concen- trate on the relationships between a given pixel and its directly adjacent pixels while overlooking the influence of non-adjacent pixels. These approaches do not fully leverage the global information in the graph, leading to sub- optimal segmentation quality. To address this limitation, we present SIT-HSS, a hierarchical superpixel segmentation method based on structural information theory. Specifically, we first design a novel graph construction strategy that incrementally explores the pixel neighborhood to add edges based on 1-dimensional structural entropy (1D SE). This strategy maximizes the retention of graph information while avoiding an overly complex graph structure. Then, we design a new 2D SE-guided hierarchical graph partitioning method, which iteratively merges pixel clusters layer by layer to reduce the graphs 2D SE until a predefined segmentation scale is achieved. Experimental results on three benchmark datasets demonstrate that the SIT-HSS performs better than state-of-the-art unsupervised superpixel segmentation algorithms. The source code is available at https://github.com/SELGroup/SIT-HSS.

Xie Minhui

Beihang University
xieminhui@buaa.edu.cn

First Name Last Name

Peng Hao, Li Pu, Zeng Guangjie
Beihang University
penghao@buaa.edu.cn, lp2024626@gmail.com,
zengguangjie@buaa.edu.cn

Wang Shuhai
y, Shijiazhuang Tiedao University
wsh36302@126.com

Wu Jia
Macquarie University
jia.wu@mq.edu.au

Yu Philip S.
University of Illinois at Chicago
syu@uic.edu

**CP6**

**Differentially Private Associative Co-Clustering**

Co-clustering is a useful tool that extracts summary information from a data matrix in terms of row and column clusters, and gives a succinct representation of the data. However, if the matrix contains data about individuals, such representations could leak their privacy-sensitive information. In terms of privacy disclosure, co-clustering is even more harmful than clustering, because of the additional information carried by the column partition. However, to the best of our knowledge, the problem of privacy-preserving co-clustering has never been studied. To fill this gap, we consider a recent co-clustering algorithm, based on a de-normalized version of the Goodman-Kruskal's $\tau$ association measure, which has a good property from a differential privacy perspective, and is supposed not to consume an excessive amount of privacy budget. This leads to a privacy-preserving co-clustering algorithm that satisfies the definition of differential privacy while providing good partitioning solutions. Our algorithm is based on a prototype-based optimization strategy that makes it fast and actionable in realistic privacy-preserving data management and analysis scenarios, as shown by our extensive experimental validation.

Elena Battaglia, Ruggero G. Pensa
University of Turin
Department of Computer Science
elenabatt.eb@gmail.com, ruggero.pensa@unito.it

**CP6**

**Multi-View Spectral Clustering for Graphs with Multiple View Structures**

Despite the fundamental importance of clustering, to this day, much of the relevant research is still based on ambiguous foundations, leading to an unclear understanding of whether or how the various clustering methods are connected with each other. In this work, we provide an additional stepping stone towards resolving such ambiguities by presenting a general clustering framework that subsumes a series of seemingly disparate clustering methods, including various methods belonging to the widely popular spectral clustering framework. In fact, the generality of the proposed framework is additionally capable of shedding light to the largely unexplored area of multi-view graphs where each view may have differently clustered nodes. In turn, we propose GenClus: a method that is simultaneously an instance of this framework and a generalization of spectral clustering, while also being closely related to k-means as well. This results in a principled alternative to the few existing methods studying this special type of multi-view graphs. Then, we conduct in-depth experiments, which demonstrate that GenClus is more computationally efficient than existing methods, while also attaining similar or better clustering performance. Lastly, a qualitative real-world case-study further demonstrates the ability of GenClus to produce meaningful clusterings.

Yorgos Tsitsikas, Evangelos E. Papalexakis
University of California, Riverside
gtsit001@ucr.edu, epapalex@cs.ucr.edu

**CP6**

**DynHAC: Fully Dynamic Approximate Hierarchical Agglomerative Clustering**

We consider the problem of maintaining a hierarchical agglomerative clustering (HAC) in the dynamic setting, when the input is subject to point insertions and deletions. We introduce DynHac – the first dynamic HAC algorithm for the popular average-linkage version of the problem which can maintain a $1 + \epsilon$ approximate solution. Our approach leverages recent structural results on $1 + \epsilon$-approximate HAC to carefully identify the part of the clustering dendrogram that needs to be updated in order to produce a solution that is consistent with what a full recomputation from scratch would have output. We evaluate DynHAC on a number of real-world graphs. We show that DynHAC can handle each update up to 423x faster than what it would take to recompute the clustering from scratch. At the same time it achieves up to 0.21 higher NMI score than the state-of-the-art dynamic hierarchical clustering algorithms, which do not provably approximate HAC.

Shangdi Yu
MIT
shangdiy@mit.edu

Laxman Dhulipala
University of Maryland, College Park, MD
laxman@umd.edu

Jakub Lacki, Nikos Parotsidis
Google Research
jlacki@google.com, nikosp@google.com

**CP7**

**Unanticipated Replenishment: Online Policy for Dynamic Service Composition in Manufacturing Cloud**

Service composition a pivotal step in allocating budgeted cloud services to fulfil requests arriving on service platforms. These requests usually arrive in an online pattern, i.e., we do not know which request comes next until the request realises itself. In addition, new services and service providers are eager to join the platform. Thus, a legitimate concern arises: how to allocate services with such service replenishment for the online setting, especially when there is no prior knowledge available for the replenishment or requests? To tackle this challenging problem, we develop a Dual-Price based Online Learning (DPOL) algorithm, and prove that DPOL can achieve a sub-linear regret bound

of $O(\sqrt{T})$ against the offline benchmark, where $T$ is the length of the planning horizon. Also, numerical experiments on synthetic datasets validate that the performance of our algorithms outperforms other baseline policies.

Yang Hu
City University of Hong Kong, College of Computing
yhu263-c@my.cityu.edu.hk

Feng Wu
Xi'an Jiaotong University, School of Management
Xi'an, Shaanxi, China
fengwu830@126.com

Xin li, Yu yang
City University of Hong Kong, College of Computing
Hong Kong S.A.R., China
xinli394@cityu.edu.hk, yuyang@cityu.edu.hk

**CP7**

**Context-Aware Frequency-Embedding Networks for Spatio-Temporal Portfolio Selection**

Recent developments in the applications of deep reinforcement learning methods to portfolio selection have achieved superior performance to conventional methods. However, two major challenges remain unaddressed in these models and inevitably lead to the deterioration of model performance. First, asset characteristics often suffer from low and unstable signal-to-noise ratios, leading to poor learning robustness of the predictive feature representations. Second, existing literature fails to consider the complexity and diversity in long-term and short-term spatio-temporal predictive relations between the feature sequences and portfolio objectives. To tackle these problems, we propose a novel Context-Aware Frequency-Embedding Graph Convolution Network (Cafe-GCN) for spatio-temporal portfolio selection. It contains three important modules: (1) frequency-embedding block that explicitly captures the short-term and long-term predictive information embedded in asset characteristics meanwhile filtering out noise; (2) context-aware block that learns multiscale temporal dependencies in the feature space; and (3) multi-relation graph convolutional block that exploits both static and dynamic spatial relations among assets. Extensive experiments on two real-world datasets demonstrate that Cafe-GCN consistently outperforms proposed techniques in the literature.

Ruirui Liu
Brunel University of London
King's College London
ruirui.liu@kcl.ac.uk

Huichou Huang
City University of Hong Kong
Bayescien Technologies
huichou.huang@exeter.oxon.org

Johannes Ruf
London School of Economics and Political Science
j.ruf@lse.ac.uk

Haoxian Liu, Qingyao Wu
South China University of Technology
241020710@qq.com, qyw@scut.edu.cn

**CP7**

**Domain-Adaptive Continual Meta-Learning for Modeling Dynamical Systems: An Application in Environmental Ecosystems**

Environmental ecosystems exhibit complex and evolving dynamics over time, making the modeling of non-stationary processes critically important. However, traditional methods often rely on static models trained on entire datasets, failing to capture the non-stationary and drastically fluctuating characteristics. Dynamically adjusting models to evolving data is challenging, as they can easily either lag behind new trends or overfit newly received data. To address these challenges, we propose Domain-Adaptive Continual Meta-Learning (DACM) method, aiming to automatically detect distribution shifts and adapt to newly emergent domains. In particular, while DACM continuously explores the sequential temporal data, it also exploits historical data that are similar in distribution to the current observations. By striking a balance between temporal exploration and distributional exploitation, DACM quickly adjusts the model to stay up-to-date with new trends while maintaining generalization ability to data with similar distributions. We demonstrate the effectiveness of DACM on a real-world water temperature prediction dataset, where it outperforms diverse baseline models and shows strong adaptability and predictive performance in non-stationary environments.

