Better Together: Text + Context

Kenneth Church

Need to Re-organize Courses (Interdisciplinary Re-Org)

Current Courses

- Machine Learning (ML)
- Vision
- Natural Language (NLP)
- Machine Translation (MT)
- Information Retrieval (IR)
- Speech

Future

- Machine Learning (ML)
- ML Applications
  - Vision/NLP/MT/IR/Speech

Soap Box: My Views

- Interdisciplinary Re-Org: ML/Vision/NLP/MT/IR/Speech
  - Current: People working on one app don’t talk to people working on other apps
  - Future: We should work together because we are all using the same ML techniques
- Responsible AI
  - Current: Bias is Bad
  - Future: Embrace Diversity
- Authors View vs. Audience Response
  - Current: Focus on Unambiguous Objective Labels (independent of context)
  - Future: More Room for Subjectivity (and richer semantics)
    - Authors’ Position ≠ Audience Response
- End-to-End Optimization vs. Modularity
  - Current: End-to-End → Better Performance
  - Future: Multiple Perspectives are Better Together

Vision: Facts → Opinions

Facts

(b) COCO: *A man and a woman holding a little kid while sitting at a table outside*

Opinions

(a) ArtEmis: *I love everything about this painting of a mother and her two children lovingly interacting with the family pet cat.*

Ambiguity: No Single Correct Gold Label

- A difference of opinion ≠ Error
  - Common Challenge in Language Apps
    - Machine Translation
    - Web Search
    - Even Part of Speech Tagging!
- Standard Solution
  - Score over multiple references (many gold standards)
    - Machine Translation: BLEU
    - Web Search: NDCG

Candidate labels: baseball cap, cap, green hat, hat, head.
Can you guess which one is in the gold standard?

Classic Challenges in Philosophy of Language

Compare & Contrast Language in Visual Genome (VG) with NLP Corpora

- Bounding Box Semantics: Too Limiting
  - <NP> <relation> <NP>
  - NPs in VG are usually rigid designators that mean the same thing in all contexts
    - Few abstract nouns (*ideas*), predicates (verbs, adj), variables (pronouns)
    - Most VG nouns are visual (and less about other senses)
  - More entropy in nouns than relations
    - 8 boring relations cover bulk of the cases
    - Linguists are more interested in predicates than arguments
  - Modifiers are limited to a single box
    - Relative Relations: *girl with green hat* vs *girl on defense*
    - Aggregations over boxes
      - Count vs. Mass: *Cloudy Sky*, *Sandy Beach*
      - Definite vs. Indefinite: *girls playing frisbee* (as opposed to many other people in picture)
      - A horse with two legs?

Verbs, Subjunctive, Focus, Perspective, etc…

- Pictures vs. Videos
  - Captions on pictures have more nouns:
    - *girl with **green hat***
  - Captions on videos have more verbs:
    - *girl **throwing** frisbee*
- Possible Worlds: Subjunctive, Hedges
  - *The girl on defense **might** block the throw,*
  - *but **probably** won't*
- Focus: ***girl on offense*** vs. ***girl on defense***
- Perspective: photographer vs. audience

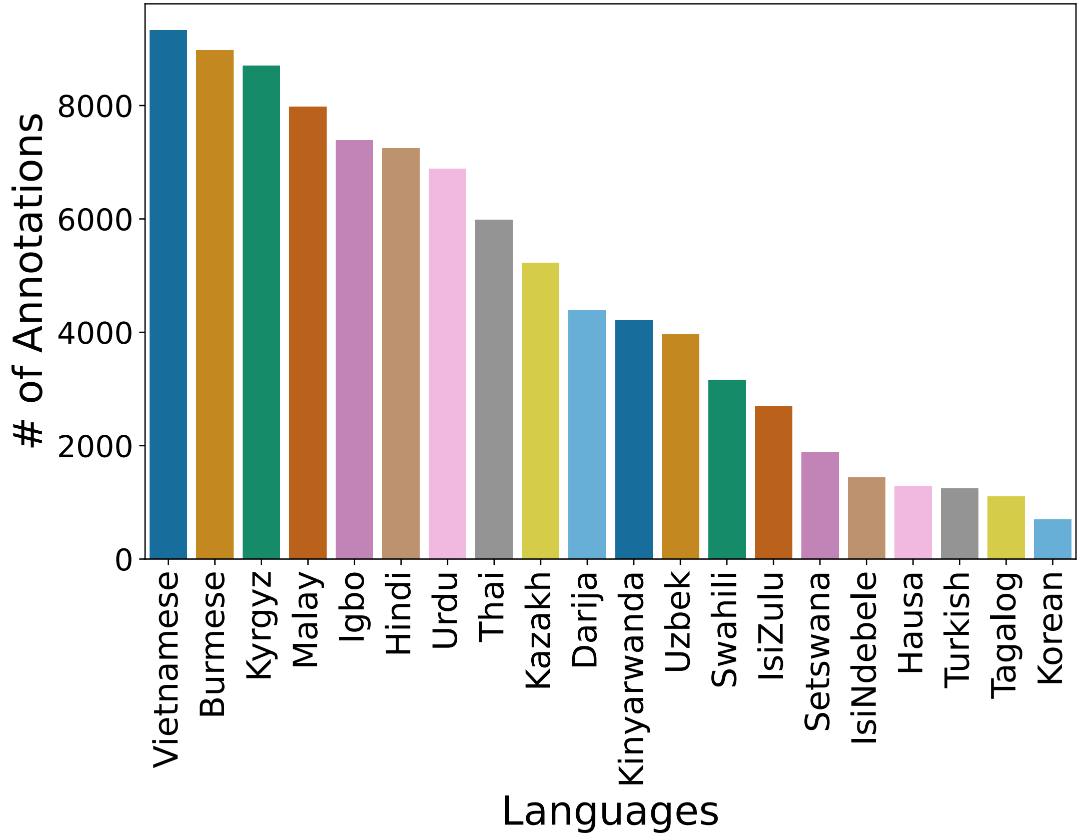
ArtELingo: A Million Emotion Annotations of WikiArt with Emphasis on Diversity over Language and Culture

Youssef Mohamed[1*] **Mohamed Abdelfattah**[1] **Shyma Alhuwaider**[1] **Feifan Li**[1]
Xiangliang Zhang[2] **Kenneth Ward Church**[3] **Mohamed Elhoseiny**[1*]
[1]KAUST [2] University of Notre Dame [3] Northeastern University
{youssef.mohamed,mohamed.abdelfattah,shyma.alhuwaider,feifan.li}@kaust.edu.sa
xzhang33@nd.edu, k.church@northeastern.edu, mohamed.elhoseiny@kaust.edu.sa

a)		شلال طبيعي جميل. مشاعر النمو والحيوية والطاقة موجودة. Translation: Beautiful natural waterfall. Feelings of growth, vitality and energy. **Excitement** **Arabic**	The water that's rushing downward looks like a bride's wedding veil. **Awe** **English**	瀑布就像四蹄生风的白马如潮水涌来，非常的壮观 Translation: The waterfall is like a white horse and wind, it is spectacular. **Contentment** **Chinese**
b)		Translation: Girls sitting with their mother outside the house, exchanging love and affection, pigeons flying over a tree. **Contentment** **Arabic**	The women relaxing while birds are flying about makes me feel relaxed and calm as well. **Contentment** **English**	Translation: Three sisters lying on a bench and watching the birds fly comfortably. **Contentment** **Chinese**
c)		Translation: The use of black and white for painting the forests with all its details brings out a feeling of satisfaction. **Contentment** **Arabic**	The trees are dead and exposing their roots due to erosion and lack of water. **Sadness** **English**	Translation: After the snow in winter, there is snow everywhere, and the dead trees look very depressed. **Sadness** **Chinese**

ArteLingo-23:
Embrace Diversity over Language and Culture

- **Africa:** Kinyarwanda, Swahili, IsiZulu, Setswana, Yoruba, Hausa, Igbo.
- **Southeast Asia:** Vietnamese, Indonesian, Thai, Burmese, Malay.
- **Sub-Indian continent:** Tagalog, Tamil, Hindi, Urdu.
- **East Asia:** Korean, *Chinese*.
- **Middle-East:** Turkish, Darija, *Arabic*.
- **Central Asia:** Uzbek
- **Europe and North America:** *English*.



