


## Co-Factor Analysis of Citation Networks

Alex Hayes & Karl Rohe

**To cite this article:** Alex Hayes & Karl Rohe (01 Oct 2024): Co-Factor Analysis of Citation Networks, Journal of Computational and Graphical Statistics, DOI: [10.1080/10618600.2024.2394464](https://doi.org/10.1080/10618600.2024.2394464)

**To link to this article:** <https://doi.org/10.1080/10618600.2024.2394464>




View supplementary material 



Published online: 01 Oct 2024.



Submit your article to this journal 



Article views: 100



View related articles 



View Crossmark data 



# Co-Factor Analysis of Citation Networks

Alex Hayes  and Karl Rohe

Department of Statistics, University of Wisconsin-Madison, Madison, WI

## ABSTRACT

One compelling use of citation networks is to characterize papers by their relationships to the surrounding literature. We propose a method to characterize papers by embedding them into two distinct “co-factor” spaces: one describing how papers send citations, and the other describing how papers receive citations. This approach presents several challenges. First, older documents cannot cite newer documents, and thus it is not clear that co-factors are even identifiable. We resolve this challenge by developing a co-factor model for asymmetric adjacency matrices with missing lower triangles and showing that identification is possible. We then frame estimation as a matrix completion problem and develop a specialized implementation of matrix completion because prior implementations are memory bound in our setting. Simulations show that our estimator has promising finite sample properties, and that naive approaches fail to recover latent co-factor structure. We leverage our estimator to investigate 255,780 papers published in statistics journals from 1898 to 2024, resulting in the most comprehensive topic model of the statistics literature to date. We find interpretable co-factors corresponding to many statistical subfields, including time series, variable selection, spatial methods, graphical models, GLM(M)s, causal inference, multiple testing, quantile regression, semi-parametrics, dimension reduction, and several more. Supplementary materials for this article are available online.

## ARTICLE HISTORY

Received August 2023  
Accepted August 2024

## KEYWORDS

Co-factor models; Matrix completion; Missing data; Spectral network analysis; Stochastic blockmodels

## 1. Introduction

Suppose we have a collection of written documents, and these documents cite each other. For example, the documents might be academic papers, judicial opinions, or patents, among other possibilities. One useful way to understand individual documents in the collection, and the collection as a whole, is to find documents that cite, and are cited, in similar ways. These documents are likely to be about the same subject, and can thus reveal information about important topics in the corpus.

We develop a network-based approach to understanding the structure in citation corpora, called *CitationImpute*. *CitationImpute* begins by representing a corpus as a network, where each document corresponds to a node, and citations between documents correspond to directed edges. Then, it uses a spectral factorization technique to embed each document into two distinct latent spaces, one characterizing how papers cite, and the other characterizing how papers get cited.

Unlike prior approaches to citation analysis, *CitationImpute* models citations from older documents to newer documents as structurally missing. As a consequence, our algorithm must estimate singular subspaces via matrix completion methods. Existing matrix completion methods are computationally prohibitive in this setting, so we develop a singular subspace estimator with reasonable time and space complexity.

After estimating singular subspaces, *CitationImpute* uses varimax rotation to identify latent factors in the network (as

opposed to k-means, or k-medians clustering). This allows each document to have a weighted membership in each cluster. The overall procedure can be understood intuitively in the context of stochastic blockmodels, but is appropriate for a much broader class of low-rank network models.

We validate the new procedure with a simulation study, finding that the new estimator recovers latent factors under a partially observed stochastic blockmodel. Finally, we analyze 255,780 statistics papers and 2.2 million citations published in journals on statistics and probability, producing a comprehensive breakdown of topics in the statistics literature. We present the keywords most associated with these topics in [Table 1](#) (factors describing how papers get cited) and [Table 2](#) (factors describing how papers cite).

*CitationImpute* is related to several lines of extant work, most notably empirical investigations of the academic statistics literature. Selby (2020) and Stigler (1994) consider relationships between statistics papers and the larger academic literature, with Selby (2020) reviewing approaches to community detection in networks and suggesting a number of diagnostic techniques for assessing model fit. Ji et al. (2022), an expansion of Ji and Jin (2016), considers a dataset with about a third as many papers as our own, and investigates undirected (and dynamic) networks of academic authors based on co-authorship and co-citation. Ji et al. (2022) estimates researcher interests by embedding researchers into a three-dimensional latent space. In contrast, we model the topics of individual manuscripts,

**Table 1.** Keywords for Y (incoming citation) factors.

Factor name	Top words	ID
Non-convex penalties	selection, variable, penalized, oracle, lasso, nonconcave	y01
Feature screening	screening, dimensional, ultrahigh, feature, independence, high	y02
Bayesian model selection	bayesian, models, complexity, disease, model, fit	y03
Post-selection inference	high, dimensional, lasso, regression, confidence, dantzig	y04
Survival analysis	survival, censored, hazards, proportional, cox, regression	y05
Information criteria	model, clustering, mixture, selection, dimension, mixtures	y06
Causal inference	propensity, causal, score, observational, treatment, effects	y07
Multiple testing	false, discovery, multiple, rate, testing, controlling	y08
Graphical models	graphical, covariance, estimation, sparse, lasso, high	y09
Bayesian non-parametrics	dirichlet, bayesian, nonparametric, mixture, mixtures, priors	y10
Supervised dimension reduction	dimension, reduction, regression, sliced, inverse, sufficient	y11
LASSO (optimization)	lasso, regularization, coordinate, descent, selection, via	y12
LASSO (classic)	lasso, selection, shrinkage, regression, via, longitudinal	y13
Kriging	spatial, gaussian, datasets, covariance, large, temporal	y14
Empirical likelihood	empirical, likelihood, confidence, ratio, intervals, regions	y15
GLM(M)s	longitudinal, data, generalized, models, estimating, binary	y16
Functional data	functional, regression, principal, data, linear, longitudinal	y17
Skew normals	skew, normal, distributions, multivariate, distribution, t	y18
Quantile regression	quantile, regression, quantiles, censored, median, estimation	y19
Bayesian model selection	bayesian, selection, variable, bayes, priors, prior	y20
Missing data	missing, imputation, data, longitudinal, nonignorable, nonresponse	y21
Adaptive clinical trials	trials, clinical, adaptive, sequential, group, multiple	y22
Splines + random effects	models, mixed, splines, smoothing, longitudinal, regression	y23
Multivariate analysis	covariance, matrices, high, dimensional, large, matrix	y24
MCMC	monte, carlo, markov, metropolis, chain, bayesian	y25
Single index models	coefficient, varying, models, index, single, partially	y26
Causal semiparametrics	missing, semiparametric, regression, sampling, data, estimation	y27
Individual/optimal treatment	treatment, regimes, individualized, learning, optimal, estimating	y28
RIDGE	ridge, regression, biased, linear, estimators, estimator	y29
Cure models	cure, survival, censored, rate, mixture, hazards	y30

**Table 2.** Keywords for Z (outgoing citation) factors.

Factor name	Top words	ID
Non-convex penalties	selection, variable, dimensional, high, penalized, lasso	z01
Experimental design	screening, dimensional, high, ultrahigh, feature, supersaturated	z02
Bayesian spatial stats	bayesian, models, spatial, model, longitudinal, hierarchical	z03
Post-selection inference	high, dimensional, lasso, recurrent, selection, regression	z04
Survival analysis	survival, hazards, censored, cox, data, proportional	z05
Mixture models	selection, clustering, model, mixture, models, mixtures	z06
Causal inference	propensity, causal, score, treatment, missing, observational	z07
Multiple testing	false, discovery, testing, multiple, rate, microarray	z08
Graphical models	graphical, high, dimensional, models, sparse, estimation	z09
Bayesian non-parametrics	bayesian, dirichlet, nonparametric, mixture, clustering, process	z10
Supervised dimension reduction	dimension, reduction, sufficient, index, inverse, sliced	z11
Times series	garch, volatility, series, models, time, change	z12
Sparse multivariate analysis	selection, lasso, high, sparse, variable, dimensional	z13
Kriging	spatial, spatio, temporal, gaussian, fields, bayesian	z14
Empirical likelihood	empirical, likelihood, inference, missing, partially, jackknife	z15
GEE	longitudinal, data, generalized, binary, estimating, clustered	z16
Functional data	functional, data, regression, longitudinal, principal, linear	z17
Skew normals	skew, normal, distributions, multivariate, distribution, t	z18
Quantile regression	quantile, regression, quantiles, censored, composite, expectile	z19
Bayesian model selection	bayesian, selection, variable, priors, prior, model	z20
Missing data	missing, imputation, data, longitudinal, with, nonignorable	z21
Adaptive clinical trials	adaptive, trials, clinical, sequential, designs, group	z22
Splines + random effects	models, mixed, splines, penalized, regression, additive	z23
Multivariate analysis	high, dimensional, covariance, matrices, matrix, factor	z24
MCMC	bayesian, carlo, monte, mcmc, metropolis, chain	z25
Single index models	varying, coefficient, models, index, single, partially	z26
Joint longitudinal/survival models	longitudinal, mixed, models, data, joint, effects	z27
Causal inference reviews	causal, treatment, effects, propensity, instrumental, effect	z28
RIDGE	ridge, regression, estimator, liu, linear, estimators	z29
Cure models	cure, censored, survival, model, rate, data	z30

and co-embed manuscripts into much more detailed thirty-dimensional “sending” and “receiving” latent spaces. Rohe and Zeng (2023) co-factor a directed network of journal-journal citation counts using varimax factor analysis, but aggregate citations over time and thus avoid the chronological missingness we consider here.

Methodologically, `CitationImpute` is an extension of the varimax rotation technique studied in Rohe and Zeng (2023), and is closely related to co-clustering methods (Choi and Wolfe 2014; Rohe, Qin, and Yu 2016; Choi 2017), as well as clustering methods for bipartite networks (Larremore, Clauset, and Jacobs 2014; Razaee, Amini, and Li 2019; Yen and Larremore 2020), some of which can be extended to handle missing data (Peixoto 2018; Zhao et al. 2022). While there is a large literature on network clustering with missing data, these techniques cannot be used for co-factoring and co-clustering. Nonetheless, some techniques similarly leverage nuclear norm penalized singular subspace estimation to handle missing edges (Chen et al. 2014; Vinayak, Oymak, and Hassibi 2014; Li, Levina, and Zhu 2020). There have also been some efforts to incorporate topic structure into preferential attachment models (Pollner, Palla, and Vicsek 2006; Hajek and Sankagiri 2019), bridging the gap between mixture modeling and more traditional bibliometric analysis (Price 1976).

