Winter Term 20/21

Graph Neural Networks

Org & Introduction

Prof. Dr. Holger Giese (holger.giese@hpi.uni-potsdam.de)

Christian Medeiros Adriano (christian.adriano@hpi.de) - **"Chris"**

Sona Ghahremani (sona.Ghahremani@hpi.de)

Lecture topics

1. **Organization and Motivation for Machine Learning on Graphs**
2. **Overview of Graph Theory and Network Science (metrics and topology)**
3. **Discovering Graph Structures (Clustering, Community)**
4. **Message Passing and Node Classification**

Descriptive models

5. **Graph Representation Learning**
6. **Link Analysis - PageRank**
7. Graph Convolutional Neural Networks
8. Graph Recurrent Networks
9. Graph Attention Networks
10. Temporal Graph Neural Networks

Prediction models

11. Deep Generative Models for Graphs
12. Cascading behavior and Failure Propagation
13. Influence maximization
14. Outbreak minimization

Intervention models

Plan for the Semester

Phase-1: Foundations and Groundwork (First 3 weeks)

90% of teaching and a few small tasks like setting up the environment, studying datasets, and learning how to use libraries. Lectures happening twice a week and individual meetings on-demand during the week.

Phase-2: Exploration and Exploitation (Short lectures)

30 min teaching and quick project status updates (5-10 min maximum). We will keep having individual meetings on demand.

Phase-3: Consolidation (Last two weeks)

writing the final report, which is a summary of the results that you would have presented in the weekly update meetings.

Project

Team size: tree, two persons or individual

Project proposal in three stages:

1- Abstract (250 words): [Context][Problem][Investigation approach] - in two weeks

2- Related work (1 page, double column) containing summary and critique – in four weeks

3- Proposal - first draft in six weeks

- Detail the problem (what is it, why should I care, why is it difficult)
- Describe the dataset (source, size, main features, cite any papers that used it)
- Determine the metrics and algorithms to be used (preliminary insights, it can change)
- Discuss how you will evaluate your results (benchmarks and null-models)

Datasets and Tools

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Datasets

- http://networkrepository.com/
- https://snap.stanford.edu/data/
- https://networkdata.ics.uci.edu/

Tools (sorted by priority)

1. cuGraph: https://github.com/rapidsai/cugraph (Strongly recommend, fast)
2. NetworkX: https://networkx.org/documentation/stable/tutorial.html (great coverage of graph algorithms)
3. Snap for Python: http://snap.stanford.edu/snappy/index.html
4. Pytorch Geometric: https://pytorch-geometric.readthedocs.io/en/latest/
5. Github project: https://github.com/orgs/hpi-sam/projects/3

Communicantion Plan

Motive	Content	Medium
Artifacts	Source code, Data Documentation, Wiki	Github - https://github.com/orgs/hpi-sam/
Papers	Copyrighted material	Bib-Admin
Messaging ad hoc	Questions, Suggestions, Sharing	Our Slack group: **graph-neural-networks.slack.com**
Official communications	Schedule, Orientations	Email
Meetings	Lectures, Status, Work meetings	Zoom, Skype
Emergency	Call, SMS, messaging	Chris mobile number (check my Slack profile)

Grading criteria

To pass the project seminar, we expect sufficient contributions to:

- the discussions in meetings (preliminary results and readings)
- the experiments
- the final report
- the final presentation

Grading:

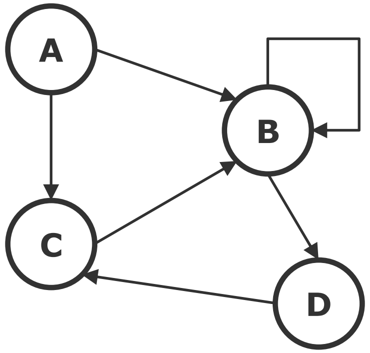
- project (60%)
- report (30%)
- presentation (10%)

