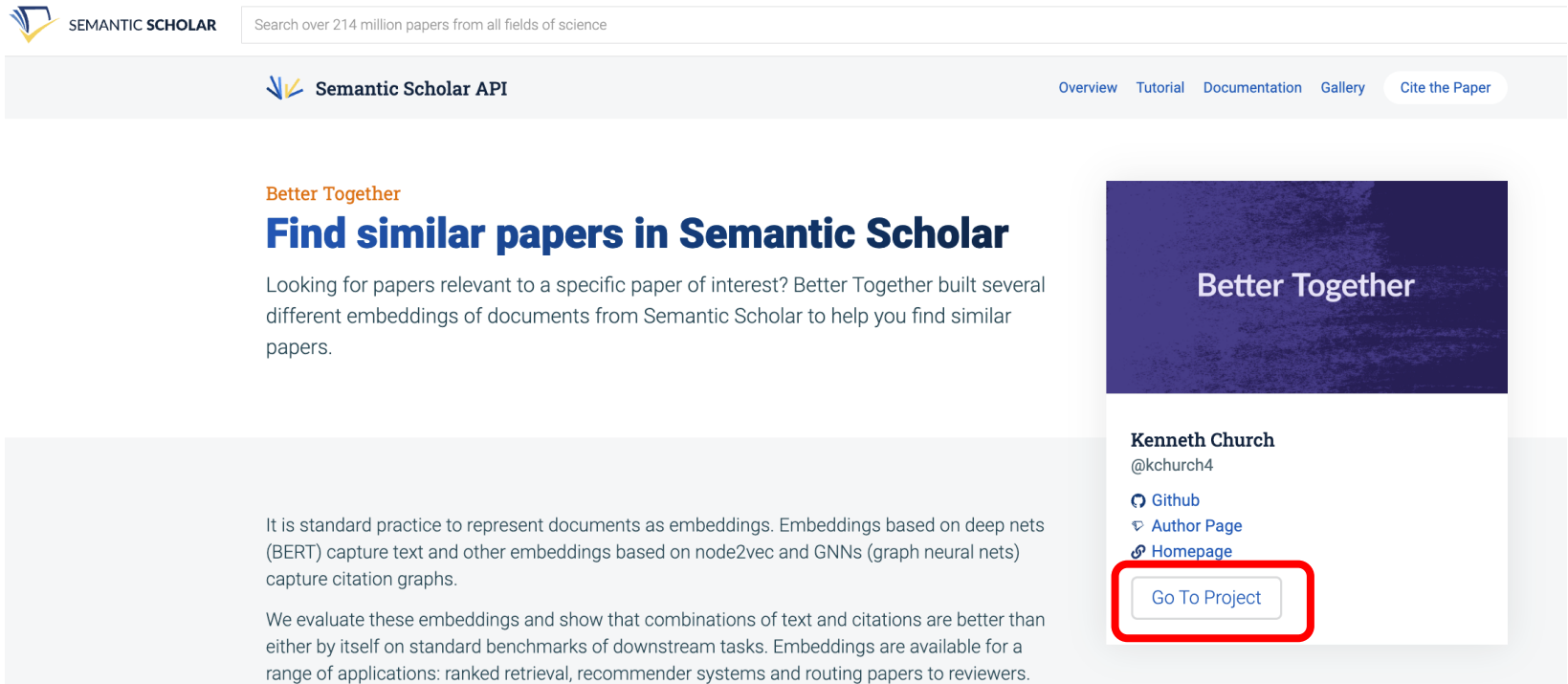
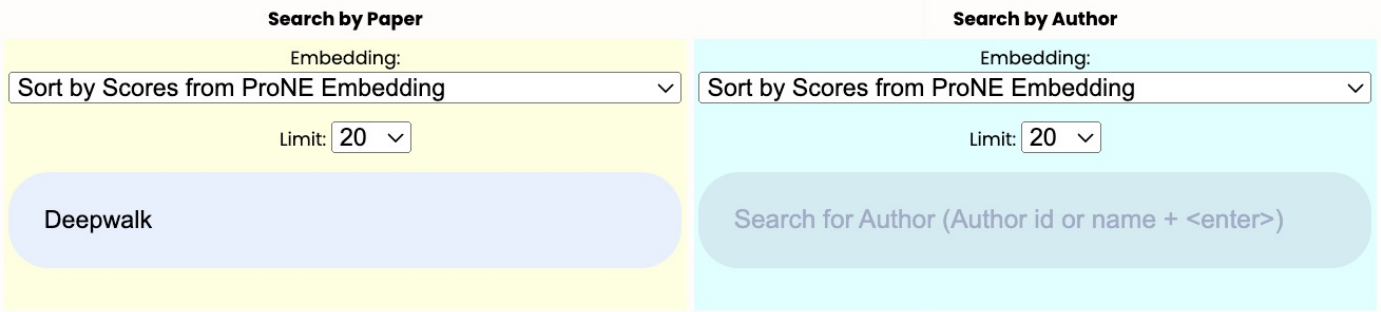
Better Together: Text + Context

Jan 18, 2024

https://www.semanticscholar.org/api-gallery/better-together



Find Similar Papers



Help GitHub Final Report (YouTube) JSALT-2023 Comments Appreciated BETA Version

Le Mans Université

JOHNS HOPKINS WHITING SCHOOL of ENGINEERING

allomedia

EAI The Institute for Experiential AI Northeastern University

Recommendation Tasks

- For readers
  - What should I read?
- For authors
  - What should I cite?
- For conference organizers
  - Who should review what?