Finally, our work is related to the general matrix completion literature, in particular nuclear norm penalization approaches for estimating partially observed matrices (Mazumder, Hastie, and Tibshirani 2010; Kim and Choi 2013; Bhojanapalli and Jain 2014; Gu et al. 2014; Klopp 2014; Shamir and Shalev-Shwartz 2014; Cui et al. 2015; Hosono, Ono, and Miyata 2016; Gu et al. 2017; Cho, Kim, and Rohe 2019; Zhang and Ng 2019; Yang, Li, and Wang 2022). While this literature has recently made impressive inroads regarding the consistency of nuclear-norm regularization for spectral recovery in deterministic and nonuniform sampling settings (Foucart et al. 2021; Zhu, Wang, and Samworth 2022), we are unaware of consistency results for the upper triangular observation pattern present in citation data, and thus validate our approach with simulations.

## Notation

Let  $\mathbf{u}_i(A)$ ,  $\lambda_i(A)$ ,  $\mathbf{v}_i(A)$  be functions that return the  $i$ th left singular vector, singular value, and right singular vector of a matrix  $A$ , respectively. Similarly, define  $\lambda_i^2(A) = (\lambda_i(A))^2$ . We use  $\langle \cdot, \cdot \rangle$  to denote the Frobenius inner product and  $\| \cdot \|_F$  the Frobenius norm. Let  $A_i$  denote the  $i$ th row of a matrix  $A$  and  $A_{\cdot j}$  denote the  $j$ th column. For a partially observed matrix  $A$ , let  $\Omega_A$  be the set  $\{(i, j) : A_{ij} \text{ is observed}\}$  and  $\tilde{\Omega}_A$  be the set  $\{(i, j) : A_{ij} \text{ is observed and nonzero}\}$ ; when  $A$  is clear from context we will omit the subscript  $A$ . By  $Y_A$  ( $Y$  when the context is clear) we denote the binary matrix such that  $Y_{ij}$  is one when  $(i, j) \in \Omega_A$  and zero otherwise.  $\odot$  indicates elementwise multiplication between two matrices with the same dimensions. We use  $P_{\Omega_A}(B) = B \odot Y_A$  to denote the projection of a matrix  $B$  onto observed support of another matrix  $A$ , and  $P_{\Omega_A}^\perp(B) = B \odot (1 - Y_A)$ . Let  $P_\ell(A)$  denote the “clipping” projection that sets the first  $\ell$  columns and the last  $\ell$  rows of  $A$  all to zero. Finally,  $g(n) = \mathcal{O}(f(n))$  means that  $\lim_{n \rightarrow \infty} g(n)/f(n) \leq M$  for some constant  $M$ . All proofs are deferred to the Appendix.

## 2. Model

### 2.1. Co-Factor Model

We use the co-factor model of Rohe and Zeng (2023) as a model for latent similarities between documents. The co-factor model is a low-rank, distributionally agnostic generalization of the stochastic co-blockmodel (Holland, Laskey, and Leinhardt 1983; Rohe, Qin, and Yu 2016), and includes sub-models such as stochastic blockmodels, degree-corrected stochastic blockmodels (Karrer and Newman 2011), (degree-corrected) mixed membership stochastic blockmodels (Airoldi et al. 2008; Jin, Ke, and Luo 2024), latent dirichlet allocation (Blei, Ng, and Jordan 2003), and (generalized) random dot product graphs (Lyzinski et al. 2014), many of which are closely related to topic models (Gerlach, Peixoto, and Altmann 2018).

In the co-factor model, each document  $i$  possesses two co-factors. One co-factor,  $Z_i \in \mathbb{R}^k$ , controls outgoing citations, or the topics that a paper is likely to cite, and the other co-factor,  $Y_i \in \mathbb{R}^k$ , controls incoming citations, or the topics that a paper is likely to be cited by. The co-factor structure of the model operationalizes the fundamental difference between citing and being cited. Mathematically, co-factor models are generalizations of factor models, and there are compelling reasons to model full co-factor structure: co-factor structure is theoretically necessary to capture key features of real world network data (Chanpuriya et al. 2020), an observation empirically verified by Rohe, Qin, and Yu (2016) and Qing and Wang (2022), among others.

**Example 2.1.** Consider Tibshirani (1996), which introduced LASSO regression. The LASSO paper builds upon a small body of statistical work on variable selection and resampling, but itself forms the basis for a large body of applied work, especially in genomics and biomedical settings. The directionality of citations is clear in the reference counts: Tibshirani (1996) cites twenty papers, but is cited by tens of thousands of papers. If we do not distinguish between papers cited and citing papers, we might fail to distinguish between the genomics literature (incoming co-topic) and the variable selection literature (outgoing co-topic), as well as differing propensities to cite and to be cited.

In the co-factor model, conditional on the latent factors, each edge  $A_{ij}$  of the network is sampled independently from a distribution with expectation  $\mathcal{A} \equiv \mathbb{E}(A \mid Z, B, Y) = ZBY^T \in \mathbb{R}^{n \times n}$  where  $B \in \mathbb{R}^{k \times k}$  is a mixing matrix that controls how the outgoing and incoming latent factors interact. In the citation setting,  $\mathcal{A}$  represents the similarities between documents in the latent topic space.  $B$  is a weighting matrix that describes how likely it is that a document  $i$  loading on outgoing factor  $Z_{\cdot k}$  forms an edge to a document  $j$  loading on incoming factor  $Y_{\cdot \ell}$ . As the  $B$ -mediated similarity between the outgoing topic of document  $i$  and the incoming similarity of document  $j$  increases, (i.e.,  $A_{ij}$  gets larger), the probability of citation  $i \rightarrow j$  goes up.

For the co-factor model to be identified, the co-factors  $Z$  and  $Y$  and the mixing matrix  $B$  must satisfy several assumptions: the mixing matrix  $B$  must be full rank, the rows of  $Z$  and  $Y$  must be independent and identically distributed (that is,  $Z_{1\cdot}, Z_{2\cdot}, \dots, Z_{n\cdot}$  must be iid, and  $Y_{1\cdot}, Y_{2\cdot}, \dots, Y_{n\cdot}$  must be iid), and the distribution of the  $Z_i$  and  $Y_i$  must be leptokurtic (i.e., skewed). Skewness is the key assumption for  $Z$  and  $Y$  to be identified.

When  $Z$  and  $Y$  come from leptokurtic distributions, the co-factors  $Z$  and  $Y$  are identified up to sign-flips and permutations of the column order.

The co-factor model is similar in form to mixed membership stochastic blockmodels, and generalizes the mixed membership stochastic blockmodel (see the supplement of Rohe and Zeng (2023) for a precise characterization). Unlike mixed membership stochastic blockmodels, the rows of  $Z$  and  $Y$  do not need to be normalized, and can take on negative values. In practice, the sign ambiguity of  $Z$  and  $Y$  can almost always be resolved by forcing the columns of  $Z$  and  $Y$  to be skew positive, in which case  $Z$  and  $Y$  typically consist of sparse, axis-aligned, positive values. The sparsity of  $Z$  and  $Y$  often enables substantive interpretations of the latent factors, as each node typically loads on a small number of factors.

## 2.2. Chronological Observation Mechanism

To specialize the co-factor model to the citation setting, we incorporate an observation mechanism.

**Definition 2.1.** Given a corpus of documents  $i = 1, \dots, n$  published at times  $T_1, \dots, T_n$ , the partially observed adjacency matrix is

$$A_{ij} = \begin{cases} 1 & \text{if } T_j \leq T_i \text{ and } i \text{ cites } j, \\ 0 & \text{if } T_j \leq T_i \text{ and } i \text{ does not cite } j, \text{ and} \\ \text{unobserved} & \text{if } T_j > T_i. \end{cases} \quad (2.1)$$

For convenience, we re-index the documents in order of publishing times, forcing  $T_1 \geq \dots \geq T_n$ , such that  $T_1$  is the most recent publishing time, and  $T_n$  is the earliest publishing time. Using this indexing scheme, the observed portion of the network is nearly upper triangular, but elements can occur in the lower triangle when  $T_i = T_j$ .

Under the citation observation mechanism, citations from older papers to newer papers are missing. This is because the lack of citation from older papers to newer papers should be uninformative about the outgoing co-factor of the older paper and the incoming co-factor of the newer paper.

If we presume that the older paper definitively cites the newer paper, or definitively does not cite the newer paper, this will force the corresponding co-factors closer together or farther apart in the latent topic space. `CitationImpute` thus treats citations forward-in-time as missing rather than precisely observed zeroes or ones. This allows the estimation procedure to spectrally infer co-factors without introducing chronological artifacts.

**Example 2.2.** Consider Hoerl and Kennard (1970), which introduced RIDGE regression. Since the RIDGE paper was published long before the LASSO paper, Hoerl and Kennard (1970) does not cite Tibshirani (1996). But, since RIDGE regression and LASSO regression are closely related, it is plausible that the two papers are close to each other in outgoing topic space. The impossibility of citation forward-in-time is uninformative about the latent similarity between the two papers.

**Remark 2.1.** The chronological observation mechanism is only relevant if citations are directed relationships. If there is no semantic information contained in the direction of a citation, we can impute the lower triangle of  $A$  based on the upper triangle of  $A$  by setting  $A_{ij} = A_{ji}$  for all missing edges.

**Remark 2.2.** In some settings, such as the scientific literature, documents might build on each other, with later documents iterating on past work. In a co-factor model, one could argue that this should be modeled as dependence amongst the latent factors. We are not aware of any approaches to handle such dependence, but believe they are an interesting topic for future work.

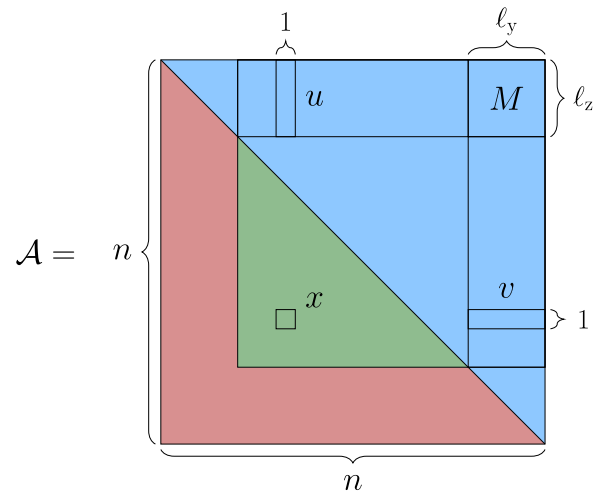
## 2.3. Statistical Identification of Latent Co-Factors

The chronological observation mechanism presents several challenges. First, it is unclear if the co-factors  $Z$  and  $Y$  are identified based on the information observed in the upper triangle of  $A$ .

In Proposition 2.1, we show that outgoing community memberships  $Z_i$  are identified for all but the very earliest documents, and that the incoming community memberships  $Y_i$  are identified for all but the most recent documents. Some co-factors are unidentified because the most recent documents have not been around long enough to possibly be cited by papers from all topics and because the oldest documents were written too early to possibly cite papers from all topics.

