Graph Machine Learning and Motivations

"*Everything connects to everything else*" - Leonardo DaVinci

Jay Urbain, PhD - 9/27/2022

Why Graphs?

Graphs are a ubiquitous data structure and a universal language for describing complex systems.

In the most general view, a graph is simply a collection of objects (i.e., nodes), along with a set of interactions (i.e., edges) between pairs of these objects.

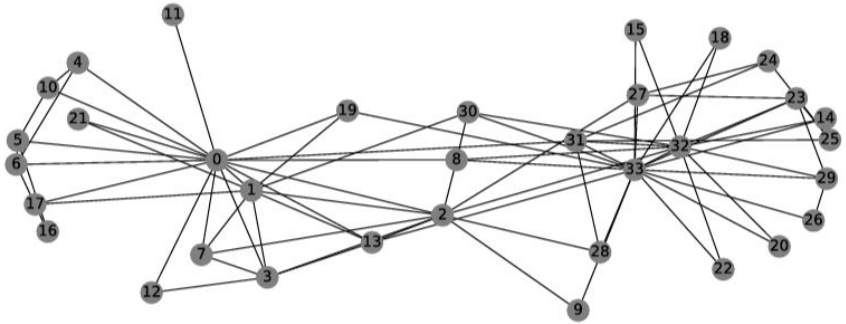
$$\mathcal{G} = (\mathcal{V}, \mathcal{E})$$

Many important real-world datasets can be represented as a graph of relationships between objects.

Graphs

Tool for modeling social networks, knowledge graphs, the Web, and biological systems such as protein-interaction networks.

For example, to encode a social network as a graph we might use nodes to represent individuals and use edges to represent that two individuals are friends

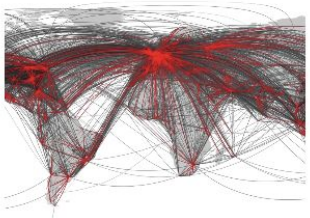


Zachary Karate Club Network represents the friendship relationships between members of a karate club studied by Wayne W. Zachary

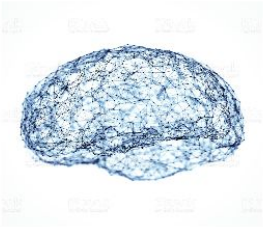
Data as graphs - explicit

Many interesting natural graph problems:

Social Graphs

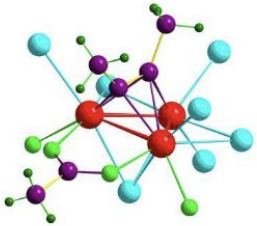


Transportation Graphs

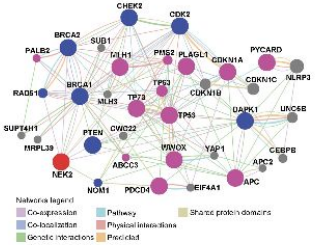


Brain Graphs

Web Graphs



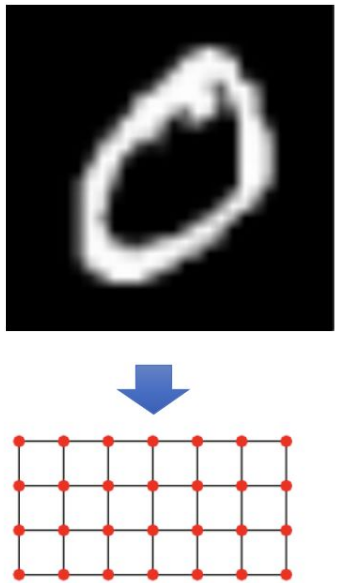
Molecular Graphs

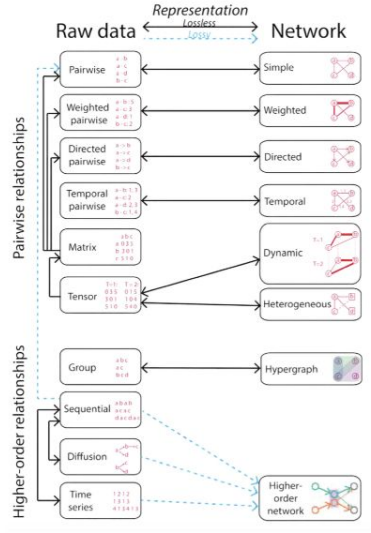


Gene Graphs

Data as graphs - implicit

Many problems use graphs implicitly:





