Improved very sparse matrix completion using an “asymmetric SVD”

Raj Rao, Charles Bordenave and Simon Coste

2005.06062.pdf (arxiv.org)

My Goals for the talk

Tell you about …

New “asymmetric eig” method that **works when SVD fails**

(very sparse regime)

Have you actively wonder (before I tell you)

What theory ⇔ computational simulation gave (crazy) idea?

Setup: (Very Sparse) Matrix Completion

- Latent (or true) **low-rank *r*** rectangular *m x n* matrix

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^H$$

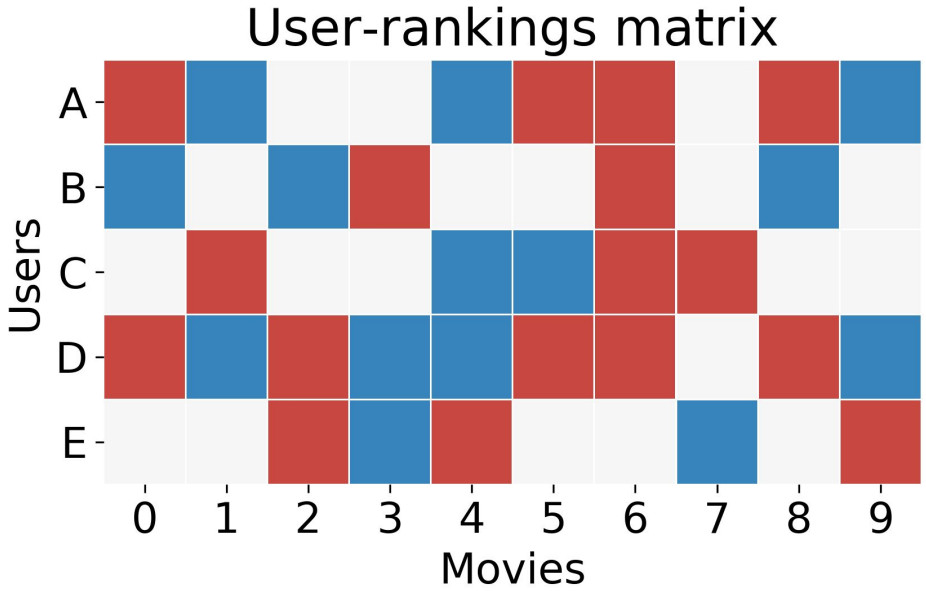
- Uniformly random missing entries w/ probability **p = d/n**

$$\tilde{A}[i,j] = A[i,j] \text{ with probability } p,$$

$$\tilde{A}[i,j] = ? \text{ with probability} 1-p.$$

- **Goal**: Reconstruct A "as accurately as possible" (knowing *r*)

Application: Recommender systems & Netflix



Extreme sparsity ubiquitous

Same setup – different domain :)

(54) **METHOD AND SYSTEM FOR IDENTIFYING PEOPLE WHO ARE LIKELY TO HAVE A SUCCESSFUL RELATIONSHIP**

(75) Inventors: **J. Galen Buckwalter**, Pasadena, CA (US); **Steven R. Carter**, Los Angeles, CA (US); **Gregory T. Forgatch**, San Marino, CA (US); **Thomas D. Parsons**, Pasadena, CA (US); **Neil Clark Warren**, Pasadena, CA (US)

(73) Assignee: **eHarmony, Inc.**, Santa Monica, CA (US)

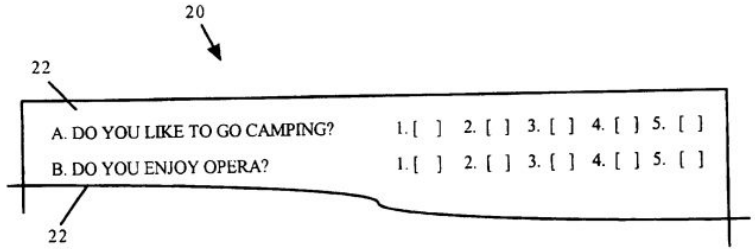


FIGURE 2

24

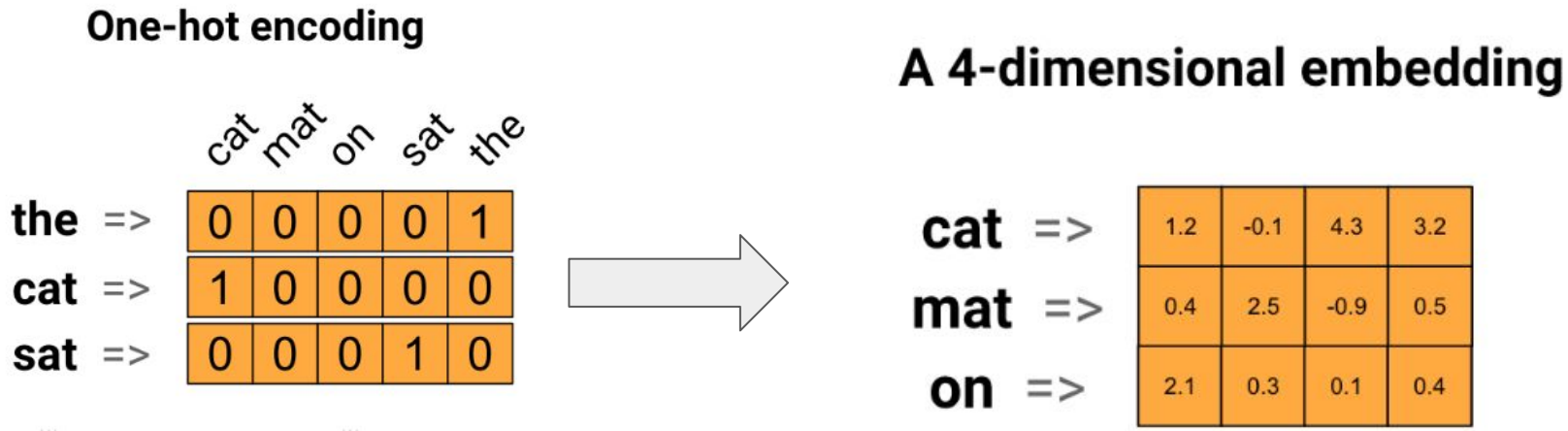
26 28 28 28

PERSON	A	B	C	D	E	F
JOHN DOE	1	1	2	5	1	3
JANE DOE	5	2	2	1	3	4
JOE SMITH	4	3	2	5	1	4
BOB BAKER	3	4	3	2	1	5
DR. BOB	2	4				

FIGURE 3

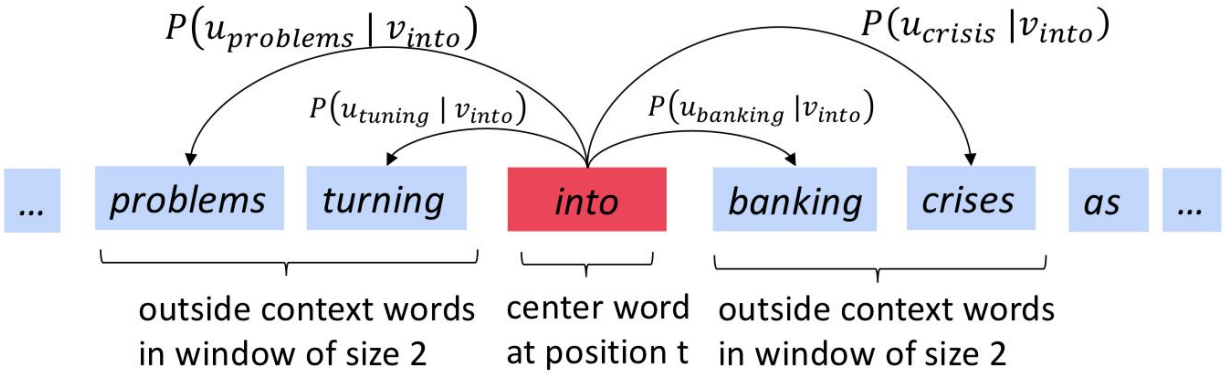
Extreme sparsity in ChatGPT like models

ChatGPT learns word embeddings from massive amounts of data



Extreme Sparsity in learning contexts

e.g Predict word from context of **4**



=> **Context Embedding Matrix** is very sparse => Opportunity for understanding

rows = # words

Sparsity = (random) length of context

The New York Times Magazine

If You Liked This, You're Sure to Love That

The first major breakthrough came less than a month into the competition. A team named Simon Funk vaulted from nowhere into the No. 4 position, improving upon Cinematch by 3.88 percent in one fell swoop. Its secret was a mathematical technique called singular value decomposition. It isn't new; mathematicians have used it for years to make sense of prodigious chunks of information. But Netflix never thought to try it on movies.

Singular value decomposition works by uncovering "factors" that Netflix customers like or don't like. Say, for example, that

Example: Single entry missing in rank one matrix

$$\tilde{A}_1 = \begin{bmatrix} 1 & 2 \\ ? & 4 \end{bmatrix}$$

Claim: Missing entry = 2 (unique). Any single entry missing is fine

Example: Single entry missing in rank one matrix

$$\tilde{A}_1 = \begin{bmatrix} 1 & 2 \\ ? & 4 \end{bmatrix}$$

Claim: Missing entry = 2 (unique). Any single entry missing is fine

Counter Claim: If A = [1 0 ; 0 0] then one entry missing could mean failure

Delocalization Assumption: Singular vectors are incoherent (or spread out)

When/why do we expect matrix completion to work?