More than word overlap

- Many systems focus on ``relevance'' (word overlap)
  - But we also want credibility
  - Don't recommend papers that are buzz-word compliant
    - But not worth reading
    - (Most papers are never cited)

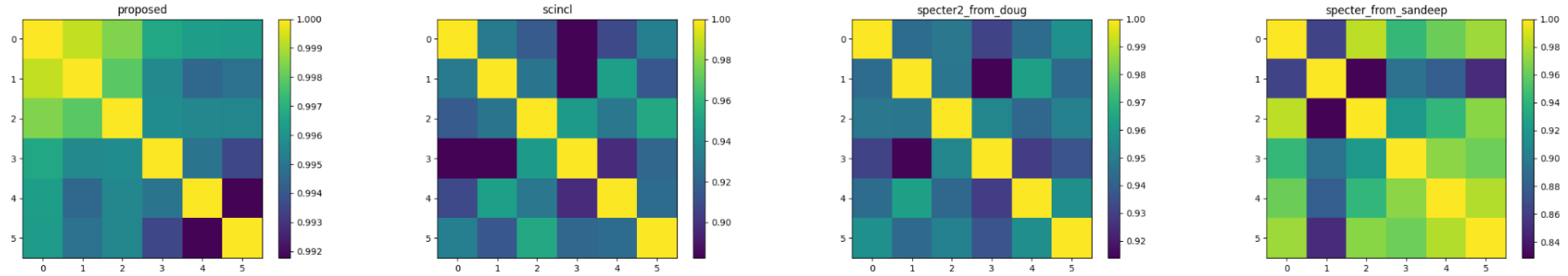
Paper Search Results

Home Page More Results S2 Recommendations bibtex
CorpusId: CorpusId:3051291
Limit: 5
Sort by: score
Embedding: proposed

Proposed Method

Table of Contents

Paper: DeepWalk: online learning of social representations



Top

score	citationCount	Paper	Authors	year	More like this	proposed	scincl	specter2_from_doug	specter_from_sandeep
	8182	DeepWalk: online learning of social representations	Bryan Perozzi, Rami Al-Rfou, S. Skiena	**2014**	**similar to this**	**1.0**	**1.0**	**1.0**	**1.0**
0.999	8779	node2vec: Scalable Feature Learning for Networks	Aditya Grover, J. Leskovec	2016	similar to this	0.999	0.931	0.944	0.864
0.998	4770	LINE: Large-scale Information Network Embedding	Jian Tang, Meng Qu, ..., Q. Mei	2015	similar to this	0.998	0.916	0.948	0.983
0.997	1511	A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications	Hongyun Cai, V. Zheng, K. Chang	2017	similar to this	0.997	0.883	0.932	0.943
0.996	1758	metapath2vec: Scalable Representation Learning for Heterogeneous Networks	Yuxiao Dong, N. Chawla, A. Swami	2017	similar to this	0.996	0.908	0.944	0.961
0.996	806	Network Embedding as Matrix Factorization: Unifying DeepWalk, LINE, PTE, and node2vec	J. Qiu, Yuxiao Dong, ..., Jie Tang	2017	similar to this	0.996	0.933	0.957	0.975

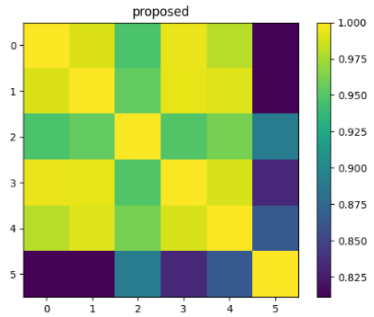
Recommendations from Semantic Scholar

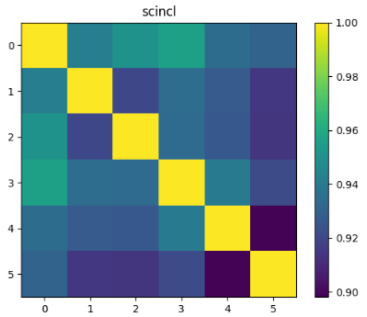
Paper Search Results

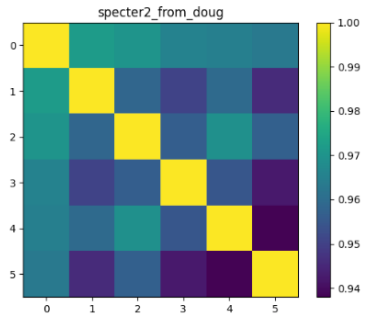
Home Page More Results Proposed bibtex
CorpusId: CorpusId:3051291
Limit: 5
Sort by: score
Embedding: s2_recommendations