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

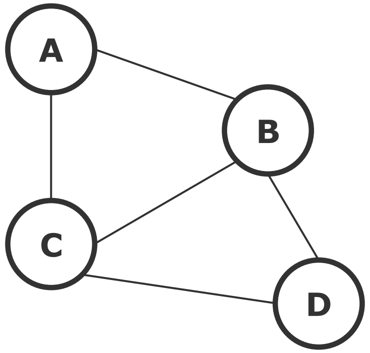
Basic Concepts

Types of graphs

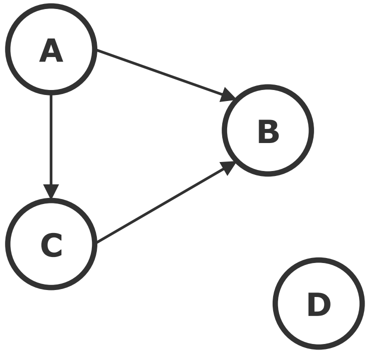
Directed



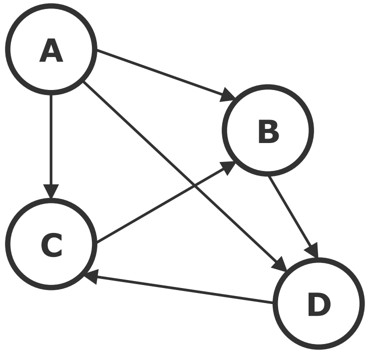
Undirected



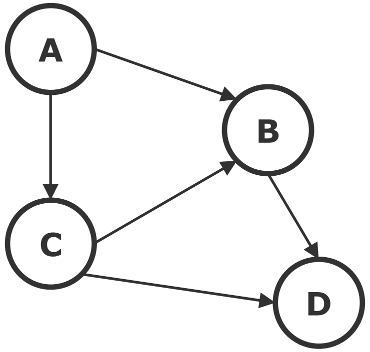
Disconnected



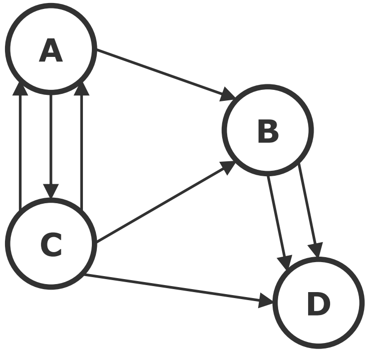
Fully connected



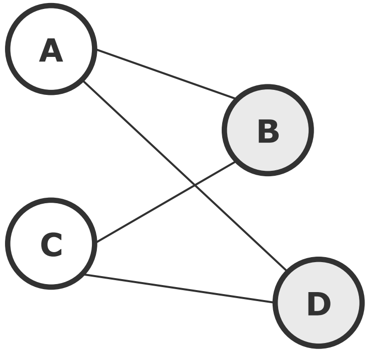
Directed Acyclic Graph



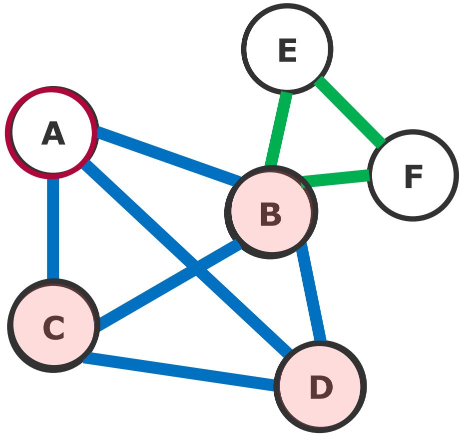
Multigraph



Bipartite



Cliques



Ego network of A

Node and Edge degrees

Node degree: number of edges of node k_i, where *i* is the node index

Indegree: number of incoming edges

Outdegree: number of outgoing edges

Average degree: $\bar{k} = \frac{1}{N} \sum_{i \in N} k_i = \frac{2E}{N}$

, where *E* = number of edges, *N*=number of nodes

Maximum number of edges: $\mathrm{E}_{\max} = \binom{N}{2} = \frac{N(N-1)}{2}$

However, most real-world networks are sparse, i.e., $E \ll E_{max}$

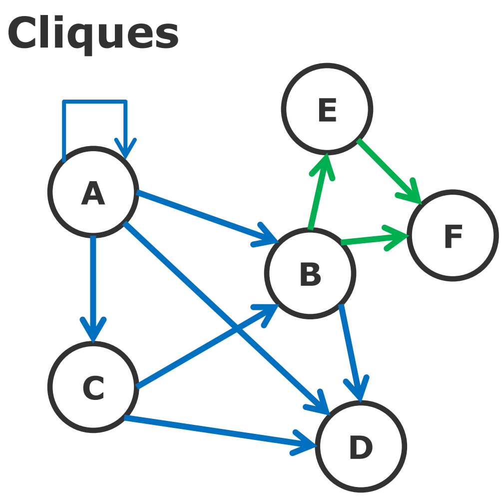
Most real-world networks are sparse

Network	N	E	N_b	E_b	$\bar{d}$	Description
Social networks						
Delicious	147,567	301,921	0.40	0.65	4.09	del.icio.us collaborative tagging social network
Epinions	75,877	405,739	0.48	0.90	10.69	Who-trusts-whom network from epinio [Richardson 03]
Flickr	404,733	2,110,078	0.33	0.86	10.43	Flickr photo sharing social network [Kumar et a
LinkedIn	6,946,668	30,507,070	0.47	0.88	8.78	Social network of professional contacts
LiveJournal01	3,766,521	30,629,297	0.78	0.97	16.26	Friendship network of a blogging community strom et al. 06]
LiveJournal11	4,145,160	34,469,135	0.77	0.97	16.63	Friendship network of a blogging community strom et al. 06]
LiveJournal12	4,843,953	42,845,684	0.76	0.97	17.69	Friendship network of a blogging community strom et al. 06]
Messenger	1,878,736	4,079,161	0.53	0.78	4.34	Instant messenger social network
Email-All	234,352	383,111	0.18	0.50	3.27	Research organization email network (all add [Leskovec et al. 07b]
Email-InOut	37,803	114,199	0.47	0.82	6.04	(all addresses but email has to be sent both [Leskovec et al. 07b]
Email-Inside	986	16,064	0.90	0.99	32.58	(only emails inside the research organiz [Leskovec et al. 07b]
Email-Enron	33,696	180,811	0.61	0.90	10.73	Enron email data set [Klimt and Yang 04]