An *m x n* matrix with incoherent s. vectors & **rank d** ⇔ O((m + n + 1) r)

- **# free parameters** = O(r (m+n+1)) = O(r m) << mn entries in matrix

Observe each element with probability $p = d/n$

- **# elements observed** = O(d m)

Regimes of difficulty:

- "Easy" ⇔ $d = O(n)$ ⇔ $p = O(1)$
- "Less easy" ⇔ $d = O(r \log n)$ ⇔ $p = O(r \log n / n)$
- **"Fun" & Challenging ⇔ $d = O(r)$ ⇔** $p = O(r / n)$

Some numerical experiments

Rank = 4

Assume: 1 entry missing

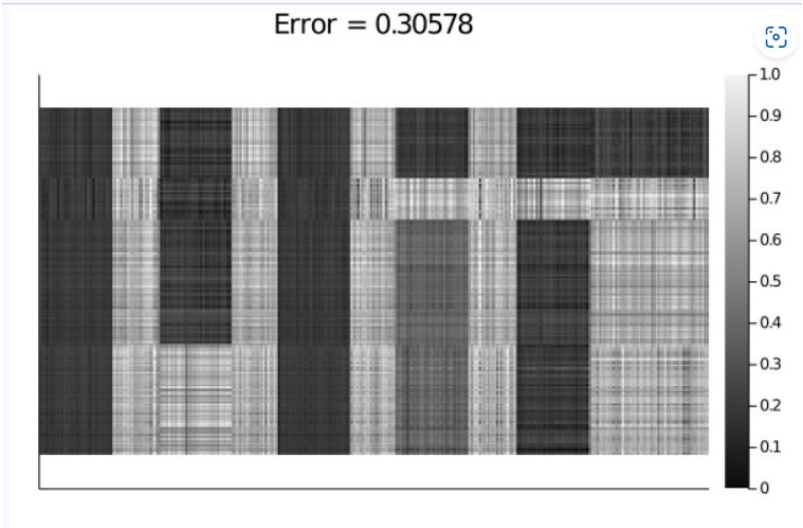
Algorithm: Fill in missing entry with zero, compute closest rank 1 matrix via truncated SVD

Claim: Frobenius Norm of Error <= Norm of removed entry

Now assume 70% entries observed

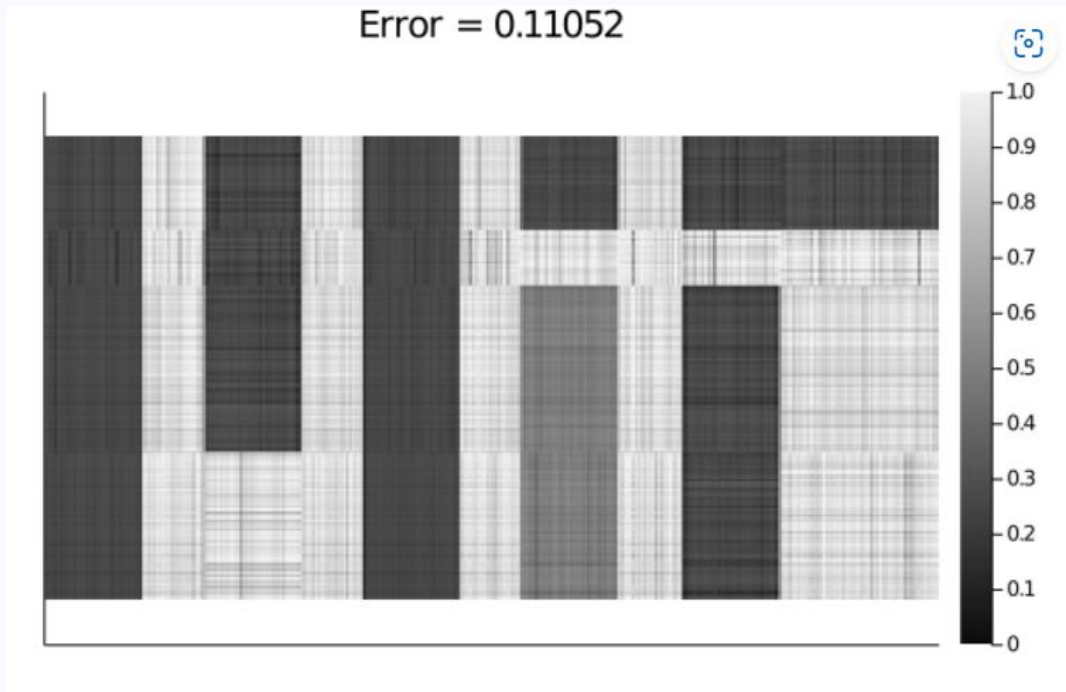
- Fill in missing entry with zero and do rank 4 truncated SVD
- **Key Question:** What does one expect? How does one reason about it?

Fill in missing entries with zeros and do rank 4 tSVD



“Not perfect” - but clear information is retained

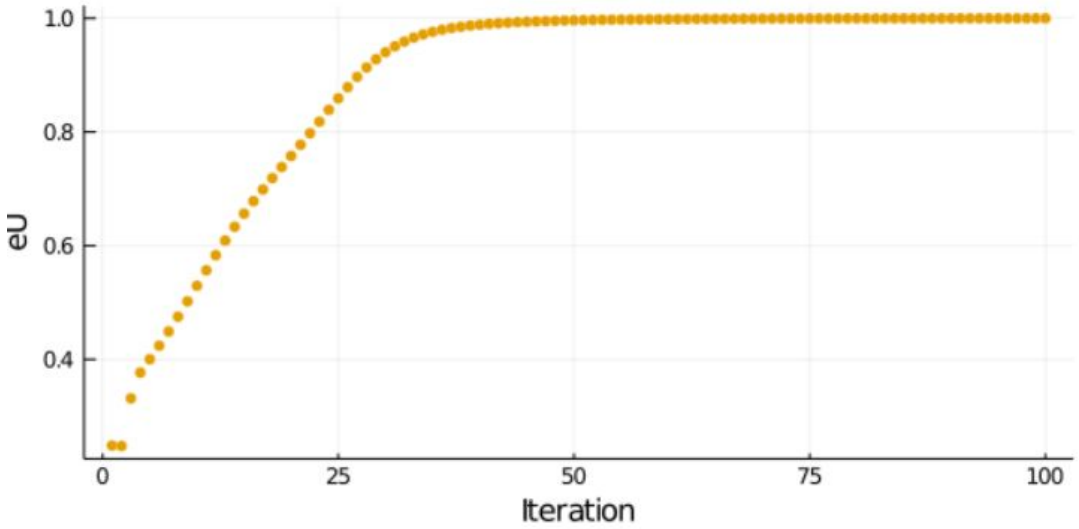
Fill in known entries and repeat rank 4 tSVD