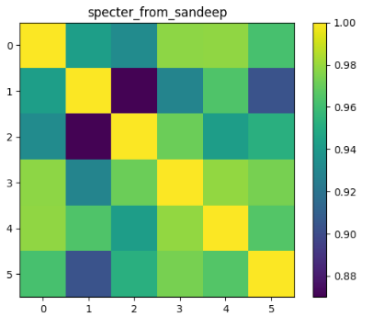
Table of Contents

Paper: DeepWalk: online learning of social representations









Top

score	citationCount	Paper	Authors	year	More like this	proposed	scincl	specter2_from_doug	specter_from_sandeep
	8182	DeepWalk: online learning of social representations	Bryan Perozzi, Rami Al-Rfou, S. Skiena	**2014**	**similar to this**	**1.0**	**1.0**	**1.0**	**1.0**
5	253	Max-Margin DeepWalk: Discriminative Learning of Network Representation	Cunchao Tu, Weicheng Zhang, ..., Maosong Sun	2016	similar to this	0.99	0.942	0.972	0.943
4	5	SimWalk: Learning network latent representations with social relation similarity	Shicheng Cui, Bin Xia, ..., Hong Zhang	2017	similar to this	0.947	0.951	0.97	0.933
3	165	Don't Walk, Skip!: Online Learning of Multi-scale Network Embeddings	Bryan Perozzi, Vivek Kulkarni, ..., S. Skiena	2016	similar to this	0.995	0.956	0.966	0.978
2	114	Discriminative Deep Random Walk for Network Classification	Juzheng Li, Jun Zhu, Bo Zhang	2016	similar to this	0.979	0.934	0.965	0.978
1	5	Learning distributed representations for large-scale dynamic social networks	Aakas Zhiyuli, Xun Liang, Zhiming Xu	2017	similar to this	0.811	0.931	0.963	0.962

Prior: More citations → More credible

Proposed

Spectral Clustering of Citation Graph

score	citationCount	Paper
	8182	DeepWalk: online learning of social representations
0.999	8779	node2vec: Scalable Feature Learning for Networks
0.998	4770	LINE: Large-scale Information Network Embedding
0.997	1511	A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications
0.996	1758	metapath2vec: Scalable Representation Learning for Heterogeneous Networks
0.996	806	Network Embedding as Matrix Factorization: Unifying DeepWalk, LINE, PTE, and node2vec

Semantic Scholar

BERT-like Deep Net

score	citationCount	Paper
	8182	DeepWalk: online learning of social representations
5	253	Max-Margin DeepWalk: Discriminative Learning of Network Representation
4	5	SimWalk: Learning network latent representations with social relation similarity
3	165	Don't Walk, Skip!: Online Learning of Multi-scale Network Embeddings
2	114	Discriminative Deep Random Walk for Network Classification
1	5	Learning distributed representations for large-scale dynamic social networks

Author Search Results

Home Page More Results S2 Recommendations
Author: David Madigan
Limit: 10
Sort by: score
Embedding: proposed

Table of Contents

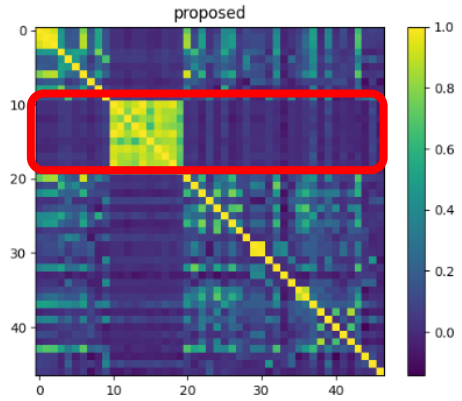
1. D. Madigan, hIndex: 62, citationCount: 19148
2. D. Madigan, hIndex: 11, citationCount: 573
3. D. Madigan, hIndex: 5, citationCount: 69
4. David Madigan, hIndex: 4, citationCount: 229
5. David Madigan, hIndex: 3, citationCount: 107
6. D. Madigan, hIndex: 2, citationCount: 23
7. David Madigan, hIndex: 2, citationCount: 204
8. David Madigan, hIndex: 2, citationCount: 28
9. David Madigan, hIndex: 1, citationCount: 8
10. David Madigan, hIndex: 1, citationCount: 16

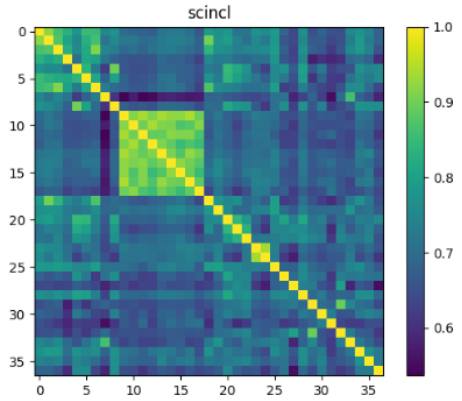
The role of flavanoid polyphenols in beer stability

