Lab Program

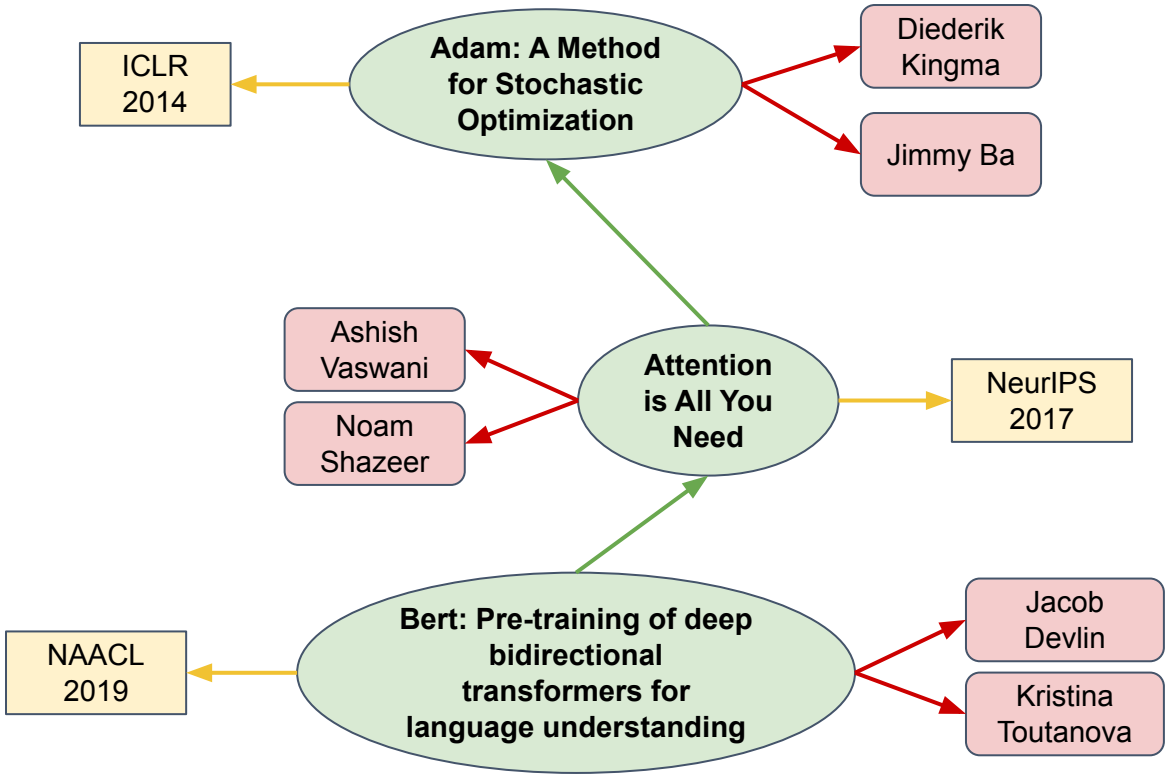
Time (EST)	Content	Speaker
10:45 - 11:00	Welcome and introduction	Filip
11:00 - 11:20	Internet Memes: *Knowledge connects culture and creativity*	Filip
11:20 - 11:40	Financial transactions: *Detecting anomalies in trading*	Ke-Thia
11:40 - 12:00	PubGraphs: *What should I read next?*	Kian & Jay
12:00 - 12:20	Morality in events: *From news to timelines and graph maps*	Gleb
12:20 - 12:30	Discussion and Closing remarks	Jay

USC Viterbi
School of Engineering
Information Sciences Institute

PubGraph:
What should I read next?