More precisely, Proposition 2.1 states that if the conditional expectation of a citation network  $\mathcal{A}$  is rank  $k$  and the  $\ell_z \times \ell_y$  submatrix in the top right of  $\mathcal{A}$  is rank  $k$ , it is possible to reconstruct all of  $\mathcal{A}$  except for the elements in the last  $\ell_z$  rows and the elements in the first  $\ell_y$  columns. Observing a full rank matrix  $M$  in the top right of  $\mathcal{A}$  ensures that no information is hidden in the lower triangle (see Figure 1).

The statement of Proposition 2.1 requires some additional notation. Let  $\mathcal{R}_{n,k}$  be the set of rank  $k$  matrices contained in



**Figure 1.** A decomposition of a conditional expectation matrix  $\mathcal{A}$ . Elements in the upper triangle are observed in the citation setting. We would like to recover elements of the lower triangle based on the information in the upper triangle. When  $\text{rank}(\mathcal{A}) = \text{rank}(M)$ , a portion of the lower triangle is identified, but the left-most columns and bottom-most rows cannot be recovered.



$\mathbb{R}^{n \times n}$ . Imagine that  $\mathcal{A}, \mathcal{B} \in \mathcal{R}_{n,k}$  are the conditional expectations of two semi-parametric factor models.  $P_U(\mathcal{A})$  and  $P_U(\mathcal{B})$  are projections of  $\mathcal{A}$  and  $\mathcal{B}$  onto the space of upper triangular matrices.  $P_U(\mathcal{A})$  and  $P_U(\mathcal{B})$  represent the conditional expectations of the observed portion of  $\mathcal{A}$  and  $\mathcal{B}$ , respectively.

**Proposition 2.1.** Let  $\mathcal{A}, \mathcal{B} \in \mathcal{R}_{n,k}$ . If  $P_U(\mathcal{A}) = P_U(\mathcal{B})$  and there exist  $\ell_z, \ell_y \in \{k, \dots, n/2\}$  such that  $M = \mathcal{A}_{[1:\ell_z, (n-\ell_y):n]}$  has rank  $k$ , then  $\mathcal{A}_{ij} = \mathcal{B}_{ij}$  for all  $i, j \in \mathbb{Z}$  satisfying  $1 < i \leq (n - \ell_z)$  and  $\ell_y < j \leq n$ .

**Remark 2.3.** Proposition 2.1 can be understood constructively as approximating  $\mathcal{A}$  from  $P_U(\mathcal{A})$  using the Nyström method (Drineas and Mahoney 2005; Gittens and Mahoney 2016). Because  $\mathcal{A}$  is rank  $k$ , the Nyström method has zero approximation error.

When the conditions of Proposition 2.1 are violated, it is possible that no elements in the lower triangle of  $\mathcal{A}$  are recoverable. That is, if  $M$  is rank  $k$  only for  $\ell_z, \ell_y > n/2$ , there is information hidden in the lower triangle of  $\mathcal{A}$  that is lost upon projecting onto the upper triangle. A concrete example where recovery is impossible is given by the following matrix  $\mathcal{A}$ , which has rank two. Let  $J_n$  denote an  $n \times n$  matrix of ones and suppose

$$\mathcal{A} = \begin{bmatrix} a J_{n/2} & a J_{n/2} \\ b J_{n/2} & a J_{n/2} \end{bmatrix}.$$

Every element in the upper triangle of  $\mathcal{A}$  is  $a$ , and thus there is no way of estimating  $b$  when the lower triangle of  $\mathcal{A}$  is missing.  $\mathcal{A}$  corresponds to a two-block stochastic blockmodel where the first  $n/2$  documents are in one block and the last  $n/2$  documents are in a separate block. Because these blocks of documents do not overlap in time (here represented by node order) and they have asymmetric citation probabilities, all information about asymmetric citation probabilities is lost. However, if  $n$  is large and the rows and columns of  $\mathcal{A}$  are permuted according to the same random permutation, then there are  $\ell_z, \ell_y \ll n/2$  that satisfies the conditions of Proposition 2.1 with high probability.

For balanced stochastic co-blockmodels, such as the one used in the simulation study, it is sufficient to take  $\ell_z = \ell_y = 2k \log k$  to achieve identifiability with high probability. This demonstrates that identification with  $\ell_z, \ell_y \ll n/2$  is reasonable in blockmodel-like settings.

**Proposition 2.2.** Suppose  $\mathcal{A}$  characterizes the expected adjacency matrix of the simulation test-bed model (Definition 4.2). Let  $\ell_z = \ell_y = 2k \log k$  and let  $M$  be as defined in Proposition 2.1. Then  $\mathbb{P}(\text{rank}(M) = k) = 1 - 2n^{-1}$ .

### 3. Spectral Estimation

Spectral clustering typically proceeds in three steps. First, the network is represented as a matrix, often the adjacency matrix, but sometimes normalized or regularized versions of the graph Laplacian. Second, the leading singular vectors of this matrix are estimated, which associates each node in the graph with a point in Euclidean space. Lastly, these node embeddings are analyzed using standard methods for Euclidean data. While this estimation strategy may seem ad hoc, spectral estimators

perform statistical inference under network models that are identified by their singular subspaces, a large class of models that includes stochastic blockmodels and many generalizations thereof (von Luxburg 2007; Lyzinski et al. 2014; Athreya et al. 2015; Jin 2015; Lei and Rinaldo 2015; Ji and Jin 2016; Rohe, Qin, and Yu 2016; Lyzinski et al. 2017; Athreya et al. 2018; Priebe et al. 2019).

### 3.1. The Algorithm

CitationImpute adapts the standard spectral estimation pipeline to the citation setting. The main difference is that we cannot estimate singular subspaces of the adjacency matrix using a singular value decomposition, due to missing data. Instead, CitationImpute uses the AdaptiveImpute algorithm of Cho, Kim, and Rohe (2019), which is a self-tuning variant of the softImpute algorithm of Mazumder, Hastie, and Tibshirani (2010).

CitationImpute accepts as input a network adjacency matrix  $A \in \mathbb{R}^{n \times n}$  where the lower triangle is assumed to be mostly missing, a desired number of co-factors  $k \in \{2, \dots, n\}$ , and clipping parameters  $\ell_z, \ell_y \in \{k, \dots, n/2\}$ . The algorithm then proceeds as follows.

1. Set all elements in the first  $\ell_y$  columns of  $A$  and last  $\ell_z$  rows of  $A$  to zero. This means that edges corresponding to the unidentified rows of  $Z$  and  $Y$  are ignored during estimation; see Proposition 2.1 for details.
2. Estimate the singular vectors and singular values of  $A \approx \widehat{U} \widehat{D} \widehat{V}^T$  using AdaptiveImpute (Cho, Kim, and Rohe 2019). In Section 3.2 we describe why a naive implementation of AdaptiveImpute is computationally infeasible, and in Section 3.3 we outline our computational contributions and a practical implementation of AdaptiveImpute for upper triangular data.
3. Compute the varimax rotations of  $\widehat{U}$  and  $\widehat{V}$  and construct rotated singular vector matrices  $\widehat{Z}, \widehat{B}$ , and  $\widehat{Y}$ , respectively. We briefly review varimax rotation in Section 3.4.

CitationImpute has several hyperparameters: the number of desired co-factors  $k$ , and the clipping parameters  $\ell_z$  and  $\ell_y$ . In simulations, we find that  $\ell_z = \ell_y = n/10$  are good default values, but recommend applying domain knowledge as appropriate, and conducting a sensitivity analysis (see Section 5.2 for an example).

### 3.2. The Computational Problem

One contribution of this article is a collection of algebraic identities (Propositions 3.1 and 3.2) that allow for an efficient implementation of AdaptiveImpute on citation matrices with hundreds of thousands of documents.

To understand why these identities are useful, we must disambiguate between two senses of sparsity. A matrix is *sparse* if most of its elements are zero. These matrices can be represented very efficiently on a computer by recording only the small number of nonzero elements and their indices. On the other hand, a matrix is *sparsely observed* if only a few of its entries are observed, regardless of the value of those entries. These two notations of

sparsity are often conflated, and sparsely observed matrices are often represented as sparse matrices, where implicit zeroes are considered missing, and the observed zeroes must be explicitly tracked.

In the citation setting, the data matrix  $A$ , as defined in (2.1), is densely observed; at least half of the entries are defined by the data. However, in the portion of the network that is observed, the data is sparse, that is, mostly zero-valued. Thus, the usual conflation of sparse and sparsely observed matrices leads to issues: there are  $n(n-1)/2$  elements in the upper triangle of  $A$  that must be explicitly tracked even if they are zero. Using this representation, even moderately sized corpora cannot be held in memory on commodity hardware. Beyond memory considerations, adding approximately  $n(n-1)/2$  explicit zeroes to a sparse matrix slows down matrix operations like matrix-vector multiplication.

This makes matrix completion algorithms infeasible in both time and space when using the naive sparse representation of  $A$ . Both `AdaptiveImpute` and `softImpute` rely on iterated singular value decompositions of a running low-rank approximation  $\tilde{A}^{(t)}$  to  $A$ . In the typical setting where the number of nodes is  $n$ , the rank of the decomposition is  $k$ , and  $n \gg k$ , naively taking a singular value decomposition of  $\tilde{A}^{(t)}$  has time complexity per iteration  $\mathcal{O}(n^2 k)$ . This high computational complexity constrains researchers to inference on networks with at most thousands of nodes.

We are able to reduce the both the time and space complexity of the matrix completion problem (Figure 2). The solution requires leveraging the fact that  $A$  is sparse, even if it is not sparsely observed. In particular, there is no need to explicitly track zeroes in the upper triangle of  $A$ , and  $A$  may be represented as a sparse matrix that records only nonzero elements of  $A$  and zeroes in the lower triangle of  $A$ . Using this representation, with some algebraic tricks, all the operations necessary for `AdaptiveImpute` are computationally feasible. In brief, by representing  $\tilde{A}^{(t)}$  as the sum of four carefully constructed matrices, we can reduce the naive time complexity from  $\mathcal{O}(n^2 k)$  down to  $\mathcal{O}(|\tilde{\Omega}| k + n k^2)$ , where  $|\tilde{\Omega}|$  is the number of observed nonzero elements of  $A$ . In real world datasets  $|\tilde{\Omega}|$  represents the number of citations between documents, and empirical evidence

suggests that each document in a citation network cites a fixed number of other documents, regardless of the overall size of the corpus. That is,  $|\tilde{\Omega}|$  is  $\mathcal{O}(n)$ . Thus the effective per-iteration runtime reduces from  $\mathcal{O}(n^2 k)$  to  $\mathcal{O}(n k^2)$ .

