Demystifying the Node-Level Link Prediction Variability of Graph Neural Networks

Tyler Derr

Network and Data Science Lab
Vanderbilt University

October 25, 2024
OARS Workshop
CIKM 2024

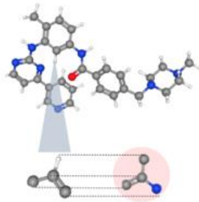
Tyler Derr

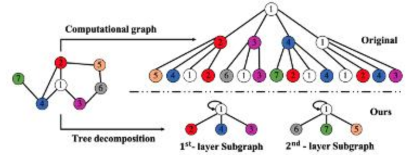
Assistant Professor of Computer Science at Vanderbilt University

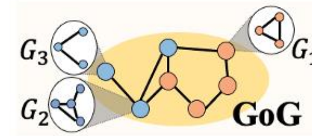
- Teaching and Affiliate Faculty Member, Data Science Institute
- Faculty Fellow, The Frist Center for Autism and Innovation

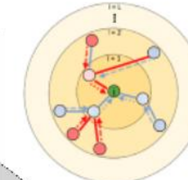
VANDERBILT COMPUTER SCIENCE DEPARTMENT

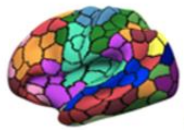
in =











Deep Learning on Graphs
- Graph Neural Networks
- Self-supervised Learning
- Knowledge Graphs
- Data Quality Issues

...

Interdisciplinary Social Good Applications
- Drug Discovery
- Political Science
- Education
- Neurodiversity

...

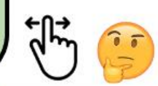
NDS

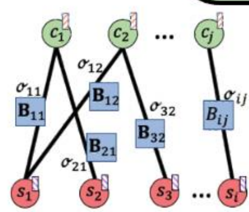
Network and Data Science Lab at Vanderbilt University

http://nds-vu.github.io

Responsible and Trustworthy AI
- Bias and Fairness
- Explainability
- Data Privacy & Machine Unlearning

...





Social Network Analysis
- Network Models
- Measurements
- Social Theories
- Applications

...

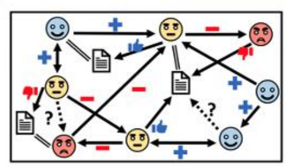
&

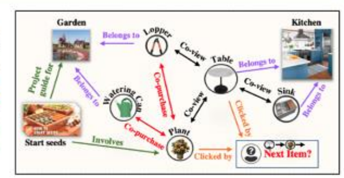
Recommender Systems
- Cold-Start Problem
- Diversity
- Session-Based
- Topological Analysis

...

Acknowledgements

Thank you for supporting our research!





Welcome to join our virtual Vanderbilt Machine Learning Seminar Series every Monday afternoon. Details at: http://vu.edu/ML

Personal Homepage: http://www.TylerDerr.com
NDS Lab Homepage: http://my.vanderbilt.edu/NDS
LinkedIn: http://www.linkedin.com/in/TylersNetwork
Twitter: http://www.twitter.com/TylersNetwork

NDS Lab, Vanderbilt, and Nashville

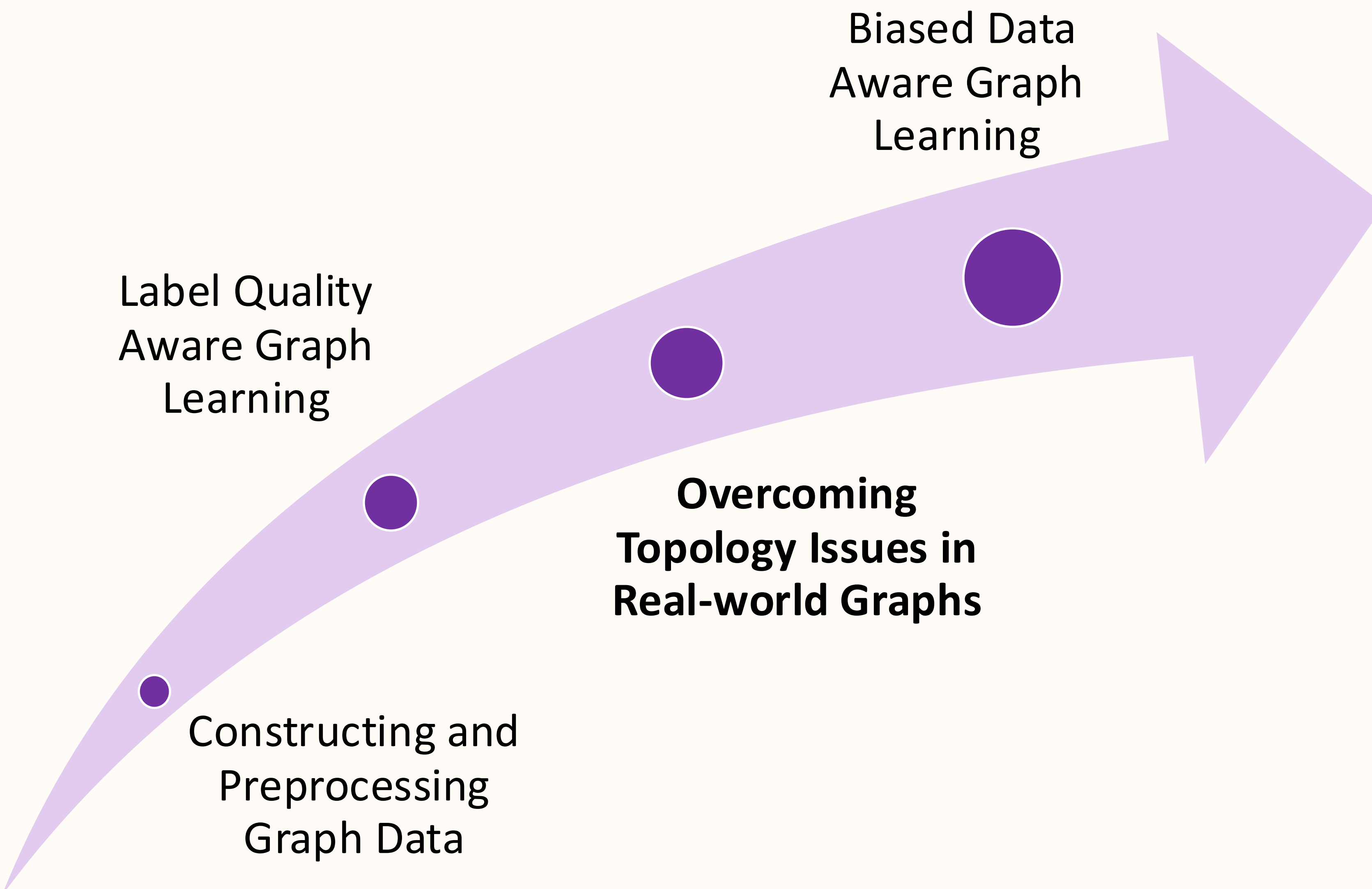
Network and Data Science Lab

Vanderbilt University

(The campus is an accredited arboretum
with ~200 species of trees and shrubs.)

Nashville, Tennessee, USA

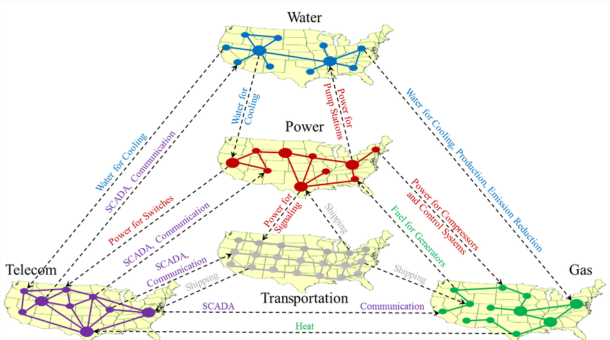
Data Quality-Aware Graph Learning



Data is Connected

Graphs are everywhere in today's connected world
…and can be constructed from (un)structured data

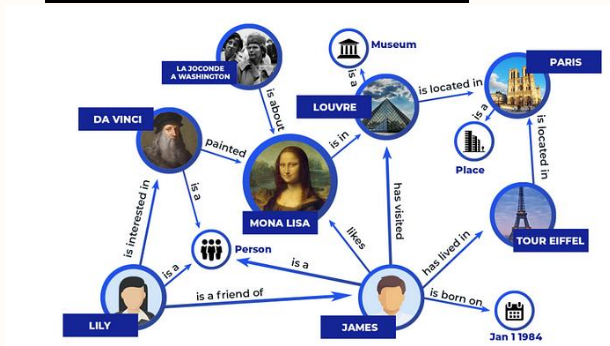
Data fusion



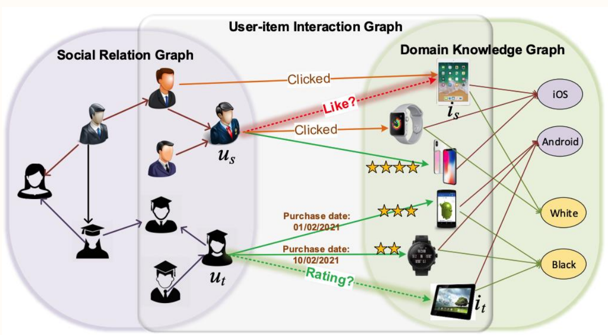
Konstantin Zuev, et al. 2017

Sunitha articlecube

Knowledge extraction

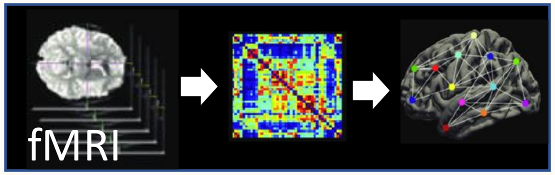


Zhiquan Liu (talk slides)

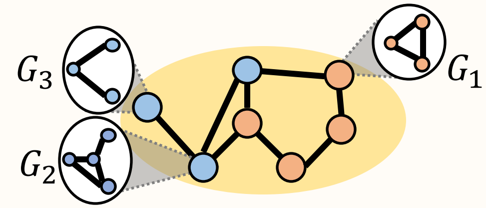


Shoujin Wang et al. 2021

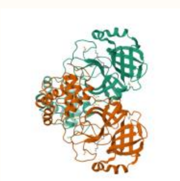
Similarity-based construction

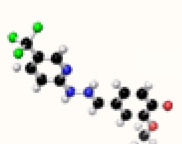


Anand Eijlers et al, 2019



Chemical structures





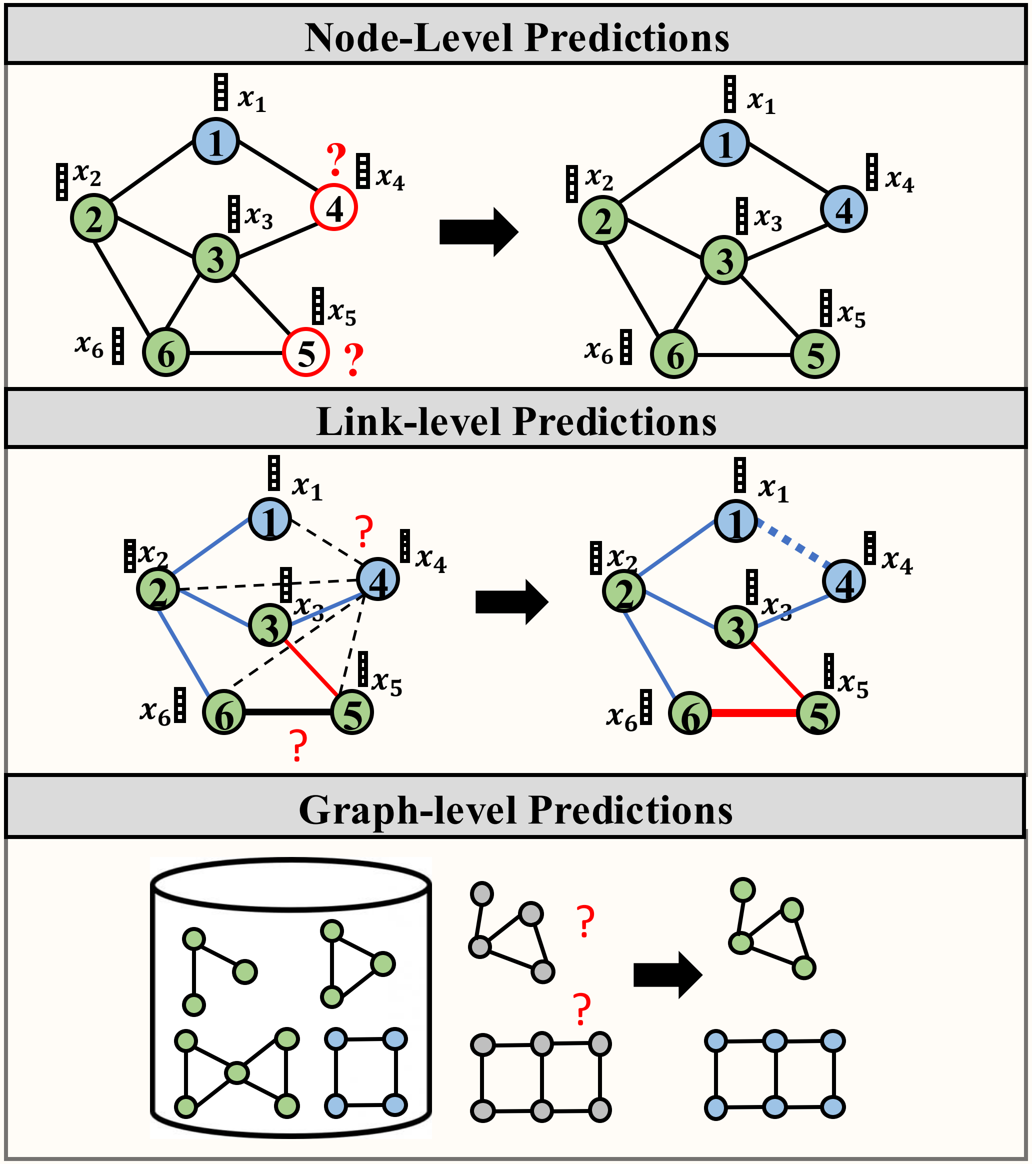
Protein Data Bank

…

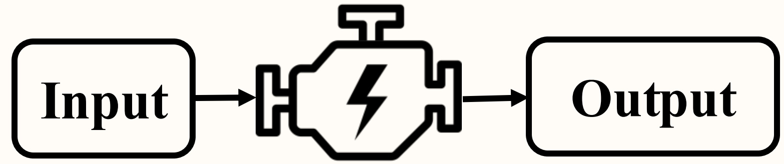
Complex social systems

Aelitta, iStock

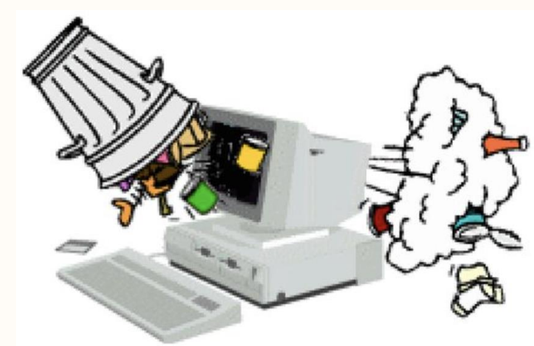
Graph Machine Learning



Graph ML Model



Real-world data can have data quality challenges…

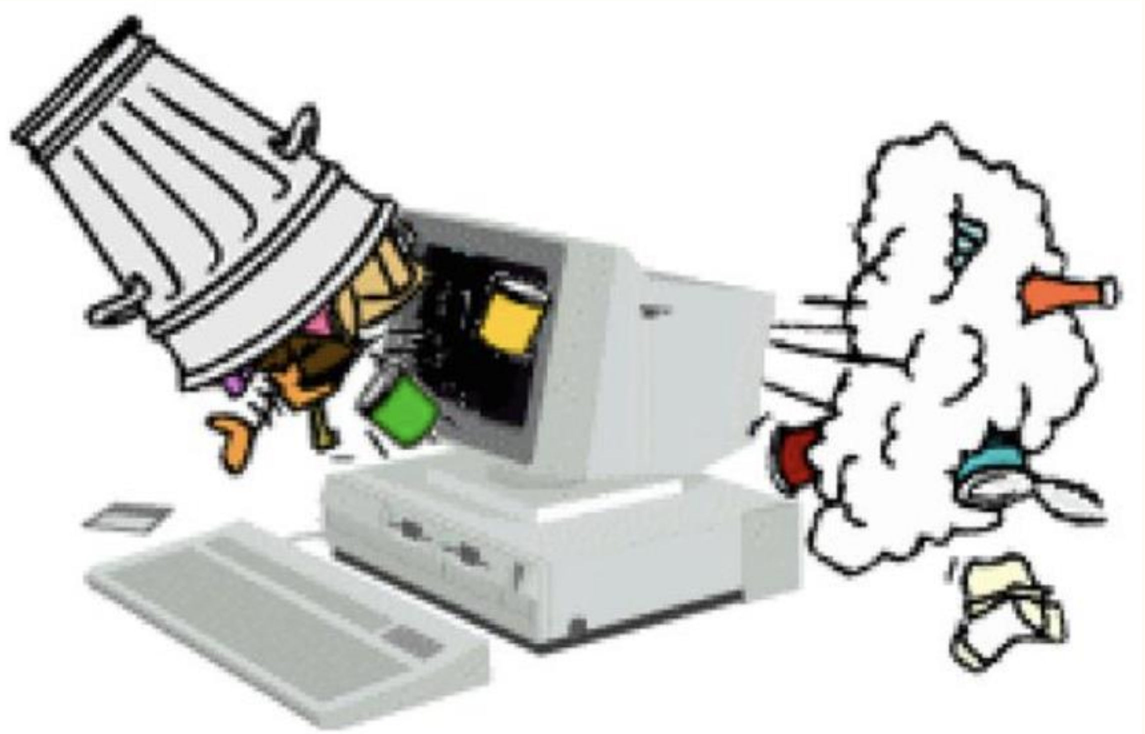


Garbage in, garbage out

Real-World Graph Data Quality Challenges

What are data quality challenges?

- Imbalanced data
- Biased data
- Noisy outliers
- Limited labels
- Missing values
- Uncertain topology
- Distribution shifts
- etc.



Garbage in, garbage out

How to mitigate these challenges?

Model-Centric vs. Data-Centric AI

Model-Centric

- Model architectures
- Loss functions/ constraints
- Hyperparameter tuning
- …
- etc.