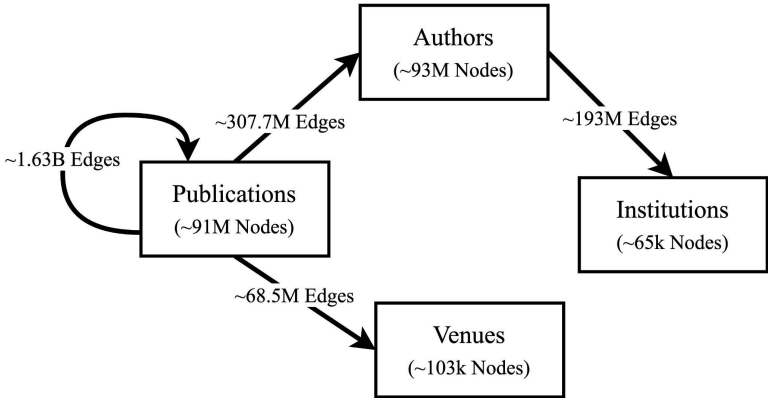
Introduction

- Scientific discoveries are built on prior research findings
- Relationships between different publications, authors, topics, & communities can help us understand new discoveries
- How can we characterize these relationships?
- … and what can they tell us about the nature of science, discovery, and how to win collaborators and influence research?



OpenAlex:

- 📄 Publications (Papers, Books, Datasets, etc.)
- 👩 Authors
- 📚 Venues
- 🏫 Institutions
- 💡 Concepts



PubGraph Dataset (arXiv)

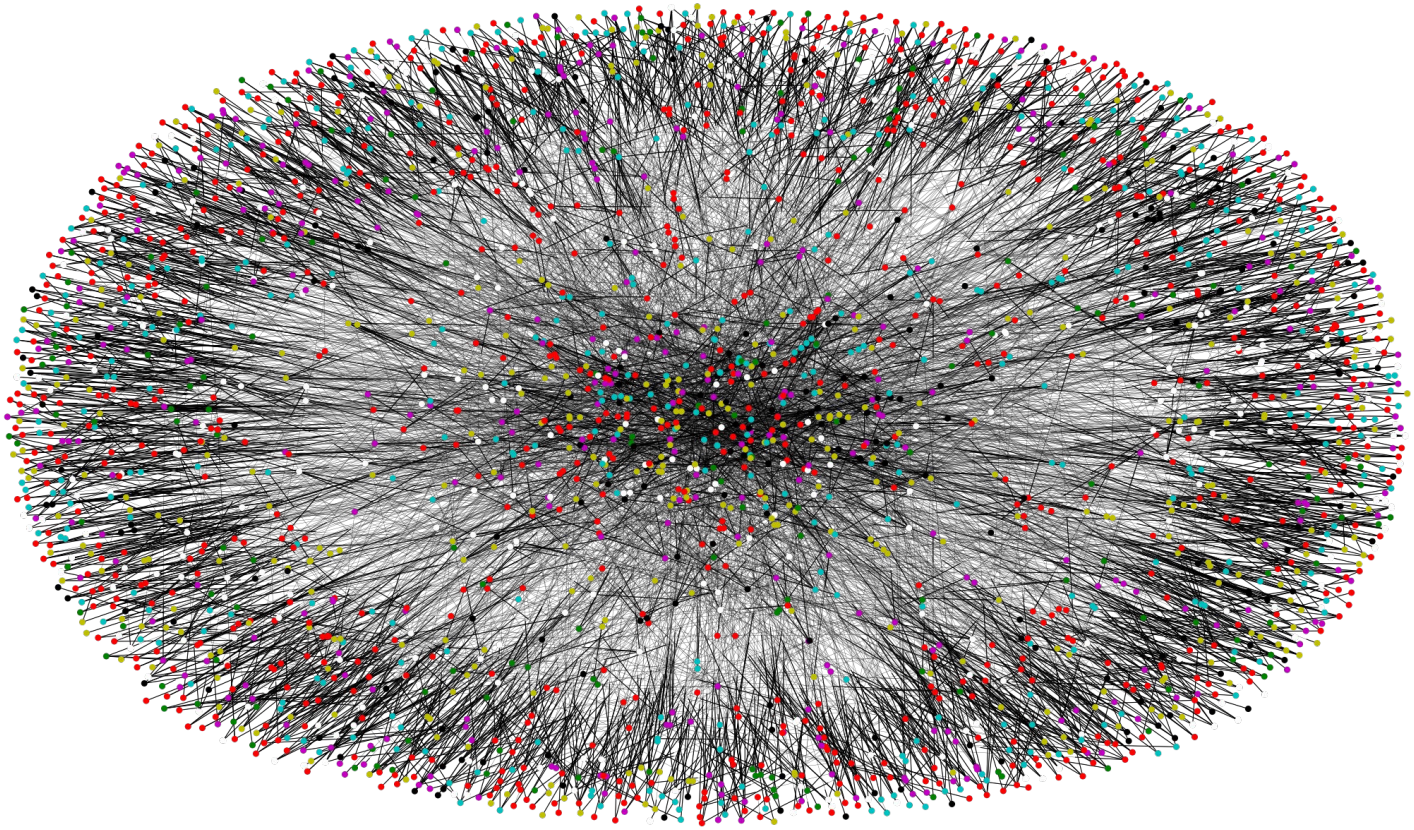
Dataset	Nodes	Edges	Rel
OGBL-Cite	3M	30M	1
Freebase	86M	339M	15000
WikiKG90Mv2	91M	601M	1000
PubGraph	432M	15B	51
PG-1M	3M	22M	4
PG-10M	25M	315M	4
PG-Full	184M	2.2B	4