Why study more languages?

- Vision:
  - Depends on both stimulus as well as context
    - Stimulus: Picture
    - Context: Language, Culture, Religion, Politics, Education, Background
- Cliché:
  - *Beauty is in the eye of the beholder*

Few Shot & Zero Shot

Seen Language

خزانة صغيرة على طاولة خشبية تحمل مزهرية بها أزهار حمراء وبيضاء وزرقاء.
Translation: A small dresser on a wooden table holds a vase with red, white, and blue flowers.
Arabic

There are a bunch of flowers in a yellow vase on a table. It looks like something from a restaurant. The table has a yellow cloth on it.
English

花瓶里白色的百合和 绿色的小花搭配着，让人感到很美。
Translation: The combination of white lilies and small green flowers in the vase looks nice.
Chinese

Unseen Language

Fleur blanche et rose par un vase très mince
Translation: White and pink flower by a very thin vase.
French

El cálido tono verde de las flores hace que la imagen sea reconfortante y agradable de contemplar.
Translation: The warm green hue of the flowers makes the image look comforting and pleasing.
Spanish

花園中花瓣色的花朵散滿在白色的花瓶上, 作者用畢生之力將花卉之美帶給世人。
Translation: The petal-colored flowers in the garden are scattered on the white vase. The author devotes his life to bringing the beauty of flowers to the world.
Chinese(Traditional)

Seen Language

تصور اللوحة امرأة وطفل في مشهد عن قرب. المرأة تمسك ذراع الطفل في ذراعها.
Translation: The painting depicts a woman and child in a close-up view. The woman is holding the child's arm in hers.
Arabic

The baby in the picture looks so calm with the mother closing her eyes and feeling peaceful and content.
English

图片描绘了母亲抚摸着她的孩子，他穿着白色短裤。母亲看起来很温柔。
Translation: The picture depicts a mother stroking her baby, who is wearing white shorts. Mother looks very gentle.
Chinese

Unseen Language

Ce tableau représente une femme vêtue d'une robe orange, assise dans le dos d'un enfant.
Translation: This painting shows a woman in an orange dress, seated behind the back of a child.
French

Esta pintura es una fotografía de un niño durmiendo con su madre.
Translation: This painting is a photograph of a boy sleeping with his mother.
Spanish

इस पेंटिंग में एक महिला मैनीक्योर किया हुआ चश्मा पहने हुए और गले में कपड़ा लपेटे हुए अपने बच्चे को देख रही है
Translation: In this painting, a woman is looking down at her child wearing manicured glasses, holding a cloth wrapped around her neck
Hindi

Anglo-Centered Baseline

- Use English as Pivot Language
- Anglo-Centered Vision Task:
  - Input: Picture
  - Output: English Caption
- If you want a caption in another language:
  - Just translate English caption to that language
- Challenge:
  - Can we do better than that?

Agreement (A) on Emotion Labels:
More A: Landscapes → Less A: Sketches

Figure 5: 8 artworks with genre. Green indicates high agreement in Table 4; red indicates high disagreement.

Genre (G)	$Pr(G\|U)$	$Pr(G\|D)$	A
landscape	0.206	0.097	-1.08
cityscape	0.071	0.036	-0.98
still life	0.043	0.042	-0.03
illustration	0.029	0.029	-0.01
misc	0.167	0.177	0.08
portrait	0.217	0.233	0.10
nude	0.030	0.032	0.11
religious	0.101	0.133	0.40
abstract	0.076	0.112	0.55
sketch	0.061	0.109	0.85

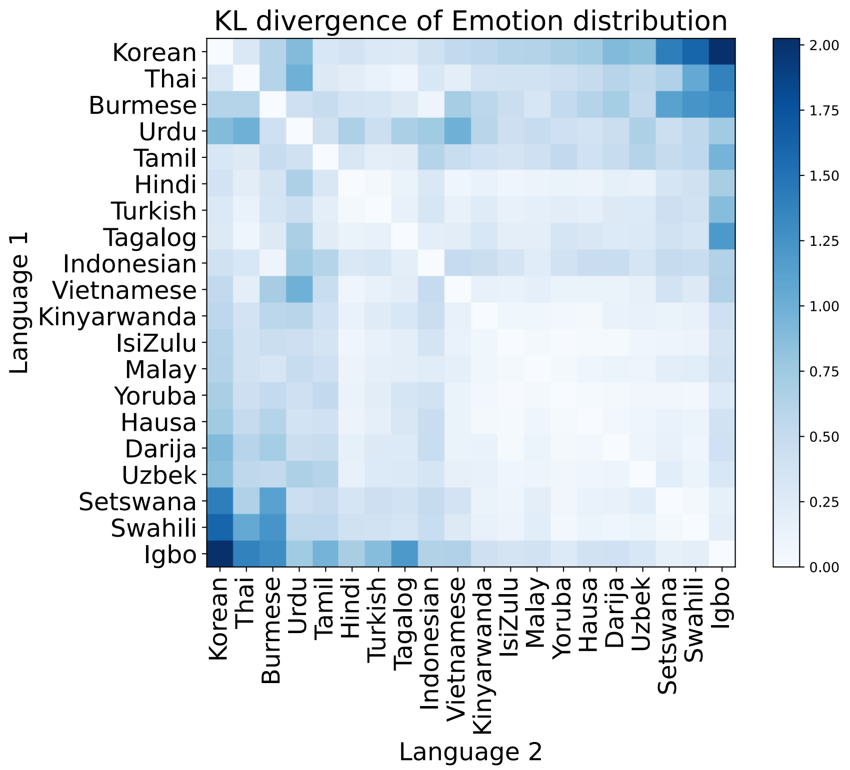
Table 4: Genre sorted by agreement (A). Most agreement: landscapes; Most disagreement: sketches.

Agreement on Emotion Labels

By Genre

Figure 5: 8 artworks with genre. Green indicates high agreement in Table 4; red indicates high disagreement.

By Language (or perhaps Education?)



Academic Search

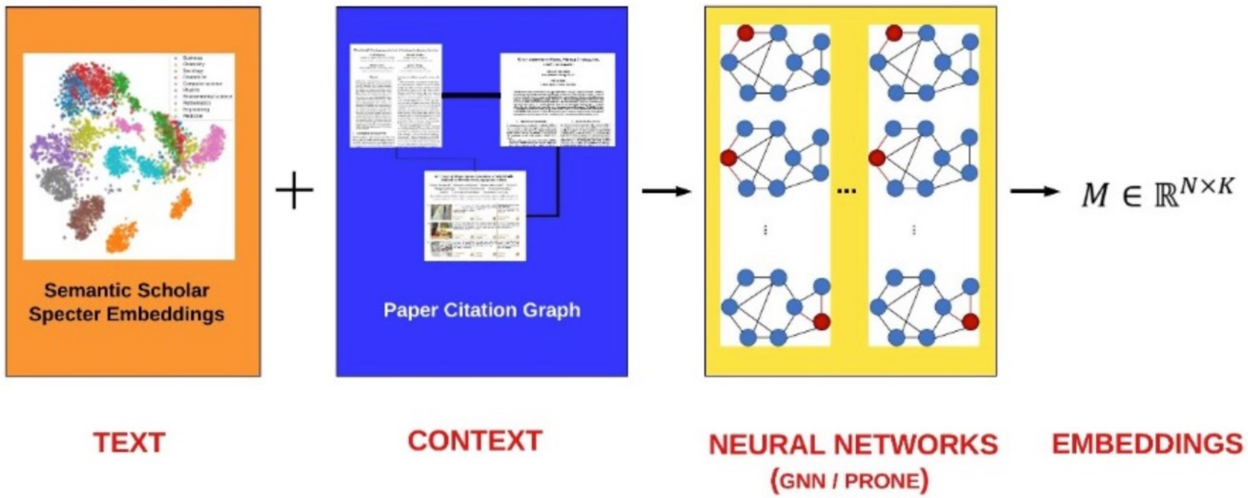
2023 Jelinek Summer Workshop on Speech and Language Technology

https://www.clsp.jhu.edu/2023-jelinek-summer-workshop/

Home > Better Together: Text + Context

Better Together: Text + Context

Abstract



Google Search: Semantic Scholar Gallery

Semantic Scholar API

Overview Tutorial Documentation Gallery Cite the Paper

Better Together

Find similar papers in Semantic Scholar

Looking for papers relevant to a specific paper of interest? Better Together built several different embeddings of documents from Semantic Scholar to help you find similar papers.

