Winter Term 21/22

Graph Neural Networks Applications & Link to Graph Queries

Org & Introduction

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

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

Matthias Barkowsky (matthias.barkowsky@hpi.de)

Key Facts

- Weekly Hours: **4**
- Credit Points: **6**
- Teaching Form: **Project Seminar**
- Enrolment Type: **Compulsory Elective Module** (“Wahlpflichtmodul”)
- Course Language: **English**
- Study Programs and Modules:
  - **IT-Systems Engineering MA**
    - Mandatory module : „*IT-Systems Engineering Analysis*"
    - Mandatory module: „*IT-Systems Engineering Design*"
    - Specialization module(s): „*Software Architecture & Modeling Technology*"
  - **Data Engineering MA**
  - **Digital Health MA**

Dates

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

- Enrollment deadline: **22.10.2021**
  - Cancellation deadline for enrollment: **30.01.2022**

- Introductory meeting: **27.10.2021** **[NOW]**

- Meetings:
  - *Lectures - scheduled*
  - *Update meetings – on demand, usually weekly*

- Final Presentations at end of the semester: **To be decided**
  - *We will be present at the lecture room, but we will also be joining via Zoom.*

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, Administrative issues	Email christian.adriano@hpi.de
Meetings	Lectures, Status, Work meetings	Zoom, Skype
Emergency	Call, SMS, messaging	Chris mobile number (check Chris' Slack profile)

Seminar Work, Deliverables and Grading

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

- Work **alone or in groups** on **one selected topic/project**.
- Each team has on-demand update meetings with teaching assistant.

Project Execution: [60% of final grade]

- Weekly update meeting
- Intermediary Presentations

Written deliverables: [30% of final grade]

- Final report on findings
  - Length: approx. 10 pages ACM Format per team participants
  - Some parts must be attributable to each individual author

Final Presentations: [10% of final grade]

- Presentation on findings
- Questions and feedback for other students' presentation

Road Map (1/2)

		Objectives
1. Intro and Course Organization	**Week-1** Organization	**Objectives** Team building, Setup, Topic
2. Graph Metrics and Random Models 3. Graph Structural Features – Clustering 4. Message Passing & Belief Propagation	**Week-2** Description and Feature models	Understand a phenomenon Extract features Establish baselines Preprocessing data
5. Graph Embeddings - Message Passing 6. PageRank & Markov Chains & Graph Queries 7. Graph Convolutional Networks	**Week-3** Basic Prediction models	Predict an outcome ML architecture and pipeline
8. Graph Attention Networks 9. Graph Evolution Networks 10. Temporal Graph Networks	**Week-4** Advanced Prediction models	Training models Evaluation models
11. Deep Graph Generative Models 12. Causal Graph Neural Networks 13. Propagation Graph Neural Networks • Network Effects, Cascading and Contagion • Outbreak Detection and Influence Maximization	**Week-5** Generative and Intervention models	Effects of interventions Risks of confounding Causal structure

Roadmap (2/2)

- **Project Phase 1: Learn fundamentals - Lectures**
  - Goal: learn fundamentals
  - Deadline: Mid-End of December
- **Project Phase 2: Present Proposal - Reading and Writing**
  - Goal: learn about the state of art of one application area
- **Project Phase 3: Apply a method - Coding and Evaluation**
  - Goal: learn to apply and evaluate a method
  - Present update in weekly meetings

- **Final Presentations** in one session in **February 2022**
- **Submission of final report** one week after the presentation

Project Proposal

Team size: up to four people.

Project proposal in two stages:

1- State-of-art (one page, double column) – in 6 weeks (First week of December)

- Each person covers at least five well-selected papers (group covers at least 20 papers)

2- Plan - first draft in 8 weeks (before New Years Break)

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

Datasets and Tools

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

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

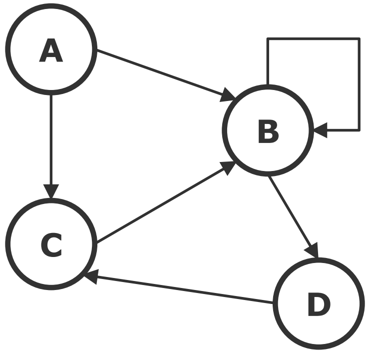
Patterns of connectivity (motifs)

Network similarity (isomorphism)