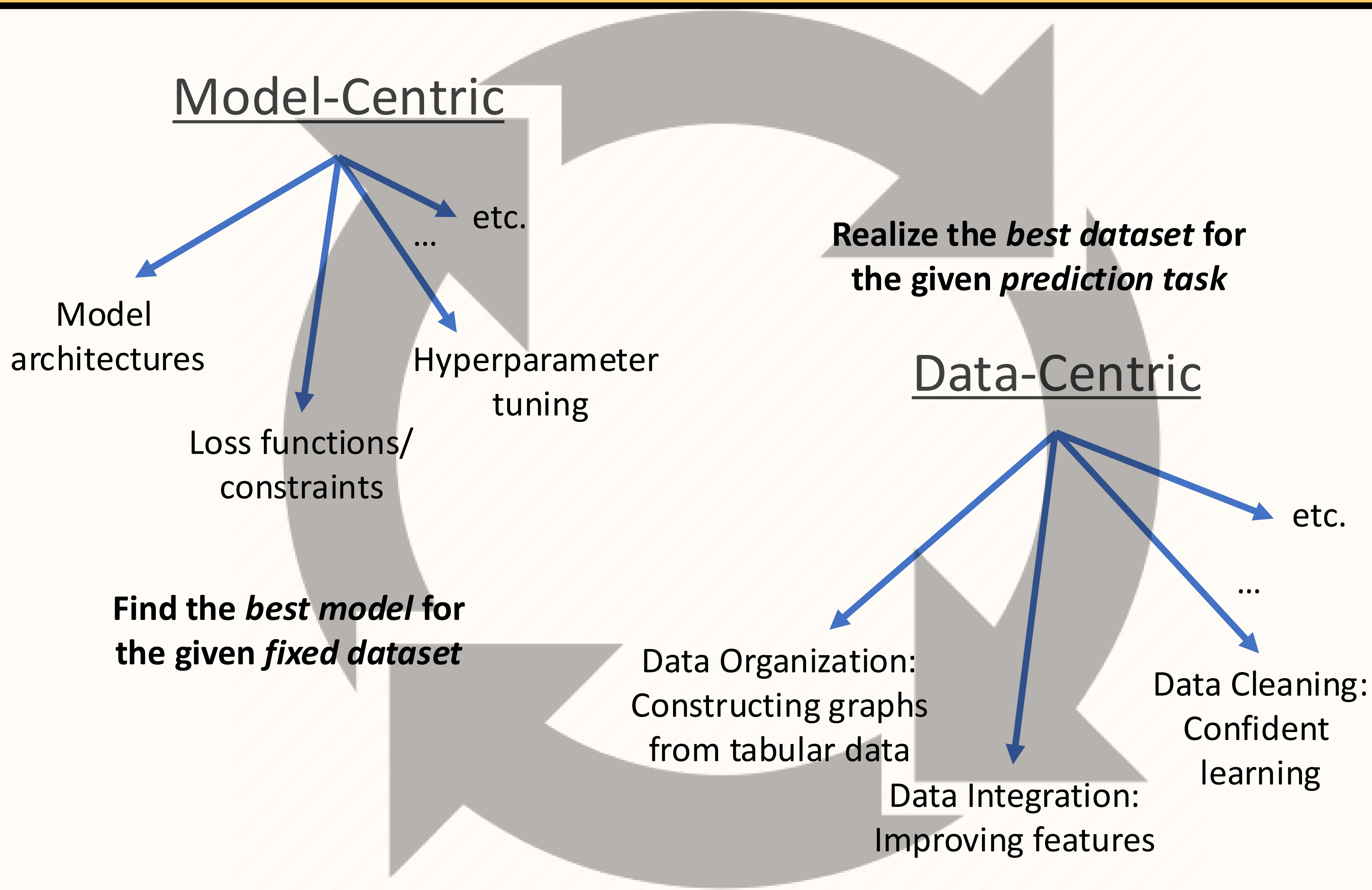
Find the *best model* for the given *fixed dataset*

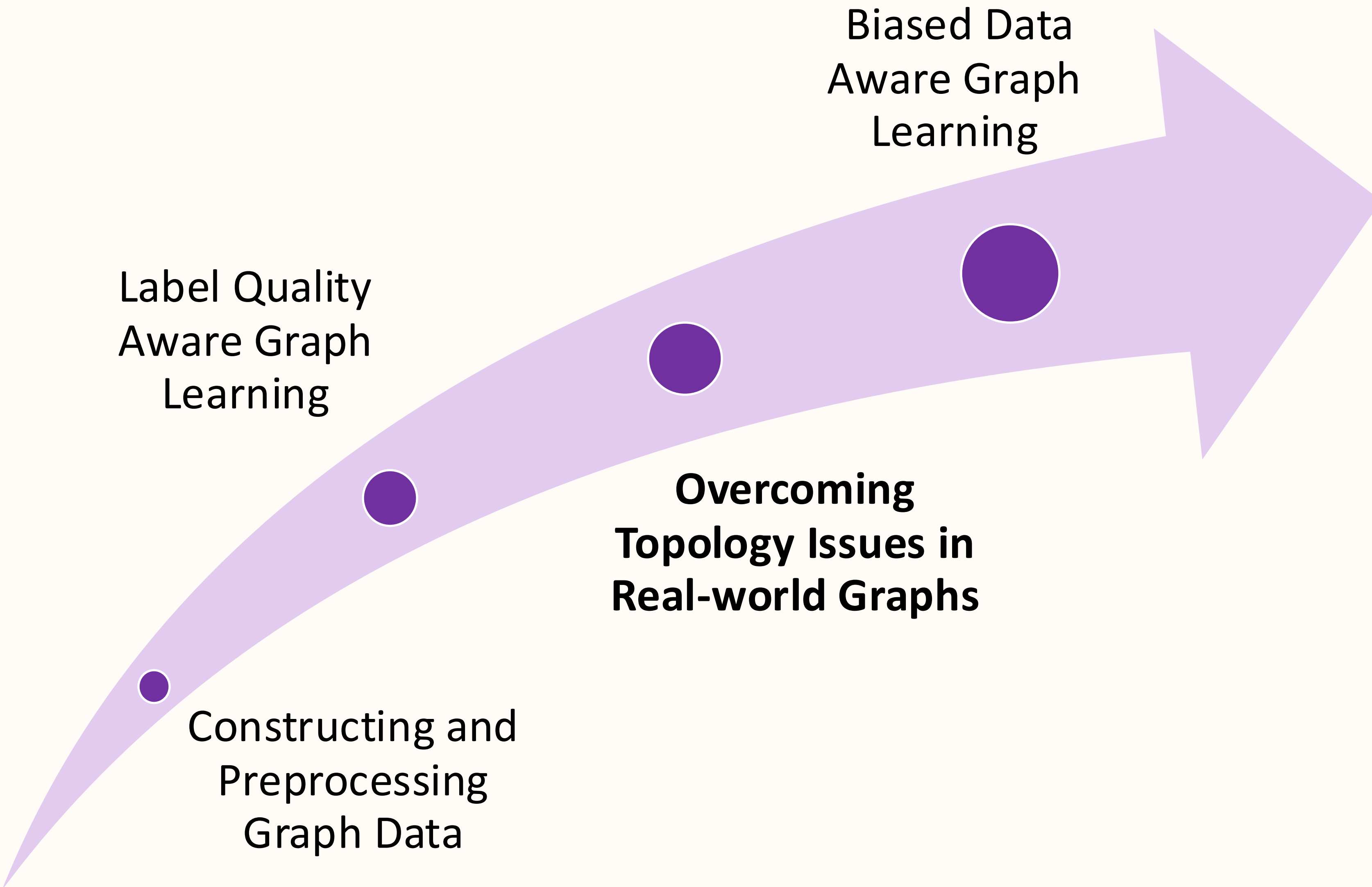
Realize the *best dataset* for the given *prediction task*

Data-Centric

- Data Organization: Constructing graphs
- Data Integration: Improving node/edge features
- Data Cleaning: Confident learning
- …
- etc.

Harmonization for Improved Ethical AI in Society




Biased Data Aware Graph Learning
Label Quality Aware Graph Learning
Overcoming Topology Issues in Real-world Graphs
Constructing and Preprocessing Graph Data

Biased Data Aware Graph Learning

Label Quality Aware Graph Learning

Overcoming Topology Issues in Real-world Graphs

Constructing and Preprocessing Graph Data

Graphs for Recommender Systems

Traditional Recommendation Problems

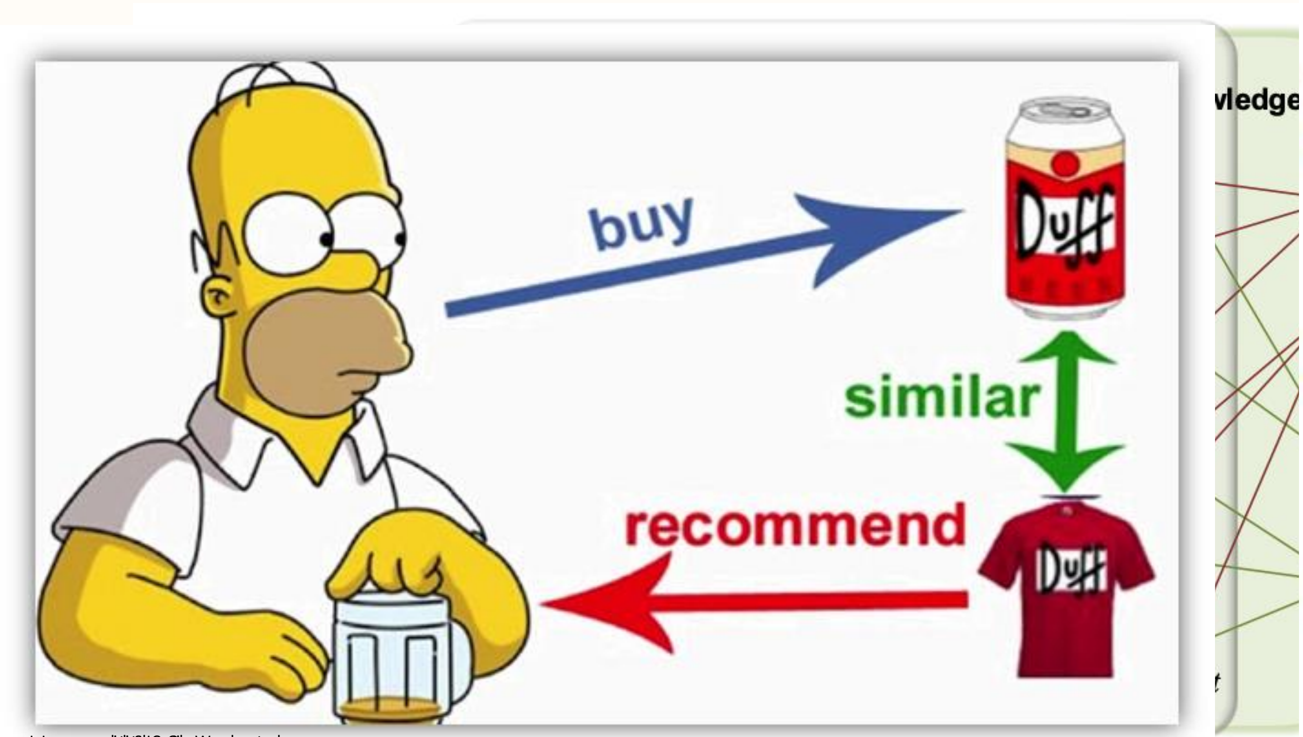


Image credit: Shoujin Wang et al.

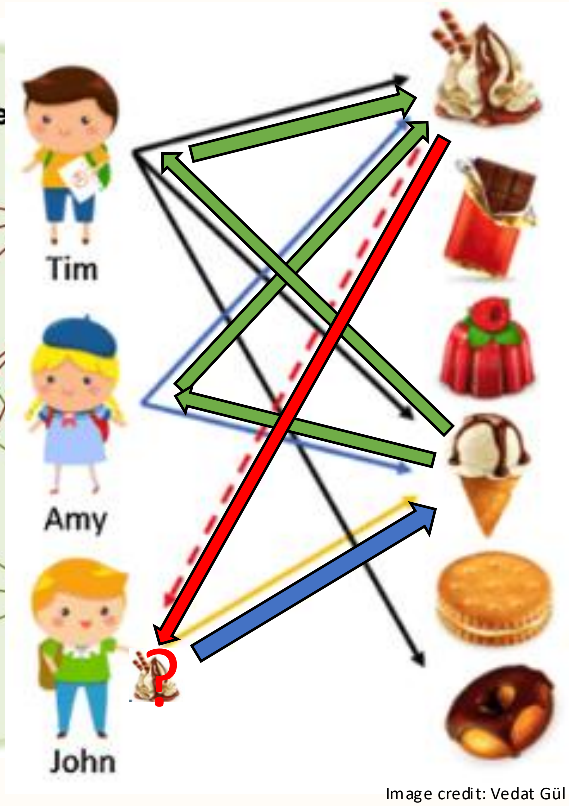
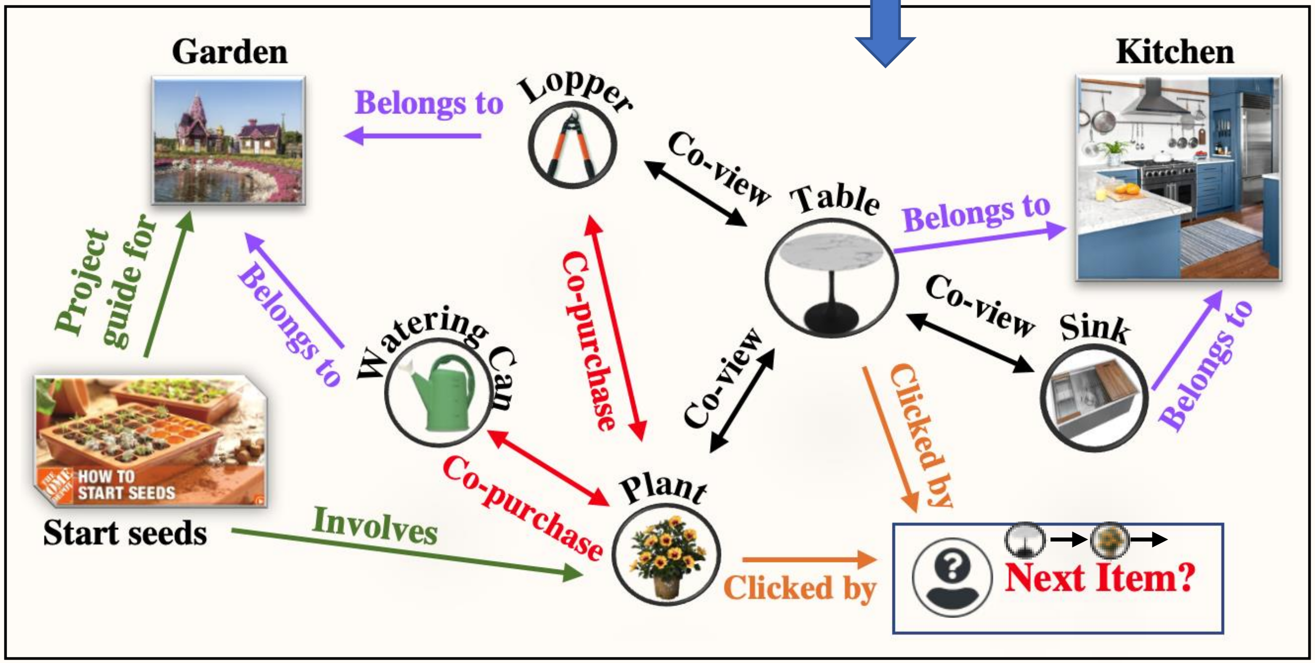


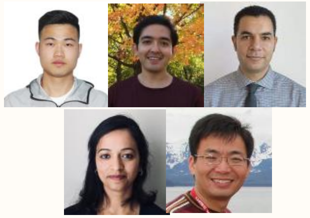
Image credit: Vedat Gül

Graphs for Session-based Recommendation

Domain Knowledge

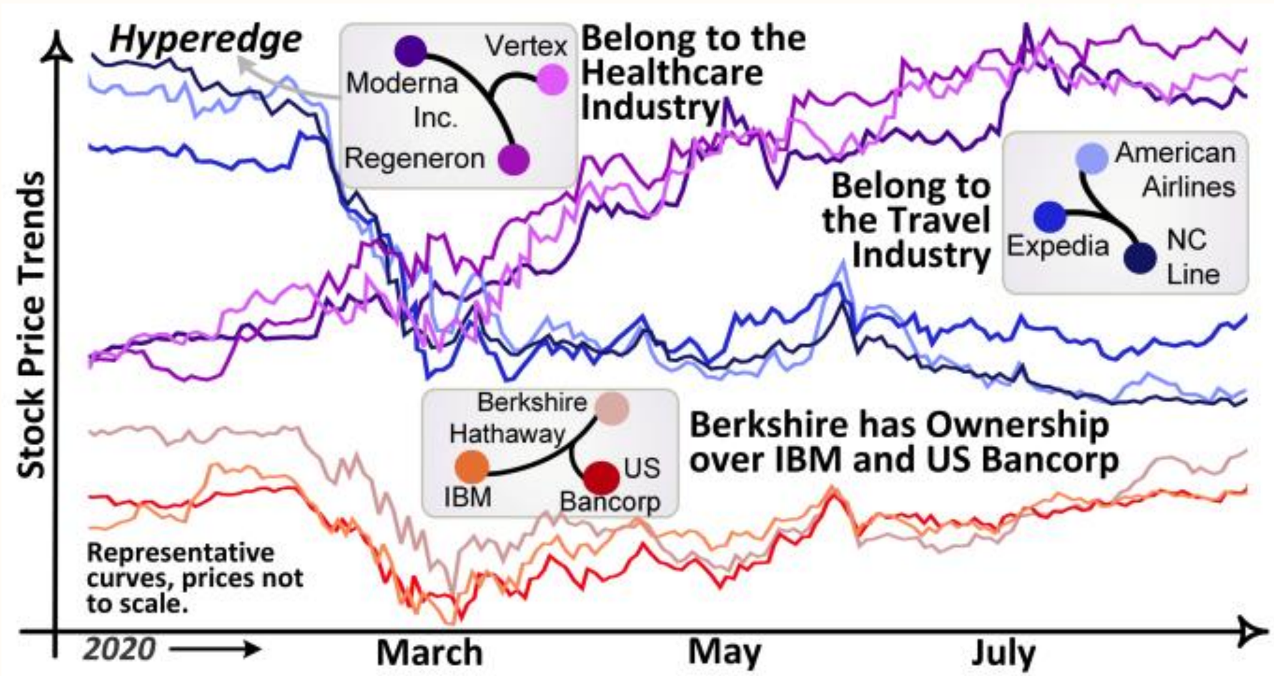
Data fusion **with domain knowledge and global information extracted across historical sessions:**



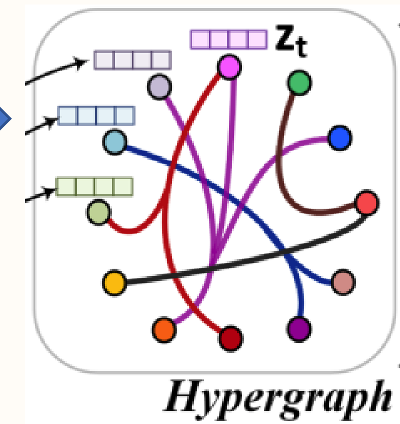


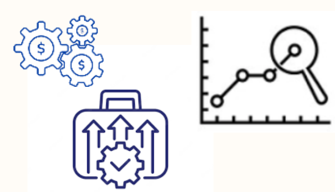
Graphs for Fintech

Constructing Relations Between Stocks

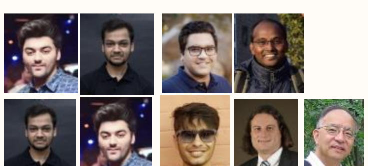


Graph Encoding Relations Between Stocks



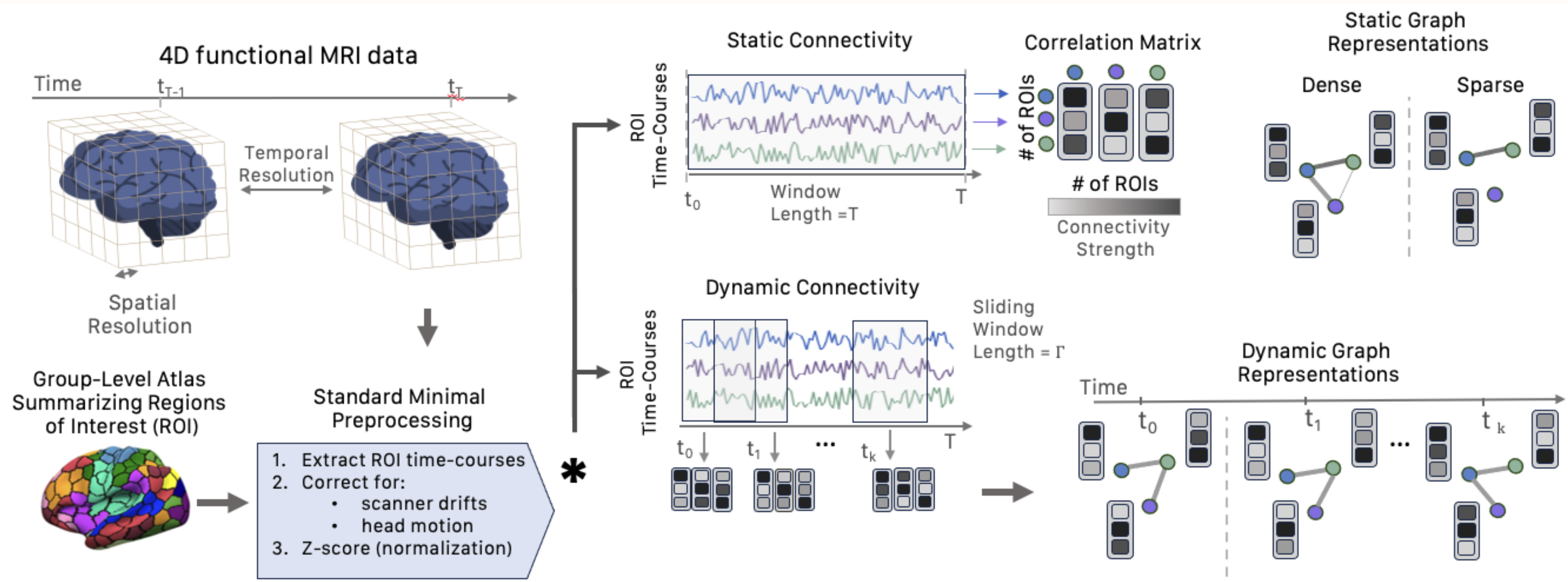


Graph-based Market Prediction & Portfolio Optimization



Graphs for Neuroimaging

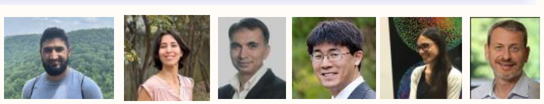
A Data-Centric AI Approach to Improved Learning in Brain Connectomics



NeuroGraph

A Python package for fMRI preprocessing and a collection of graph-based Neuroimaging datasets for graph machine learning applications

Install: pip install NeuroGraph



Datasets, code, and documentation is publicly available!
https://neurograph.readthedocs.io/

	Dataset	$\|G\|$	$\|N\|_{avg}$	$\|E\|_{avg}$	d_{max}	d_{avg}	K	$\|X\|$	$\|Y\|$	Task
Static	HCP-Activity	7443	400	7029.18	153	19.40	0.41	400	7	Graph Classification
	HCP-Gender	1078	1000	45578.61	413	45.78	0.46	1000	2	Graph Classification
	HCP-Age	1065	1000	45588.40	413	45.78	0.46	1000	3	Graph Classification
	HCP-FI	1071	1000	45573.67	413	45.78	0.46	1000	-	Graph Regression
	HCP-WM	1078	1000	45578.61	413	45.78	0.46	1000	-	Graph Regression
Dynamic	DynHCP-Activity	7443	100	843.04	992	6.22	0.427	100	7	Graph Classification
	DynHCP-Gender	1080	100	874.88	992	9.26	0.439	100	2	Graph Classification
	DynHCP-Age	1067	100	875.42	992	9.26	0.439	100	3	Graph Classification
	DynHCP-FI	1073	100	874.82	992	9.26	0.438	100	-	Graph Regression
	DynHCP-WM	1080	100	874.88	992	9.26	0.439	100	-	Graph Regression

NeuroGraph

Five new benchmark datasets for graph classification/regression!

