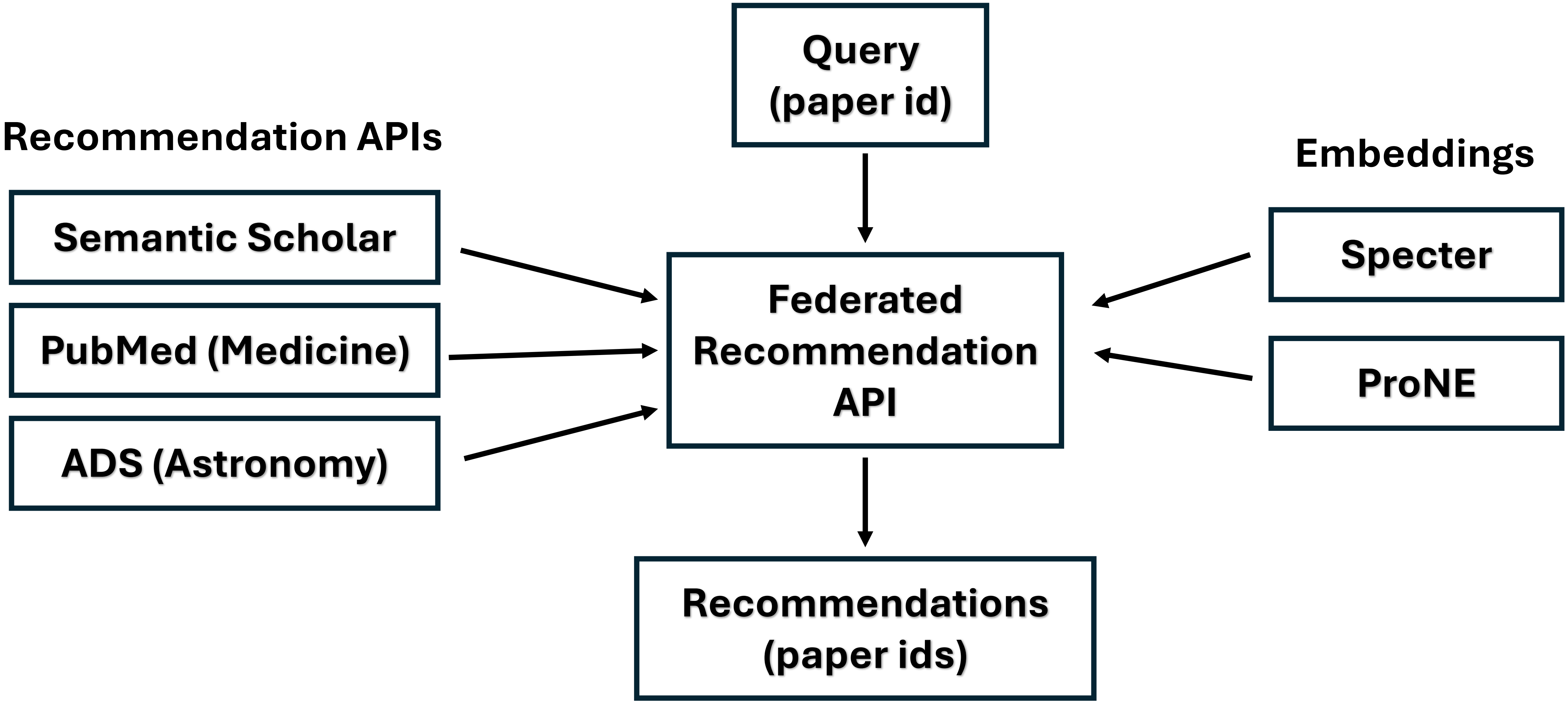
Better Together API

Ken Church

Town Hall

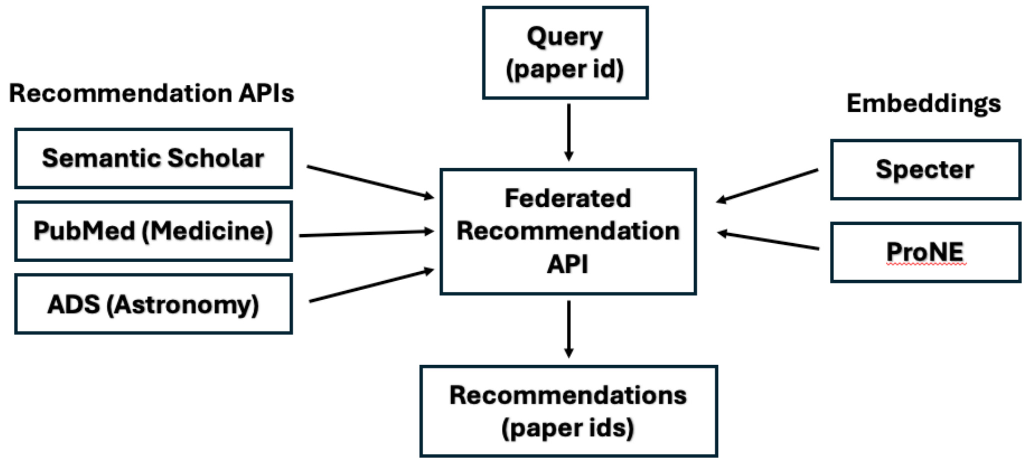
Aug 20, 2024

Federated API

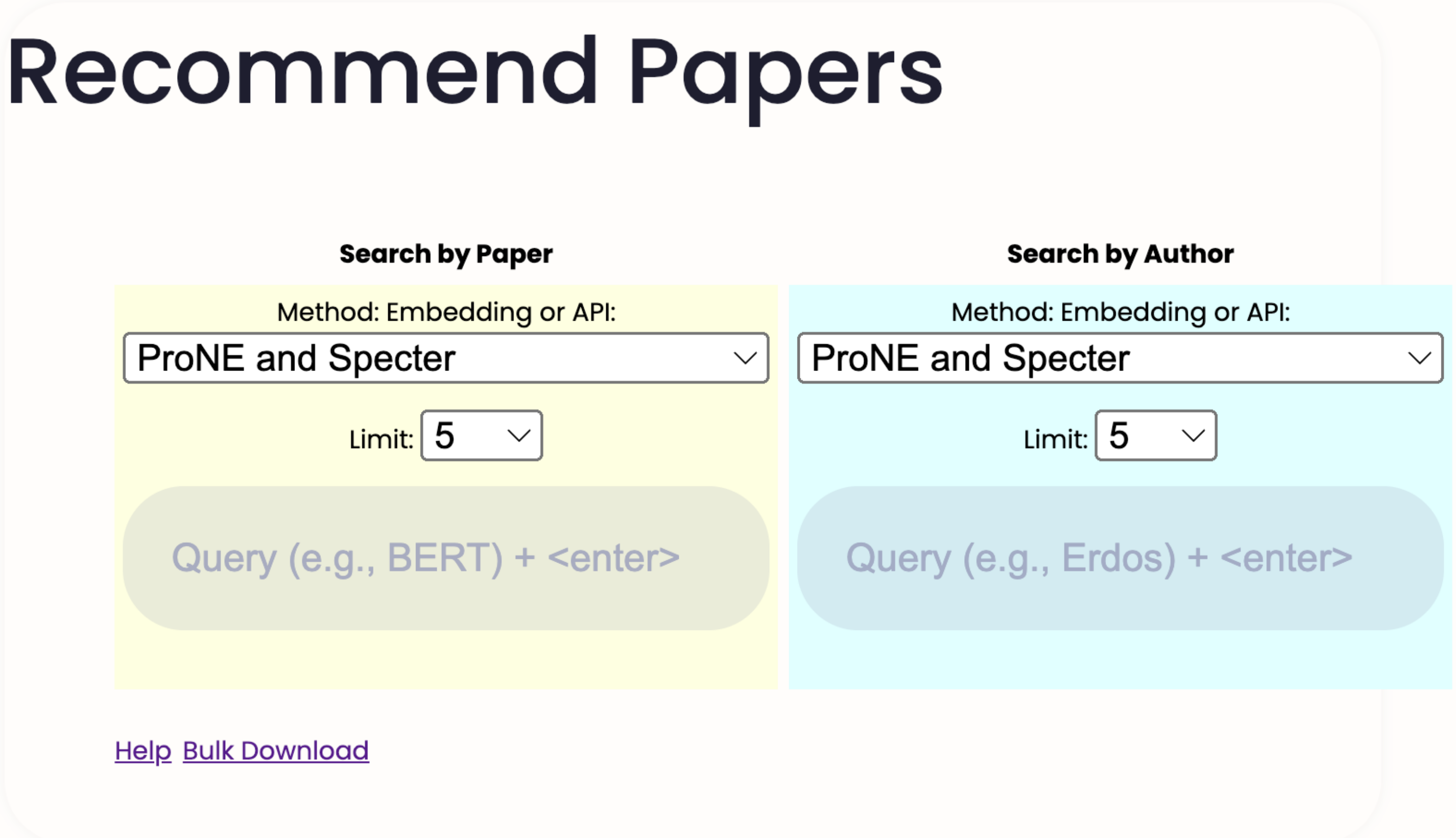


Motivation for a Federated API

- Easy to create web sites
  - https://recommendpapers.xyz
- To compare and contrast
  - Different APIs/Embeddings
- Different recommendations are different
  - Almost no overlap



https://recommendpapers.xyz



id: df2b0e26d0599ce3e70

recommend_method: ProNE

limit: 10

Submit

query: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
authors: Jacob Devlin, Ming-Wei Chang, ..., Kristina Toutanova
tldr: A new language representation model, BERT, designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers, which can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks.
citations: 79686

Score	Citations	Year	Title	Authors	Like this	Specter	ProNE	SciNCL	GNN
0.983	1258	2020	Pre-trained models for natural language processing...	Xipeng Qiu, Tianxiang Sun, ..., Xuanjing Huang	like this	0.938	0.983	0.0	0.963