Metric	PG-1M	PG-10M	PG-Full
Mutual Citations (%)	0.03	0.04	0.06
Authorship Completeness (%)	99.97	99.97	99.92
Venue Completeness (%)	92.37	90.25	75.34
Institution Completeness (%)	81.45	71.21	45.77

Table 2: Validity and completeness metrics of sampled KGs.

WikiData Properties					
OpenAlex Metadata	**WikiData Property**	**OpenAlex Metadata**	**WikiData Property**	**OpenAlex Metadata**	**WikiData Property**
OpenAlex Id	P10283 (OpenAlex ID)	Type	P2308 (class)	Country Code	P299 (ISO 3166-1 numeric code)
MAG	P6366 (Microsoft Academic ID)	Updated Date	P5017 (last update)	City	P131 (located in the administrative territorial entity)
Orcid	P496 (ORCID ID)	Year	P585 (point in time)	Geo Region	P276 (location)
Grid Id	P2427 (GRID ID)	Created Date	P571 (inception)	Country	P17 (country)
Scopus Id	P1153 (Scopus author ID)	Date of Publication	P577 (publication date)	Published in	P1433 (published in)
ROR	P6782 (ROR ID)	Author	P50 (author)	Latitude & Longitude	P625 (coordinate location)
Wikidata Id	P1687 (Wikidata property)	Title	P1476 (title)	Description	P10358 (original catalog description)
Geonames Id	P1566 (GeoNames ID)	Publisher	P123 (publisher)	Ancestor	P1038 (relative)
DOI	P356 (DOI)	Related to	P1659 (related properties)	Level/Position	P1352 (ranking)
ISSN	P236 (ISSN)	Works Count	P3740 (number of works)	Twitter Handle	P2002 (Twitter username)
ISSNL	P7363 (ISSN-L)	Volume	P478 (volume)	MeSH Qualifier Id	P9341 (MeSH qualifier ID)
PMID	P698 (PubMed ID)	Issue	P433 (issue)	MeSH Descriptor Id	P486 (MeSH descriptor ID)
PMCID	P932 (PMCID)	Homepage URL	P856 (official website)	Score	P4271 (number of points/goals/set scored)
UMLS CUI	P2892 (UMLS CUI)	OA status	P6954 (online access status)	Relationship	P2309 (relation)
Instance of	P31 (instance of)	Affiliation	P1416 (Affiliation)	Concept	P921 (main subject)
Wikipedia	P5178 (Wikipedia Library partner ID)	Cites work	P2860 (Cites work)	Display Name	P2561 (name)
Artificial Properties					
OpenAlex Metadata	**Artificial Property**	**OpenAlex Metadata**	**Artificial Property**	**OpenAlex Metadata**	**Artificial Property**
Cited by count	P_cited_by_count	Is retracted	P_is_retracted	UMLS AUI	P_umls_aui

