Alex Hayes
PhD Candidate in Statistics

Research interests:

- (Social) networks
- Causal inference
- Peer effects
- Social processes in science
- Software design

Why the Observatory on Social Media?

Want to do more applied work

Past experience with social media:

- Large-scale analysis of Twitter data
- Collaborations with journalism scholars
- Two internships at Facebook

Curious about social dynamics in science

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2024, VOL. 00, NO. 0, 1–14
https://doi.org/10.1080/10618600.2024.2394464

Taylor & Francis
Taylor & Francis Group

Check for updates

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, GLMM/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 `CitationInput`. `CitationInput`e 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, `CitationInput`e 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, `CitationInput`e 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).

`CitationInput`e 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,

CONTACT Alex Hayes alex.hayes@wisc.edu Department of Statistics, University of Wisconsin-Madison, 1300 University Ave, Madison, WI 53706.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

My background is in statistical methods development for networks

Projects so far:

- PCA + neural nets to embed networks
- Embeddings for causal inference
- Contagion and peer effects
- Causal machine learning (a little)

Estimating network-mediated causal effects via principal components network regression

Alex Hayes ALEX.HAYES@WISC.EDU
Department of Statistics
University of Wisconsin–Madison
Madison, WI, USA

Mark M. Fredrickson MFREDRIC@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI, USA

Keith Levin KDLEVIN@WISC.EDU
Department of Statistics
University of Wisconsin–Madison
Madison, WI, USA

Abstract

We develop a method to decompose causal effects on a social network into an indirect effect mediated by the network, and a direct effect independent of the social network. To handle the complexity of network structures, we assume that latent social groups act as causal mediators. We develop principal components network regression models to differentiate the social effect from the non-social effect. Fitting the regression models is as simple as principal components analysis followed by ordinary least squares estimation. We prove asymptotic theory for regression coefficients from this procedure and show that it is widely applicable, allowing for a variety of distributions on the regression errors and network edges. We carefully characterize the counterfactual assumptions necessary to use the regression models for causal inference, and show that current approaches to causal network regression may result in over-control bias. The structure of our method is very general, so that it is applicable to many types of structured data beyond social networks, such as text, areal data, psychometrics, images and omics.

Keywords: causal mediation, latent mediators, network regression, principal components regression, random dot product graph, spectral embedding

1 Introduction

Recent years have seen a concerted effort to study causal effects on networks, motivated by striking claims about contagions in social networks (Christakis and Fowler, 2007). One of the key ideas to emerge from this push is the need to account for clustering in networks (Shalizi and Thomas, 2011). Sociologists have long known that people in social networks are mostly connected to other people like themselves, which is often expressed informally as "birds of a feather flock together," and more formally called "homophily". To identify and estimate causal effects in social settings, it is thus fundamental to model how social groups form in networks, as well any downstream effects of social group membership. This is challenging, as social groups in a network are typically unobserved.

I write a lot of code

- Background in open-source development
- Nine R packages on CRAN
- Thousands of lines of `Python` running daily at Facebook
- Use Github and build system everyday
- Some C++ experience (speeding up slow linear algebra)

JOSS
The Journal of Open Source Software

Welcome to the Tidyverse

Hadley Wickham[1], **Mara Averick**[1], **Jennifer Bryan**[1], **Winston Chang**[1], **Lucy D'Agostino McGowan**[2], **Romain François**[3], **Garrett Grolemund**[1], **Alex Hayes**[2,7], **Lionel Henry**[1], **Jim Hester**[1], **Max Kuhn**[1], **Thomas Lin Pedersen**[1], **Evan Miller**[4,5], **Stephan Milton Bache**[6], **Kirill Müller**[7], **Jeroen Ooms**[7,8], **David Robinson**[9], **Dana Paige Seidel**[10], **Vitalie Spinu**[4], **Kohske Takahashi**[11], **Davis Vaughan**[1], **Claus Wilke**[5], **Kara Woo**[1], **and Hiroaki Yutani**[12]

1 RStudio **2** cynkra **3** Redbubble **4** Erasmus University Rotterdam **5** Flatiron Health **6** Department of Integrative Biology, The University of Texas at Austin **7** Sage Bionetworks **8** Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health **9** Chukyo University, Japan **10** Department of Environmental Science, Policy, & Management, University of California, Berkeley **11** LINE Corporation **12** University of Wisconsin, Madison **13** None **14** University of California, Berkeley

DOI: 10.21105/joss.01686

Software
- Review
- Repository
- Archive

Editor: Karthik Ram
Reviewers:
- @desiccco-USGS
- @jeffreyhanson

Submitted: 09 August 2019
Published: 21 November 2019

Summary

At a high level, the tidyverse is a language for solving data science challenges with R code. Its primary goal is to facilitate a conversation between a human and a computer about data. Less abstractly, the tidyverse is a collection of R packages that share a high-level design philosophy and low-level grammar and data structures, so that learning one package makes it easier to learn the next.

The tidyverse encompasses the repeated tasks at the heart of every data science project: data import, tidying, manipulation, visualisation, and programming. We expect that almost every project will use multiple domain-specific packages outside of the tidyverse; our goal is to provide tooling for the most common challenges; not to solve every possible problem. Notably, the tidyverse doesn't include tools for statistical modelling or communication. These toolkits are critical for data science, but are so large that they merit separate treatment. The tidyverse package allows users to install all tidyverse packages with a single command.

There are a number of projects that are similar in scope to the tidyverse. The closest is perhaps Bioconductor (Gentleman et al., 2004; Huber et al., 2015), which provides an ecosystem of packages that support the analysis of high-throughput genomic data. The tidyverse has similar goals to R itself, but any comparison to the R Project (R Core Team, 2019) is fundamentally challenging as the tidyverse is written in R, and relies on R for its infrastructure; there is no tidyverse without R! That said, the biggest difference is in priorities: base R is highly focused on stability, whereas the tidyverse will make breaking changes in the search for better interfaces. Another closely related project is data.table (Dowle & Srinivasan, 2019).

My vision for the future

Continue ongoing research:

- Estimating peer effects in noisy networks
- Estimating peer effects in dynamic networks

Big picture goals:

- Develop new vein of research via postdoc
- Tenure-track faculty position

2024 **Alex Hayes** and Karl Rohe. Oct. 2024. **"Co-Factor Analysis of Citation Networks".** In: *Journal of Computational and Graphical Statistics*, pp. 1–14. DOI: 10.1080/10618600.2024.2394464.

Alex Hayes and Keith Levin. Oct. 2024. ***Peer Effects in the Linear-in-Means Model May Be Inestimable Even When Identified.*** arXiv: 2410.10772 [stat].

Alex Hayes, Mark M. Fredrickson, and Keith Levin. Sept. 2024. ***Estimating Network-Mediated Causal Effects via Principal Components Network Regression.*** arXiv: 2212.12041 [stat].

2019 Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, **Alex Hayes**, Lionel Henry, Jim Hester, Max Kuhn, Thomas Pedersen, Evan Miller, Stephan Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Nov. 2019. **"Welcome to the Tidyverse".** In: *Journal of Open Source Software* 4.43, p. 1686. DOI: 10.21105/joss.01686.