0.981	1123	2019	What Does BERT Learn about the Structure of Langua...	Ganesh Jawahar, Benoît Sagot, Djamé Seddah	like this	0.946	0.981	0.891	0.947
0.977	557	2020	Revisiting Pre-Trained Models for Chinese Natural ...	Yiming Cui, Wanxiang Che, ..., Guoping Hu	like this	0.943	0.977	0.895	0.891
0.976	398	2019	To Tune or Not to Tune? Adapting Pretrained Repres...	Matthew E. Peters, Sebastian Ruder, Noah A. Smith	like this	0.914	0.976	0.79	0.962
0.971	967	2019	A Structural Probe for Finding Syntax in Word Repr...	John Hewitt, Christopher D. Manning	like this	0.918	0.971	0.86	0.973
0.971	581	2020	LUKE: Deep Contextualized Entity Representations w...	Ikuya Yamada, Akari Asai, ..., Yuji Matsumoto	like this	0.933	0.971	0.819	0.962
0.97	296	2020	PhoBERT: Pre-trained language models for Vietnames...	Dat Quoc Nguyen, A. Nguyen	like this	0.9	0.97	0.782	0.92
0.97	674	2019	Linguistic Knowledge and Transferability of Contex...	Nelson F. Liu, Matt Gardner, ..., Noah A. Smith	like this	0.946	0.97	0.868	0.96
0.967	783	2019	What do you learn from context? Probing for senten...	Ian Tenney, Patrick Xia, ..., Ellie Pavlick	like this	0.933	0.967	0.864	0.969

Home
Help
Bulk Download

API Documentation

Documentation for Recommend Papers

Table of Contents

APIs

These APIs provide a federated view over three APIs: Semantic Scholar (S2), PubMed and Astrophysics Data System (ADS). In addition, the APIs support recommendations based on cached embeddings from Specter and ProNE. Specter is a BERT-like model, fine-tuned on a few million citations. ProNE uses spectral clustering to embed 2 billion citations. Recommendations from ProNE tend to have more citations, whereas recommendations from Specter tend to be more recent (papers with more citations tend to be older because it takes time for papers to accumulate citations). The API provides access to 4 embeddings: Specter, ProNE, SciNCL and GNN; see discussion of score1, score2 and embeddings.

All of the APIs use GET HTTP requests, and return json objects. Click on the examples below to see the input GET request and the output json objects.

