Spectral Methods for Social Media Data Analysis

by

Fan Chen

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2021

Date of final oral examination: 05/05/2021

The dissertation is approved by the following members of the Final Oral Committee:
Karl Rohe, Associate Professor, Statistics
Sündüz Keleş, Professor, Statistics, Biostatistics & Medical Informatics
Sébastien Roch, Professor, Mathematics
Michael A. Newton, Professor, Statistics, Biostatistics & Medical Informatics
Po-Ling Loh, Associate Professor, Statistics, Electrical & Computer Engineering

To my grandma, in loving memory.

ACKNOWLEDGMENTS

This dissertation comes a long way, and it would be impossible without the great help and support from my advisors, colleagues, and family.

It was Sündüz Keleş who made it possible for me to come to UW. I could never forget there was a quiet afternoon in my senior year of college when I was still desperately searching funding resources for graduate school. I had that life-changing video chat with Sündüz. It was a warm and casual conversation—we went over some past experiences and future opportunities in the lab, but it really was one of the most decisive moments in my life. I feel so fortunate that I joined UW Statistics, and I believe the choice was proven right. The transition from college to graduate school was unimaginable, but Sündüz patiently guided me through it. She taught me how to face whatever challenges ahead—I remember she told me just one week before the Qualifier exam: "just study 10 hours a day, and you will be fine." It magically eased my nerves. (I mean) What a simple yet upright attitude! Sündüz has always been so inspiring and such a role model to me; she just does not give up. In my second year, our RBP project got stuck in a modeling obstacle, and I was gradually losing my hope. What had been so enlightening to me was that she started to try out new modeling ideas and write codes herself. I deeply appreciate the kind of persistence she has, and it has always been an encouragement to me moving forward.

I could not be more grateful to have Karl Rohe as my thesis advisor, who is super careful about every little detail when it comes to research. In the meantime, he always gives out a pleasant and freestyle vibe in hosting students. The level of flexibility I had while working with him is precious to me. Karl can always come up with so many fun research topics that are neatly directed to the core. I really enjoyed the process of exploring the unknown and not having to worry about anything else. Once we discover something interesting, we write a paper! Such a smart advisor as Karl is, he also writes intriguing paragraphs (and tweets). I am so glad that I had the opportunity to learn about writing, research, and much more from him.

A big thank you to Michael Cammilleri who provided me with endless technical support in high-throughput computing. Mike is a musician and plays in band. Thank you for telling so many goodwill jokes and making my computation work literally fun and playful.

I would like to thank several faculty members at UW, including Sébastien Roch, Michael A. Newton, Po-Ling Loh, Garvesh Raskutti, Murray Clayton, Jun Zhu, Kris Sankaran, Keith Levin, Zhengjun Zhang, Grace Wahba, Yingyu Liang, Stephen J. Wright, Christina Kendziorski, Marvin Wickens, Judith Kimble, Heidi Dvinge, Chris Wells, Lew Friedland, and Dhavan Shah. Thank you for being so generous in sharing and for kindly helping me navigate through so many questions.

I want to also thank my colleagues who work in or collaborate with the Rohe Lab, including Yini Zhang, Josephine Lukito, Shuqi Yu, Yilin Zhang, Muzhe Zeng, Alex Hayes, E. Auden Krauska, Sijia Fang, Jitian Zhao, Hannah Wang, Jocelyn Ostrowski, Christopher Kardatzke, Caleb Stevens, Xiyu Yang, and Sara Stoudt. Thank you for the affable company along the journey and for all the deep/happy conversations that are related/unrelated to research. Those afternoon meetings in the Orton Park were truly beautiful and memorable. I hope we could do this more.

I am thankful to all the colleagues who work in or collaborate with the Keles group, including Brian Carrick, Ye Zheng, Kailei Chen, Rene Welch, Chengyang Dong, Cony Rojo, Daniel Conn, Pixu Shi, Sunyoung Shin, Peng Liu, Shan Lu, Siqi Shen, Jared Brown, Ziyue Wang, Zijian Ni, and Daniel Panyard. Thank you for sharing your brilliant ideas and for the pleasant time. I am sure that I will miss those productive Sunday (and the laughter) that we spent in the office.

I proudly thank my wonderful army of friends in Madison: Rui Chen, Tzu-Hsiang Hung, Jie Song, Yifan Mei, Kunling Huang, Yongfeng Wu, Lili Zheng, Yuchen Zhou, Muxuan Liang, Yifei Liu, Xiuyu Ma, Ting Ye, Lan Luo, Muhong Gao, Yuetian Luo, Linquan Ma, Yue Gao, Rungang Han, Fan Yang, Yicheng Hu, Yaqing Wu, Jun Li, Zhuohan Li, Zening Duan, Sijie Chen, Mengguo Jin, Jia-Rung Wu, Shao-Chun Wang, Chien-Fu Chen, Tun Lee Ng, Jacob Maronge, Michael Johnson, Behzad Aalipur, Jonathan Spaw, Jurijs Nazarovs, Michael

Liou, Srikanth Aravamuthan, Stephen Berg, Michael Kutzler, Kyler Westerfeldt, and Jesse Kroll. Thank you for making my life in Madison so pleasurable and filled with so many new adventures the whole time.

I must also express my thank words to my people back in China: Yilan He, Ruihong Huang, Jinyuan Jiang, Guoxian Xiao, Zhiyu Dong, Xuebo Huang, Ziming Fang, Junjie Chen, Shengnan Lin, Yiyun Ye, Chang Liu, Linhai Zhuo, Boyuan Huang, Xiang Li, Jiaxiang Jiang, Siyao Zheng, Xihan Li, Yueyue Song, Yina Zhang, Tianyi Xia, Yu Wang, Fan Wu, Hui Pang, Zhe Zhu, Kang Ying, Xutai Ma, Tao Jin, Yining An, Sheng Zhang, Tianran Liu, Renxuan Liu, Yandi Shen, Yuqing Ni, Qian Xu, Peng Zhang, Shun Cao, Zhe Wu, Xin Zhang, Yu Jin, Liujia He, Yingzhu Tao, and Qixiang Zhang. Thank you for always being there and for all the sweet and cheering conversations. I could never accomplish so much without your constant support.

Finally, I am deeply thankful to my dear parents, Lirong and Jianhua, and my dearest grandpa, Wenda, without whom I could not possibly start this adventure. Thank you for your endless, enormous, and loving support and for continuously offering advice. I would also like to thank my young cousins, Chen and Shenyi, who always embrave me to fight on. I wish you both very happy and accomplishing careers ahead. This dissertation is dedicated to my grandma who did not witness the final completion of it.

Fan Chen
May 2021

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Online social media have come to record an ever-increasing share of human communication and social interaction, making available vast quantities of data. This enables studies of individuals and society at large while presenting numerous computational and statistical challenges throughout the process. This dissertation introduces a set of spectral methods for social media data analysis and provides their statistical justifications and theoretical performance guarantees. These methods are motivated by the following questions that emerged in social media data analysis:

Q1: How to obtain a targeted sample of accounts?

Q2: How to select the number of communities?

Q3: How to find the underlying community structure?

Q4: How to collect some topic-specific documents?

For Q1, we study personalized PageRank (PPR), a popular technique that samples a small community from a massive network. Under the degree-corrected stochastic block model, we provide a simple and interpretable form for the PPR vector, highlighting its biases towards high degree nodes outside of the target block. We examine a simple adjustment based on node degrees and establish the consistency results for PPR clustering.

For Q2, we introduce a notion of cross-validated eigenvalues. Under a large class of random graph models, we provide a simple estimation procedure, a central limit theorem that gives a p-value for the statistical significance of each sample eigenvector, and a proof of consistency for estimating the number of communities in a network.

For Q3, we propose a new basis for sparse principal component analysis which can be applied on a graph adjacency matrix to identify the community structure. We provide evidence showing that for the same level of sparsity, the proposed method is more stable and can explain more variance compared to alternative methods.

For Q4, we study a local word embedding technique that measures both the frequency and exclusivity of words to a targeted topic. Under the popular latent Dirichlet allocation, we provide the statistical consistency for this embedding.

Finally, we introduce the “murmuration” framework which integrates the statistical methods and tracks the public political opinion expressions on Twitter, demonstrating a new way of imagining and measuring opinions on social media.

1 INTRODUCTION

1.1 Background

Spectral methods hold a central place in statistical data analysis. In a nutshell, spectral methods refer to a collection of algorithms built upon the eigenvectors (resp. singular vectors) and eigenvalues (resp. singular values) of some properly designed matrices of data. Classical spectral methods include principal components analysis (PCA), in which a low-dimensional subspace that explains most of the variance in the data is sought (Pearson, 1901; Hotelling, 1933); Fisher's discriminant analysis, which aims to determine a separating hyperplane for data classification (Fisher, 1936); and multidimensional scaling, used to realize metric embeddings of the data (Kruskal, 1964a,b). Recent developments of spectral methods have highlighted its strengths on handling large-scale, high-dimensional, and noisy data (Belabbas and Wolfe, 2009; Chen et al., 2020b), including community detection in networks (McSherry, 2001; Newman et al., 2006; Rohe et al., 2011; Abbe, 2017), sampling (Heckathorn and Cameron, 2017; Rohe et al., 2019; Chen et al., 2020a), clustering (Ng et al., 2001; Von Luxburg, 2007; Rohe and Zeng, 2020), dimensionality reduction (Belkin and Niyogi, 2003; Chen and Rohe, 2020), low-rank matrix estimation (Achlioptas and McSherry, 2007; Keshavan et al., 2010), among others. This dissertation covers some of the methodological developments that are largely motivated by social media data analysis.

As lots of human activities now taking place online (e.g., social media), digital media have come to record an ever-increasing share of human communication and social interaction, making available vast quantities of big data, in the forms of text, audio, and video (Gentzkow et al., 2019; Golder and Macy, 2014). Such rich and large-scale data offer an unprecedented opportunity for social scientists to study individuals and society at large unobtrusively (Salganik, 2019). For example, Twitter have emerged as one key battleground of public discourse, where people from different backgrounds actively comment on current events and public issues, and strive to exert influence (Conway et al., 2015; Tufekci, 2013; Kim et al., 2015). This leads to naturally occurring, temporally sensitive, and inherently social opinions (Anstead and O'Loughlin, 2014; Boyd, 2010; McGregor, 2019). Another key aspect of social media is the various homogeneous networks that it is embedded in, where like-minded individuals interact with each other and reinforce opinions (Colleoni et al., 2014; Conover et al., 2011; Barberá et al., 2015; Sunstein, 2018). More importantly, blended into individual day-to-day practice and the social world (Becker et al., 2010; Couldry, 2012; Tufekci and Wilson, 2012; McGregor, 2020), social media are an important public opinion domain in and of itself.

Existing studies that leverage social media data have primarily use natural language processing of text to identify patterns of communities of opinion expressions, such as sentiments or topics (Bollen et al., 2011; Tumasjan et al., 2011; Cody et al., 2015), and to compare the results with survey-based opinion polls (O'Connor et al., 2010). However, as the conception of public opinion is deeply intertwined with tools of opinion measurement (Herbst, 2001; Zaller, 1994), a blunt comparison between social media and survey-based opinion polls can be misleading. Furthermore, the text-centric approach fails to take full advantage of social media data to reveal the networks that opinions are embedded in and the social and conversational aspects of public opinion (Anstead and O'Loughlin, 2014), such as "Who talks about which events?" and "How are they talking about these events?" The "who" and "how" questions are especially important to address as a myriad of actors, ranging from social movement activists to propagandists, use social media to influence public opinion (Freelon et al., 2018; Tucker et al., 2018). Some studies have moved beyond text to account for characteristics of social media users, which can be detected with high accuracy (Pennacchiotti and Popescu, 2011; Kosinski et al., 2013). For example, Twitter accounts have been selected as "computational focus groups" based on hashtag use to map shared attention (Lin et al., 2014) or classified into hierarchical groups based on Twitter lists to trace opinion flow (Wu et al., 2011).

1.2 Motivating questions

We introduce a framework called "murmuration" for the study of public opinion on social media (Chapter 6). Given homophily driving friendship formation (McPherson et al., 2001; De Choudhury, 2011) and abundant empirical evidence for the effectiveness of social network structure (i.e., friendship relations) in predicting individual characteristics (Pennacchiotti and Popescu, 2011; Al Zamal et al., 2012; Grabowicz et al., 2012; Barberá et al., 2015; Pan et al., 2019), this framework uses social network structure to computationally identifies focus groups, which we call "flocks" (drawing on the idiom "birds of a feather flock together"). In this work, we consider the following series of motivating questions emerged in the design of the "murmuration."

Q1: How to obtain a targeted sample of accounts?

One of the key difficulties in studying social media is to gather subjects that are relevant to the scientific objective. A motivating example is to sample the Twitter friendship graph for accounts that report and discuss current political events. For this task, Chapter 2 provides statistical theory and intuition for Personalized PageRank (PPR), a popular technique that samples a small community from a massive network. We study a setting where the entire network is expensive to thoroughly obtain or maintain, but we can start from a seed node of interest and "crawl" the network to find other nodes through their connections. By crawling the graph in a designed way, the PPR vector can be approximated without querying the entire massive graph, making it an alternative to snowball sampling. Using the degree-corrected stochastic block model, we study whether the PPR vector can select nodes that belong to the same block as the seed node. We provide a simple and interpretable form for the PPR vector, highlighting its biases towards high degree nodes outside of the target block. We examine a simple adjustment based on node degrees and establish consistency results for PPR clustering that allows for directed graphs. These results are enabled by recent technical advances showing the element-wise convergence of eigenvectors. We illustrate the method with the massive Twitter friendship graph, which we crawl using the Twitter API. We find that (i) the adjusted and unadjusted PPR techniques are complementary approaches, where the adjustment makes the results particularly localized around the seed node and (ii) the bias adjustment greatly benefits from degree regularization.

Q2: How to select the number of communities?

In applied multivariate statistics, estimating the number of latent dimensions k is a fundamental and recurring problem. One common diagnostic is the scree plot, which plots the largest sample eigenvalues in decreasing order; the user searches this plot for a "gap" or "elbow" in the decaying eigenvalues. This diagnostic has two key limitations. First, the eigenvalues often have multiple gaps and elbows. Second, in statistical models with k true dimensions, bias differentially effects the k and $k+1$ sample eigenvalues, and this bias blurs any gap or elbow between them. A more general problem is that a useful theory and methodology must confront the possibility that only some of the leading k population eigenvectors are estimable. In this situation, the "correct" choice of k is the number of statistically useful dimensions. To confront these problems, Chapter 3 introduces a notion of cross-validated eigenvalues. Under a large class of random graph models, we provide (1) a simple estimation procedure, (2) a central limit theorem that gives a p-value for the statistical significance of each sample eigenvector, and (3) a proof of consistency. This approach can be used to estimate the number of statistically useful sample eigenvectors, naturally adapting to the complexity of the estimation task. In simulations and a data example, the proposed estimator compares favorably to alternative approaches in both computational and statistical performance.

Q3: How to find the underlying community structure?

Spectral clustering applies PCA on a graph adjacency matrix (with or without normalization) then runs the K-means algorithm on the loading coefficients to obtain clusters (or communities). We consider a simplification by solely using sparse PCA. Previous versions of sparse PCA have presumed that the eigen-basis (a $p \times k$ matrix) is approximately sparse. In Chapter 4, we propose a method that presumes the $p \times k$ matrix becomes approximately sparse after a $k \times k$ rotation. The simplest version of the algorithm initializes with the leading k principal components. Then, the principal components are rotated with an $k \times k$ orthogonal rotation to make them approximately sparse. Finally, soft-thresholding is applied to the rotated principal components. This approach differs from prior approaches because it uses an orthogonal rotation to approximate a sparse basis. One consequence of this rotation is that a sparse component need not to be a leading eigenvector, but rather a mixture of them. In this way, we propose a new (rotated) basis for sparse PCA, which can be applied to a graph adjacency matrix to identify the underlying community structure. In addition, our approach avoids "deflation" and multiple tuning parameters required for that. Our sparse PCA framework is versatile; for example, it extends naturally to a two-way analysis of a data matrix for simultaneous dimensionality reduction of rows and columns. We provide evidence showing that for the same level of sparsity, the proposed sparse PCA method is more stable and can explain more variance compared to alternative methods. Through three applications—sparse coding of images, analysis of transcriptome sequencing data, and large-scale clustering of social networks, we demonstrate the modern usefulness of sparse PCA in exploring multivariate data.

Q4: How to collect some topic-specific documents?

In "text-as-data" research, the critical first step is to assemble a set of documents relevant to the topic of interest from a large corpus. In Chapter 5, we address a setting where the entire corpus is computationally expensive and time consuming to thoroughly examine, but it is cheaper to query documents by keyword search, where the inclusion of a document is determined by whether it contains the keyword(s). To select keywords that can yield targeted documents with high precision and relevance, we propose to an approach called "WordPPR" that (1) constructs a word co-occurrence graph, in which two words form an edge if they co-occur in one document; (2) ranks words in the graph by the personalize PageRank (PPR) and its degree-adjusted version.

We study WordPPR under the latent Dirichlet allocation and provide statistical consistency results for the algorithm, highlighting that the PPR vector and its degree-adjusted version measure words' popularity and exclusivity to the targeted topic respectively. These results are enabled by recent technical advances showing the element-wise convergence of eigenvectors. We illustrate the method with simulation studies and the assemblage of tweets related to the "#MeToo" movement. We find that (i) WordPPR is robust to a wide choice of the teleportation constant and (ii) WordPPR benefits from using the documents collected from searching the initial keywords, rather than a random sample of documents.

1.3 The "murmuration"

As people tend to connect with like-minded others and express opinions in response to current events on social media, their expression is naturally occurring, temporally sensitive, and inherently social. Chapter 6 presents a new way of imagining and measuring opinions that emerge from various homogeneous networks on social media. Our framework for large-scale measurement of social media public opinion (1) samples targeted nodes from a large social graph, (2) identifies homogeneous, interactive, and stable networks of accounts, which we call "flocks," based on social network structure, and (3) measures and (4) presents their opinions with flocks. We apply this framework to Twitter and provide empirical evidence that the social network structure encoded in flocks accurately and consistently predicts opinion expression. We further show that this framework captures the intensity and temporal dynamics of opinion expression by various flocks as well as their opinion contestation in response to real-world events. Taken together, the results demonstrate one way that social media can be leveraged to examine the social dynamics of public opinion in the digital media environment.

2 TARGETED SAMPLING FROM MASSIVE BLOCK MODEL GRAPHS WITH PERSONALIZED PAGERANK

2.1 Introduction

Much of the literature on graph sampling has treated the entire graph, or all of the people in it, as the target population. However, in many settings, the target population is a small community in the massive graph. For example, a key difficulty in studying social media is to gather data that is sufficiently relevant for the scientific objective. A motivating example for this paper is to sample the Twitter friendship graph for accounts that report and discuss current political events.[1] This corresponds to sampling and identifying multiple different communities, each a potentially small part of the massive network. In such an application, the graph is useful for two primary reasons. First, via link tracing, we can find potential members of the target population. Second, the graph connections are informative for identifying community membership. Throughout, we presume that the sampling is initiated around a "seed node" that belongs to the target community of interest.

Personalized PageRank (PPR) can be thought of as an alternative to snowball sampling, a popular technique for gathering individuals close to the seed node. For some $d \geqslant 0$, snowball sampling gathers all individuals who are d friends away from the seed. This process has two competing flaws for our application which are addressed by PPR. First, snowball sampling fails to account for the density of common friendships. For example, perhaps i and j are both one friend removed from the seed, but i has 10 friends in common with the seed, while j only has 1 friend in common. It seems natural to suppose that i is closer than j to the seed. Hence, the metric for snowball sampling can be misleading. Second, the snowball sample size grows very quickly with d. For example, under the "six degrees of separation" phenomenon (Watts and Strogatz, 1998; Newman et al., 2006), snowballing gathers the entire graph if $d \geqslant 6$.

[1] See our website `http://murmuration.wisc.edu` which does this.

PPR gives a sample that is more localized around the seed node. The PPR vector is defined as the stationary distribution of what we call a *personalized random walk* (Page et al., 1998). At each step of the personalized random walk, the random walker returns to the seed node with probability α, called the teleportation constant, and with probability $1 - \alpha$, the random walker goes to an adjacent node that is chosen uniformly at random. Consider the stationary distribution of this process as giving the inclusion probability for a sample of size 1. This is the PPR vector. PPR naturally leads to a clustering algorithm, where the cluster is made up of the nodes with a large inclusion probability. To quickly approximate the PPR vector, Berkhin (2006) proposed an algorithm that only examines nodes with large inclusion probabilities (i.e. nodes near the seed). As such, PPR is particularly useful for its computational efficiency – the running time and the amount of data it requires is nearly linear in the size of the output cluster, which is typically much smaller than that of the entire graph. Due to the local nature of the algorithm, it can be used to study large graphs such as Twitter where the entire graph is not available, but where one can query to find the connections to any small set of nodes.

One way to conduct local clustering is by exploring and ranking the nearby nodes of a seed node. (Andersen and Lang, 2006; Andersen and Peres, 2009; Alamgir and von Luxburg, 2010; Gharan and Trevisan, 2012). Spielman and Teng (2004) pioneered local clustering by defining nearness as the landing probability of a random walk starting from the seed node. Their algorithm's guarantee was improved in follow-up work by Andersen et al. (2006) which proposed using an approximate PPR vector. Local algorithms can be applied recursively to solving more complicated problems such as graph partitions (k-way partitions) (Spielman and Teng, 1996; Karypis and Kumar, 1998), and has many fruitful applications (Jeh and Widom, 2003; Macropol et al., 2009; Liao et al., 2009; Gupta et al., 2013; Gleich, 2015), particularly when it comes to sampling and studying massive graphs.

Along with the widespread use of PPR, there has been recent work to study its statistical estimation properties under a statistical model with latent community structure. Beyond the scope of local clustering, Kloumann et al. (2017) showed that the PPR vector is asymptotically equivalent to optimal linear discriminant analysis under the stochastic block model (SBM) (Holland et al., 1983a), assuming a symmetry condition on the block structure. We add to this statistical understanding of PPR by providing a simple and more general representation for PPR vectors that allows for different block sizes, more than two blocks, degree heterogeneity, and directed edges. In order to understand the effects of heterogeneous node degrees, this paper uses the degree-corrected stochastic block model (DC-SBM) (Karrer and Newman, 2011b) and examines when the PPR clustering recovers nodes within the same block as the seed node (local cluster). Breaking the symmetry that is imposed by Kloumann et al. (2017) reveals additional insight. In particular, given a seed node in the first block, we show that PPR is likely to contain high degree nodes outside of that block. We study an adjustment that was previously proposed in Andersen et al. (2006). We show how this adjustment can correct for the bias. We illustrate these ideas with examples from the Twitter friendship graph.

An illustrative example in social media

Local clustering using PPR is particularly well suited to studying current political events on Twitter because (i) the accounts that discuss politics or current events are a small part of the entire Twitter graph, (ii) it is reasonable to believe that the accounts in our target population are well connected to one another in the Twitter friendship graph, and (iii) while the entire Twitter graph is not publicly available, the way that PPR (Algorithm 2.1 and 2.3) queries the graph matches the Twitter API protocol which is the primary mode of access for researchers.

While we do not suppose that the Twitter friendship graph is sampled from a DC-SBM, Twitter does have all of the heterogeneities that our results identify as important. The Twitter friendship graph is composed of users who can freely follow others but will not necessarily be followed back, or friended. Such asymmetry between following and friending forms a directed graph where follower count indicates status – some popular/high-status nodes command millions of followers while the majority of nodes are followed by far fewer.

The theoretical results in this paper suggest that such degree heterogeneities

Table 2.1: Top 15 handles by PPR clustering. Column names represent seed nodes, and the sampled nodes are ranked by PPR values, with teleportation constant $\alpha = 0.15$ uniformly.

	@CNN	**@BreitbartNews**	**@dailykos**
1	CNN Breaking News	Alex Marlow	Hillary Clinton
2	CNN International	AndrewBreitbart	Stephen Colbert
3	Wolf Blitzer	Big Hollywood	Rachel Maddow MSNBC
4	Anderson Cooper	Big Government	Jake Tapper
5	Christiane Amanpour	James O'Keefe	Joy Reid
6	Pope Francis	Sean Hannity	Chris Hayes
7	Dr. Sanjay Gupta	Raheem	Emma Gonzlez
8	CNNMoney	Joel B. Pollak	Markos Moulitsas
9	Jake Tapper	Ann Coulter	Maggie Haberman
10	Brian Stelter	Allum Bokhari	Sarah Silverman
11	CNN Newsroom	Ben Kew	Lin-Manuel Miranda
12	Dana Bash	Brandon Darby	Elizabeth Warren
13	CNN Politics	Noah Dulis	Jon Favreau
14	BBC Breaking News	Michelle Malkin	Michelle Obama
15	Brooke Baldwin	Nate Church	Bill Clinton

Through the PPR vector, the top 15 handles returned to each of the three seed nodes fit well with the characteristics of the seed nodes. They are popular/high-status handles either directly related to the seed nodes or align with their political leanings. This shows the effectiveness of clustering via the PPR vector. It also shows the PPR vector's preference for highly connected nodes.

will make the PPR vector biased for detecting block memberships (Theorem 2.4). We propose a way to adjust for this bias (Algorithm 2.2) and show that it is a consistent estimator (Corollary 2.7). Not surprisingly, this section demonstrates that PPR with and without the bias adjustment give fundamentally different results on the Twitter graph. However, depending on the application, the biases in the PPR vector might be advantageous. In this way, PPR with and without the bias adjustment are complementary, not competing, approaches.