Better Together

Kenneth Church

@kchurch4

- Github
- Author Page
- Homepage

Go To Project

It is standard practice to represent documents as embeddings. Embeddings based on deep nets (BERT) capture text and other embeddings based on node2vec and GNNs (graph neural nets) capture citation graphs.

We evaluate these embeddings and show that combinations of text and citations are better than either by itself on standard benchmarks of downstream tasks. Embeddings are available for a range of applications: ranked retrieval, recommender systems and routing papers to reviewers.

Find Similar Papers

Search by Paper	**Search by Author**
Embedding (or API): ProNE-s	Embedding: ProNE-s
Limit: 20	Limit: 20
Query by Paper (Paper id or keywords + <enter>)	Query by Author (Author id or name + <enter>)

Help Bulk Download
GitHub Final Report (YouTube)
JSALT-2023 Contact us (by email)
BETA Version

The Institute for Experiential AI

Northeastern University

Le Mans Université

Johns Hopkins Whiting School of Engineering

allomedia

Google Search: Gallery Semantic Scholar

Find Similar Papers

Search by Paper

Embedding (or API):

ProNE-s | ProNE-s

Limit: 20

Query by Paper (Paper id or keywords + <enter>) | Query

Help | Bulk Download

GitHub | Final Report (YouTube)

JSALT-2023 | Contact us (by email)

BETA Version

The Institute for Experienti

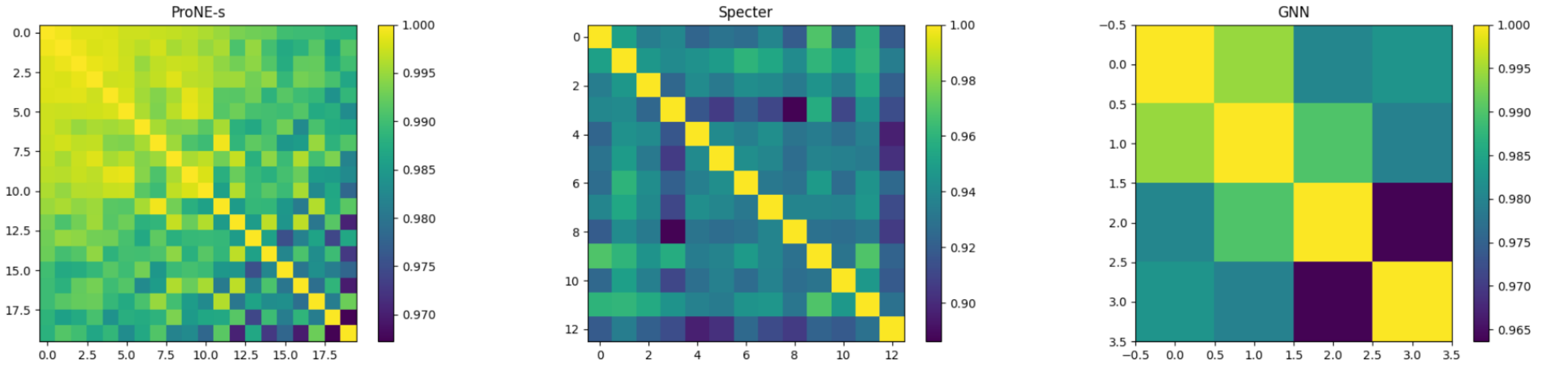
Northeastern University

Le Mans Université

allomedia

Johns Hopkins Whiting School of Engineering

Paper: InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning



Top

score	citationCount	Paper	Authors	year	More like this	Compare & Contrast	ProNE
	662	InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning	Wenliang Dai, Junnan Li, ..., Steven C. H. Hoi	**2023**	Similar to this	Compare & Contrast	**1.0**
0.999	431	mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality	Qinghao Ye, Haiyang Xu, ..., Feiyan Huang	2023	Similar to this	Compare & Contrast	0.999
0.998	152	MultiModal-GPT: A Vision and Language Model for Dialogue with Humans	T. Gong, Chengqi Lyu, ..., Kai Chen	2023	Similar to this	Compare & Contrast	0.998
0.998	300	Otter: A Multi-Modal Model with In-Context Instruction Tuning	Bo Li, Yuanhan Zhang, ..., Ziwei Liu	2023	Similar to this	Compare & Contrast	0.998
0.998	225	MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models	Chaoyou Fu, Peixian Chen, ..., Rongrong Ji	2023	Similar to this	Compare & Contrast	0.998
0.998	142	SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension	Bohao Li, Rui Wang, ..., Ying Shan	2023	Similar to this	Compare & Contrast	0.998
0.997	546	Improved Baselines with Visual Instruction Tuning	Haotian Liu, Chunyuan Li, ..., Yong Jae Lee	2023	Similar to this	Compare & Contrast	0.997
0.997	818	MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models	Deyao Zhu, Jun Chen, ..., Mohamed Elhoseiny	2023	Similar to this	Compare & Contrast	0.997
0.997	154	OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models	Anas Awadalla, Irena Gao, ..., Ludwig Schmidt	2023	Similar to this	Compare & Contrast	0.997
0.997	187	MMBench: Is Your Multi-modal Model an All-around Player?	Yuanzhan Liu, Haodong Duan, ..., Dahua Lin	2023	Similar to this	Compare & Contrast	0.997

APIs

API	Examples	Arguments	Description
Paper Search	example	help, query, fields	• Find papers matching input query (a string); output fields from S2 for each paper. • See documentation on fields for more information on fields in S2. • A common use case is to request paper ids from titles of papers since many of the APIs below are based on ids in Semantic Scholar (and other sources).
Author Search	example	help, query, fields	• Find authors matching input query (a string); output fields from S2 for each author. • See documentation on fields for more information on fields in S2. • Note: author fields are different from paper fields.
Lookup Paper	simple example, more challenging example	help, id, fields, embeddings	• Input one or more comma separated paper id and output fields from S2, as well as embeddings. • If embeddings argument is specified, then output embedding vectors for each input paper (missing values will have vectors of 0). • See documentation on embeddings for details on how to specify combinations of different embeddings to return.
Lookup Author	example	help, id, fields	• Input author id and output author fields from S2. • Note: author ids are different from paper ids and author fields are different from paper fields.
Lookup Citations	example	help, offset (defaults to 0), limit (defaults to 100; max is 1000), id, fields	• Lookup Citations for paper id and output fields from S2 for each citation. • A useful field to request is contexts; that field returns citing sentences, sentences from other papers that cite the input paper id. • For papers with more than 1000 citations, call this API multiple times with different offsets.
Coauthors	example	help, query, after_year	• Input query (a string); for each matching author ids, returns a list of coauthors filtered by after_year (a 4 digit number). • Note: since Semantic Scholar may have multiple author ids for the same author, the json object contains a list of coauthors for each author matching the input query
Recommend Papers	example	help, id, limit, method, fields, sort_by, score1, score2	• Recommend papers similar to paper id using method. • See documentation on method for choices of methods that are currently supported. • Output fields from S2 for each recommended paper. • The optional arguments, score1 and score2, score recommendations one at a time (for score1) and pairwise (for score2), using one or more of four embeddings.
Recommend Authors	example	help, id, limit, method, fields, sort_by, score1, score2	• Recommend authors near paper id using method • Output fields from S2 for each recommended author.
Compare and Contrast	example1, example2 example2	help, ids (two or more ids, separated by commas)	• Use RAG to compare and contrast the first id with the rest.
Compare and Contrast Texts	example	help, text1, text2	• Use RAG to compare and contrast text1 with text2, where both texts are strings.