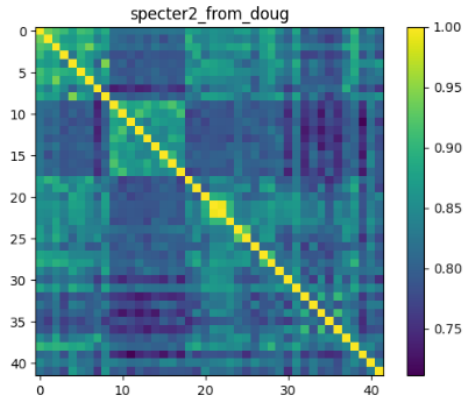
Determination of proanthocyanidins and catechins in beer and barley by high-performance liquid chromatography with dual-electrode electrochemical detection.

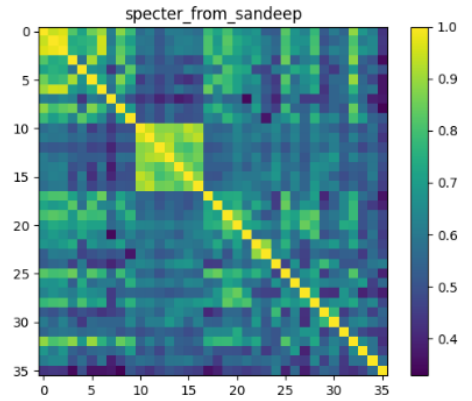
CONTROL OF FERULIC ACID AND 4-VINYL GUAIACOL IN BREWING

Semipreparative chromatographic procedure for the isolation of dimeric and trimeric proanthocyanidins from barley









Semantic Scholar (S2): Significant Effort (slide from Dan Weld)

50 person team

7 year project

207M+ scientific paper index

8M+ monthly active users

https://github.com/kwchurch/JSALT_Better_Together

Better Together Team Github

JSALT (Jelinek Summer Workshop on Speech and Language Technology):

Useful links

1. Final Report (YouTube) and JSALT-2023 Team Page
2. Documentation of web service (on Amazon AWS), with examples
3. Slides
4. large datasets from Globus and What's Where (on Globus)
5. Deliverables
6. Reading List
7. Notation
8. SciRepEval Baselines
9. Zoom Link and Meeting Notes
10. Status of Production Runs
11. Similar Venues

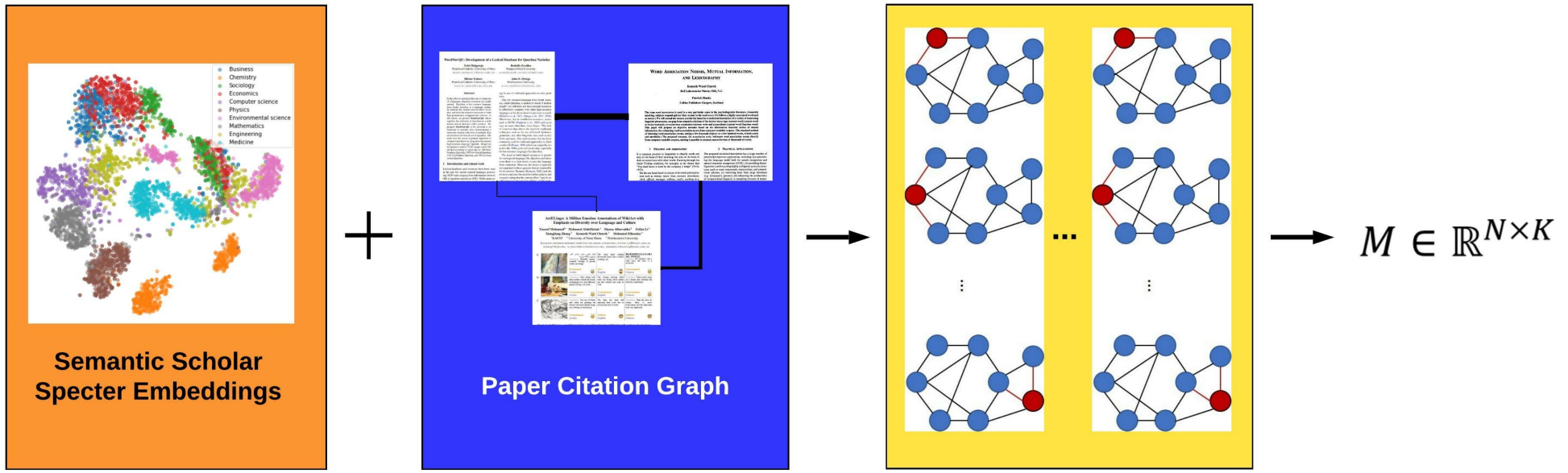
YouTube

Web Page

including these slides

Better Together:

Text (Titles, Abstracts, Body); Context (Citations)