### 3.3. AdaptiveImpute

The `AdaptiveImpute` algorithm is similar to `softImpute` (Hastie et al. 2015), with two key differences. First, `AdaptiveImpute` initializes with a debiased singular value decomposition. Second, on each iteration, `AdaptiveImpute` adaptively varies the `softImpute` thresholding parameter. This procedure is defined in Algorithm 1, which is identical to the algorithm as defined in (Cho, Kim, and Rohe 2019) but with

---

#### Algorithm 1: ADAPTIVEIMPUTE

---

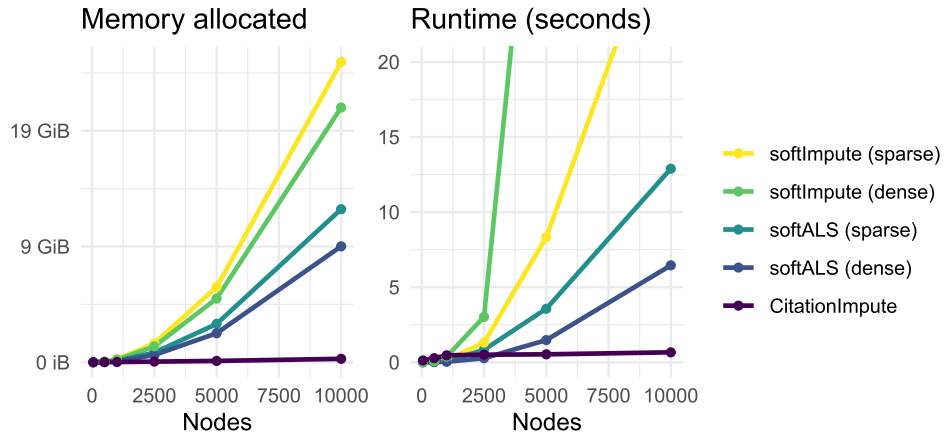
**Input:** partially observed matrix  $A \in \mathbb{R}^{n \times n}$ , rank

$k \in \{2, \dots, n\}$ , convergence tolerance  $\varepsilon > 0$ , and maximum allowable iterations  $T \in \mathbb{Z}^+$ .

```

1  $Z^{(1)} \leftarrow \text{AdaptiveInitialize}(A, k)$ 
2 repeat
3    $\tilde{A}^{(t)} \leftarrow P_{\Omega}(A) + P_{\Omega}^{\perp}(Z^{(t)})$ 
4    $\hat{V}_i^{(t)} \leftarrow \mathbf{v}_i(\tilde{A}^{(t)})$  for  $i = 1, \dots, k$ 
5    $\hat{U}_i^{(t)} \leftarrow \mathbf{u}_i(\tilde{A}^{(t)})$  for  $i = 1, \dots, k$ 
6    $\tilde{\alpha}^{(t)} \leftarrow \frac{1}{n-k} \sum_{i=k+1}^n \lambda_i^2(\tilde{A}^{(t)})$ 
7    $\hat{\lambda}_i^{(t)} \leftarrow \sqrt{\lambda_i^2(\tilde{A}^{(t)}) - \tilde{\alpha}^{(t)}}$  for  $i = 1, \dots, k$ 
8    $Z^{(t+1)} \leftarrow \sum_{i=1}^k \hat{\lambda}_i^{(t)} \hat{U}_i^{(t)} \hat{V}_i^{(t)T}$ 
9    $t \leftarrow t + 1$ 
10 until  $\|Z^{(t+1)} - Z^{(t)}\|_F^2 / \|Z^{(t+1)}\|_F^2 < \varepsilon$  or  $t \geq T$ 
11 return  $\hat{\lambda}_i^{(t)}, \hat{U}_i^{(t)}, \hat{V}_i^{(t)}$  for  $i = 1, \dots, k$ 
```

---



**Figure 2.** Comparison of memory and time complexity of `CitationImpute` with some existing options for low-rank matrix completion when applied to our simulation test-bed model (Definition 4.2). Some run-times are truncated at 20 sec in the right panel. Each estimator is iterative; we compare time and memory use for five iterations. Existing implementations for both sparse and dense data representations are memory bound and do not scale to networks with more than several thousand nodes. Our implementation, although un-optimized, uses less memory and is faster for large networks.

some minor notation changes and the introduction of a maximum number of iterations  $T$ .

The initializer is given by running Algorithm 2, which we defer to the appendix. If we compute  $Z^{(1)}$  by taking a rank  $k$  singular value decomposition of  $P_{\Omega}(A)$  and fix  $\alpha^{(t)} = \lambda$  for all  $t$  (note that  $\tilde{\alpha}^{(t)}$  is the data adaptive thresholding parameter), `AdaptiveImpute` reduces to `softImpute`. This implies that a naive implementation of `AdaptiveImpute` inherits the per-iteration time complexity of `softImpute`, which is  $\mathcal{O}(|\Omega|k + nk^2)$ , plus the cost of evaluating  $\tilde{\alpha}^{(t)}$ .

### 3.3.1. Feasible Implementation

In practice, the runtime for each iteration of `AdaptiveImpute` and `softImpute` is dominated by the singular value decomposition, which is computed using an algorithm such as the implicitly restarted Lanczos bidiagonalization algorithm. The time complexity of this decomposition depends fundamentally on an underlying bidiagonalization subroutine (Algorithm 1 in the appendix), and the time complexity of the bidiagonalization subroutine in turn depends on cost of left and right matrix-multiplication of  $\tilde{A}^{(t)}$  with an appropriately sized vector (Baglama and Reichel 2005).

When  $A$  is sparsely observed,  $\tilde{A}^{(t)}$  can be expressed as a sparse matrix plus a low-rank matrix

$$\tilde{A}^{(t)} = \underbrace{P_{\Omega}(A - Z^{(t)})}_{\text{sparse}} + \underbrace{Z^{(t)}}_{\text{low-rank}}, \quad \text{sparsely observed setting}$$

and matrix-vector multiplication has time complexity  $\mathcal{O}(|\Omega|k)$  for the sparse part and  $\mathcal{O}(nk^2)$  for the low-rank part. In the citation setting, naively reusing this decomposition in the bidiagonalization subroutine is inefficient since  $|\Omega| \approx n^2/2$ .

However, a similar trick can improve the time complexity of multiplication with  $\tilde{A}^{(t)}$ : we can drop observed zeroes from consideration if we partition  $\tilde{A}^{(t)}$  carefully. Since  $P_{\tilde{\Omega}}(A) = P_{\Omega}(A)$ , we can compute only on  $\tilde{\Omega}$ . Let  $U = \{(i, j) : i < j\}$  denote the indices of the upper triangle of  $A$  and  $L$  denote the indices of the observed elements of  $A$  on the lower triangle, such that  $\Omega = U \cup L$ . Then

$$\begin{aligned} \tilde{A}^{(t)} &= P_{\Omega}(A) + P_{\Omega}^{\perp}(Z^{(t)}) && \text{citation setting} \\ &= P_{\Omega}(A) - P_{\Omega}(Z^{(t)}) \\ &\quad + P_{\Omega}(Z^{(t)}) + P_{\Omega}^{\perp}(Z^{(t)}) \\ &= P_{\tilde{\Omega}}(A) - P_{\Omega}(Z^{(t)}) + Z^{(t)} \\ &= \underbrace{P_{\tilde{\Omega}}(A)}_{\text{sparse}} - \underbrace{P_L(Z^{(t)})}_{\text{sparse}} \\ &\quad - \underbrace{P_U(Z^{(t)})}_{\text{low-rank until projection}} + \underbrace{Z^{(t)}}_{\text{low-rank}}. \end{aligned}$$

Efficient implementation strategies for matrix-vector multiplications with the sparse and low-rank terms are well known. This leaves the  $P_U(Z^{(t)})$  term, which is low-rank until it is projected onto the upper triangle. There one can use the same implementation strategy as for the low-rank component, but summing over fewer indices.

**Proposition 3.1.** Let  $Z^{(t)} \in \mathbb{R}^{n \times n}$  be a rank  $k$  matrix with singular value decomposition  $Z^{(t)} = UDV^T$  and let  $x \in \mathbb{R}^n$ . Then

$$\left[ P_U(Z^{(t)}) x \right]_i = \langle U_i, \tilde{W}_i \rangle,$$

where  $\tilde{W}_{ki} = \sum_{j=i+1}^n W_{kj}$  and  $W_{\cdot j} = (DV^T)_{\cdot j} \cdot x_j$ .

We defer the proof to the Appendix. **Proposition 3.1** is a straightforward result that suggests a computational scheme for evaluating the term  $P_U(Z^{(t)})x$ . In particular, it suggests constructing  $W$ , then  $\tilde{W}$ , and then obtaining elements of  $P_U(Z^{(t)})x$  element by element. This procedure requires  $\mathcal{O}(nk^2)$  flops as opposed to the  $\mathcal{O}(n^2k)$  flops of a naive implementation. The left-multiplication case is analogous.

The last requirement to implement `AdaptiveImpute` is a similarly efficient calculation of  $\alpha^{(t)}$ .

**Proposition 3.2.** Let  $\tilde{A}^{(t)}$ ,  $Z^{(t)}$  and  $\alpha^{(t)}$  be as defined in **Algorithm 1**. Recall that  $Z^{(t)}$  is a low-rank matrix of the form  $UDV^T$  with  $U, V \in \mathbb{R}^{n \times k}$  orthonormal and  $D \in \mathbb{R}^{k \times k}$  diagonal. Then

$$\begin{aligned} \alpha^{(t)} &= \frac{1}{n-k} \left[ \|P_{\tilde{\Omega}}(A)\|_F^2 + \|Z^{(t)}\|_F^2 - \|P_L(Z^{(t)})\|_F^2 \right. \\ &\quad \left. - \|P_U(Z^{(t)})\|_F^2 - \sum_{i=1}^k \lambda_i^2(\tilde{A}^{(t)}) \right]. \end{aligned}$$

Additionally, define  $U^{rq} \in \mathbb{R}^n$  and  $V^{rq\Delta} \in \mathbb{R}^n$  such that

$$\begin{aligned} U_i^{rq} &= U_{ir} U_{iq}, \quad \text{and} \\ V_i^{rq\Delta} &= \sum_{j=i+1}^n (DV)_{rj}^T (DV)_{qj}^T \quad \forall i = 1, \dots, n. \end{aligned}$$

Then

$$\|P_U(Z^{(t)})\|_F^2 = \sum_{r=1}^k \sum_{q=1}^k \langle U^{rq}, V^{rq\Delta} \rangle.$$

To understand the computational complexity of this expression we proceed term by term. First, consider the  $\sum_{i=1}^k \lambda_i^2(\tilde{A}^{(t)})$  term. Each iteration of `AdaptiveImpute` computes a truncated singular value decomposition of  $\tilde{A}^{(t)}$  of rank  $k$  before computing  $\alpha^{(t)}$ , so evaluating this term is a trivial  $\mathcal{O}(k)$  summation since  $\lambda_i(\tilde{A}^{(t)})$  is available for  $i = 1, \dots, k$ . Next, observe that  $\|P_{\tilde{\Omega}}(A)\|_F^2$  and  $\|P_L(Z^{(t)})\|_F^2$  are collectively  $\mathcal{O}(|\tilde{\Omega}|k)$ . This leaves the terms  $\|Z^{(t)}\|_F^2$  and  $\|P_U(Z^{(t)})\|_F^2$ , both of which require  $\mathcal{O}(nk^2)$  flops. As in **Proposition 3.1**, the idea is that evaluating  $\|P_U(Z^{(t)})\|_F^2$  is essentially the same as evaluating  $\|Z^{(t)}\|_F^2$ , modulo some care while indexing. The time complexity to compute  $\alpha^{(t)}$  is then  $\mathcal{O}(|\tilde{\Omega}|k + nk^2)$  flops. Using this scheme to evaluate  $\alpha^{(t)}$ , the overall time complexity of each iteration of `AdaptiveImpute` is  $\mathcal{O}(|\tilde{\Omega}|k + nk^2)$ .