To illustrate, Table 2.1 displays the top 15 handles ranked by the PPR vector (without adjustment) for three different seed nodes: @CNN, @BreitbartNews, and @dailykos, the Twitter accounts of three different types of media outlets that exhibit distinct political leanings (legacy broadcast news, online right-wing and online left-wing). For @CNN, all top 15 handles ranked by the PPR vector are its subsidiary accounts and its celebrity reporters and anchors (like Wolf Blitzer and Anderson Cooper), except for one account, Pope Francis, who

Table 2.2: Top 15 handles by adjusted PPR (with regularization) sampling. Column names represent seed nodes, and the sampled nodes are ranked by adjusted PPR values, with teleportation constant $\alpha = 0.15$ uniformly.

	@CNN	**@BreitbartNews**	**@dailykos**
1	PowerZ	Robert	Two Thanks
2	Elissa Weldon	Lee Peace	Catherine Daligga
3	Tess Eastment	Wynn Marlow	exmearden
4	Chris_Dawson	Logan Churchwell	Faith Gardner
5	carol kinstle	Peter Schweizer	Andrew Thornton
6	erinmclaughlin	Breitbart Sports	UnreasonableFridays
7	Taylor Ward	Jon Fleischman	DKos Top Comments
8	Jennifer Z. Deaton	Nate Church	2016 relitigator
9	Pam Benson	Daniel Nussbaum	Daily Kos
10	amy entelis	Noah Dulis	Walter Einenkel
11	Grace Bohnhoff	Jon David Kahn	Candelaria Vargas
12	kate lazarus	Breitbart California	Mara Schechter
13	Newstron	Ken Klukowski	Emi Feldman
14	Becky Brittain	pam key	The Soulful Negress
15	CNN Ballot Bowl	Auntie Hollywood	Kim Soffen

After adjustment, PPR returns a more localized cluster. Instead of the highly visible public faces of the three seed organizations, the individuals in this table serve a central role to the internal organization (e.g. editors and writers). Depending on the application, one might prefer the results in Table 2.1 or Table 2.2.

enjoys an extremely larger following. The top 15 handles for @BreitbartNews are a mixed bag of influential conservatives (like Sean Hannity and Ann Coulter) and Breitbart's editors/writers. However, the top 15 handles returned to @dailykos by the PPR vector are all famous liberal personalities not directly affiliated with Daily Kos, but one, its founder Markos Moulitsas. Those people range from democratic politicians to liberal media personalities and journalists, such as Hillary Clinton, Stephen Colbert, and Rachel Maddow. All the handles align with the characteristics of their respective media outlets, attesting to the clustering effectiveness. However, it is worth noting that the top handles ranked by the PPR vector tend to be popular handles with millions of followers. This shows that the PPR vector's preference for high in-degree nodes.

In contrast, for each of the three seeds, adjusted PPR finds accounts that are more central to the internal functioning of these organizations. Table 2.2 lists those accounts. The bias adjustment also greatly benefits from a degree regularization (Qin and Rohe, 2013). For @CNN, those handles include primarily its own staff/producers/journalists (like Elissa Weldon, Chris_Dawson, and Grace Bohnhoff), a freelance journalist (Tess Eastment). The pattern is similar for @BreitbartNews and @dailykos, their top 15 handles including their own journalists, editors as well as related writers/campaigners/activists. The general pattern is that the adjustment returns editors, journalists and staff working within each media outlet. As such, the adjustment is useful for identifying a more localized cluster.

Main contributions

The main contributions of the paper are (a) a simple and interpretable form for the PPR vector and (b) a statistical guarantee for clustering with the adjusted PPR vector.

(a) This paper reveals a simple two-stage form of the PPR vector under the population (expectation) DC-SBM. Consider the v-th element of the PPR vector as the probability of sampling node v in a sample of size 1 from the stationary distribution of the personalized random walk. This inclusion probability is akin to stratified sampling:

> *The inclusion probability for node v is the product of two separate probabilities. First, the probability that the personalized random walk samples any node in v's block. Second, the probability that the personalized random walk selects node v, conditional on sampling that block.*

Both of these probabilities have simple expressions. If there are K blocks in the graph, then the block-wise probability comes from the PPR vector of a graph with K vertices, with edge weights specified by the "block connection matrix" in the DC-SBM. The second probability is proportional to the degree of node v. In addition to the population results, Theorem 2.6 demonstrates that when the graph is random, the PPR vector concentrates around its population (expectation) under certain conditions.

(b) This paper identifies two sources of bias of using a PPR vector for local clustering under the DC-SBM – the ancillary effects of heterogeneous node degrees and block degrees. With this finding, the paper examines a simple bias adjustment that remedies the two biases simultaneously and suggests conditions when the adjusted PPR can be used to return the correct local cluster. In other words,

> *PPR clustering with the adjustment achieves the precise identification of the local cluster, provided the graph is sufficiently dense.*

These results establish statistical performance (consistency) of PPR clustering under the DC-SBM, in the sparse regime where the minimum expected degree grows logarithmically with the number of nodes in the network. Our results provide an element-wise perturbation bound for PPR vectors, that allows the number of clusters to grow with the size of graphs, and generalize to a directed graph setting as PageRank does.

The rest of the paper proceeds as follows. Section 2.2 formally introduces the PPR method and some of the known results. Section 2.2 also introduces the degree-corrected stochastic block model. Section 2.3 gives a population analysis of the PPR clustering under directed block model graphs. Section 2.4 provides concentration results for the PPR vector when the graph is random and provides a statistical guarantee on the PPR local clustering method. Section 2.5 presents several numerical results showing the effectiveness of the PPR clustering. Section 2.6 illustrates the PPR clustering through the massive Twitter friendship graph and demonstrates the benefits of a smoothing step in the PPR adjustment.

2.2 Preliminaries

Throughout this paper, let $G = (V, E)$ denotes an unweighted and connected graph, where E is the edge set and V is the set of vertices indexed by $1, ..., N$. When G is an undirected and unweighted graph, encode E into a binary *adjacency* matrix $A \in \{0, 1\}^{N \times N}$ with $A_{uv} = A_{vu} = 1$ if and only if edge (u, v)

appears in E. Define a diagonal matrix $D = \operatorname{diag}(d_1, ..., d_N)$ and the *graph transition* matrix P as follows:

$$d_u = \sum_{v \in V} A_{uv} \quad \text{and} \quad P = D^{-1}A.$$

When G is a directed graph, the adjacency matrix $A \in \{0,1\}^{N \times N}$ accordingly becomes asymmetric with $A_{uv} = 1$ if and only if edge $(u, v) \in E$, and the graph transition matrix is defined as

$$P = [D^{\text{out}}]^{-1}A,$$

where $D^{\text{out}} = \operatorname{diag}\left(d_1^{\text{out}}, ..., d_N^{\text{out}}\right)$ and $d_u^{\text{out}} = \sum_{v \in V} A_{uv}$ is the number of edges leaving from node u. In addition, define $D^{\text{in}} = \operatorname{diag}\left(d_1^{\text{in}}, ..., d_N^{\text{in}}\right)$ where $d_v^{\text{in}} = \sum_{u \in V} A_{uv}$ is the number of edges pointing to node v.

Personalized PageRank and the local clustering algorithm

The personalized PageRank (PPR) is an extension of Google's PageRank (Brin and Page, 1998; Haveliwala, 2003). To illustrate, consider a personalized random walk (or originally called "surfing") on the graph $G = (V, E)$ with a *seed node* $v_0 \in V$. At each step, the random walker either restarts from the seed node v_0 with probability α (called the *teleportation constant*) or continues the random walk from the current node to a neighbor uniformly at random. The *personalized PageRank vector* $p \in [0,1]^N$ is the stationary distribution of this process, thus the solution to the equation

$$p^T = \alpha\pi^T + (1 - \alpha)p^T P, \tag{2.1}$$

where P is the graph transition matrix, and π is the elementary unit vector in the direction of seed node v_0. Here p is a column vector normalized by a positive scalar such that its elements sum to 1, and without loss of generality, we set $v_0 = 1$ and thus $\pi = (1, 0, ..., 0)^T$.

In general, the *preference vector* π does not have to be an elementary unit vector, but any probability distribution on V. For example, when $\pi = (1/N, ..., 1/N)^T$, PPR is equivalent to ordinary PageRank. Moreover, the PPR vector is a linear function of the preference vector. That is, let $p(\pi_1)$ and $p(\pi_2)$ be two PPR vectors corresponding to two preference vectors π_1 and π_2 respectively. Then, for a new preference vector that is a convex combination of π_i, the resulting PPR vector is constructive of $p(\pi_i)$,

$$p(w_1\pi_1 + w_2\pi_2) = w_1 p(\pi_1) + w_2 p(\pi_2),$$

where $w_i \geqslant 0$ and $w_1 + w_2 = 1$. Define Π to be an $N \times N$ matrix with repeating rows π^T, and let $Q = \alpha\Pi + (1-\alpha)P$, then Q is the Markov transition matrix for the stochastic process and Equation (2.1) becomes $p^T = p^T Q$. Below are some useful properties of the PageRank vector (also see Haveliwala (2003); Jeh and Widom (2003) and Appendix A.1).

Proposition 2.1. *For any fixed* $\alpha \in (0, 1]$*, the PPR vector* p *is*

(a) *the left leading eigenvector of* Q*, associated with the simple eigenvalue 1; and*

(b) *the infinite sum of landing probability* $\{(P^s)^T \pi\}_{s=0}^{\infty}$ *with weights* $\phi = \{\alpha(1-\alpha)^s\}_{s=0}^{\infty}$*,*

$$p^T = \alpha \sum_{s=0}^{\infty} (1-\alpha)^s \pi^T P^s. \tag{2.2}$$

Berkhin (2006) gives an iterative algorithm based on Proposition 2.1 to approximate the PPR vector (that scales to large graphs); each update requires only neighborhood information of one visited vertex. A few lines of linear algebra show that the PPR vector is equivalent to the solution to the linear system

$$p^T = \alpha' \pi^T + (1-\alpha') p^T W,$$

where $W = (I + P)/2$ is the lazy graph transition matrix and $\alpha' = \alpha/(2-\alpha)$. Using this fact, Algorithm 2.1 approximates the PPR vector in running time of order $\mathcal{O}\left(\frac{1}{\varepsilon\alpha}\right)$, by reaching at most $\frac{2}{\varepsilon(1-\alpha)}$ vertices. The following proposition gives a guarantee on the approximation error for this algorithm in terms of the *tolerance* parameter and the degrees of visited nodes.

Proposition 2.2 (Entrywise approximation error (Andersen et al., 2006))**.** *Let* p *be a PPR vector, and let* $p^{\varepsilon} \in [0,1]^{N}$ *be an approximate PPR vector computed by Algorithm 2.1 with a tolerance* $\varepsilon > 0$*. For any vertex* u *that is sampled in Algorithm 2.1,*

$$|p_u - p_u^{\varepsilon}| \leqslant \varepsilon d_u.$$

Proposition 2.2 ensures that for any fixed graph, the approximate PPR vector is arbitrarily close to the exact PPR vector, as long as the tolerance $\varepsilon > 0$ is sufficiently small. Appendix A.1 contains a proof of this proposition for completeness. Given a seed node in the graph, Algorithm 2.2 uses the approximate PPR vector from Algorithm 2.1 and returns a set of nodes with the largest corresponding values in the *adjusted personalized PageRank* (aPPR) vector, which is defined as

$$p_v^* = \frac{p_v}{d_v}, \text{ for } v = 1, 2, ..., N.$$

The aPPR vector was previously proposed in Andersen et al. (2006). Algorithm 2.1 and 2.2 operate on undirected graphs. We will generalize them to directed graphs in Section 2.3 thanks to a simplified and interpretable form for the PPR vector.

```
Input: Undirected graph $G$, preference vector $\pi$, teleportation constant
       $\alpha$, and tolerance $\varepsilon$.
Procedure:
    Initialize $p \leftarrow 0$, $r \leftarrow \pi$, $\alpha' \leftarrow \alpha/(2-\alpha)$.
    while $\exists u \in V$ such that $r_u \geqslant \varepsilon d_u$ do
        Uniformly sample a vertex $u$ satisfying $r_u \geqslant \varepsilon d_u$.
        $p_u \leftarrow p_u + \alpha' r_u$.
        for $v : (u, v) \in E$ do
            $r_v \leftarrow r_v + (1-\alpha') r_u / (2 d_u)$.
        $r_u \leftarrow (1-\alpha') r_u / 2$.
Output: $\varepsilon$-approximate PPR vector $p$.
```

Algorithm 2.1: Approximate PPR Vector (undirected) (Andersen et al., 2006)

```
Input: Undirected graph $G$, seed node $v_0$, and the desired size of local
    cluster $n$.
Procedure:
    1. Calculate the approximate PPR vector $p$ (Algorithm 2.1).
    2. Adjust the PPR vector $p$ by node degrees, $p_v^* \leftarrow p_v/d_v$.
    3. Rank all vertices according to the adjusted PPR vector $p^*$.
Output: Local cluster – $n$ top-ranking nodes.
```

Algorithm 2.2: PPR Clustering (undirected)

Stochastic block model

In the stochastic block model (SBM), each node belongs to one of K blocks. The presence of each edge corresponds to an independent Bernoulli random variable, where the probability of an edge between any two nodes depends only on the block memberships of two nodes (Holland et al., 1983a). The formal definition is as follows.

Definition 2.3. *For a vertex set* $V = \{1, 2, ..., N\}$*, let* $z : \{1, 2, ..., N\} \to \{1, 2, ..., K\}$ *partition the* N *nodes into* K *blocks, so* $z(v)$ *is the block membership of vertex* v*. Let* $\mathbf{B}$ *be a* $K \times K$ *matrix with all entries range in* $[0, 1]$*. Under the SBM, the probability of an edge between* u *and* v *is* $\mathbf{B}_{z(u)z(v)}$*. That is,* $A_{uv} \mid z(u), z(v) \overset{\text{ind.}}{\sim}$ *Bernoulli* $\left(\mathbf{B}_{z(u)z(v)}\right)$*, for any* $u, v \in \{1, 2, ..., N\}$*.*

Under the ordinary SBM, nodes in the same block have the same expected degree. One extension is the degree-corrected stochastic block model (DC-SBM), which adds a series of parameters ($\theta_v > 0$ for every vertex v) to create more heterogeneous node degrees (Karrer and Newman, 2011b). Let $\mathbf{B}$ be a $K \times K$ matrix with $\mathbf{B}_{ij} > 0$ for any i and j. Then the probability of an edge between u and v is $\theta_u \theta_v \mathbf{B}_{z(u)z(v)}$. That is,

$$A_{uv} \mid z(u), z(v) \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\theta_u \theta_v \mathbf{B}_{z(u)z(v)}\right),$$

for $u, v \in \{1, 2, ..., N\}$. Since θ_v's are arbitrary to a multiplicative constant which can be absorbed into $\mathbf{B}$, Karrer and Newman (2011b) suggest imposing the constraint that the θ_v's sum to 1 within each block. That is, $\sum_{v:z(v)=i} \theta_v = 1$

for all $i = 1, 2, .., K$. With this constraint, $\mathbf{B}_{ij}$ represents the expected number of edges between block i and j if $i \neq j$, and twice of that if $i = j$. Throughout this paper, we presume $\mathbf{B}$ is positive definite [2] and all blocks are connected (we ignore any blocks that are isolated from the seed). The DC-SBM can be generalized to directed graphs by giving each node two parameters, θ_v^{in} and θ_v^{out}, controlling its in-degree and out-degree respectively (Zhu et al., 2013). Then, the presence of an directed edge from u to v, given the block memberships, corresponds to an independent Bernoulli random variable,

$$A_{uv} \mid z(u), z(v) \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\theta_u^{\text{out}} \theta_v^{\text{in}} \mathbf{B}_{z(u)z(v)}\right).$$

In order to make the model identifiable, we need to impose a structural constraint on θ^{in}'s and θ^{out}'s, that both of them sum up to 1 within each block,

$$\sum_{v:z(v)=i} \theta_v^{\text{in}} = \sum_{v:z(v)=i} \theta_v^{\text{out}} = 1, \text{ for any } i = 1, 2, ..., K.$$

Because the off-diagonal elements of $\mathbf{B}$ can be interpreted as the expected number of edges between blocks, we define the block in-degree and block out-degree to be the total number of incoming edges and outgoing edges respectively, that is, $\mathbf{d}_j^{\text{in}} = \sum_{i=1}^K \mathbf{B}_{ij}$, and $\mathbf{d}_i^{\text{out}} = \sum_{j=1}^K \mathbf{B}_{ij}$.

2.3 Population analysis of PageRank

In this section, we analyze the PPR vector of the expected adjacency matrix under the DC-SBM. This provides a simple representation of the PPR vector that motivates (1) the bias adjustment and (2) the generalization of Algorithm 2.1 and 2.2 to directed graphs.

We use three distinct typefaces to denote three classes of objects. Calligraphic typeface is given to the population version of any observable quantities in random graphs, such as graph adjacency matrix and node degrees (e.g.

[2] This prevents scenarios where edges are unlikely within blocks and more likely between blocks. In such scenarios, local clustering needs to be reimagined cautiously. See Appendix A.2 for additional details about generalizations.

Equation (2.3)). Normal typeface is given to unobserved model parameters, such as block membership and degree parameters θ_i. Bold face is given to all block-level quantities and parameters like $\mathbf{B}$ and $\mathbf{d}_i^{\text{out}}$.

Define the population graph adjacency matrix,

$$\mathscr{A} = \mathbb{E}\left(A \mid z(1), z(2), ..., z(N)\right), \tag{2.3}$$

to be the expectation of random adjacency matrix A. Let $Z \in \{0,1\}^{N\times K}$ be the block membership matrix with $Z_{vi} = 1$ if and only if vertex v belongs to block i, and define diagonal matrices Θ^{in} and Θ^{out} with entries θ^{in}'s and θ^{out}'s respectively. Then, under the directed DC-SBM with K blocks and parameters $\{\mathbf{B}, Z, \Theta^{\text{in}}, \Theta^{\text{out}}\}$, $\mathscr{A} \in \mathbb{R}^{K\times K}$ can be compactly expressed as

$$\mathscr{A} = \Theta^{\text{out}} Z \mathbf{B} Z^T \Theta^{\text{in}}.$$

Accordingly, we define the population node degrees and the population transition matrix, $\mathscr{d}_u^{\text{in}} = \sum_{v\in V} \mathscr{A}_{uv}$, $\mathscr{d}_v^{\text{out}} = \sum_{u\in V} \mathscr{A}_{uv}$, and $\mathscr{P} = [\mathscr{D}^{\text{out}}]^{-1}\mathscr{A}$, where $\mathscr{D}^{\text{in}}$ and $\mathscr{D}^{\text{out}}$ are the diagonal matrices of the population node in-degrees $\mathscr{d}_u^{\text{in}}$'s and out-degrees $\mathscr{d}_v^{\text{out}}$'s respectively. Let $\mathscr{p}$ be the population PPR vector (i.e., the solution to equation $\mathscr{p}^T = \alpha\pi^T + (1-\alpha)\mathscr{p}^T\mathscr{P}$) and let $\mathscr{p}^* = \left[\mathscr{D}^{\text{in}}\right]^{-1}\mathscr{p}$ be the population aPPR vector.

In addition, define the *block transition matrix* $\mathbf{P} \in \mathbb{R}^{K\times K}$ as

$$\mathbf{P} = \left[\mathbf{D}^{\text{out}}\right]^{-1}\mathbf{B}, \tag{2.4}$$

where $\mathbf{D}^{\text{in}} \in \mathbb{R}^{K\times K}$ and $\mathbf{D}^{\text{out}} \in \mathbb{R}^{K\times K}$ are diagonal matrices of the block in-degrees $\mathbf{d}_i^{\text{in}}$'s and out-degrees $\mathbf{d}_i^{\text{out}}$'s.

A representation of PPR vectors

This section provides a simple and interpretable form for PPR vectors under the population DC-SBM. To this end, we define the *"block-wise" PPR vector* $\mathbf{p} \in \mathbb{R}^K$ to be the unique solution to linear system

$$\mathbf{p}^T = \alpha\boldsymbol{\pi}^T + (1-\alpha)\mathbf{p}^T\mathbf{P}, \tag{2.5}$$

where $\boldsymbol{\pi} = Z^T\pi \in \mathbb{R}^K$ is the block-wise preference vector and $\mathbf{P}$ is the block transition matrix in Equation (2.4). This treats the block connectivity matrix $\mathbf{B}$ as a weighted adjacency matrix of blocks and the block of seed node as a seed block. To build up the relationship between PPR and the block-wise PPR, the next theorem gives an explicit form for PPR vectors which also reveals the sources of bias for local clustering.

Theorem 2.4 (Explicit form of PPR vectors)**.** *Under the population directed DC-SBM with* K *blocks and parameters* $\left\{\mathbf{B}, Z, \Theta^{\text{in}}, \Theta^{\text{out}}\right\}$*,*

(a) *the population PPR vector* $\mathscr{p} \in \mathbb{R}^N$ *has elements*

$$\mathscr{p}_u = \theta_u^{\text{in}} \mathbf{p}_{z(u)}$$

where $\mathbf{p}$ *is the block-wise PPR vector in Equation* (2.5)*,*

(b) *and the population aPPR vector* $\mathscr{p}^* \in \mathbb{R}^N$ *has elements*

$$\mathscr{p}^*_u = \mathbf{p}^*_{z(u)} \tag{2.6}$$

where $\mathbf{p}^* = \left[\mathbf{D}^{\text{in}}\right]^{-1}\mathbf{p}$.

Theorem 2.4 demonstrates that the PPR vector $\mathscr{p}$ decomposes into block-related information ($\mathbf{p}$) and node specific information (Θ). Within each block, the PPR values are proportional to the node degree parameters θ_v's and sum up to the block-wise PPR value of the block. The proof of Theorem 2.4 (Appendix A.1) relies on a key observation (Appendix A.1) that the powers of population transition matrix, $\mathscr{P}^s$ for $s = 1, 2, \ldots$, have a similarly simple form and the node specific information components (i.e., $z(v)$ and θ_v) are invariant in s.

In order to justify the adjustment (Step 2.2) in Algorithm 2.2, we observe that the seed always has the highest population aPPR score. This turns out to be a key feature that facilitates the aPPR vector to recover a local cluster correctly, so we state it in the following lemma.

Lemma 2.5 (The largest entry of aPPR vector)**.** *Under the population DC-SBM, assume that the minimum expected degree is positive, that is,* $\min_{v \in V} \mathscr{d}_v > 0$*. Then,* *for any fixed $\alpha > 0$, the population aPPR vector $\mathscr{p}^*$ has the strictly largest entry corresponding to the seed node,*

$$\mathscr{p}^*_{v_0} > \mathscr{p}^*_v, \textit{ for any } v \neq v_0.$$

On the other hand, this is not generally true for a PPR vector.

When $\alpha = 0$ (i.e., no teleportation), the PPR vector becomes the limiting distribution of a standard random walk and all entries of the aPPR vector are equal (Appendix A.1). Lemma 2.5 (applied to block-wise PPR vectors) and Theorem 2.4 together identify two sources of bias for PPR vectors and suggest a justification for the degree adjustment, which we discuss in order:

(i) Both node degree heterogeneity (Θ) and block size imbalance ($\mathbf{D}$) confound the identification of local cluster by the PPR vector. In particular, suppose vertex v belongs to a block $z(v) = i$ other than 1. PPR vector assigns it a score $\theta_v \mathbf{p}_i$, where $\mathbf{p}_i$ is the block-wise PPR of block i, and θ_v is the parameter specifically controlling the degree of v. Then, node v may rank at the top, if θ_v is large enough. Furthermore, Lemma 2.5 implies that $\mathbf{p}_1$ is not necessarily the largest due to block degree heterogeneity. Specifically, if block i has an exceedingly high block degree, it is likely that $\mathscr{p}$ fails to down-rank node v vis-a-vis those nodes of block 1.

(ii) Adjusted personalized PageRank removes the node and the block degree heterogeneity simultaneously, and perfectly recovers the local cluster. To see this, note that $\mathbf{p}^*$ is the adjusted version of block-wise PPR vector. From Lemma 2.5, $\mathbf{p}^*_1$ is the largest entry of $\mathbf{p}^*$. From Equation 2.6, the aPPR vector assigns any vertex v a score $\mathbf{p}^*_{z(v)}$. Hence, nodes with the highest value of $\mathscr{p}^*$ belong to block 1, which is precisely the desired local cluster.

Note that the PPR vector can still be biased for local clustering even under the classic SBM. To see this, set the matrix Θ to the identity matrix in Theorem 2.4. In this case, the heterogeneous block degrees still confound the PPR vector (Section 2.5); there is generally no guarantee for $\mathbf{p}_1$ to appear on the top (due to Lemma 2.5), unless there are further symmetry conditions. Kloumann et al. (2017) uses such one scenario. As a byproduct of our analysis, we extend their results under the DC-SBM with the symmetric conditions (see Appendix A.3 to the paper).