NeuroGraph

A Python package for fMRI preprocessing and a collection of graph-based Neuroimaging datasets for graph machine learning applications

Install: pip install NeuroGraph

NeuroGraphDataset

PyG

Now available in PyG!

class **NeuroGraphDataset** (**root**: str, **name**: str, **transform**: Optional[Callable] = **None**, **pre_transform**: Optional[Callable] = **None**, **pre_filter**: Optional[Callable] = **None**) [source]

Bases: `InMemoryDataset`

The NeuroGraph benchmark datasets from the "NeuroGraph: Benchmarks for Graph Machine Learning in Brain Connectomics" paper. `NeuroGraphDataset` holds a collection of five neuroimaging graph learning datasets that span multiple categories of demographics, mental states, and cognitive traits. See the documentation and the Github for more details.

Key Insight: Better performance with ***larger***, ***(sparser) graphs*** with ***correlation node features.***

Dataset		k-GNN	GCN	SAGE	UniMP	ResGCN	GIN	Cheb	GAT	SGC	General	Avg.
100ROIs	CORR	65.65	68.98	68.70	68.33	66.06	68.24	63.94	69.49	68.43	64.95	**67.30**
	BOLD	49.58	50.97	51.67	51.30	51.34	55.09	53.19	49.95	51.90	51.11	51.11
	CORR+BOLD	52.78	51.02	50.28	50.79	50.60	54.91	49.44	50.37	51.57	51.30	51.36
400ROIs	CORR	72.21	74.10	61.66	68.57	70.09	71.89	58.94	69.35	**75.99**	**73.09**	**69.56**
	BOLD	51.16	51.62	53.94	51.39	52.31	55.09	49.07	50.46	53.24	53.94	52.22
	CORR+BOLD	51.53	51.90	52.96	51.57	52.36	55.56	50.63	52.13	52.08	52.61	53.33
1000ROIs	CORR	**78.80**	**75.19**	**71.71**	**75.14**	**78.75**	**77.22**	**64.77**	**71.34**	73.75	63.13	**72.98**
	BOLD	48.15	46.99	49.31	50.93	47.92	56.48	47.22	50.93	49.31	51.62	49.89
	CORR+BOLD	51.30	51.81	51.25	51.11	49.86	54.35	49.66	51.22	51.34	51.37	51.33

Dataset			$k-$GNN	GCN	SAGE	UniMP	ResGCN	GIN	Cheb	GAT	SGC	General
Gender Classification	100ROIs	Sparse	63.33	72.96	69.35	69.72	68.06	69.72	63.70	70.28	70.37	67.22
		Medium	65.65	68.98	68.70	68.33	66.06	68.24	63.94	69.49	68.43	64.95
		Dense	64.44	68.52	65.00	68.06	63.70	66.39	64.26	69.72	68.43	61.76
	400ROIs	Sparse	69.95	**77.14**	69.86	67.56	71.43	69.4	66.45	72.72	**78.25**	76.13
		Medium	65.65	68.98	68.70	68.33	66.06	68.24	63.94	69.49	68.43	64.95
		Dense	71.61	76.13	62.58	61.20	69.77	73.27	61.84	67.83	74.19	72.44
	1000ROIs	Sparse	**82.13**	75.46	77.69	**76.67**	78.33	75.56	59.07	**76.2**	76.48	**78.89**
		Medium	78.80	75.19	71.71	75.14	78.75	77.22	71.43	71.34	73.75	63.13
		Dense	61.57	73.80	**78.86**	72.50	**78.89**	**78.70**	**76.67**	71.67	75.25	72.69

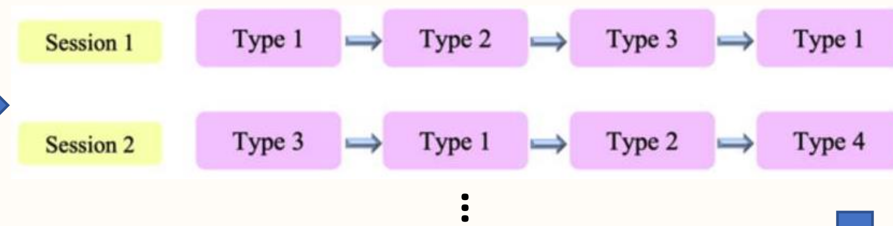
Graphs for Healthcare

Electronic Health Records

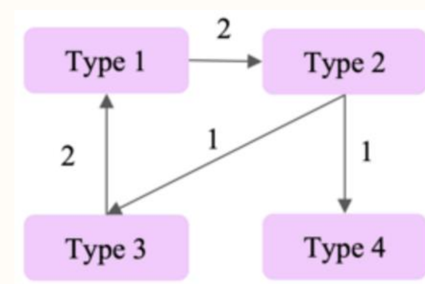
Audit log events performed on a patient

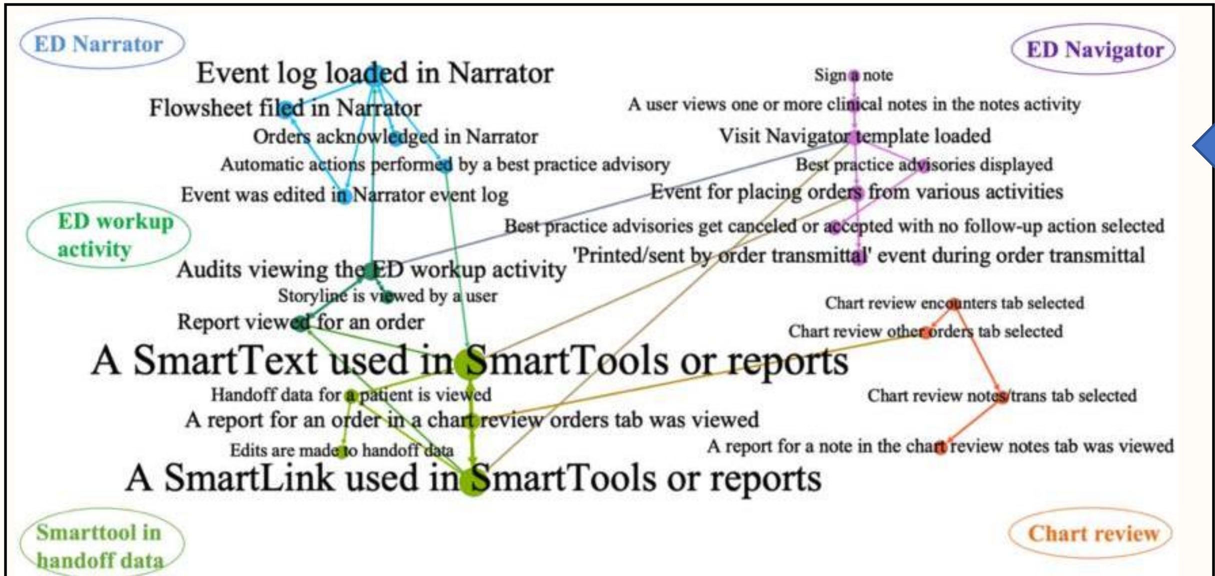
Patient ID	User ID	Timestamp	User-EHR interaction type
1000	A	10:41:00	Measurements reviewed
1000	B	10:41:10	Medication prescribed
1000	A	10:42:00	Signed an order
1000	C	10:46:00	Lab test results exported
...	...	...	...
1000	B	12:46:50	Medication list exported

Clinician Task Workflows



Global Task Interaction Graph





Applications:
Identify bottlenecks, cluster tasks, identify inconsistencies, …

Graphs for Biomedical

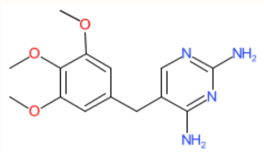
Molecule Representation

SMILES String Representation

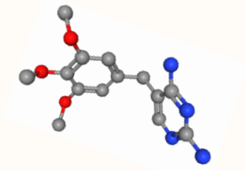
COc1cc(cc(c1OC)OC)Cc2cnc(nc2N)N

Graph Representations

2D Graph

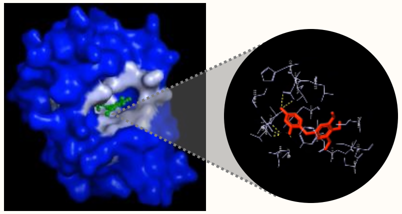


3D Graph



Virtual Screening
with Graph Machine Learning

Prediction Task:
Active or Inactive

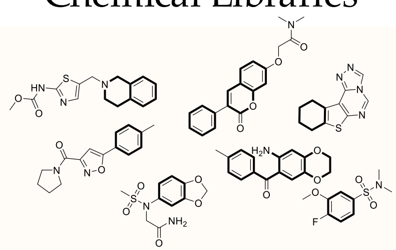


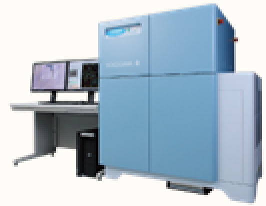
Labeled Data

Drug Discovery

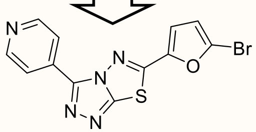
Experimental Screening

Chemical Libraries





High Throughput Screening (HTS) Equipment

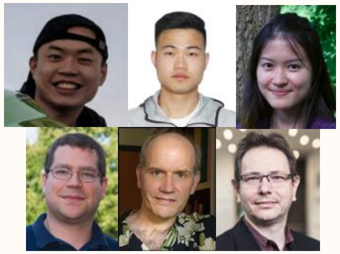


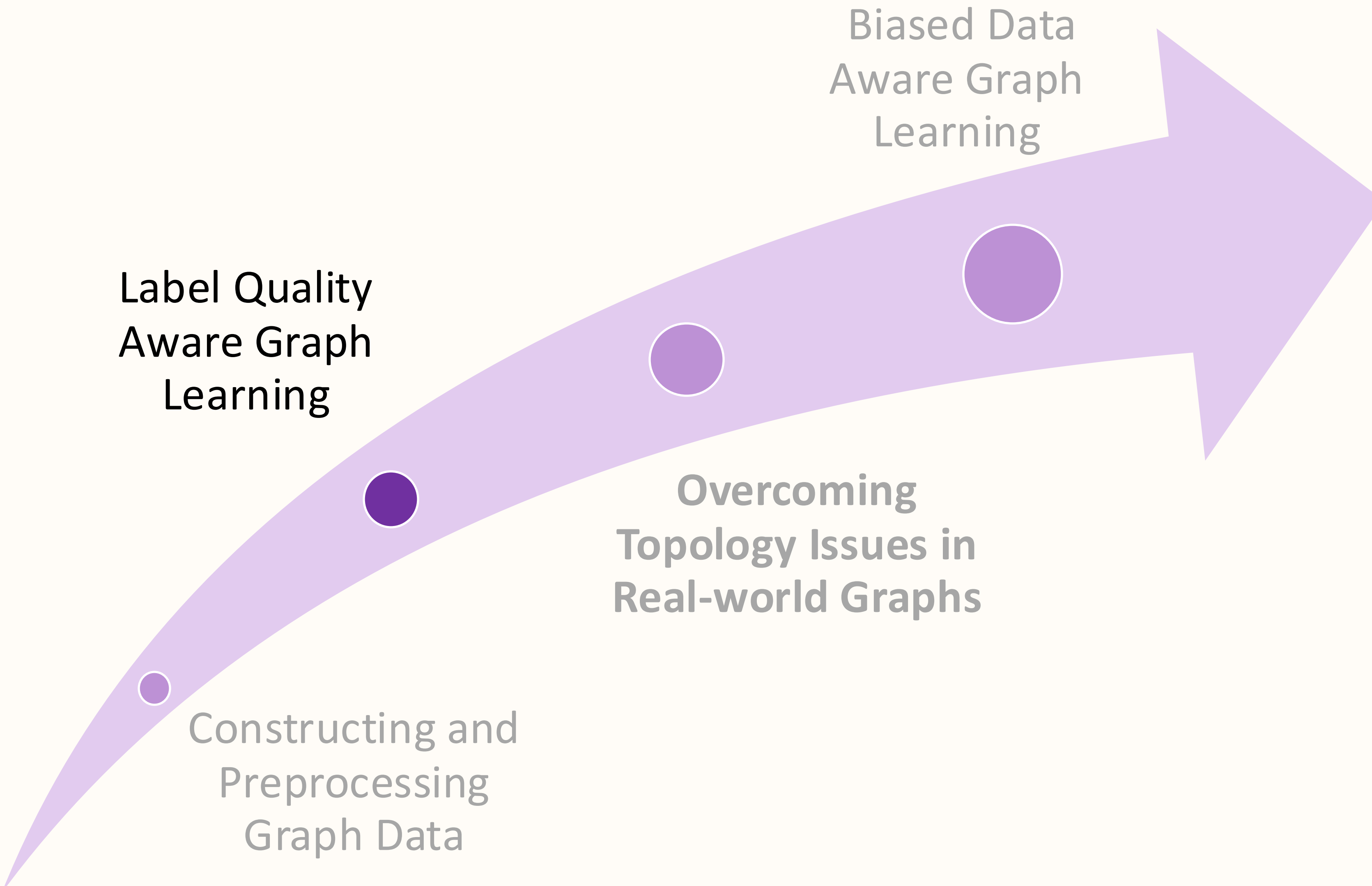
Hit Rate: 0.05%-0.5%

WelQrate
Defining the Gold Standard in Small Molecule Drug Discovery Benchmarking

Coming soon!
Just accepted at NeurIPS'24

HTS generates *inherently highly imbalanced* labeled data




Constructing and Preprocessing Graph Data
Label Quality Aware Graph Learning
Overcoming Topology Issues in Real-world Graphs
Biased Data Aware Graph Learning

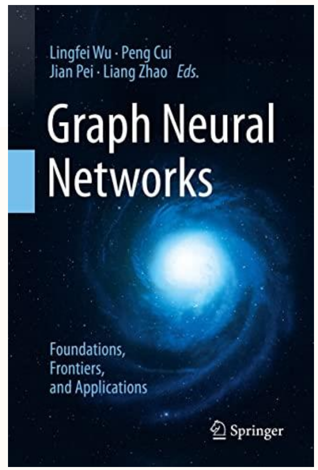
Self-Supervised Learning on Graphs

Less/no labeled data? Can leverage SSL on Graphs.

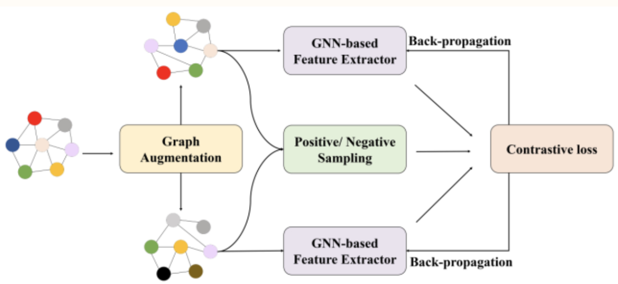
Chapter 18

Graph Neural Networks: Self-supervised Learning

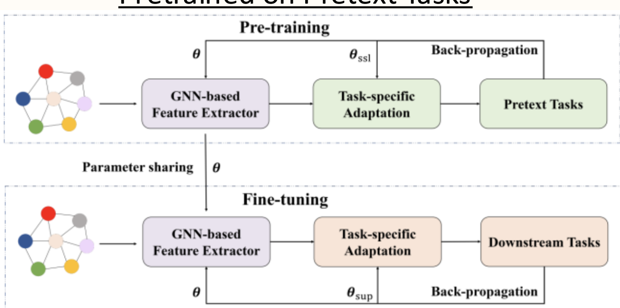
Yu Wang, Wei Jin, and Tyler Derr



Contrastive Learning



Pretrained on Pretext Tasks



⋮

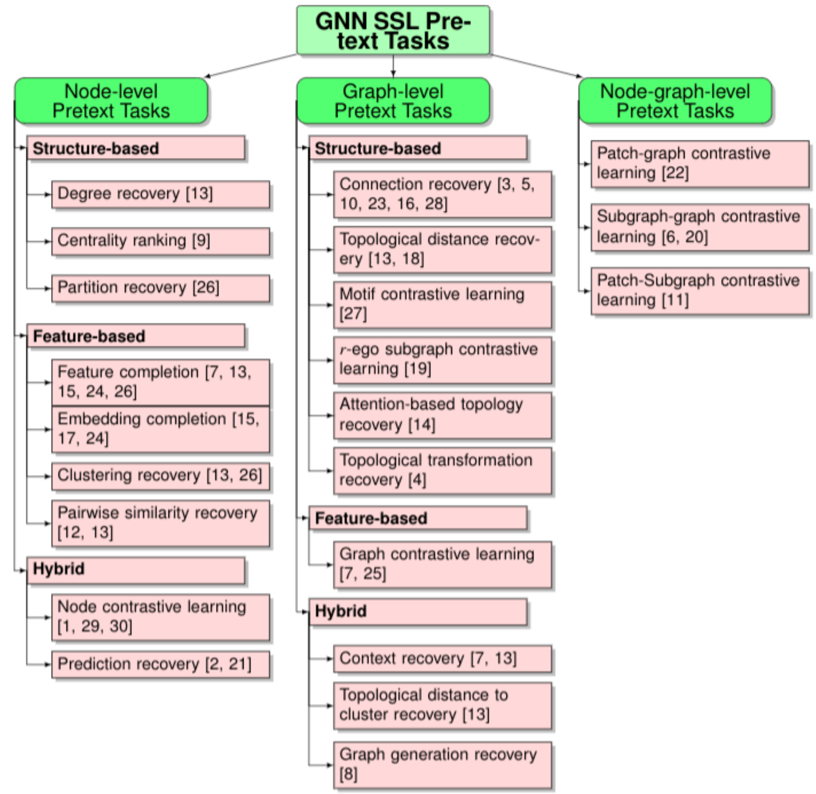
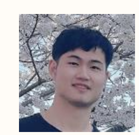
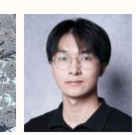


Figure 1: A categorization of SSL pretext tasks used in GNNs.

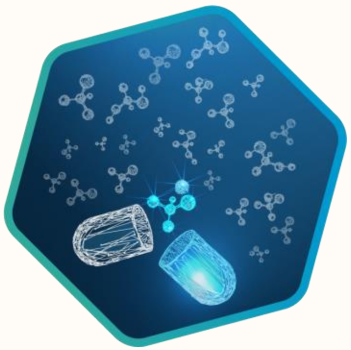
https://github.com/NDS-VU/GNN-SSL-chapter





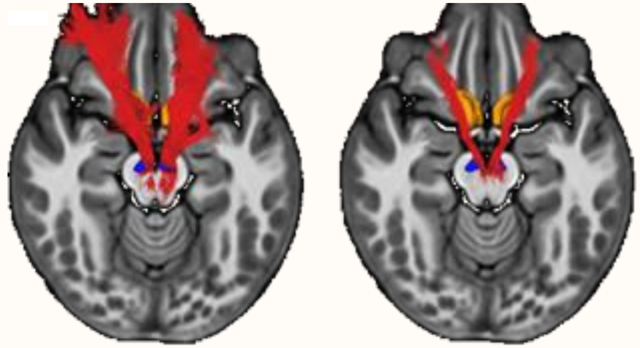
Imbalanced Graph Datasets

Drug Discovery



HTS Hit Ratio
0.05% to 0.5%
Bajorath et al. 2002

Brain Classification



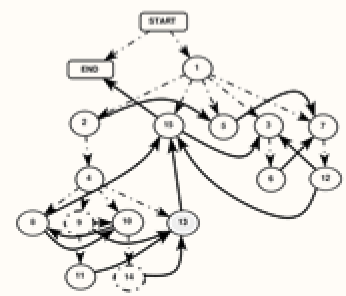
Typical		Autism
36	:	1

Autism Statistics. 2023

Fake News Detection

0.15%
Dou et al. 2021

Malware Detection

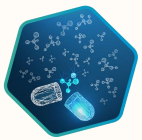


0.01% to 2% Android
Oak et al. 2019

Classification on Imbalanced Graph Datasets

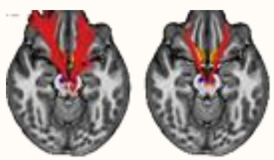
Problem

Drug Discovery

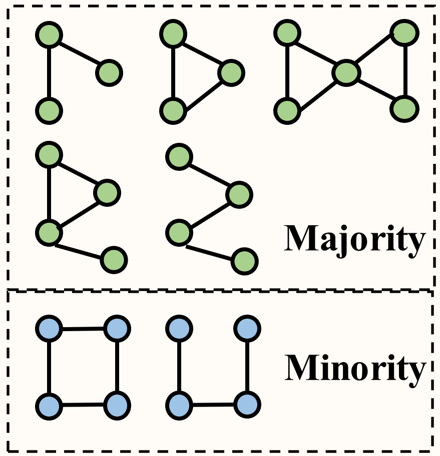


HTS: Hit Ratio
0.05% to 0.5%

ASD Brain Classification

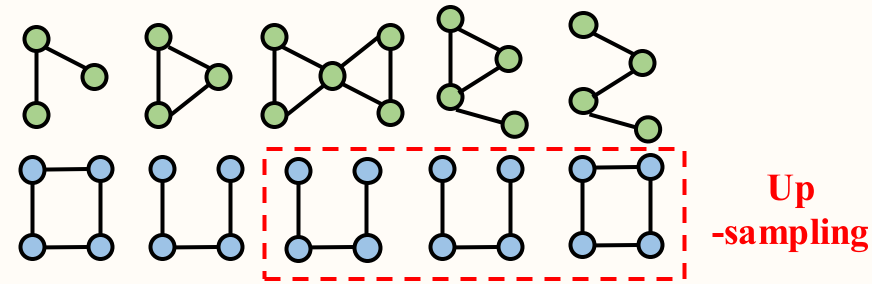


Normal **Autism**
36 : 1



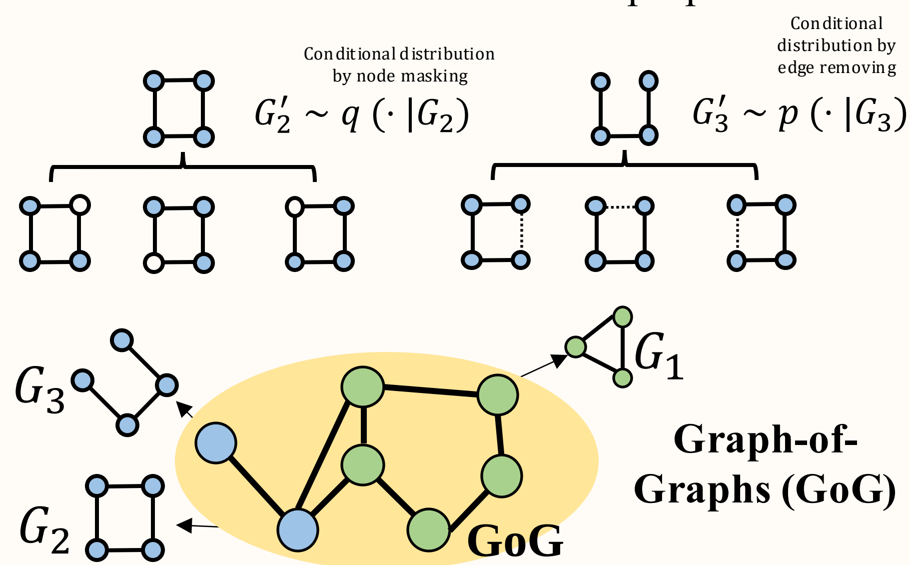
Method

Quantity Augmentation



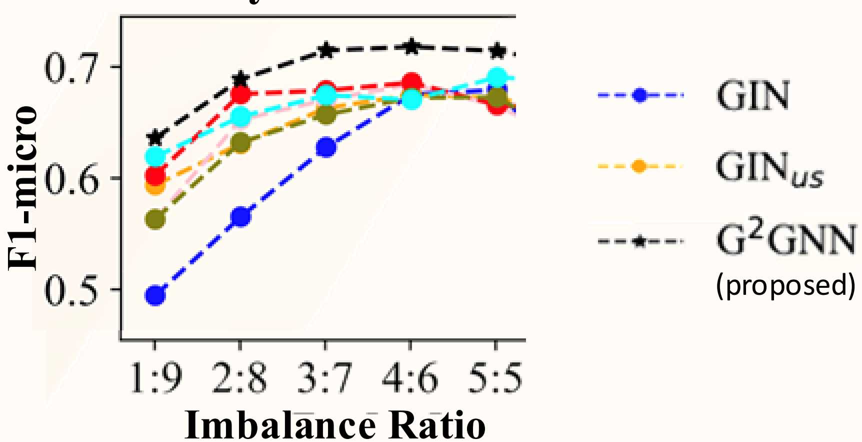
Structure Augmentation

Similar Property Principle - Structurally similar molecules tend to have similar properties