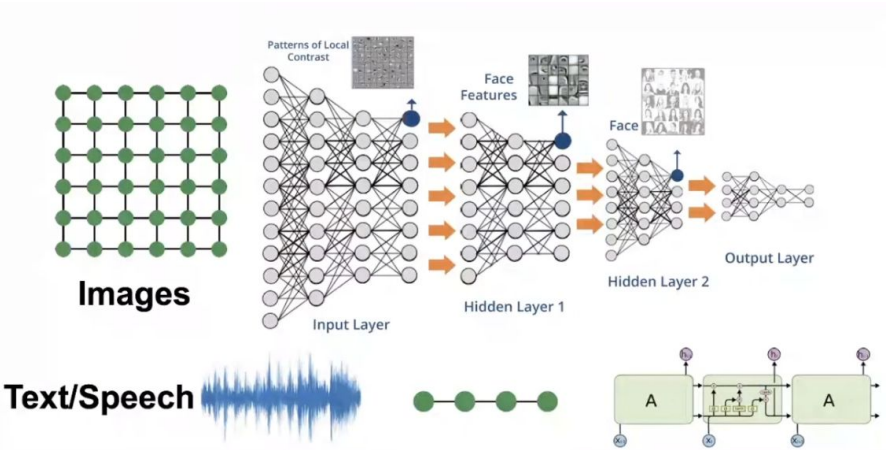
Jian Xu. Representing Big Data as Networks. PhD Dissertation, University of Notre Dame

Traditional ML methods

Traditional machine learning methods are designed for IID data, and are either tabular, simple sequences, or grids.

This simplification of data structure can introduce incorrect modeling bias.

	Total defects	A	B	C	D	E
A4636	131	37	21	28		45
A2524	86	20	24	21	1	20
A3713	75	17	13	18		27
A4452	73	5	33	17		18
A4088	72	14	16	12	2	28
A2103	68	14	13	14	1	26
A2156	68	16	13	19	2	18
A3681	66	12	16	9	1	28
A1366	50	11	15	12		12
A2610	39	5	7	12		15
Total	728	151	171	162	7	237



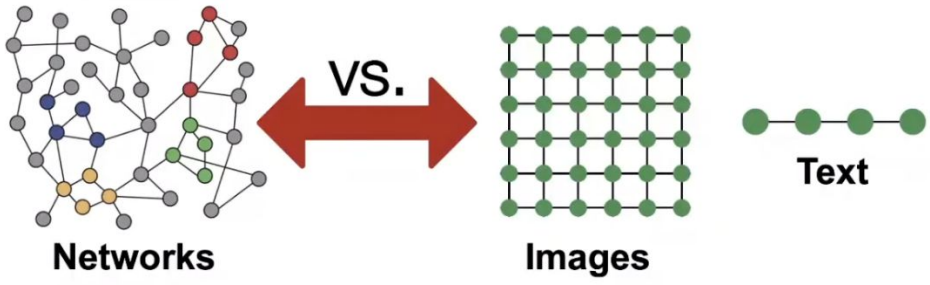
Why are graphs challenging for traditional ML methods?

Networks are complex.

Arbitrary size and complex topological structure - no spatial locality like sequences or grids.

No fixed node ordering or reference point.