Local clustering on directed graphs

In light of the clean form of PPR vectors under the DC-SBM, one can modify Algorithm 2.1 and 2.2 to operate on a directed graph accordingly. To this end, note that the transition matrix of a directed graph requires node out-degrees, hence Algorithm 2.1 examines only the edges leaving visited nodes. Consequently it suffices to replace d_u's in Algorithm 2.1 by d_u^{out}'s (Algorithm 2.3). Proposition 2.2 applies to Algorithm 2.3 as well, and one can approximate the PPR vector provided the out-degrees of visited nodes can be observed and the tolerance parameter $\varepsilon > 0$ is sufficiently small.

To perform local clustering on a directed graph, Algorithm 2.4 adjusts the approximate PPR vectors from Algorithm 2.3 by node in-degrees, that is,

$$p_v^* = \frac{p_v}{d_v^{\text{in}}}, \text{ for } v = 1, 2, ..., N.$$

Another option is regularized adjustment, which produces the *regularized* PPR (rPPR) vector,

$$p_v^\tau = \frac{p_v}{d_v^{\text{in}} + \tau}, \text{ for } v = 1, 2, ..., N,$$

where $\tau > 0$ is the regularization parameter. The regularized adjustment greatly stabilize the PPR clustering in practice, by removing nodes with extremely low in-degrees (see Section 2.6 for more details). Adjusted PPR for directed graphs is a local algorithm so long as d^{in} is available with a local query, for example, the Twitter friendship graph.

2.4 Personalized PageRank in random graphs

This section establishes several concentration results for the local clustering algorithm using the adjusted PPR vector (Algorithm 2.2 and 2.4) under the

```
Input: Directed graph $G$, preference vector $\pi$, teleportation constant $\alpha$,
       and tolerance $\varepsilon$.
Procedure:
    Initialize $p \leftarrow 0$, $r \leftarrow \pi$, $\alpha' \leftarrow \alpha/(2-\alpha)$.
    while $\exists u \in V$ such that $r_u \geqslant \varepsilon d_u^{\mathrm{out}}$ do
        Sample a vertex $u$ uniformly at random, satisfying $r_u \geqslant \varepsilon d_u^{\mathrm{out}}$.
        $p_u \leftarrow p_u + \alpha' r_u$.
        for $v : (u, v) \in E$ do
            $r_v \leftarrow r_v + (1-\alpha') r_u/(2 d_u^{\mathrm{out}})$.
        $r_u \leftarrow (1-\alpha') r_u/2$.
Output: $\varepsilon$-approximate PPR vector $p$.
```

Algorithm 2.3: Approximate PPR Vector (directed)

```
Input: Directed graph $G$, seed node $v_0$, the desired size of local cluster
       $n$, and an optional regularization parameter $\tau$.
Procedure:
    1. Calculate the approximate PPR vector $p$ (Algorithm 2.1).
    2. Adjust the PPR vector $p$ with:
        Option (a): node in-degrees, $p_v^* \leftarrow p_v/d_v^{\mathrm{in}}$,
        Option (b): regularized node in-degrees, $p_v^\tau \leftarrow p_v/(d_v^{\mathrm{in}} + \tau)$.
    3. Rank all vertices according to the aPPR vector $p^*$ or $p^\tau$
Output: Local cluster – $n$ top-ranking nodes.
```

Algorithm 2.4: PPR Clustering (directed)

DC-SBM. The results show that if the graph is generated from the DC-SBM, then PPR clustering returns the desired local cluster with high probability. Since in Algorithm 2.4, the calculation for PPR vectors only relies on node out-degrees and the adjustment step solely utilizes node in-degrees, it is not difficult to distinguish d^{in} and d^{out}. Thus, we state the results in undirected graphs for simplicity. One can draw the analogous conclusions for directed graphs by tracing the proof step by step.

We first present a useful tool that controls the entrywise errors of a PPR vector in random graphs. Recall that $\mathscr{p}$ is the stationary distribution of probability transition matrix $\mathscr{Q} = \alpha\Pi + (1-\alpha)\mathscr{P}$. For any vector $x \in \mathbb{R}^n$, define the vector infinity norm as $\|x\|_\infty = \max_i |x_i|$. The following theorem bounds the entrywise error of the stationary distribution of $\mathscr{Q}$.

Theorem 2.6 (Concentration of the PPR vectors)**.** *Let* $G = (V, E)$ *be a graph of* N *vertices generated from the DC-SBM with* K *blocks and parameters* $\{\mathbf{B}, Z, \Theta\}$*. Let* p *and* $\mathscr{p}$ *be the PPR vector corresponding to random transition matrix* P *and its population version* $\mathscr{P}$ *respectively, with the same teleportation constant* α*. Let* $p^*, \mathscr{p}^* \in [0, 1]^N$ *be the adjusted PPR vector of* p *and* $\mathscr{p}$*. Let* δ *be the average expected node degrees, that is,* $\delta = \frac{1}{N}\sum_{v \in V} \mathscr{d}_v$*. Assume that* $\rho = \frac{\max_{v\in V} \mathscr{d}_v}{\min_{v \in V} \mathscr{d}_v}$ *is bounded by some finite constant and that*

$$\delta > c_0(1-\alpha)^2 \log N, \tag{2.7}$$

for some sufficiently large constant $c_0 > 0$*. Then, with probability at least* $1 - \mathscr{O}(N^{-5})$*,*

$$\frac{\|p - \mathscr{p}\|_\infty}{\|\mathscr{p}\|_\infty} \leqslant c_1(1-\alpha)\sqrt{\frac{\log N}{\delta}}, \quad \textit{and} \quad \frac{\|p^* - \mathscr{p}^*\|_\infty}{\|\mathscr{p}^*\|_\infty} \leqslant c_2(1-\alpha)\sqrt{\frac{\log N}{\delta}},$$

for some sufficiently large constant $c_1, c_2 > 0$*.*

The proof of Theorem 2.6 invokes the elementary eigenvector perturbation bound for asymmetric matrices, an analog to the celebrated Davis-Kahan $\sin\Theta$ theorem (Davis and Kahan, 1970), and the novel leave-one-out technique due to Chen et al. (2019). The detailed proof is given in Appendix A.1.

Theorem 2.6 demonstrates that if the expected average degree δ exceeds $(1-\alpha)^2 \log N$ to some sufficiently large extent, then with high probability, the random aPPR vector concentrates around the population aPPR vector in terms of all entries. In fact, the concentration statement holds for any valid preference vector π. Hence, the classic PageRank vector and some other variants also enjoy the entrywise error bounds, so long as they can be written as the solution to the linear system (2.1).

Next, we introduce a separation measure of the DC-SBM. Recall that one can conduct a local clustering task by selecting nodes ranked by the adjusted PPR vector p^*. In the population version, it is equivalent to distinguishing between $\mathbf{p}_1^*$ and $\mathbf{p}_k^*$, for all $k = 2, 3, ..., K$, which also characterizes the distance from the desired local cluster (block 1) to its complement set (the other blocks). Only if they are sufficiently separated, can the local cluster be identifiable in the sample. Due to Lemma 2.5, we assume without loss of generality that the second block has the second highest value in the "block-wise" aPPR vector, that is, $\mathbf{p}_1^* > \mathbf{p}_2^* \geqslant \mathbf{p}_k^*$ for $k = 3, 4, ..., K$. Then, we define the *separation measure* $\Delta_\alpha \in (0, 1]$,

$$\Delta_\alpha = \frac{\mathbf{p}_1^* - \mathbf{p}_2^*}{\mathbf{p}_1^*},$$

which turns out to be crucial in determining the sample complexity required to guarantee the exact recovery. We remark that Δ_α is an increasing function of the teleportation constant, hence the subscript α.

With Theorem 2.6 and the separation measure, we then give following corollary that bounds the accuracy of Algorithm 2.2, in terms of graph edge density.

Corollary 2.7 (Exact recovery by adjusted PPR vector)**.** *For any seed nodes, let $C \subset V$ be the local cluster of n nodes returned by Algorithm 2.2 with teleportation constant α and tolerance ε, and $\mathscr{C} \subset V$ be the nodes in the seed node's block. Assume that $\rho < c_0$, $\varepsilon \leqslant c_1(1-\alpha)\mathbf{p}_1^* \sqrt{\log N/\delta}$, and that*

$$\delta > 16c_2 \left(\frac{1-\alpha}{\Delta_\alpha}\right)^2 \log N, \tag{2.8}$$

for some sufficiently large constants $c_0, c_1, c_2 > 0$. If the desired size of the local cluster $n = |\mathscr{C}|$, then with probability at least $1 - \mathcal{O}(N^{-5})$, we have $C = \mathscr{C}$.

The proof of Corollary 2.7 is presented in Appendix A.1. We make a few remarks:

(i) Corollary 2.7 demonstrates that Algorithm 2.2 works under a sparse scenario, where the number of edges is exceedingly small in proportion to the number of possible edges in the network. To reach the entrywise control of the aPPR vector and the sufficient separation of local cluster from others, the theorem calls for the expected node degree δ to grow with only a fraction (for any fixed teleportation constant α) of the logarithm of the size of the network, $\log N$. In other words, Algorithm 2.2

requires a sample complexity (the number of edges) of order

$$\left(\frac{1-\alpha}{\Delta_\alpha}\right)^2 N \log N.$$

(ii) The results show that α leverages between the sampling complexity and statistical performance of PPR clustering. To see this, rearrange condition (2.8),

$$\left(\frac{1-\alpha}{\Delta_\alpha}\right)^2 < \frac{c'\delta}{\log N},$$

for some small enough constant $c' > 0$. As α increases, the left hand side is decreasing to zero thus making the condition more likely to hold. On the other hand, as α increases, the tolerance ε must decrease at rate $\mathcal{O}(1-\alpha)$ in order to guarantee an entrywise control of p^ε analogous to the form in Theorem 2.6 (Appendix A.1). More intuitively, if ε does not decrease, then as α goes to one, Algorithm 2.1 may terminate early without reaching all vertices in the desired local cluster. In sum, Algorithm 2.1 and 2.3 need at least $\mathcal{O}\left(\frac{1}{\alpha(1-\alpha)}\right)$ queries (see Appendix A.2 for an example). This implies that one can approach the conditions in Corollary 2.7 by setting the teleportation constant sufficiently large, while the computational burden can increase as $\alpha \to 1$.

2.5 Simulation studies

This section compares the PPR vector and the aPPR vector. The results show the effectiveness and robustness of aPPR vector in detecting a local cluster. Experiment 1 utilizes the DC-SBM with a power-law degree distribution and investigates the effects of heterogeneous node degrees. Experiment 2 uses the SBM with unequal block sizes to study the influences of heterogeneous block degrees. Experiment 3 generates networks from the SBM with equal block sizes and varying edge density to examine the efficacy of PPR methods in sparse graphs.

In all simulations, we employee the block connectivity matrix **B** with homogeneous diagonal elements, $\mathbf{B}_{ii} = b_1$, and homogeneous off-diagonal elements, $\mathbf{B}_{ij} = b_2$ for any $i \neq j$. Define the signal-to-noise ratio (SNR) to be the expected number of in-block edges divided by the expected number of out-block edges, that is, $b_1/(b_2(K-1))$, where K is the number of blocks. In particular, we set the SNR to 1.5 and choose the teleportation constant of $\alpha = 0.15$ throughout the section. Additional simulation results (illustrating the Theorem 2.6) are available in Appendix A.2.

Experiment 1

This experiment illustrates how node degree heterogeneity affects the discriminant power in identifying local cluster using a PPR vector or an aPPR vector. The results also illustrate the advantages of having multiple seed nodes. The Θ parameters from the DC-SBM are drawn from the power law distribution with lower bound $x_{\min} = 1$ and shape parameter $\beta = 2.5$. A random networks were sampled from the DC-SBM with $K = 3$, $N = 1500$ and equal block sampling proportions,

$$z(v) \overset{\text{i.i.d.}}{\sim} \text{Multinomial}\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right),$$

for vertex $v = 1, 2, ..., N$, whose expected average degree (δ) is set to 105. The PPR vector is calculated with one or ten seeds randomly chosen from block one.

Figure 2.1 plots PPR values (left two panels) and aPPR values (right two panels) of a random graph generated from the DC-SBM, excluding seed node(s). The upper two panels in Figure 2.1 contrast PPR and aPPR when there is only one seed node and the bottom two panels compare two vectors when ten seed nodes are used. The vertices from the local block in the SBM are colored in blue and the others are in yellow. The nodes are ordered first by block, then by node degree parameters θ (left is larger). A horizontal line is drawn for each block indicating the median of the aPPR values within that block.

With one seed node (upper two panels), the scatter plots has two clouds within each block. The upper cloud contains the immediate neighbors of the

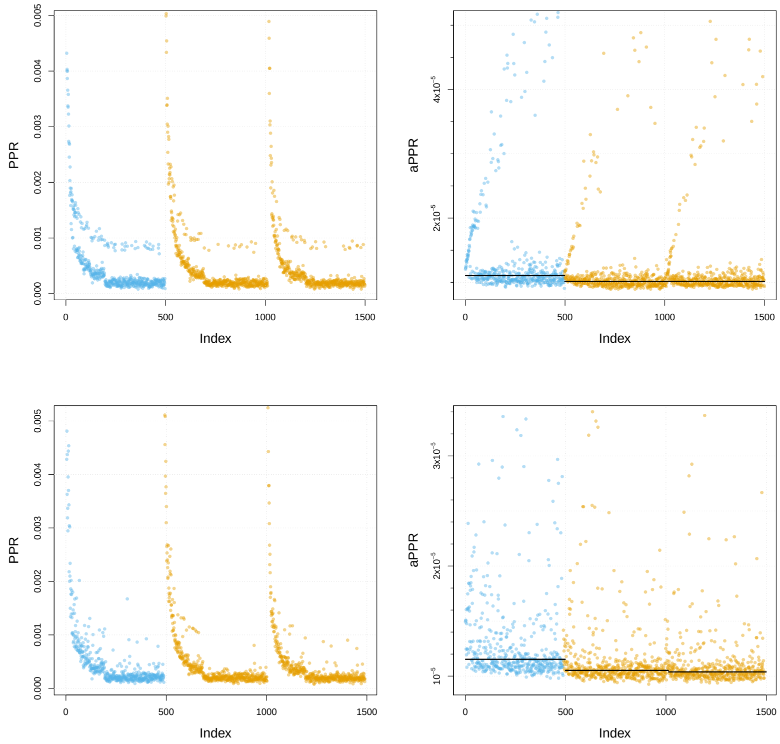


Figure 2.1: Comparison of PPR (left two panels) and aPPR (right two panels) under the DC-SBM with one seed node (upper two panels) and ten seed nodes (bottom two panels). Local cluster is in blue and other clusters are colored in yellow. Solid horizontal lines on right panels indicate the median of aPPR values within each cluster.

seed node. This separation disappears when multiple seed nodes are used (bottom two panels). To see the effect of node heterogeneity, the skewed distribution of PPR values in each block demonstrates its bias towards high degree node inside and outside of the seed nodes block in the SBM. In contrast, aPPR values are evenly distributed within blocks, verifying that aPPR vector removes the effects of node degree heterogeneity.

Experiment 2

This experiment compares PPR and aPPR under the SBM with block degree heterogeneity. A number of random networks were sampled from the SBM with $K = 3$, $N = 900$, and geometric block sampling proportions,

$$z(v) \overset{\text{i.i.d.}}{\sim} \text{Multinomial}\left(1, b, b^2\right), \tag{2.9}$$

where $b \in \{1.0, 1.2, 1.4, 1.6, 1.8, 2.0\}$. When b is larger, the population of nodes in each block becomes more unbalanced and thus inducing greater block degree heterogeneity. The block connectivity matrix $\mathbf{B}$ is configured as described in the beginning of this section. The expected average degree (δ) is set to 70. For each sampled network, the size of the first block is assumed known to Algorithm 2.2. The PPR vector is calculated exactly in place of the approximation PPR vector (Step 1), with one seed randomly chosen from the first block.

The top panels of Figure 2.2 displays the PPR vector on an example network with $b = 1.4$, demonstrating its preference toward the high degree block (the third block) over local cluster. Given the size of the first block, we measure the accuracy by the proportion of vertices belonging to the first block in the returned cluster. The bottom left panel of Figure 2.2 shows the accuracy of PPR and aPPR for six different values of b (i.e., the geometric ratio in distribution (2.9)) where each point is the average of 100 sampled network. The comparison demonstrates that the adjusted PPR vector corrects the bias of PPR caused by block heterogeneity. Moreover, block degree heterogeneity degrades the performance of both PPR and aPPR. Note that aPPR outperforms PPR even when $b = 1$; this is likely due to the fact that even when nodes have equal expected degrees in the SBM, the actual node degrees will be heterogeneous due to the randomness in the sampled graph. In a finite graph, this variability is enough to give aPPR an advantage over PPR. Asymptotically, this advantage should fade away (Kloumann et al., 2017).

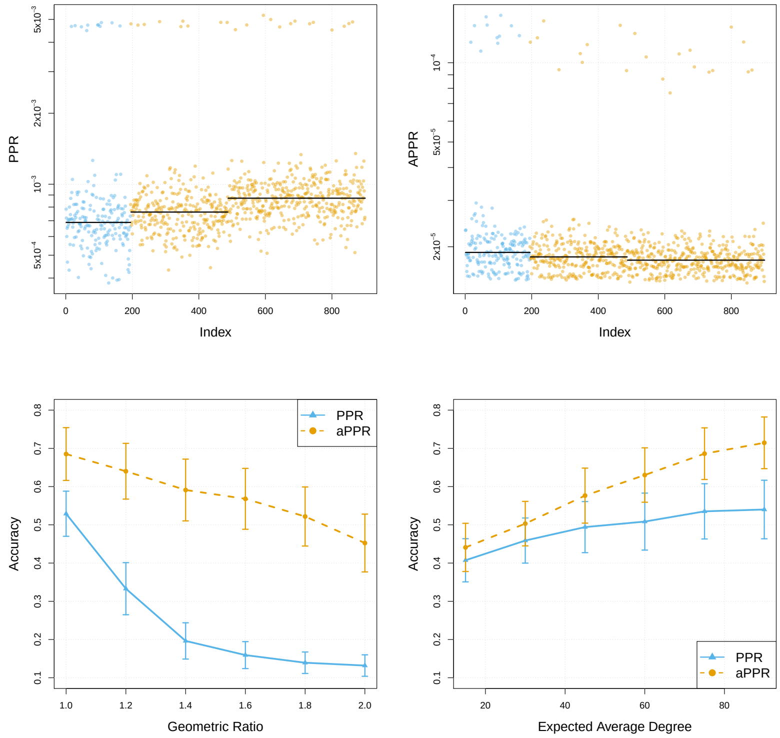


Figure 2.2: (Top) Simulated network generated from the classic SBM of 3 blocks with block degree heterogeneity. Three horizontal lines indicate the median of PPR and aPPR values within each cluster. (Bottom Left) Comparison of performance for PPR (triangles with solid line) and aPPR (circles with dashed line) under the SBM with different levels of block degree heterogeneity. (Bottom Right) Comparison of performance for PPR and aPPR under the four-parameters SBM with different sparsity. Error bars are drawn using standard deviation.

Experiment 3

This experiment investigates the performance of PPR and aPPR under the SBM where there is no heterogeneity in the expected node degrees or block degrees. A number of random networks were sampled from the four-parameter stochastic block model, $\text{SBM}(K = 3, N = 900, b_1 = 0.6, b_2 = 0.2)$ (Rohe et al., 2011). Under the four-parameter SBM, each of K blocks has equal size in expectation, N/K, and the probability of a connection between two nodes is b_2 if they are in two separate blocks, or b_1 if in the same one. In addition, the expected average degree varies, $\delta \in \{15, 30, 45, 60, 75, 90\}$. For every setting, the results are averaged over 100 samples of the network. The PPR vector is calculated with one seeds randomly chosen from block one. The bottom right panel of Figure 2.2 contrasts the accuracy of PPR and aPPR against six different values of expected average degree, showing that when the sampled graph has minimal degree heterogeneity, the adjusted PPR vector has only slightly higher accuracy than the PPR vector.

2.6 A sample of Twitter

In this section, we provide a more detailed case study to illustrate the properties of different PPR vectors. We obtain a local cluster of nodes around the seed node @NBCPolitics (NBC Politics) in the Twitter friendship graph. In the Twitter graph, the nodes are called handles or accounts (e.g. @NBCPolitics) and if Twitter handle i follows Twitter handle j, then we define this as a directed edge (i, j) pointing from i to j. Affiliated with NBC news, NBC Politics specializes in political news coverage and has over 470k followers on Twitter (in-degree) and follows 145 handles (out-degree) as of December 2018. A brief look through @NBCPolitics' following list reveals that it follows a wide range of accounts, from TV programs, reporters and editors affiliated with NBC, to media accounts and journalists of other news outlets as well as politicians.

Data on following and handle profile information were collected through the Standard Twitter Search API. We queried the Twitter friendship graph starting from the seed node @NBCPolitics, using Algorithm 2.3 with teleportation constant $\alpha = 0.15$ and termination parameter $\varepsilon = 10^{-7}$, ending up with 5840 surrounding handles. Through this exercise, we intend to illustrate the properties and applications of local clustering using PPR, aPPR and rPPR vectors, where we set the regularization parameter τ to 100.

We first present the results of PPR. As Table 2.3 shows, the top 30 handles (except @NBCPolitics) with the highest PPR values are a combination of (i)

Table 2.3: Top 30 handles of PPR with seed node @NBCPolitics and the teleportation constant $\alpha = 0.15$ in December 2018.

	Name	Followers	Description
1	Melania Trump	11242283	This account is run by the Office of First Lady Melania Trump...
2	The White House	17625630	Welcome to @WhiteHouse! Follow for the latest from President...
3	Chuck Todd	2032038	Moderator of @meetthepress and @nbcnews political director; ...
4	NBC News	6280551	The leading source of global news and info for more than 75 ...
5	NBC Nightly News	962290	Breaking news, in-depth reporting, context on news from ...
6	Andrea Mitchell	1737764	NBC News Chief Foreign Affairs Correspondent/anchor, Andrea ...
7	Savannah Guthrie	881669	Mom to Vale & Charley, TODAY Co-Anchor, Georgetown Law. ...
8	Joe Scarborough	2521215	With Malice Toward None
9	MSNBC	2261911	The place for in-depth analysis, political commentary and ...
10	Rachel Maddow MSNBC	9498076	I see political people...
11	Breaking News	9223158	
12	NBC News First Read	53847	The first place for news and analysis from the @NBCNews Poli...
13	TODAY	4276453	America's favorite morning show | Snapchat: todayshow
14	Meet the Press	566713	Meet the Press is the longest-running television show in history ...
15	The Wall Street Journal	16188842	Breaking news and features from the WSJ.
16	Pete Williams	70062	NBC News Justice Correspondent. Covers US Supreme Court, ...
17	Mark Murray	97571	Mark Murray is the senior political editor for NBC News, ...
18	POLITICO	3695835	Nobody knows politics like POLITICO. Got a news tip for us? ...
19	Katy Tur	587474	MSNBC anchor @2pm, NBC News correspondent, author of NYT ...
20	Bill Clinton	10697521	Founder, Clinton Foundation and 42nd President of the United...
21	Kasie Hunt	381704	@NBCNews Capitol Hill Correspondent. Host, @KasieDC, Sundays...
22	TIME	15584815	Breaking news and current events from around the globe. Host...
23	Kelly O'Donnell	195765	White House Correspondent @NBCNews Veteran of Cap Hill & ...
24	John McCain	3181773	Memorial account for U.S. Senator John McCain, 1936-2018. To...
25	Peter Alexander	283522	@NBCNews White House Correspondent / Weekend @TODAYshow ...
26	Hallie Jackson	359099	Chief White House Correspondent / @NBCNews / @MSNBC Anchor ...
27	Kristen Welker	182244	@NBCNews White House Correspondent. Links and retweets ...
28	Carrie Dann	37119	.@NBCNews / @NBCPolitics. RTs not endorsements.
29	Willie Geist	807536	Host @NBC #SundayTODAY, Co-Host @Morning_Joe, "Sunday ...
30	Morning Joe	563650	Live tweet during the show! Links to must-read op-eds ...

Through the PPR vector, the top 30 handles returned to @NBCPolitics include NBC's news related programs and celebrity reporters, comparable mainstream media outlets, as well as prominent political and public figures and institutions. Such results line up with its status as a mainstream political news source, demonstrating clustering effectiveness. Those Twitter handles tend to have millions of followers, showing the PPR vector's bias toward high in-degree.