### 3.4. Varimax Rotation

After obtaining an estimated singular value decomposition  $A \approx \widehat{U}\widehat{D}\widehat{V}^T$  from `AdaptiveImpute`, `CitationImpute` varimax rotates the estimates to obtain latent factors for each node.



Given an  $n \times k$  matrix orthonormal matrix  $U$ , varimax rotation finds a  $k \times k$  orthogonal matrix  $R$  that maximizes

$$v(R, U) = \sum_{\ell=1}^k \frac{1}{n} \sum_{i=1}^n \left( [UR]_{i\ell}^4 - \left( \frac{1}{n} \sum_{j=1}^n [UR]_{j\ell}^2 \right)^2 \right)$$

over the set of  $k \times k$  orthonormal matrices. In particular, it computes  $\hat{R}_U$  that maximizes  $v(\cdot, \hat{U})$  and  $\hat{R}_V$  that maximizes  $v(\cdot, \hat{V})$  where  $\hat{R}_U$  and  $\hat{R}_V$  are  $k \times k$  orthonormal matrices. Calculating these rotation matrices is a routine operation available in many statistical packages. After the rotation matrices  $\hat{R}_U$  and  $\hat{R}_V$  have been found, the latent factors are estimated as

$$\begin{aligned} \hat{Z} &= \sqrt{n} \hat{U} \hat{R}_U, & \hat{Y} &= \sqrt{n} \hat{V} \hat{R}_V, & \text{and} \\ \hat{B} &= \hat{R}_U^T \hat{D} \hat{R}_V / n. \end{aligned} \quad (3.1)$$

Rohe and Zeng (2023) show that, for  $\hat{U}, \hat{D}, \hat{V}$  obtained from the singular value decomposition in the fully observed case, varimax rotated estimates  $\hat{Z}, \hat{B}$ , and  $\hat{Y}$  are consistent for population terms  $Z, B$ , and  $Y$ .

#### 4. Simulation Study

To assess the performance of `CitationImpute`, we perform a simulation study using a co-stochastic blockmodel, a sub-model of the co-factor model. In the simulation study, `CitationImpute` recovers singular subspaces of  $\mathcal{A}$  and the latent factors  $Z$  and  $Y$  at the same rate as an oracle estimator that has access to all of  $A$ . The simulations also show that naive imputation of missing data leads to inconsistent estimates.

For the simulations, we use a Poisson degree-corrected stochastic co-blockmodel subject to lower triangular missingness.

**Definition 4.1 (Degree-corrected stochastic co-blockmodel).** The *degree-corrected stochastic co-blockmodel* is random graph model on  $n$  nodes. Each node  $i$  is assigned an incoming community  $z(i) \in \{1, \dots, k\}$  and an outgoing community  $y(i) \in \{1, \dots, k\}$  according to parameters  $\pi^{\text{in}} \in [0, 1]^k$  and  $\pi^{\text{out}} \in [0, 1]^k$ , such that  $\mathbb{P}(z(i) = j) = \pi_j^{\text{in}}$  and  $\mathbb{P}(y(i) = j) = \pi_j^{\text{out}}$  for  $j \in \{1, \dots, k\}$ . Each node  $i$  is also assigned a propensity  $\theta_i^{\text{out}} \in \mathbb{R}_+$  to send edges, and a propensity  $\theta_i^{\text{in}} \in \mathbb{R}_+$  to receive edges. Conditional on community memberships and edge formation propensities, integer-valued edges occur independently according to a Poisson distribution with expectation.

$$\mathbb{E}(A_{ij} | z(i), y(j)) = \theta_i^{\text{out}} B_{z(i), y(j)} \theta_j^{\text{in}}.$$

where  $B \in [0, 1]^{k \times k}$  is a rank  $k$  mixing matrix denoting propensities of edge formation between communities.  $B$  can be rescaled by a constant to enforce that the expected density of edges in the network is  $\rho$ .

The idea behind the simulation model is to mimic the behavior we expect in citation networks, where papers in a given field will primarily cite papers from that same field (strong diagonal structure in  $B$ ), but will intermittently cite papers from other fields (some active elements of  $B$  on the off-diagonal). This is motivated by the observation that the topics that Tibshirani (1996) cites and the topics that cite Tibshirani (1996) are distinct.

**Definition 4.2 (simulation model).** The simulation model is a degree-corrected stochastic co-blockmodel with  $n$  nodes,  $k$  co-communities, and expected density  $\rho = 0.15$ . Let  $\pi_j^{\text{in}} = \pi_j^{\text{out}} = 1/k$  for  $j = 1, \dots, k$ , such that the co-communities are balanced. Let  $\theta^{\text{in}}$  and  $\theta^{\text{out}}$  be generated by sampling  $n$  independent realizations from an exponential distribution with mean eight, and then adding one to each realization, inducing some degree-heterogeneity. The diagonal elements of  $B$  are set to  $B_{\text{within}} = 0.8$ .  $k$  elements of the off-diagonal to  $B_{\text{between}} = (B_{\text{within}}/3 - (k-2)B_{\text{inactive}})$  (in particular, the off-diagonal values in the first row of  $B$ , and the last element of the second column of  $B$ ). The remaining elements of the off diagonal to  $B_{\text{inactive}} = 0.01$ . This ensures that  $B$  is rank  $k$  and that there is strong assortative structure in the network. In the simulations, we use  $k \in \{3, 6, 9\}$ , with corresponding values of  $B_{\text{between}} = 0.257, 0.227, 0.197$ .

We compare the `CitationImpute` to an oracle estimator with access to the full data  $A$ , and also two imputation estimators. In total, we compare four estimators:

1. `CitationImpute`, with  $\ell_z = \ell_y = n/10$ ,
2. singular value decomposition applied after imputing all missing data as zeros (call this the *zero-imputed* estimator),
3. singular value decomposition applied after imputing all missing data by symmetrizing the observed data (call this the *symmetrized* estimator), and
4. oracle singular value decomposition applied to a fully observed similarity data (call this the *fully observed* estimator).

For the last three estimators, after estimating singular subspace, the singular vectors are varimax rotated according to (3.1) to obtain co-factor estimates.

To measure how well various estimators recover the singular subspaces of  $\mathcal{A}$ , we compute the  $\sin \Theta$  distance between the subspaces spanned by  $U$  and  $\hat{U}$  (Vu and Lei 2013; Bhatia 1997), for identified rows only. Given two orthonormal bases  $U \in \mathbb{R}^{n \times k}$  and  $\hat{U} \in \mathbb{R}^{n \times k}$ , the singular values  $\sigma_1, \dots, \sigma_k$  of  $U^T \hat{U}$  are the cosines of the principal angles  $\cos \theta_1, \dots, \cos \theta_k$  between the span of  $U$  and the span of  $\hat{U}$ . Define  $\sin \Theta(U, \hat{U})$  to be a diagonal matrix containing the sine of the principle angles of  $U^T \hat{U}$ . Then the  $\sin \Theta$  distance between the subspaces spanned by  $U$  and  $\hat{U}$  is given by

$$d(U, \hat{U}) = \|\sin \Theta(U, \hat{U})\|_F.$$

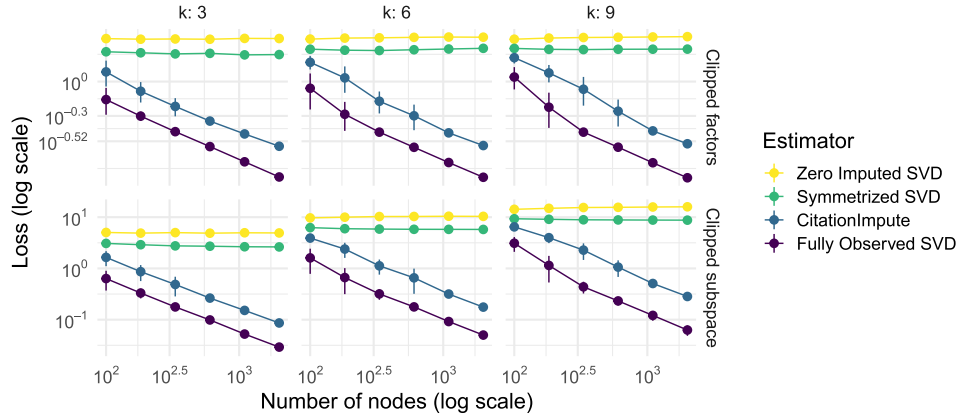
We aggregate error across identified rows of the estimates  $\hat{U}$  and  $\hat{V}$  and report a single metric

$$\mathcal{L}_{\text{subspace}}(U, \hat{U}, V, \hat{V}) = \|\sin \Theta(U, \hat{U})\|_F + \|\sin \Theta(V, \hat{V})\|_F.$$

To measure how well the estimators recover the latent factors  $Z$  and  $Y$ , we report root mean squared error on individual elements of identified rows  $\hat{Z}$  and  $\hat{Y}$ . Since varimax estimates  $\hat{Z}$  and  $\hat{Y}$  are only determined up to sign-flips and column reordering, this requires an alignment step to match  $\hat{Z}$  with  $Z$ , and  $\hat{Y}$  with  $Y$ . Let  $\mathcal{P}(k)$  be the set of  $k \times k$  orthogonal matrices whose entries  $P_{ij}$  are elements of  $\{-1, 0, 1\}$ . Define

$$P_Z = \arg \min_{P \in \mathcal{P}(k)} \|Z - \hat{Z}P\|_F \quad (4.1)$$

$$P_Y = \arg \min_{P \in \mathcal{P}(k)} \|Y - \hat{Y}P\|_F. \quad (4.2)$$



**Figure 3.** Average estimation error as a function of sample size, on log – log scale. The top row of panels visualizes estimation error of the factors  $Z$  and  $Y$ , excluding clipped factors. The bottom row of panels visualizes estimation error of the singular subspaces, again excluding clipped rows of  $U$  and  $V$ . Each column of panels represents a simulation model with a different number of latent communities. Within each panel, each line corresponds to the loss of a single estimator. Average loss plus and minus one standard deviation are shown as a dotplot; in most cases the standard deviations are too small to see.