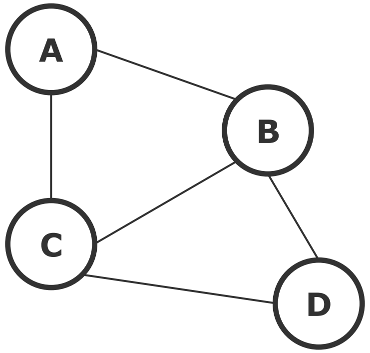
Basic Concepts

Types of graphs

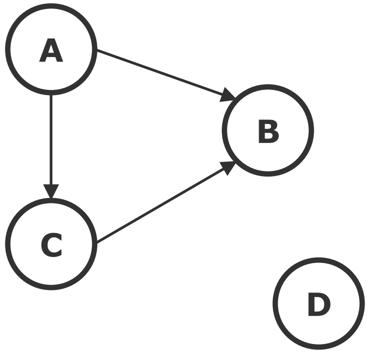
Directed



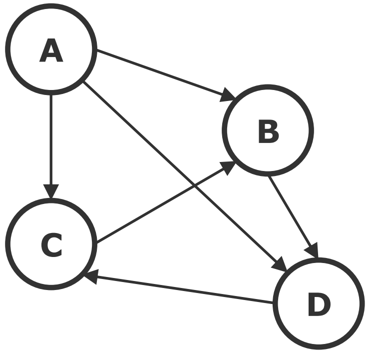
Undirected



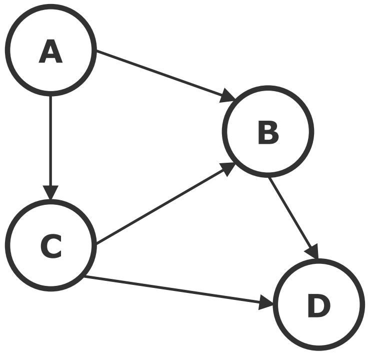
Disconnected



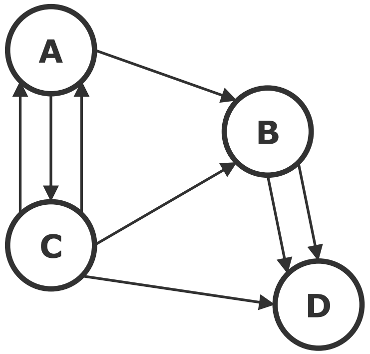
Fully connected



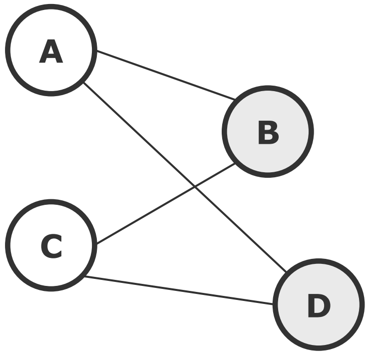
Directed Acyclic Graph



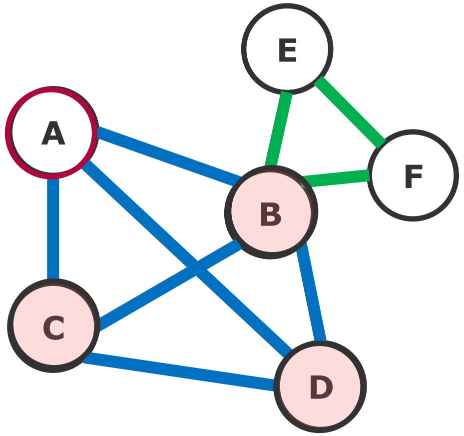
Multigraph



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

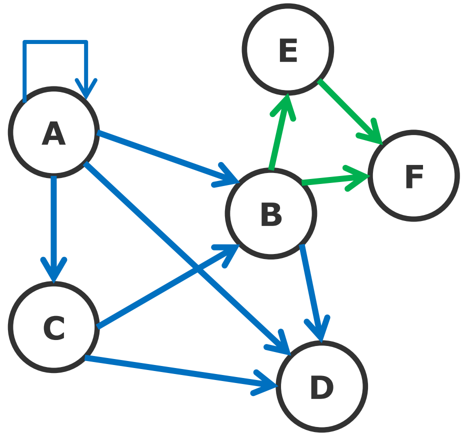
, where *E* = number of edges, *N*=number of nodes

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

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

Adjacency matrix

Cliques



	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

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]

N = number of nodes

E= number of edges

N_b = fraction nodes not in whiskers (size of largest biconnected component)

E_b = fraction of edges in biconnected component

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

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.

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

END