NBC's news related programs such as NBC News, TODAY and Meet the Press; (ii) NBC's political reporters, anchors and editors, from well-known figures like Chuck Todd and Andrea Mitchell to less-known ones like Pete Williams (justice correspondent) and Mark Murray (senior political editor); (iii) other mainstream news outlets such as The Wall Street Journal, POLITICO, and TIME; and (iv) prominent public figures and politicians like Melania Trump, Bill Clinton and John McCain. In light of NBC's status as a mainstream news outlet and the political focus of @NBCPolitics, such results make sound sense. It must also be noted that all the top 30 handles are direct friends of @NBCPolitics's and have at least tens of thousands of followers. The median follower count is 1.4 million, suggesting high in-degrees. In fact, the pattern observed in the top 30 extends to the top 200 handles with the highest PPR values, which include NBC's own programs, journalists, editors and staff; fellow mainstream media outlets and their staff; and prominent public figures, politicians and government institutions (see Appendix A.4). The median in-degree of top 200 handles is around 184k, though there are four handles with less than one thousand followers. One important thing to notice is that among the top 200 handles, the first 139 are all directly followed by @NBCPolitics, with handles having high in-degrees generally ranked higher than those having low in-degrees (although @NBCPolitics follows 145 handles, 6 of them might have privacy protection that has prevented us from accessing their information). The remaining handles on the list, although not directly followed by @NBCPolitics, include five handles associated with NBC, from its news anchor Lester Holt to its News International President. However, the majority of those indirectly followed by @NBCPolitics are mainly high profile political and public figures (like President Trump, Vice President Pence, Hillary Clinton, and Stephen Colbert), government organizations (like WhiteHouse Office of Cabinet Affairs and National Security Council), and mainstream news outlets (like New York Times, CNN and AP) and well-known journalists (like John Dickerson and Anderson Cooper). We can thus conclude that the PPR vector is biased toward popular accounts followed directly by the seed node or indirectly by its friends, reflecting the popular Twitter handles followed by them. This property of the PPR vector can be harnessed by researchers interested in identifying the upstream of a handle, i.e., those Twitter elites who are followed by and might influence the seed node and by extension its followers.

In contrast, the aPPR vector up-weights handles that are much less popular (i.e., those with low in-degrees). As shown in Table 2, the 30 handles with the highest aPPR values include NBC's reporters, writers, editors, producers, and programs, all of whom have a few hundred to a few thousand followers. The 30 handles also include those unaffiliated with NBC, such as director of

Table 2.4: Top 30 handles of aPPR with seed node @NBCPolitics and the teleportation constant $\alpha = 0.15$ in December 2018.

	Name	Followers	Description
1	Stephanie Palla	198	Enroll America National Regional Director...
2	Jennifer Sizemore	386	
3	Alissa Swango	441	Director of Digital Programming at @natgeo. All things ...
4	Making a Difference	670	@NBCNightlyNews' popular feature profiles ordinary ...
5	Ron Whittemore	1	
6	Svante Stockselius	3	
7	Greg Martin	1161	Political Booking Producer at @nbcnews @todayshow
8	Area Man	1	I am Area Man. I pwn your news feed.
9	CELESTIA ROBINSON	2	
10	NBC Field Notes	1390	NBC News correspondents and http://t.co/1eSopOQt8s ...
11	rob adams	2	
12	JL	2	
13	David Kelsey	1	
14	Hank Morris	1	
15	Jesse Marks	1	
16	Brayden Rainey	1	
17	child of the tiger	3	yet another activist twitter, fighting all those fun...
18	Julie Swango	4	
19	Author Dianne Kube	7	Dianne Kube is an Author with a passion, for family,...
20	Consider the Source	7	
21	Adam Edelman	2341	Political reporter @nbcnews. Wisconsin native, ...
22	Phil McCausland	2519	@NBCNews Digital reporter focused on the rural-urban...
23	Corky Siemaszko	2538	Senior Writer at NBC News Digital (former NY Daily ...
24	Sam Petulla	2588	Editor @cnnpolitics • Usually looking for datasets. ...
25	Ken Strickland	2693	NBC News Washington Bureau Chief
26	Mike Mullen	7	
27	Elyse PG	2697	White House producer @nbcnews |@USCAnnenberg alum ...
28	A. Johnson	2	Change your thoughts & you change your world. -Normal...
29	Steve Fenton	4	
30	Dobe Pitty Mami	13	

Through the aPPR vector, the top 30 handles returned to @NBCPolitics include some relevant handles (NBC's news team and their counterparts in other mainstream news organizations) and many obscure ones (handles with few followers and no profile descriptions). This results from the aPPR vector's bias toward extreme low degree and introduces noise to the clustering results.

a non-profit (Enroll America), director of digital programming at National Geographic, and @CNNPolitics' editor. All of them are professionally related to the seed node. This testifies to the applicability of aPPR for locating an idiosyncratic local cluster around a seed node. However, more than half (17) of the 30 handles are obscure and not directly followed by @NBCPolitics. The reason they appear on the list is probably that they have just one and at most a dozen followers (recall that aPPR divides by in-degree). In fact, 160 of the top 200 handles are not direct friends of @NBCPolitics; the median in-degree

Table 2.5: Top 30 handles of rPPR with seed node @NBCPolitics and the teleportation constant $\alpha = 0.15$ in December 2018.

	Name	Followers	Description
1	Stephanie Palla	198	Enroll America National Regional Director http://t.co/X6jJIE...
2	Jennifer Sizemore	386	
3	Alissa Swango	441	Director of Digital Programming at @natgeo. All things food....
4	Making a Difference	670	@NBCNightlyNews' popular feature profiles ordinary people do...
5	Greg Martin	1161	Political Booking Producer at @nbcnews @todayshow
6	NBC Field Notes	1390	NBC News correspondents and http://t.co/1eSopOQt8s reporters...
7	Adam Edelman	2341	Political reporter @nbcnews. Wisconsin native, Bestchester ...
8	Phil McCausland	2519	@NBCNews Digital reporter focused on the rural-urban divide....
9	Corky Siemaszko	2538	Senior Writer at NBC News Digital (former NY Daily News ...
10	Sam Petulla	2588	Editor @cnnpolitics • Usually looking for datasets. You can ...
11	Ken Strickland	2693	NBC News Washington Bureau Chief
12	Elyse PG	2697	White House producer @nbcnews |@USCAnnenberg alum | LA kid ...
13	Hasani Gittens	3002	Level 29 Mage. Senior News Ed. @NBCNews. Sheriff of Nattahna...
14	Scott Foster	3464	Senior Producer, Washington @NBCNEWS @TODAYshow
15	Zach Haberman	3693	Lead Breaking News Editor, @NBCNews. Previously had other jobs...
16	Emmanuelle Saliba	4004	Head of Social Media Strategy @Euronews | Launched #THECUBE ...
17	Alex Johnson	4371	News, data and analysis for @NBCNews; data geek; ...
18	Savannah Sellers	4637	News junkie. Host of NBC's "Stay Tuned" on Snapchat. Storyte...
19	NYC Clothing Bank	154	We distribute new, never-worn clothing and merchandise...
20	Shaquille Brewster	5362	@NBCNews Producer/Politics | @HowardU Alum| Journalist | Pol...
21	Joey Scarborough	6277	NBC News Social Media Editor. New York Daily News Alum. RTs ...
22	Jane C. Timm	6478	@nbcnews political reporter and fact checker. More fun than ...
23	Anthony Terrell	6827	Emmy Award winning journalist. Political observer. Covered ...
24	NBC News Videos	7838	The latest video from http://t.co/xPyvMOTEF6
25	Libby Leist	7946	Executive Producer @todayshow
26	Voices United	310	Voices United is a non profit educational organization ...
27	Social Headlines	344	Daily roundup of top social media and networking stories.
28	James Miklaszewski	337	Writer, Photographer, Editor, Director, Producer, Newshound ...
29	Courtney Kube	9494	NBC News National Security & Military Reporter...
30	Bob Corker	10042	Serving Tennesseans in the U.S. Senate

Through the rPPR vector, the top 30 handles returned to @NBCPolitics include much fewer low in-degree and obscure ones and many more moderately connected nodes that are relevant to @NBCPolitics, including its reporters and editors and media professionals from other organizations.

of the top 200 handles is merely 8 (Appendix A.4). Those handles might have ended up on the list due to a combination of luck and, more importantly, their extremely low in-degrees. In this regard, "noise" can be introduced by the aPPR vector because it prioritizes handles with extremely low in-degrees that are possibly several degrees separated from the seed node.

To reduce noise, we applied a regularization step to the aPPR vector to remove those "distant" and small nodes while preserving the close and relevant ones. In Table 3, the majority of the top 30 handles with the highest regularized aPPR (i.e., rPPR) values have three- or four-digit numbers of followers. Similar to the aPPR results, they include NBC's news crew. But the difference is that the overwhelming majority (18) of the top 30 handles work at NBC. Some handles who work for other news organizations (e.g., Sam Petulla at @cnnpolitics and Emmanuelle Saliba at @Euronews) might have previously worked at NBC or have close connection with its news team. Even the four handles that are not directly followed by @NBCPolitics are interesting – they are non-profit organizations (NYC Clothing Bank and Voices United) and news-related individual or organization (James Miklaszewski and Social Headlines). This pattern can also be observed in the top 200 handles, 72 of whom are directly followed by @NBCPolitics. The overwhelming majority of those directly followed by it are affiliated with NBC, comprising its day-to-day news team, who enjoy much less publicity than the celebrity reporters. The remaining 128 of them, who are not directly followed by @NBCPolitics, actually also include 20 NBC's journalists and staff, such as Ray Farmer (NBC News photographer) and Jim Miklaszewski (chief Pentagon correspondent for NBC News). Others are non-profits like Vets Helping Heroes and professionals from other news organizations or companies such as WSJ, NFL Network, and Microsoft, who might have worked for NBC or have close connection with it. Although there still appear to be obscure handles with few followers, they decrease significantly in number – the median in-degree of the top 200 handles is 340 (Appendix A.4), a precipitous drop from that of the top PPR handles yet not too small as compared to that of the top aPPR handles. We thus conclude that the regularized aPPR vector returns a local cluster with little noise, reflecting a seed node's close circles, either directly or indirectly related.

In order to evaluate the influence of the desired cluster size n on the results based on different PPR vectors, we compare the local clusters of PPR, aPPR, and rPPR by varying sample size. Define the *in-and-out ratio* of local cluster $C \subset V$ as the proportion of edges inside C among all edges connected to C,

$$\frac{2 \times \sum_{u,v \in C} A_{uv}}{\sum_{u \in C} d_u^{\mathrm{in}} + d_u^{\mathrm{out}}}.$$

A higher in-and-out ratio indicates a more internally connected sample. Figure 2.3 (Right) shows the effectiveness of aPPR and rPPR in producing a compact local cluster. When the sample size is bigger than 100, the connectedness of the local cluster produced by rPPR stabilizes; the greater the sample size, the more densely connected a cluster aPPR would produce. However, PPR is easily susceptible to the inclusion of popular nodes. In this case, a sharp drop of in-and-out ratio for PPR when the sample size reaches around 140 is caused by inclusions of highly popular accounts @POTUS (President Trump) and @realDonaldTrump (Donald J. Trump).

The PPR clustering is fairly robust to the choice of teleportation constant, despite the size of local cluster. To illustrate this, we also performed the same pipeline of analysis with the seed @NBCPolitics while varying the value of α (e.g., 0.05, 0.25, and 1/3) in parallel. We observed that those local clusters returned by Algorithm 2.4 all share a great portion of members in common. For example, there are 280 (93.3%) overlapping members between two targeted samples of size $n = 300$, using $\alpha = 0.15$ and 0.25 respectively. These suggest a low sensitivity to the teleportation constant (see Appendix A.2).

The left panel of Figure 2.3 depicts the behaviors of PPR, aPPR and rPPR. Each handle queried in this sampling is displayed as a dot, with y-axis representing the PPR value and x-axis the number of followers (i.e., in-degree). Top handles with the highest PPR values are above blue dashed line, which tend to concentrate on the right end of the x-axis and thus are biased toward high in-degrees. Top handles with highest aPPR values are dots to the left of the yellow dotdash line, which gather on the left end of the x-axis and thus in favor of low in-degrees. Regularized aPPR, by purple dots, excludes the very low degree nodes and very high degree nodes. As the empirical results show, these three vectors can be thought of as lenses through which we view the local structure of a given Twitter handle with varying foci, rendering high, moderate, and low in-degree blocks and serving different needs and purposes.

2.7 Discussion

This paper studies the PPR vector under the degree-corrected stochastic block model and PPR clustering in massive block model graphs. We establish some consistency results for this method, and examine its performance through

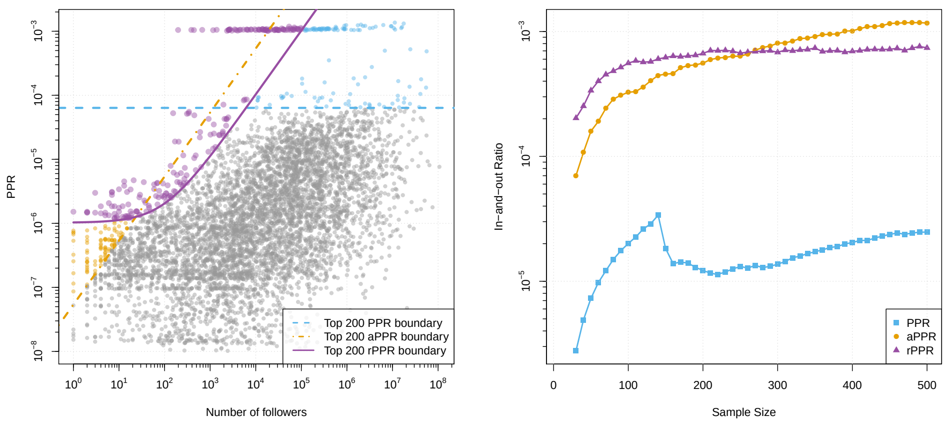


Figure 2.3: Left: an illustration of 5840 Twitter handles examined by Algorithm 2.3 and three samples of size 200 by PPR, aPPR, and rPPR. Each dot represents a user in Twitter. The blue dashed line delimits the top 200 handles by PPR vector; vertices above the line are PPR's sample. Similarly, the yellow dotdash line determines the sample returned by Algorithm 2.4 given $n = 200$; vertices above this boundary correspond to aPPR's sample. In particular, dots in purple stand for the sample of rPPR; the purple solid line shows the boundary of this sample. Right: The in-and-out ratio of local clusters identified by PPR, aPPR, and rPPR, as the sample sizes vary. A higher in-and-out ratio indicates a more internally connected cluster.

analysis of Twitter friendship graph. As shown in the results, the PPR vectors with and without adjustment have distinct properties and can be used to effectively sample a massive graph for various purposes. However, there are limitations worthy of future investigations.

In Section 2.3, we provide a representation of the PPR vector under the DC-SBM and its extension into directed graphs. The result does not impose extra structural restrictions on the model parameters, except that **B** corresponds to a strongly connected "block-wise" graph. We consider a positive definite connectivity matrix particularly so that it is intuitive to conceive the notion of local cluster. In practice (and many of our experiments, see Appendix A.2), however, a PPR-type algorithm appears to continue working for a broader range of **B** (e.g., singular or indefinite), provided that the teleportation constant is sufficiently large (e.g. $\alpha > 0.1$). It is unclear yet what is the minimum constraint needed on $\mathbf{B}$ in order for the PPR clustering to function. In addition, DC-SBM does have its limits. For example, the model fails to capture either mixed block membership or popularity features which are potentially informative in real world networks. The behavior of a PPR vector under other extensions of stochastic block model, such as mixed membership stochastic block model and popularity-adjusted block model, remains unknown (Airoldi et al., 2008; Sengupta and Chen, 2018). Future studies on the PPR vector under these models could shed further light on the PPR clustering and offer more practical guidelines on their application.

In Section 2.4, we proved the consistency of the PPR clustering, requiring the average expected node degree to grow in order of $\log N$, which hits the boundary between the theoretical guarantees and the realistic observation. In contrast, scale-free networks such as the preferential attachment model (Barabási and Albert, 1999) have finite expected node degrees. Future investigations into variants of PPR that could possibly overcome this limitation yet ensure a fine local cluster discovery would be particularly interesting and useful.

In Section 2.6, we introduce the regularized version of adjusted PPR (rPPR) vector, with a series of empirical evidence showing its efficacy in targeted sampling. While the results appear promising, theoretical guarantees for this technique remain unexplored. In order for some mathematical analyses, one may resort to the techniques used in Le et al. (2016). It is previously shown that the regularized graph Laplacian (or transition matrix) enjoys "nice" finite sample properties, which facilitate the consistency of many regularized spectral methods. It thus is reasonable conjecture that rPPR vectors are also suitable for local clustering.

An R implementation of the PPR clustering is available at author's GitHub (`https://github.com/RoheLab/aPPR`).

3 ESTIMATING GRAPH DIMENSION WITH CROSS-VALIDATED EIGENVALUES

3.1 Introduction

In network analysis, many recent community detection methods assume the number k of communities as known a priori (Karrer and Newman, 2011a; Rohe et al., 2011; Zhao et al., 2011; Amini et al., 2013; Gao et al., 2018; Xu et al., 2020). However, k is rarely available in the data. As such, the user is required to choose k. (the performance of these approaches are fundamentally associated with how well we select k.) This paper proposes a way to estimate k for a large-scale network (or graph).

We model the network to have independent random edges and have rank k expectation. Several named distributions for random graphs have the same assumptions, including Erdős-Rényi (Erdős and Rényi, 1960), Chung Lu (Chung and Lu, 2006), Stochastic Blockmodels (SBM) (Holland et al., 1983b), Degree-Corrected SBM (Karrer and Newman, 2011a), and Mixed Membership SBM (Airoldi et al., 2008). Under these random graph models, numerous methods have been proposed to estimate k (Bordenave et al., 2015; Bickel and Sarkar, 2016; Lei, 2016; Wang and Bickel, 2017; Chen and Lei, 2018; Ma et al., 2019; Le and Levina, 2019; Liu et al., 2019; Li et al., 2020; Jin et al., 2020). These methods roughly fall into one of the three categories: spectral, cross-validation, and (penalized) likelihood based approaches.

Methods based on likelihood or cross-validation are actively researched, yet the majority of them are commonly restrained by the scale of networks. Among the likelihood based approach, Wang and Bickel (2017) proposed to estimate k by solving a BIC type optimization problem, where the objective function sums the log-likelihood and the model complexity. The computation is not feasible because the likelihood contains exponentially many terms. In Ma et al. (2019), a pseudo-likelihood ratio is used to compare goodness-of-fit of models with differing ks that have been estimated using spectral clustering with regularization (Rohe et al., 2011; Qin and Rohe, 2013; Joseph and Yu, 2016;

Su et al., 2019), speeding up the computation. However, the two methods allow little node degree heterogeneity. Related to the goodness-of-fit technique, Jin et al. (2020) presents a stepwise testing based on the number of quadrilaterals in the networks. The statistic (or the counting) requires at least n^2 times of multiplication operations, regardless of the sparsity of the graph, thus is infeasible when n scales (here, n is the number of nodes in the graph). More recently, cross-validation (Picard and Cook, 1984; Arlot and Celisse, 2010) has also been adapted in the context of choosing k. For example, in Chen and Lei (2018), a block-wise node-pair splitting technique is introduced. In each fold, a block of rows of the adjacency matrix are held out from the SBM fitting (including the community memberships), then the left-out rows are used to calculated a predictive loss. In Li et al. (2020), they propose to hold out a random fraction of node-pairs, instead of nodes (thus all the incidental node-pairs). In addition, they suggest using a general low-rank matrix completion (e.g., a singular value thresholding approach (Chatterjee, 2015)) to calculate the loss on the left-out node-pairs. Theoretical conditions for not under-estimating k were established in both cross-validation based methods (Chen and Lei, 2018; Li et al., 2020). Due to the need of calculating loss on either held-out rows or scatters in the adjacency matrix regardless of sparsity, each fold requires about $O(n^2)$ computations thus is also intractable for large networks.

Spectral methods are highly scalable for estimating k in large networks, although their rigorous analyses require delicate, highly technical random matrix arguments (Ajanki et al., 2017; Benaych-Georges et al., 2019; Chakrabarty et al., 2020; Dumitriu and Zhu, 2019; Benaych-Georges et al., 2020; Hwang et al., 2020). In Bickel and Sarkar (2016); Lei (2016), hypothesis tests using the top eigenvalue or singular value of a properly normalized adjacency matrix are proposed, based on edge universality and other related results for general Wigner ensembles (Tracy and Widom, 1994; Soshnikov, 1999; Erdős et al., 2012, 2013; Alex et al., 2014). The analyses of these hypothesis tests assume dense graphs. In Liu et al. (2019), a version of the "elbow in the scree plot" approach (see, e.g., Zhu and Ghodsi (2006) for a discussion of this approach) is analyzed rigorously under the Degree-Corrected SBM, also in the dense case. For sparser graphs, the spectral properties of other matrices associated to graphs have been used to estimate k, including the non-backtracking matrix (Krzakala et al., 2013; Bordenave et al., 2015; Le and Levina, 2019) and the Bethe-Hessian matrix (Le and Levina, 2019). However, their theoretical analysis allow little node degree heterogeneity in the very sparse case.

One way to conceptualize the analytical challenges of estimating k on a more intuitive level is that sample eigenvectors (and singular vectors) are likely to "overfit" to the noise in large-scale graphs and this overfitting makes the scree plot biased. Estimators overfit when they corresponds too closely to a particular set of data in a way that does not generalize to future observations. In practice, cross-validation addresses this problem; first, compute an estimator on "fitting data," then examine the estimators performance on "testing data." For example, in ordinary least squares regression, the regression coefficients are estimated by computing the coefficients that minimize the mean squared error (MSE) on the training data. On independent testing data, the MSE for these coefficients will likely be larger. In particular, larger models tend to have a larger difference. Examining the validation MSE helps to correct the bias that is presented in the training MSE. In spectral analysis, analogously, we hope the sample eigenvectors to correlate well with a secondary, independent graph sampled from the same probability distribution as the currently observed graph. Because of the noise in real data, especially when $n \gg k$, the $(k+1)$th sample eigenvalue is often greater than 0, which blurs the "elbow" in the scree plot. Interestingly, there is no analogue to the "validation" estimate for eigenvalue estimation, which predicts how well sample eigenvectors correlate with a secondary graph's eigenspace.

In this paper, we exploit a notion of cross-validated eigenvalues as a new approach to estimating k. We demonstrate that unlike sample eigenvalues, the cross-validated eigenvalues could avoid "overfitting" the data. Under a large class of random graph models, we provide a simple procedure to compute cross-validated eigenvalues. The estimation of cross-validation eigenvalues is made possible by a simple edge splitting idea (Abbe and Sandon, 2015; Abbe et al., 2016). Basically, edges are bipartitioned at random; one set of edges is used to perform spectral analysis, and the other set of edges is for "cross-validation." This holdout approach was previously explored in the econometrics literature (Abadir et al., 2014; Lam, 2016), although the emphasis was on estimating sample covariance matrices rather than hypothesis testing of graphs. For cross-validated eigenvalues, we provide an intuitive central limit theorem, which leads to a p-value for the statistical significance of each sample eigenvector. This can be used to estimate the number of statistically useful sample eigenvectors, thus the number k of communities in a graph. In addition, we provide the consistency results for the proposed estimation, allowing weighted and very sparse graphs. Finally, through simulations and real data applications, we show that this estimator compares favorably to alternative approaches in both computational and statistical performance.

Further related work Rank estimation has also been studied in the context of certain Poisson reduced-rank models (Jentsch et al., 2020). There is also related work on bootstrapping (Snijders and Borgatti, 1999; Thompson et al., 2016; Green and Shalizi, 2017; Levin and Levina, 2019; Lin et al., 2020a), jackknife resampling (Lin et al., 2020b) and subsampling (Bhattacharyya and Bickel, 2015; Lunde and Sarkar, 2019; Naulet et al., 2021) in network analysis. In particular, in (Lunde and Sarkar, 2019), subsampling schemes are applied to the nonzero eigenvalues of the adjacency matrix under low-rank graphon models. Weak convergence results are established under some technical conditions, including sufficient edge density (i.e., average degree growing asymptotically faster than $\sqrt{n}$); simulation results also indicate that sparsity leads to poor performance for the estimators considered, especially in the case of the eigenvalues closer to the bulk.

3.2 Sample and cross-validated eigenvalues

We consider a connected multigraph $G = (V, E)$ consisting of the set of nodes $V = \{1, \ldots, n\}$ and edges E, where we allow multiple edges and self-loops. The adjacency matrix $A \in \mathbb{N}^{n \times n}$ records the number of edges between i and j in element A_{ij}. We presume that A is symmetric (i.e., edges are undirected) for simplicity. In this paper, we focus on the following random graph model.

Definition 3.1 (Poisson graph)**.** *We consider random graph models where the elements of* A *are independent Poisson random variables and*

$$\mathbb{E}(A) = U\Lambda U^T \tag{3.1}$$

for $U \in \mathbb{R}^{n\times k}$ *with orthogonal columns* $u_1, \ldots, u_k \in \mathbb{R}^n$*, and diagonal matrix* $\Lambda \in \mathbb{R}^{k\times k}$ *with positive elements down the diagonal in non-increasing order.*

Like the model above, many other models have independent edges and low-rank expectation. For example, when the edges are Bernoulli random variables, this model class has been referred to as the eigenmodel and the random dot product graph model (Hoff, 2008; Cai et al., 2016; Athreya et al., 2017). We study Poisson edges for their convenience. In the sparse graph setting, the Poisson graph can be tightly coupled to Bernoulli edge graphs (see Rohe et al. (2018) for a further discussion of these points).