Yiming Sun, Runlong Yu, Runxue Bao
University of Pittsburgh
yimingsun@pitt.edu, ruy59@pitt.edu,
runxue.bao@pitt.edu

Yiqun Xie
University of Maryland
xie@umd.edu

Ye Ye
University of Pittsburgh
yey5@pitt.edu

Xiaowei Jia
U. of Pittsburgh
xiaowei@pitt.edu

**CP7**

**Inter-Well Active Magnetic Ranging with Temporal and Interaction Network**

Inter-well distance measurement is crucial for safety in oil and gas drilling operations, especially during blowout emergencies where rescue wells mitigate risks. Traditional methods, like GPS and ultrasonic ranging, are ineffective in downhole environments. Magnetic ranging methods, such as Passive and Active Magnetic Ranging (AMR), offer solutions but face challenges like interference and handling complex well structures. We propose a novel method integrating AMR with a Temporal and Interaction Network (TINet). Using active magnetic ranging tools, we collected and calibrated a dataset comprising magnetic and non-magnetic data. This dataset trains TINet, which employs enhanced feature extraction via improved sample convolution and interaction networks, combined with a long short-term memory unit to capture temporal-spatial relationships and utilize historical data. To further enhance accuracy, we introduce Adaptive Error Scaling Loss, balancing relative and absolute errors. Experiments show our method significantly outperforms traditional approaches across evaluation metrics and distance scales.

Hao Zelong, Haitao Zhang

Beijing University of Posts and Telecommunications
hzzll@bupt.edu.cn, zht@bupt.edu.cn

Yang Che
CNPC Engineering Technology R&D Company Ltd
cheyangdri@cnpc.com.cn

Wang Liu
Beijing University of Posts and Telecommunications
lw2055610463@163.com

**CP7**

**Bridging Numbers and Narratives: Enhancing Financial Market Risk Predictions Through Numerical Information from Financial Documents**

Modeling financial documents is crucial for financial data mining and practice, offering insights into market dynamics and improving risk management. However, traditional approaches focus on narrative content while neglecting numerical data in financial texts due to language models' limited numerical understanding, reducing risk prediction effectiveness. We propose a novel framework integrating numerical and narrative information to enhance financial document modeling. By identifying, enriching, and combining numerical data, our method improves financial market risk prediction accuracy. Validated with Russell 3000 company data, our approach outperforms state-of-the-art techniques. Ablation studies confirm numerical data's role in boosting performance. This work bridges the gap between traditional financial document modeling and real-world practices by combining advanced NLP with numerical analysis, providing researchers and practitioners with tools for better information extraction and investment opportunity identification.

Yu Qin, Chengshang Zhang
Renmin University of China
yu.qin@ruc.edu.cn, seoms@ruc.edu.cn

Wei Xu
Renmin University of China, Information school
Weixu@ruc.edu.cn

**CP8**

**Heterogeneous Multi-Agent Framework for Dynamic Generalized Category Discovery**

In the fast-paced realm of open-world machine learning, Generalized Category Discovery (GCD) has emerged as a crucial task for identifying new classes within ever-evolving datasets. With the rise of multimodal data that includes text, images, audio, and video, traditional GCD methods, which often rely on parametric classifiers and single-modality inputs, face significant limitations. These approaches can lead to overfitting and hinder the ability to generalize to new categories effectively. This paper highlights the pressing need for innovative strategies that harness the richness of multimodal data to enhance contextual understanding and facilitate real-time category identification. We aim to establish a foundational framework for future GCD research, promoting a more agile and resilient approach to data classification in todays complex information landscape. To achieve this, we propose a dynamic framework that integrates heterogeneous multi-agent systems, combining Large Language Models (LLMs) with diverse non-LLM methodologies. This approach not only enhances the adaptability and robustness of GCD solutions but also opens up transformative possibilities across critical fields such as autonomous driving, medical diagnostics, and social media analysis.

Fatimah D. Alotaibi
Virginia Tech
falotaibi@vt.edu

Adithya Kulkarni
Iowa State University
aditkulk@vt.edu

Dawei Zhou
Virginia Tech
zhoud@vt.edu

**CP8**

**Explainable Ai for Real-Time Video Anomaly Anticipation**

With computational power expanding on the edge and deep learning models now capable of real-time inference, it is time to rethink our approach to anomalies. Instead of waiting for anomalies to happen and then detecting them, why not predict them before they occurand prevent them altogether? Imagine a system that can look at the data at time $t$ and warn us that something might go wrong at $t + h$. If $h$ is long enough, we can act—automatically or semi-automatically—to stop the anomaly in its tracks. Of course, that might mean changing some people's plans, and when the anomaly does not happen (because it was prevented), they might wonder why those changes were necessary. That is why these systems need to explain themselves—showing us, visually or descriptively, what they thought was about to go wrong. In this paper, we explore the challenges and opportunities of building real-time video anomaly anticipation systems and share a vision for how these tools could make a real-world impact.

David C. Anastasiu
Santa Clara University, U.S.
danastasiu@scu.edu

**CP8**

**Blue Sky: Expert-in-the-Loop Representation Learning Framework for Audio Anti-Spoofing: Multimodal, Multilingual, Multi-Speaker, Multi-Attack (4M)**

Audio spoofing has surged with the rise of generative artificial intelligence, posing a serious threat to online communication. Recent studies have shown promising avenues in detecting spoofed audio specifically those that use human expert knowledge in representation learning, but more work is needed to evaluate performance across various realistic scenarios that tend to pose challenges in spoofed audio detection. In this paper, we introduce a comprehensive framework for expert-in-the-loop representation learning for audio anti-spoofing that is robust enough to address four specific challenging scenarios. Multimodal, Multilingual, Multi-speaker, and Multi-attack ($4M$). Preliminary results demonstrate the framework's potential effectiveness in audio anti-spoofing.

Sara Khanjani
University Of Maryland Baltimore County
zkhanja1@umbc.edu

Vandana Janeja
Information Systems Department
University Of Maryland Baltimore County
vjaneja@umbc.edu

First Name Last Name

Sanjay Purushotham
University of Maryland, Baltimore County
psanjay@umbc.edu

**CP8**

**Blue Sky: Reducing Performance Gap Between Commercial and Open-Source Llms**

The performance gap between commercial and open-source large language models (LLMs) poses a critical challenge in achieving equitable access to advanced AI technologies, particularly for underfunded institutions. As commercial entities like OpenAI invest substantial resources into proprietary models, open-source alternatives struggle with limitations such as a lack of access to high-quality datasets and feedback, restricting opportunities for research and innovation. We propose strategies needed to democratize AI technology, emphasizing collaboration and knowledge sharing within the community. By fostering a more inclusive environment, we can develop versatile, user-focused models that empower diverse stakeholders and expand the horizons of AI research across various sectors. This paper calls for a holistic approach to bridging this gap through behavioral modeling, leveraging techniques such as reinforcement learning and scenario-based testing to enhance the capabilities of open-source LLMs.