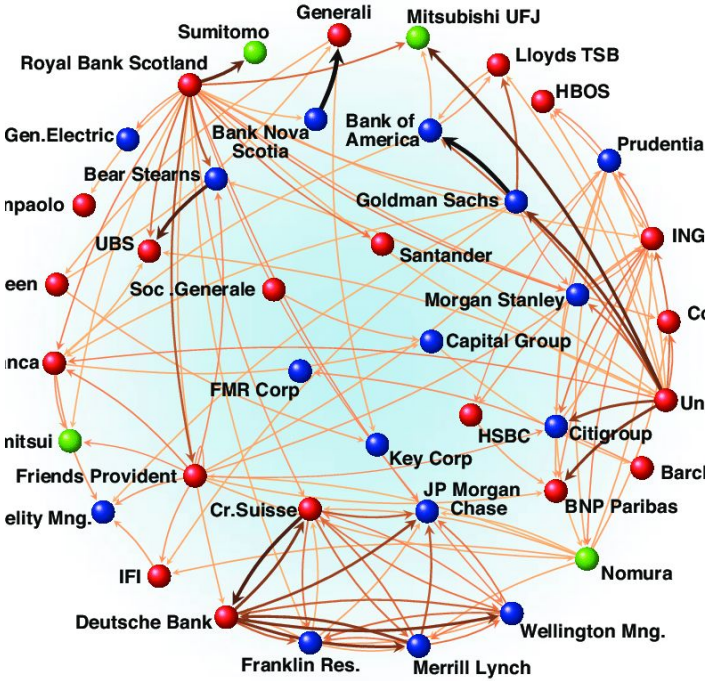
Often dynamic, and multimodal.



Economic Networks

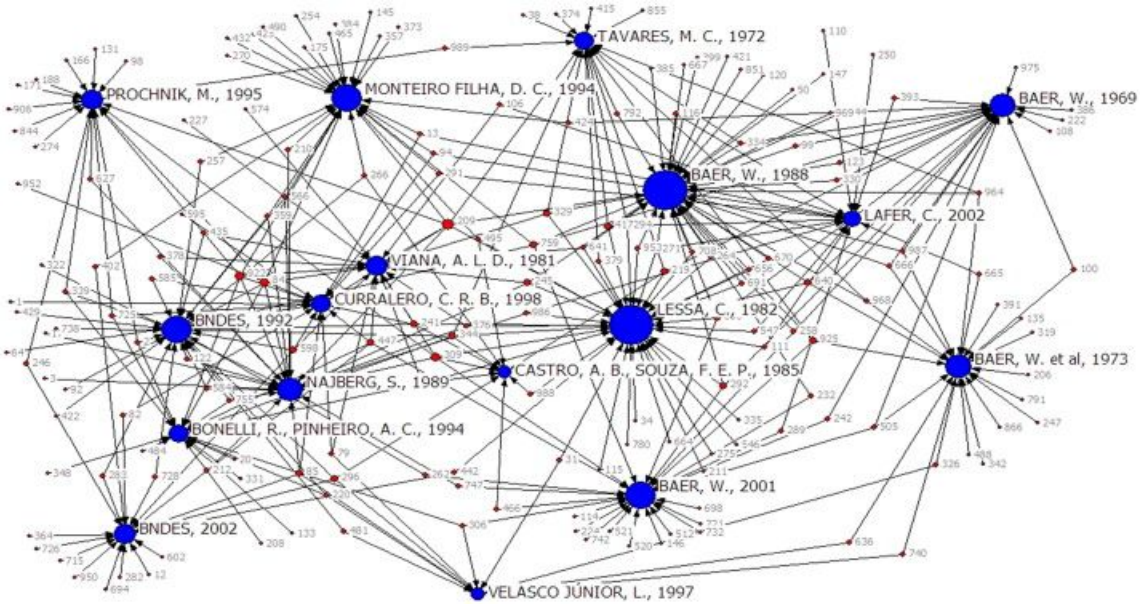
What are financial pathways that could cripple the world economy?

Are there anomalies?



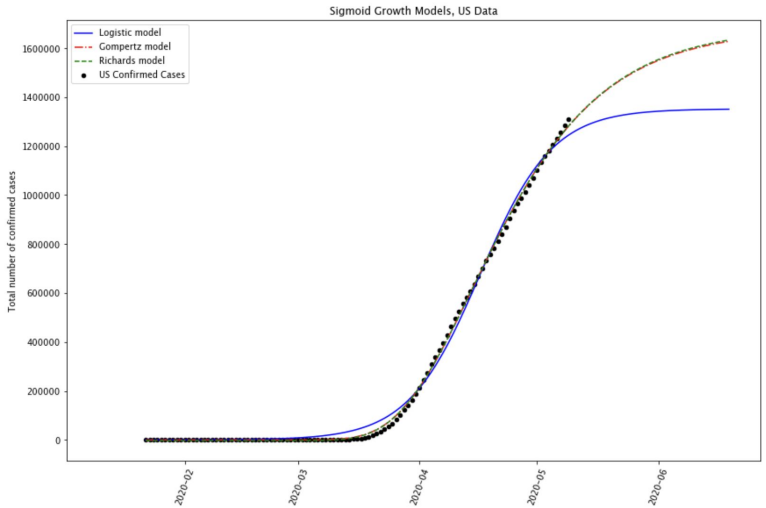
Citation Networks

Who has the cross disciplinary skills necessary to develop biological weapons?