After 10 iterations …

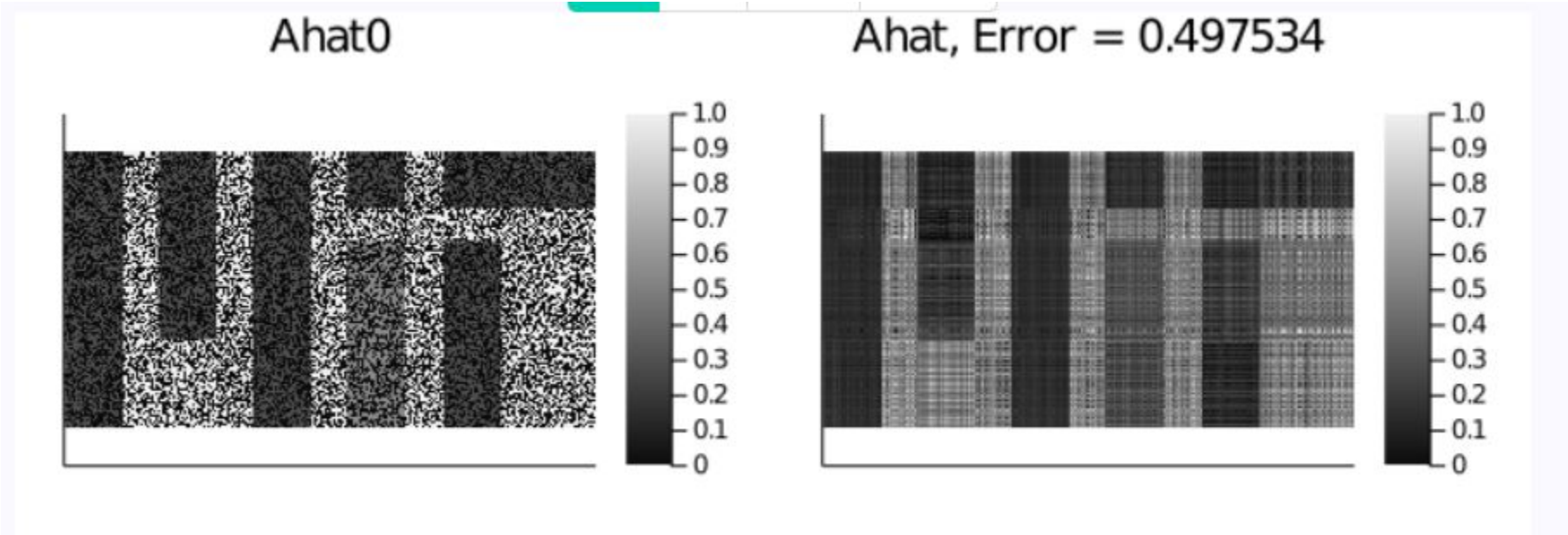
Error tends to 0

Inner product with true “U” w/ iterations

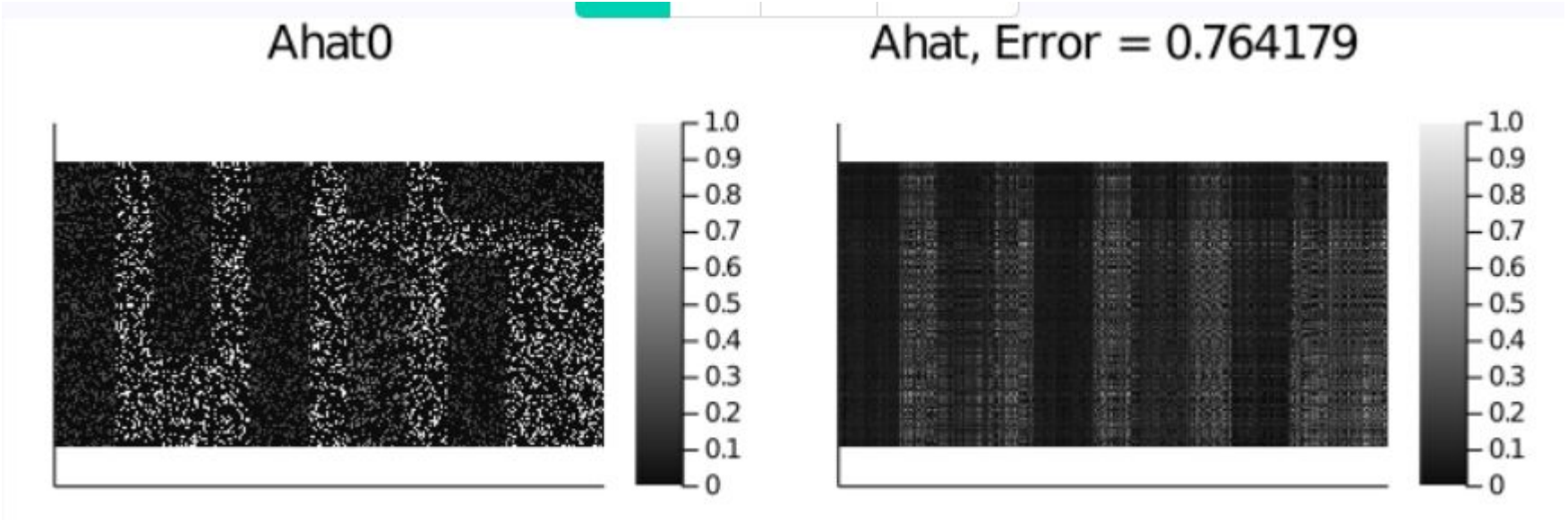


Key insight: **“some” noisy information @** iteration 1 ⇔ **hope**!

What is we make it sparser? p = 0.5

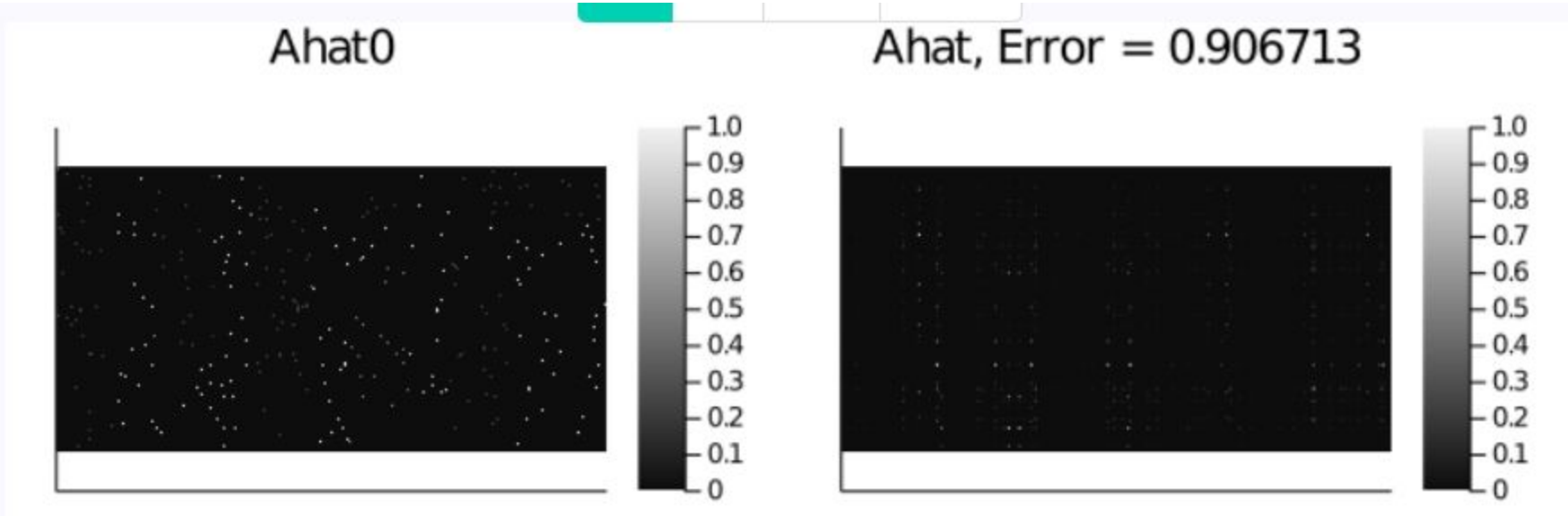


What if even sparser?



p = 0.2

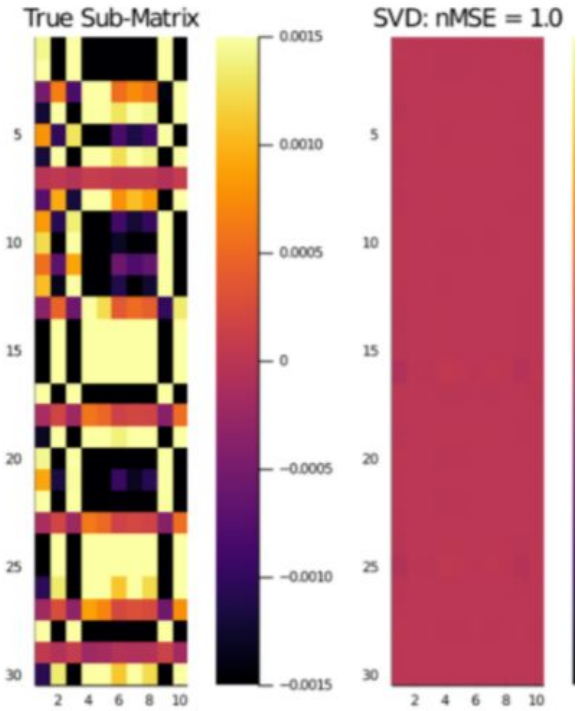
Even sparser



p = 0.01

Expectation: “no” noisy information @ iteration 1 ⇔ **failure** ?

What happens to the SVD in (very) sparse regime



-

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^H$$

- Entries observed with probability p = d/n
- When d = O(log n), perfect reconstruction
- **When d = O(1),** can't expect perfect
- **SVD breaks down completely**
  - Singular vectors are localized

-

RMT Reasoning about difficulty of different regimes

Each matrix element observed with probability p :

Observed A = E[A] + **(A - E[A])** = p A + **Δ**

$$\text{Error in singular vectors of low-rank } A + \Delta \leq C(A) \cdot \sigma_1(\Delta),$$

Expectation:

Well known!

- **Perfect reconstruction** if *vanishing perturbation* ⇔ d = O(n log n)
- **Imperfect/Noisy reconstruction** if *bounded perturbation* ⇔ d = O(1) ? ⇐
- **Junk reconstruction** if *unbounded perturbation* ⇔ d = O(1) ? ⇐ This talk

(Classical) Matrix Completion Strategies

$$A = \left(\frac{n}{d}\right) P \odot M,$$

- **Nuclear norm regularization**
  - Find matrix with smallest nuclear norm that matches revealed entries
  - Iterative algorithm involving truncated SVD in first step
- **Alternating minimization**
  - Replace missing entries of *A* with zeros, do SVD
  - Truncate to rank-r
  - Replace (known) revealed entries, repeat

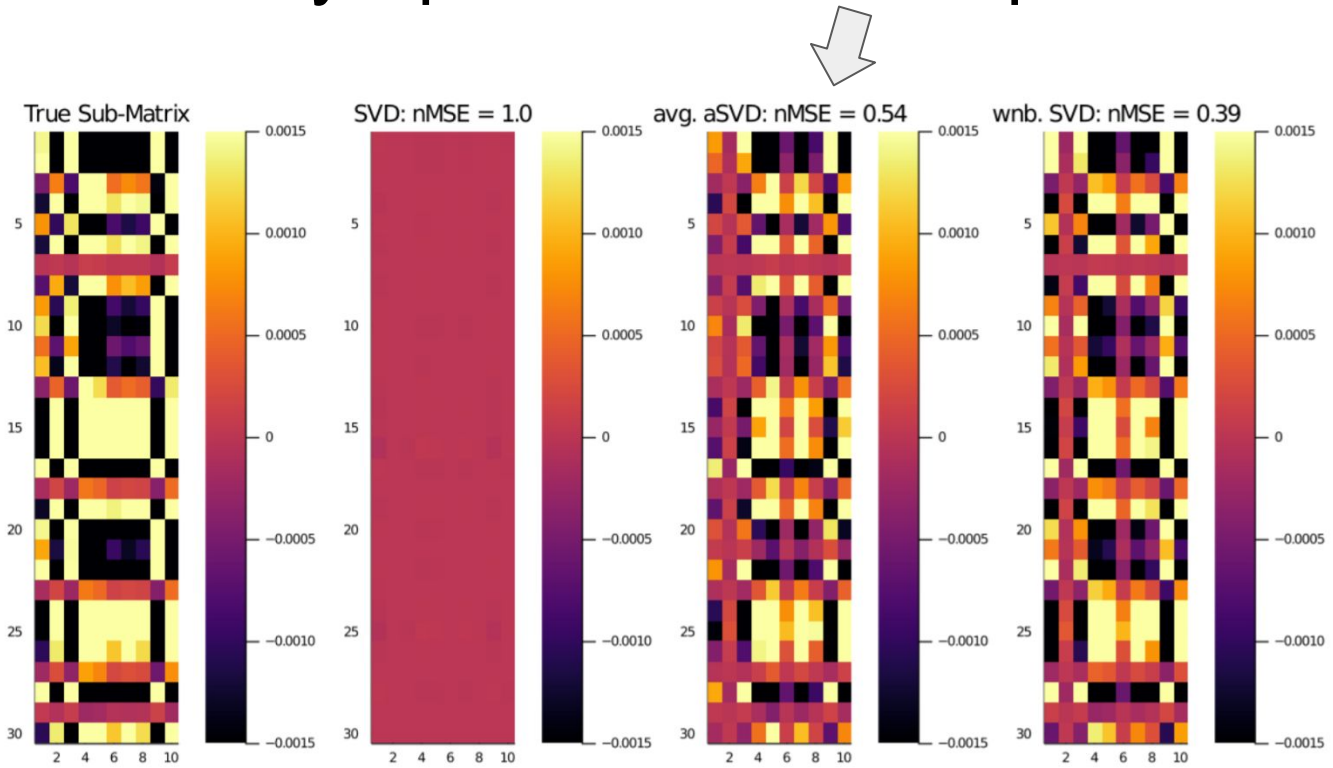
Questions motivating this talk

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^H$$

Questions: When A entries observed with p = *d/n with* ***d = O(1):***

- Possible to reconstruct reliably with error (i.e. not perfectly)?
- Polynomial time algorithm(s)?
- Fundamental limit(s) of reliable reconstruction?