API	Examples	Arguments	Description
Paper Search	• example • more challenging example with get_pdfs, get_bibtex and sort_by	help, query, fields, sort_by, limit, get_pdfs, get_bibtex	• Find papers matching input query (a string); output fields from Semantic Scholar for each paper. • See documentation on fields for more information on fields in Semantic Scholar. • 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). • sort_by can be any of the fields that can be converted to integers
Author Search	• simple example • more challenging example	help, query, fields, sort_by, limit	• Find authors matching input query (a string); output fields from Semantic Scholar for each author. • See documentation on fields for more information on fields in Semantic Scholar. • sort_by can be any of the fields that can be converted to integers • Limit argument will truncate results (after sorting) • Note: author fields are different from paper fields.
	• simple example • more challenging example (with score2)		• Input one or more comma separated paper id and output fields from Semantic Scholar, as well as embeddings.

10 Entry Points

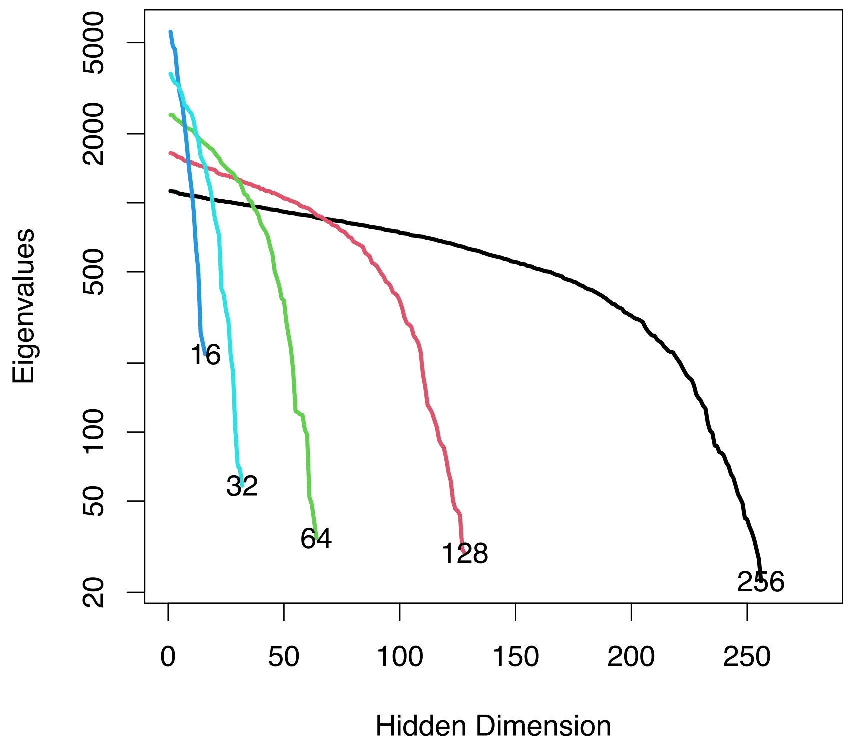
- Paper Search:
  - Input query (strings); output papers ids
- Author Search:
  - Input query (strings); output author ids
- Lookup Paper:
  - Input paper ids; output fields (titles, abstracts, embeddings, etc.)
- Lookup Author:
  - Input author ids; output fields (titles, abstracts, embeddings, etc.)
- Lookup Citations:
  - Lookup Citations for paper id and output fields from Semantic Scholar for each citation
- Coauthors:
  - Input a query (a string); output a list of coauthors (filtered by after_year) for each matching author id.
- Recommend Papers:
  - Input a paper id and a recommendation method;
  - output paper ids (with fields, scores, etc.)
- Recommend Authors:
  - Input an author id and a recommendation method;
  - output paper ids (with fields, scores, etc.)
- Compare and contrast papers with chatbot/RAG
- Compare and Contrast texts with chatbot/RAG