We find  $P_Z$  and  $P_Y$  by using the Hungarian algorithm to match columns of the estimates  $\hat{Z}, \hat{Y}$  to the corresponding population values  $Z, Y$ . Then the elementwise factor root mean squared error is

$$\mathcal{L}_{\text{factor}}(Z, \hat{Z}, Y, \hat{Y}) = \sqrt{\frac{1}{nk} (\|Z - \hat{Z}P_Z\|_F^2 + \|Y - \hat{Y}P_Y\|_F^2)}.$$

To perform the simulation, we evaluate the subspace loss and the factor loss 200 times for every estimator, every  $k \in \{3, 6, 9\}$ , and every  $n \in \{100, 182, 331, 603, 1099, 2000\}$ . In Figure 3, we report the average subspace loss and the average factor loss for these combinations. Estimation error for `CitationImpute` decreases at approximately  $\sqrt{n}$ -rates, suggesting that `CitationImpute` is a consistent estimator of the singular subspaces of  $A$  and also of the latent factors  $Z$  and  $Y$ . The rate for `CitationImpute` parallels that of the oracle estimator with access to all of  $A$ , although it unsurprisingly advantageous to observe the full data.

In contrast, the symmetric imputation strategy and the zero-imputation strategies are not reliable ways to estimate singular subspaces or latent factors. Estimation error for both imputation strategies is constant as a function of  $n$ , suggesting that estimators based on naive imputation approaches are inconsistent. The symmetric imputation strategy is always better than treating the unobserved entries as zeroes, which makes sense as the model has some underlying symmetry. Some additional simulation results investigating the imputation estimators are available in Appendix B.

## 5. Analysis of the Statistics Literature

We next leveraged `CitationImpute` to analyze of the academic statistics literature.

### 5.1. Data

We used proprietary Web of Science data that we obtained through an institutional agreement with Clarivate Analytics. The complete Web of Science corpus contains hundreds of millions of documents, which amount to nearly a terabyte of

data. We considered only papers published in a subset of 125 journals focused on probability and statistics (see Appendix D for a list of the journals). The node-induced subgraph formed by considering only these papers and the citations between them had 281,883 nodes, 2,224,775 edges, and 24,051 weakly connected components (a weakly connected component is a subgraph where there is a path between every pair of nodes, ignoring the direction of edges). Most of the 24,051 weakly connected components were singletons. The largest weakly connected component contained 255,780 nodes and 2,222,363 edges. From this point onward, when we refer to the “citation network” or “citation graph” we are referring exclusively to this largest connected component. For each document we additionally knew the authors, publication date, and the abstract text, although some of this information was missing.

Papers in the citation network were published between 1898 and 2024. The number of citations received from other papers in the largest connected component (i.e., in-degree) ranged from 0 to 4759 and the number of citations sent to other papers in the largest connected component (i.e., out-degree) ranged from 0 to 603. There are several articles in the sample that cite hundreds of other papers; these articles are typically bibliographies or reviews. A small number of papers mutually cited each other.

### 5.2. Methods

First we constructed the partially observed adjacency matrix of the citation graph. We ordered nodes chronologically, and then clipped data using  $\ell_z = 100,000$  and  $\ell_y = 50,000$ . This amounted to discarding outgoing citations for papers published before 2004, and incoming citations for papers published after 2018. These clipping parameters were primarily selected on the basis of domain knowledge—we supposed that the topics in the modern statistics literature were present by 2004. We did not estimate incoming co-factors for papers published after 2018, because it can take several years to publish an academic paper, and we believed papers in 2018 were the latest papers that reasonably had the chance to be discovered, cited, and included in the Web of Science dataset.

We then ran `AdaptiveImpute` to obtain a low-rank decomposition  $A \approx \widehat{U}\widehat{D}\widehat{V}^T$ . Here we report the results for a rank  $k = 30$  decomposition. After computing a low-rank decomposition  $A \approx \widehat{U}\widehat{D}\widehat{V}^T$ , we performed varimax rotation of  $\widehat{U}$  and  $\widehat{V}$  to obtain a final low-rank decomposition  $A \approx \widehat{Z}\widehat{B}\widehat{Y}^T$ , as described in Section 3. The rows of  $\widehat{Z}$  and the rows of  $\widehat{Y}$  thus correspond to document-level latent co-factors (Rohe and Zeng 2023; Rohe, Qin, and Yu 2016). The rows of  $\widehat{Z}$  contained outgoing-citation factors, and the rows of  $\widehat{Y}$  contained incoming-citation factors. Both  $\widehat{Z}$  and  $\widehat{Y}$  were relatively sparse. To interpret the co-factors  $\widehat{Y}$  and  $\widehat{Z}$ , we took several approaches.

First, we found keywords most associated with each factor by examining the words in paper titles following the “best features” approach of Zhang, Chen, and Rohe (2021) and Chen (2021). We constructed a document-term matrix from the manuscript title. Letting  $X \in \mathbb{Z}^{255,780 \times 11298}$ ,  $X_{i\ell}$  indicates the number of times word  $\ell$  appears in manuscript title  $i$ . We restricted our analysis to words that appeared in at least five manuscript titles. Then, for each factor  $j$ , define the sets  $in(j) = \{i : \widehat{Y}_{ij} \geq 0\}$  and  $out(j) = \{i : \widehat{Y}_{ij} < 0\}$ . Then the importance of word  $\ell$  to factor  $j$  is

$$\text{b f f}(j, \ell) = \sqrt{\frac{\sum_{i \in in(j)} \widehat{Y}_{ij} X_{i\ell}}{\sum_{j \in in(i)} \widehat{Y}_{ij}}} - \sqrt{\frac{\sum_{i \in out(j)} X_{i\ell}}{|out(j)|}},$$

and in Tables 1 and 2 we report the six words most important to each factor. To complement this keyword analysis, we found the papers with the largest loadings for each dimension of  $\widehat{Y}$  and  $\widehat{Z}$ , which we refer to as hub papers (see Tables 2 and 3 in the Appendix).

### 5.3. Results

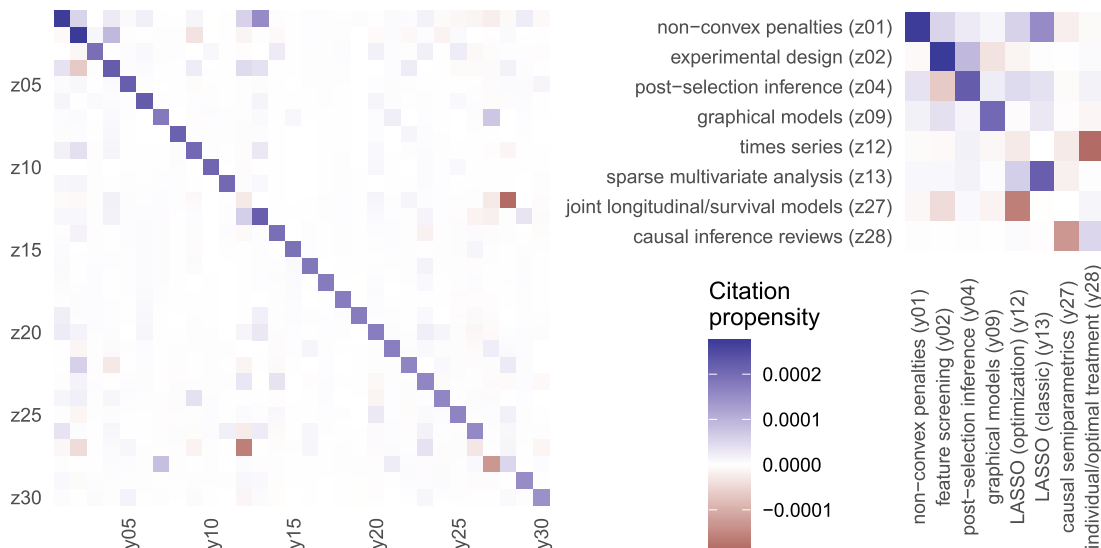
The incoming and outgoing co-factors were interpretable and associated with meaningful statistical sub-fields. We found co-factors corresponding to statistical sub-field such as GLMM(s), GEE, multiple testing, feature selection, post-selection inference, survival analysis, MCMC, causal inference, clinical trial

design, experimental design, functional data, multivariate analysis, graphical models, semiparametrics, kriging, model selection (both Bayesian and frequentist).

One particularly interesting feature of the  $\widehat{Y}$  co-factors was the presence of numerous incoming dimensions related to penalized regression. These factors covered the LASSO proper (y13), optimization methods for  $L_1$  penalization (y14), non-convex penalties (y01), post-selection inference (y04), feature screening (y02), graphical models (y09), and RIDGE regression (y29). Several other incoming  $\widehat{Y}$  co-factors were interesting because they corresponded to more niche statistical subfields. For example, we found incoming factors corresponding to empirical likelihood (y15), supervised dimension reduction (y11), and skew normals (y18). We suspect these co-factors emerged due to strong assortative structure in the sub-field: that is, a tendency to cite heavily within the factor while citing limited papers outside the factor. The tendency for spectral methods to find assortative clusters is widely known within spectral clustering literature, and it makes sense that they would pick up smaller but self-contained topics.

Most of the incoming co-factors  $\widehat{Y}$  correspond closely with an outgoing co-factor  $\widehat{Z}$  on the same topic. For instance, there is an incoming survival analysis co-factor (y05) and also an outgoing survival analysis co-factor (z05). The hubs for the incoming co-factor are highly cited methods papers such as Cox (1972) and Andersen and Gill (1982). The hubs for the outgoing co-factor are review papers that cite many of these works while receiving few citations themselves, such as Guo and Zeng (2014) and Kalbfleisch and Schaubel (2023). To investigate correspondences between  $\widehat{Y}$  and  $\widehat{Z}$  factors, we plotted the mixing matrix  $\widehat{B}$  in the left panel of Figure 4.

We found several  $\widehat{Z}$  factors that did not correspond closely with any incoming  $\widehat{Y}$  factor. For instance, z12 is a co-factor describing propensity to cite papers on times series analysis, while y12 is a co-factor related to the LASSO and optimization. Similarly, z27, a co-factor about joint longitudinal models, and z28, a causal inference co-factor, did not exhibit topical corre-



**Figure 4.** Left: The varimax estimate  $\widehat{B}$ . Each entry  $\widehat{B}_{ij}$  denotes the estimated citation propensity from papers loading on  $i$ th outgoing co-factor  $Z_i$  to the  $j$ th incoming co-factor  $Y_j$ . Right: A labeled sub-matrix of  $\widehat{B}$  considering the co-factors exhibiting off-diagonal structure.

spondence with y27, on causal semiparametrics, and y28, on individualized treatment rules. We visualized the relationships between the unmatched factors in the right panel of Figure 4, where it is clear that some co-factors are not in one-to-one correspondence with one another.