Improved Very Sparse Matrix Completion



(A) Normally distributed singular vectors.

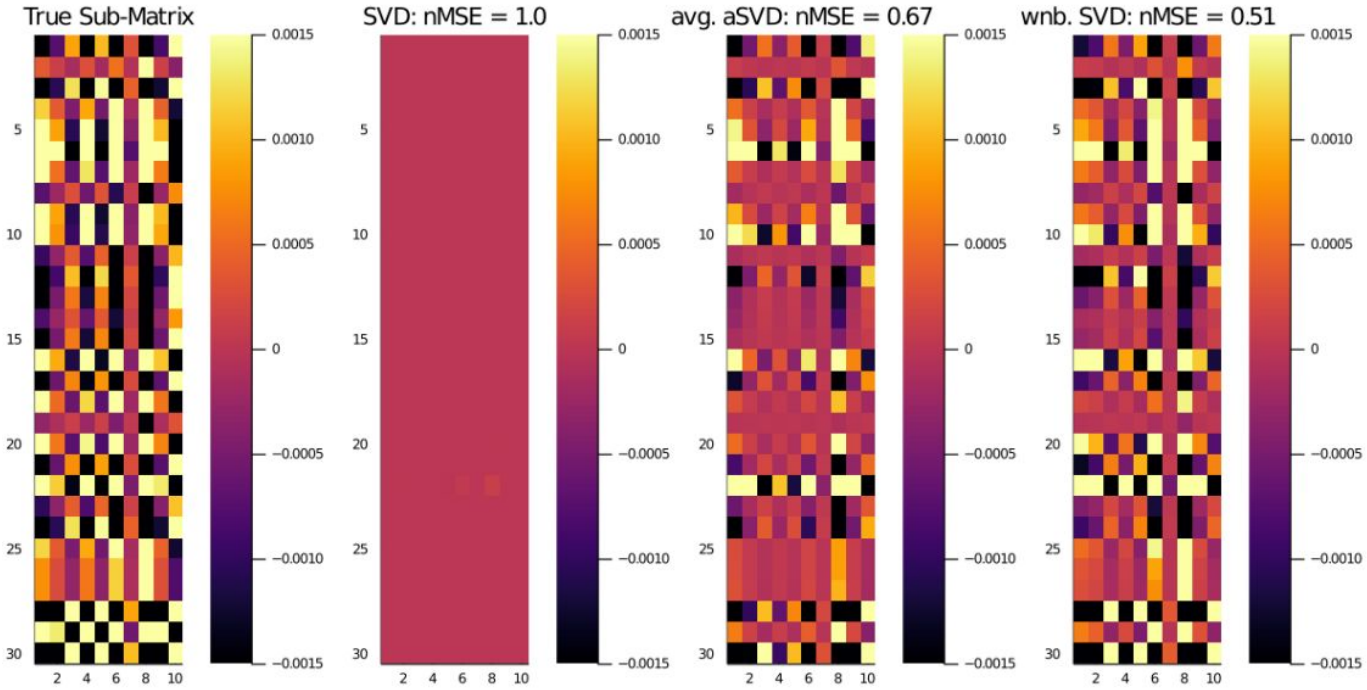
Questions answered in this talk

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^H$$

Questions answered: When *d* = *O(1):*

- Possible to reconstruct reliably with error (i.e. not perfectly)? **YES!**
  - Keshavan, Montanari and Oh (2009) proved reliable matrix completion possible in this regime by pruning rows/columns with large degree
  - **This work - no thresholding or pruning**
- Polynomial time algorithm(s)? **Two new algorithms**
- Fundamental limit(s) of reliable reconstruction? **Two new limits**

Improved Very Sparse Matrix Completion



(B) Hyperbolic secant distributed singular vectors.

Core idea behind the algorithm

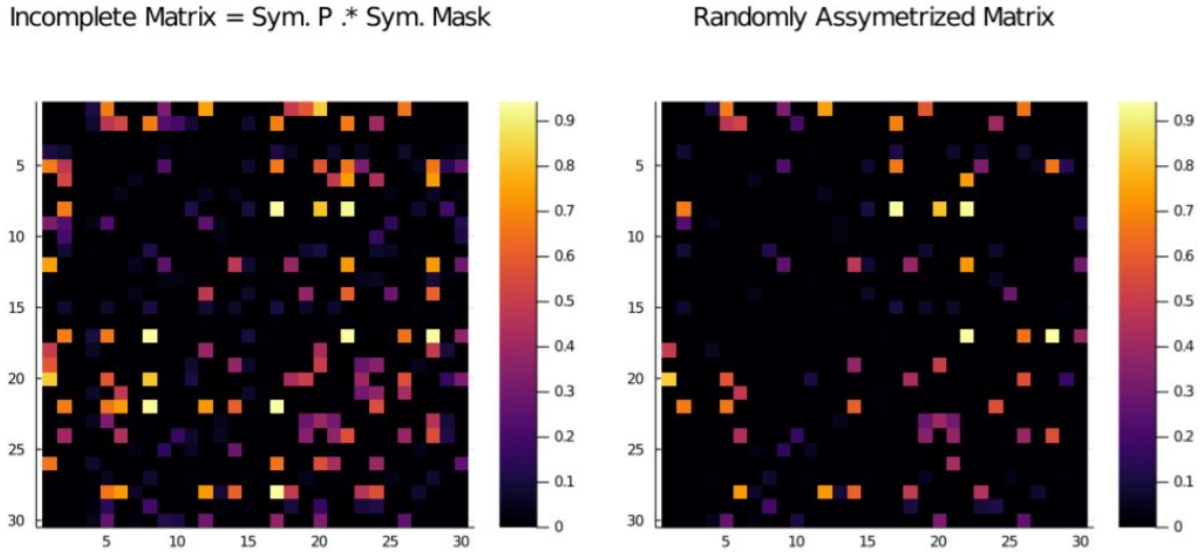
(Related) Setup: Low rank square symmetric matrix with symmetric masking

$$P = \sum_{k=1}^{n} \mu_k \varphi_k \varphi_k^*,$$

$$A = \left(\frac{n}{d}\right) P \odot M,$$

- **(Symmetric) Eig** breaks down in the very sparse regime
- **Our idea**: Intentionally asymmetrize A
  - With probability ½ take elements either above or below the diagonal

Power of random asymmetrization + (asym) eig.



Theorem (BCN'20): Above explicit threshold:

1) Eigenvectors of rand. asym. matrix **are** reliably estimated
2) **Best estimator** averages left and right eigenvectors

Theory: Detection threshold in very sparse regime

Detection threshold ϑ: *any number* ϑ *such that*

$$\vartheta \geqslant \max\{\vartheta_1, \vartheta_2\},$$

where the 'theta parameters' are defined by

$$\vartheta_2 = \sqrt{\frac{\rho}{d}} \quad \text{and} \quad \vartheta_1 = \frac{L}{d}.$$

Variance matrix Q:

$$Q_{x,y} = n|P_{x,y}|^2 \qquad \rho = \|Q\|.$$

Amplitude parameter L:

$$L = n \max_{x,y} |P_{x,y}|.$$

Equivalently, it is the scaled L^1 *to* L^∞ *norm of* P.