Adithya Kulkarni
Iowa State University
aditkulk@vt.edu

Mohna Chakraborty
University of Michigan
mohna@umich.edu

**CP8**

**Evaluating Time Series Models with Knowledge Discovery**

Time series data is one of the most ubiquitous data modality existing in diverse critical domains such as healthcare, seismology, manufacturing, and energy. In recent years, there are increasing interest in the data mining community to develop time series deep learning model to pursue better performance. The models performance are often evaluated by certain evaluation metrics such as RMSE, Accuracy, and F1-score. Yet time series data are often hard to interpret and is collected with unknown environment factors, sensor configuration, latent physic mechanisms, and non-stationary evolving behavior. As a result, a model that is better on standard metric-based evaluation may not always perform better in real-world tasks. In this blue-sky paper, we aim to explore the challenge that exists in the metric-based evaluation framework for time series data mining and propose a potential blue-sky idea — developing a knowledge-discovery-based evaluation framework, which aims to effectively utilize domain-expertise knowledge to evaluate models. We demonstrate that an evidence-seeking explanation can potentially have stronger persuasive power than metric-based evaluation and obtain better generalization ability for time series data mining tasks.

Li Zhang
University of Texas Rio Grand Valley, U.S.
li.zhang@utrgv.edu

**CP9**

**Optimizing Transit Network Expansion with Gated Attentive Graph Reinforcement Learning**

Transit network expansion is a challenging urban planning task that requires sophisticated decision-making to meet growing travel demands and improve urban mobility. This paper proposes the Gated Attentive Graph Reinforcement Learning (GAGRL) framework to optimize transit network expansion. GAGRL models the urban environment as a heterogeneous graph, where nodes represent urban regions and multiple edge types capture diverse relationships. By formulating the network expansion task as a Markov decision process within an expanding partial subgraph, GAGRL leverages a specially designed graph neural network encoder with gated message passing to effectively model urban features such as spatial connectivity and mobility flows. An attentive policy network ensures its efficient exploration of the solution space while adhering to budget constraints and transportation engineering requirements. Extensive experiments on real-world transit networks demonstrate that GAGRL outperforms state-of-the-art methods, achieving an average 25.95% improvement in total served origin-destination demand across various budget scenarios in the Beijing metro network. The superior performance of GAGRL, particularly in larger and more complex urban environments, highlights its potential as a powerful tool for automated transit network design.

Fanglan Chen
Virginia Tech
fanglanc@vt.edu

Dongjie Wang
University of Kansas
wangdongjie@ku.edu

Jianfeng He, Shuo Lei
Virginia Tech
jianfenghe@vt.edu, slei@vt.edu

Chang-Tien Lu
Computer Science, Virginia Tech
ctlu@vt.edu

**CP9**

**Feature Deviation Embedding Improves Graph Structure Learning for Spatial Interpolation**

The graph structures generated by natural or simple heuristics often fail to represent the spatial correlations influenced by complex factors. Therefore, introducing graph structure learning (GSL) can enhance the graph neural network-based spatial interpolation models. However, the input features of the GSL module are systematically unbalanced in spatial interpolation tasks. For example, many natural variables follow Gaussian- or gamma-distribution, and sensor spatial distributions are generally uneven. Thus, the GSL module must systematically incorporate corresponding solutions to avoid negatively impacting its generalization ability and degrading model performance. Our proposed model utilizes two encoders to embed feature deviations of node readings and

centroid distance from preset distributions, respectively. Notably, these encoders are jointly optimized with other model components, and their generalization ability is improved through an adaptively adjustable information bottleneck. Consequently, the GSL module can offer a more robust graph structure by explicitly perceiving feature deviations in the input. Experimental results demonstrate that our model outperforms existing state-of-the-art baselines across multiple real-world datasets with diverse characteristics.

Chaofan Li, Till Riedel, Michael Beigl
Karlsruhe Institute of Technology
chaofan.li@kit.edu, riedel@teco.edu, michael@teco.edu

**CP9**

**REGE: A Method for Incorporating Uncertainty in Graph Embeddings**

Machine learning models for graphs in real-world applications are prone to two primary types of uncertainty: (1) those that arise from incomplete and noisy data and (2) those that arise from uncertainty of the model in its output. These sources of uncertainty are not mutually exclusive. Additionally, models are susceptible to targeted adversarial attacks, which exacerbate both of these uncertainties. In this work, we introduce Radius Enhanced Graph Embeddings (REGE), an approach that measures and incorporates uncertainty in data to produce graph embeddings with radius values that represent the uncertainty of the model's output. REGE employs curriculum learning to incorporate data uncertainty and conformal learning to address the uncertainty in the model's output. In our experiments, we show that REGE's graph embeddings perform better under adversarial attacks by an average of 1.5% (accuracy) against state-of-the-art methods.

Zohair Shafi, Germans Savcisens, Tina Eliassi-Rad
Northeastern University
shafi.z@northeastern.edu, g.savcisens@northeastern.edu, t.eliassirad@northeastern.edu

**CP9**

**Equipping Graph Autoencoders: Revisiting Masking Strategies from a Robustness Perspective**

Masked Graph Autoencoders (MGAEs), represented by GraphMAE and GraphMAE2, which utilize masked feature (or structure) reconstruction strategies, have demonstrated the potential to surpass contrastive learning. However, current masked reconstruction strategies primarily rely on random strategies, only prove effective on reliable graph data. Therefore, these popular methods face immediate robustness deficiencies issues. **Firstly**, when the graph is unreliable or under adversarial attacks, the selection of nodes for masked reconstruction has a significant impact on downstream tasks. **Secondly**, the reconstructed features contains redundant components. In this paper, to overcome the non-robustness caused by randomness, we provide a theoretical analysis and evaluation of the robustness of state-of-the-art MGAEs. Additionally, we design two lightweight plug-and-play tools: **Box-Based Weighted Reliability Ranking Masking Strategy** and **Decoupled Feature Reconstruction**. Without incurring additional time overhead, these tools provide a defense armor against adversarial attacks for MGAEs, significantly boosting the robustness performance of downstream tasks. Extensive experiments on real-world graphs attacked by various attacks demonstrate our designs have a considerable robust expressive ability. Especially on datasets with large perturbations, the defense performance could even be improved by up to 20%.

Shuhan Song, Ping Li, Ming Dun, Yuan Zhang, Huawei Cao, Xiaochun Ye
State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
songshuhan19s@ict.ac.cn, liping20b@ict.ac.cn, dunming@ict.ac.cn, zhangyuan-ams@ict.ac.cn, caohuawei@ict.ac.cn, yexiaochun@ict.ac.cn

**CP10**

**Protecting Privacy Against Membership Inference Attack with Llm Fine-Tuning Through Flatness**

The privacy concerns associated with the use of Large Language Models (LLMs) have grown dramatically with the development of pioneer LLMs such as ChatGPT. Differential Privacy (DP) techniques that utilize DP-SGD are explored in existing work to mitigate their privacy risks at the cost of generalization degradation. Our paper reveals that the flatness of DP-SGD trained models' loss landscape plays an essential role in the trade-off between their privacy and generalization. We further propose a holistic framework Privacy-Flat to enforce appropriate weight flatness, which substantially improves model generalization with promising privacy protection. It innovates from three coarse-to-grained levels: Perturbation-aware min-max optimization within a layer, flatness-guided sparse prefix-tuning across layers, and weight knowledge distillation between private & non-private weights copies. We empirically demonstrate that our framework Privacy-Flat outperforms vanilla private training baseline while protecting privacy from membership inference attacks (MIA). Comprehensive experiments of both black-box and white-box scenarios are conducted to demonstrate the effectiveness of our proposal in enhancing generalization.

Tiejin Chen, Longchao Da
Arizona State University
tchen169@asu.edu, longchao@asu.edu

Huixue Zhou
University of Minnesota Twin Cities
zhou1742@umn.edu

Pingzhi Li
University of North Carolina at Chapel Hill
pingzhi@cs.unc.edu

Kaixiong Zhou
North Carolina State University
zhou22@ncsu.edu

Tianlong Chen
The University of North Carolina at Chapel Hill
tianlong@cs.unc.edu

Hua Wei
Arizona State University
hua.wei@asu.edu

**CP10**

**CoT-Decoding: Complex Reasoning Via Chain-of-Thought Decoding**

Complex reasoning is a challenging task, and Chain-of-

Thought (CoT) reasoning has become a research hotspot. However, existing methods face two key limitations: lack of fine-grained analysis at the reasoning step level and reliance on greedy decoding, often leading to local optima and limiting reasoning performance. To address these issues, we propose the CoT-Decoding framework, consisting of two modules: faithful reasoning step generation and reasoning chain decoding. In the first model, we designed a relevance determination module based on in-context learning and small Long Short-Term Memory (sLSTM). Subsequently, a large language model (LLM) was used to decompose complex questions into sub-questions, and their relevance was determined. Relevant evidence was then retrieved through cross-referencing. The second module integrates in-context learning with a Siamese network for logic step scoring and uses contrastive search strategies to ensure logical coherence. By jointly considering model probability, contrastive scores, and logical consistency, CoT-Decoding selects the most reliable reasoning steps to construct the final reasoning chain. CoT-Decoding outperformed baseline models across five benchmarks: HOTPOTQA (62.1), 2WIKIMQA (66.7), MUSIQUE (31.0), FERMI (41.7), and STRATEGYQA (82.2) in F1 scores. Even with lightweight LLMs, it demonstrated strong performance, highlighting its potential for applications in resource-constrained environments.