APIs

- Paper Search:
  - find paper ids matching input query
- Author Search:
  - find author ids matching input query
- Lookup Paper:
  - lookup fields and embeddings by id
- Lookup Author:
  - lookup fields by id
- Recommend Papers:
  - input paper id and output more paper ids
- Recommend Authors:
  - input paper id and output author ids
- Compare and Contrast:
  - use RAG to compare and contrast one paper id with more paper id(s)

- paper ids:
  - Includes ids from
    - Semantic Scholar (S2)
    - PubMed
    - ACL
    - arXiv
    - MAG (Microsoft Academic Graph)
- embeddings: Specter, ProNE, etc
- fields (from S2): properties of ids
  - title, authors, tldr, abstract, bibtex, references, citations

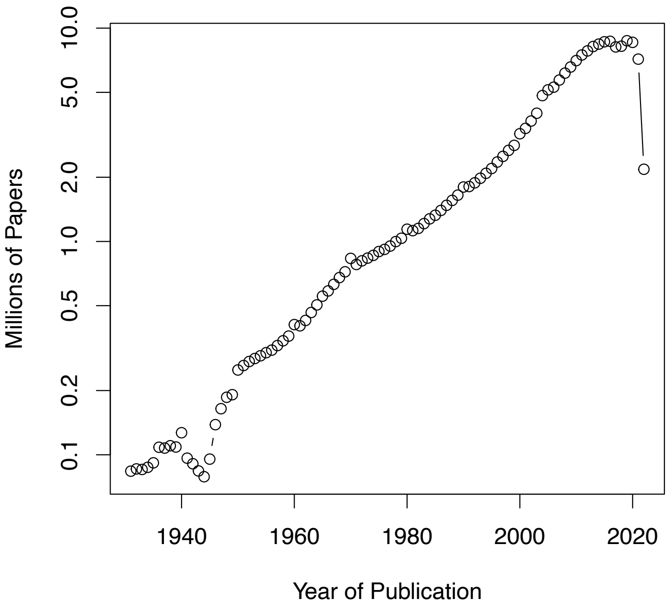
Surveys on Academic Search

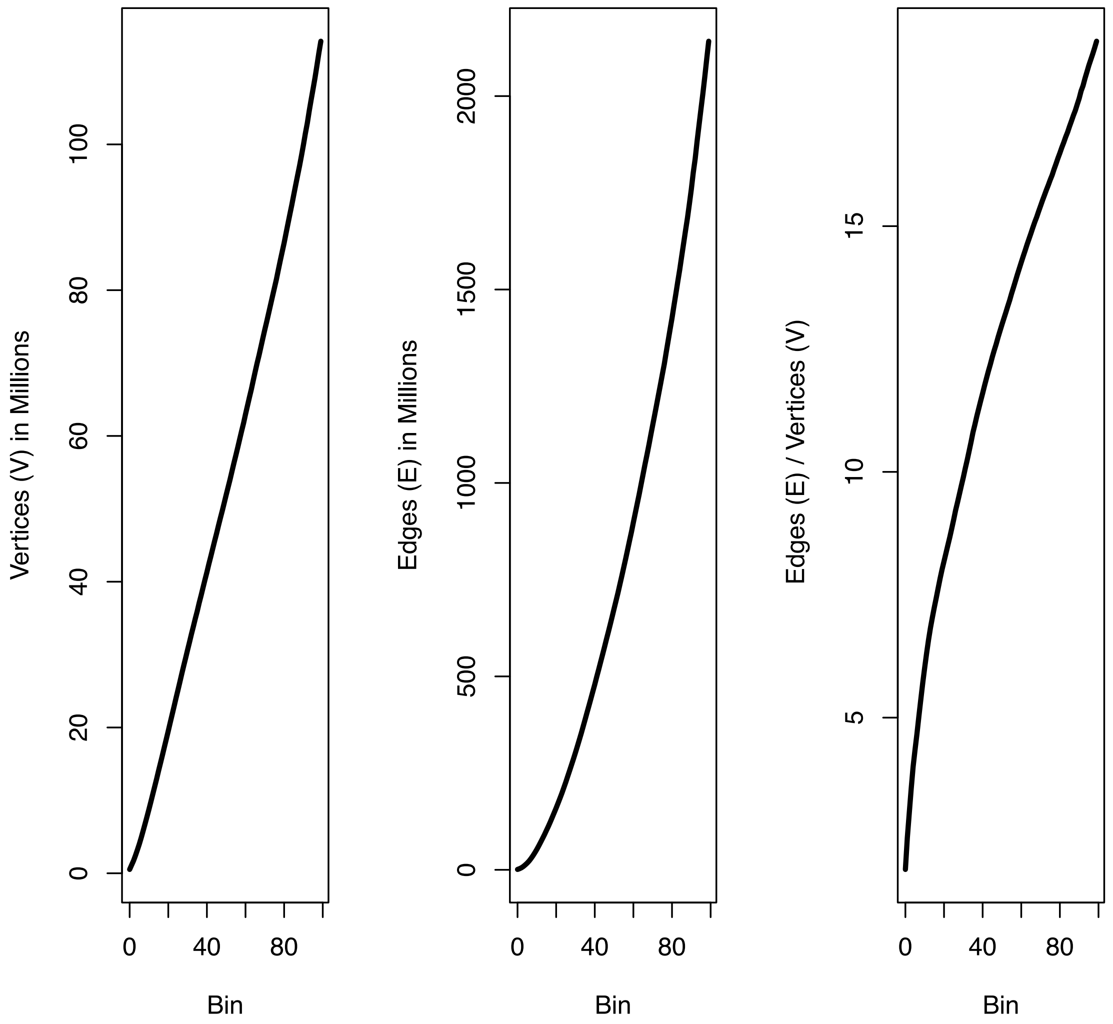
- Content-Based Filtering (CBF): Abstracts (Specter)
- Graph-Based Methods (GB): Citations (ProNE)
- Collaborative Filtering (CF): Clicks
- Better Together: Hybrids/Ensembles of above

- Why study academic search?
  - Academic search is like many important recommendation tasks
    - eCommerce (Product Recommendation), Traffic Analysis (Defense)
  - But data is less sensitive and available
- We will have little to say about CF
  - Because behavioral signals (clicks) are sensitive

Materials

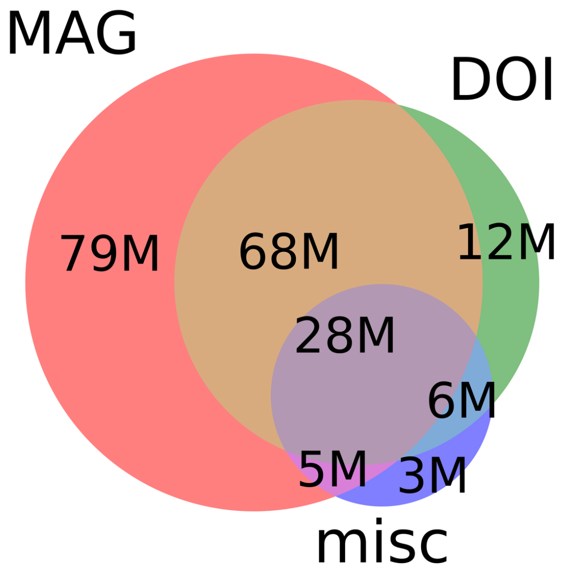
- Large and Growing
  - Semantic Scholar (S2)
    - 200+ Million Papers (nodes)
    - and 2+ Billion Citations (edges)
  - Literature doubles every 9 years!





More on Materials

- Seven Sources:
  - Two Big Sources: MAG, DOI
  - Five More (*misc*): PubMed, PubMedCentral, DBLP, arXiv, ACL
    - arXiv and ACL are tiny
- Many fields of study:
  - Medicine (45M), Chemistry (13M), CS (13M), Biology (13M), Materials Science (10M), Engineering (8M), Physics (7M), Psychology (7M), Mathematics (5M), Political Science (4M), Business (4M), Sociology (3M), Geography (3M), Economics (3M), Environmental Science (3M), Geology (3M), History (2M), Art (2M), Philosophy (1M)
  - Not just CS (Computer Science)

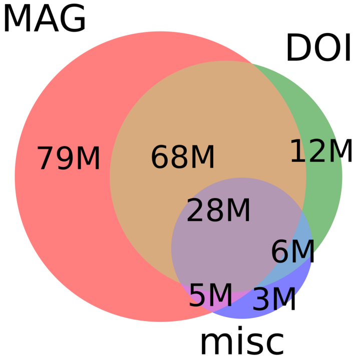


Semantic Scholar: Significant Effort (source: Dan Weld)