Results

DHFR Enzyme Classification



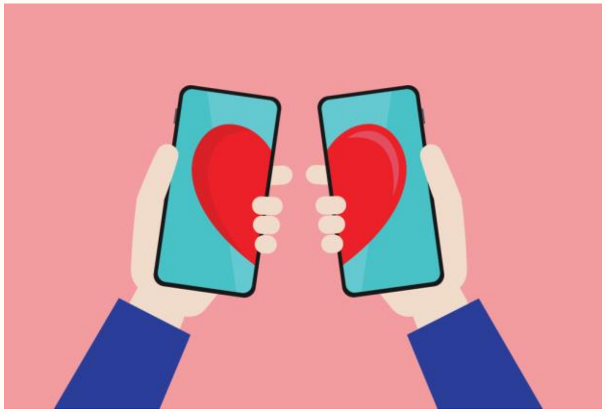
Constructing and Preprocessing Graph Data

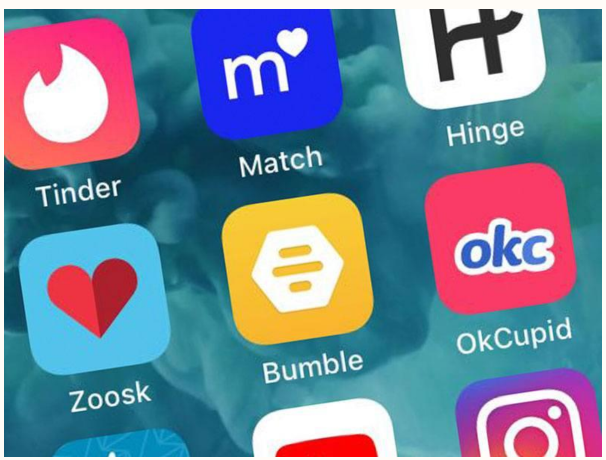
Label Quality Aware Graph Learning

Overcoming Topology Issues in Real-world Graphs

Biased Data Aware Graph Learning

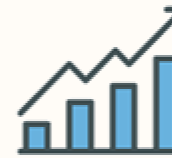
Online Dating

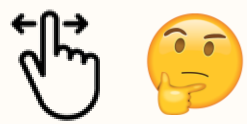




Online dating:
- 15% of Americans (2013)
- Increase to 30% (2019)

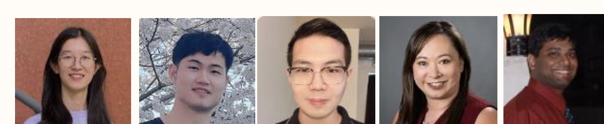
Increasing demand on online dating



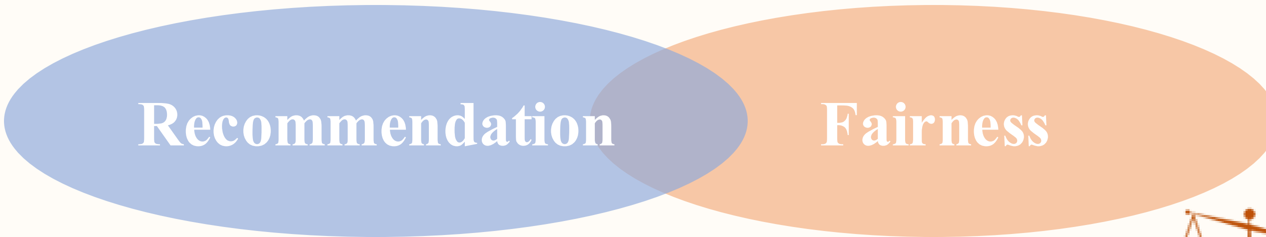


Information overload

Online Dating Recommender Systems



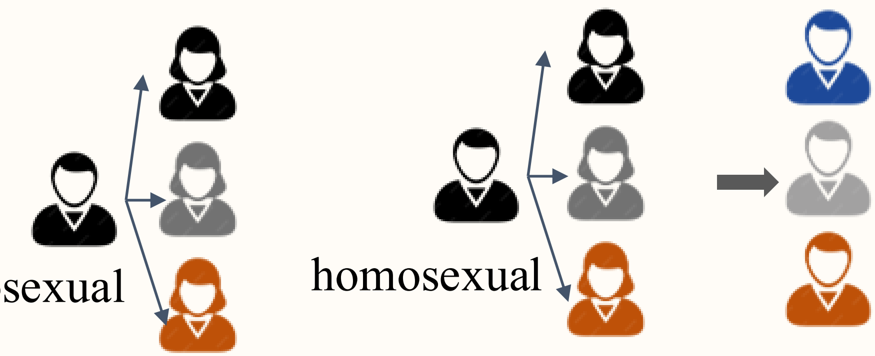
Ethics of AI in Online Dating



Are users in diverse groups treated fairly?

Prior Works
- Gender
- Race
- Religion
- Subscription

User Sexual Orientation

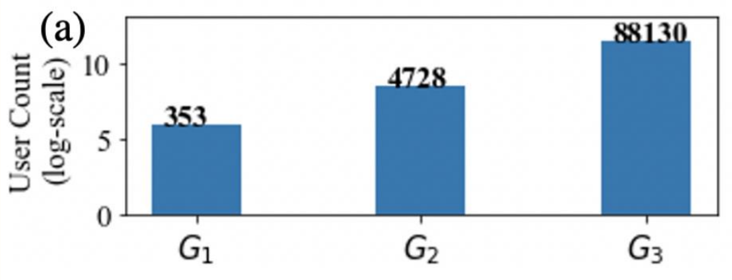


Do users of varying sexual orientation get treated fairly?

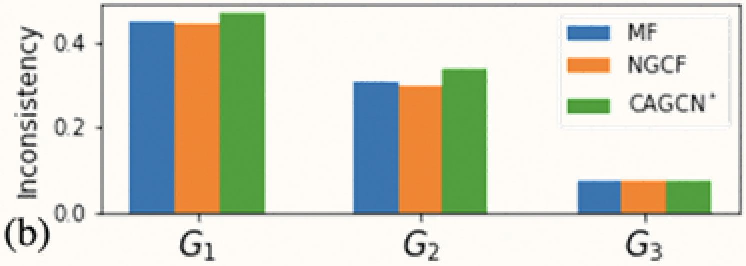
Potential Reasons for The Performance Gap

Reasons

Group Data Quantity Imbalance



Gender Inconsistency Imbalance
(train/recommendation)



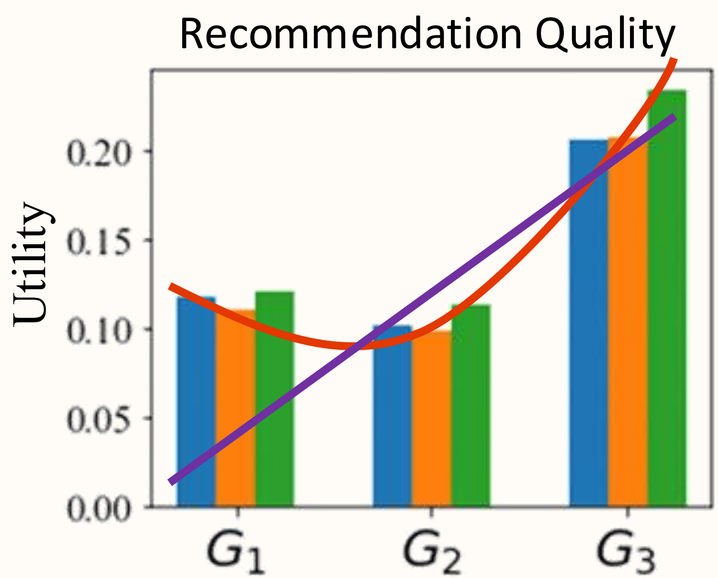
Increasingly deviates away from historical interaction behaviors

Solutions

Re-weighting
- In-processing
- Adjust the weights during optimization

Re-ranking
- Post-processing
- Calibration to mitigate inconsistency

Quantity Imbalance vs. User Interest Diversity



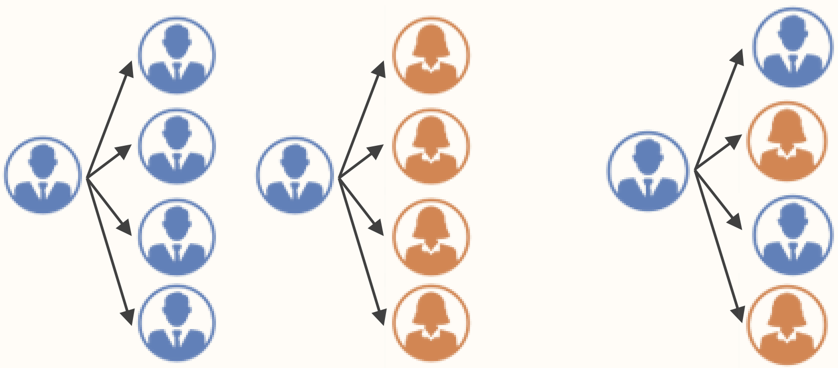
Specific Interests to the same gender

Purple showing expected performance based on quantity imbalance.

Broad Interests to both genders

Specific Interests to the opposite gender

Online Dating Recommendation

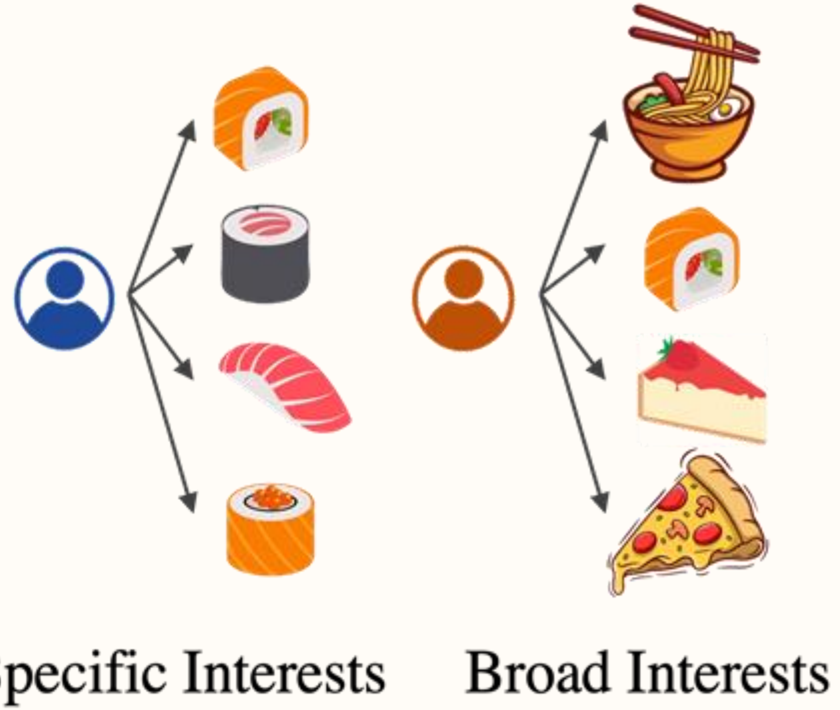


Specific Interests

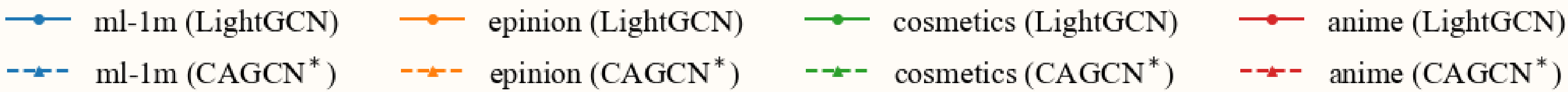
Broader Interests

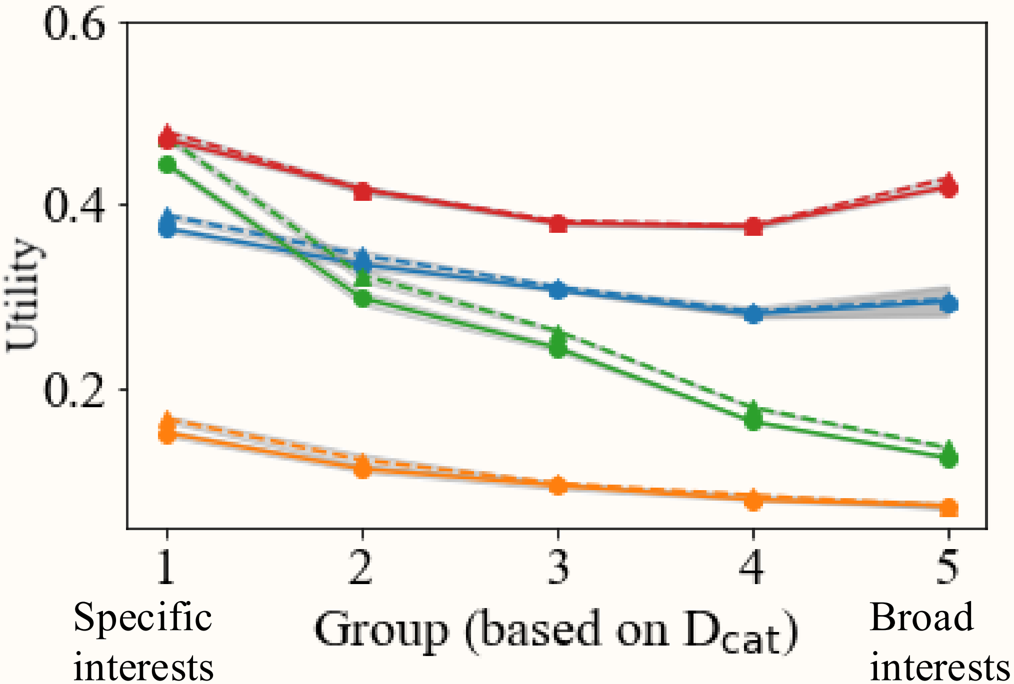
Research Question

Are users of varied interest diversity treated fairly in Recommender Systems?

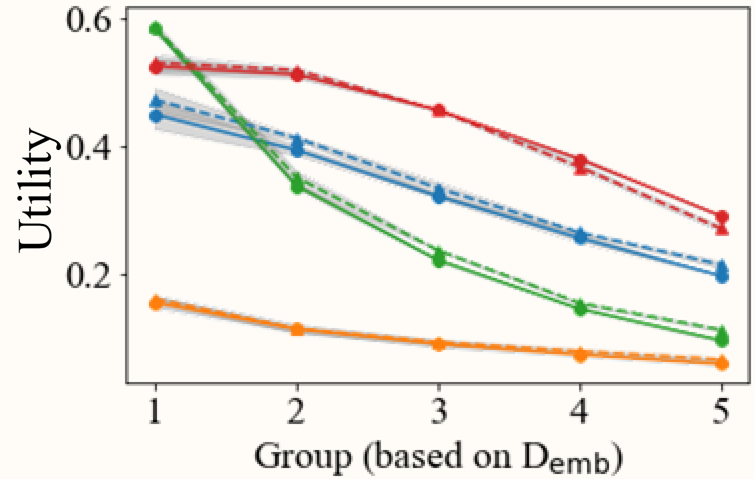


(in terms of item categories)

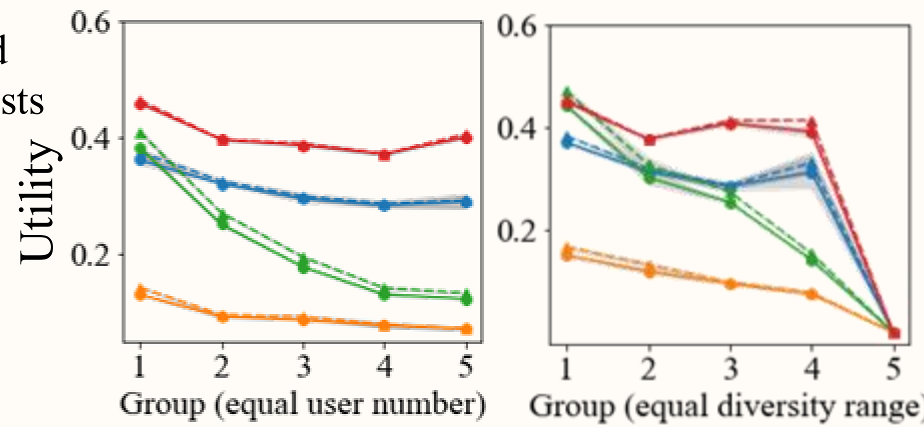




(A) Across datasets & models



(B) Across diversity metrics



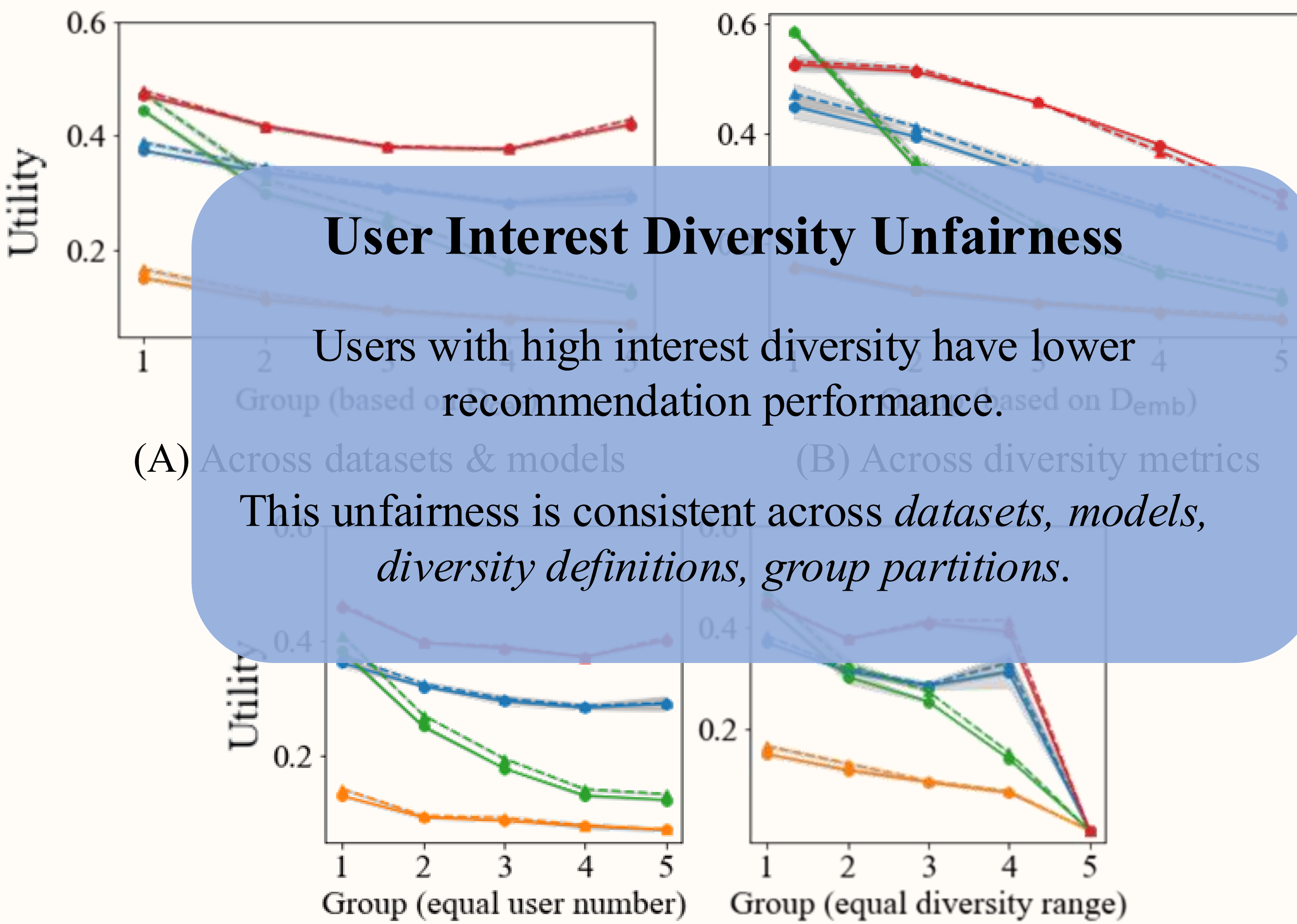
(C) Across group partitions

Different colors: various datasets
Solid and dashed lines: two RS backbones

User Interest Diversity Unfairness

Users with high interest diversity have lower recommendation performance.