TEXT **CONTEXT** **NEURAL NETWORKS (GNN / PRONE)** **EMBEDDINGS**

(a) (b) (c) (d)

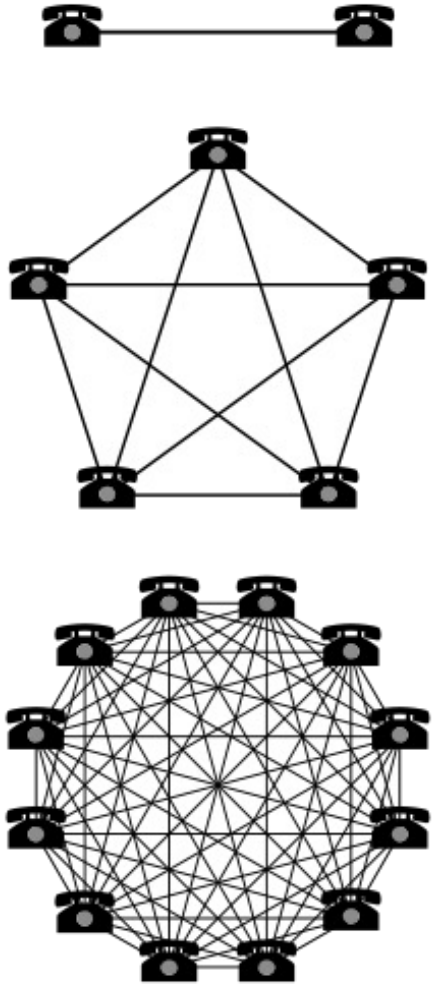
Specter, SciNCL

Proposed (ProNE)

Standard benchmarks are small, static and clean

Challenge	Standard	Proposed
Volume	Small	Large (Metcalfe's Law)
Velocity	Static	Growing (double in 9 years)
Variety	Clean	Dirty (missing/bad values)
Fore-casting	Random splits	Causal splits (bin papers by time)

Metcalfe's Law (Network Effects)

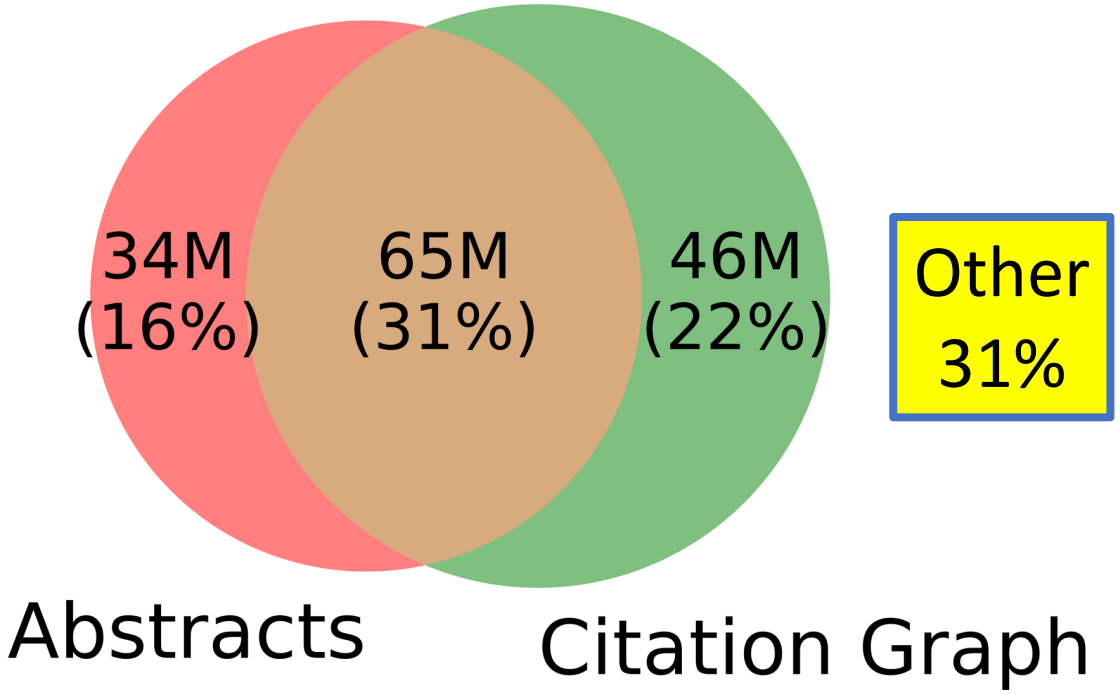


- History: 3Com was selling small networks
  - 3 = 1 printer + 2 computers
  - Metcalfe argued they should sell bigger networks
    - (and more 3Com products)
    - because of economies of scale
- Economy of Scale:
  - Benefits scale faster than costs
    - Benefits: $\sim n^2$
    - Costs: $\sim n$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