WikiData Ontology Mapping



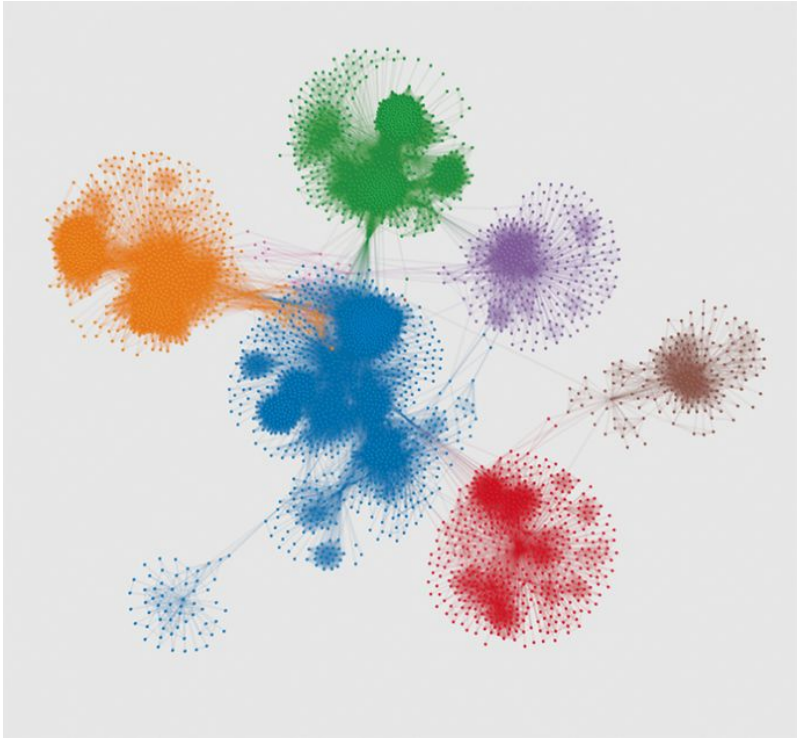
What should I read next?

Knowledge Graph Embedding Using KGTK

- Find papers that are similar to the current paper that we are reading
- How? **Representation Learning**
  - Learn an n-dimensional representation for each paper using link structure
  - Find similar papers through a similarity measure (e.g., euclidean distance or cosine similarity)

$$f_r(\theta_x, \theta_y) = c(\theta_x, g_r(\theta_y))$$

- Let's try out this scenario!