Li Li, Guoquan Lu
College of Computer Information Science
Southwest University
lily@swu.edu.cn, luguoquan@email.swu.edu.cn

Lin Peng
College of Big Data
Yunnan Agricultural University
2007013@ynau.edu.cn

**CP10**

**Language Models Are Explorers for Join Discovery on Data Lakes**

Yaohua Wang
tbd
xiachen.wyh@taobao.com

Bolin Ding
Alibaba Group
bolin.ding@alibaba-inc.com

**CP10**

**Comal: Collaborative Multi-Agent Large Language Models for Mixed-Autonomy Traffic**

The integration of autonomous vehicles into urban traffic has great potential to improve efficiency by reducing congestion and optimizing traffic flow systematically. In this paper, we introduce **CoMAL** (**Co**llaborative **M**ulti-**A**gent **L**LMs), a framework designed to address the *mixed-autonomy traffic* problem by collaboration among autonomous vehicles to optimize traffic flow. CoMAL is built upon large language models and operates in an interactive traffic simulation environment. Specifically, It utilizes a Perception Module to observe surrounding agents and a Memory Module to store strategies for each agent. The overall workflow includes a Collaboration Module that encourages autonomous vehicles to discuss the effective strategy and allocate roles, a reasoning engine to determine optimal behaviors based on assigned roles, and an Execution Module that controls vehicle actions using a hybrid approach combining rule-based models. Experimental results demonstrate that CoMAL achieves superior performance on the Flow benchmark. Additionally, we evaluate the impact of different language models and compare our framework with reinforcement learning approaches. It highlights the strong cooperative capability of LLM agents and presents a promising solution to the mixed-autonomy traffic challenge. The code is available at https://github.com/Hyan-Yao/CoMAL

Huaiyuan Yao, Longchao Da, Vishnu Nandam, Justin Turnau
Arizona State University
huaiyuanyao@gmail.com, longchao@asu.edu, vnandam@asu.edu, jturnau@asu.edu

Zhiwei Liu
Salesforce AI Research
zhiweiliu@salesforce.com

Linsey Pang
Salesforce
panglinsey@gmail.com

Hua Wei
Arizona State University
hua.wei@asu.edu

**CP11**

**Data Mining the Functional Architecture of the Brain's Circuitry**

The brain is a complex organ consisting of a myriad of subsystems that flexibly interact and adapt to enable perception, cognition, and behavior. Understanding the multiscale nature of the brain, i.e., how circuit- and molecular-level interactions build up to create brain function, holds incredible potential for developing interventions for neurodegenerative and psychiatric diseases, as well as understanding our very nature. Historically technological limitations have forced systems neuroscience to study limited biological quantities in localized, small neural populations in single brain areas during constrained behaviors. New developments in neural recording technology and behavioral monitoring now provide the data needed to break free of local neuroscience to global neuroscience: i.e., understanding how the brain's many subsystem interact, adapt, and change across the many behaviors animals and humans must perform. Specifically, while we have much knowledge of the anatomical architecture of the brain (the hardware), we finally are approaching the data needed to find the functional architecture and discover the fundamental properties of the software that runs on the hardware. We must take this opportunity to bridge between the vast amounts of data to discover this functional architecture which will face numerous challenges from low-level data alignment up to high level questions of interpretable models of brain activity.

Adam Charles
Johns Hopkins University
adamsc@jhu.edu

**CP11**

**Better AI For Understanding Life on Earth: Predict First, Design Later**

Generative AI is generating much enthusiasm on potentially advancing biological design in computational biology. In this paper we take a somewhat contrarian view, arguing that a broader and deeper understanding of existing biological sequences is essential before undertaking the design of novel ones. We draw attention, for instance, to current protein function prediction methods which currently face significant limitations due to incomplete data and inherent challenges in defining and measuring function. We propose a "blue sky" vision centered on both comprehensive and precise annotation of existing protein and DNA sequences, aiming to develop a more complete and precise understanding of biological function. By contrasting recent studies that leverage generative AI for biological design with the pressing need for enhanced data annotation, we underscore the importance of prioritizing robust predictive models over premature generative efforts. We advocate for a strategic shift toward thorough sequence annotation and predictive understanding, laying a solid foundation for future advances in biological design.

Amarda Shehu
George Mason University
ashehu@gmu.edu

Yana Bromberg
Emory University
yana.bromberg@emory.edu

**CP11**

**What We Talk About When We Talk About AI for Science**

'AI for Science' has become a prominent yet controversial research frontier, eliciting both acclaim and criticism. While celebrated for its potential to revolutionize scientific discovery, concerns persist regarding the reliability, interpretability, and validation of AI-generated knowledge. This paper explores core challenges, including the opacity of AI insights, the difficulty of evaluating unverifiable outputs, and the inadequacy of traditional frameworks like the Turing Test. Although addressing these issues may seem like a distant goal, this paper proposes a Blue Sky Idea aimed at redefining AIs role in scientific exploration and paving the way for transformative progress.

Runlong Yu
University of Pittsburgh
ruy59@pitt.edu

Yiqun Xie
University of Maryland
xie@umd.edu

Xiaowei Jia
U. of Pittsburgh
xiaowei@pitt.edu

**CP11**

**Optimizing External and Internal Knowledge of Foundation Models for Scientific Discovery**

In the emerging landscape of AI-driven scientific discovery, foundation models hold significant promise for enhancing research ideation and overall scientific advancement. This paper explores a future where foundation models should be able to effectively utilize both external and internal knowledge sources to maximize their role in scientific discovery. The core challenge lies in optimizing two knowledge types: external knowledge, drawn from diverse data sources, and internal knowledge, the parametric understanding acquired during training. We propose a dual-framework solution for this optimization, including X-augmented generation and in-context X learning. X-augmented generation approaches, such as retrieval-augmented generation, knowledge graph-augmented generation, and third-party tool integration, enhance external knowledge processing. In-context X learning methods, including in-context adversarial learning and in-context reinforcement learning, improve

models' internal knowledge adaptation and utility for scientific tasks. We aim to inspire the research community by proposing a bold pathway toward leveraging foundation models as active participants in scientific discovery, tackling the inherent complexity of optimizing vast, multimodal knowledge sources. By addressing this challenge, we envision a future where foundation models catalyze breakthroughs across disciplines, ultimately leading to a more dynamic, collaborative, and insight-driven scientific process.

Aidong Zhang
University of Virginia
aidong@virginia.edu

**CP12**

**AnchorDrug: A System for Drug-Induced Gene Expression Prediction in New Contexts Through Active Learning**

Large pre-trained models have been extensively explored for numerous biomedical tasks. However, the diversity and complexity of biological systems often make zero-shot learning in a new context challenging. In many instances, the budget allows for the acquisition of a small number of labeled data through experiments for few-shot learning. Yet, the methodology for selecting the optimal set of samples for these experiments remains underexplored. In this work, we present an application focused on drug-induced gene expression prediction to demonstrate a data-driven approach for sample selection. We developed a system named AnchorDrug, which predicts drug-induced gene expression changes in new cell lines after fine-tuning with experimental data from a limited number of drugs. Initially, we built a pre-trained model with a large dataset of drug-induced gene expressions. We then adopted active learning to identify an optimal set of drugs for experiments, aiming to ensure that the experimental data used for subsequent fine-tuning would maximize model performance. Several acquisition functions are customized and incorporated into our pipeline. Compared with knowledge-based drug selection, our customized active learning methods proved more effective in selecting anchor drugs. We further provided insights into the reasons behind its superior performance. Our system is designed to mimic real-world scenarios, enabling its easy application to real biomedical research projects.