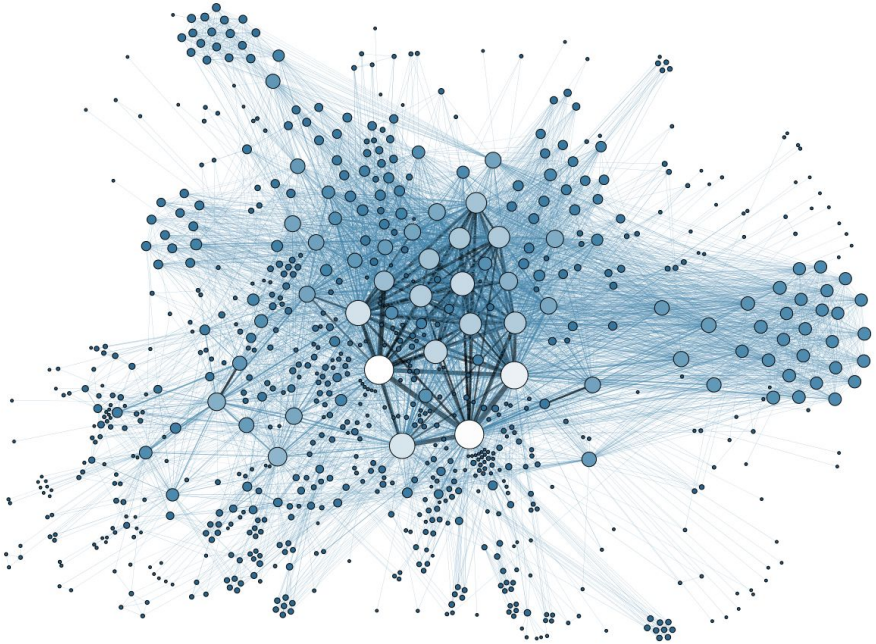
Viral spread

Can we build better models to track and predict viral spread?



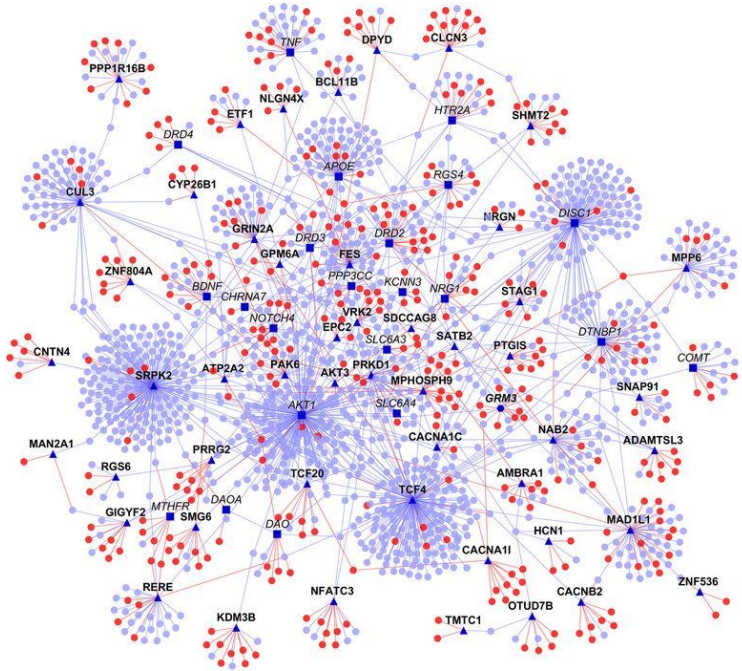
Social Networks

Who should you know? What should you watch? Where should you work? Who will you vote for? What will you buy? Who should you date?