Use Case: Assign Papers to Reviewers

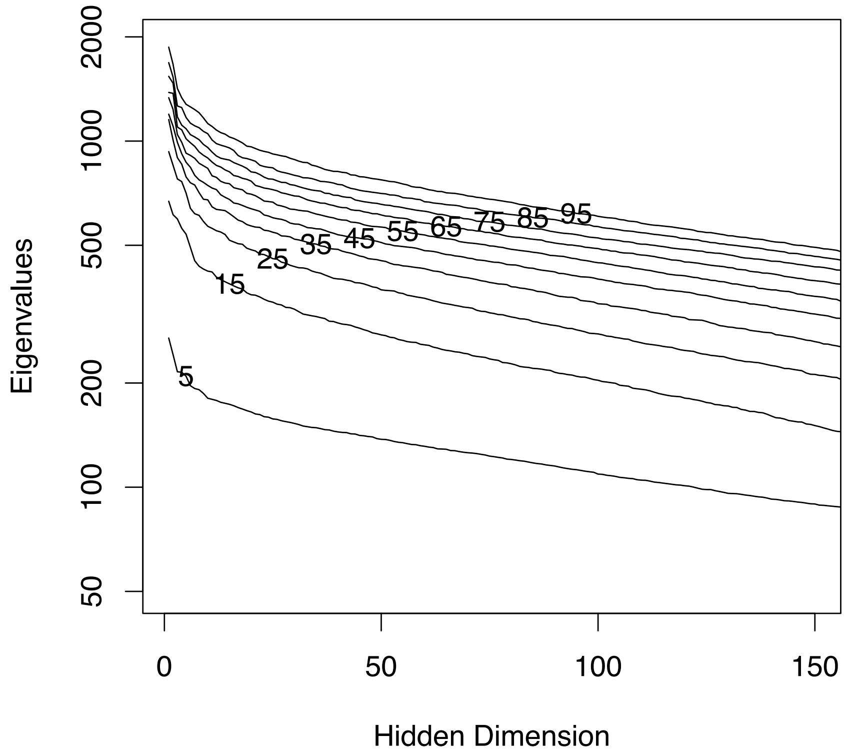
- Had a very promising call with a NASA PM Megan Ansdell
- It takes her about a week to put together a panel
- She needs to find reviewers that
  - are qualified
  - diverse over experience, etc.
  - avoid conflicts of interest
- She can program in Python (and use these APIs)
  - She wrote a tool to check proposals for anonymity (pronouns)
  - Many of her peers are more senior than she is
    - So she runs her tool on their proposals
  - We talked about a sole source grant, so we could work together on this use case
- More generally, I would like to scale this up to large conferences

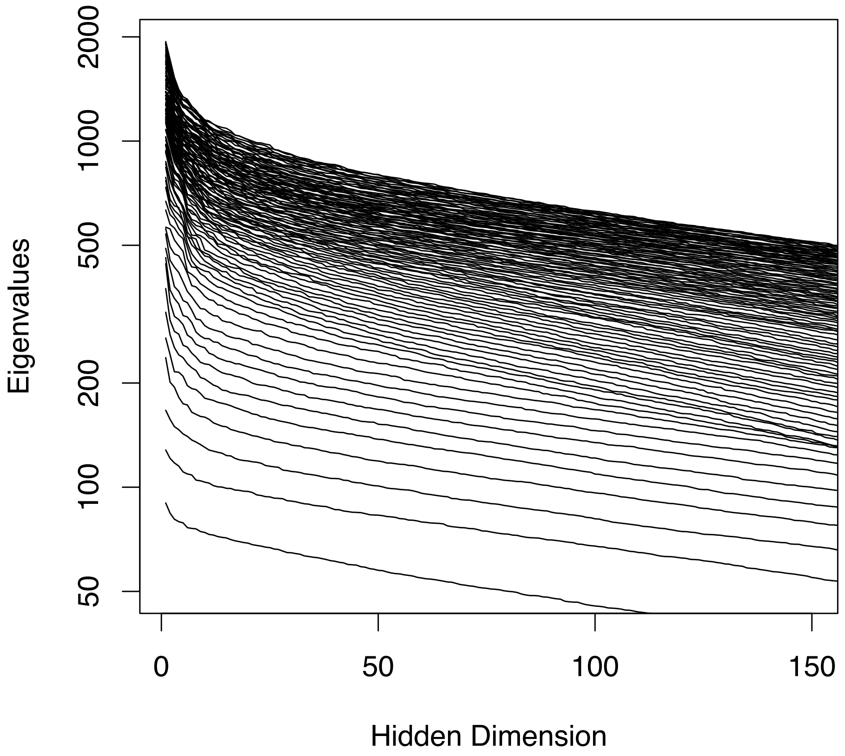
Eigenvalues by K (# Hidden Dimensions)

- Start with a large graph of 250M papers
- Compute ProNE embeddings
  - with K = 16, 32, 64, …, 256
- For each embedding, E
  - $U, D, Vt = svd(E)$
- Plot D
- Why does D have this structure?