Theory: Overlap between eigenvectors

Inner product between true and estimated *i-th* evector

$$\gamma_i = \sum_{s=0}^{\ell} \left(\frac{\vartheta_2}{\mu_i} \right)^{2s} = \frac{1 - (\vartheta_2/\mu_i)^{2(\ell+1)}}{1 - (\vartheta_2/\mu_i)^2}$$

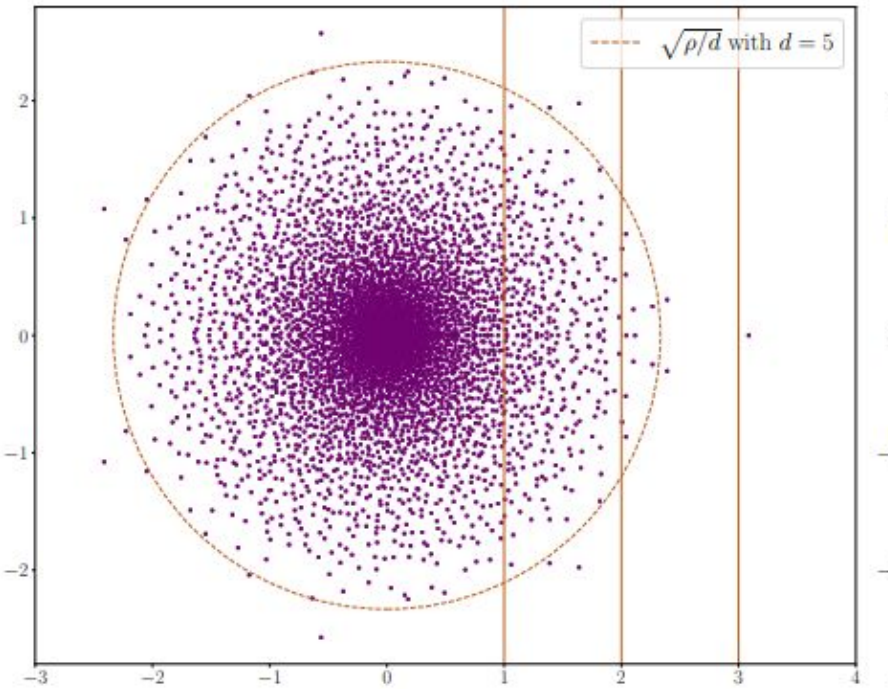
$$\ell = \lfloor (1/8) \log_D(n) \rfloor$$

Theory: Det. Threshold & Overlap in rank 1 setting

$$\vartheta_2 = \sqrt{\frac{n|\varphi|_4^4}{d}} \quad \text{and} \quad \gamma = \frac{1-\vartheta_2^{2(\ell+1)}}{1-\vartheta_2^2}$$

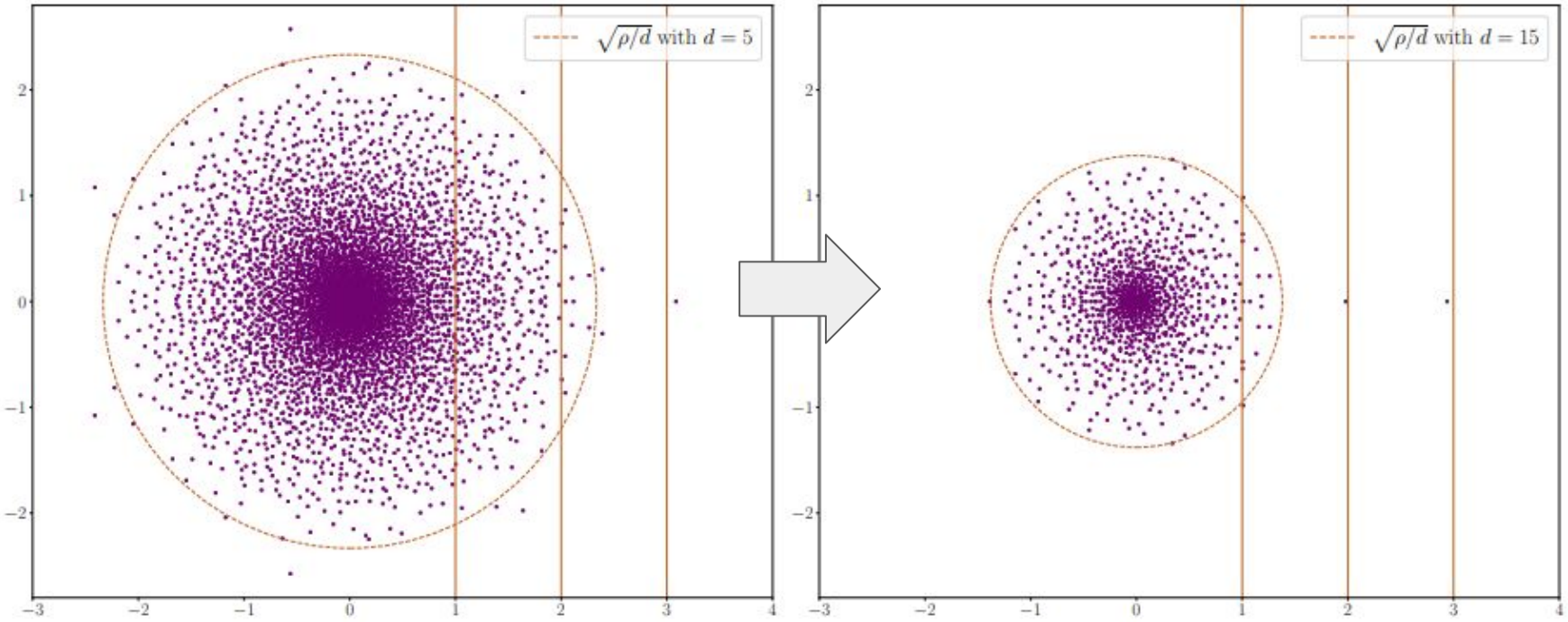
Non-universal dependence on kurtosis!

Emergence of **real eigs** above det. threshold



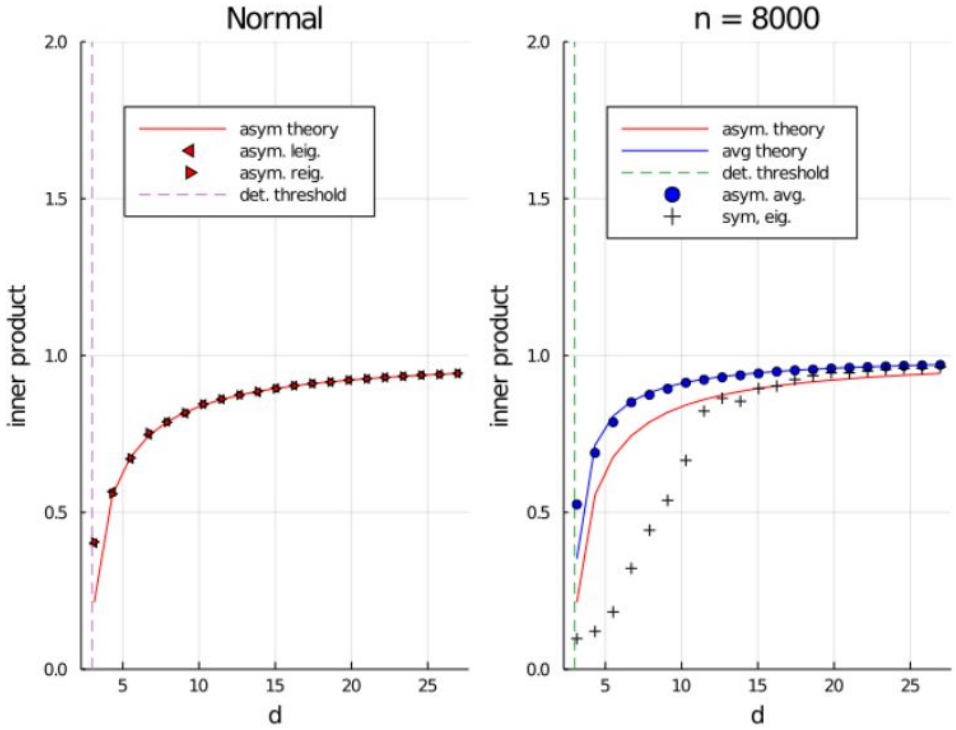
Rank 3 latent matrix – 1 eigenvalue above phase transition

Emergence of **real eigs** above det. threshold



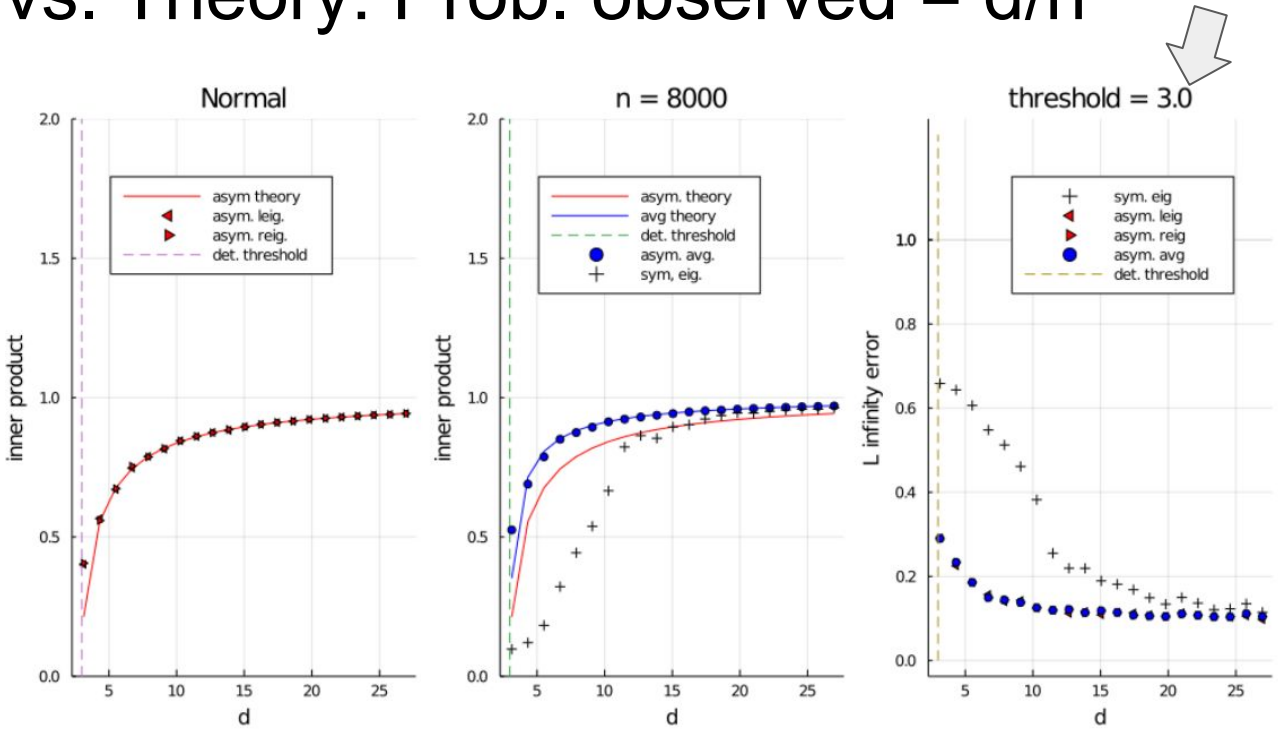
Rank 3 latent matrix – **now 2 eigenvalues** above phase transition