50 person team; 7 year project

207M+ papers; 2B Citations

8M+ monthly active users

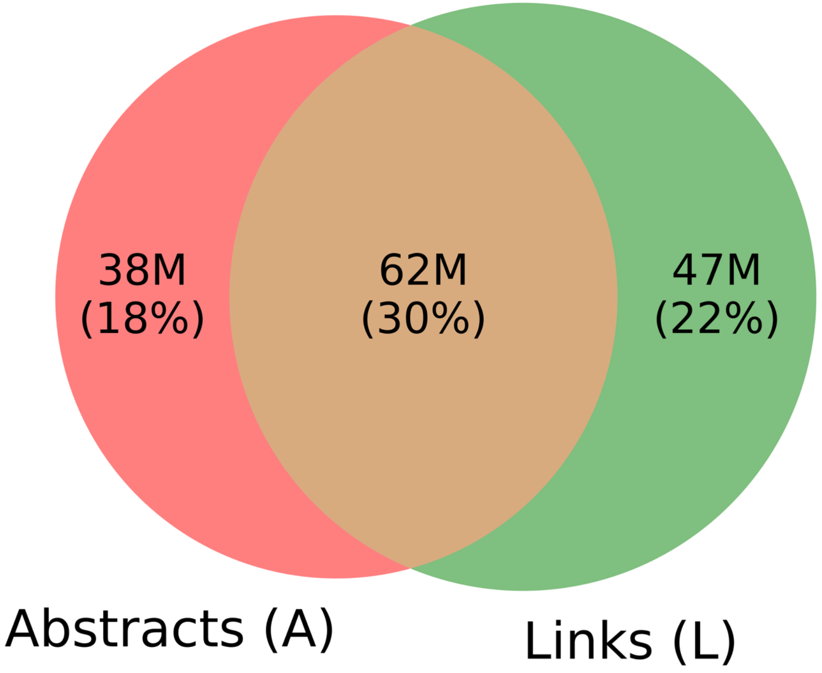


SCALE

	Source	Papers (millions)
1	CorpusId	207.80
2	MAG (Microsoft Academic Graph)	182.18
3	DOI	113.54
4	PubMed	35.03
5	DBLP	6.06
6	PubMedCentral	4.86
7	ArXiv	2.15
8	ACL	0.08

Why Better Together?

- Answer: Coverage
- GNNs assume papers have both
  - Abstracts (A), and
  - Links (L) in Citation Graph
- But not in Semantic Scholar
  - where $|A \cap L| \approx 30\%$
- Too many unrealistic benchmarks:
  - where $|A \cap L| \gg 30\%$



Motivating Scenario: Help Authors Write Sections on Related Work

- Four subtasks
  - **candidate list generation**: list papers to discuss,
  - **organization**: organize list by topic, time, etc.,
  - **summarization**: summarize papers, and
  - **connecting the dots**: explain how papers are relevant to the present discussion

Task: Find papers on ...	Query	CBF	GB
Recommender systems	[9]	[10–14]	[15–19]
Who should review what?	[20]	[21–26]	[27–31]
Citation Recommendation	[32]	[33–37]	[38–41]
RAG	[42]	[43–47]	[48–52]

Table 1: Complementary Recommendations.

Similarities and Differences: CBF & GB

Similarities of CBF & GB

- Embeddings
  - Both CBF and GB
    - represent papers as vectors
  - $sim(a, b) \approx \cos(vec(a), vec(b))$
  - Recommendation ≈ ANN
    - Approx nearest Neighbors
      - Query is a paper (or vector)
      - Output nearby papers in S2

Differences between CBF & GB

Feature	CBF	GB
Inputs	Titles and abstracts	Citation graph
Perspective	Authors' position	Audience response
Technology	Deep Nets & LLMs	Spectral Clustering
Discipline	Computer Science	Linear Algebra (Math)
Motivation	Use cases in NLP	Traffic Analysis
Embedding	Specter [6]	ProNE [53]
Interpretation of large cosines	Similar abstracts	Nearby in terms of random walks
Bottleneck	Cycles	Memory
Hardware	GPUs	Terabytes of RAM
Scale	Favor smaller graphs	Favor larger graphs
Invariance	Abstracts are invariant after publication	Citations accumulate after publication
Priors	More recent	More impact (cites)
Corner Cases	Non-English abstracts	Few links in graph

Table 3: Feature table for comparing CBF and GB methods for academic paper recommendation.

Differences

- Inputs:
  - Titles and abstracts for CBF;
  - citations for GB.
- Interpretations:
  - For CBF, large cosines indicate similar abstracts
  - for GB, large cosines indicate similarity in terms of random walks on citation graph.
- History:
  - Deep networks evolved from NLP
  - whereas spectral clustering has roots in Linear Algebra
    - and was inspired by use cases such as traffic analysis in Applied Math.

Feature	**CBF**	**GB**
Inputs	Titles and abstracts	Citation graph
Perspective	Authors' position	Audience response
Technology	Deep Nets & LLMs	Spectral Clustering
Discipline	Computer Science	Linear Algebra (Math)
Motivation	Use cases in NLP	Traffic Analysis
Embedding	Specter [6]	ProNE [53]
Interpretation of large cosines	Similar abstracts	Nearby in terms of random walks
Bottleneck	Cycles	Memory
Hardware	GPUs	Terabytes of RAM
Scale	Favor smaller graphs	Favor larger graphs
Invariance	Abstracts are invariant after publication	Citations accumulate after publication
Priors	More recent	More impact (cites)
Corner Cases	Non-English abstracts	Few links in graph

Table 3: Feature table for comparing CBF and GB methods for academic paper recommendation.

Differences

- Implementation Details:
  - We use Specter (a deep network) for CBF and ProNE (spectral clustering) for GB.
- Computational Bottlenecks:
  - Deep networks are limited by computational cycles,
    - whereas spectral clustering is limited by memory.
  - We use GPUs for deep networks,
    - and terabytes of RAM to compute SVDs for spectral clustering

Feature	CBF	GB
Inputs	Titles and abstracts	Citation graph
Perspective	Authors' position	Audience response
Technology	Deep Nets & LLMs	Spectral Clustering
Discipline	Computer Science	Linear Algebra (Math)
Motivation	Use cases in NLP	Traffic Analysis
Embedding	Specter [6]	ProNE [53]
Interpretation of large cosines	Similar abstracts	Nearby in terms of random walks
Bottleneck	Cycles	Memory
Hardware	GPUs	Terabytes of RAM
Scale	Favor smaller graphs	Favor larger graphs
Invariance	Abstracts are invariant after publication	Citations accumulate after publication
Priors	More recent	More impact (cites)
Corner Cases	Non-English abstracts	Few links in graph

Table 3: Feature table for comparing CBF and GB methods for academic paper recommendation.

More Differences

- Scale:
  - Larger graphs favor GB because of network effects.
- Time Invariance:
  - CBF embeddings are time invariant because abstracts do not change after publication;
  - GB embeddings improve as papers accumulate citations over time.
- Priors:
  - GB recommendations have more citations,
  - but are less recent
    - (because it takes time to accumulate citations).
- Corner Cases and Missing Values:
  - Multiple perspectives create opportunities to improve robustness and coverage with error detection and imputing missing values.

Feature	CBF	GB
Inputs	Titles and abstracts	Citation graph
Perspective	Authors' position	Audience response
Technology	Deep Nets & LLMs	Spectral Clustering
Discipline	Computer Science	Linear Algebra (Math)
Motivation	Use cases in NLP	Traffic Analysis
Embedding	Specter [6]	ProNE [53]
Interpretation of large cosines	Similar abstracts	Nearby in terms of random walks
Bottleneck	Cycles	Memory
Hardware	GPUs	Terabytes of RAM
Scale	Favor smaller graphs	Favor larger graphs
Invariance	Abstracts are invariant after publication	Citations accumulate after publication
Priors	More recent	More impact (cites)
Corner Cases	Non-English abstracts	Few links in graph

Table 3: Feature table for comparing CBF and GB methods for academic paper recommendation.

Differences in More Detail

Feature	CBF	GB
Inputs	Titles and abstracts	Citation graph
Perspective	Authors' position	Audience response
Technology	Deep Nets & LLMs	Spectral Clustering
Discipline	Computer Science	Linear Algebra (Math)
Motivation	Use cases in NLP	Traffic Analysis
Embedding	Specter [6]	ProNE [53]
Interpretation of large cosines	Similar abstracts	Nearby in terms of random walks
Bottleneck	Cycles	Memory
Hardware	GPUs	Terabytes of RAM
Scale	Favor smaller graphs	Favor larger graphs
Invariance	Abstracts are invariant after publication	Citations accumulate after publication
Priors	More recent	More impact (cites)
Corner Cases	Non-English abstracts	Few links in graph

Table 3: Feature table for comparing CBF and GB methods for academic paper recommendation.