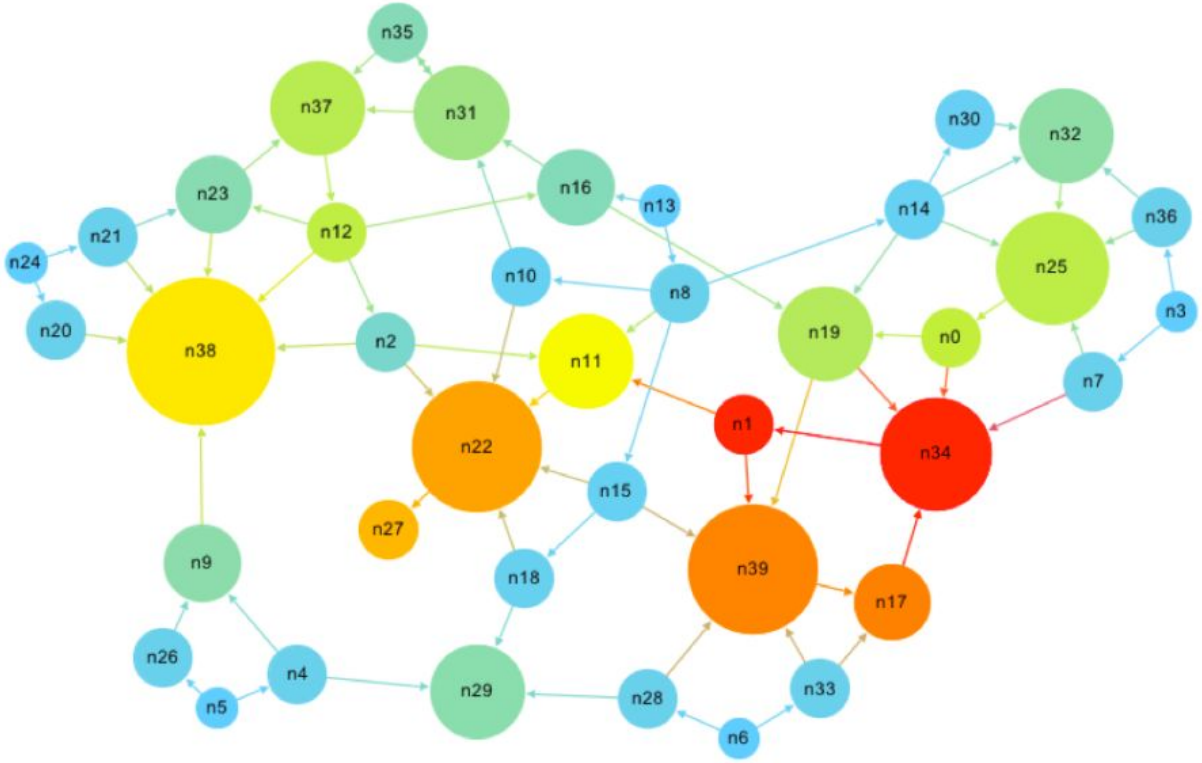


20 years of network community detection (Fortunato and Newman, 2022)

Who should we we invite for a talk?

Community Detection Using KGTK

- Find researchers that have similar interests
- How? **Community Detection**
  - Find “communities” a subset of papers that have strong internal citation connections and weak external citation connections
  - Assume publications are good representatives of authors’ interests
- Let’s see how this can be done!