Sims vs. Theory: Prob. observed = d/n



(B) Normal

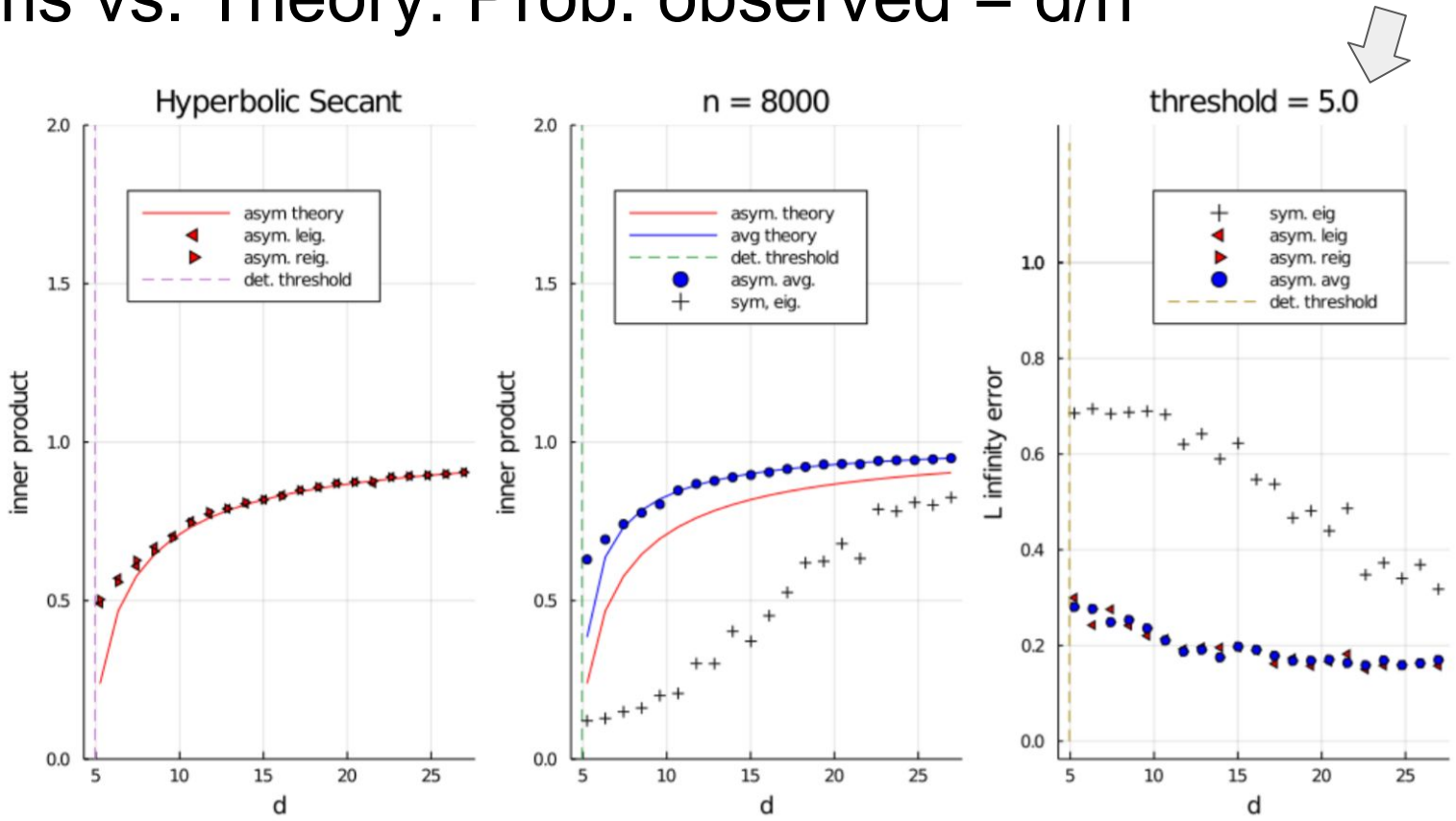
Sims vs. Theory: Prob. observed = d/n



(B) Normal

Averaging left + right evecs improves accuracy, inner product between left and right provides estimate of evec overlap

Sims vs. Theory: Prob. observed = d/n



- **Non universal distribution dependent threshold**

Very Sparse Matrix Completion w/ asym. eig

- Latent (or true) **low-rank *r*** rectangular *m x n* matrix

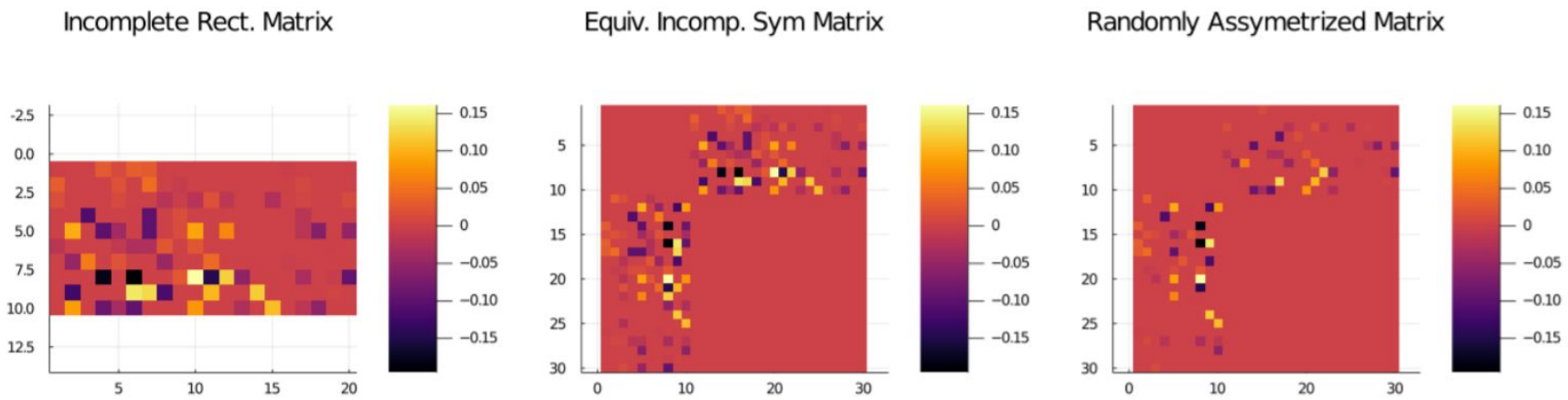
$$P = \sum_{k=1}^{r} \sigma_k \zeta_k \xi_k^*,$$

$$A = \left(\frac{n}{d}\right) P \odot M,$$

- **Fact:**
  - SVD of A can be obtained from eig of embedding $\begin{pmatrix} 0 & A^* \\ A & 0 \end{pmatrix}$
-
- **“Asymmetric SVD”** : Embed A + randomly asymmetrize + (asym) eigs

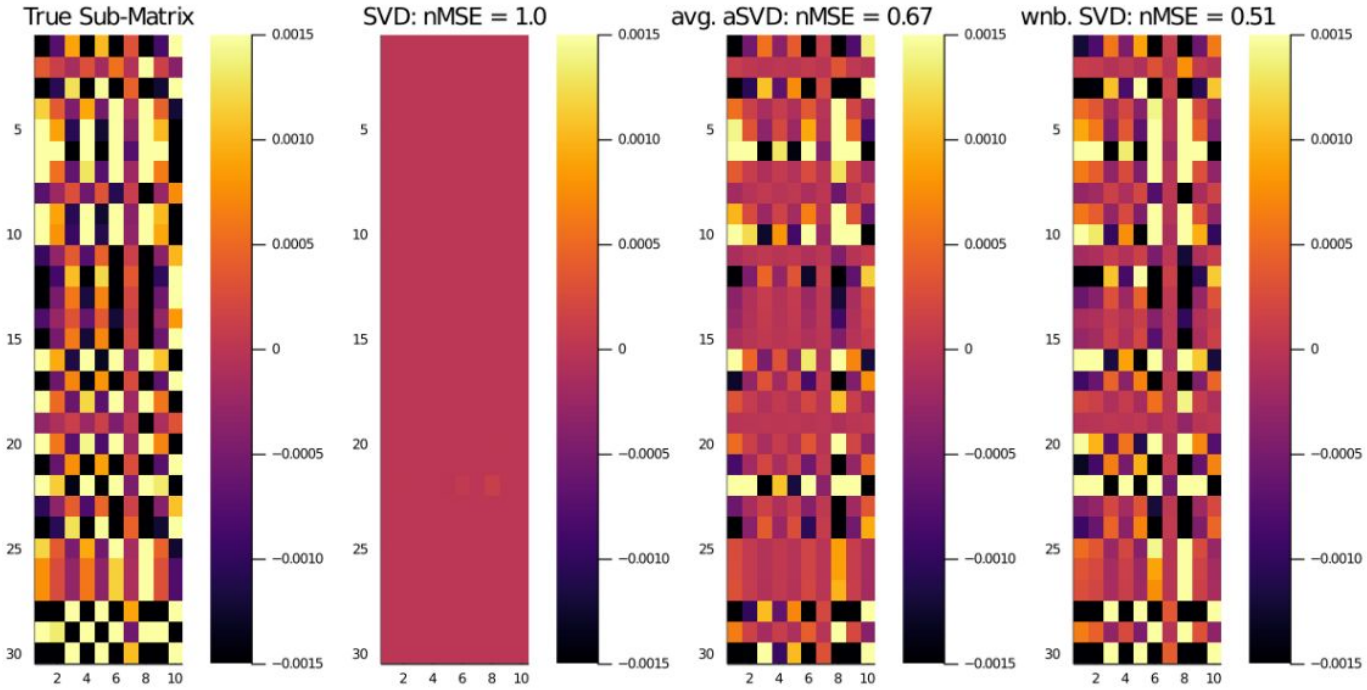
The randomized “asymmetric SVD”

sition.