Inputs and Perspectives

- Inputs
  - CBF is based on abstracts
  - and GB is based on citations
- Many of the differences
  - are consequences of inputs
- Perspectives
  - Abstracts:
    - authors’ perspective
  - Citations:
    - responses from audience

Feature	**CBF**	**GB**
Inputs	Titles and abstracts	Citation graph
Perspective	Authors’ position	Audience response
Technology	Deep Nets & LLMs	Spectral Clustering
Discipline	Computer Science	Linear Algebra (Math)
Motivation	Use cases in NLP	Traffic Analysis
Embedding	Specter [6]	ProNE [53]
Interpretation of large cosines	Similar abstracts	Nearby in terms of random walks
Bottleneck	Cycles	Memory
Hardware	GPUs	Terabytes of RAM
Scale	Favor smaller graphs	Favor larger graphs
Invariance	Abstracts are invariant after publication	Citations accumulate after publication
Priors	More recent	More impact (cites)
Corner Cases	Non-English abstracts	Few links in graph

Table 3: Feature table for comparing CBF and GB methods for academic paper recommendation.

History

- CBF and GB come from
  - different disciplines
  - with different motivations
- Disciplines
  - CBF: Deep Nets & LLMs from CS
  - GB: Eigenvectors (like Page Rank)
    - Linear Algebra
    - Spectral Clustering
    - Applied Math
- Motivations (Use Cases)
  - GB: Traffic Analysis
    - Know who is talking to who
    - But not what they are saying
  - CBF: NLP use cases
    - Situation is reversed
      - Know the content, but not the context of
      - how documents are connected to one another

Feature	**CBF**	**GB**
Inputs	Titles and abstracts	Citation graph
Perspective	Authors' position	Audience response
Technology	Deep Nets & LLMs	Spectral Clustering
Discipline	Computer Science	Linear Algebra (Math)
Motivation	Use cases in NLP	Traffic Analysis
Embedding	Specter [6]	ProNE [53]
Interpretation of large cosines	Similar abstracts	Nearby in terms of random walks
Bottleneck	Cycles	Memory
Hardware	GPUs	Terabytes of RAM
Scale	Favor smaller graphs	Favor larger graphs
Invariance	Abstracts are invariant after publication	Citations accumulate after publication
Priors	More recent	More impact (cites)
Corner Cases	Non-English abstracts	Few links in graph

Table 3: Feature table for comparing CBF and GB methods for academic paper recommendation.

Interpretation

- Large cosines suggest papers are similar to one another
  - But for different reasons
- CBF: large cosines →
  - abstracts use similar words
  - in terms of LLMs
- GB: large cosines →
  - papers are near one another
  - in terms of random walks

Feature	CBF	GB
Inputs	Titles and abstracts	Citation graph
Perspective	Authors' position	Audience response
Technology	Deep Nets & LLMs	Spectral Clustering
Discipline	Computer Science	Linear Algebra (Math)
Motivation	Use cases in NLP	Traffic Analysis
Embedding	Specter [6]	ProNE [53]
Interpretation of large cosines	Similar abstracts	Nearby in terms of random walks
Bottleneck	Cycles	Memory
Hardware	GPUs	Terabytes of RAM
Scale	Favor smaller graphs	Favor larger graphs
Invariance	Abstracts are invariant after publication	Citations accumulate after publication
Priors	More recent	More impact (cites)
Corner Cases	Non-English abstracts	Few links in graph

Table 3: Feature table for comparing CBF and GB methods for academic paper recommendation.

Implementation Details

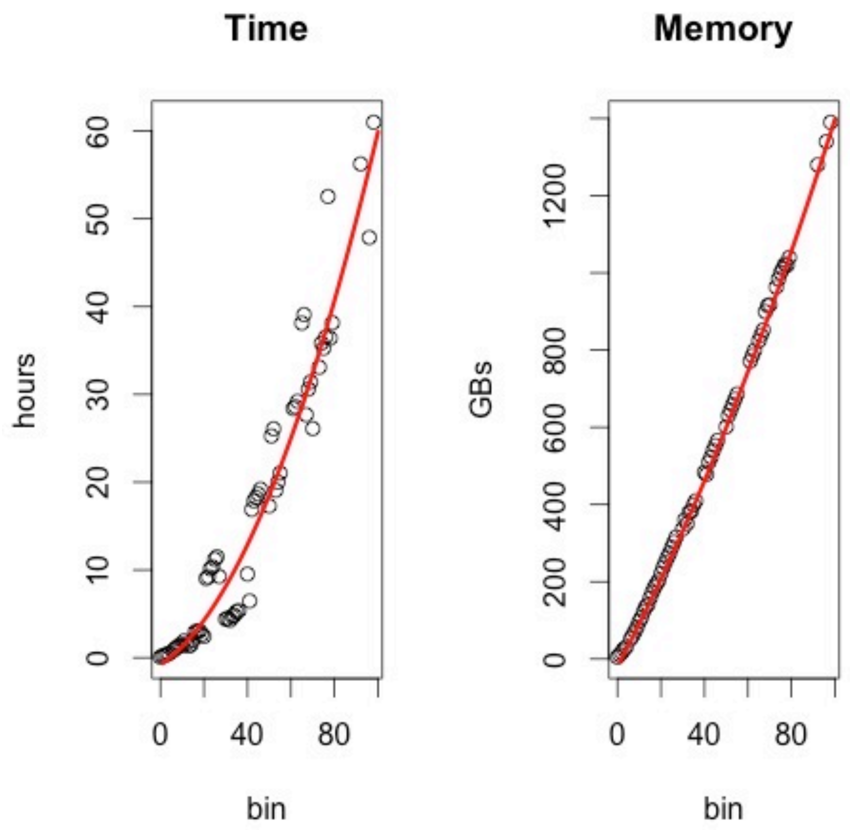
- Embeddings
  - CBF: Specter, a BERT-like deep net
  - GB: ProNE, spectral clustering
- For Specter,
  - no need for training or inference
    - because S2 distributes models and vectors
  - citations are not used for inference,
    - but are used for fine-tuning
- For ProNE, we had to compute them ourselves (heavy lifting)

Feature	CBF	GB
Inputs	Titles and abstracts	Citation graph
Perspective	Authors' position	Audience response
Technology	Deep Nets & LLMs	Spectral Clustering
Discipline	Computer Science	Linear Algebra (Math)
Motivation	Use cases in NLP	Traffic Analysis
Embedding	Specter [6]	ProNE [53]
Interpretation of large cosines	Similar abstracts	Nearby in terms of random walks
Bottleneck	Cycles	Memory
Hardware	GPUs	Terabytes of RAM
Scale	Favor smaller graphs	Favor larger graphs
Invariance	Abstracts are invariant after publication	Citations accumulate after publication
Priors	More recent	More impact (cites)
Corner Cases	Non-English abstracts	Few links in graph

Table 3: Feature table for comparing CBF and GB methods for academic paper recommendation.