One question was how to interpret co-factors exhibiting one-to-one incoming-to-outgoing correspondence. For example, what was the difference between the outgoing survival analysis factor (z10) and the incoming survival analysis factor (y10)? To answer this question, we looked at the hub papers for each co-factor. For the survival analysis factor, for example, the top incoming hub was Cox (1972), which introduced the proportional hazards model, and the top outgoing hub was Guo and Zeng (2014), a survey of semiparametric models in survival analysis. Incoming  $\hat{Y}$  hub papers were typically highly cited, important papers in each sub-field. In contrast, the outgoing  $\hat{Z}$  hub papers were typically review articles, retrospectives, tutorials, and papers with good literature reviews that summarized the past literature. Put differently, statistical papers tended to either: (a) perform important synthesis of past work but be cited very little, (b) cite a limited number of papers while receiving many citations, or (c) cite and be cited very little.

This distinct behavior from the  $\hat{Y}$  paper hubs and  $\hat{Z}$  paper hubs is evidence of co-factor structure in the statistics literature, and more broadly, evidence that papers do indeed cite and get cited in fundamentally different ways.

#### 5.4. Sensitivity to Choice of Rank and Clipping Parameters

We repeated our analysis for  $k \in \{5, 10, 20, 30, 40\}$ , holding  $\ell_z = 100,000$  and  $\ell_y = 50,000$  fixed. Factor keywords, hubs, and mixing matrices for these analyses can be found in the supplemental material. We obtained qualitatively consistent results across values of  $k$ , and found that the co-factors can be coherently interpreted at all values of  $k$  that we explored. In practice, increasing  $k$  revealed additional, finer-grain factor structure. We chose to analyze  $k = 30$  co-factors because those co-factors revealed rich structure in the statistics literature while remaining interpretable and digestible.

We additionally explored  $\ell_z = \ell_y \in \{1, 25,000, 50,000, 70,000\}$ , holding  $k = 30$  constant. Factors keywords, hubs, and mixing matrices for these analyses can also be found in the supplemental

material. We found that the  $Y$  keywords, the  $Y$  factor hubs, and the  $Z$  factor hubs remained fairly stable across choices of  $\ell_z$  and  $\ell_y$ . However,  $Z$  keywords and factor identities were more varied, and mixing matrices  $\hat{B}$  also exhibited substantial variation. For smaller clipping parameters,  $\hat{B}$  exhibited substantial off-diagonal structure. As the clipping parameters increased,  $\hat{B}$  became more and more diagonal. The results for  $\ell_z = 100,000$  and  $\ell_y = 50,000$  had the least off-diagonal structure in  $\hat{B}$ .

Altogether, the sensitivity analysis for the clipping parameters indicated that  $\hat{B}$ , and to a lesser extent, the outgoing co-factors  $\hat{Z}$ , were somewhat unstable across hyperparameter values. Ultimately, our choice of  $\ell_z = 100,000$  and  $\ell_y = 50,000$  was based on domain knowledge: we assumed that it would take until 2004 for all outgoing co-topics to appear in the statistical literature, and that papers published after 2018 would not have the chance to be cited by papers from each incoming co-topic, due to the lengthy academic publication process. Regardless, since we do not definitely know how to select  $\ell_z$  and  $\ell_y$ , results should be treated as somewhat tentative.

#### 5.5. How the Past Would Cite the Future

One of the interesting features of our missing data framework is that it allows us to impute latent similarities from older documents to newer documents, or, with conceptual abuse, citations forward in time. In particular, if a paper  $i$  was published before paper  $j$ , we can estimate the latent similarity from paper  $i$  to paper  $j$  via the real-valued imputation  $\hat{A}_{ij} \approx \hat{Z}_i \hat{B} \hat{Y}_j^T$ . We suggest interpreting these imputed similarities as you would interpret probability estimates from a linear probability model; as in the linear probability model, we have no guarantee that  $\hat{A}_{ij} \in [0, 1]$ , such that  $\hat{A}_{ij}$  represents a valid probability of “citation.” However, we can still think of  $\hat{A}_{ij}$  as indicative of probability of citation, had citation been possible.

In particular, for each paper, we calculated all of these imputed similarities from prior papers. Summing over these imputations, we obtained an estimate of the number of times papers from the past would have cited papers from the future on the basis of topical similarity, were they so able. We computed these estimates for each of the papers in our citation network and report the 15 papers with the highest imputed in-degree

Table 3. Imputed incoming citations (identified edges only).

Title	Imputed	Cited by
On asymptotically optimal confidence regions and tests for high-dimensional models (2014)	1632	360
Confidence intervals for low dimensional parameters in high dimensional linear models (2014)	1564	350
Sure independence screening for ultrahigh dimensional feature space (2008)	1387	905
Estimating individualized treatment rules using outcome weighted learning (2012)	1215	280
Regularization paths for generalized linear models via coordinate descent (2010)	1135	1124
Feature screening via distance correlation learning (2012)	1094	327
A robust method for estimating optimal treatment regimes (2012)	1014	210
Model-free feature screening for ultrahigh-dimensional data (2011)	871	246
Sure independence screening in generalized linear models with np-dimensionality (2010)	847	305
Double/debiased machine learning for treatment and structural parameters (2018)	831	222
Nonparametric independence screening in sparse ultra-high-dimensional additive models (2011)	812	262
Exact post-selection inference, with application to the lasso (2016)	757	188
Performance guarantees for individualized treatment rules (2011)	753	219
Simultaneous analysis of lasso and dantzig selector (2009)	728	617
Sparse inverse covariance estimation with the graphical lasso (2008)	717	754



**Table 4.** Imputed outgoing citations (identified edges only).

Title	Imputed	Cites
Bayesian statistics in medicine: a 25 year review (2006)	754	511
Joint modeling of longitudinal and time-to-event data: an overview (2004)	172	36
Joint longitudinal-survival-cure models and their application to prostate cancer (2004)	165	34
Methodological issues with adaptation of clinical trial design (2006)	155	41
Adaptive statistical analysis following sample size modification based on interim review of effect size (2005)	143	26
Semiparametric regression during 2003-2007 (2009)	137	219
A 25-year review of sequential methodology in clinical studies (2007)	135	85
Group sequential and adaptive designs - a review of basic concepts and points of discussion (2008)	134	76
Maximum likelihood estimation in semiparametric regression models with censored data (2007)	131	53
Adaptive seamless designs: selection and prospective testing of hypotheses (2007)	128	61
A regulatory view on adaptive/flexible clinical trial design (2006)	128	30
An overview of statistical approaches for adaptive designs and design modifications (2006)	124	35
An investigation of two-stage tests (2006)	122	30
Efficient group sequential designs when there are several effect sizes under consideration (2006)	122	28
Joint modeling of longitudinal and survival data via a common frailty (2004)	118	21

in Table 3 and the 15 papers with highest imputed out-degree in Table 4. Most of the papers with high imputed in-degree are related to feature screening, the graphical LASSO, or some form of high dimensional regression. Most of the papers with high imputed out-degree are review articles published in *Biometrika*.

## 6. Discussion

We proposed a new method to co-factor documents in citation networks. The method is motivated by the observation that factors should be based on similarity measurements, and citations are only partially observed similarity measurements. Factoring a partially observed network complicated standard spectral clustering procedures and required use of matrix completion methods to estimate singular subspaces of the graph adjacency matrix. Here we found computational difficulties due to the precise observation pattern of citation data, which we resolved via a careful new implementation of the `AdaptiveImpute` algorithm. Because of dependence in the observation mechanism in the citation setting, existing theoretical results for `AdaptiveImpute`, and nuclear norm minimization more generally, were not applicable, and we validated our approach to matrix completion via a simulation study.

Our work suggests several avenues for methodological and theoretical exploration. Methodologically, it may be interesting to propose computationally efficient estimation procedures for other matrix completion methods in the upper triangular observation setting, or more generally in settings where sparse data is densely observed. Methods designed for independent but general sampling distributions, such as weighted nuclear norm minimization, may perform particularly well in the citation setting. Alternatively, further computational improvements would allow for larger scale bibliometric exploration of scientific citation networks. Current bibliometric databases contain hundreds of millions of papers and billions of references, more data than our method can handle. While our analysis of the statistics literature is one of the most extensive to date, incorporating additional papers could illuminate the relationships between statistical methodology and scientific practice at large. Another open question is how to extend our approach to the tensor, or multi-layer, citation network case, which would be appropriate for data like U.S. Court Opinions, where there are several distinct and explicitly labeled types of citation that documents may

use when referencing each other. Finally, it may be of significant practical use to develop a better theoretical understanding of how matrix completion methods perform in settings with dependent observation mechanisms.

## Supplementary Materials

**Appendix:** Contains pseudocode for important sub-routines, additional simulation results, proofs, the list of journals considered in the data analysis, tables of factor hubs, and a sensitivity analysis. ([appendix.pdf](#))

**fastadi R package:** A proof-of-concept implementation of the `AdaptiveImpute` estimator specialized to the citation setting. We have included the package source in the supplemental material, but it is also on CRAN and at <https://github.com/RoheLab/fastadi>. ([fastadi.zip](#))

**Replication package:** Code to reproduce the simulations and performance comparisons. Due to licensing agreements, we cannot publish the Web of Science data, but the replication package contains the code we used to analyze the Web of Science data. See `README.md` for details about contents. Also available online at <https://github.com/alexphayes/citation-cofactoring-replication/>. ([replication-package.zip](#))

## Acknowledgments

We thank Steve Meyer at UW-Madison Libraries for assistance with the Web of Science dataset; Keith Levin, Vivak Patel and several anonymous reviewers for feedback on this manuscript; Yunyi Shen for a code contribution; and Alexander Tahk, Ben Bolker, Mark Padgham, Noam Ross, Max Kuhn, Dan Simpson, Sam Power, Patrick Girardet, and Cannon Lewis for generative discussions throughout the course of the project.

## Disclosure Statement

The authors report there are no competing interests to declare.

## Funding

This project was supported by the NSF under grants DMS-1916378, DMS-1612456, and DMS-2023239, and by the ARO under grant W911NF-15-1-0423.