Realities (Dirty): Missing Values

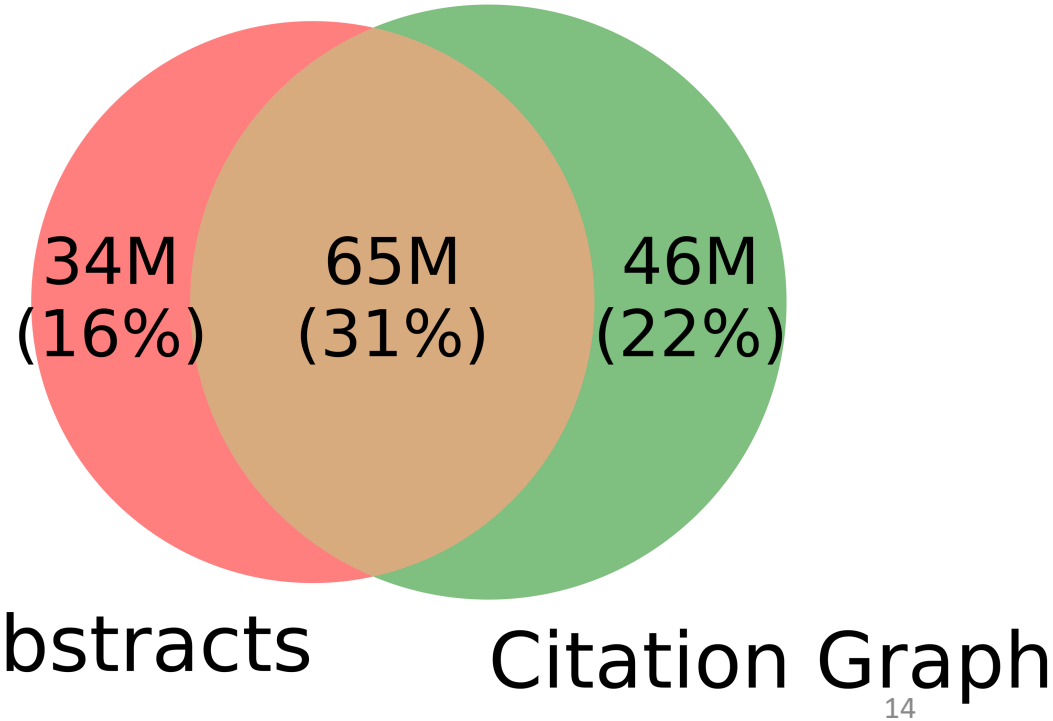
- A: papers with abstracts, a
- L: papers with links, l

- Opportunity: $A \cup L$
  - Prior work targets small subset
    - Specter: designed for A
    - GNNs: designed for $A \cap L$



Better-Together Conjectures

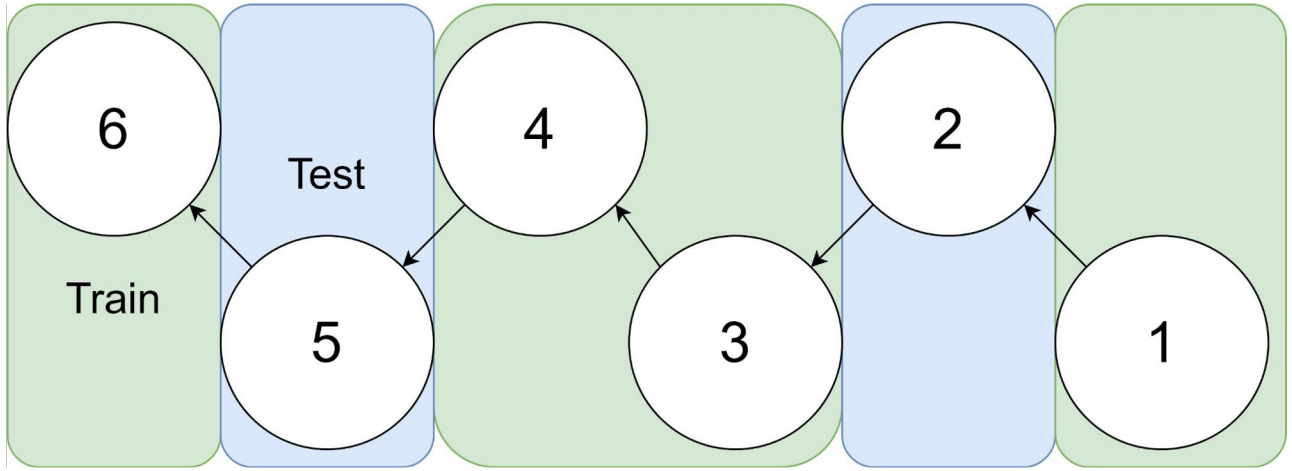
- Multiple representations are helpful
  - for missing values (and bad values)
    - If abstracts are missing, use links
    - If links are missing, use abstracts
  - and for graphs of different sizes
    - When graphs are small:
      - Text >> Links
    - when graphs are large:
      - Links >> Text (Metcalfe's Law)