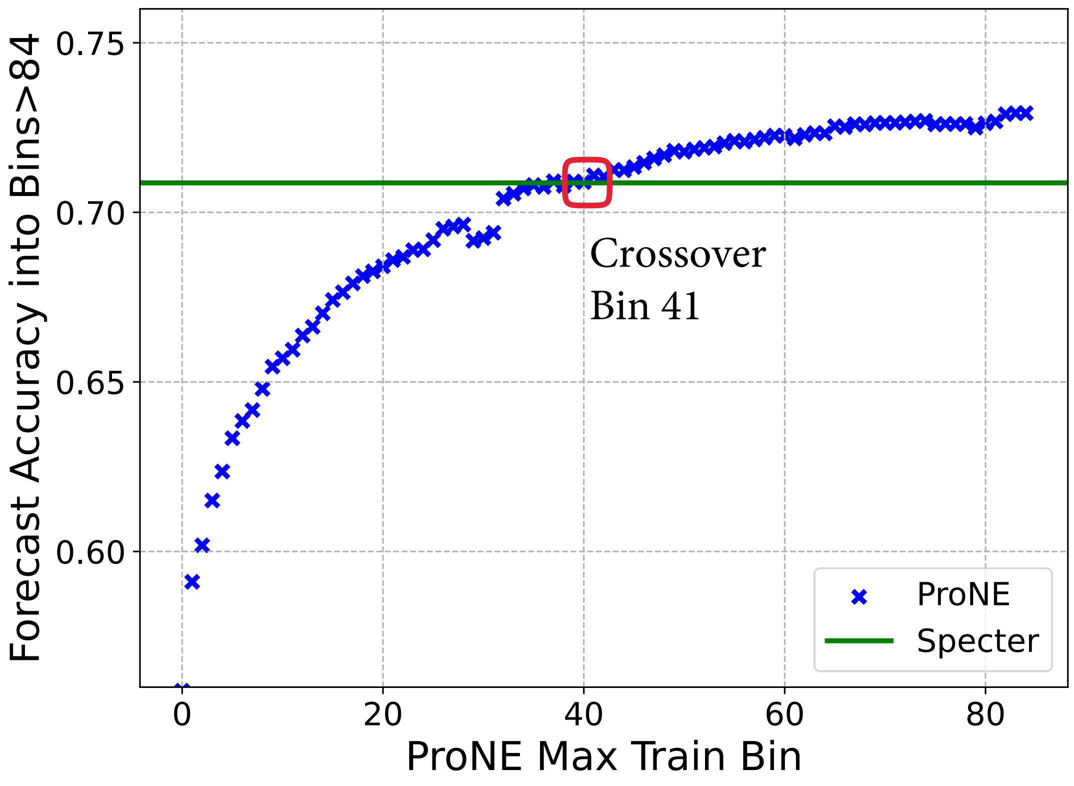
More Implementation Details

- ProNE heavy lifting
  - It takes a few days and a few TBs
  - to compute SVD for larger graphs
    - 100 bins $\approx 100\%$ of S2
- Specter fine-tuning
  - Start with SciBERT
  - fine-tune with a few million triples
    - $< query, pos, neg >$
- ProNE trains on billions
  - as opposed to millions



Scale

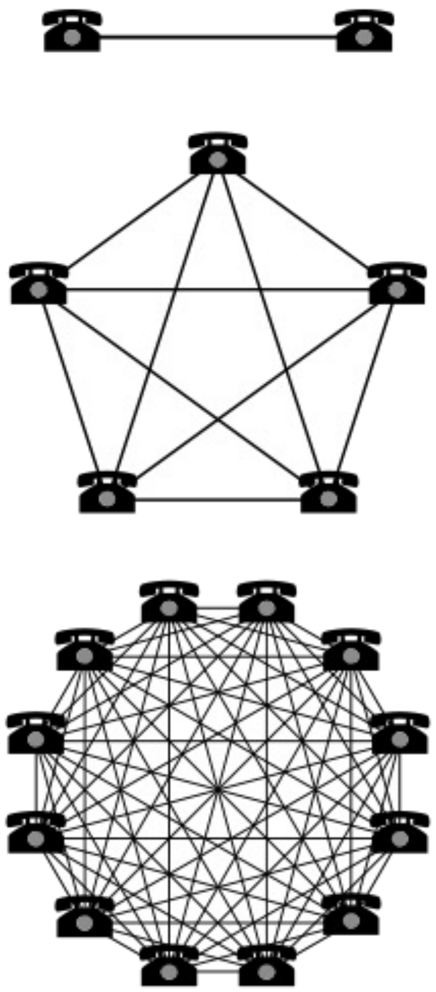
- Scale favors ProNE (GB)
  - because of network effects
- Network effects (Metcalfe's Law)
  - nodes: n
  - edges: n^2
  - paths: 2^n
- Citation Prediction Task
  - Does paper a cite paper b?
  - Given pairs: a, b that are 1-4 hops apart
  - Classification: is this pair 1 hop apart?
- Binning:
  - Train 100 ProNE models



	Vertices (Papers)	Edges (Links)
ogbl-citation2	3M	31M
crossover (bin 41)	42M	499M

Table 2: $|\text{OGB}| \ll$ crossover

Metcalfe's Law (Network Effects)



- History: 3Com was selling small networks
  - 3 = 1 printer + 2 computers
  - Metcalfe argued they should sell bigger networks
    - (and more 3Com products)
    - because of economies of scale
- Economy of Scale:
  - Benefits scale faster than costs
    - Benefits: $\sim n^2$
    - Costs: $\sim n$
- Law has been good for AT&T, Google, Social Media
- Hypo: also good for Academic Search
  - Consequently, we should experiment with large graphs

Time Invariance and Priors

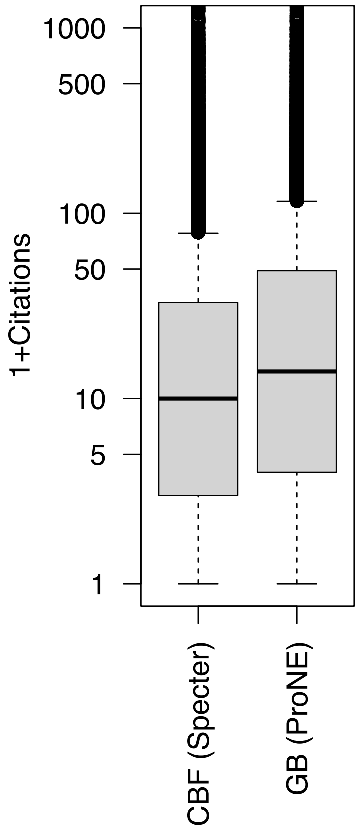
- Time Invariance
  - Authors cannot change abstracts after publication
    - But audience perspective evolves over time
    - Citations accumulate years after publication
- Priors
  - GB returns papers that have more citations but less recent
    - Because it can take time for papers to accumulate citations

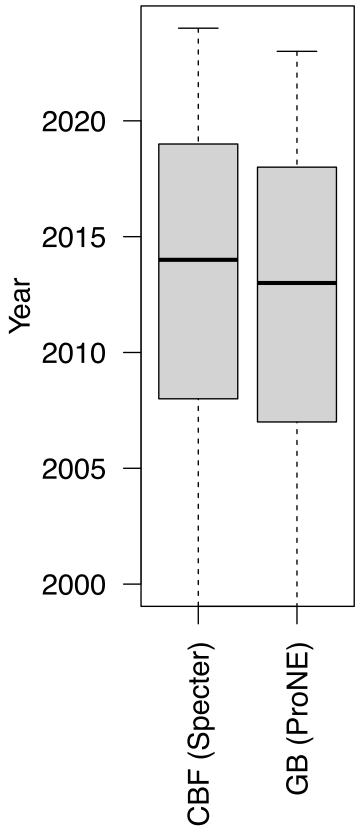
Feature	CBF	GB
Inputs	Titles and abstracts	Citation graph
Perspective	Authors' position	Audience response
Technology	Deep Nets & LLMs	Spectral Clustering
Discipline	Computer Science	Linear Algebra (Math)
Motivation	Use cases in NLP	Traffic Analysis
Embedding	Specter [6]	ProNE [53]
Interpretation of large cosines	Similar abstracts	Nearby in terms of random walks
Bottleneck	Cycles	Memory
Hardware	GPUs	Terabytes of RAM
Scale	Favor smaller graphs	Favor larger graphs
Invariance	Abstracts are invariant after publication	Citations accumulate after publication
Priors	More recent	More impact (cites)
Corner Cases	Non-English abstracts	Few links in graph

Table 3: Feature table for comparing CBF and GB methods for academic paper recommendation.