This unfairness is consistent across *datasets, models, diversity definitions, group partitions*.

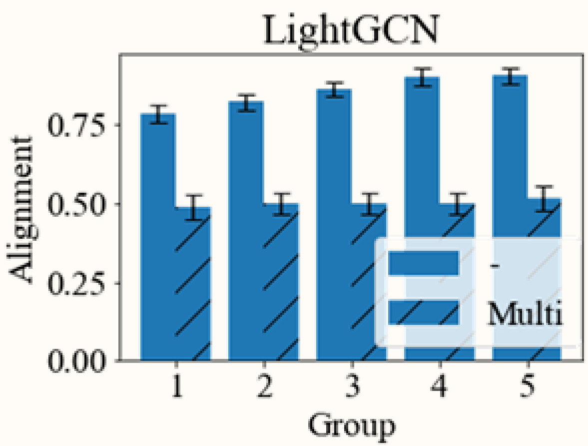


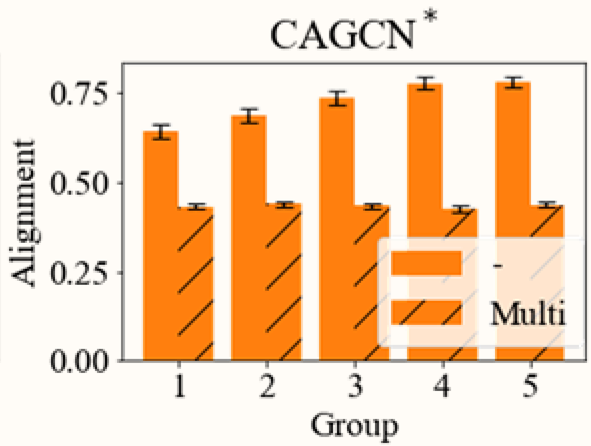
(A) Across datasets & models

(B) Across diversity metrics

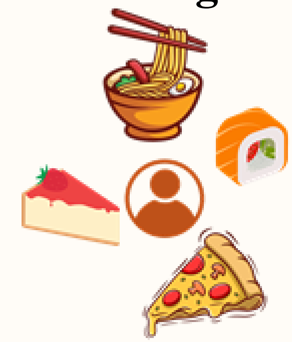
(C) Across group partitions

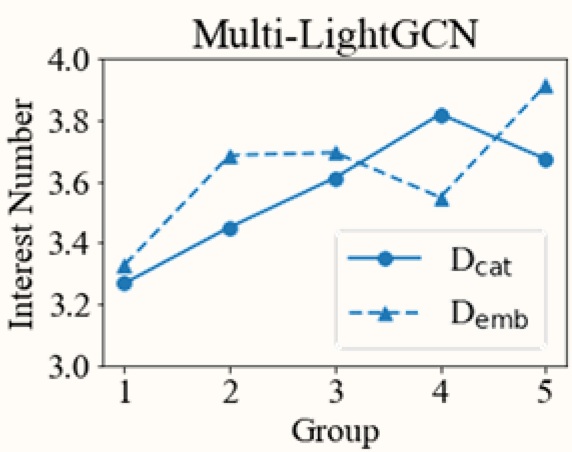
User Interest Diversity Fairness

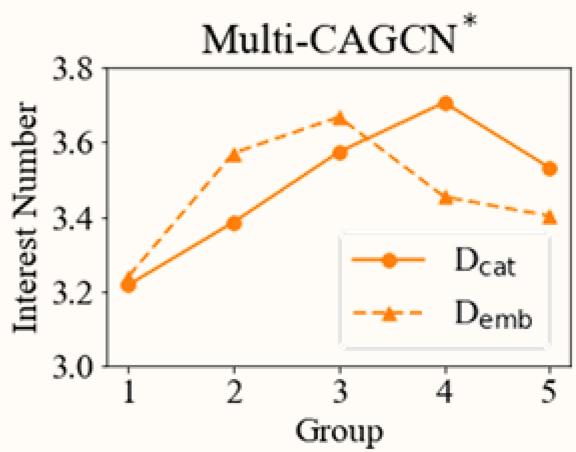




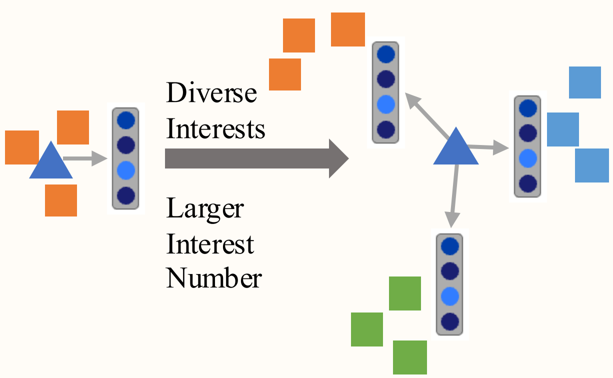
Better Alignment





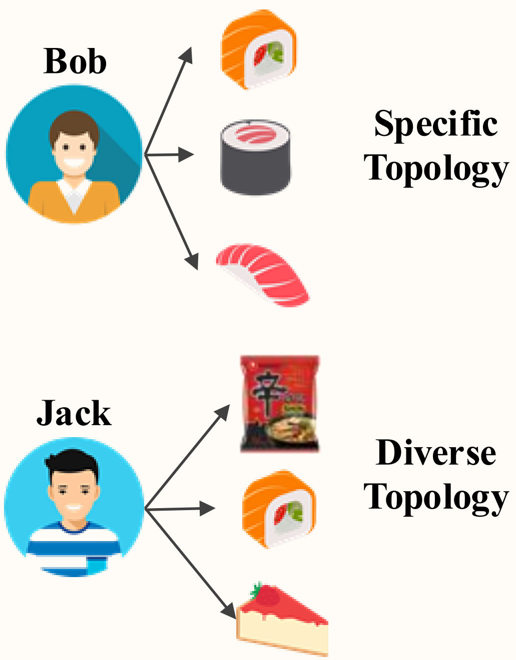


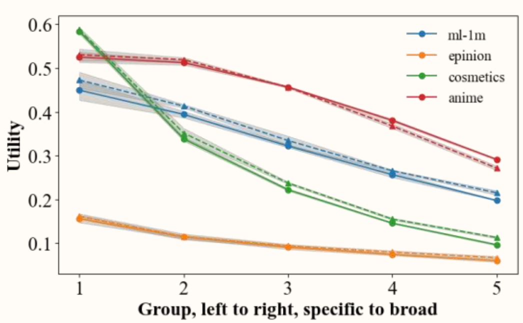
Implicit Interest Matching



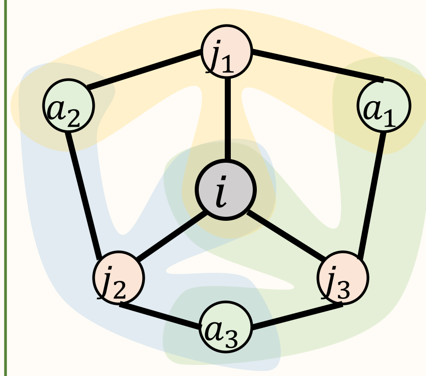
Diverse Topology Issue

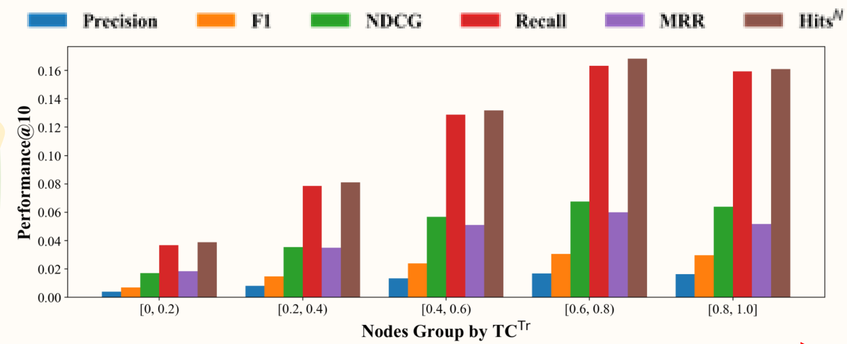
Problem





Metric

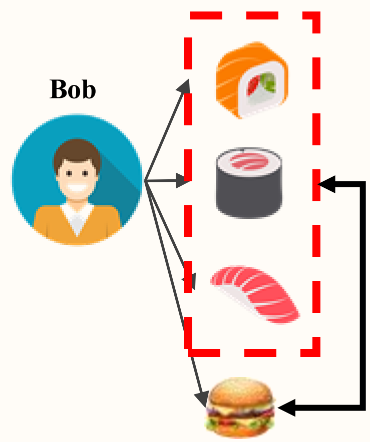




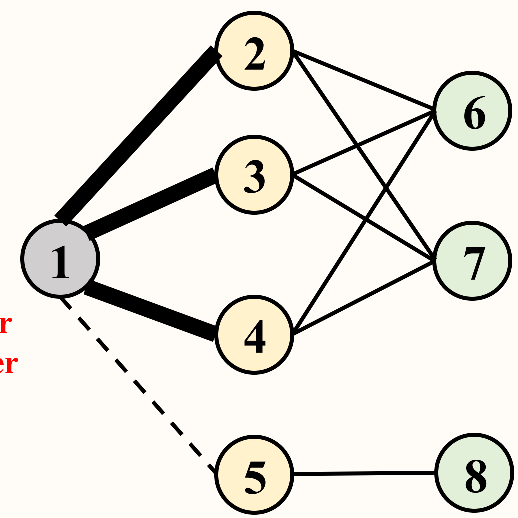
Less and Less Diverse, More and More Overlap

Our original version $\mathrm{TC}_i \rightarrow \mathrm{ATC}_i$ Proposed approximated version