In population (or expectation), define $\lambda_1, \ldots, \lambda_d \in \mathbb{R}$ to be the eigenvalues of $P = \mathbb{E}(A)$. In (3.1), the diagonal of Λ contains the leading k eigenvalues and their corresponding eigenvectors are in the columns of U. Moreover, for $j > k$, $\lambda_j = 0$. So, P has exactly k non-zero eigenvalues. Thus, we can estimate k by estimating the number of non-zero eigenvalues.

Sample eigenvalues: a poor diagnostic

The symmetric matrix $A \in \mathbb{N}^{n\times n}$ has eigenvectors $\hat{x}_1, \ldots, \hat{x}_n \in \mathbb{R}^n$ that are the solution to

$$\hat{x}_j \quad = \quad \underset{x\in\hat{S}_j}{\operatorname{argmax}} \ \ x^T A x, \tag{3.2}$$

where $\hat{S}_j = \{x \in R^n : \|x\|_2 = 1 \text{ and } x^T\hat{x}_\ell = 0 \text{ for } \ell = 1, \ldots, j-1\}$. The eigenvalues $\hat{\lambda}_j$ for $j = 1, \ldots, n$ are defined as

$$\hat{\lambda}_j = \hat{x}_j^T A \hat{x}_j.$$

Note that the quadratic form in Equation (3.2) that defines the eigenvectors and eigenvalues is equivalent to

$$x^T A x = \mathrm{tr}(x^T A x) = \mathrm{tr}(A x x^T) = \langle A, x x^T \rangle.$$

If the elements of A and xx^T were centered to have mean zero and scaled to have standard deviation one, then $\langle A, xx^T \rangle$ would be the correlation between the elements in the two matrices. As such, the maximization problem in Equation (3.2) boils down to finding a rank one matrix that has maximum "correlation" with A and the value of that "correlation" is the eigenvalue.

The most common approach to estimating the eigenvalues of P is to use a plug-in approach, i.e., estimating the *population* eigenvalues of P with the *sample* eigenvalues of A. The eigenvalues of A, $\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3, \dots$, are often plotted against their index $1, 2, 3, \dots$. This is called a scree plot and it is often used as a diagnostic to estimate k. In this scree plot, there might be a "gap" or an "elbow" at the kth eigenvalue.

However, there is a fundamental problem with the plug-in estimator for the population eigenvalues which can make the "gap" or "elbow" in the scree plot more difficult to observe. The leading eigenvalue estimates $\hat{\lambda}_1, \dots, \hat{\lambda}_k$ are asymptotically unbiased, so long as their corresponding population eigenvalues $\lambda_1, \dots, \lambda_k$ are large enough (see, e.g., Chakrabarty et al. (2020) for related results). However, when n is much larger than k, $\hat{\lambda}_{k+1}$ is a biased estimate of $\lambda_{k+1} = 0$, with $\mathbb{E}(\hat{\lambda}_{k+1}) > \lambda_{k+1} = 0$ (see, e.g., Benaych-Georges et al. (2020) for related results). So, if λ_k is not large enough, then this bias diminishes the appearance of a "gap" or "elbow" between $\hat{\lambda}_k$ and $\hat{\lambda}_{k+1}$ in the scree plot. For example, due to the bias, it could happen that $\hat{\lambda}_{k+1} > \lambda_k$.

Cross-validated eigenvalues: the signal strength of sample eigenvectors

Even if an oracle were to tell us that the population eigenvector x_j has population eigenvalue $\lambda_j \neq 0$, we should only use $\hat{x}_j$ for statistical inference if $\hat{x}_j$ has estimated some signal. In the realistic setting where the signal is not overwhelming for every sample eigenvector $\hat{x}_j$ for $j < k$, the following quantity is more methodologically useful for statistical inference. We measure the signal strength of a sample eigenvector $\hat{x}_j$ as

$$\lambda_P(\hat{x}_j) = \hat{x}_j^T P \hat{x}_j = \langle P, \hat{x}_j \hat{x}_j^T \rangle.$$

While this quantity is unknown, it can be estimated via cross-validation. The cross-validated estimator is unbiased and asymptotically normal with an estimable variance.

For population eigenvector x_j, $\lambda_P(x_j)$ is the corresponding population eigenvalue λ_j. As such, if sample eigenvector $\hat{x}_j$ is close to its population counterpart, then $\lambda_P(\hat{x}_j)$ is close to the population eigenvalue. However, if $\hat{x}_j$ is nearly orthogonal to the eigenvectors of P that have non-zero eigenvalues, then $\lambda_P(\hat{x}_j) \approx 0$. Notably, and most importantly, this can happen even if $\hat{x}_j$'s corresponding population eigenvector x_j has a non-zero eigenvalue. For example, this happens when the estimation problem is too difficult. The next proposition, proved in Section B.1, makes this intuition more rigorous.

Proposition 3.2 (Lam (2016))**.** *Let $\hat{X} \in \mathbb{R}^{n \times q}$ contain the leading q sample eigenvectors $\hat{x}_1, \ldots, \hat{x}_q$ in its columns. The solution to*

$$\min_{\Gamma \text{ is diagonal}} \|P - \hat{X} \Gamma \hat{X}^T\|_F$$

contains $\lambda_P(\hat{x}_1), \ldots, \lambda_P(\hat{x}_q)$ down the diagonal.

The notion of $\lambda_P(\hat{x}_j)$ was originally proposed and studied in Abadir et al. (2014); Lam (2016) for optimal estimation of eigenvalue shrinkage. In this section, we develop a statistical testing procedure for the null hypothesis

$$H_0 : \lambda_P(\hat{x}_j) = 0 \tag{3.3}$$

conditionally on $\hat{x}_j$.

The "plug-in" estimator for $\lambda_P(\hat{x}_j) = \hat{x}_j^T P \hat{x}_j$ replaces P with A; $\hat{x}_j^T A \hat{x}_j = \hat{\lambda}_j$. The difficulty of studying this quantity is that $\hat{x}_j$ is determined by A. This dependence makes $\hat{\lambda}_j$ and its distributional relationship to λ_P or λ_j difficult to study. The rest of this section shows how we can estimate this quantity via cross validation instead; we will split the graph into two so that we can create an independent test adjacency matrix A_{test}. The proposed estimator is unbiased and satisfies a central limit theorem with an estimable variance. As such, we can test the null hypothesis (3.3).

Edge splitting and a central limit theorem

```
Input: Adjacency matrix $A \in \mathbb{N}^{n\times n}$ and edge splitting probability
        $\varepsilon \in (0, 1)$.
Procedure ES$(A, \varepsilon)$:
    1. Convert $A$ into $G = (V, E)$, where $\{i, j\}$ is repeated in the edge set
       $E$ potentially more than once if $A_{ij} > 1$.
    2. Initiate $\tilde{E}_{\text{test}}$ and $\tilde{E}$, two empty edge sets on $V$.
    3. for each copy of edge $\{i, j\} \in E$ do
           assign it to $\tilde{E}_{\text{test}}$ with probability $\varepsilon$. Otherwise, assign it to $\tilde{E}$.
    4. Convert $(V, \tilde{E}_{\text{test}})$ into an adjacency matrix $\tilde{A}_{\text{test}} \in \mathbb{N}^{n\times n}$ and
       $(V, \tilde{E})$ into an adjacency matrix $\tilde{A} \in \mathbb{N}^{n\times n}$
Output: $\tilde{A}$ and $\tilde{A}_{\text{test}}$.
```

Algorithm 3.1: Edge splitting

Algorithm 3.1 splits the edges of a graph into two graphs and outputs the two adjacency matrices $\tilde{A}$ and $\tilde{A}_{\text{test}}$. The splitting is administered by the edge splitting probability $\varepsilon \in (0, 1)$. Under the Poisson random graph model, the matrices $\tilde{A}$ and $\tilde{A}_{\text{test}}$ are conveniently independent and the spectral properties of $P = \mathbb{E}(A)$, $\mathbb{E}(\widetilde{A})$ and $\mathbb{E}(\widetilde{A}_{\text{test}})$ are closely related. These key observations, which are spelled out in the next proposition proved in Appendix B.1, will enable the estimation of λ_P.

Proposition 3.3. *If $A \in \mathbb{N}^{n\times n}$ is generated from the Poisson random graph model and $\tilde{A}$ and $\tilde{A}_{\text{test}}$ are generated from splitting A with the probability ε as defined above, then*

1. *$\tilde{A}$ and $\tilde{A}_{\text{test}}$ are independent Poisson random graphs.*

2. *The population eigenvectors of $A, \tilde{A}$, and $\tilde{A}_{\text{test}}$ are identical.*

3. *If* λ_j *is a population eigenvalue of* A*, then* $(1-\varepsilon)\lambda_j$ *is a population eigenvalue of* $\tilde{A}$ *and* $\varepsilon\lambda_j$ *is a population eigenvalue of* $\tilde{A}_{\text{test}}$*.*

Because $\tilde{A}$ and $\tilde{A}_{\text{test}}$ are independent, we can compute an eigenvector $\hat{x}_j$ on $\tilde{A}$ and test its signal strength on $\tilde{A}_{\text{test}}$ via

$$\lambda_{\text{test}}(\hat{x}_j) = \hat{x}_j^T \tilde{A}_{\text{test}} \hat{x}_j.$$

Conditionally on $\hat{x}_j$, we then get

$$\mathbb{E}\lambda_{\text{test}}(\hat{x}_j) = \hat{x}_j^T \mathbb{E}(\tilde{A}_{\text{test}})\hat{x}_j = \varepsilon \hat{x}_j^T P \hat{x}_j = \varepsilon \lambda_P(\hat{x}_j).$$

Most importantly, if $\lambda_P(\hat{x}_j) = 0$, then $\mathbb{E}\lambda_{\text{test}}(\hat{x}_j) = 0$. Not only is $\lambda_{\text{test}}(\hat{x}_j)/\varepsilon$ an unbiased estimator of $\lambda_P(\hat{x}_j)$, the following central limit theorem (CLT) shows that it is also asymptotically normal, with an easy to estimate standard error.

To state the theorem formally, we consider a sequence of random adjacency matrices $B^{(n)} \in \mathbb{N}^{n\times n}$ from Poisson random graphs with mean matrix $Q^{(n)} \in \mathbb{R}^{n\times n}$ satisfying $\max_{ij} Q_{ij}^{(n)} \leqslant 1$, and a sequence of unit vectors $x^{(n)} \in \mathbb{R}^n$. To simplify the notation, we suppress the explicit dependence on n. We will impose the following delocalization condition on x:

$$\|x\|_\infty^2 = o(\sigma), \tag{3.4}$$

where

$$\sigma^2 = (x^2)^T Q (x^2),$$

with x^2 being the vector x with entries squared. Similarly, we also define

$$\hat{\sigma}^2 = (x^2)^T B (x^2).$$

In the next section, we will apply the theorem to $B := \tilde{A}_{\text{test}}$, $Q := \varepsilon P$ and x an eigenvector of $\tilde{A}$.

Theorem 3.4 (CLT for cross-validated eigenvalue)**.** *Let* B*,* Q*,* σ *and* $\hat{\sigma}$ *be as* *above. Assume that* x *satisfies Condition* (3.4)*. Then,*

$$\frac{\lambda_B(x) - \lambda_Q(x)}{\hat{\sigma}} \Rightarrow N(0,1). \tag{3.5}$$

The proof of Theorem 3.4 is in Appendix B.1.

Remark 3.5. *The theorem assumes a Poisson graph. In fact, we can similarly consider the case when edges are unweighted. In fact, suppose the edges are Bernoulli instead of Poisson, then so long as the graph is sparse, the formulas for* σ^2 *and* $\hat{\sigma}^2$ *can still be used. However,* $\tilde{A}$ *and* $\tilde{A}_{\text{test}}$ *are no longer perfectly independent; the reason is that if edge* (i,j) *appears in* A*, then only* $\tilde{A}$ *or* $\tilde{A}_{\text{test}}$ *can inherit the edge. In simulations, the procedure still appears to perform well.*

Remark 3.6. *Regarding the delocalization condition* (3.4)*, when all entries of* Q *are of the same order* $\rho = o(1)$*, then* $\sigma = \Theta(\rho^{1/2})$ *and the condition boils down to* $\|x\|_\infty = o(\rho^{1/4})$*. In Appendix B.1* (*Corollary B.2*)*, we discuss a sufficient condition for* $\|x\|_\infty^2 = o(\sigma)$ *to hold in terms of* m *and the expected number of edges in* B*.*

3.3 Cross-validated eigenvalue estimation

In this section, we use Theorem 3.4 to test the null hypothesis $H_0 : \lambda(\hat{x}_j) = 0$, where $\hat{x}_j$ is an eigenvector of A.

The algorithm

We propose an eigenvalue cross-validation algorithm to estimate whether each eigenvector of A is correlated with the population eigenspace. The algorithm reports a p-value for each eigenvector, which can then be used to determine k. For input, in addition to the splitting probability ε, the algorithm takes two more parameters: (i) the maximum number $k_{\max}$ of eigenvectors to consider and (ii) the significance level α. We describe the algorithm for an undirected unweighted graph with the adjacency matrix $A \in \{0,1\}^{n \times n}$ in Algorithm 3.2; the algorithms for directed or Poisson graphs can be devised analogously.

A few remarks on the theory and the implementation are in order.

```
Input: Adjacency matrix $A \in \mathbb{N}^{n \times n}$, edge splitting probability
        $\varepsilon \in (0, 1)$, and significance level $\alpha \in (0, 1)$
Procedure EigCV($A, \varepsilon, k_{\max}$):
    1. Obtain $\tilde{A}$, $\tilde{A}_{\text{test}} \leftarrow$ES($A, \varepsilon$) from splitting $A$.  // Algorithm 3.1
    2. (Optional) Substitute $A$, $\tilde{A}$ and $\tilde{A}_{\text{test}}$ by their regularized
      symmetric Laplacians as follows. For a symmetric adjacency
      matrix $M \in \{0, 1\}^{n \times n}$, the regularized symmetric Laplacian
      $L \in \mathbb{R}^{n \times n}$ is define with

          $$L_{ij} = \frac{M_{ij}}{\sqrt{(d_i + \tau)(d_j + \tau)}},$$

      where $d_i = \sum_j M_{ij}$ and $\tau = \sum_i d_i / n$, for $i, j = 1, 2, ..., n$.
    3. for $k = 2, \ldots, k_{\max}$ do
        compute the test statistic

          $$T_k = \frac{\tilde{\lambda}_{\text{test}}(\tilde{x}_k)}{\tilde{\sigma}_k},$$

        where $\tilde{\lambda}_{\text{test}}(x) = x^T \tilde{A}_{\text{test}} x$, and $\tilde{\sigma}_k = \sqrt{\varepsilon (\tilde{x}_k^2)^T A \tilde{x}_k^2}$ is the
        standard error evaluated using the full graph. Here, $\tilde{x}_k^2 \in \mathbb{R}^n$
        is the vector $\tilde{x}_k$ with each element squared.
    4. Compute the one-sided p-value $p_k = 1 - \Phi(T_k)$, where $\Phi$ is the
      cumulative distribution function of the standard normal
      distribution.
Output: The graph dimensionality estimate:
        $\operatorname{argmin}_{k \leqslant k_{\max}} \{p_k \geqslant \alpha\} - 1$.
```

Algorithm 3.2: Eigenvalue cross-validation

Remark 3.7. *We further study, in Appendix B.2, the appropriate range of ε in order to satisfy given type I and II error of the test. We use a Bonferroni correction for the k_{max} tests and perform a power calculation for a Stochastic Blockmodel near the reconstruction threshold (Mossel et al., 2015). As with any power calculation, a user may specify different inputs and find a different value of ε. Throughout the simulations and data applications, we default ε to 0.05 for simplicity.*

Remark 3.8. *In practice, we recommend to use the regularized Laplacian (Step 2), which helps reduces localization of eigenvectors (Le et al., 2017; Zhang and Rohe, 2018). Our theory focuses on the direct use of A. In this case, it is necessary to check the delocalization of eigenvectors (see Section 3.3 below). In addition, we allow repeats of Step 1-4 (e.g., for 10 times) and take the average of T_k across the replicates, which helps to reduce the randomness caused by the edge splitting. Across our limited experiments, we found that these algorithmic options lead to more accurate estimations of graph dimensionality.*

Remark 3.9. *If the p-values p_k are used to select eigenvectors, then only the eigenvectors $\tilde{x}_k$ should be used (not $\hat{x}_k$). This is because the p-value p_k is only associated with the eigenvector $\tilde{x}_k$. It is tempting to compute the eigenvectors of A or L with all of the edges and then give jth eigenvector $\hat{x}_k$ the p-values p_k. However, when the left-out edges are also used to compute the eigenvectors, this alters the eigenvectors. In addition to slightly changing the elements of the eigenvectors, it is common for the order of the eigenvectors to also change. Or, for the new eigenvectors to be a more general rotation of the subsampled eigenvectors. It is an area for future research to understand if and how the p-values can be extended. By making ε small we can ensure that the subsampled eigenvectors $\tilde{x}_k$ are nearly as good as $\hat{x}_k$.*

Statistical consistency

This section states a consistency result for a modified version of the algorithm stated in Appendix B.1. The main modification is the addition of a delocalization test.

We will make some further assumptions. Let $P = \rho_n P^0$, where $0 < \rho_n < 1$ controls the sparsity of the network, and $P^0 = U\Lambda^0 U^T$ is a matrix of rank k with

$P^0_{ij} \leqslant 1$ for all i, j. Here, $\Lambda^0 = \mathrm{diag}(\lambda^0_1, \cdots, \lambda^0_k)$ is the diagonal matrix of its non-increasing eigenvalues, and $U = (u_1, \cdots, u_k)$ contains the corresponding eigenvectors. We first consider the signal strength in the population adjacency matrix. The magnitude of the leading eigenvalues characterize the useful signal in the data; only if they are sufficiently large is it possible to identify them from a finite graph sample. As such, the first assumption requires that the leading eigenvalues of the population graph are of sufficient and comparable magnitude. We also include necessary assumptions on the sparsity of the graph.

Assumption 3.10 (Signal strength and sparsity)**.** *We assume that there exist positive constants* ψ_1, ψ'_1 *such that*

$$\kappa := \lambda^0_1/\lambda^0_K \in (0, \psi_1), \qquad \lambda^0_1 \geqslant \psi'_1 n.$$

In addition, we assume that $P^0_{ij} \leqslant 1$ *for all* i, j *and that the network sparsity satisfies* $c_0 \frac{\log^{\xi_0} n}{n} \leqslant \rho_n \leqslant c'_0 n^{-\xi_1}$*, for some constants* $\xi_0 > 2$*,* $\xi_1 \in (0, 1)$*,* $c_0, c'_0 > 0$*.*

Observe that Assumption 3.10 implies in particular that $\psi_1^{-1}\psi'_1 n\rho_n \leqslant \lambda_K \leqslant \lambda_1 \leqslant n\rho_n$ since $\lambda_1 \leqslant \mathrm{tr}(P) \leqslant n\rho_n$. Assumption 3.10 is less strict than the assumptions in Li et al. (2020). This is because we do not require a minimum gap between distinct eigenvalues, which is hard to satisfy in reality.

Next, we consider a property of the population eigenvectors. The notion of coherence was previously introduced by Candés & Recht (Candès and Recht, 2009). Under the parametrization of Assumption 3.10, the coherence of U is defined as

$$\mu(U) = \max_{i \in [n]} \frac{n}{K} \|U^T e_i\|^2 = \frac{n}{K} \|U\|^2_{2,\infty}.$$

A lower coherence indicates that the population eigenvectors are more spread-out—that is, they are not concentrated on a few coordinates.

Assumption 3.11 (Coherence)**.** *We assume* $\mu(U) \leqslant \mu_0$*, for some constant* $\mu_0 > 1$*.*

Our main theoretical result asserts the consistency of our cross-validated eigenvalue estimator for estimating K. The proof of Theorem 3.12 is in Appendix B.1.

Theorem 3.12 (Consistency). *Suppose $A \in \mathbb{R}^{n \times n}$ is a Poisson graph satisfying Assumptions 3.10 and 3.11. Let K be the true latent space dimension, and let $\hat{K}$ be the output of Algorithm B.1 (see Appendix B.1) with edge splitting probability ε. Then*

$$\mathbb{P}\left(\hat{K} = K\right) \to 1 \quad \textit{as} \quad n \to \infty.$$

3.4 Simulation and real data application

This section compares the proposed method (EigCV) with some existing graph dimensionality estimators using both simulated and real graph data. For this, we selected (1) BHMC, a spectral method based on the Bethe-Hessian matrix with correction (Le and Levina, 2019); (2) LR, a likelihood ratio method adapting a Bayesian information criterion (Wang and Bickel, 2017); (3) ECV, an edge cross-validation method with an area under the curve criterion (Li et al., 2020); (4) NCV, a node cross-validation using an binomial deviance criterion (Chen and Lei, 2018); and (5) StGoF (with $\alpha = 0.05$), a stepwise goodness-of-fit estimate (Jin et al., 2020). We performed all computations in R. For (1)-(4), we invoked the R package `randnet`, and for (5), we implemented the original Matlab code (shared by the authors) in R.[1]

Numerical experiments

This section presents several simulation studies that compare our method with other approaches to graph dimensionality. We set the graph splitting probability ε to 0.05 and set the significance level cut-off at $\alpha = 0.05$. We sampled random graphs with $n = 2,000$ nodes and $k = 10$ blocks from the degree-corrected stochastic block model (DCSBM). Specifically, for any

[1] We provide an R package that contains the proposed method, `eigcv`, and an implementation of StGoF, `stgof`. The source code is is available at `https://github.com/RoheLab/gdim`.

$i, j = 1, 2, ..., n,$

$$A_{ij} \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\theta_i \theta_j B_{z(i)z(j)}\right),$$

where $z(i) \in \{1, 2, ..., k\}$ is the block membership of node i, and $B \in \mathbb{R}^{k \times k}$ is the block connectivity matrix, with $B_{ii} = 0.28$ and $B_{ij} = 0.08$ for $i, j = 1, 2, ..., k$, and $\theta_i > 0$ is the degree parameters of node i. We investigated the effects of degree heterogeneity by drawing θ_i's from three distribution (before scaling to unit sum): (i) a point mass distribution, (ii) an Exponential distribution with rate 5, (iii) a Pareto distribution with location parameter 0.5 and dispersion parameter 5. From (i) to (iii), the node degrees become more heterogeneous. Finally, to examine the effects of sparsity, we chose the expected average node degree in $\{25, 30, ..., 60\}$. For each simulation setting, we evaluated all methods 100 times.

Figure 3.1 displays the accuracy of all graph dimensionality methods. Here, the accuracy is the fraction of times an estimator successfully identified the true underlying graph dimensionality (which is 10).[2] From the results, both BHMC and ECV offered satisfactory estimation when the graph is degree-homogeneous and the average degree becomes sufficiently large, while they were affected drastically by the existence of degree heterogeneity. The LR estimate was affected by degree heterogeneity as well (although less than BHMC or ECV) and also required a relatively large average node degree to estimate the graph dimensionality. The NCV methods failed to estimate the graph dimensionality under most settings. The StGoF estimate worked better for degree-heterogeneous graphs but required a larger average node degree for accuracy. It is also worth pointing out that, the LR and StGoF methods tended to over-estimate the graph dimensionality when the average degree is large, especially for the power-law graphs (see supporting Figure B.2). Finally, our method provided a much more accurate dimensionality estimate overall, requiring smaller average node degree and allowing degree heterogeneity. In addition, our testing approach also enjoys a strong advantage of reduced computational cost. To show this, Figure 3.2 depicts the average runtime

[2] Besides comparison of accuracy, we also compared the deviation of the estimation by each method, for which similar results hold consistently (see supporting Figure B.2).

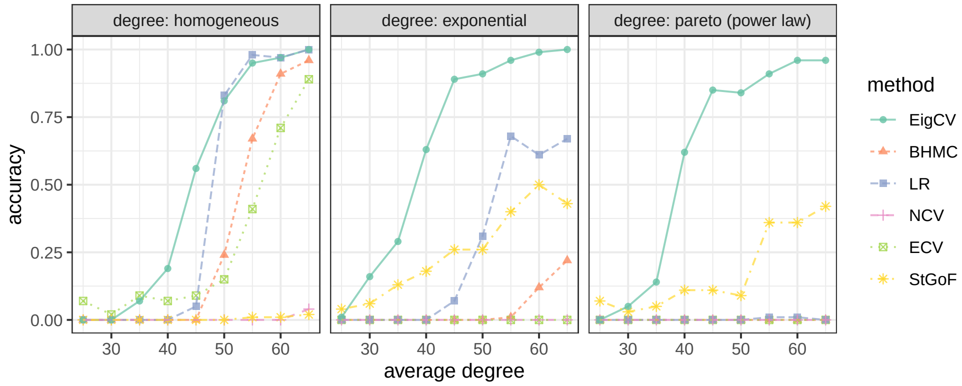


Figure 3.1: Comparison of accuracy for different graph dimensionality estimates under the DCSBM. The panel strips on the top indicate the node degree distribution used. Within each panel, each colored line depicts the relative error of each estimation method as the average node degree increases. Each point on the lines are averaged across 100 repeated experiments.

for each method. It can be seen that the proposed method and BHMC are faster than competing methods by several magnitudes. The computational complexity of each StGoF iteration (or test) is at least $O(n^2)$, regardless of whether the graph is sparse or not. Consequently, StGoF requires the longest runtime.