Ruoqiao Chen
michigan state university
chenruo4@msu.edu

Han Meng
college of william and mary
miss.menghan@gmail.com

Bin Chen
michigan state university
chenbi12@msu.edu

Jiayu Zhou
university of michigan
jiayuz@umich.edu

**CP12**

**Domain Knowledge Augmented Contrastive Learning on Dynamic Hypergraphs for Improved Health Risk Prediction**

Accurate health risk prediction is crucial for making informed clinical decisions and assessing the appropriate allocation of medical resources. While recent deep learning approaches have shown promise in risk prediction, they focus on modeling the sequential information in electronic health records (EHRs) and fail to leverage the rich mobility interactions among health entities, resulting in unsatisfactory performance in downstream risk prediction tasks, especially Clostridioides difficile Infection (CDI) incidence prediction that primarily spreads through mobility interactions. To address this issue, our work leverages hypergraphs to explicitly model mobility interactions to improve predictive performance. Unlike graphs that are limited to modeling pairwise relationships, hypergraphs can effectively characterize complex, high-order relationships between patients. Moreover, we introduce a new contrastive learning strategy that exploits domain knowledge to generate meaningful homologous and heterologous augmentations. This strategy boosts the power of contrastive learning by generating feature representations that are both robust and aligned with domain knowledge. Experiments on 2 real-world datasets show the advantage of our approach in time-varying risk prediction tasks. Our framework obtains gains in performance up to 29.49% for UIHC, 30.64% for MIMIC-IV for MICU transfer prediction, 13.17% for UIHC, and 4.45% for MIMIC-IV for CDI incidence prediction.

Akash Choudhuri
University of Iowa
Department of Computer Science
akash-choudhuri@uiowa.edu

Hieu Vu
Department of Computer Science,
University of Iowa
hieu-vu@uiowa.edu

Kishlay Jha
Department of Electrical and Computer Engineering,
University of Iowa
kishlay-jha@uiowa.edu

Bijaya Adhikari
Department of Computer Science,
University of Iowa
bijaya-adhikari@uiowa.edu

**CP12**

**Accurately Estimating Unreported Infections using Information Theory**

One of the most significant challenges in combating against the spread of infectious diseases was the difficulty in estimating the true magnitude of infections. Unreported infections could drive up disease spread, making it very hard to accurately estimate the infectivity of the pathogen, therewith hampering our ability to react effectively. Despite the use of surveillance-based methods such as serological studies, identifying the true magnitude is still challenging. This paper proposes an information theoretic approach for accurately estimating the number of total infections. Our approach is built on top of Ordinary Differential Equations (ODE) based models, which are commonly used in epidemiology and for estimating such infections. We show how we can help such models to better compute the number of total infections and identify the parametrization by which we

need the fewest bits to describe the observed dynamics of reported infections. Our experiments on COVID-19 spread show that our approach leads to not only substantially better estimates of the number of total infections but also better forecasts of infections than standard model calibration based methods. We additionally show how our learned parametrization helps in modeling more accurate what-if scenarios with non-pharmaceutical interventions. Our approach provides a general method for improving epidemic modeling which is applicable broadly.

Jiaming Cui
Georgia Institute of technology
jiamingcui@vt.edu

Bijaya Adhikari
Department of Computer Science,
University of Iowa
bijaya-adhikari@uiowa.edu

Arash Haddadan
Amazon
ahaddada@amazon.com

A S M Ahsan-Ul Haque
University of Virginia
ah3wj@virginia.edu

Jilles Vreeken
CISPA Helmholtz Center for Information Security
vreeken@cispa.de

Anil Vullikanti
University of Virginia
vsakumar@virginia.edu

B. Aditya Prakash
College of Computing, Georgia Institute of Technology
badityap@cc.gatech.edu

**CP12**

**Spatially-Delineated Domain-Adapted Ai Classification: An Application for Oncology Data**

Given multi-type point maps from different place-types (e.g., tumor regions), our objective is to develop a classifier trained on the source place-type to accurately distinguish between two classes of the target place-type based on their point arrangements. This problem is societally important for many applications, such as generating clinical hypotheses for designing new immunotherapies for cancer treatment. The challenge lies in the spatial variability, the inherent heterogeneity and variation observed in spatial properties or arrangements across different locations (i.e., place-types). Previous techniques focus on self-supervised tasks to learn domain-invariant features and mitigate domain differences; however, they often neglect the underlying spatial arrangements among data points, leading to significant discrepancies across different place-types. We explore a novel multi-task self-learning framework that targets spatial arrangements, such as spatial mix-up masking and spatial contrastive predictive coding, for spatially-delineated domain-adapted AI classification. Experimental results on real-world datasets (e.g., oncology data) show that the proposed framework provides higher prediction accuracy than baseline methods.

Majid Farhadloo
University of Minnesota, Twin Cities
farha043@umn.edu

Arun Sharma
University of Minnesota, Twin Cities
Department of Computer Science and Engineering
sharm485@umn.edu

Alexey Leontovich, Svetomir Markovic
Mayo Clinic
leontovich.alexey@mayo.edu,
markovic.svetomir@mayo.edu

Shashi Shekhar
University of Minnesota
shekhar@umn.edu

**CP12**

**Mexa-Ctp: Mode Experts Cross-Attention for Clinical Trial Outcome Prediction**

Clinical trials are the gold standard for assessing the effectiveness and safety of drugs for treating diseases. Given the vast design space of drug molecules, elevated financial cost, and multi-year timeline of these trials, research on clinical trial outcome prediction has gained immense traction. Accurate predictions must leverage data of diverse modes such as drug molecules, target diseases, and eligibility criteria to infer successes and failures. Previous Deep Learning approaches for this task, such as HINT, often require wet lab data from synthesized molecules and/or rely on prior knowledge to encode interactions as part of the model architecture. To address these limitations, we propose a light-weight attention-based model, MEXA-CTP, to integrate readily-available multi-modal data and generate effective representations via specialized modules dubbed "mode experts", while avoiding human biases in model design. We optimize MEXA-CTP with the Cauchy loss to capture relevant interactions across modes. Our experiments on the Trial Outcome Prediction (TOP) benchmark demonstrate that MEXA-CTP improves upon existing approaches by, respectively, up to 11.3% in F1 score, 12.2% in PR-AUC, and 2.5% in ROC-AUC, compared to HINT. Ablation studies are provided to quantify the effectiveness of each component in our proposed method.

Yiqing Zhang, Xiaozhong Liu
Worcester Polytechnic Institute
yzhang37@wpi.edu, xliu14@wpi.edu

Fabricio Murai
Worcester Polytechnic Institute, Worcester, MA
fmurai@wpi.edu

**CP13**

**Metrics for Inter-Dataset Similarity with Example Applications in Synthetic Data and Feature Selection Evaluation**

Measuring inter-dataset similarity is an important task in machine learning and data mining with various use cases and applications. Existing methods for measuring inter-dataset similarity are computationally expensive, limited, or sensitive to different entities and non-trivial choices for parameters. They also lack a holistic perspective on the entire dataset. In this paper, we propose two novel metrics for measuring inter-dataset similarity. We discuss the mathematical foundation and the theoretical basis of our proposed metrics. We demonstrate the effectiveness of the

proposed metrics by investigating two applications in the evaluation of synthetic data and in the evaluation of feature selection methods. The theoretical and empirical studies conducted in this paper illustrate the effectiveness of the proposed metrics.