Application

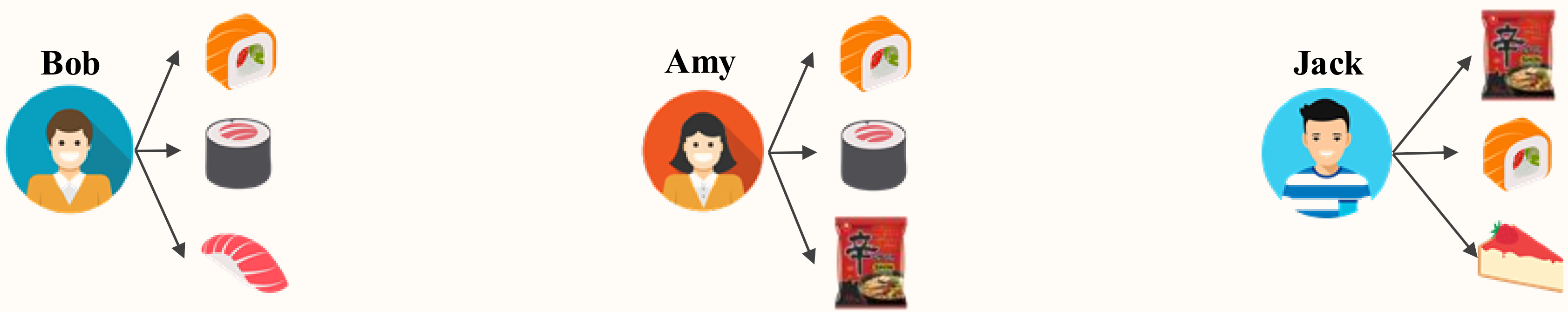


The burger is an outlier

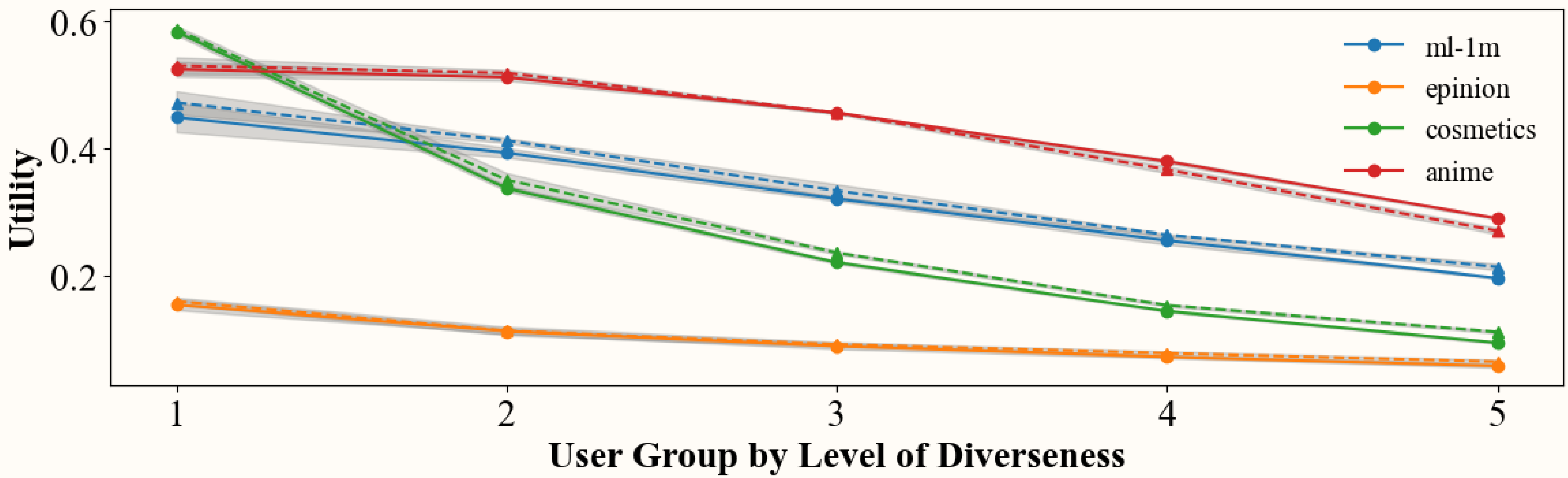


Diverse Topology - Motivation

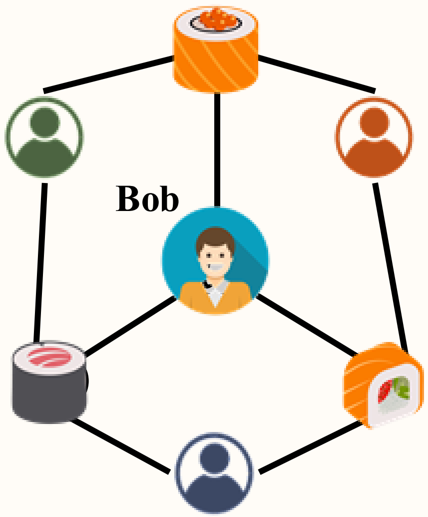
Specific → **Diverse**

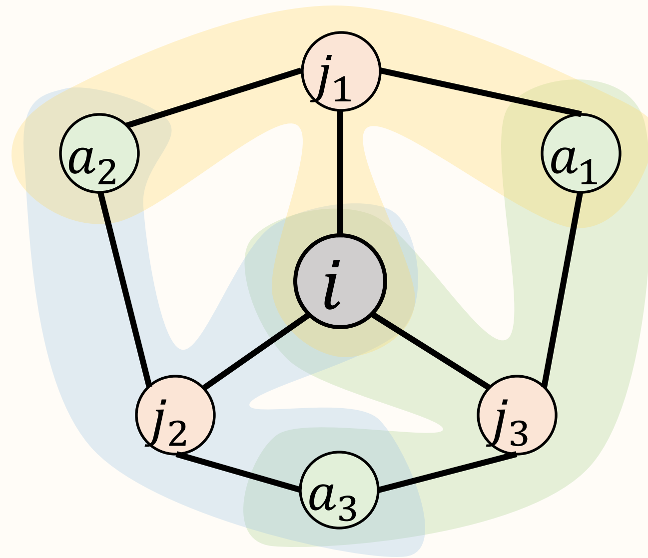


Sushi Lover | Sushi & Ramen Lover | Not Sure



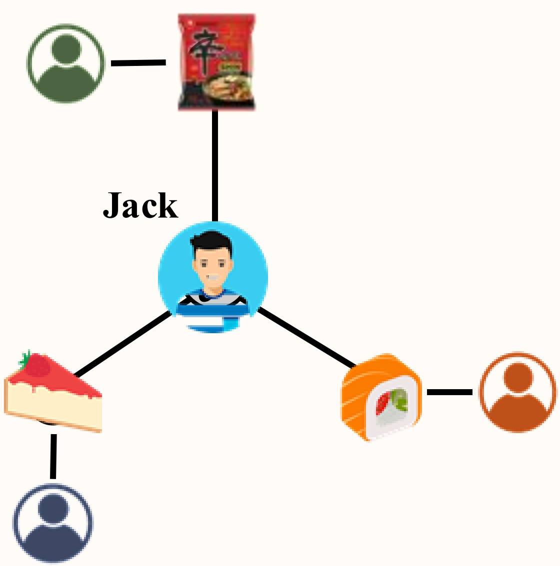
Diverse Topology - Quantification

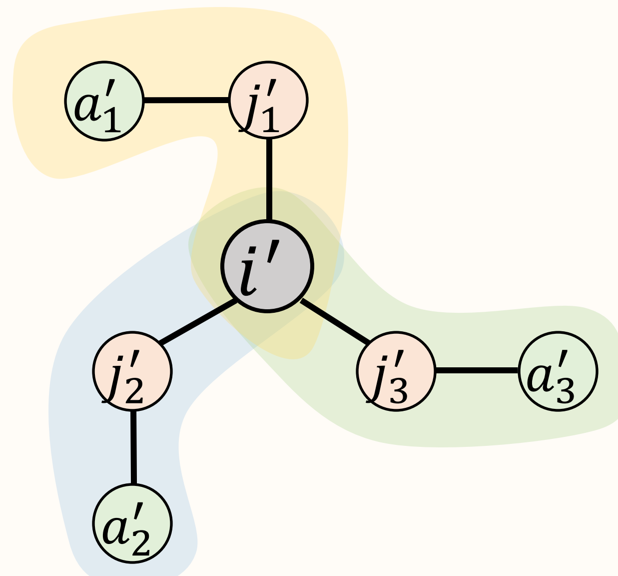




Low Diversity

High Overlap

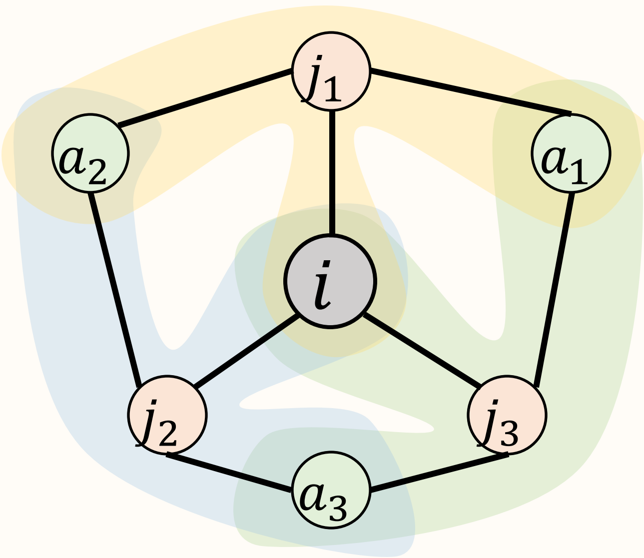




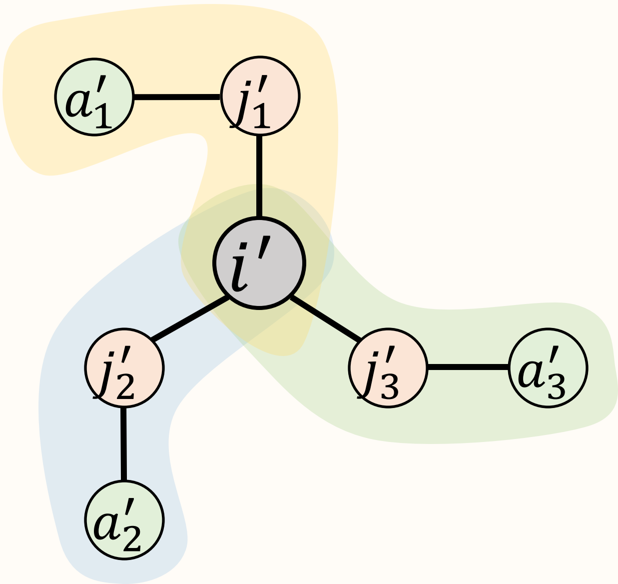
High Diversity

Low Overlap

Diverse Topology - Quantification



High Overlap



Low Overlap

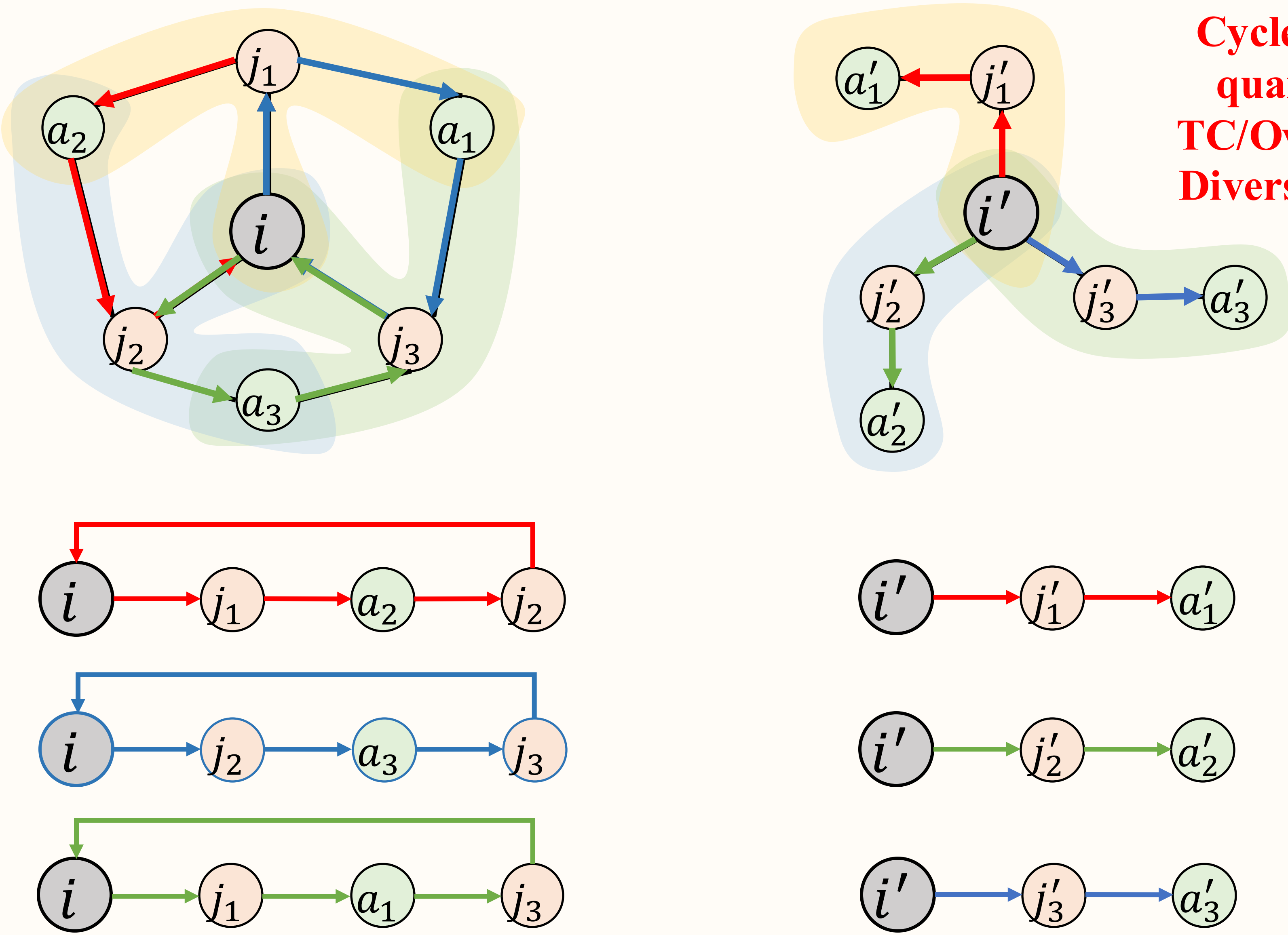
Can we mathematically measure this overlap?

TC_i
High

Topological Concentration (TC)

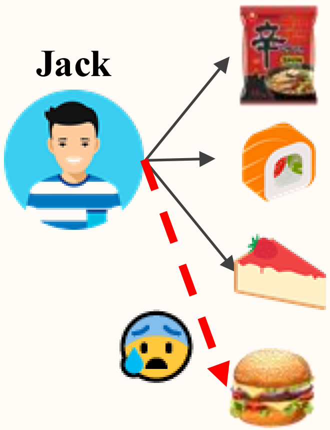
$\mathrm{TC}_{i'}$
Low

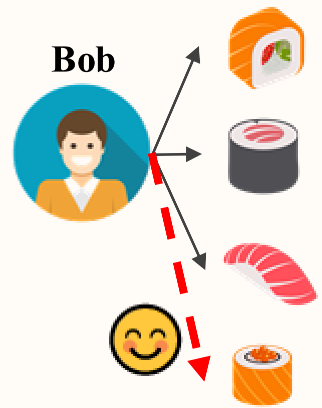
Diverse Topology - Quantification



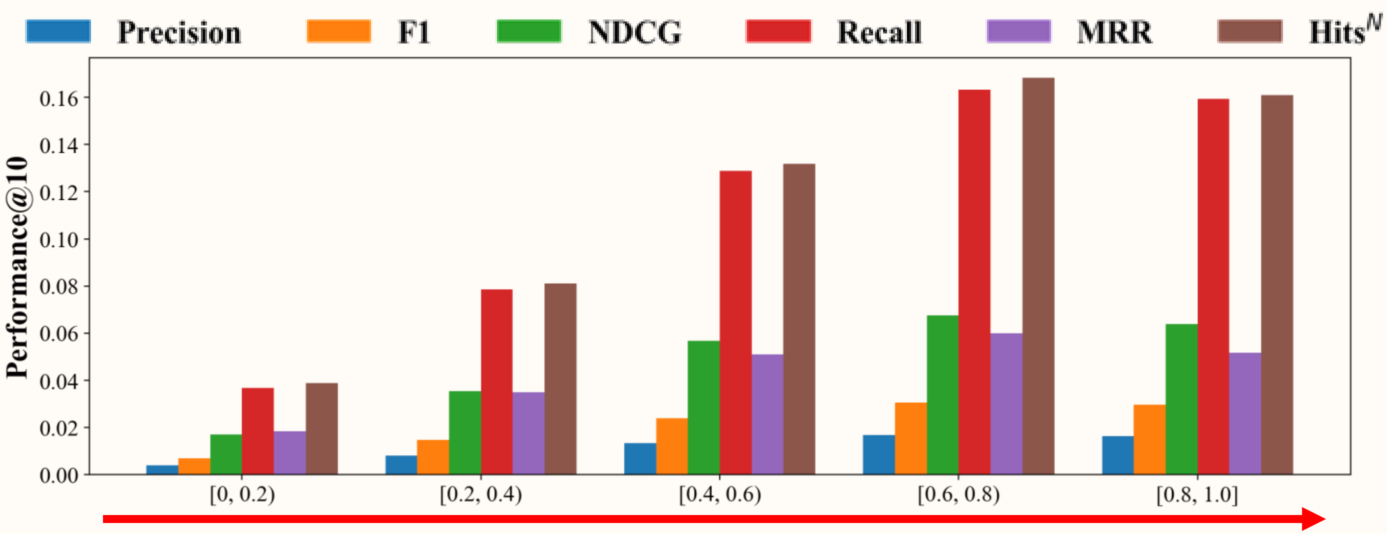
Cycles can quantify TC/Overlap/ Diverseness!

Diverse Topology - Analysis





Hard to predict preference of People with diverse interests

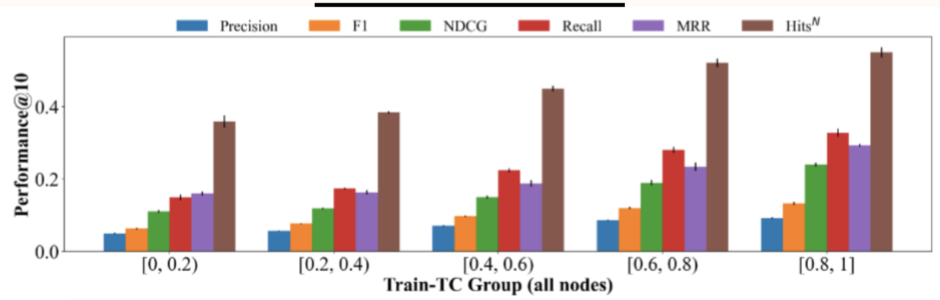


Higher TC, More Overlap, Less Diverseness

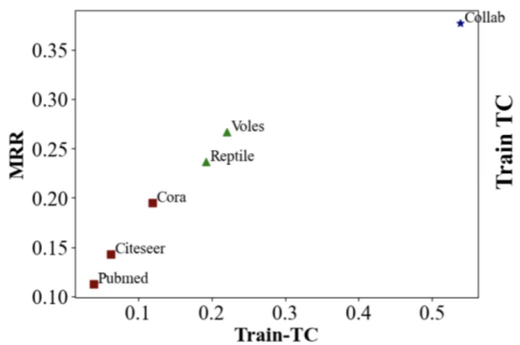
Nodes with diverse topology have worse recommendation performance

Diverse Topology - Analysis

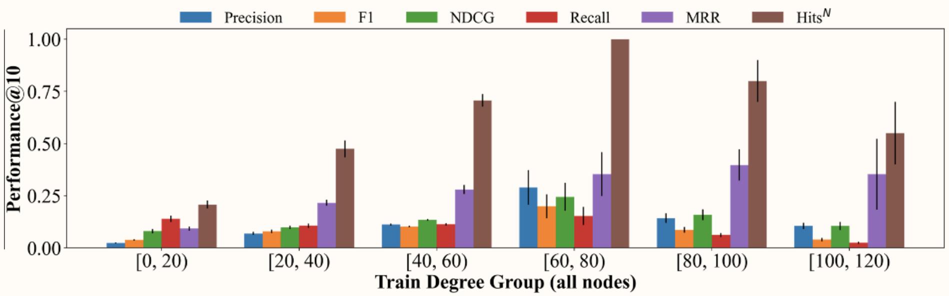
Within Network



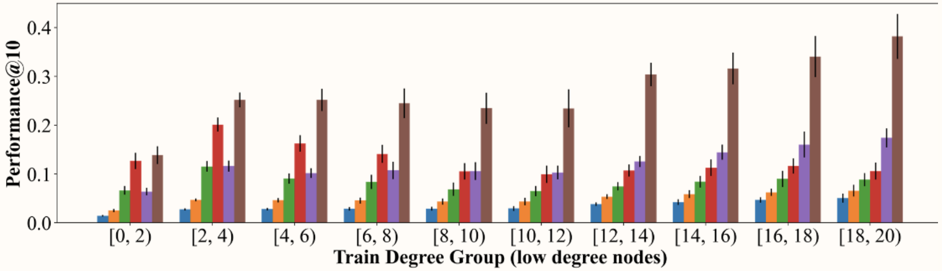
Across Datasets



Key insight: TC better defines node-centric LP difficulty than node degree

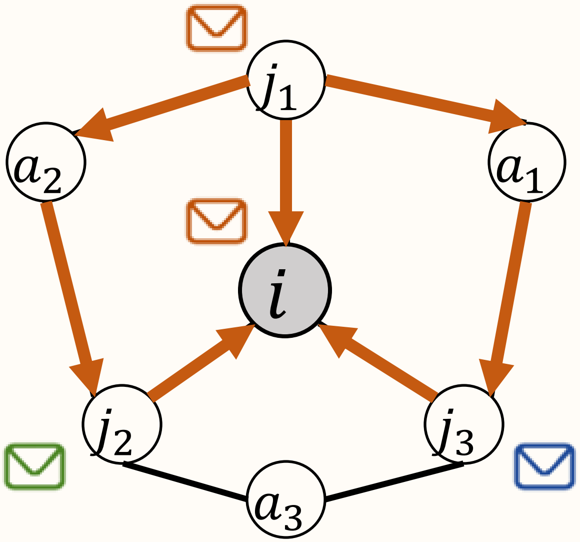


(c) Pubmed – All nodes



(d) Pubmed – Low degree nodes

Diverse Topology – Optimizing Computation



TC_i based on cycle counting

K: size of the cycle

$|\mathcal{V}|$: # of Nodes

Quadratic! $\mathcal{O}(K^2|\mathcal{V}||\mathcal{E}|)$ $|\mathcal{E}|$: # of Edges

$$\mathbf{R} \sim \mathcal{N}(\mathbf{0}^d, \boldsymbol{\Sigma}^d) \qquad \mathbf{N} = \sum_{k=1}^{K} \alpha_k \tilde{\mathbf{A}}^k \mathbf{R}$$

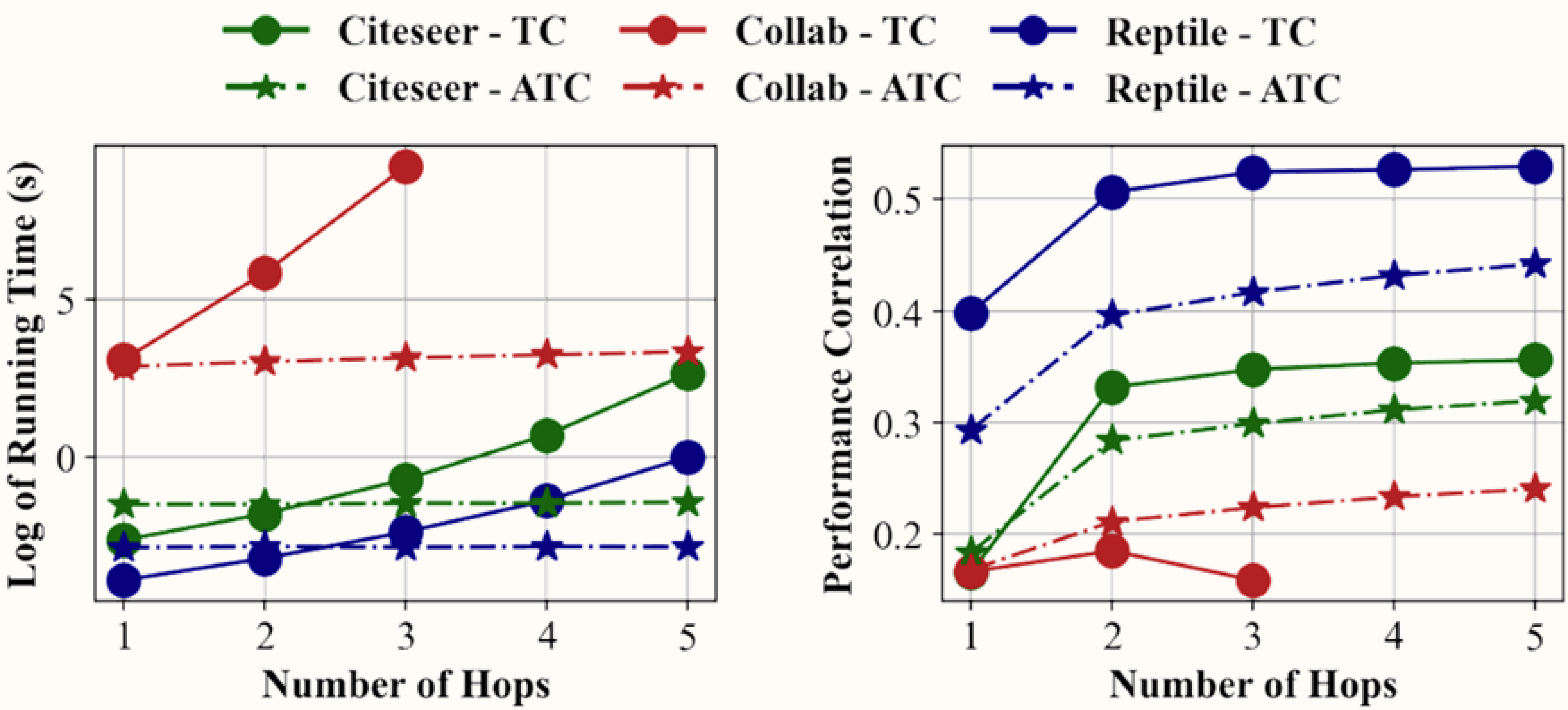
$$\text{ATC}_i = \mathbb{E}_{v_j \sim \mathcal{N}_i} \phi(\mathbf{N}_i, \mathbf{N}_j) \qquad \phi\text{: cosine similarity}$$

1. Initialize node embeddings from the d-dimensional Multivariate Gaussian Distribution

2. Perform message-passing

3. Embedding similarity computation: how much percentage of message does i receives?

Diverse Topology – Optimizing Computation



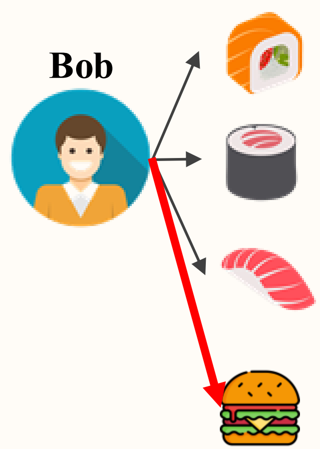
Running time of ATC is much shorter than TC

ATC still maintains a good correlation to performance!

Diverse Topology – Denoising

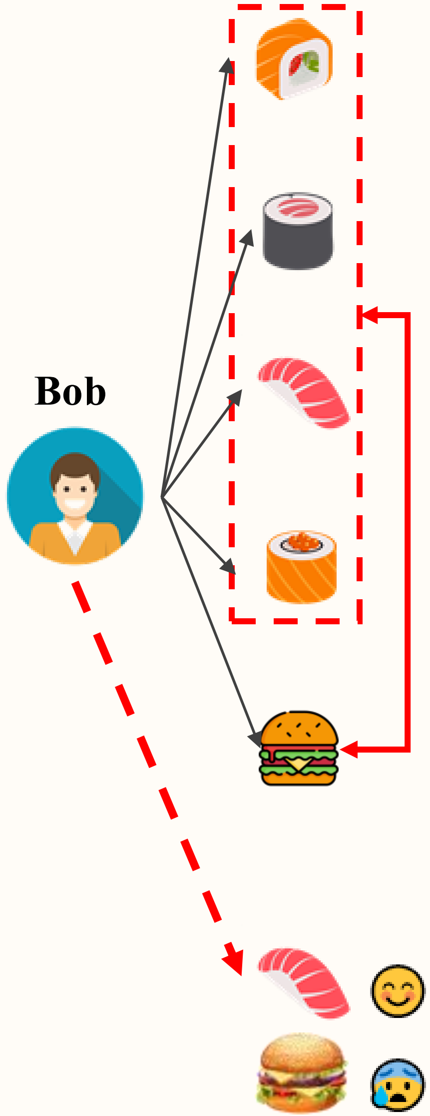
Across Different Users

Within one User



Maybe someday Bob orders a burger……

Diverse Topology – Denoising



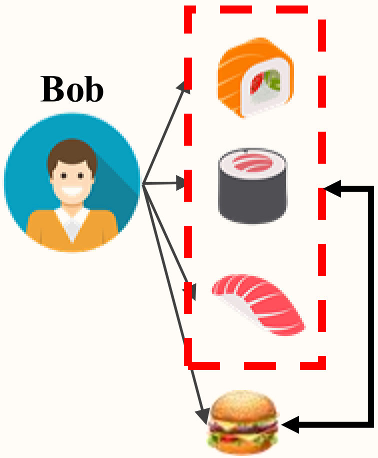
One day Bob bought a burger for his friend.

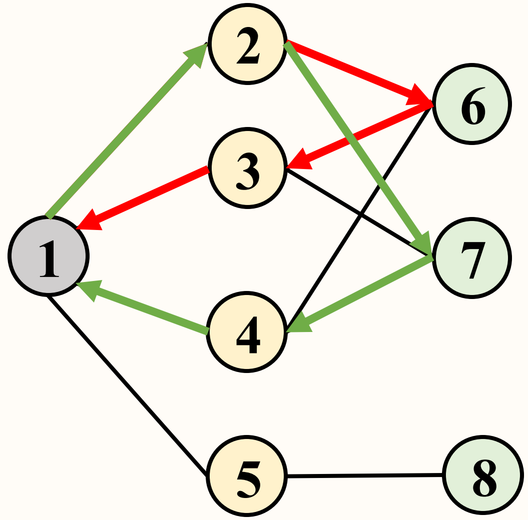
However, the burger cannot represent the eating behavior of Bob, as its an outlier of the whole neighborhood of Bob.

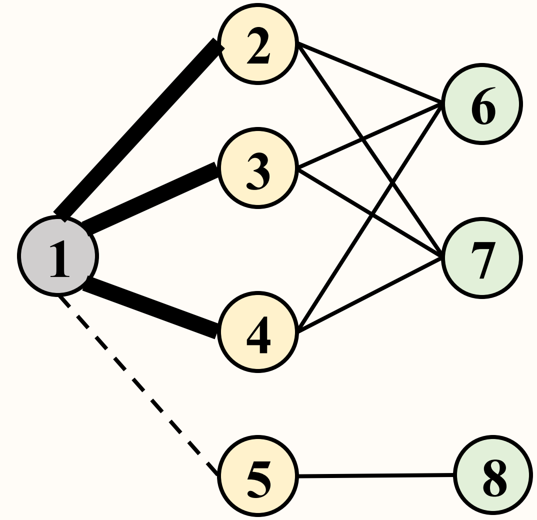
If Bob wants to order food using Uber Eats, it's highly likely he will order more sushi rather than a burger.