Email network

A real data network was generated using email data within a large European research institution, with each node representing one of the 1005 core members (Leskovec et al., 2007). There is an edge from node i to node j, if i sent at least one email to j. The dataset also contains 42 "ground-truth" community memberships of the nodes. That is, each individual belongs to exactly one of 42 departments at the research institute. For simplicity, we removed the 14 small departments that consist of less than 10 members (see supporting Table B.1 for similar results without the filter). This resulted in a directed and unweighted network with a total of 936 nodes from 28 communities.

We applied the graph dimensionality methods to estimate the number of

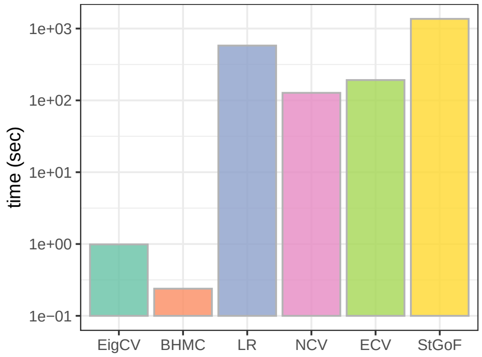


Figure 3.2: Comparison of runtime for the different graph dimensionality methods. Each colored bar indicates the runtime of applying each method on a DCSBM graph with 2000 nodes and 10 blocks. The maximum graph dimensionality is set to 15 for all methods. The runtime was averaged across 100 repeated experiments.

clusters in the network. For the randomized methods (including ECV, NCV, and our proposed method), we ran them 25 times and report the mean and standard deviation of the estimates. For the methods that report a p-value (including StGoF and our proposed method), we use a significance level of $\alpha = 0.01$, followed by a multiplicity correction using the procedure of Benjamini & Hochberg (Benjamini and Hochberg, 1995). We set $k_{\max} = 50$. Finally, we chose the splitting probability to be 0.05, as the network is sparse with an average node degree being 23.5. Table 3.1 lists the inferences made by each method. As shown, our method provided an estimate that is close to the true number of departments within the institute. BHMC, LR, NCV, and ECV all estimated small numbers of clusters, while StGoF went significant larger ($\geqslant 50$). These observations were consistent with the simulation results (see supporting Figure B.2). Among all the others, only the proposed method provided a close estimate (≈ 28) to the true number of departments. Similarly to the simulation results, the BHMC method and our method are more computationally efficient, with much shorter runtime than the others.

Table 3.1: Comparison of graph dimensionality estimates using the email network among members in a large European research institution. Each members belongs to one of 28 departments.

Method	Estimate (mean)	Runtime (second)
EigCV	28.3	0.68
BHMC	14	0.02
LR	17	85.19
NCV	6.5	204.97
ECV	16.5	41.07
StGoF	$\geqslant 50$	397.47

3.5 Discussion

In this paper, we proposed a concept of cross-validated eigenvalues to estimate the number k of communities in a graph. Through edge splitting and thanks to a simple central limit theorem, the estimation of cross-validated eigenvalues is efficient for very large graph. The paper also provides theoretical justification showing that the estimator is consistent in finite graphs. Our simulations and empirical data application validate the theory and further demonstrate the efficacy of the proposed method.

The problem of estimating k is also related to selecting the number of factors (resp. components) in factor analysis (resp. principal component analysis), although usually studied under different statistical models, such as the factor model (resp. the spiked covariance model) (Donoho et al., 2018; Rohe and Zeng, 2020). Rank estimation has also been studied in the context of certain Poisson reduced-rank models (Jentsch et al., 2020). Besides the popular parallel analysis of Horn's (Horn, 1965), there are some encouraging methodological developments recently (Dobriban, 2020; Dobriban and Owen, 2019; Hong et al., 2020). It is of future interest to investigate whether the introduced cross-validation eigenvalue estimator can be adapted under alternative models.

4 A NEW BASIS FOR SPARSE PRINCIPAL COMPONENT ANALYSIS

4.1 Introduction

Principal component analysis (PCA), introduced in the early 20th century (Pearson, 1901; Hotelling, 1933), is one of the most prevalent tools in exploratory multivariate data analysis. PCA projects higher-dimensional data into a lower-dimensional space that is spanned by some uncorrelated principal components (PCs), with the vast majority of the variance in the data kept. It is, however, commonly conceived that PCs are difficult to interpret (e.g., Jeffers, 1967), as each PC is a linear combination of many, if not all, original variables. To remedy such disadvantage, sparse PCA estimates "sparse" PCs, each of which consists of a small subset of original variables (Zou and Xue, 2018).

Sparse PCA is originally formulated as an optimization problem over the loading coefficients with a cardinality constraint. Such non-convex constraint results in an NP-hard problem in the strong sense (Tillmann and Pfetsch, 2014). In order to circumvent the obstacle, various methods have been proposed, such as the iconic regression-based approach by Zou et al. (2006), a convex relaxation to semidefinite programming (d'Aspremont et al., 2007), the penalized matrix decomposition framework of Witten et al. (2009), and the generalized power method due to Journée et al. (2010). More recently, theoretical developments of sparse PCA have covered the consistency (Johnstone and Lu, 2009; Shen et al., 2013), variable selection properties (Amini and Wainwright, 2009), rates of convergence, the minimaxity over some Gaussian or sub-Gaussian classes (Vu and Lei, 2013; Cai et al., 2013), and the statistical-computational trade-offs under the restricted covariance concentration condition (Berthet and Rigollet, 2013; Wang et al., 2016).

Despite the extensive literature of sparse PCA, there are two enigmas. First, sparse PCA often explains far less variance in the data than PCA does (Figure 4.1). While this may appear to be a trade-off for sparsity, our results show that a substantial improvement is possible. Second, the most common formulations of sparse PCA only estimate a single component at a time and thus rely on a matrix deflation after estimating each component. This deflation

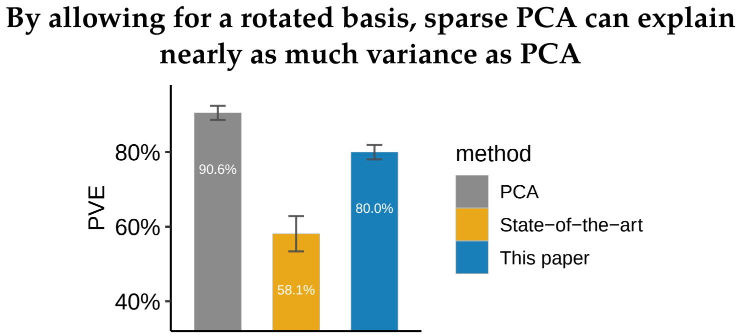


Figure 4.1: Comparison of the explanatory power of sparse PCA methods. Each bar shows the proportion of variance explained (PVE) by 16 PCs. For two sparse PCA methods, an error bar (based on the three-sigma rule) depicts the variation of PVE over 30 replicates. More details about the simulated data and settings (e.g., sparsity constraints) are described in Section 4.4.

entails complications of multiple tuning parameters, non-orthogonality, and sub-optimality (Mackey, 2008). Identifiability and consistency present more subtle issues; there is no reason to assume a priori distinct eigenvalues or that the gaps between the eigenvalues are small (Vu et al., 2013). Estimating the subspace spanned by multiple sparse PCs at once overcomes this dilemma (Vu et al., 2013).

There are two distinct notions of subspace sparsity: row sparsity and column sparsity (Vu and Lei, 2013). Contemporary approaches to sparse PCA primarily focus on row sparsity, which implies that the eigenvectors of the covariance matrix themselves are sparse (e.g., Moghaddam et al., 2006). The second notion, column sparsity, is an alternative. A column sparse subspace "*is one which has some orthogonal basis consisting of sparse vectors. This means that the choice of basis is crucial; the existence of a sparse basis is an implicit assumption behind the frequent use of rotation techniques by practitioners to help interpret principal components*" (Vu and Lei, 2013). Row sparsity is the most prevalent notion of sparsity used in contemporary sparse PCA, yet it does not appear to describe many contemporary parametric multivariate models; conversely, many contemporary parametric models in multivariate statistics can be estimated with the sparse PCA approaches that can identify column sparsity (Rohe and

Zeng, 2020).

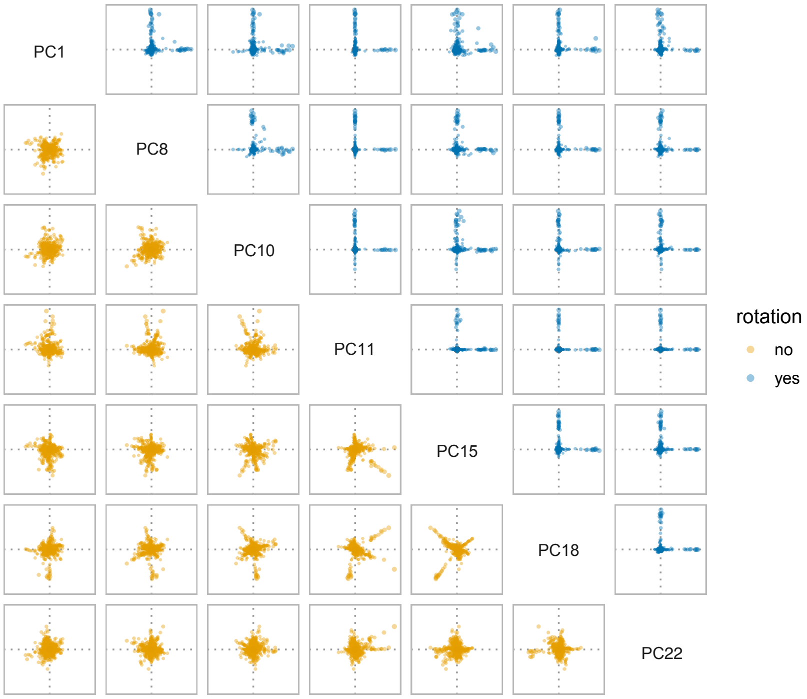
In high-dimensional regression, sparse penalties such as the Lasso resolve an invariance; there is an entire space of solutions b which exactly interpolate the data $Y = Xb$ and presuming that the solution b is sparse can make the solution unique. Interestingly, there is no analogue to "sparsity resolving an invariance" for the estimation of row sparse subspace, but there is a very clear analogue in estimating column sparse subspace; the basis is determined by the one that provides the most sparse representation of data.

Our contributions

In this work, we propose a new method, sparse component analysis (SCA), to estimate multiple PCs that are column sparse. The column sparsity is achieved by allowing an orthogonal rotation to PCs prior to imposing any sparsity constraints. The algorithm is motivated by two facts. First, an orthogonal rotation does not affect the total variance explained by a given set of PCs. Second, by choosing the orthogonal rotation carefully, PCs can be aligned closely with the coordinate axes, making them approximately sparse (Figure 4.2). This technique has been commonly adapted in factor analysis, a close cousin of PCA (Thurstone, 1931; Kaiser, 1960; Jolliffe, 1995). For example, the varimax rotation (Kaiser, 1958) is a popular choice in the psychology literature. SCA incorporates the orthogonal rotation and sparsity constraints to find the sparse and orthogonal basis in a subspace (i.e., column sparse PCs). We show in Proposition 4.2 that

> *column sparse PCs can explain more variance in the data than row sparse PCs.*

We validated this with numerical experiments. Additionally, the simulations suggest that SCA is more stable and robust across tuning parameters than existing sparse PCA methods. Our framework of SCA generalizes naturally to a two-way analysis of a data matrix for simultaneous row and column dimensionality reductions. For this, we introduce a low-rank matrix approximation method called sparse matrix approximation (SMA). The SMA builds on the



The same data in seven dimensions, before and after rotation. After the sparse rotation, each PC uses only a small subset of the original variables.

Figure 4.2: Loadings of seven principal components (PCs) from a large scale social network matrix. Each (off-diagonal) panel shows the loadings of two PCs on the original variables (displayed as points). The lower-triangular panels (yellow) depict the PCs before a rotation. The upper-triangular panels (blue) display the PCs after an orthogonal rotation. The PCs before and after the rotation have no special or corresponding relationship. In each panel, two perpendicular dotted lines (grey) indicate the coordinate axes. See Section 4.5 for details about the data analyzed.

penalized matrix decomposition previously proposed by Witten et al. (2009). Furthermore, the SMA provides a unified view of sparse PCA and other modern multivariate data analysis, including sparse independent component analysis (see, e.g., Comon, 1994). Finally, we demonstrate our sparse PCA methods with various high-dimensional data applications, including sparse coding of images, blind source separation, analysis of single-cell transcriptome data, and large-scale clustering of social networks. We find compelling evidence for the usefulness of our approach, despite concerns about the consistency of PCA in high-dimensions.

Organization

The rest of this paper goes as follows. Section 4.2 describes the methods. Section 4.3 compares SCA to existing methods. Section 4.4 compares different sparse PCA methods using simulated data. Section 4.5 applies SCA to several high-dimensional datasets. Section 4.6 concludes the paper with some discussions.

Notations

In this paper, we discuss the *entrywise* matrix norm only. For any matrix $A \in \mathbb{R}^{m \times n}$, its entrywise ℓ_p-norm is defined as $\|A\|_{p,p} = (\sum_{i=1}^{m} \sum_{j=1}^{n} |A_{ij}|^p)^{1/p}$. For simplicity, we also use the notation $\|A\|_p$ for entrywise norm, rather than the norm induced by a vector norm. In particular, the Frobenius norm (or the Hilbert-Schmidt norm) is then an alias of entrywise ℓ_2-norm, $\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2} = \|A\|_2$. Throughout, the following sets of matrices are frequently considered. $\mathscr{U}(n) = \{U \in \mathbb{R}^{n \times n} \mid U^T U = U U^T = I_n\}$ denotes all orthogonal (unitary) matrices in $\mathbb{R}^n$. $\mathscr{V}(n,k) = \{V \in \mathbb{R}^{n \times k} \mid V^T V = I_k\}$ represents the Stiefel manifold in $\mathbb{R}^n$, and $\mathscr{B}(n,k) = \{V \in \mathbb{R}^{n \times k} \mid V^T V \preceq I_k\}$ is its convex hull (Gallivan and Absil, 2010).

4.2 The methods

Consider the data matrix $X \in \mathbb{R}^{n \times p}$ of n observations (or samples) on p variables. Without loss of generality, we assume that each column of X is centered (i.e. mean-zero) unless otherwise noted. PCA finds some (say k) uncorrelated linear transformations of the original variables such that after the linear transformations, the most variance is kept. That is,

$$\underset{Y}{\text{maximize}} \quad \|XY\|_F \quad \text{subject to } \; Y \in \mathscr{V}(p, k), \tag{4.1}$$

where the feasible set is the Stiefel manifold, $\mathscr{V}(p, k) = \{Y \in \mathbb{R}^{p \times k} \mid Y^T Y = I_k\}$. The jth PC is the linear combination of original variables whose coefficients are in the jth columns of Y. The coefficients are often called *loadings* (or loading coefficients). Note that loadings are usually non-zero (i.e., Y is usually not sparse). The transformed data $S = XY \in \mathbb{R}^{n \times k}$ contains the *scores*. That is, S_{ij} is the score of the ith sample on the jth PC.

In PCA, PCs are often defined sequentially. That is, in order to find the kth PCs, we fix the previous $k - 1$ PCs and solve (4.1); repeat this for $k = 1, 2, ...$ in order. Such definition ensures the first k PCs together always explain the most variance in the data. By contrast, for sparse PCA, we reason in the following that it is sufficient to solve the optimization problem for all PCs at once. Note first that the solution to (4.1) is a subspace, because if Y^* is an optimizer of (4.1), then for any orthogonal matrix $R \in \mathscr{U}(k)$, $Y^* R$ is also an optimizer. The solution to (4.1) being a rotation-invariant subspace is desirable because it allows a sparsity-enabling orthogonal rotation to any given solution. Importantly, such rotation exists under the assumption of *column sparsity* (see Section 4.2 and Vu and Lei, 2013). We thereby propose a new method for sparse PCA.

Sparse component analysis

For sparse PCA, we impose an ℓ_1-norm constraint[1] on the loadings and formulate the following minimization of matrix reconstruction error:

$$\begin{aligned} \underset{Z,B,Y}{\text{minimize}} \quad & \left\|X - ZBY^T\right\|_F \\ \text{subject to} \quad & Z \in \mathscr{V}(n,k),\ Y \in \mathscr{V}(p,k),\ \|Y\|_1 \leqslant \gamma, \end{aligned} \tag{4.2}$$

where $\gamma > 0$ is the sparsity controlling parameter, and the columns of Y are PC loadings. ZBY^T is an approximation of X.

The fundamental difference between formulation (4.2) and previous sparse PCA formulations is that the middle B matrix is not necessarily diagonal. Compared to the diagonal B case, this added flexibility has two merits—(i) it allows PCs to be column sparse and (ii) it allows sparse PCs to explain more variance in the data.

Column sparsity

Our formulation (4.2) presumes the PCs are column sparse. That is, given the subspace of PCs, there exists a orthogonal rotation, such that after the rotation, the PCs are approximately sparse.

Let UDV^T be the low-rank singular value decomposition (SVD) of X, where $U \in \mathscr{V}(n,k)$ and $V \in \mathscr{V}(p,k)$ contain singular vectors, and $D \in \mathbb{R}^{k \times k}$ is a diagonal matrix with the diagonal entries in decreasing order, and $k \leqslant \min\{n,p\}$ is the rank. For any two orthogonal matrices $O, R \in \mathscr{U}(k)$, define $Z = UO$, $B = O^TDR$, and $Y = VR$. With these definitions,

$$X \approx UDV^T = (UO)(O^TDR)(VR)^T = ZBY^T.$$

As such, ZBY^T approximates X as well as UDV^T. In particular, the middle B matrix is not diagonal because it absorbs the orthogonal matrices (O and R). Z and Y are orthogonally rotated from U and V, and both matrices still have orthogonal columns. Hence, by imposing an ℓ_1-norm constraint on Y to make

[1] The ℓ_1-norm constraint could be replaced by other sparsity constraints, e.g., the ℓ_0-norm analogue.

it approximately sparse, we presume that there exists at least one orthogonal basis for the column space of V (i.e., the eigenvectors' subspace), which is not necessarily the original coordinate basis, such that the PCs are sparse under that basis.

Remark 4.1. *The formulation of SCA does not implicitly order sparse PCs. This is because permuting the columns of* Y*, which can be absorbed by the orthogonal matrix* R*, does not change the approximation of* ZBY^T*. As such, the solution to* (4.2) *is not unique. In practice* (*see Section 4.4*)*, we sort sparse PCs by the explained variance* (*EV*) *of individual PCs, which is defined as* $\|Xy\|_2^2$*, where* $y \in \mathbb{R}^p$ *contains the loadings of a PC.*

Explained variance in the data

A non-diagonal middle B matrix facilitates the more general formulation of column sparse PCA. Specially, if B is restricted to diagonal, the formulation reduces to row sparse PCA.[2] Row sparse PCA presumes that given the subspace of PCs (i.e., the subspace spanned by some singular vectors of X), the PC loadings are approximately sparse by themselves (i.e., the singular vectors align closely with the natural coordinate axes). The next proposition compares column and row sparse PCA in terms of the matrix reconstruction error (the proof is provided in Appendix C.1).

Proposition 4.2 (Comparison of row and column sparsity)**.** *Let* $X \in \mathbb{R}^{n \times p}$ *be any matrix. Suppose* $S_Z \subseteq \mathbb{R}^{n \times k}$ *and* $S_Y \subseteq \mathbb{R}^{p \times k}$ *are the feasible sets for* Z *and* Y *respectively, where* $k \leqslant \min(n, p)$*. Then, subject to* $Z \in S_Z$*,* $Y \in S_Y$*, and* D *is diagonal, it holds that*

$$\min_{Z,B,Y} \|X - ZBY^T\|_F \leqslant \min_{Z,D,Y} \|X - ZDY^T\|_F.$$

Proposition 4.2 says that the solution to column sparse PCA has smaller reconstruction error of the data matrix than row sparse PCA. Since the squared matrix reconstruction error here is the unexplained variance in the data, it

[2] This restricted formulation is essentially a low-rank SVD with an additional sparsity constraint on the right singular vectors.

follows that the solution to column sparse PCA can capture more variance in the data than row sparse PCA.

Remark 4.3. *From a parametric perspective, SCA explains more variance because it uses* $k^2 - k$ *more parameters in the* B *matrix. Relative to the total number of parameters, this is typically a small increase; the* Z *and* Y *matrices contain roughly* $(n+p)k$ *parameters, and typically* k *is much smaller than* $n + p$*. Whether these additional parameters in B are statistically justified must be addressed in a case-by-case basis. In our limited experience with these techniques, the additional parameters are easily justified because the proportion of variance explained dramatically increases (see Section 4.4); the output becomes more stable across initializations, perturbations, and tuning parameters (see Section 4.4); and the estimated factors are easily interpretable (see Section 4.5 and 4.5).*

An algorithm for SCA

To solve SCA, the following lemma translates (4.2) into an equivalent and more convenient form (the proof can be found in Appendix C.1).

Lemma 4.4 (Bilinear form of SCA)**.** *Solving the minimization in* (4.2) *is equivalent to solving the following maximization problem,*

$$\begin{aligned} \underset{Z,Y}{\text{maximize}} \quad & \left\|Z^T X Y\right\|_F \\ \text{subject to} \quad & Z \in \mathscr{V}(n,k),\ Y \in \mathscr{V}(p,k),\ \|Y\|_1 \leqslant \gamma. \end{aligned} \tag{4.3}$$

In particular, for the optimizer in (4.2), $B = Z^T X Y$.

Due to the non-convexity of ℓ_2-equality constraints ($Z \in \mathscr{V}(n,k)$ and $Y \in \mathscr{V}(p,k)$), the feasible set in (4.3) is not convex in general. We replace the feasible set with its convex hull using some ℓ_2-inequality constraints for simplicity,

$$\begin{aligned} \underset{Z,Y}{\text{maximize}} \quad & \left\|Z^T X Y\right\|_F \\ \text{subject to} \quad & Z \in \mathscr{B}(n,k),\ Y \in \mathscr{B}(p,k),\ \|Y\|_1 \leqslant \gamma. \end{aligned} \tag{4.4}$$

Due to the Karush-Kuhn-Tucker conditions (see, e.g., Nocedal and Wright, 2006), one could expect the solution to fall on the boundary (i.e., $Z \in \mathscr{V}(n, k)$, $Y \in \mathscr{V}(p, k)$, and $\|Y\|_1 = \gamma$) so long as the sparsity parameters are chosen such that $k \leqslant \gamma \leqslant k\sqrt{p}$.[3] As such, local optima are not necessarily global optima. We discuss a data-driven method of tuning the sparsity parameters in Appendix C.2.

Next, we describe an algorithm that computes sparse PCs as formulated in (4.4). The input includes a data matrix X, the desired number of sparse PCs k, and optionally the sparsity controlling parameters γ. In our experiences, a default value of $\gamma = \sqrt{pk}$ appears to generate robust and interpretable sparse PCs (see, e.g., Section 4.4). The algorithm outputs the loadings of k sparse PCs. The SCA algorithm initializes $Z \in \mathscr{V}(n, k)$ and $Y \in \mathscr{V}(p, k)$ with the top k left and right singular vectors of X respectively. Once initialized, the algorithm alternatively updates Z and Y; fixing one and optimizing the other until convergence. The iteration is because the objective function is bilinear in Z and Y, allowing for fast updates. Specifically, with Y fixed, (4.4) takes the form

$$\underset{Z}{\text{maximize}} \quad \left\|Z^T X Y\right\|_F \quad \text{subject to } \; Z \in \mathscr{B}(n, k). \tag{4.5}$$

With Z fixed, (4.4) takes the form

$$\underset{Y}{\text{maximize}} \quad \left\|Z^T X Y\right\|_F \quad \text{subject to } \; Y \in \mathscr{B}(p, k), \; \|Y\|_1 \leqslant \gamma. \tag{4.6}$$

Update Z fixing Y

The update of Z fixing Y in (4.5) is algebraic. The following lemma provides a set of solutions to (4.5), which is extended from Theorem 7.3.2 in Horn and Johnson (1985) (the proof is included in Appendix C.1 for completeness).

Lemma 4.5 (Maximization without sparsity constraint)**.** *Given a full-rank matrix* $X \in \mathbb{R}^{n \times p}$*, with* $p \leqslant n$*, let the singular values of* X *be* σ_i *for* $i = 1, 2, ..., p$.

[3] This is for the set $\{Y \in \mathbb{R}^{p \times k} \mid \|Y\|_1 = \gamma\}$ to intersect with the Stiefel manifold $\mathscr{V}(p, k)$.

Then,

$$\max_{Y \in \mathscr{V}(n,p)} \quad \left\|X^T Y\right\|_F \;=\; \sum_{i=1}^{p} \sigma_i$$

with the maximizer $Y^* = \operatorname{polar}(X)$*, up to any orthogonal rotation from the right. Here,* $\operatorname{polar}(X) = X(X^T X)^{-1/2}$ *is the polar of* X.