Muhammad Rajabinasab
Syddansk Universitet / University of Southern Denmark
Syddansk Universitet / University of Southern Denmark
rajabinasab@imada.sdu.dk

Anton Lautrup
Syddansk Universitet / University of Southern Denmark
lautrup@imada.sdu.dk

Arthur Zimek
University of Southern Denmark,
zimek@imada.sdu.dk

**CP13**

**Parameter-Efficient Interventions for Enhanced Model Merging**

Model merging integrates knowledge from task-specific models into a single multi-task model, eliminating the need for joint training on all task data. Existing approaches encounter challenges related to representation bias, which can impact task performance. This presentation will discuss IntervMerge, a method that applies task-specific interventions to mitigate representation bias. Additionally, the concept of mini-interventions will be introduced, modifying only a portion of the representation to reduce the number of additional parameters while maintaining performance.

Marcin Osial
Jagiellonian University
IDEAS NCBR
marcin.osial@doctoral.uj.edu.pl

Daniel Marczak
Warsaw University of Technology
IDEAS NCBR
daniel.marczak@ideas-ncbr.pl

Bartosz Zielinski
Jagiellonian University
IDEAS NCBR
bartosz.zielinski@uj.edu.pl

**CP13**

**VisTabNet: Adapting Vision Transformers for Tabular Data**

Although deep learning models have had great success in natural language processing and computer vision, we do not observe comparable improvements in the case of tabular data, which is still the most common data type used in biological, industrial and financial applications. In particular, it is challenging to transfer large-scale pre-trained models to downstream tasks defined on small tabular datasets. To address this, we propose VisTabNet – a cross-modal transfer learning method, which allows for adapting Vision Transformer (ViT) with pre-trained weights to process tabular data. By projecting tabular inputs to patch embeddings acceptable by ViT, we can directly apply a pre-trained Transformer Encoder to tabular inputs. This approach eliminates the conceptual cost of designing a suitable architecture for processing tabular data, while reducing the computational cost of training the model from scratch. Experimental results on multiple small tabular datasets (less than 1k samples) demonstrate VisTabNet's superiority, outperforming both traditional ensemble methods and recent deep learning models. The proposed method goes beyond conventional transfer learning practice and shows that pre-trained image models can be transferred to solve tabular problems, extending the boundaries of transfer learning. We share our example implementation as a GitHub repository available at `https://github.com/wwydmanski/VisTabNet`.

Marek Smieja, Witold Wydmanski, Ulvi Movsum-Zada, Jacek Tabor
Jagiellonian University
marek.smieja@uj.edu.pl, wwydmanski@gmail.com, ulvi.movsumzade@gmail.com, jacek.tabor@uj.edu.pl

**CP13**

**Agent Reinforcement Learning Via Coalition Labeling and Structural Entropy**

Multi-agent cooperation is essential for tasks that require collaboration to achieve optimal performance or cannot be completed by individual agents alone. These tasks often necessitate a divide-and-conquer strategy, where subgoals are allocated to individual agents or groups. By integrating coalition formation concepts from cooperative game theory, we demonstrate the implicit learning of coalition formation and task assignments, resulting in emergent cooperative behavior. We propose a novel **CO**a**L**ition **LAB**eling technique for Multi-Agent Reinforcement Learning (**COLLAB**-MARL) to encourage coalition formation and introduce a structural entropy measure to detect the emergence of coalitions and cooperative behavior. Compared to classical MARL methods, COLLAB-MARL is more effective, explainable, and easier to implement. Experiments on state-of-the-art cooperative MARL benchmarks show that our method's mean return outperforms the strongest baselines by 8.4% on average. Additionally, visualization and structural entropy analysis reveal that COLLAB-MARL effectively learns meaningful cooperative behavior. The source code is available at `https://github.com/SELGroup/collab`.

Dingli Su
Beihang University
sudingli@buaa.edu.cn

Hao Peng
Beihang University, China
penghao@buaa.edu.cn

Guangjie Zeng
Beihang University
zengguangjie@buaa.edu.cn

Pu Li
Kunming University of Science and Technology
lip@stu.kust.edu.cn

Angsheng Li, Yicheng Pan
Beihang University
angsheng@buaa.edu.cn, yichengp@buaa.edu.cn

**CP14**

**Avatar: Adversarial Autoencoders with Autore-**

**gressive Refinement for Time Series Generation**

Data augmentation can significantly enhance the performance of machine learning tasks by addressing data scarcity and improving generalization. However, generating time series data presents unique challenges. A model must not only learn a probability distribution that reflects the real data distribution but also capture the conditional distribution at each time step to preserve the inherent temporal dependencies. To address these challenges, we introduce AVATAR, a framework that combines Adversarial Autoencoders (AAE) with Autoregressive Learning to achieve both objectives. Specifically, our technique integrates the autoencoder with a supervisor and introduces a novel supervised loss to assist the decoder in learning the temporal dynamics of time series data. Additionally, we propose another innovative loss function, termed distribution loss, to guide the encoder in more efficiently aligning the aggregated posterior of the autoencoder's latent representation with a prior Gaussian distribution. Furthermore, our framework employs a joint training mechanism to simultaneously train all networks using a combined loss, thereby fulfilling the dual objectives of time series generation. We evaluate our technique across a variety of time series datasets with diverse characteristics. Our experiments demonstrate significant improvements in both the quality and practical utility of the generated data, as assessed by various qualitative and quantitative metrics.

Mohammadreza Eskandarinasab, Shah Muhammad Hamdi, Soukaina Filali Boubrahimi
Utah State University
reza.eskandarinasab@usu.edu, s.hamdi@usu.edu, soukaina.boubrahimi@usu.edu

**CP14**

**End-To-End Self-Tuning Self-Supervised Time Series Anomaly Detection**

Time series anomaly detection (TSAD) finds many applications such as monitoring environmental sensors, industry KPIs, patient biomarkers, etc. A two-fold challenge for TSAD is a versatile and unsupervised model that can detect various *different types* of time series anomalies (spikes, discontinuities, trend shifts, etc.) *without any labeled data*. Modern neural networks have outstanding ability in modeling complex time series. Self-supervised models in particular tackle unsupervised TSAD by transforming the input via various augmentations to create pseudo anomalies for training. However, their performance is sensitive to the choice of augmentation, which is hard to choose in practice, while there exists no effort in the literature on data augmentation tuning for TSAD without labels. Our work aims to fill this gap. We introduce TSAP for *TSA "on autoPilot'*, which can *(self-)tune* augmentation hyperparameters end-to-end. It stands on two key components: a differentiable augmentation architecture and an unsupervised validation loss to effectively assess the alignment between augmentation type and anomaly type. Case studies show TSAP's ability to effectively select the (discrete) augmentation type and associated (continuous) hyperparameters. In turn, it outperforms established baselines, including SOTA self-supervised models, on diverse TSAD tasks exhibiting different anomaly types.

Meng-Chieh Lee
Carnegie Mellon University
mengchil@andrew.cmu.edu

Boje Deforce
Carnegie Mellon University
KU Leuven
boje.deforce@kuleuven.be

Bart Baesens, Estefanía Serral Asensio
KU Leuven
bart.baesens@kuleuven.be,
estefania.serralasensio@kuleuven.be

Jaemin Yoo
Korea Advanced Institute of Science & Technology
jaemin@kaist.ac.kr

Leman Akoglu
Carnegie Mellon University
lakoglu@andrew.cmu.edu

**CP14**

**Autostdiff: Autoregressive Spatio-Temporal Denoising Diffusion Model for Asynchronous Trajectory Generation**

Large-scale trajectory data is crucial for applications like human mobility prediction and pandemic intervention. However, concerns over data privacy have limited access to real-world datasets. Advanced generative models offer a promising alternative for creating synthetic yet realistic trajectory data, but most existing methods focus on synchronous trajectories with fixed time intervals, which fail to capture the complexities of asynchronous trajectories, such as Point of Interest (POI) check-ins with uncertain intervals and varying lengths. To address this gap, we propose AutoSTDiff, a novel **Auto**regressive **S**patio-**T**emporal denoising **Diff**usion model for asynchronous trajectory generation. AutoSTDiff includes two key components: (i) a hybrid embedding module that captures comprehensive spatio-temporal patterns considering human behavior and varying trajectory lengths, and (ii) a spatio-temporal diffusion model with a spatial status conversion module and a conditional spatio-temporal generation module for autoregressive trajectory generation. Extensive experiments on two public trajectory datasets show that AutoSTDiff outperforms state-of-the-art models, e.g., an increase of 51.2% and 43.8% on the length and G-rank metrics.