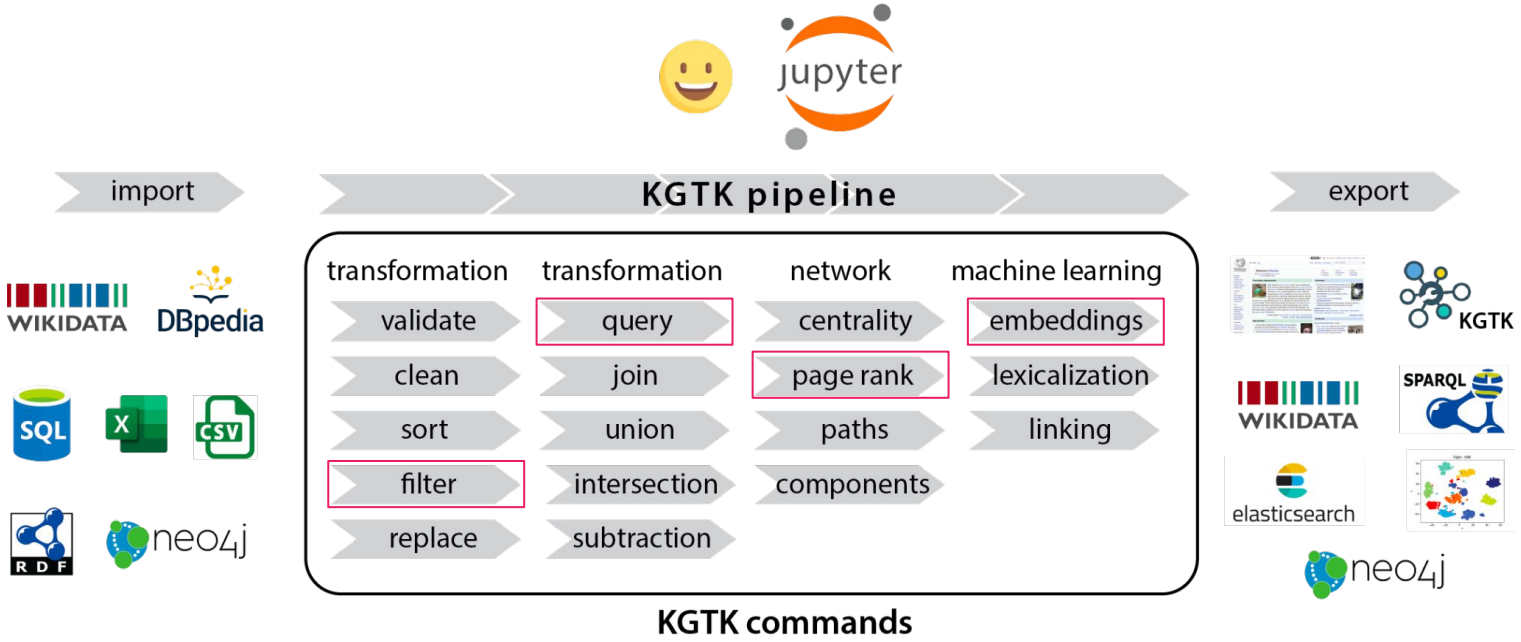


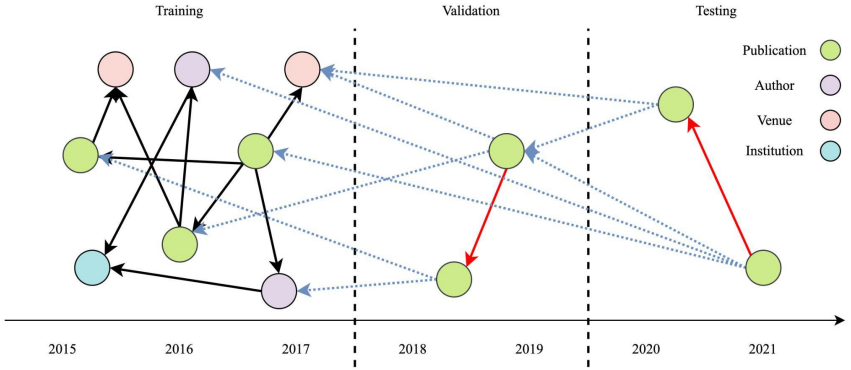
What should I read first?

PageRank Using KGTK

- Find papers that have been influential in a specific domain that we are interested in.
- How? **Node Importance**
  - Assign an importance score to each paper based on all the citations and references in the graph
- Let’s do this on our data!

Using KGTK at Scale





Model	MRR	Hits@1	Hits@10	Hits@50
ComplEx	0.007	0.000	0.008	0.071
GraphSAGE-D	0.149	0.062	0.324	0.664
RGCN-D	0.166	0.071	0.364	0.709
GraphSAGE-E	0.779	0.697	0.917	0.968
RGCN-E	0.706	0.609	0.872	0.939
GraphSAGE-H	**0.784**	**0.697**	**0.928**	0.973
RGCN-H	0.746	0.645	0.919	**0.977**

Variation	#Negative Samples	MRR	Hits@1	Hits@10	Time (Seconds)
Random	1000	0.723	0.608	0.918	588 (CPU)
Entity Type	1000	0.560	0.418	0.826	655 (CPU)
Time Constrained	1000	0.577	0.449	0.817	601 (CPU)
Community	1000	0.076	0.023	0.167	1008 (CPU)
Full	~3.38M	0.015	0.000	0.036	81987 (GPU)

Variation	MRR	Hits@1	Hits@10	Hits@50
Full	**0.725**	**0.609**	**0.920**	**0.992**
- V	0.713	0.598	0.909	0.969
- I	0.722	0.606	0.919	0.978
- V - I	0.715	0.600	0.909	0.970
- A - I	0.536	0.394	0.802	0.922
- V - A - I	0.518	0.380	0.780	0.900

PubGraph: A Large Scale Scientific Temporal Knowledge Graph (arXiv)

Impact / Future Work

Impact:

- **PubGraph**: Massive new resource with 15B edges
- New methods and benchmarks for **inductive learning**
- Characterizing **researcher dynamism** and why it matters
- 2 workshop papers at AAAI, 2 papers in submission to IJCAI

What's next?

- Identifying **surprising** connections
- Extending analysis **researcher dynamism**
- Using more **paper content** alongside PubGraph