(B) Rectangular matrix completion via asymmetrization and (non-symmetric) eigendecomposition.

Improved Very Sparse Matrix Completion



(B) Hyperbolic secant distributed singular vectors.

RMT Reasoning about difficulty of different regimes

Each matrix element observed with probability p :

Observed A = E[A] + **(A - E[A])** = p A + **Δ**

$$\text{Error in singular vectors of low-rank } A + \Delta \leq C(A) \cdot \sigma_1(\Delta),$$

Expectation:

- **Perfect reconstruction** if *vanishing perturbation* ⇔ d = O(n log n)
- **Imperfect/Noisy reconstruction** if *bounded perturbation* ⇔ d = O(1)
- **Junk reconstruction** if *unbounded perturbation* ⇔ d = O(1) ?

So far ..

Told you about …

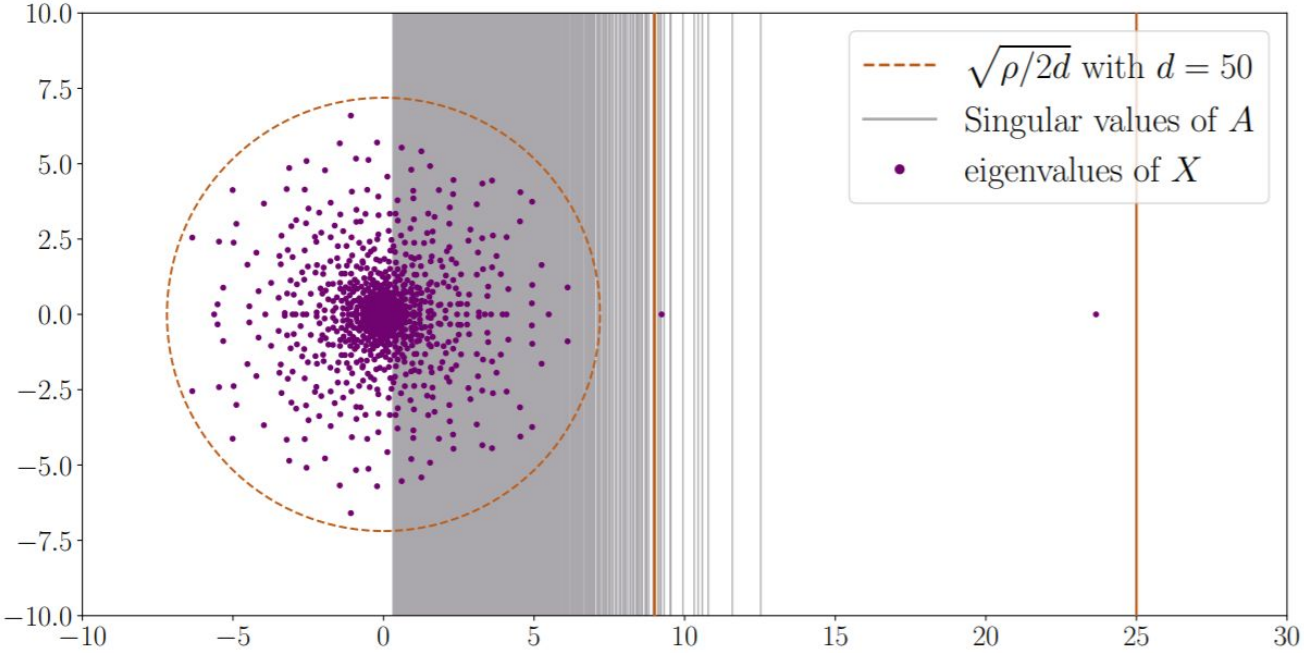
New “asymmetric eig” method that **works when SVD fails**

(very sparse regime)

Have you actively wonder (before I tell you)

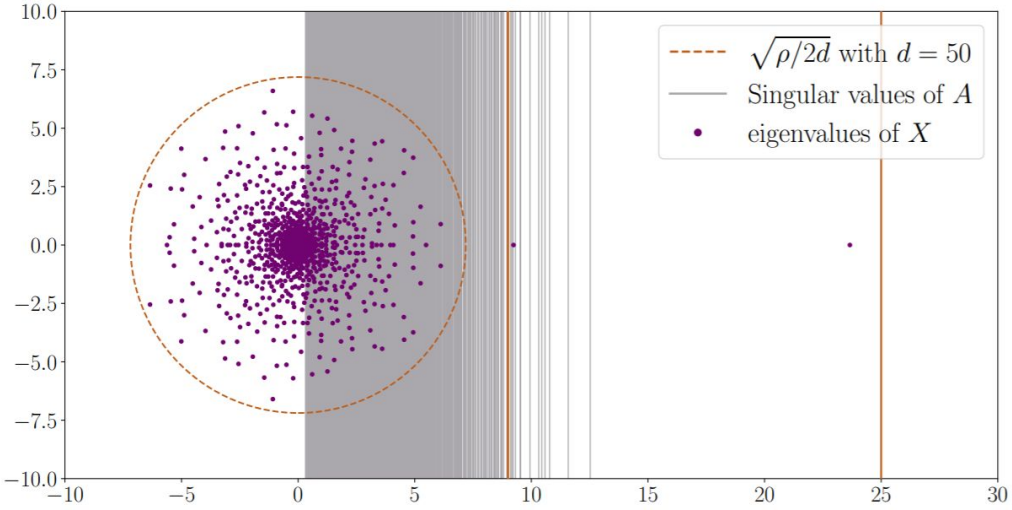
What theory ⇔ computational simulation gave (crazy) idea?

The simulation that led to the idea



- Singular values are spread out (grey lines)
- Complex eigenvalues as dots on plane

The simulation that led to the idea



- Singular values are spread out (grey lines)
- **RMT insight** (Chafai and others): Asymptotically for Erdos-Renyi (ER) graph
  - *undirected* E-R **operator norm** unbounded
  - BUT
  - *directed* E-R **spectral radius** is bounded

Weighted non-backtracking based matrix completion

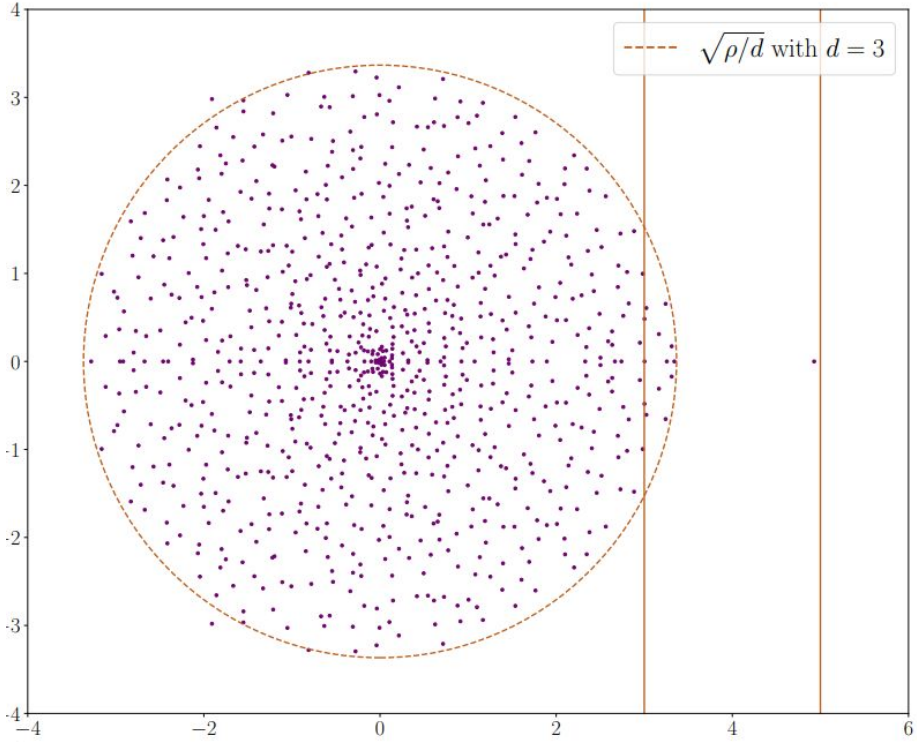
Definition:

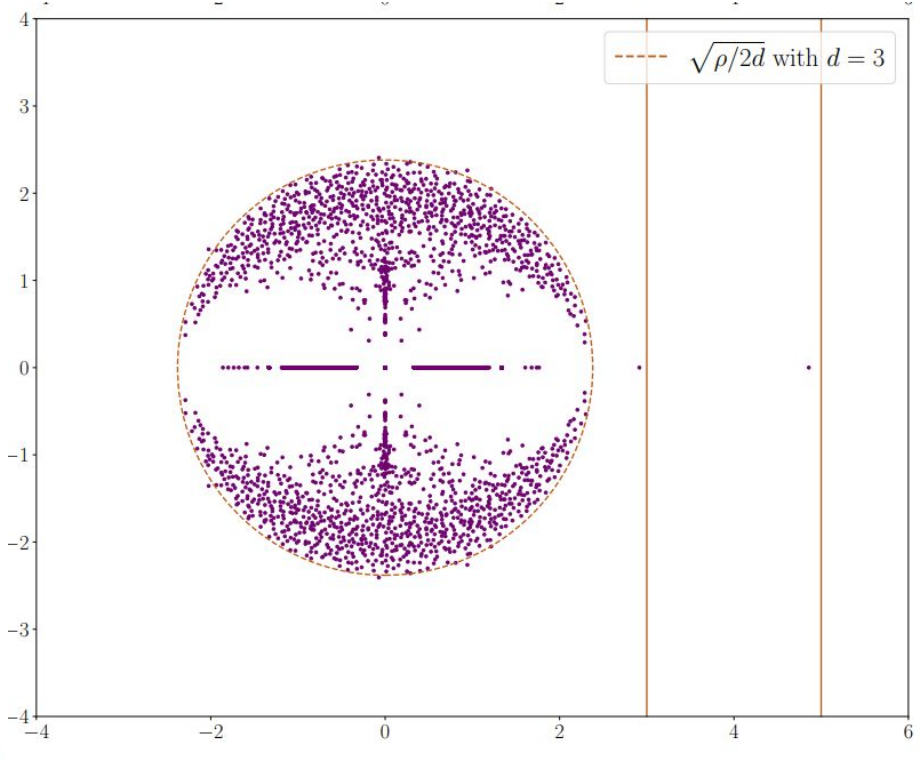
The *weighted non-backtracking matrix* $B \in \mathscr{M}_E(\mathbb{R})$ is the non-symmetric matrix indexed by E with entries, for $e = (x, y) \in E$ and $f = (a, b) \in E$ (those are directed edges):

$$B_{e,f} = \frac{n}{\bar{d}} \mathbf{1}_{a=y} \mathbf{1}_{x \neq b} P_{a,b}.$$

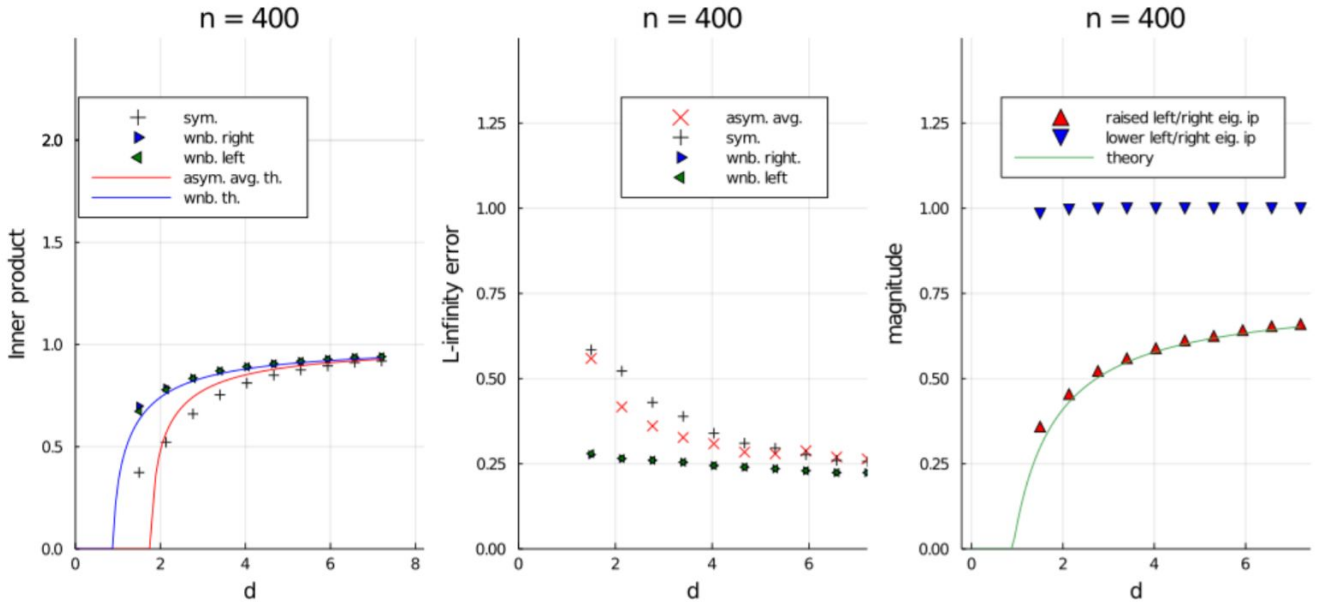
- |E| x |E| sized matrix
- Eigenvectors need to be ‘lowered’ by summing over edge/vertex pairs (see paper)
- **Theorem [BCN’20]: Lower threshold by $\frac{1}{\sqrt{2}}$ (uses all info)**

Asvd (left) vs Non-backtracking (right) spectrum





Weighted non-backtracking vs Asym eigs.



- Lower limit
- Computationally more expensive by $O(d^3)$

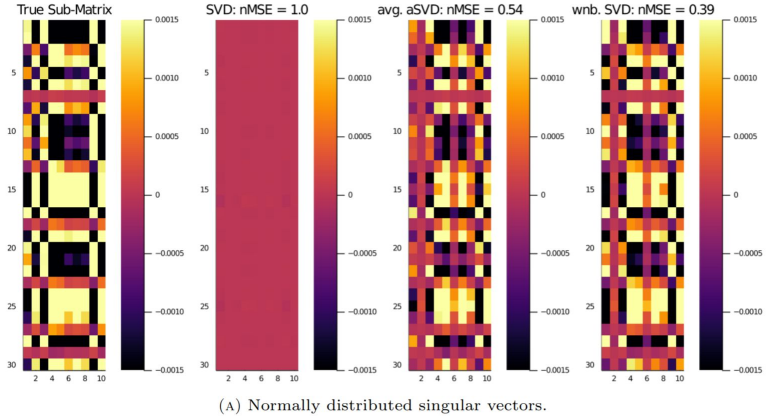
Proof -- arxiv.org/pdf/2005.06062.pdf

- Relies on understanding of eigenvectors of rooted Galton-Watson trees
- Deeply understanding spectrum non-backtracking operator
-
- Hoffman-Wielandt identity plays critical role in bounding perturbation

  (Charles Bordenave + Simon Coste are masters of this)

Summary: Very sparse matrix completion

$$A = \left(\frac{n}{d}\right) P \odot M,$$



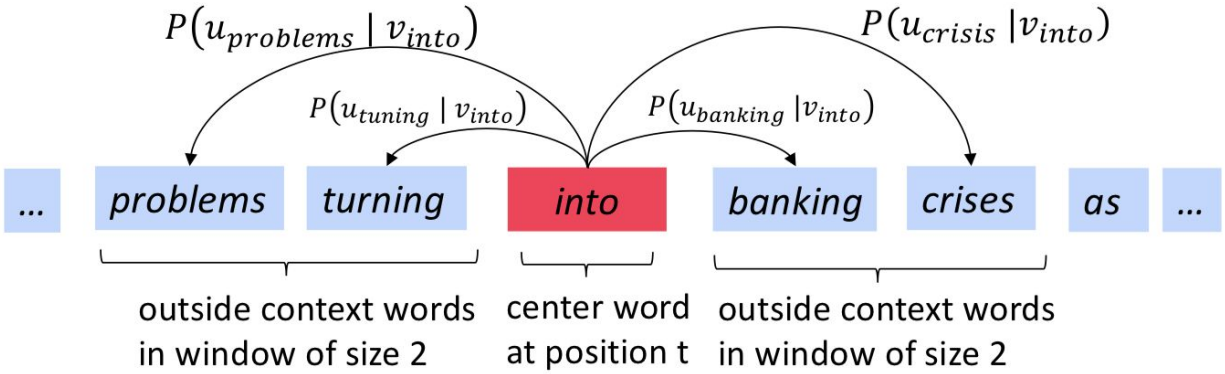
(A) Normally distributed singular vectors.

Questions answered: When *d* = *O(1):*

- Possible to reconstruct reliably with error (i.e. not perfectly) and no thresholding/pruning (ala Keshavan, Montanari, Oh - 2009) ? **Yes**
- Polynomial time algorithm(s)? **Two algorithms**
- Fundamental limit(s) of reliable reconstruction? **Two limits**
  - **Non-universal limit(s) depend on fourth moment of elements of the singular vectors**
  - **More powerful algorithm is O(d^2) more computationally expensive**

Math + RMT opportunities in Emergent AI systems

e.g Predict word from context of **d** words



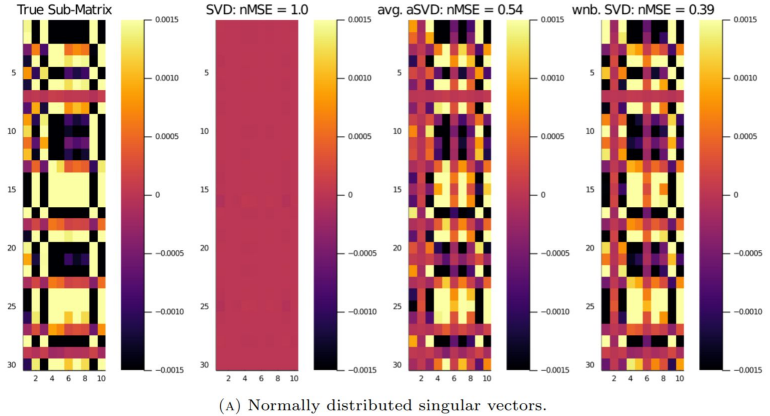
=> **Context Embedding Matrix** is **d** sparse

Application Goal: How to select d?

Creative Math opportunity: How to infuse RMT insights into modern AI systems?

Summary: Very sparse matrix completion

$$A = \left(\frac{n}{d}\right) P \odot M,$$



(A) Normally distributed singular vectors.

Questions answered: When *d* = *O(1)*:

- Possible to reconstruct reliably with error (i.e. not perfectly)? **YES!**
- Polynomial time algorithm(s)? **Two algorithms!**
- Fundamental limit(s) of reliable reconstruction? **Two limits!**
  - **Non-universal limit(s) depend on fourth moment of elements of the singular vectors**
  - **More powerful algorithm is $O(d^2)$ more computationally expensive**