Due to Lemma 4.5, the SCA algorithm updates Z with the polar of XY, $\hat{Z} = \operatorname{polar}(XY)$, which can be computed in $\mathscr{O}(nk)$ time (Journée et al., 2010).

Update Y **fixing** Z

To update Y fixing Z, we start by solving the non-sparse version of (4.6) (i.e., remove the sparsity constraint $\|Y\|_1 \leqslant \gamma$),

$$\underset{Y}{\text{maximize}} \quad \left\|Z^T X Y\right\|_F \quad \text{subject to } \; Y \in \mathscr{B}(p, k). \tag{4.7}$$

Let $\tilde{Y} = \operatorname{polar}(X^T Z)$. Then, $\tilde{Y}$ is one element in the subspace of the solutions to (4.7). Before imposing the sparsity constraint, we look for an orthogonal rotation R to $\tilde{Y}$ to minimize $\|\tilde{Y}R\|_1$. However, $\|Y\|_1$ is not a smooth function of Y if it contains at least one zero entry, entailing the complications of defining subgradients. Alternatively, the SCA algorithm minimizes a smoother criterion based on the $\ell_{4/3}$ norm:

$$\underset{R}{\text{minimize}} \quad \left\|\tilde{Y}R\right\|_{\frac{4}{3}} \quad \text{subject to } \; R \in \mathscr{U}(k). \tag{4.8}$$

This sub-problem leads to the varimax rotation (see Section 4.2) that is widely applied in factor analysis (Kaiser, 1958). We denote $Y^* = \tilde{Y}R^*$ to be the orthogonally rotated solution to (4.7), where R^* is the solution to (4.8). Finally, considering the ℓ_1-norm sparsity constraint, we apply the element-wise soft-thresholding of Y^* with the sparsity parameter γ, which is defined as (Donoho, 1995; Tibshirani, 1996)

$$[T_\gamma(Y^*)]_{ij} = \operatorname{sign}(Y^*_{ij}) \cdot \left(|Y^*_{ij}| - t\right)_+, \tag{4.9}$$

where $t > 0$ is the threshold determined by the equation $\|T_\gamma(Y^*)\|_1 = \gamma$, and x_+ equals x if $x > 0$ or 0 otherwise. We discuss several properties of soft-thresholding in Appendix C.3. In summary, the update of Y given Z consists of three steps that we call "Polar-Rotate-Shrink" (PRS, Algorithm 4.1)—first, compute a solution to the unconstrained problem (4.7); second, rotate with varimax; third, soft-threshold all of the elements. Algorithm 4.2 summarizes the algorithm of SCA.

```
Input: $A \in \mathbb{R}^{p \times k}$,
    sparsity parameter $\gamma$ (optional, default to $\sqrt{pk}$)   // Section C.2
Procedure PRS($A$):
    $\tilde{Y} \leftarrow$ left singular vectors of $A$
    $Y^* \leftarrow$ rotate $\tilde{Y}$ with varimax                  // Section 4.2
    $\hat{Y} \leftarrow$ soft-threshold $Y^*$ with parameter $\gamma$        // Appendix C.3
Output: $\hat{Y}$
```

Algorithm 4.1: Polar-Rotate-Shrink (PRS)

```
Input: Data matrix $X$ and a number of components $k$
Procedure SCA ($X, k$):
    Initialize $\hat{Z}$ and $\hat{Y}$ with the top $k$ left and right singular vectors of $X$
    repeat
        $\hat{Y} \leftarrow$ PRS($X^T\hat{Z}$)                        // Algorithm 4.1
        $\hat{Z} \leftarrow$ polar($X\hat{Y}$)                           // Lemma 4.5
    until convergence
Output: Sparse loadings $\hat{Y}$
```

Algorithm 4.2: Sparse Component Analysis (SCA)

The varimax rotation

For any matrix $A \in \mathbb{R}^{p \times k}$, the *varimax criterion* is defined as the sum of column (sample) variance of squared elements (A_{ij}^2) (Kaiser, 1958):

$$C_{\text{varimax}}(A) = \sum_{j=1}^{k} \left[\frac{1}{p} \sum_{i=1}^{p} A_{ij}^4 - \frac{1}{p^2} \left(\sum_{i=1}^{p} A_{ij}^2 \right)^2 \right].$$

For a fixed matrix $Y \in \mathbb{R}^{p \times k}$, the *varimax rotation* seeks an orthogonal rotation $R \in \mathbb{R}^{k \times k}$ to maximize the varimax criterion evaluated at YR,

$$\underset{R}{\text{maximize}} \quad C_{\text{varimax}}(YR) \quad \text{subject to } \; R \in \mathscr{U}(k). \tag{4.10}$$

It is commonly used in factor analysis for producing nearly sparse and interpretable loadings of PCs, especially in the psychology literature. The varimax rotation is easy to compute; for example, the base function `varimax` in `R` implements a gradient projection algorithm of it (Bernaards and Jennrich, 2005). Jennrich (2001) showed that the gradient projection algorithm converges to a local optimum from any starting point and enjoys geometric (or linear) convergence rate.

The varimax criterion naturally links to the $\ell_{4/3}$-norm objective function in (4.8). Since $Y \in \mathscr{V}(p,k)$, the columns of Y have unit length. Hence, $\sum_{i=1}^{p} Y_{ij}^2 = 1$, and the varimax criterion reduces to a simpler form (also known as the *quartimax* criterion as introduced by Carroll (1953)) up to an additive constant:

$$C_{\text{quartimax}}(Y) = \sum_{i=1}^{p} \sum_{j=1}^{k} Y_{ij}^4 = \|Y\|_4^4,$$

which is the ℓ_4-norm of Y to the power of 4. Next, by the Hölder's inequality (using the Hölder conjugates 4/3 and 4) and the power mean inequality (and that $\|Y\|_F = \sqrt{k}$),

$$\|Y\|_{\frac{4}{3}} \|Y\|_4 \geqslant \|Y\|_1 \geqslant \|Y\|_F = \sqrt{k}.$$

This implies that maximizing the varimax criterion is the dual problem of minimizing the $\ell_{4/3}$-norm objective. Hence, to update Y in the algorithm of SCA, we invoke the varimax rotation in (4.10) as a proxy of (4.8).

Remark 4.6. *Besides varimax, we experimented the orthogonal rotation that directly minimizes the ℓ_1 norm, which we call the "absmin" rotation:*

$$\underset{R}{\text{minimize}} \quad \|YR\|_1 \quad \text{subject to } \; R \in \mathscr{U}(k). \tag{4.11}$$

However, the objective function is not smooth at those R where YR contains at least one *zero element; this posts challenges to solving* (4.11). *For example, we tried a gradient projection algorithm using the gradient direction* $Y^T \operatorname{sign}(YR)$, *where* $\operatorname{sign}(\cdot)$ *is the element-wise sign function, yet the algorithm hardly converges. It is worth noting that in our limited experiments, where we used the absmin rotation but only allowed fifteen iterations of this gradient projection algorithm, we obtained marginally better solutions, in terms of explained variance, than using the varimax rotation* (*see Section 4.4*). *It is of future interest to investigate alternative orthogonal rotations that are easy to compute and can generate approximately sparse structure.*

Sparse matrix approximation

In the SCA algorithm above, a sparsity constraint can also be applied to Z, in addition to Y. We call this sparse matrix approximation (SMA). We define SMA as the solution to a matrix reconstruction error minimization problem:

$$\begin{aligned} \underset{Z,B,Y}{\text{minimize}} \quad & \left\|X - ZBY^T\right\|_F \\ \text{subject to} \quad & Z \in \mathscr{B}(n,k),\ P_1(Z) \leqslant \gamma_z, \\ & Y \in \mathscr{B}(p,k),\ P_2(Y) \leqslant \gamma_y, \end{aligned} \tag{4.12}$$

where $\gamma_z > 0$ and $\gamma_y > 0$ are the sparsity controlling parameters, and P_1 and P_2 are some *penalty* functions that promote sparsity. If γ_Z is so large that $P_1(Z) \leqslant \gamma_z$ is always satisfied, then (4.12) is equivalent to SCA. Similar to Lemma 4.4, we transform (4.12) into an equivalent and more convenient form (the proof is almost identical to that of Lemma 4.4 thus is omitted),

$$\begin{aligned} \underset{Z,Y}{\text{maximize}} \quad & \left\|Z^T XY\right\|_F \\ \text{subject to} \quad & Z \in \mathscr{B}(n,k),\ P_1(Z) \leqslant \gamma_z, \\ & Y \in \mathscr{B}(p,k),\ P_2(Y) \leqslant \gamma_y. \end{aligned} \tag{4.13}$$

The two criteria in (4.12) and (4.13) are equivalent if and only if $B = Z^T XY$. We interpret B as the "*score*" of SMA, since the solution to (4.12) maximizes the sum of squares of its elements, $\sum_{i,j} B_{ij}^2$. It is also worth noting that the squared matrix reconstruction error equals to $\|X\|_F^2 - \|B\|_F^2$ (see the proof of

Lemma 4.4).

Since SMA is a simple extension from SCA, we extend Algorithm 4.2 for SMA in Algorithm 4.3, where we apply PRS to Z in addition to Y. The output includes the estimated Z, B, and Y.

```
Input: data matrix $X \in \mathbb{R}^{n \times p}$ and the approximation rank $k$
Procedure SMA $(X, k)$:
    Initialize $\hat{Z}$ and $\hat{Y}$ with the top $k$ left and right singular vectors of $X$
    repeat
        $\hat{Z} \leftarrow$ PRS$(X\hat{Y})$                              // Algorithm 4.1
        $\hat{Y} \leftarrow$ PRS$(X^T\hat{Z})$                            // Algorithm 4.1
    until convergence
    $\hat{B} \leftarrow \hat{Z}^T X \hat{Y}$
Output: $\hat{Z}$, $\hat{B}$, and $\hat{Y}$
```

Algorithm 4.3: Sparse Matrix Approximation (SMA) with $P_1(A) = P_2(A) = \|A\|_1$.

We highlight that SMA generalizes the popular penalized matrix decomposition (PMD) proposed by Witten et al. (2009), which is also similar to the method of Shen and Huang (2008). The PMD also approximates a data matrix $X \in \mathbb{R}^{n \times p}$ by the product of three matrices, ZDY^T, where $Z \in \mathscr{V}(n, k)$ and $Y \in \mathscr{V}(p, k)$ are presumed sparse, and $D \in \mathbb{R}^{k \times k}$ is a diagonal matrix whose diagonal entries are in decreasing order, and k is the rank of the matrix approximation. For sparsity, PMD applies penalty functions to Z and Y, leading to the matrix reconstruction error minimization formulation of PMD:[4]

$$\begin{aligned} \underset{U,D,V}{\text{minimize}} \quad & \|X - ZDY^T\|_F \\ \text{subject to} \quad & Z \in \mathscr{B}(n, k),\ P_1(Z) \leqslant \gamma_z, \\ & Y \in \mathscr{B}(p, k),\ P_2(Y) \leqslant \gamma_y, \\ & D \text{ is diagonal}, \end{aligned}$$

where $\gamma_z, \gamma_y > 0$ are parameters that control the sparsity of Z and Y, and P_1

[4] The paper originally considers the PMD with $k = 1$. The PMD finds multiple factors sequentially using a deflation technique.

and P_2 are some convex penalty function (e.g. ℓ_1-norm).

The single difference between SMA and PMD is the the diagonal constraint on the middle matrix. In this way, SMA generalizes PMD, because, SMA estimates $k^2 - k$ more parameters in B than PMD (see Remark 4.3). Proposition 4.2 suggests that the reconstruction error of SMA is less or equal to that of PMD (see also Remark C.1 in Appendix C.1). Algorithmically, in order to compute PMD, Witten et al. (2009) proposed to find the solution by sequentially maximizing B_{ii} for $i = 1, 2, ..., k$ (recall that $B = Z^T XY$). By contrast, solving the SMA in (4.13) amounts to maximizing the entirety of the score matrix, that is, $\|B\|_F$.

4.3 Connections to existing methods

In this section, we compare SCA with several existing methods of sparse PCA and discuss two variants and one extension of SCA.

Existing sparse PCA methods

The formulation of SCA is akin to multiple existing sparse PCA formulations. However, the possibility of orthogonal rotations has not been explored thoroughly, despite the plethora of available methods. In this section, we elucidate these connections and point to some differences.

SPCA (Zou et al., 2006) SPCA is motivated to maximize the explained variance in the data (Jolliffe et al., 2003). The formulation of SPCA minimizes a “residual sum of squares plus penalties” type of criterion,

$$\begin{array}{ll} \underset{U,V}{\text{minimize}} & \left\|X - XVU^T\right\|_F^2 + \lambda_1 \|V\|_F^2 + \lambda_2 \|V\|_1 \\ \text{subject to} & U \in \mathscr{V}(p, k), \end{array}$$

where $V \in \mathbb{R}^{p \times k}$ is the sparse loadings of interest, and λ_1 and λ_2 are tuning parameters. We note that the first term in the objective function is also invariant to any orthogonal rotation applied to U and V, because $\left\|X - XVU^T\right\|^2 = \left\|X - X(VR)(UR)^T\right\|^2$ for any $R \in \mathscr{U}(k)$. However, the algorithm of SPCA for U and V does not use orthogonal rotations to search over the solution space, as it is adapted from the elastic net (Zou and Hastie, 2005). Explicitly searching for a sparsity-enabling rotation R could help to find a smaller objective value in SPCA.

SPC (Witten et al., 2009) SPC finds one sparse PC at a time,

$$\begin{aligned} \underset{u,v}{\text{maximize}} \quad & u_i^T X v_i \\ \text{subject to} \quad & \|u_i\|_2 = 1,\ \|v_i\|_2 = 1,\ \|v_i\|_1 \leqslant \gamma, \end{aligned} \tag{4.14}$$

where $v_i \in \mathbb{R}^p$ contains the loadings of the ith sparse PC, for $1 \leqslant i \leqslant k$. When $k = 1$, our formulation of SCA in (4.3) takes the same form as the SPC formulation, where an orthogonal rotation is unnecessary. When $k > 1$, however, SPC searches for sparse PCs sequentially and does not rotate PCs, unlike SCA, which computes k sparse PCs simultaneously. SPC is similar to the rSVD proposed by Shen and Huang (2008) and the TPower proposed by Yuan and Zhang (2013) in that all the three methods rely on a deflation technique for multiple PCs. This technique entails complications of, for example, non-orthogonality and sub-optimality (Mackey, 2008). More generally, these methods can each be viewed as a special case of the following GPower formulation.

GPower (Journée et al., 2010) GPower has a "block version" that computes multiple sparse PCs simultaneously by considering a linear combination of individual sparse PCA (as formulated in SPC),

$$\begin{aligned} \underset{U,V}{\text{maximize}} \quad & \sum_{j=1}^{k} \mu_j u_j^T X v_j - \sum_j \lambda_j \|v_j\|_1 \\ \text{subject to} \quad & U \in \mathscr{B}(n,k),\ V \in \mathscr{V}(p,k), \end{aligned}$$

where V contains the PC loadings, and u_j and v_j are the jth column of U and V respectively, and μ_j is the weight for the jth sparse PC, and λ_j is the sparsity tuning parameter for the jth sparse PC. The algorithm of GPower fundamentally deals with sparse PCs individually, which prohibits orthogonal rotations (on V).

SPCArt (Hu et al., 2016) SPCArt is the first (to our knowledge) sparse PCA method that concerns orthogonal rotations in its formulation. It searches for sparse PCs by directly approximating the singular vectors (as opposed to minimizing the reconstruction error or maximizing the explained variance),

$$\begin{array}{ll} \underset{Y,R}{\text{minimize}} & \|V - YR\|_F^2 + \lambda\|Y\|_1 \\ \text{subject to} & Y \in \mathscr{V}(p,k),\ R \in \mathscr{U}(k), \end{array}$$

where $V \in \mathscr{V}(p,k)$ contains the top k singular vectors of X, and Y contains the sparse loadings. Conceptually, introducing an orthogonal rotation (R) allows a larger searching space for Y. However, the algorithm of SPCArt does not specifically update R to promote sparsity (e.g., minimize $\|Y\|_1$ as in SCA); instead, SPCArt simply computes R so as to align the polar of V and Y (i.e., $\hat{R} = \text{polar}(Y^T V)$). As such, the performance of SPCArt could be sensitive to the initialization of Y. Empirically, SPCArt yields results that are nearly comparable to the GPower based method, as concluded by the authors.

Sparse coding and independent component analysis

Sparse coding concerns low-rank representations of individual samples. We view it as a variant of PCA, where we presume the component scores to be sparse. Recall that the scores are the representations of individual data points in $\mathbb{R}^k$, where k is the number of PCs. In particular, presuming sparse scores implies that each data point is correlated with only a small subset of PCs. Sparse coding is useful to generate simple representations of individual date points, and the basis of such representations (i.e., PCs) usually provide scientific insights. For example, sparse coding of natural images recovers the common understanding of how the primary visual cortex in mammalian perceives scenes (see Section 4.5 for an example).

The SCA algorithm can be used to solve sparse coding. This is because, similar to SCA, sparse coding can be viewed as a special case of the SMA problem. To see this, simply omit the sparsity constraint on Y in (4.12),

$$\begin{aligned} \underset{Z,B,Y}{\text{minimize}} \quad & \left\|X - ZBY^{T}\right\|_{F} \\ \text{subject to} \quad & Z \in \mathscr{B}(n,k),\ Y \in \mathscr{B}(p,k),\ P_1(Z) \leqslant \gamma_z \end{aligned}$$

Here, Z contains the sparse scores, and BY^{T} contains the basis of sparse coding. To solve sparse coding, we apply the SCA algorithm (Algorithm 4.2) to the transposed data matrix, X^{T}. In doing this, the output of the algorithm is actually an estimate of sparse component scores for the original data matrix.

More broadly, independent component analysis (ICA) is widely applied for sparse coding in the signal processing literature. Despite the different motivations, sparse PCA on a transposed data matrix appears to perform very similarly to sparse ICA on the original data. We elaborate on this in Appendix C.4 and apply SCA to blind source separation of images.

4.4 Simulation studies

In this section, we compare several sparse PCA methods using simulated data. Specifically, we focused on (1) their ability of explaining variance in the data, (2) the robustness against varying sparsity parameters, and (3) the computational speed. We selected SPCA, SPC, GPower, the SPCAvRP method recently proposed by Gataric et al. (2020), SCA, and another variant of SCA which deploys the absmin rotation (SCA-absmin, see Remark 4.6 of Section 4.2). For SCA and SCA-absmin, we implemented the algorithms in `R`.[5] For SPCA, SPC, and SPCAvRP, we invoked the original `R` packages `elasticnet`, `PMA`, and `SPCAvRP` respectively. The implementation of GPower (in `MATLAB`) was obtained from the authors' website. For all the iterative methods, we specified maximum number of iterations to 1,000 and the stopping

[5] We provide an `R` package `epca`, for **e**xploratory **p**rincipal **c**omponent **a**nalysis, which implements SCA and SMA with various algorithmic options. The package is available from CRAN (`https://CRAN.R-project.org/package=epca`).

(convergence) criterion to 10^{-5}. Overall, our numerical experiments showed that the SCA algorithm converges faster and produces more robust sparse PCs that capture a larger amount of variance in the data.

Proportion of variance explained

In this simulation, we compared the abilities of sparse PCA methods in explaining variance in the data. To this end, we simulated 30 data matrices with $n = 100$ observations and $p = 100$ variables from the following low-rank generative model:

$$X = SY^T + E,$$

where $S \in \mathbb{R}^{100\times 16}$ contains the component scores, and $Y \in \mathbb{R}^{100\times 16}$ contains the loadings of sparse PCs, and $E \in \mathbb{R}^{100\times 100}$ is some noise. To generate S, we randomly sampled $U \in \mathscr{V}(100, 16)$ and $V \in \mathscr{U}(16)$ and set $S = U\Sigma V^T$, where Σ is a diagonal matrix with the diagonals $\sigma_l = 10 - \sqrt{l}$ for $l = 1, 2, ..., 16$. To simulate a sparse Y, we took a random element from $\mathscr{V}(100, 16)$, then soft-threshold its elements with sparsity parameter $\gamma = 20$ (i.e., T_{20} as defined in Equation (4.9)). Note that, it is unnecessary to re-scale the columns of loadings to unit length, because the column of S can absorb these scalars. Lastly, the elements in E were drawn independently from the normal distribution, $E_{ij}\sim N(0, 0.1^2)$.

We applied the six sparse PCA methods to each simulated data matrix X with $k = 2, 4, 6, ..., 16$. For each k, we imposed the same ℓ_1-norm constraint on the sparse loadings for all methods. Specifically, for SCA, and SPC, we directly configured the sparsity controlling parameters to 2.5k. As for SPCA, GPower and SPCAvRP, to ensure a fair comparison, we tuned the parameters such that the returned loadings all have the same ℓ_1 norm of 2.5k. To evaluate sparse PCs, we define the cumulative proportion of variance explained (PVE) by the first k sparse PCs as $\|X_Y\|_F^2$, where $X_Y = XY(Y^TY)^{-1}Y^T$ (Shen and Huang, 2008). Note that the PVE by sparse PCs is upper bounded by that of traditional PCs (no sparsity constraint). Therefore, we also applied PCA to X for comparison. Figure 4.3 displays the mean PVE for different PCA methods, varying the requested number of PCs from 2 to 16. It can be seen that SPCAvRP and SPCA

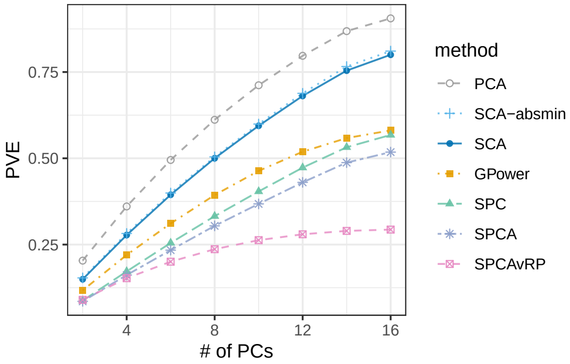


Figure 4.3: Comparisons of sparse PCA methods using simulated data. The proportion of variance explained (PVE) by sparse principal components (PCs) with the number of targeted PCs varying from 2 to 16.

explained less than half of the PVE by PCA, and that GPower and SPC both exhibited some improvements over SPCA. For GPower, we tested both the single-unit and the block versions, but the block version often converged to a defective solution with some columns decaying to all zeros. This happened when the number of targeted PCs went above 5 in this simulation. Overall, SCA performed the best among sparse PCA methods and were the closest to PCA. In addition, the SCA algorithm converged with fewer iterations than the other sparse PCA methods (see Table 4.1 for a comparison when $k = 16$). We also observed that using the varimax rotation (SCA), the algorithm was more computationally efficient than using the absmin rotation (SCA-absmin).

Robustness against tuning parameters

This simulation study investigates the robustness of sparse PCA to the choice of sparsity parameters. For this, we applied sparse PCA to detect communities in networks (or graph partitioning) (see, e.g., Fortunato, 2010), using the graph adjacency matrix (see the definition below) as input. This application is possible thanks to the recent consistency results (Rohe and Zeng, 2020) showing that under the stochastic block model (SBM, see for example Holland et al., 1983a), the support of each sparse PC estimates the membership (indicator)

Method	# of iterations	Mean run time (s)	Environment
SCA	10 ~ 65 (all PCs)	0.96	R
SPC	25 ~ 1,000 (each PC)	1.21	R
GPower	30 ~ 150 (each PC)	0.19	MATLAB
SPCA	470 ~ 920 (all PCs)	56.30	R
SPCAvRP	/	28.67	R
SCA-absmin	/	23.5	R

Table 4.1: Comparison of the computational efficiency of sparse PCA methods. Each method is tasked to find 16 PCs on a single CPU (2.50GHz). SPCAvRPs is not iterative (yet is parallelizable), hence the number of iterations is not applicable. The absmin rotation is less efficient, so we halted the algorithm of SCA-ABSMIN after the 15th iteration.

of one community. Hence, we could evaluate sparse PCs by examining their support.

We simulated 30 undirected graphs with $n = 900$ nodes and four equally sized blocks from the SBM. Under the SBM, the edge between node i and j is sampled from the Bernoulli distribution, Bernoulli$(B_{z(i),z(j)})$, where $z(i) \in \{1, 2, 3, 4\}$ is the membership of node i, and

$$B = 0.05 \times \begin{bmatrix} 0.6 & 0.2 & 0.1 & 0.1 \\ 0.2 & 0.7 & 0.05 & 0.05 \\ 0.1 & 0.05 & 0.6 & 0.25 \\ 0.1 & 0.05 & 0.25 & 0.6 \end{bmatrix}$$

is the block connectivity matrix. Under this setting, the expected number of edges connected to each node is 45. For each simulated graph, we defined the adjacency matrix $A \in \{0, 1\}^{n \times n}$ with $A_{ij} = 1$ if and only if i and j are connected.