Rongchao Xu
Florida State University
rx21a@fsu.edu

Zhiqing Hong
Rutgers University
zhiqing.hong@rutgers.edu

Guang Wang
Florida State University
guang@cs.fsu.edu

**CP14**

**Fine-Grained Spatio-Temporal Event Prediction with Self-Adaptive Anchor Graph**

Event prediction tasks often handle spatio-temporal data distributed in a large spatial area. Different regions in the area exhibit different characteristics while having latent correlations. This spatial heterogeneity and correlations greatly affect the spatio-temporal distributions of event occurrences, which has not been addressed by state-of-the-art models. Learning spatial dependencies of events in a con-

tinuous space is challenging due to its fine granularity and a lack of prior knowledge. In this work, we propose a novel Graph Spatio-Temporal Point Process (GSTPP) model for fine-grained event prediction. It adopts an encoder-decoder architecture that jointly models the state dynamics of spatially localized regions using neural Ordinary Differential Equations (ODEs). The state evolution is built on the foundation of a novel Self-Adaptive Anchor Graph (SAAG) that captures spatial dependencies. By adaptively localizing the anchor nodes in the space and jointly constructing the correlation edges between them, the SAAG enhances the models ability of learning complex spatial event patterns. The proposed GSTPP model greatly improves the accuracy of fine-grained event prediction. Extensive experimental results show that our method greatly improves the prediction accuracy over existing spatio-temporal event prediction approaches.

Wang-Tao Zhou, Zhao Kang
University of Electronic Science and Technology of China
wtzhou@std.uestc.edu.cn, zkang@uestc.edu.cn

Sicong Liu, Lizong Zhang
School of Computer Science and Engineering
University of Electronic Science and Technology of China
202321081425@std.uestc.edu.cn, l.zhang@uestc.edu.cn

Ling Tian
University of Electronic Science and Technology of China
lingtian@uestc.edu.cn

**CP15**

**Staleness-Alleviated Distributed Gnn Training Via Online Dynamic-Embedding Prediction**

Despite the recent success of Graph Neural Networks (GNNs), it remains challenging to train GNNs on large-scale graphs due to neighbor explosions. As a remedy, distributed computing becomes a promising solution by leveraging abundant computing resources. However, the node dependency of graph data increases the difficulty of achieving high concurrency in distributed GNN training, which suffers from the massive communication overhead. To address this, historical value approximation is deemed a promising class of distributed training techniques. It utilizes an offline memory to cache historical information as an affordable approximation of the exact value and achieves high concurrency. However, such benefits come at the cost of involving dated training information, leading to staleness, imprecision, and convergence issues. To overcome these challenges, this paper proposes SAT, a novel and scalable distributed GNN training framework that reduces the embedding staleness adaptively. The key idea of SAT is to model the GNN's embedding evolution as a temporal graph and build a model upon it to predict future embedding. We propose an online algorithm to train the embedding predictor and the distributed GNN alternatively and further provide a convergence analysis. Empirically, we demonstrate that SAT can effectively reduce embedding staleness and thus achieve better performance and convergence speed on multiple large-scale graph datasets.

Guangji Bai, Ziyang Yu
Emory University
guangji.bai@emory.edu, zyu31@emory.edu

Zheng Chai, Yue Cheng
University of Virginia
dub6yh@virginia.edu, mrz7dp@virginia.edu

Liang Zhao
Emory University
liang.zhao@emory.edu

**CP15**

**FedGrAINS: Personalized SubGraph Federated Learning with AdaptIve Neighbor Sampling**

Graphs are crucial for modeling relational and biological data. As datasets grow larger in real-world scenarios, the risk of exposing sensitive information increases, making privacy-preserving training methods like federated learning (FL) essential to ensure data security and compliance with privacy regulations. Recently proposed personalized subgraph FL methods have become the de-facto standard for training personalized Graph Neural Networks (GNNs) in a federated manner while dealing with the missing links across clients' subgraphs due to privacy restrictions. However, personalized subgraph FL faces significant challenges due to the heterogeneity in client subgraphs, such as degree distributions among the nodes, which complicate federated training of graph models. To address these challenges, we propose *FedGrAINS*, a novel data-adaptive and sampling-based regularization method for subgraph FL. FedGrAINS leverages Generative Flow Networks (GFlowNets) to evaluate node importance in relation to client tasks, dynamically adjusting the message-passing step in clients' GNNs. This adaptation reflects task-optimized sampling aligned with a trajectory balance objective. Experimental results demonstrate that the inclusion of *FedGrAINS* as a regularizer consistently improves the FL performance compared to baselines that do not leverage such regularization.

Emir Ceyani
University of Southern California
ceyani@usc.edu

Han Xie
Emory University
han.xie@emory.edu

Baturalp Buyukates
University of Birmingham
b.buyukates@bham.ac.uk

Carl Yang
Department of Computer Science
Emory University
j.carlyang@emory.edu

Salman Avestimehr
University of Southern California
avestime@usc.edu

**CP15**

**Unveiling the Impact of Local Homophily on GNN Fairness: In-Depth Analysis and New Benchmarks**

Graph Neural Networks (GNNs) struggle to generalize when graphs exhibit both homophily and heterophily. Specifically, GNNs tend to underperform for nodes with local homophily levels that differ significantly from the global homophily level. This issue poses a risk in user-centric applications where underrepresented homophily levels are present. Concurrently, fairness within GNNs has received substantial attention due to the amplification of biases via message passing. However, the connection between local homophily and fairness in GNNs remains underexplored.

In this work, we move beyond global homophily and explore how local homophily levels can lead to unfair predictions. We begin by formalizing the challenge of fair predictions for underrepresented homophily levels as an out-of-distribution (OOD) problem. We then conduct a theoretical analysis that demonstrates how local homophily levels can alter predictions for differing sensitive attributes. We additionally introduce three new GNN fairness benchmarks, as well as a novel semi-synthetic graph generator, to empirically study the problem. Across extensive analysis we find that two factors can promote unfairness: (a) OOD distance, and (b) heterophilous nodes situated in homophilous graphs. When these two conditions are met, fairness drops by up to 24% on real world data, and 30% in semi-synthetic data. Collectively, our insights unveil a previously overlooked source of unfairness rooted in the graph's homophily information.

Donald Loveland, Danai Koutra
University of Michigan, Ann Arbor
dlovelan@umich.edu, dkoutra@umich.edu

**CP15**

**Gaim: Attacking Graph Neural Networks Via Adversarial Influence Maximization**

Recent studies show that well-devised perturbations on graph structures or node features can mislead trained Graph Neural Network (GNN) models. However, these methods often overlook practical assumptions, over-rely on heuristics, or separate vital attack components. In response, we present GAIM, an integrated adversarial attack method conducted on a node feature basis while considering the strict black-box setting. Specifically, we define an adversarial influence function to assess the adversarial impact of node perturbations, thereby reframing the GNN attack problem into the adversarial influence maximization problem. In our approach, we unify the selection of the target node and the construction of feature perturbations into a single optimization problem, ensuring a unique and consistent feature perturbation for each target node. We use a surrogate model to transform this problem into a solvable linear programming task, streamlining the optimization process. Moreover, we extend our method to accommodate label-oriented attacks, broadening its applicability. Evaluations on five benchmark datasets across three popular models underscore the effectiveness of our method.

Xiaodong Yang
Visa Research
sheldon.uestc@gmail.com

**CP15**

**Evidence-Based Out-of-Distribution Detection on Multi-Label Graphs**

The Out-of-Distribution (OOD) problem in graph-structured data is becoming increasingly important in various areas of research and applications, including social network recommendation, protein function detection, etc. Furthermore, owing to the inherent multi-label properties of nodes, multi-label OOD detection remains more challenging than in multi-class scenarios. A lack of uncertainty modeling in multi-label classification methods prevents the separation of OOD nodes from in-distribution (ID) nodes. Existing uncertainty-based OOD detection methods on graphs are not applicable for multi-label scenarios because they are designed for multi-class settings. Therefore, node-level OOD detection on multi-label graphs becomes desirable but rarely touched. In this paper, we propose a novel Evidence-Based Out-of-Distribution Detection method on multi-label graphs. The evidence for multiple labels, which indicates the amount of support to suggest that a sample should be classified into a specific class, is predicted by Multi-Label Evidential Graph Neural Networks (ML-EGNNs). The joint belief is designed for multi-label opinions fusion by a comultiplication operator. Additionally, we introduce a Kernel-based Node Positive Evidence Estimation (KNPE) method to reduce errors in quantifying positive evidence. Experimental results prove both the effectiveness and efficiency of our model for multi-label OOD detection on 7 multi-label benchmarks.