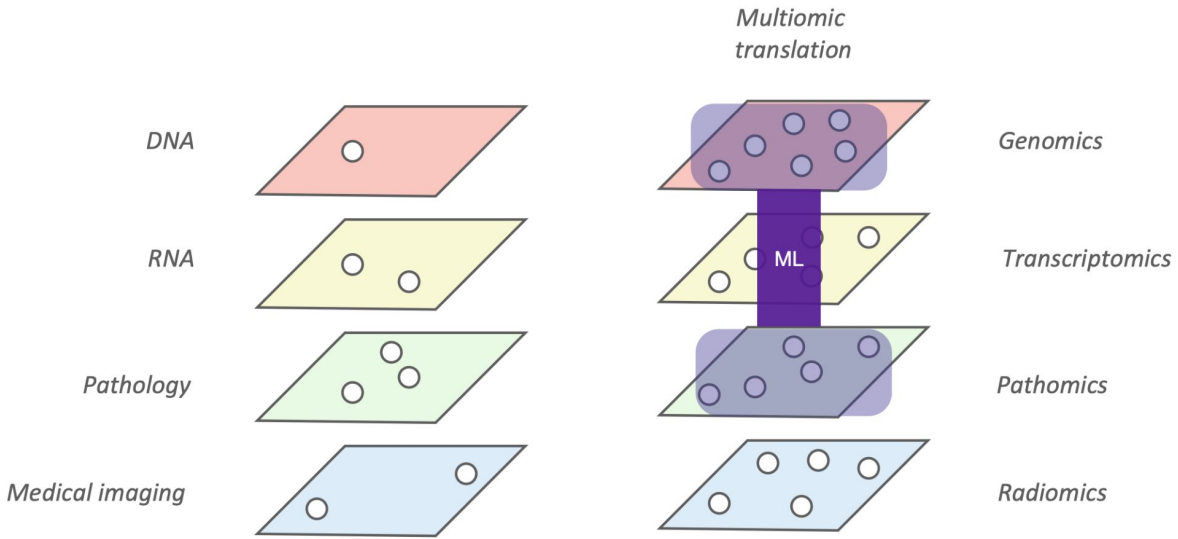
Schizophrenia protein-protein interaction network

Can we discover pathways for understanding and treatment?



Multi-omic translation networks

Can we learn relationships of complex multi-omic cancer translation networks?



Biological Knowledge Network

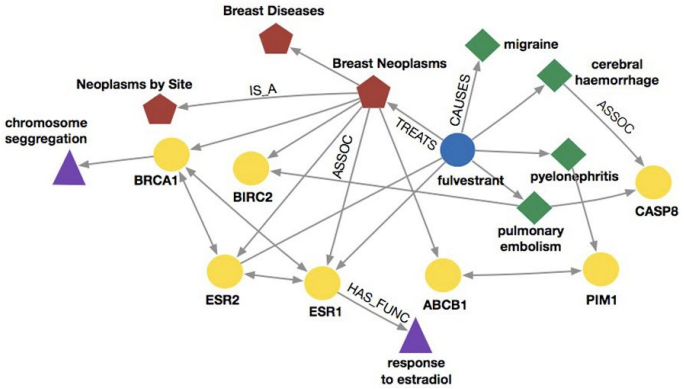
Big idea: Construct knowledge graph that models known biology and learn to reason over it.

Represent facts as triples (h, r, t)