Eigenvalues by Bin





https://en.wikipedia.org/wiki/Eigenvalues_and_eigenvectors

Additional properties of eigenvalues [edit]

Let A be an arbitrary $n \times n$ matrix of complex numbers with eigenvalues $\lambda_1, \ldots, \lambda_n$. Each eigenvalue appears $\mu_A(\lambda_i)$ times in this list, where $\mu_A(\lambda_i)$ is the eigenvalue's algebraic multiplicity. The following are properties of this matrix and its eigenvalues:

- The trace of A, defined as the sum of its diagonal elements, is also the sum of all eigenvalues,[29][30][31]

$$\operatorname{tr}(A) = \sum_{i=1}^{n} a_{ii} = \sum_{i=1}^{n} \lambda_i = \lambda_1 + \lambda_2 + \cdots + \lambda_n.$$

- The determinant of A is the product of all its eigenvalues,[29][32][33]

$$\det(A) = \prod_{i=1}^{n} \lambda_i = \lambda_1 \lambda_2 \cdots \lambda_n.$$

- The eigenvalues of the kth power of A; i.e., the eigenvalues of A^k, for any positive integer k, are $\lambda_1^k, \ldots, \lambda_n^k$.
- The matrix A is invertible if and only if every eigenvalue is nonzero.
- If A is invertible, then the eigenvalues of A^{-1} are $\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_n}$ and each eigenvalue's geometric multiplicity coincides. Moreover, since the characteristic polynomial of the inverse is the reciprocal polynomial of the original, the eigenvalues share the same algebraic multiplicity.
- If A is equal to its conjugate transpose A^*, or equivalently if A is Hermitian, then every eigenvalue is real. The same is true of any symmetric real matrix.
- If A is not only Hermitian but also positive-definite, positive-semidefinite, negative-definite, or negative-semidefinite, then every eigenvalue is positive, non-negative, negative, or non-positive, respectively.
- If A is unitary, every eigenvalue has absolute value $|\lambda_i| = 1$.
- If A is a $n \times n$ matrix and $\{\lambda_1, \ldots, \lambda_k\}$ are its eigenvalues, then the eigenvalues of matrix $I + A$ (where I is the identity matrix) are $\{\lambda_1 + 1, \ldots, \lambda_k + 1\}$. Moreover, if $\alpha \in \mathbb{C}$, the eigenvalues of $\alpha I + A$ are $\{\lambda_1 + \alpha, \ldots, \lambda_k + \alpha\}$. More generally, for a polynomial P the eigenvalues of matrix $P(A)$ are $\{P(\lambda_1), \ldots, P(\lambda_k)\}$.

https://en.wikipedia.org/wiki/Determinant

Upper and lower bounds [edit]

For a positive definite matrix A, the trace operator gives the following tight lower and upper bounds on the log determinant

$$\operatorname{tr}\left(I - A^{-1}\right) \leq \log\det(A) \leq \operatorname{tr}(A - I)$$

with equality if and only if $A = I$. This relationship can be derived via the formula for the Kullback-Leibler divergence between two multivariate normal distributions.

Also,

$$\frac{n}{\operatorname{tr}(A^{-1})} \leq \det(A)^{\frac{1}{n}} \leq \frac{1}{n}\operatorname{tr}(A) \leq \sqrt{\frac{1}{n}\operatorname{tr}\left(A^2\right)}.$$

These inequalities can be proved by expressing the traces and the determinant in terms of the eigenvalues. As such, they represent the well-known fact that the harmonic mean is less than the geometric mean, which is less than the arithmetic mean, which is, in turn, less than the root mean square.