Ruomeng Ding
Tianjin University
rmding@tju.edu.cn

Xujiang Zhao
NEC Laboratories America
xuzhao@nec-labs.com

Chen Zhao
Baylor University
chen_zhao@baylor.edu

Minglai Shao
Tianjin University
shaoml@tju.edu.cn

Zhengzhang Chen
NEC Laboratories America
zhengzhang.chen@gmail.com

Haifeng Chen
NEC Laboratories America, Inc.
haifeng@nec-labs.com

**CP16**

**An Interpretable Measure for Quantifying Predictive Dependence Between Continuous Random Variables**

A fundamental task in statistical learning is quantifying the joint dependence or association between two continuous random variables. We introduce a novel, fully nonparametric measure that assesses the degree of association between continuous variables $X$ and $Y$, capable of capturing a wide range of relationships, including non-functional ones. A key advantage of this measure is its interpretability: it quantifies the expected relative loss in predictive accuracy when the distribution of $X$ is ignored in predicting $Y$. This measure is bounded within the interval [0,1] and is equal to zero if and only if $X$ and $Y$ are independent. We evaluate the performance of our measure on more than 90,000 pairs of variables extracted from a large real dataset, as well as on multiple synthetic datasets, benchmarking it against leading alternatives. Our results demonstrate that the proposed measure provides valuable insights into underlying relationships, particularly in cases where

Renato M. Assuncao
ESRI Inc.
DCC/UFMG - Brazil
rassuncao@esri.com

Flávio Figueiredo, Francisco Tinoco JR., Leo S-FREIRE
Universidade Federal de Minas Gerais

flaviovdf@dcc.ufmg.br, francisconeves@dcc.ufmg.br, leomartins@dcc.ufmg.br

Fabio Silva
CEFET-MG, Brazil
fabiorochadasilva@cefetmg.br

**CP16**

**Meta-Learning of Class Knowledge in Zero-Shot Learning**

Zero-shot learning is a promising approach to generalizing a model to categories unseen during training, and various methods have been proposed. However, they assume that class knowledge, the semantic information of the classes, is available as prior knowledge and thus fails to support domains whose class knowledge is unavailable. We propose a meta-learning method that allows us to apply the zero-shot learning method even if class knowledge is unavailable. We assume multiple zero-shot learning tasks where some classes are missing for each task, but they appear in other tasks. Our method simultaneously estimates the appropriate class knowledge and classification model using a meta-learning approach that extracts task features. We can use the trained models to perform zero-shot classification on unseen tasks without class knowledge. In experiments on datasets where true class knowledge is available, class knowledge is unavailable, and class knowledge is provided but imprecise, we show that the proposed method performs better than existing zero-shot learning methods.

Yuta Nambu, Masahiro Kohjima
NTT Human Informatics Laboratories, NTT Corporation
yuta.nambu@ntt.com, masahiro.kohjima@ntt.com

Tomoharu Iwata
NTT Communication Science Laboratories, NTT Corporation
tomoharu.iwata@ntt.com

Ryuji Yamamoto
NTT Human Informatics Laboratories, NTT Corporation
ryuji.yamamoto@ntt.com

**CP16**

**Hybrid Bayesian Optimization with Direct**

We present a hybrid Bayesian optimization method that incorporates the well-known DIRECT algorithm for black-box optimization. The DIRECT algorithm originates from Lipschitz optimization for global optimization and performs searches using all possible values for Lipschitz constants by balancing global and local search. It has a good ability to locate promising regions of the feasible space; however, it can have slow asymptotic convergence. As a result, DIRECT can require extensive function evaluations to obtain a high-quality solution, which is not practical in many machine learning applications. Bayesian Optimization (BO) uses some prior knowledge about the function; for example, the unknown function is assumed to be a Gaussian process. In a small sub-region, the BO assumption is more likely to be true, which allows BO to converge quickly in the confined feasible domain identified by DIRECT. The new hybrid algorithm, Bayesian DIRECT (BD), combines strengths from both methods for better convergence to the global optimum. We provide a convergence result, and experiments show the efficiency of the new hybrid algorithm compared with Bayesian optimization.

Hongsheng Liu
University of North Carolina, Chapel Hill
hsliu@live.unc.edu

Dzung Phan
IBM T.J. Watson Research Center
phandu@us.ibm.com

**CP16**

**Approximating Splits for Decision Trees Quickly in Sparse Data Streams**

Decision trees are one of the most popular classifiers in the machine learning literature. While the most common decision tree learning algorithms treat data as a batch, numerous algorithms have been proposed to construct decision trees from a data stream. A standard training strategy involves augmenting the current tree by changing a leaf node into a split. Here we typically maintain counters in each leaf which allow us to determine the optimal split, and whether the split should be done. In this paper we focus on how to speed up the search for the optimal split when dealing with sparse binary features and a binary class. We focus on finding splits that have the approximately optimal information gain or Gini index. In both cases finding the optimal split can be done in $O(d)$ time, where $d$ is the number of features. We propose an algorithm that yields $(1+\alpha)$ approximation when using conditional entropy in amortized $O(\alpha^{-1}(1+m\log d)\log\log n)$ time, where $m$ is the number of 1s in a data point, and $n$ is the number of data points. Similarly, for Gini index, we achieve $(1+\alpha)$ approximation in amortized $O(\alpha^{-1}+m\log d)$ time. Our approach is beneficial for sparse data where $m \ll d$. In our experiments we find almost-optimal splits efficiently, faster than the baseline, overperforming the theoretical approximation guarantees.

Nikolaj Tatti
University of Helsinki
nikolaj.tatti@helsinki.fi

**MT1**

**See Tutorial Webpage for Tutorial Information**

Tutorial Attendees
,
NA

**MT2**

**Supervised Algorithmic Fairness in Distribution Shifts**

Chen Zhao
Baylor University
chen_zhao@baylor.edu

**MT3**

**See Tutorial Webpage for Tutorial Information**

Tutorial Attendees
,

NA

**MT4**

**See Tutorial Webpage for Tutorial Information**

Tutorial Attendees
.
N/A

**MT5**

**See Tutorial Webpage for Tutorial Information**

Tutorial Attendees
,
NA

**MT6**

**See Tutorial Webpage for Tutorial Information - 6**

Tutorial Attendees
,
NA

**MT7**

**See Tutorial Webpage for Tutorial Information**

Tutorial Attendees
,
NA

**MT8**

**Empowering Retrieval-augmented Generation with Graph-structured Knowledge**

¡ahref=INSERT WORKSHOP URL HERE¿INSERT WORKSHOP URL HERE¡/a¿

Yu Wang
University of Oregon
yuwang@uoregon.edu

Haoyu Han
Michigan State University
hanhaoy1@msu.edu

Harry Shomer
Michigan State University
USA
tbd

Kai Guo
Michigan State University
kaiguo@umich.edu

Yongjia Lei
University of Oregon
yongjia@uoregon.edu

Xianfeng Tang, Qi He
Amazon
xut10@psu.edu, tbd@amazon.com

Jiliang Tang
Michigan State University
tangjili@msu.edu

**MT9**

**TBD-1**

Xi Li
University of Alabama at Birmingham
XiLiUAB@uab.edu

**MT10**

**Mining Temporal Graphs: Algorithms and Applications**

Aristides Gionis
KTH Royal Institute of Technology
argioni@kth.se

**MT11**

**Unifying Spectral and Spatial Graph Neural Networks**

Zhiqian Chen
Mississippi State University
zchen@cse.msstate.edu