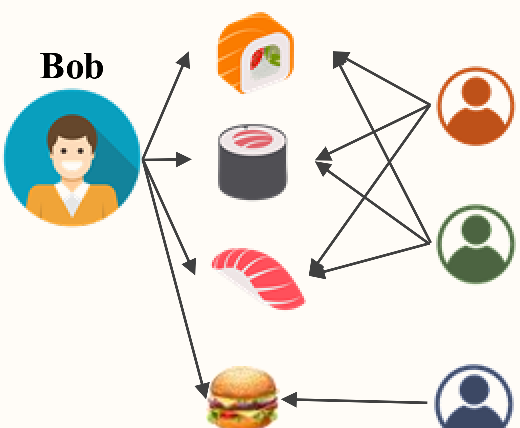
Therefore, adding this burger would diversify Bob's interest and it is a noisy interaction.

Diverse Topology – Denoising

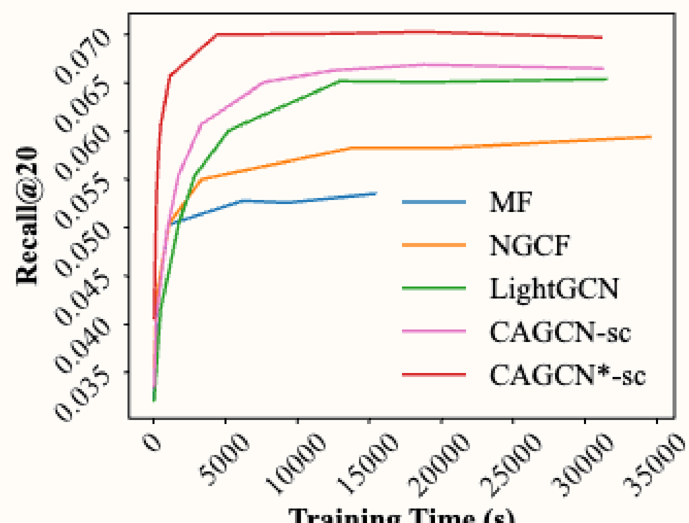




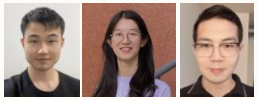




# of Cycles to decide Edge Weights

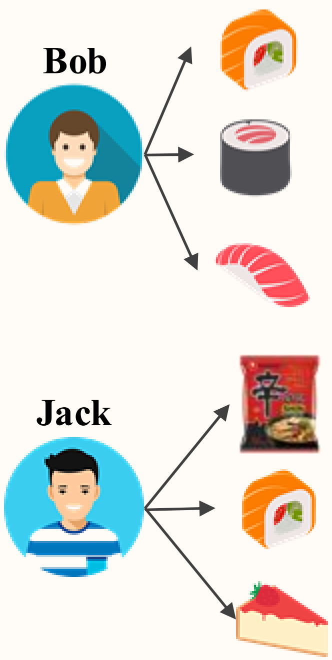


Up 10% in Recall@20
80% speedup



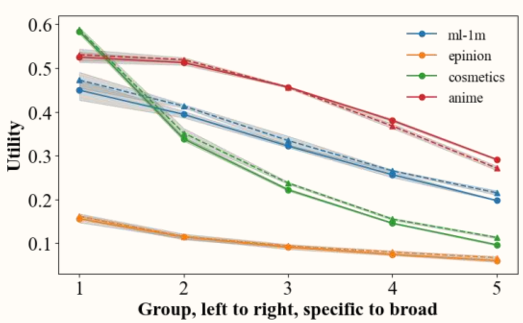
Diverse Topology – Summary

Problem

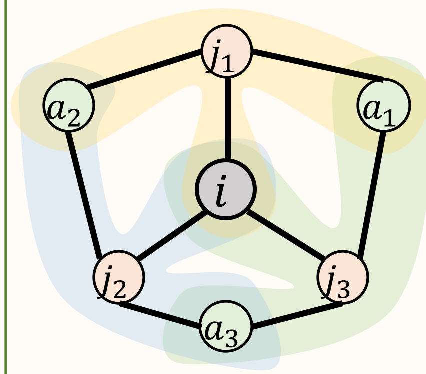


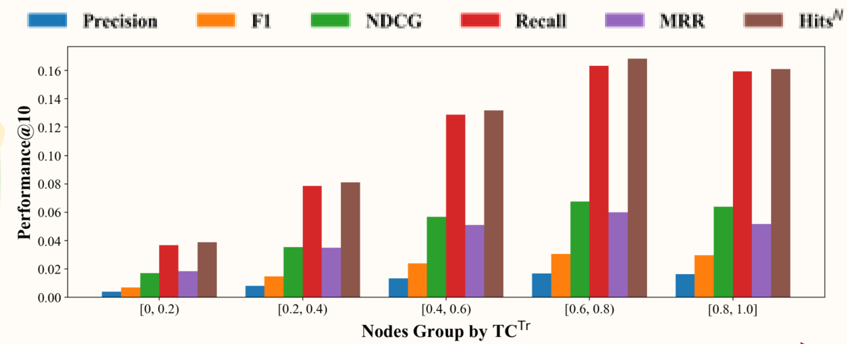
Specific Topology

Diverse Topology



Metric

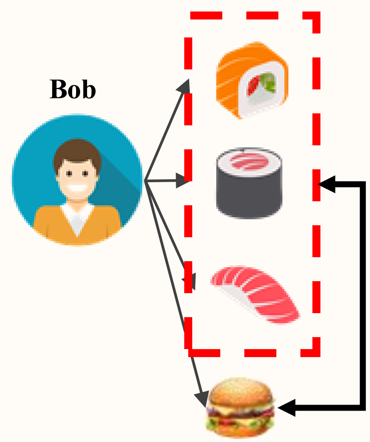




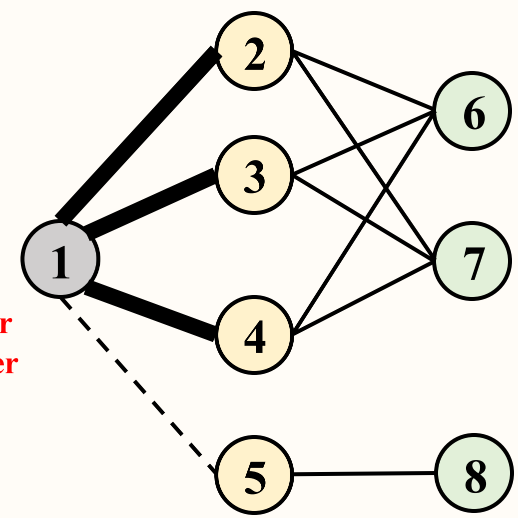
Less and Less Diverse, More and More Overlap

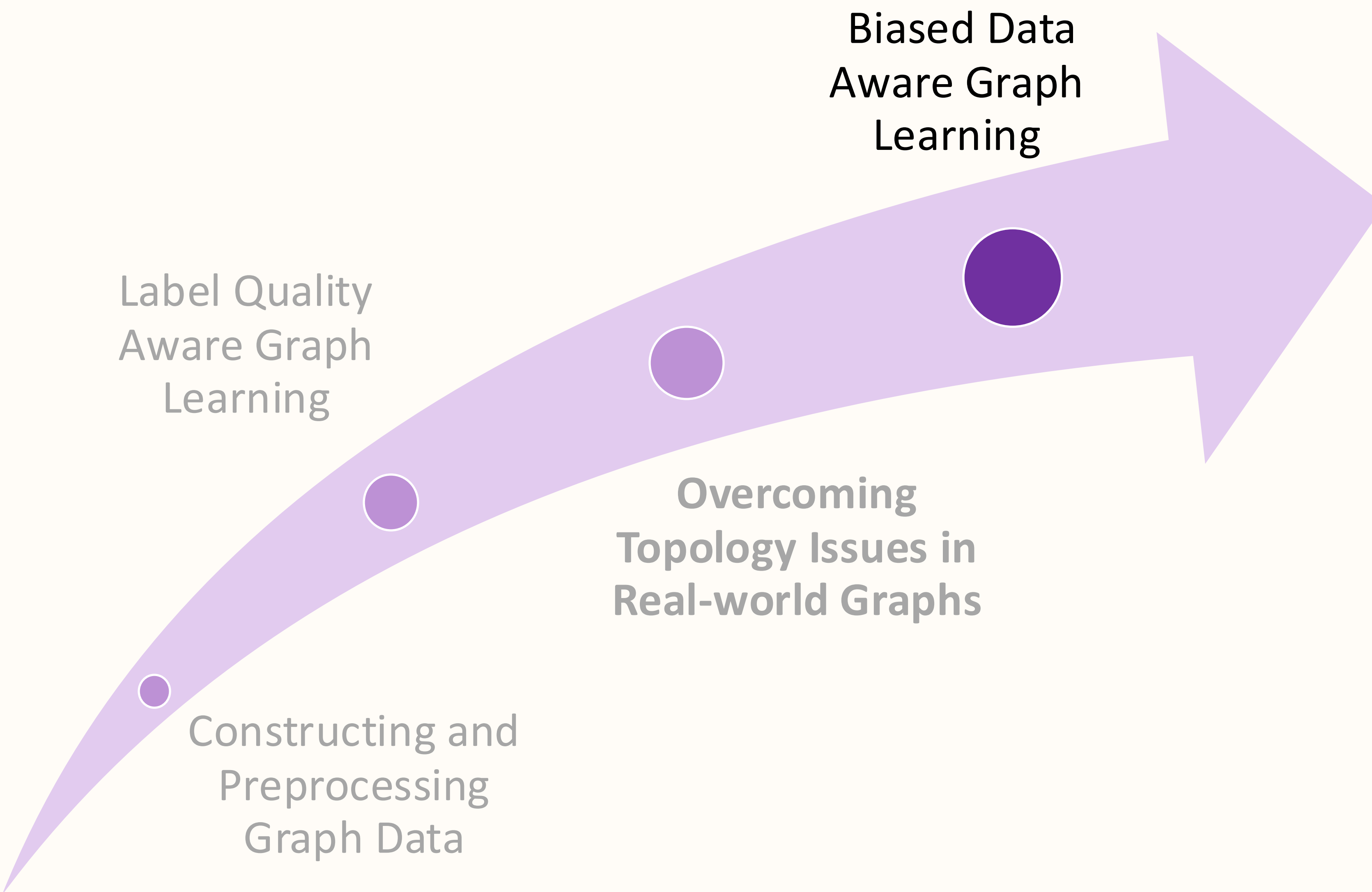
Our original version $\mathrm{TC}_i \rightarrow \mathrm{ATC}_i$ Proposed approximated version

Application



The burger is an outlier




Biased Data Aware Graph Learning
Label Quality Aware Graph Learning
Overcoming Topology Issues in Real-world Graphs
Constructing and Preprocessing Graph Data