## ORCID

Alex Hayes  <http://orcid.org/0000-0002-4985-5160>

## References

- Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008), “Mixed Membership Stochastic Blockmodels,” *Journal of Machine Learning Research*, 9, 1981–2014. [3]



- Andersen, P. K., and Gill, R. D. (1982), "Cox's Regression Model for Counting Processes: A Large Sample Study," *The Annals of Statistics*, 10, 1100–1120. [10]
- Athreya, A., Fishkind, D. E., Tang, M., Priebe, C. E., Park, Y., Vogelstein, J. T., Levin, K., Lyzinski, V., Qin, Y., and Sussman, D. L. (2018), "Statistical Inference on Random Dot Product Graphs: A Survey," *Journal of Machine Learning Research*, 18, 1–92. [5]
- Athreya, A., Priebe, C. E., Tang, M., Lyzinski, V., Marchette, D. J., and Sussman, D. L. (2015), "A Limit Theorem for Scaled Eigenvectors of Random Dot Product Graphs," *Sankhya A: The Indian Journal of Statistics*, 78, 1–18. [5]
- Baglama, J., and Reichel, L. (2005), "Augmented Implicitly Restarted Lanczos Bidiagonalization Methods," *SIAM Journal on Scientific Computing*, 27, 19–42. [7]
- Bhatia, R. (1997), *Matrix Analysis*, New York: Springer. [8]
- Bhojanapalli, S., and Jain, P. (2014), "Universal Matrix Completion," in *Proceedings of the 31st International Conference on Machine Learning*. [3]
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022. [3]
- Chanpuriya, S., Tsourakakis, C. E., Musco, C., and Sotiropoulos, K. (2020), "Node Embeddings and Exact Low-Rank Representations of Complex Networks," in *34th Conference on Neural Information Processing Systems*, Vancouver, Canada. [3]
- Chen, F. (2021), "Spectral Methods for Social Media Data Analysis," Ph. D. thesis, University of Wisconsin-Madison. [10]
- Chen, Y., Jalali, A., Sanghavi, S., and Xu, H. (2014), "Clustering Partially Observed Graphs via Convex Optimization," *Journal of Machine Learning Research*, 15, 2213–2238. [3]
- Cho, J., Kim, D., and Rohe, K. (2019), "Intelligent Initialization and Adaptive Thresholding for Iterative Matrix Completion: Some Statistical and Algorithmic Theory for Adaptive-Impute," *Journal of Computational and Graphical Statistics*, 28, 323–333. [3,5,6]
- Choi, D. (2017), "Co-Clustering of Nonsmooth Graphons," *The Annals of Statistics*, 45, 1488–1515. [3]
- Choi, D., and Wolfe, P. J. (2014), "Co-Clustering Separately Exchangeable Network Data," *The Annals of Statistics*, 42, 29–63. [3]
- Cox, D. R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society, Series B*, 34, 187–202. [10,11]
- Cui, Z., Zhang, D., Wang, K., Zhang, H., Li, N., and Zuo, W. (2015), "Weighted Nuclear Norm Minimization Based Tongue Specular Reflection Removal," *Mathematical Problems in Engineering*, 2015, 1–15. [3]
- Drineas, P., and Mahoney, M. W. (2005), "On the Nystrom Method for Approximating a Gram Matrix for Improved Kernel-Based Learning," *Journal of Machine Learning Research*, 6, 2153–2175. [5]
- Foucart, S., Needell, D., Pathak, R., Plan, Y., and Wootters, M. (2021), "Weighted Matrix Completion From Non-Random, Non-Uniform Sampling Patterns," *IEEE Transactions on Information Theory*, 67, 1264–1290. [3]
- Gerlach, M., Peixoto, T. P., and Altmann, E. G. (2018), "A Network Approach to Topic Models," *Science Advances*, 4, 1–11. [3]
- Gittens, A., and Mahoney, M. W. (2016), "Revisiting the Nystrom Method for Improved Large-scale Machine Learning," *Journal of Machine Learning Research*, 17, 1–65. [5]
- Gu, S., Xie, Q., Meng, D., Zuo, W., Feng, X., and Zhang, L. (2017), "Weighted Nuclear Norm Minimization and Its Applications to Low Level Vision," *International Journal of Computer Vision*, 121, 183–208. [3]
- Gu, S., Zhang, L., Zuo, W., and Feng, X. (2014), "Weighted Nuclear Norm Minimization with Application to Image Denoising," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, pp. 2862–2869, IEEE. [3]
- Guo, S., and Zeng, D. (2014), "An Overview of Semiparametric Models in Survival Analysis," *Journal of Statistical Planning and Inference*, 151–152, 1–16. [10,11]
- Hajek, B., and Sankagiri, S. (2019), "Community Recovery in a Preferential Attachment Graph," *IEEE Transactions on Information Theory*, 65, 6853–6874. [3]
- Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R. (2015), "Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares," *Journal of Machine Learning Research*, 16, 3367–3402. [6]
- Hoerl, A. E., and Kennard, R. W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55–67. [4]
- Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983), "Stochastic Blockmodels: First Steps," *Social Networks*, 5, 109–137. [3]
- Hosono, K., Ono, S., and Miyata, T. (2016), "Weighted Tensor Nuclear Norm Minimization for Color Image Denoising," in *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, USA, pp. 3081–3085, IEEE. [3]
- Ji, P., and Jin, J. (2016), "Coauthorship and Citation Networks for Statisticians," *The Annals of Applied Statistics*, 10, 1779–1812. [1,5]
- Ji, P., Jin, J., Ke, Z. T., and Li, W. (2022), "Co-Citation and Co-authorship Networks of Statisticians," *Journal of Business & Economic Statistics*, 40, 469–485. [1]
- Jin, J. (2015), "Fast Community Detection by SCORE," *The Annals of Statistics*, 43, 57–89. [5]
- Jin, J., Ke, Z. T., and Luo, S. (2024), "Mixed Membership Estimation for Social Networks," *Journal of Econometrics*, 239, 105369. [3]
- Kalbfleisch, J. D., and Schaubel, D. E. (2023), "Fifty Years of the Cox Model," *Annual Review of Statistics and Its Application*, 10, 1–23. [10]
- Karrer, B., and Newman, M. E. J. (2011), "Stochastic Blockmodels and Community Structure in Networks," *Physical Review E*, 83, 016107. [3]
- Kim, Y.-D., and Choi, S. (2013), "Variational Bayesian View of Weighted Trace Norm Regularization for Matrix Factorization," *IEEE Signal Processing Letters*, 20, 261–264. [3]
- Klopp, O. (2014), "Noisy Low-Rank Matrix Completion with General Sampling Distribution," *Bernoulli*, 20, 282–303. [3]
- Larremore, D. B., Clauset, A., and Jacobs, A. Z. (2014), "Efficiently Inferring Community Structure in Bipartite Networks," *Physical Review E*, 90, 012805. [3]
- Lei, J., and Rinaldo, A. (2015), "Consistency of Spectral Clustering in Stochastic Block Models," *The Annals of Statistics*, 43, 215–237. [5]
- Li, T., Levina, E., and Zhu, J. (2020), "Network Cross-Validation by Edge Sampling," *Biometrika*, 107, 257–276. [3]
- Lyzinski, V., Sussman, D. L., Tang, M., Athreya, A., and Priebe, C. E. (2014), "Perfect Clustering for Stochastic Blockmodel Graphs via Adjacency Spectral Embedding," *Electronic Journal of Statistics*, 8, 2905–2922. [3,5]
- Lyzinski, V., Tang, M., Athreya, A., Park, Y., and Priebe, C. E. (2017), "Community Detection and Classification in Hierarchical Stochastic Blockmodels," *IEEE Transactions on Network Science and Engineering*, 4, 13–26. [5]
- Mazumder, R., Hastie, T., and Tibshirani, R. (2010), "Spectral Regularization Algorithms for Learning Large Incomplete Matrices," *Journal of Machine Learning Research*, 11, 2287–2322. [3,5]
- Peixoto, T. P. (2018), "Reconstructing Networks with Unknown and Heterogeneous Errors," *Physical Review X*, 8, 041011. [3]
- Pollner, P., Palla, G., and Vicsek, T. (2006), "Preferential Attachment of Communities: The Same Principle, but a Higher Level," *Europhysics Letters (EPL)*, 73, 478–484. [3]
- Price, D. D. S. (1976), "A General Theory of Bibliometric and Other Cumulative Advantage Processes," *Journal of the American Society for Information Science*, 27, 292–306. [3]
- Priebe, C. E., Park, Y., Vogelstein, J. T., Conroy, J. M., Lyzinski, V., Tang, M., Athreya, A., Cape, J., and Bridgeford, E. (2019), "On a Two-Truths Phenomenon in Spectral Graph Clustering," *Proceedings of the National Academy of Sciences*, 116, 5995–6000. [5]
- Qing, H., and Wang, J. (2022), "Directed Mixed Membership Stochastic Blockmodel," arXiv:2101.02307. [3]
- Razaei, Z. S., Amini, A. A., and Li, J. J. (2019), "Matched Bipartite Block Model with Covariates," *Journal of Machine Learning Research*, 20, 1–44. [3]
- Rohe, K., Qin, T., and Yu, B. (2016), "Co-Clustering Directed Graphs to Discover Asymmetries and Directional Communities," *Proceedings of the National Academy of Sciences*, 113, 12679–12684. [3,5,10]
- Rohe, K., and Zeng, M. (2023), "Vintage Factor Analysis with Varimax Performs Statistical Inference," *Journal of the Royal Statistical Society, Series B*, 85, 1037–1060. [3,4,8,10]
- Selby, D. A. (2020), *Statistical Modelling of Citation Networks*, Research Influence and Journal Prestige, Ph. D. thesis, University of Warwick. [1]

- Shamir, O., and Shalev-Shwartz, S. (2014), “Matrix Completion with the Trace Norm: Learning, Bounding, and Transducing,” *Journal of Machine Learning Research*, 15, 3401–3423. [3]
- Stigler, S. M. (1994), “Citation Patterns in the Journals of Statistics and Probability,” *Statistical Science*, 9, 94–108. [1]
- Tibshirani, R. (1996), “Regression Shrinkage and Selection Via the Lasso,” *Journal of the Royal Statistical Society, Series B*, 58, 267–288. [3,4,8]
- Vinayak, R. K., Oymak, S., and Hassibi, B. (2014), “Graph Clustering With Missing Data : Convex Algorithms and Analysis,” in *Advances in Neural Information Processing Systems*. [3]
- von Luxburg, U. (2007), “A Tutorial on Spectral Clustering,” *Statistics and Computing*, 17, 395–416. [5]
- Vu, V. Q., and Lei, J. (2013), “Minimax Sparse Principal Subspace Estimation in High Dimensions,” *The Annals of Statistics*, 41, 2905–2947. [8]
- Yang, M., Li, Y., and Wang, J. (2022), “Feature and Nuclear Norm Minimization for Matrix Completion,” *IEEE Transactions on Knowledge and Data Engineering*, 34, 2190–2199. [3]
- Yen, T.-C., and Larremore, D. B. (2020), “Community Detection in Bipartite Networks with Stochastic Block Models,” *Physical Review E*, 102, 032309. [3]
- Zhang, X., and Ng, M. K. (2019), “A Corrected Tensor Nuclear Norm Minimization Method for Noisy Low-Rank Tensor Completion,” *SIAM Journal on Imaging Sciences*, 12, 1231–1273. [3]
- Zhang, Y., Chen, F., and Rohe, K. (2021), “Social Media Public Opinion as Flocks in a Murmuration: Conceptualizing and Measuring Opinion Expression on Social Media,” *Journal of Computer-Mediated Communication*, 27, zmab021. [10]
- Zhao, J., Sun, M., Chen, F., and Chiu, P. (2022), “Understanding Missing Links in Bipartite Networks With MissBiN,” *IEEE Transactions on Visualization and Computer Graphics*, 28, 2457–2469. [3]
- Zhu, Z., Wang, T., and Samworth, R. J. (2022), “High-Dimensional Principal Component Analysis with Heterogeneous Missingness,” *Journal of the Royal Statistical Society, Series B*, 84, 2000–2031. [3]