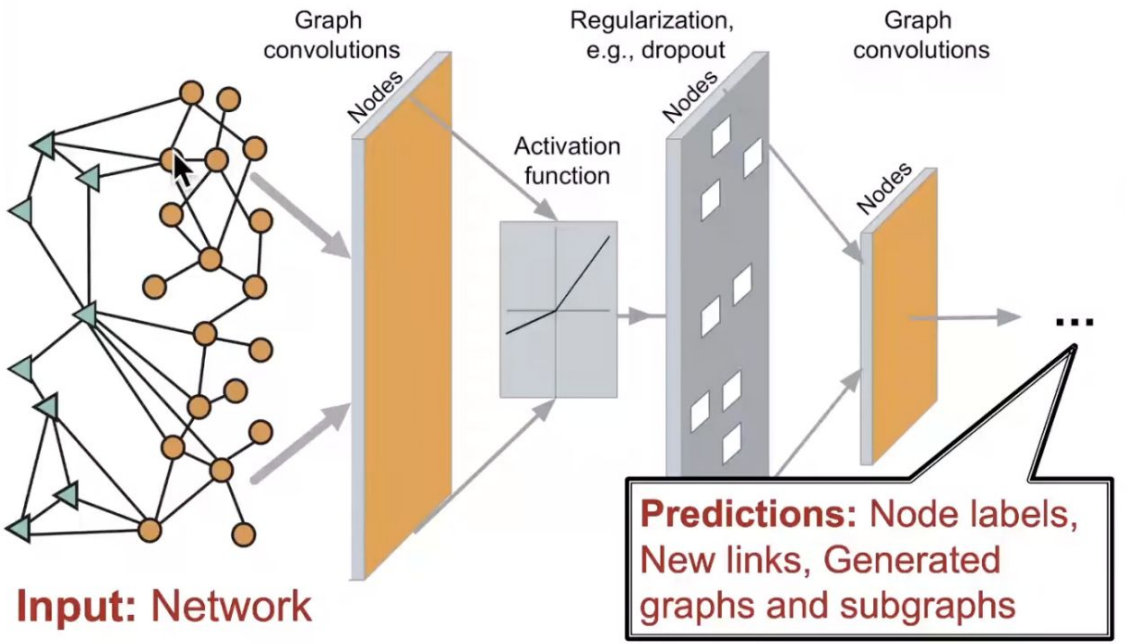
- ('BRCA1', 'associated_with', 'Breast_Neoplasms')
- ('Breast_Neoplasms', 'is_a', 'Breast_Disease')
- …

Node types: drug, disease, adverse event, protein, functions, …

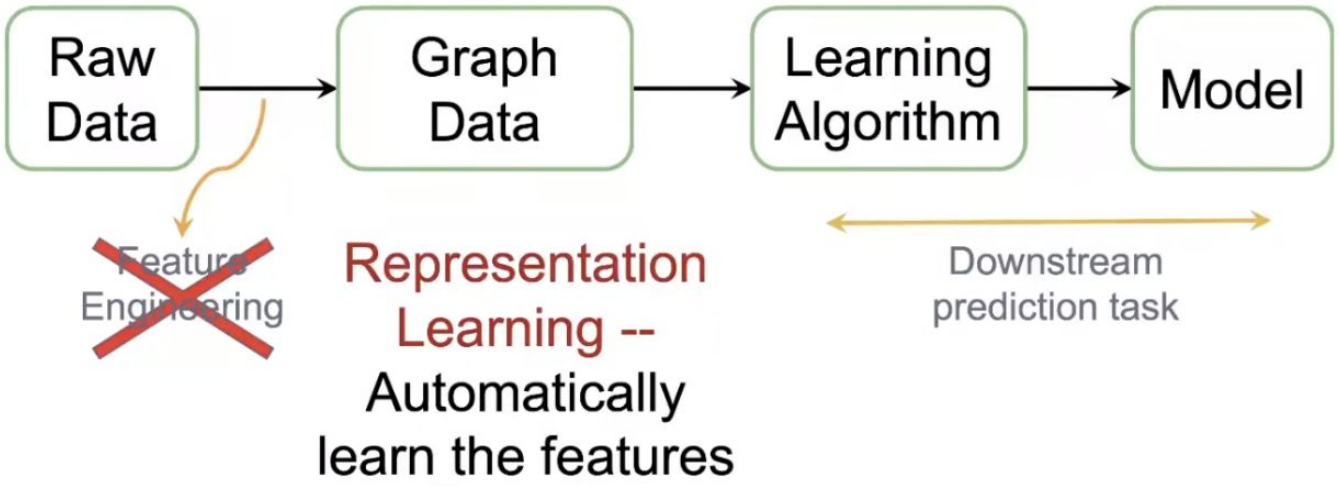
Relation types: causes, assoc, treat, interact, …