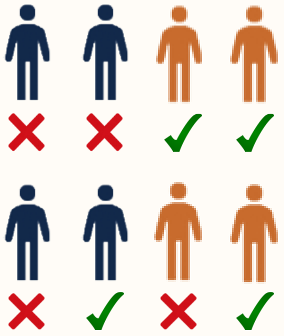
Potential Bias in Graph Data

Problem

✗ No-Bail
✓ Bail
Group 1
Group 2

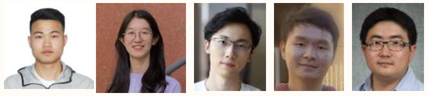
Statistical Parity

$$\Delta_{\text{sp}} = |P(\hat{y} = 1|s = 0) - P(\hat{y} = 1|s = 1)|$$

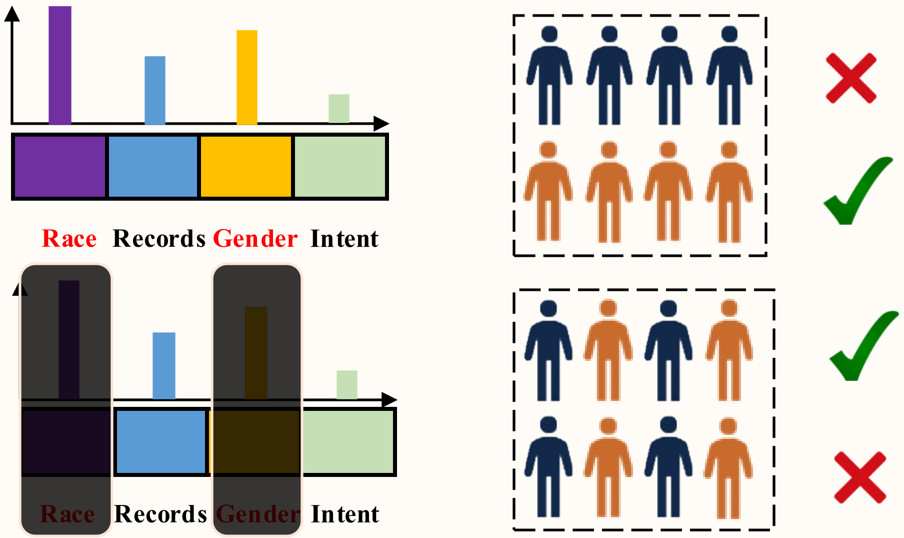


Biased Decision

Fair Decision

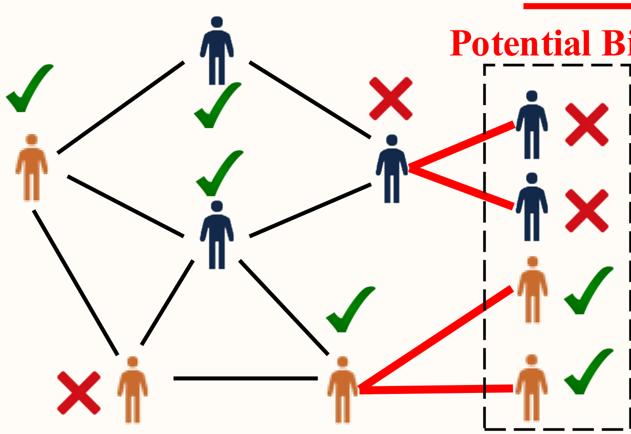


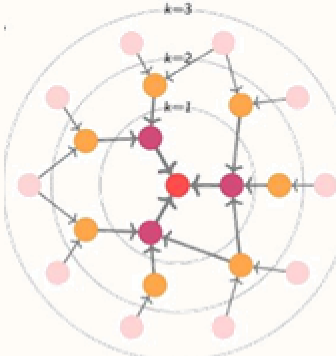
Discriminative Feature Bias



Social Interaction Bias

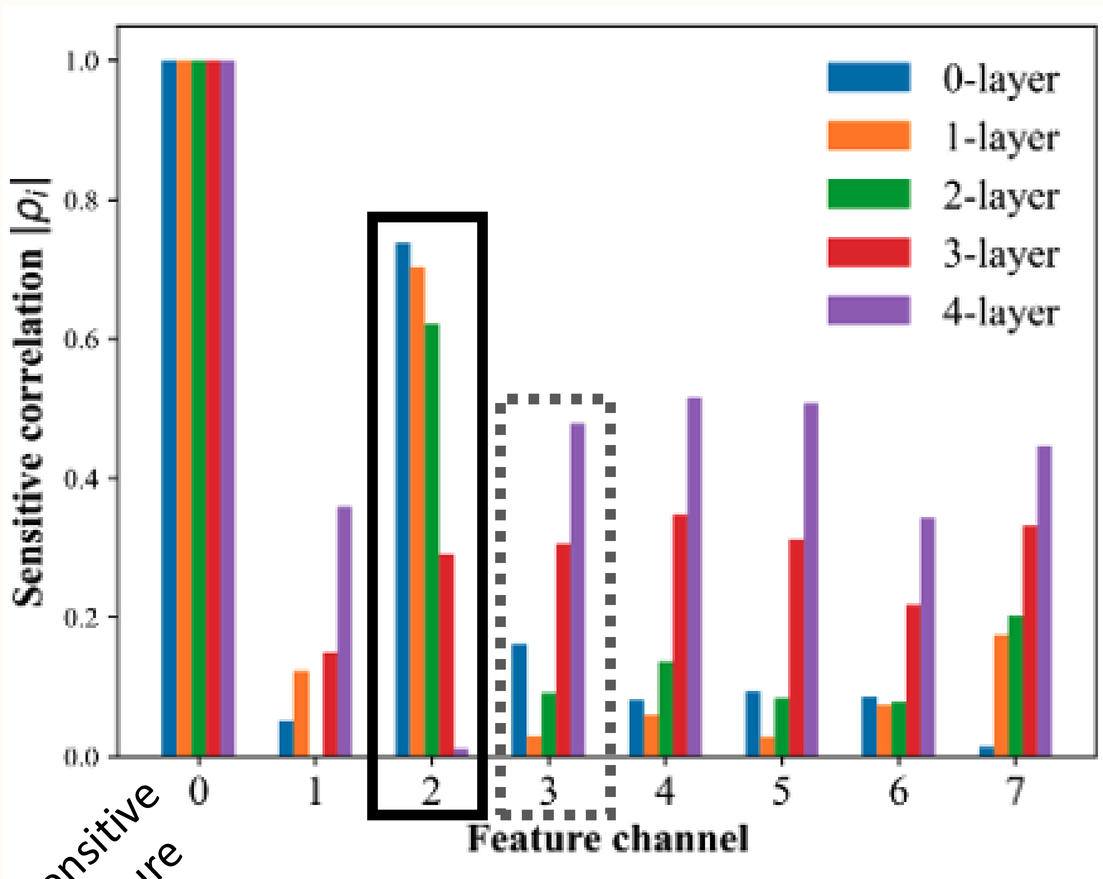
Criminal Associate Network



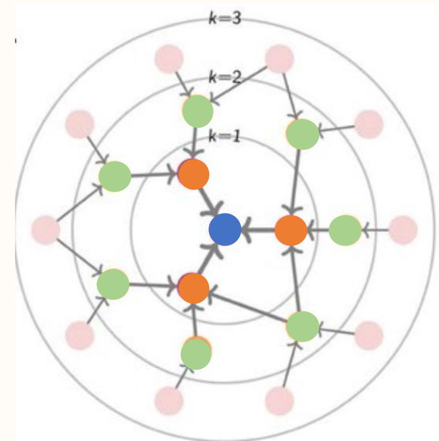


Feature Correlation Variation

Motivation



Sensitive Feature



Feature aggregation can cause feature correlation variation

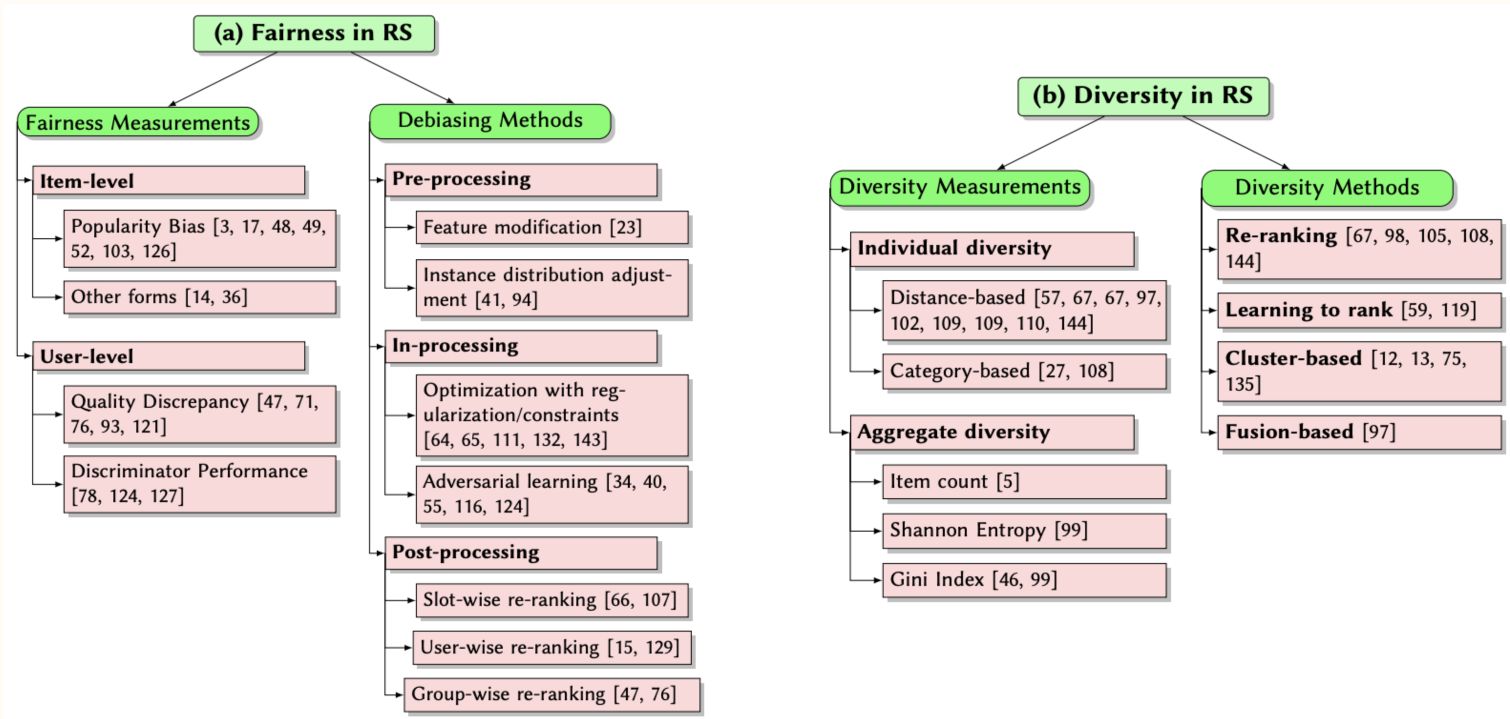
Feature 2: continuous decrease

Feature 3: decrease then increase

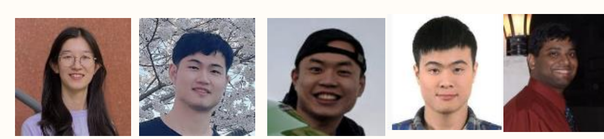
⋮

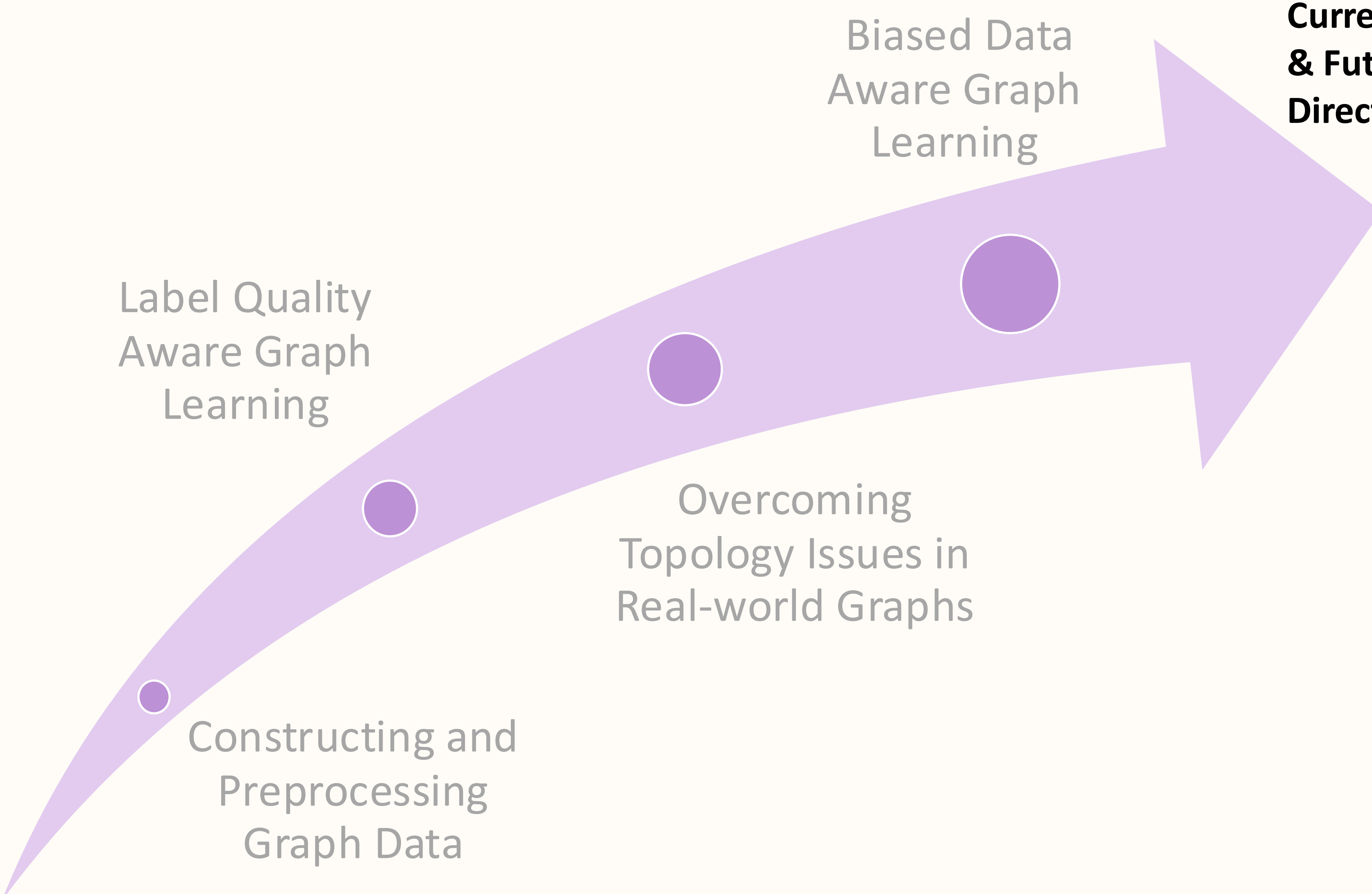
Fairness and Diversity in Online Recommendations

https://github.com/NDS-VU/Fair-Online-Dating-Recommendation



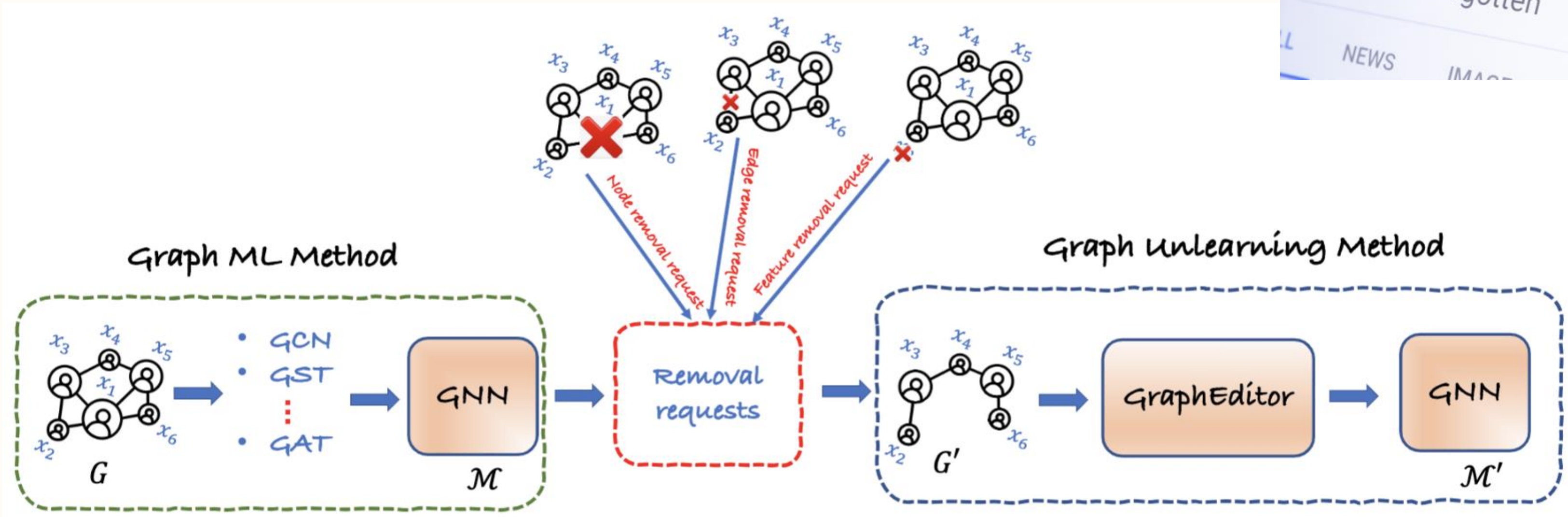
- Multi-objective scenarios
- Personalized sensitive attributes
- …

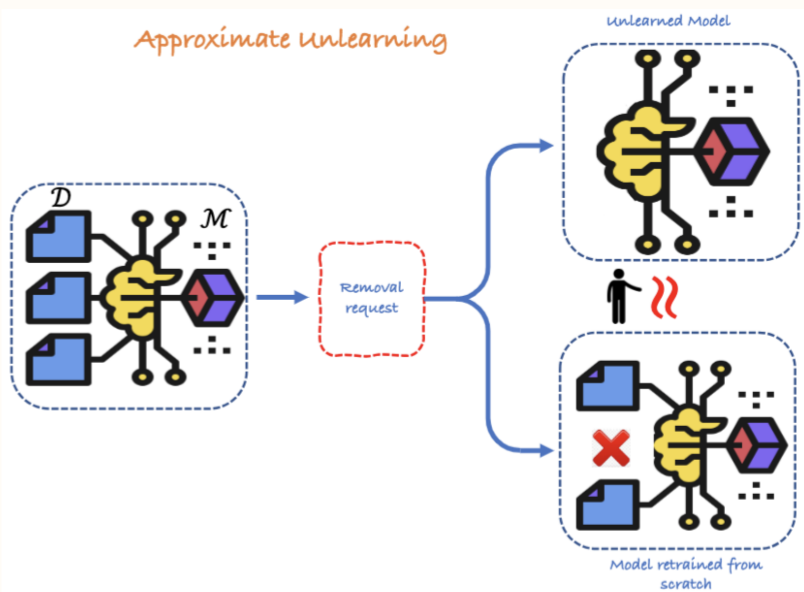




Current & Future Directions

Graph Machine Unlearning

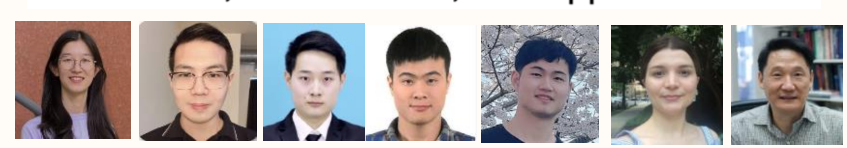




How effective can adversaries leverage unlearning tactics within online social media?

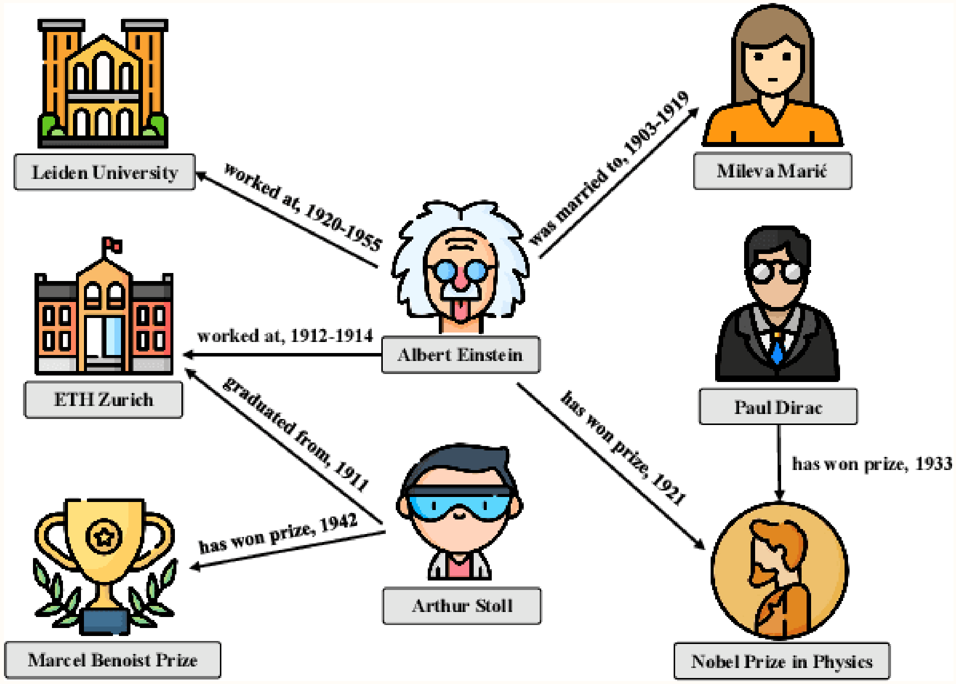
Fairness in machine unlearning…

A Survey on Privacy in Graph Neural Networks: Attacks, Preservation, and Applications



Temporal Knowledge Graphs

Most work on KGs focus on completion via link prediction



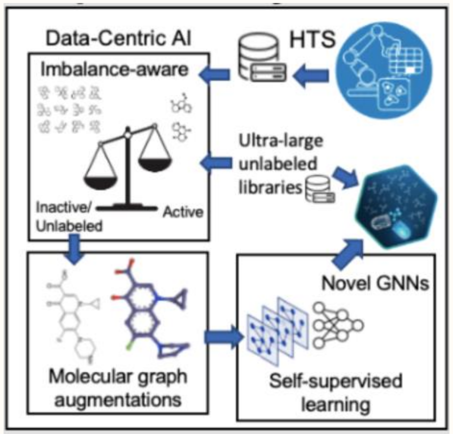
However, …

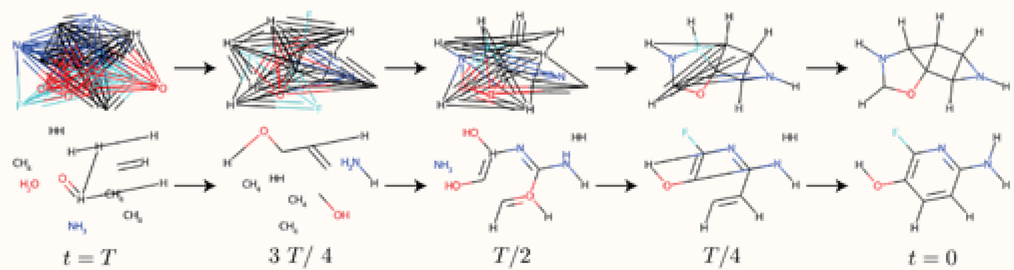
- Some facts/relations inherently have a limited lifetime
- KG quality is not always perfect and may require unlearning

… and working on linkages with LLMs

Generative Graph Models for Science

From large virtual screening to direct molecular generation

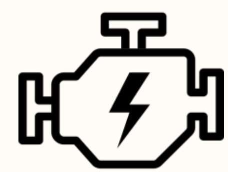




Foundational Graph Generator

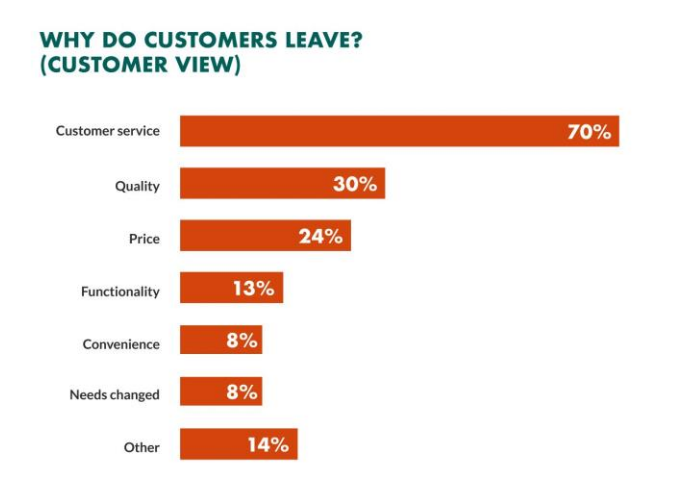
Data & Network Collections. Find and interactively VISUALIZE and EXPLORE hundreds of network data

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



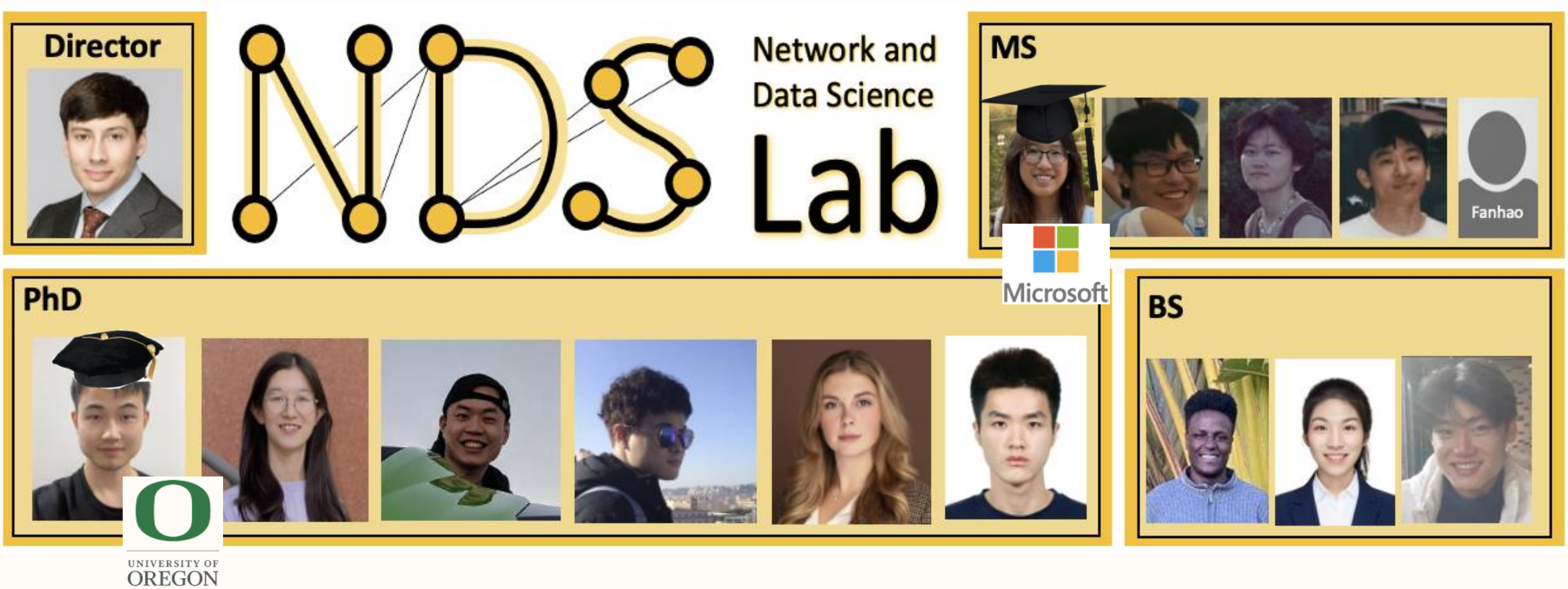
Graph Diffusion

Minimizing User Churn in Online Platforms

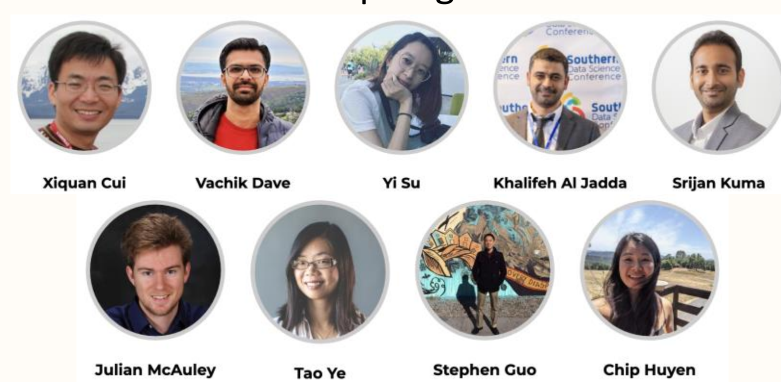


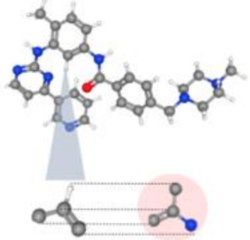
Acknowledgements

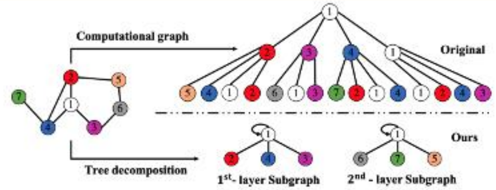
Thank you!

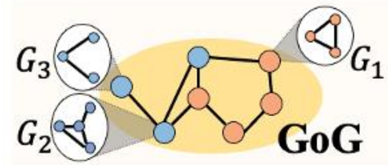


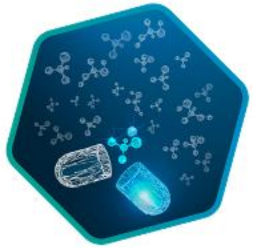
Special thanks to the CIKM’24 OARS Workshop Organizers!

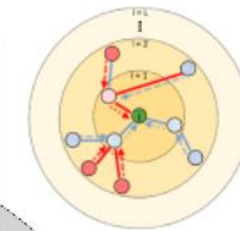








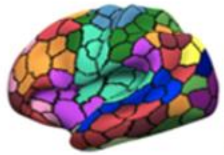




Deep Learning on Graphs

- Graph Neural Networks
- Self-supervised Learning
- Knowledge Graphs
- Data Quality Issues

...



Interdisciplinary Social Good Applications

- Drug Discovery
- Political Science
- Education
- Neurodiversity

...

NDS

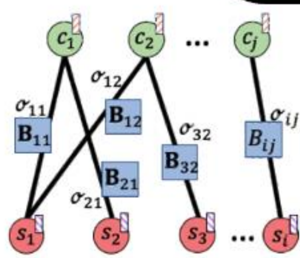
Network and Data Science Lab at Vanderbilt University

http://nds-vu.github.io

Responsible and Trustworthy AI

- Bias and Fairness
- Explainability
- Data Privacy & Machine Unlearning

...



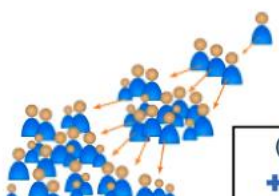
Social Network Analysis & **Recommender Systems**

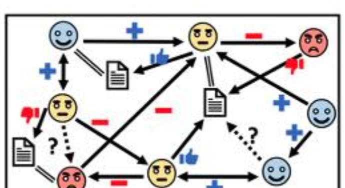
- Network Models
- Measurements
- Social Theories
- Applications

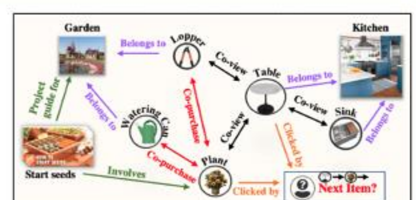
...

- Cold-Start Problem
- Diversity
- Session-Based
- Topological Analysis

...







Acknowledgements

THE HOME DEPOT NVIDIA Snap Inc.

Thank you for supporting our research!

Welcome to join our virtual Vanderbilt Machine Learning Seminar Series every Monday afternoon. Details at:
http://vu.edu/ML

ML V SS