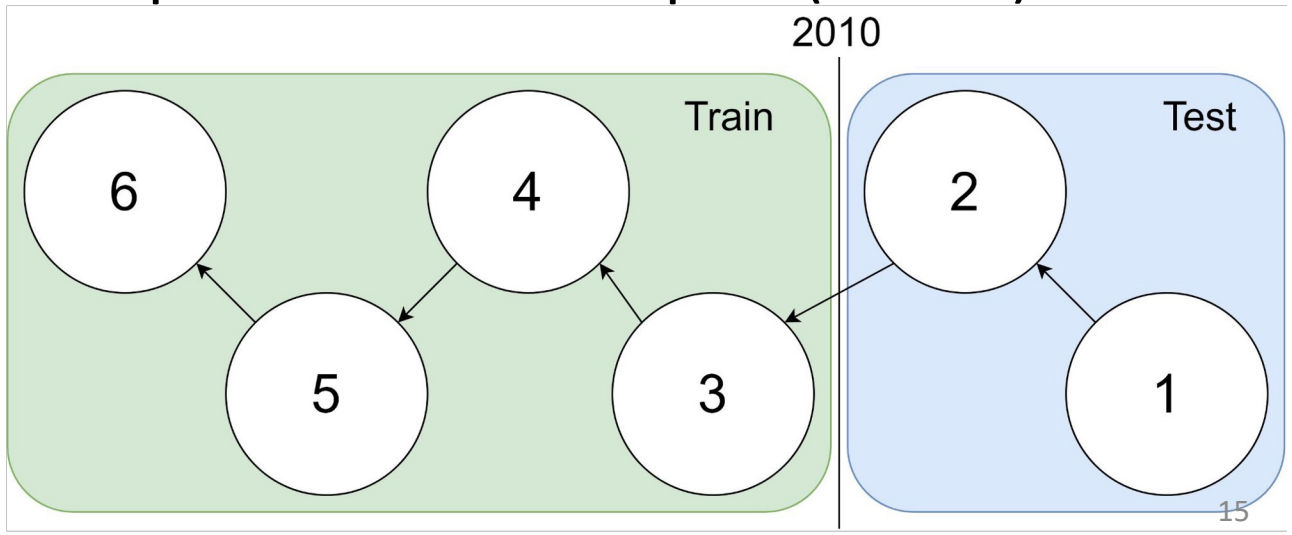
Example: 1 cites 2, 2 cites 3, … 5 cites 6

idx	Pub Year	Paper Title
1	2018	[...] Photofragment imaging
2	2016	Convenient probe of S(1D2)[...]
3	2005	Megapixel ion imaging [...]
4	2003	Direct current slice imaging [...]
5	1995	profiles of Cl(2Pj) photofragments [...]
6	1988	Adiabatic dissociation of [...]

Traditional Train-Test Split (Non-Causal)



Proposed Train-Test Splits (Causal)



Opportunities

Scale

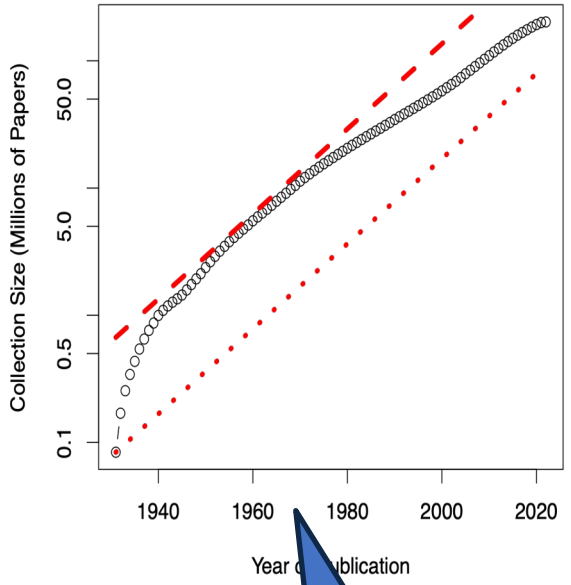
- Many benchmarks return a single number (figure of merit)
  - How does performance scale
  - with the size of the problem?

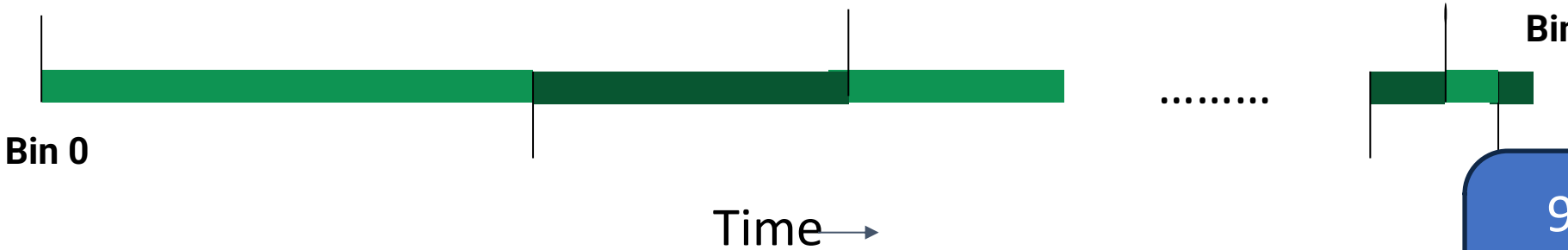
Forecasting

- Standard test/train splits
  - More relevant to interpolation
  - than extrapolation
- Time is asymmetric
  - *It is difficult to make predictions*
  - *especially about the future*
    - -- Yogi Berra