Time Invariance and Priors

- Time Invariance
  - Authors cannot change abstracts after publication
    - But audience perspective evolves over time
    - Citations accumulate years after publication
- Priors
  - GB returns papers that have more citations but less recent
    - Because it can take time for papers to accumulate citations





Corner Cases

- Multiple perspectives create opportunities for robustness
  - Detect corner cases by looking for large differences in cosines
    - Duplicate docs:
      - large CBF cosines,
      - but small GB cosines
  - Common corner cases for CBF
    - Duplicate documents
    - Missing/bogus abstracts
    - Non-English (Chinese)
  - GB does not suffer from these corner cases
    - But GB has other corner cases
    - Few (if any) links in citation graph

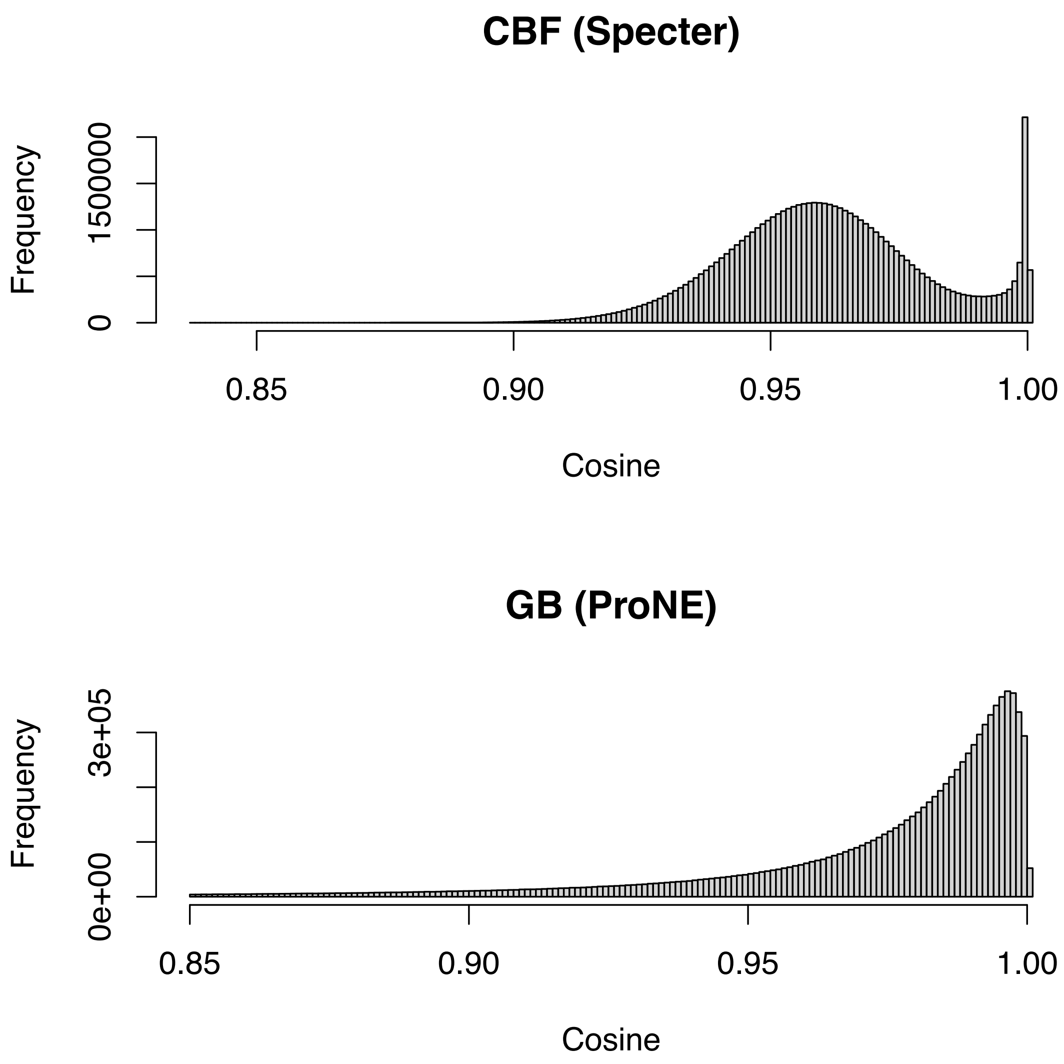
Feature	CBF	GB
Inputs	Titles and abstracts	Citation graph
Perspective	Authors' position	Audience response
Technology	Deep Nets & LLMs	Spectral Clustering
Discipline	Computer Science	Linear Algebra (Math)
Motivation	Use cases in NLP	Traffic Analysis
Embedding	Specter [6]	ProNE [53]
Interpretation of large cosines	Similar abstracts	Nearby in terms of random walks
Bottleneck	Cycles	Memory
Hardware	GPUs	Terabytes of RAM
Scale	Favor smaller graphs	Favor larger graphs
Invariance	Abstracts are invariant after publication	Citations accumulate after publication
Priors	More recent	More impact (cites)
Corner Cases	Non-English abstracts	Few links in graph

Table 3: Feature table for comparing CBF and GB methods for academic paper recommendation.

$$cos(q, cand_1)$$

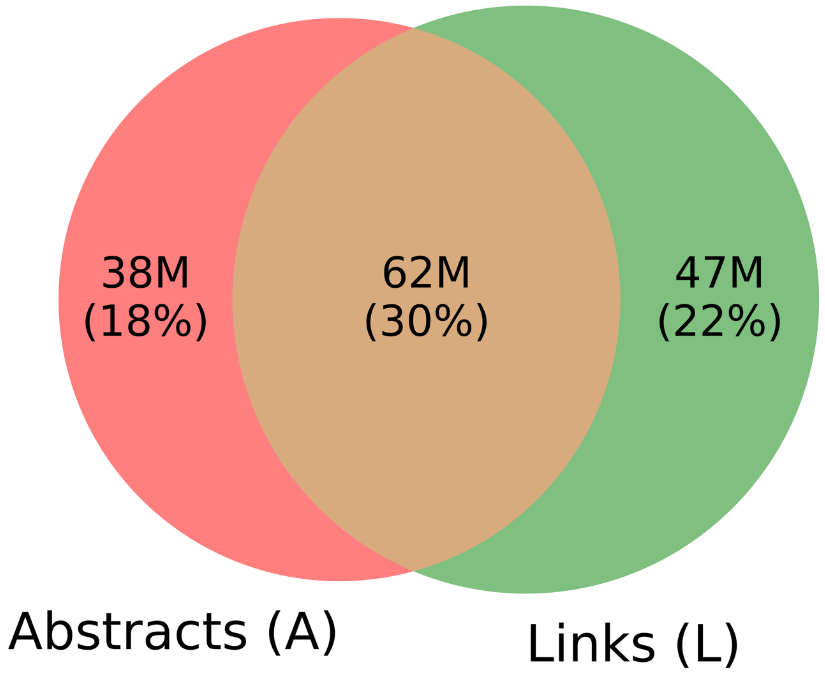
Corner Cases

- Multiple perspectives create opportunities for robustness
  - Detect corner cases by looking for large differences in cosines
    - Duplicate docs:
      - large CBF cosines,
      - but small GB cosines
  - Common corner cases for CBF
    - Duplicate documents
    - Missing/bogus abstracts
    - Non-English (Chinese)
  - GB does not suffer from these corner cases
    - But GB has other corner cases
    - Few (if any) links in citation graph



Corner Cases: Imputing Missing Values

- Centroid Approximation
  - Infer a missing vector from
  - the average of its references
- Better Together Approximation
  - Infer a missing vector from
  - the average of papers
  - nearby in another embedding
- Synergies between CBF and GB



Conclusions: Better Together

- Soap Box:
  - Interdisciplinary Re-Org: ML/Vision/NLP/MT/IR/Speech
  - Responsible AI: Embrace Diversity
  - Multiple Perspectives:
    - Authors vs. Audience Response
    - Better Together
- Vision: More perspectives
  - subjectivity, diversity, semantics
- Academic Search Deliverables:
  - APIs, Website, Embeddings
- CBF & GB: Better Together
  - Complementary
  - Synergistic

Feature	CBF	GB
Inputs	Titles and abstracts	Citation graph
Perspective	Authors' position	Audience response
Technology	Deep Nets & LLMs	Spectral Clustering
Discipline	Computer Science	Linear Algebra (Math)
Motivation	Use cases in NLP	Traffic Analysis
Embedding	Specter [6]	ProNE [53]
Interpretation of large cosines	Similar abstracts	Nearby in terms of random walks
Bottleneck	Cycles	Memory
Hardware	GPUs	Terabytes of RAM
Scale	Favor smaller graphs	Favor larger graphs
Invariance	Abstracts are invariant after publication	Citations accumulate after publication
Priors	More recent	More impact (cites)
Corner Cases	Non-English abstracts	Few links in graph

Table 3: Feature table for comparing CBF and GB methods for academic paper recommendation.