Answers	488,484	1,240,189	0.45	0.78	5.08	Yahoo Answers social network
Answers-1	26,971	91,812	0.56	0.87	6.81	Cluster 1 from Yahoo Answers
Answers-2	25,431	65,551	0.48	0.80	5.16	Cluster 2 from Yahoo Answers
Answers-3	45,122	165,648	0.53	0.87	7.34	Cluster 3 from Yahoo Answers
Answers-4	93,971	266,199	0.49	0.82	5.67	Cluster 4 from Yahoo Answers
Answers-5	5,313	11,528	0.41	0.73	4.34	Cluster 5 from Yahoo Answers
Answers-6	290,351	613,237	0.40	0.71	4.22	Cluster 6 from Yahoo Answers
Information (citation) networks						
Cit-Patents	3,764,105	16,511,682	0.82	0.96	8.77	Citation network of all US patents [Leskovec et a
Cit-hep-ph	34,401	420,784	0.96	1.00	24.46	Citations between physics (ArXiv hep-th) [Gehrke et al. 03]
Cit-hep-th	27,400	352,021	0.94	0.99	25.69	Citations between physics (ArXiv hep-ph) [Gehrke et al. 03]
Blog-nat05-6m	29,150	182,212	0.74	0.96	12.50	Blog citation network (6 months of data) [Lesk al. 07c]

$$\bar{d} = \bar{k} = average\ degree$$

source :Leskovec, J., et al. "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters." *Internet Mathematics* 6.1 (2009): 29-123.

Adjacency matrix



	A	C	D	B	E	F
A	1	1	1	1	0	0
C	1	0	1	1	0	0
D	1	1	0	1	0	0
B	1	1	1	0	1	1
E	0	0	0	1	0	1
F	0	0	0	1	1	0

However, adjacency matrix of real-world networks are full of zeros

Motivation for Learning on Graphs and GNNs

Scenarios and Network Types

Network Types

- Event graphs
- Disease pathways
- Knowledge-graphs
- Scene graphs
- Heterogeneous graphs (different types of nodes and edges)

Scenarios

- Clustering in Social network
- Protein interaction
- Cell similarity networks
- Failure propagation in infrastructure networks
- Fake news detection
- Side-effects of drugs
- Network attacks
- Traffic jams

Types of Predictions

Node classification

What type of node is this?

Link prediction

Are these two nodes connected?

With which strength?

Graph Classification

Patterns of connectivity (motifs)

Network similarity (isomorphism)

Next steps

Lecture-2: Overview of Graph Theory and Network Science (Wed 13h30 - 15h00)

- Degree Distribution
- Path length
- Clustering Coefficient
- Null-Models for Graphs

First task:

- Accept Slack invitation
- Study datasets
- See examples of use of Snap or NetworkX
- Think of scenarios

END