Assign 200M papers to 100 bins (by time)

- Sort 200M papers by publication date
- Output: Split papers into 100 bins, with 2M papers/bin
- Because of exponential growth
  - (literature doubles every 9 years)
  - older bins span more time, and
  - newer bins span less time





90% of literature published after I went to grad school

Experimental Results

- Performance improves
  - when training set is larger and spans more time
- Performance degrades
  - with predictions further and further into the future
- Because of Metcalfe's Law
  - Text/Context trade-off
    - Depends on scale
  - Larger graphs favor context
    - ProNE: Spectral clustering on citations
    - Specter: BERT-like encoding of text
- Better-Together
  - Ensembles >> Either by itself

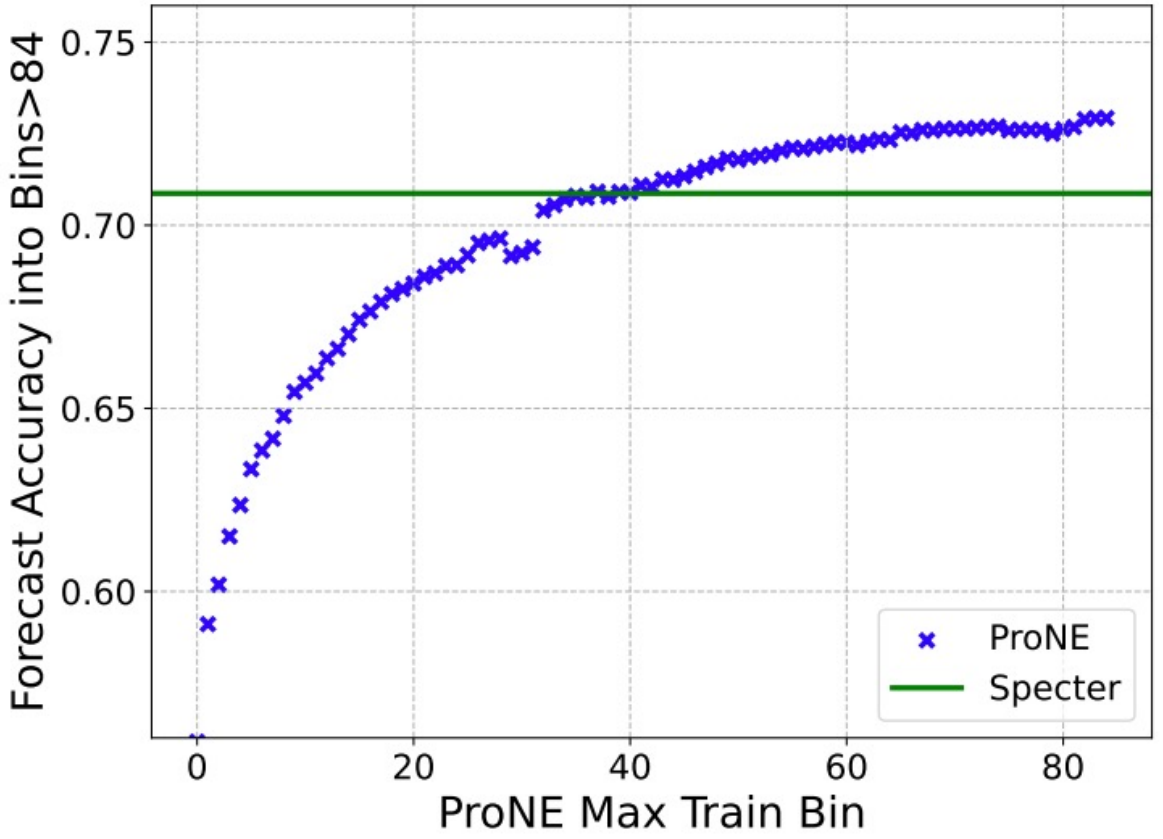


Figure 9: ProNE-Specter Crossover: Metcalfe's Law favors larger citation graphs (more than 82M papers).

Conclusion: Deliverables

- GitHub:
  - massive (but needs editing)
  - slides, code, reading list
  - pointers to large files on Globus
- Web Server
  - Recommendations
    - Input: queries (docs)
    - Output: recommendations (docs)
      - What Should I read? Cite?
- Resources
  - Data behind web server:
    - citation graph
    - embeddings
    - indexes for approx. nearest neighbors
    - pairs of corpusIds with large cosines
    - all of the above for
      - Specter1, Specter2, SciNCL, ProNE
      - x100 bins for ProNE
- Evaluations & Benchmarks
  - Intrinsic: Perplexity, Random Walks
  - Extrinsic: Local Citation Prediction