End-to-end deep learning



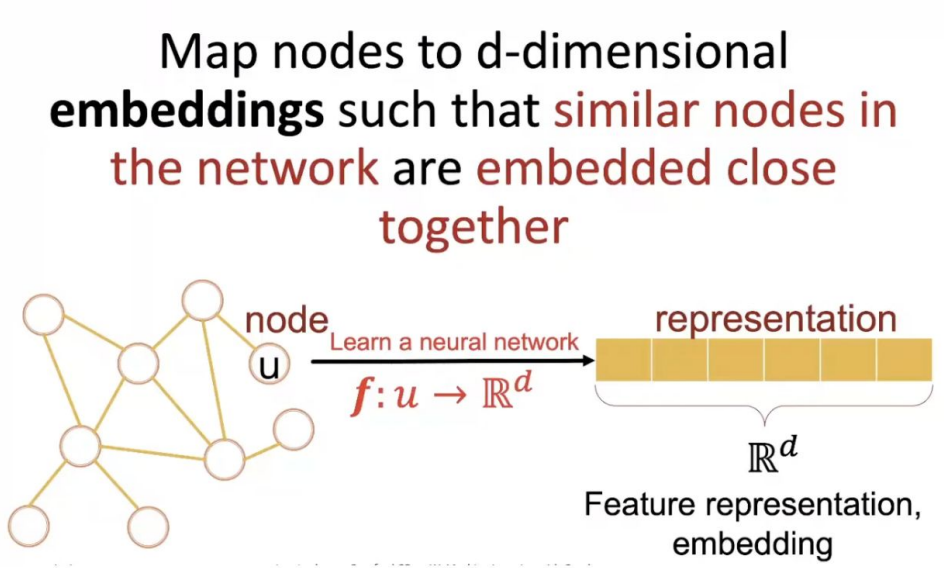
Representation learning



Representation learning

Map nodes to *d-dimensional* embeddings

Similar nodes in the network are embedded close.



Examples for key challenges in GML

Node classification: KSM is the uncle of Ramzi Yousef. KSM met with Osama bin Laden in the Tora Bora mountains. Ramzi attended a Madrasa. Is Ramzi Yousef a terrorist?

Link prediction: how likely is a cancer patient with a specific genetic phenotype who is being treated with chemotherapy likely to have an adverse drug reaction?

Subgraph analysis: Can we detect functional modules in protein interaction networks?

Generative graph modeling: given the structure and properties of a virus can we generate new molecules that would disable the virus?

Clustering, classification, and regression: Can we uncover fraudulent groups of users in financial transaction networks?

Course hypothesis

Until recently, very little attention has been devoted to the generalization of neural network models to richly structured datasets.

By capturing the relational structure in graphs, we can learn better representations of data.

Having learned better representations of we can build better models that are much more broadly applicable. And we can solve problems outside of traditional ML.

Graphs are a new frontier of deep learning.

Course

1. Intro to Machine Learning on Graphs
2. Traditional Graph Learning
3. Node Embeddings and Link Analysis
4. Label Propagation for Node Classification
5. Graph Neural Networks: GNN Model
6. Applications of Graph Neural Networks
7. Theory of Graph Neural Networks
8. Knowledge Graphs, Link Detection
9. Subgraph Mining with GNNs
10. Generative Graph Models
11. Advanced Topics, Project

- Weekly assignments - 50%
- Reading + research
- Midterm - 20%
- Final project + Final Exam - 30%

Note: New course and quickly evolving topic. Reserve right to adjust topics and topic ordering may change.

Secondary objective: learn how to do a research project.

Demo?

https://colab.research.google.com/drive/1TdZOh9eZkmazvKc5scF_bj0o8AMaWA1l#scrollTo=i4WcGtt1SP88

SKIP - Course hypothesis

Graph-structured data is ubiquitous throughout the natural and social sciences, from telecommunication networks to quantum chemistry.

Building relational inductive biases into deep learning architectures is crucial if we want systems that can learn, reason, and generalize from this kind of data.

Advances in graph representation learning have led to new state-of-the-art results in numerous domains: chemical synthesis, 3D-vision, recommender systems, question answering, and social network analysis.

Plentiful graph datasets - challenge is unlocking the potential of this data. How can we use ML to tackle this challenge?