We applied SCA, SPC, and GPower[6] to each of the 30 simulated adjacency matrices with $k = 4$. We varied the sparsity parameter γ to take value in $\{18, 24, 36, 48, 60, 66\}$. For SPC, we required each of the four PCs to have ℓ_1 norm $\gamma/4$. As for GPower, we tuned its parameters such that the returned

[6] Since SPCA and SPCAvRP performs worse than SPC and GPower (Zou and Xue, 2018), we excluded the two methods in this simulation for simplicity.

loading matrix has the ℓ_1 norm of γ. Figure 4.4 depicts the estimated loadings returned by SCA and SPC. On the left two columns of panels ($\gamma = 48$ and 36), the supports of the four sparse PCs were well separated and indicated block memberships. This suggested that we could use the loadings to cluster nodes and quantitatively assessed the quality of sparse PCA methods. Specifically, we assigned node i to cluster j if Y_{ij} is the largest absolute value in the ith row of Y, that is, $|Y_{ij}| > |Y_{il}|$ for all $l \neq j$. In the case of ties or all-zero rows, the cluster label is randomly assigned. For each estimate, let $C \in \{1, 2, 3, 4\}^n$ contain the assigned cluster labels and $C^* \in \{1, 2, 3, 4\}^n$ contain the true labels. Define the accuracy as

$$\text{Accuracy}(C, C^*) = \max_{\pi \in \mathscr{P}(4)} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left(\pi(C_i) = C_i^*\right) \right\},$$

where $\mathscr{P}(4)$ contains all the possible permutation functions of the set $\{1, 2, 3, 4\}$, and $\mathbb{1}(x)$ is the indicator function of x. We used the accuracy to assess the quality of the sparse PCA solutions. Figure 4.5 depicts the accuracy of the three methods with varying sparsity parameters. It can be seen that the performance of GPower and SCA were less affected by the changing of sparsity parameter, while SPC was profoundly influenced. As γ became smaller, SPC quickly lost its power in community detection, suggesting that SPC is more sensitive to the choices of tuning parameter. Although less sensitive to the change in γ, GPower produced poorer estimation of sparse PCs, with the accuracy slightly better than random guesses (accuracy $= 0.25$). Overall, SCA yielded higher accuracy with smaller deviation compared to the others, suggesting that SCA is less dependent on the choice of sparsity parameters.

In this example, SCA outperforms SPC because it finds a better optimization solution. This comparison could be made difficult by the fact that they have different objective functions. However, in this case, even though SCA is optimizing a different objective function, it outperforms SPC at *optimizing the SPC objective function*. Table 4.2 lists the objective values of SPC (Equations (4.14)) evaluated using the solutions of the SCA and SPC algorithms with various γ. When $\gamma \in \{36, 48, 60, 66\}$, the SCA algorithm outputs a solution that

	$\gamma = 18$	$\gamma = 24$	$\gamma = 36$	$\gamma = 48$	$\gamma = 60$	$\gamma = 66$
Using SCA sol.	191.47	323.36	**1135.03**	**1906.25**	**2554.86**	**2783.73**
Using SPC sol.	**544.81**	**705.01**	1029.04	1195.91	1334.67	1423.95

Table 4.2: Comparison of the SPC objective values, $\sum_{i=1}^{4}(u_i^T A v_i)^2$ (see Equation (4.14)), evaluated using the output of the SCA and SPC algorithms with various sparsity parameter (γ).

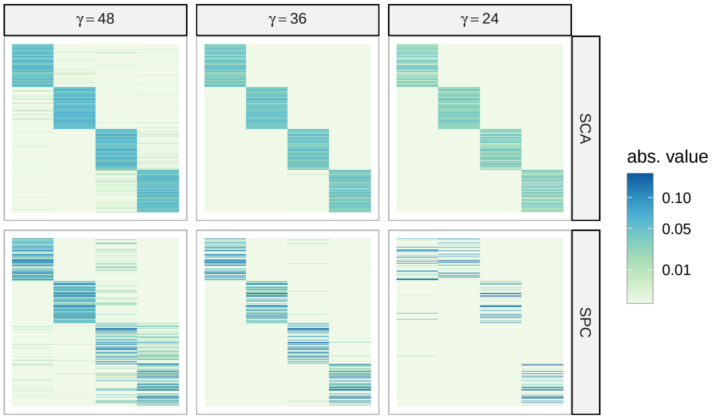


Figure 4.4: Comparisons of SCA and SPC using simulated network data. Heat maps of the loadings (900×4 matrices) returned by SCA and SPC using three different sparsity parameters ($\gamma = 24, 36, 48$). In each heat map, rows correspond to nodes, which are grouped by the true community membership, and each column corresponds to one sparse PC. The color shade indicates the absolute of loadings.

achieves a higher value of the SPC objective, suggesting that the SPC algorithm is likely to return local optima.

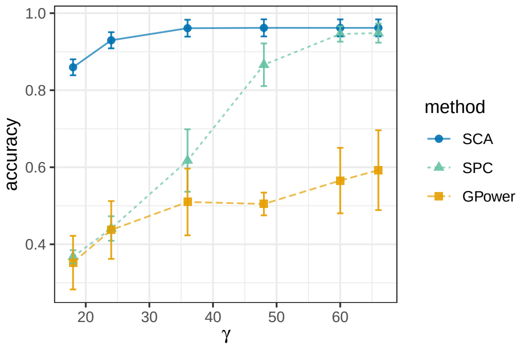


Figure 4.5: Comparisons of sparse PCA methods using simulated network data. The accuracy of SCA, GPower, and SPC in community detection using various sparsity parameters (γ). Each point indicates the mean accuracy across 30 replicates, and the error bar indicates the standard deviation of the evaluated accuracy.

4.5 Applications

In this section, we applied SCA to real data. The first application is the sparse coding of natural images. It illustrates the utility of sparse PCA as independent component analysis. Appendix C.4 contains another application of SCA to blind source separation of images. Next, we demonstrate the ability of SCA in handling high-dimensional problems (i.e., $p > n$) through a transcriptome sequencing dataset and a targeted sample of Twitter friendship network. These datasets are of large scale. To our knowledge, no other current implementations of sparse PCA can efficiently handle a large matrix at the scale. As such, we will restrict our discussion to SCA.

Sparse coding of images

Low-level visual layers, such as retina, the lateral geniculate nucleus, and the primary visual cortex (V1) are shared processing components in mammalian. The receptive fields in the V1 can be characterized as being spatially localized, oriented and bandpass (i.e., selective to structure at different spatial scales). To understand V1, one line of research focuses on finding sparse and linearly independent codes for natural images, which provides an efficient representation for later stages of processing (Field, 1994; Olshausen and Field, 1996; Bell and Sejnowski, 1997). This type of research is based on the hypothesis of sparse coding, that is, any perceived scenes can be synthesized via the linear combination of some small subsets of basis images (Lee et al., 2006; Gregor and LeCun, 2010)). In this application, we show that sparse PCA produces a set of bases for natural images that resembles those found in Olshausen and Field (1996).

We utilized ten natural images from Olshausen and Field (1996), each of which contains 512 × 512 pixels. We followed the same whitening process as described by the authors. Next, we randomly sampled a total of $12,000$ small image patches the ten images, where each patch contains 16 × 16 pixels. This was followed by a centering step that subtracts each pixel by the mean of all 256 pixels. We vectorized each patch of image and put them into the rows of a data matrix, $X \in \mathbb{R}^{n \times p}$, where $n = 12,000$ and $p = 256$. Finally, we applied SCA to the transposed data matrix, X^T, to find 49 sparse PCs ($k = 49$) with the default sparsity parameter, $\gamma = \sqrt{pk}$ (Note that this is sparse coding). In particular, for the varimax rotation, we normalized the rows to unit length rescaled them afterward, as recommended by Kaiser (1958). In the output of SCA, the estimated scores $S \in \mathbb{R}^{p \times k}$ contains the basis images, and the estimated sparse loadings $Y \in \mathbb{R}^{n \times k}$ encodes how the basis images are linearly combined to form each image patch (i.e., Y contains the linear coefficients).

Figure 4.6 displays the 49 image bases returned by PCA and SCA, where each image represents one column of S (transformed into a 16 × 16 array). For SCA, all of the basis images appeared to exhibit simple patterns, such as lines and edges. As for PCA, the oriented structure in the first few basis images does not arise as a result of the oriented structures in natural images, yet more likely because of the existence of those components with low spatial frequency (Field, 1987).

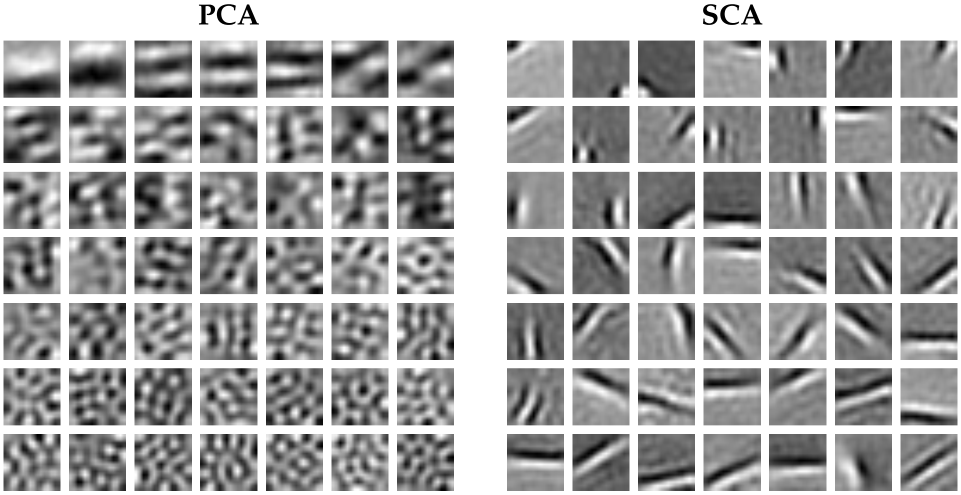


Figure 4.6: Sparse image encoding using PCA (left) and SCA (right). For both method, shown are the 49 image bases (i.e., component scores) extracted from natural images. Each image basis is in 16 × 16 pixel.

Analysis of single-cell gene expression data

Single-cell transcriptome sequencing (scRNA-seq) provides high-throughput transcriptome expression quantification at individual cell level. It has been widely used across biological disciplines. For example, patterns of gene expression can be identified through clustering analysis. This helps uncover the existence of rare cell types within a cell population that have never been seen (Plasschaert et al., 2018; Montoro et al., 2018). In this application, we aimed to use SCA to extract the sparse PCs of genes that characterize some known cell types.

For this application, we used the human pancreatic islet cell data from Baron et al. (2016). We removed the genes that do not exhibit variation across all cells (i.e., zero standard deviation) and removed the cell types that contain fewer than 100 cells. This resulted in a data matrix $X \in \mathbb{R}^{n \times p}$ of $n = 8,451$ cells across nine cell types and $p = 17,499$ genes, with X_{ij} measuring the expression level of gene j in cell i. X is sparse; it contains 10.8% non-zero elements. We applied SCA on X to find $k = 9$ sparse gene PCs. We set the

PC	# of genes	Gene name(s)
1	1	INS
2	1	SST
3	1	GCG
4	8	CTRB2, REG1A, REG1B, REG3A, SPINK1 ...
5	15	CELA3A, CPA1, CTRB1, PRSS1, PRSS2 ...
6	1	IAPP
7	1	PPY
8	3	CLU, GNAS, TTR
9	61	ACTG1, EEF1A1, FTH1, FTL, TMSB4X ...

Table 4.3: Sparse gene PCs estimated by SCA. For each gene PC, the number of genes (i.e., the number of non-zeros in the loadings) and the top 5 genes according to the absolute loadings are reported.

sparsity parameter to $\gamma = \log(pk) \approx 12$, as we aimed for particularly sparse PCs (i.e., each PC is consist of a small number of genes). The algorithm took about 5 minutes (24 iterations) to complete on a single processor (3.3GHz). As a result, each column of the loading matrix contains a small number of non-zero elements, suggesting that most of the gene PCs consist of one or a few genes. Table 4.3 lists the names of these genes for each PCs. For example, the PC 2 consists of only one gene, SST. Despite the simple structure of PCs, these PCs picked up informative gene markers for individual cell types. To see this, we calculated the scores for each cell using the 9 PCs (That is, each cell gets 9 scores, each of which corresponds to one of the nine PCs). Figure 4.7 displays the box plots of the scores stratified by cell type. For example, the expression of the SST gene (which solely composes the 2nd PC) identifies the "delta" cells. This result highlights the power of scRNA-seq in capturing cell-type specific information and suggests the applicability of our methods to high-dimensional biological data.

Clustering of Twitter friendship network

This application serves in a grand efforts of ours to study political communication on social media, like Twitter. The information on Twitter is organized so that users primarily read the tweets of their "friends." In order to select

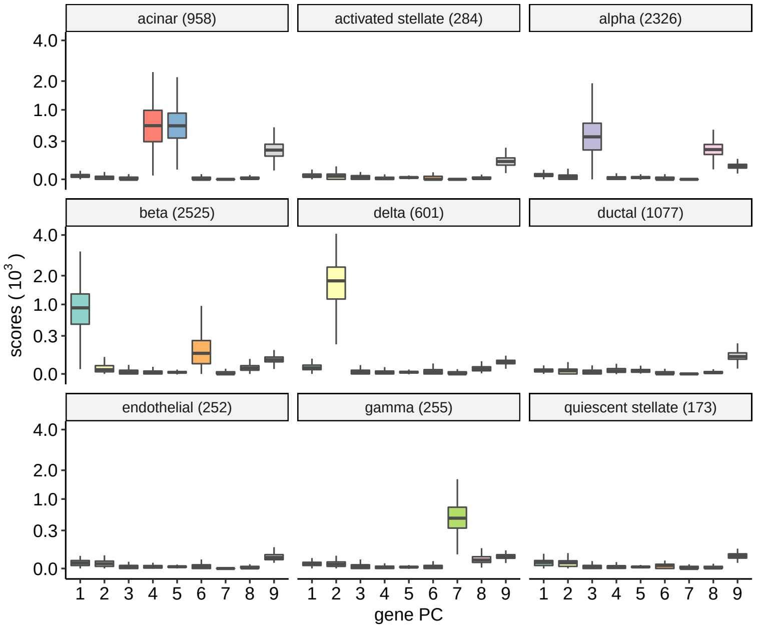


Figure 4.7: Scores of sparse gene principal components (PCs) stratified by cell types. Each panel displays one of nine cell types with the names of cell types and the number of cells reported on the top strips. For each cell type, a box depicts the component scores for nine sparse gene PCs.

content, a user can freely "follow" (and "unfollow") any other accounts, and we call these other accounts the friends of it. Thanks to this design, the communication on Twitter can be contextualized by the friendship network. As such, we hypothesize that user's community membership in the network offers the context of user's opinion expression on social media (Zhang et al., 2021). To study the hypothesis, a key step is to cluster Twitter accounts using their friendship network. In this section, we demonstrate large-scale network clustering using sparse PCA.

For this application, we collected a targeted sample from the Twitter friendship network in August 2018 (Chen et al., 2020a). In this sample, there are $n = 193,120$ Twitter accounts who follow a total of $p = 1,310,051$ accounts, after filtering out the accounts with few followers or followings. We defined the graph adjacency matrix $A \in \{0,1\}^{n \times p}$ with $A_{ij} = 1$ if and only if account i

follows account j.[7] This resulted in a sparse A with about 0.02% entries being 1. We applied SMA to A with $k = 100$ and default sparsity parameters. This analysis was computationally tractable; one iteration of the SMA algorithm took about 54 minutes on a single processor (2.5GHz), thanks to the efficient algorithm that computes the sparse SVD (Baglama and Reichel, 2005). Figure 4.2 displays seven example columns of Y. Using the output $Z \in \mathbb{R}^{n \times k}$ and $Y \in \mathbb{R}^{p \times k}$ from SMA, the clusters of Twitter accounts were determined as follows (same as in Section 4.4): the ith row account of A was assigned to the lth row cluster if Z_{il} was the greatest in the ith row of Z, that is, $|Z_{il}| \geqslant |Z_{il'}|$ for all $l' = 1, 2, ..., k$, and the jth column account of A was assigned to the lth column cluster if Y_{jl} was the greatest in the jth row of Y, $|Y_{jl}| \geqslant |Y_{jl'}|$ for all $l' = 1, 2, ..., k$. Upon detailed evaluation of these clusters, we showed that our clustering of Twitter accounts formed homogeneous, connected, and stable social groups (Zhang et al., 2021). For example, we found that a user is more likely to retweet the content that originated from another member in the same clusters (p-value $< 10^{-16}$ in a χ^2 test). More interestingly, the estimated row clusters and column clusters are matched (Rohe et al., 2016), that is, the kth row cluster tends to follow the accounts in the kth column cluster. To illustrate this, we quantified the number of followings from the row clusters to the corresponding column clusters. Figure 4.8 displays the results for 50 selected clusters that are related to U.S. politics. It can be seen that the number of followings between each paired row and column clusters (i.e., the diagonals in Figure 4.8) showed marked enrichment. These results suggest the efficacy of our methods for analysis of social network data.

4.6 Discussion

In this paper, we introduced SCA, a new method for sparse PCA, and SMA, an extension for two-way matrix analysis. SCA differs from the existing sparse

[7] The columns of A are not centered nor scaled. One alternative is to use the normalized version of A. For example, define the regularized graph Laplacian as $L \in \mathbb{R}^{n \times p}$ with $L_{ij} = A_{ij}/\sqrt{(r_i + \bar{r})(c_j + \bar{c})}$, where $r_i = \sum_j A_{ij}$ is the sum of the ith row of A, $c_j = \sum_i A_{ij}$ is the sum the jth column of A. Here, $\bar{r}$ and $\bar{c}$ are the means of r_i's and c_j's respectively. (Zhang and Rohe, 2018).

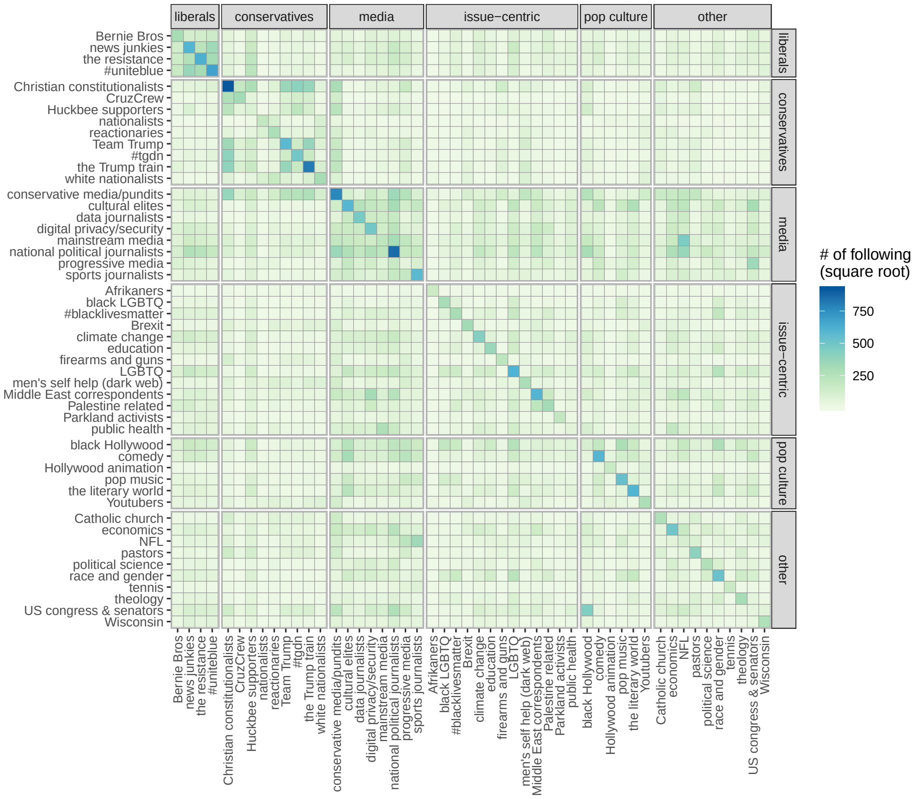


Figure 4.8: Heat map of friend counts between row and column clusters of Twitter accounts. Each row and column corresponds to a cluster. The row and column panels indicate cluster category, with the category names shown in the top and right strips. The color shades indicate the number of followings from the row cluster to the column cluster, after the square root transformation.

PCA methods in that it estimates column sparse PCs, that is PCs that are sparse in an orthogonally rotated basis. This is particularly useful when the singular vectors of a data matrix (or the eigenvectors of the covariance matrix) are not readily sparse. We demonstrated that it explains more variance in the data than the state-of-the-art methods of sparse PCA. In addition, the algorithm is also stable and robust against a wide choices of tuning parameters. In practice, SCA is advantageous when multiple PCs are desired because it does not require the deflation.

5 LOCAL WORD EMBEDDING FOR TARGETED DOCUMENT ASSEMBLING

5.1 Introduction

As lots of human activities now taking place online, digital media have come to record an ever-increasing share of human communication and social interaction, making available vast quantities of big data, in the forms of text, audio and video (Gentzkow et al., 2019; Golder and Macy, 2014). Such rich data offers an unprecedented opportunity for social scientists to study individuals and society at large unobtrusively (Salganik, 2019). Particularly, recent years have seen an explosion of "text-as-data" research in almost every discipline of social science (Grimmer and Stewart, 2013). Given that text data from digital media like social media platforms are not custom-made for research and often stored in vast databases (e.g., Twitter's global stream of tweets), the first order of business for a social scientist researching on a topic of interest (e.g., a political issue or a news event) is to assemble a comprehensive and relevant subset of documents by querying the database. And this can be an early-stage challenge for research using big data.

In an ideal situation, the research topic is accurately characterized by attributes of documents, such as a given group of authors and a specific journal venue in bibliometric research. In this case, some basic database queries would be sufficient to produce a consistent assembling of documents efficiently, thanks to matured data management technologies. However, in most cases, social scientists have to collect or assemble a relevant subset of data for their research subjects from a particular vast database that stores digital media big data. The targeted assembling of documents is critical for "text-as-data" research, because the assembled documents can greatly influence the research results. Any mistakes made at this initial stage will propagate into the rest of the research (Kim et al., 2016). And these mistakes are often difficult to identify after the fact.

Example: Querying a Twitter archive. The School of Journalism and

> Mass Communication maintains a massive archive of tweets from Twitter. Twitter is a social networking and microblogging service, enabling registered users to read and post short messages called tweets. The vast majority (88%) of tweets are shorter than 140 characters and only 2% of them are longer than 190 characters (Rosen, 2017; Boot et al., 2019). Despite the length of tweets being short, the amount of tweet text is arguably tremendous. As of 2021, Twitter has on average 187 million daily active users (80% are outside the US) sending around 500 million tweets every day (Aslam, 2021). The database (using the Hadoop architecture) streams in 1% of the tweets daily; each day contributes about 5 million additional tweets to the Twitter archive. The enormous database provides an unprecedented opportunity for studies on civic culture, politics, social media, and mass communication, while bringing up multiple statistical challenges in collecting and analyzing vast amount of text-as-data. For example, to study the discourse centered around #MeToo on Twitter in the first four months of the movement, we can query the database using "#metoo" as keyword. This query returns a set of tweets within the time frame that contains the keyword at least once. The downstream analysis and the conclusions of the study will be based on this sample of tweets.

Targeted document assembling is to estimate the topical distribution of documents. In particular, the key is to decide for each document whether it covers the targeted topic or it covers other topics. There are multiple available methods to classify a text document by topics, varying in levels of time intensity. (i) The most expensive option is to read and manually label each document as targeted or not, based upon some criteria or human judgment. (ii) An unsupervised approaches is topic modeling by, for example, fitting the latent Dirichlet allocation (LDA) (Blei et al., 2003). This approach entitles a complication of matching the targeted topic to one of the topics identified from LDA. (iii) Supervised LDA has the potential to align the topics with the target documents (Mcauliffe and Blei, 2008), which then requires a labeled sample of documents. However, this approach requires continuous labors of labeling documents, because the database itself is growing constantly. The emerging of new topics and new usage of words presents more indefinite issues, because the targeted topics of interest often appear as new social phenomena/events or new linguistic interpretations/developments. More importantly, both (ii) and (iii) compute with the full set of text, which is fundamentally infeasible in application to a large corpus. By contrast, (iv) the keyword search approach can be think of as a semi-supervised method and is more commonly used due to its simplicity.

Keyword search

The keyword search approach represents a class of techniques for targeted document assembling. In *keyword search*, a list of keywords is determined through some process, then any documents that contain at least one keyword are included as part of the assembled documents for analysis. This approach is a popular technique for identifying which document to include, for example, in the systematic review of literature (O'Mara-Eves et al., 2015). The simple design of keyword search enjoys three obvious advantages. (i) It is very fast thanks to efficient database queries. (ii) This approach does not require topical labeling of documents in the database, thus avoiding the ambiguity of interpreting topics. (iii) In social sciences, it is usually easy for researchers to identify an initial set of "good" keywords (see below) for the targeted topic (although the set may not necessarily be complete). It is worth noting that in practice, the use of keyword search is usually coupled with other filters on the known attributes of documents, such as authors or languages. In this paper, we presume these filters to be fixed.

The quality of keywords fundamentally determines the quality of targeted document assembling. The goal of keyword search is to assemble as much targeted documents as possible while including as few other documents as possible. Qualitatively, the objective implies two characters of keywords (Airoldi and Bischof, 2016):

Frequency The keyword is widely used in the targeted topic discussion.[1] For any document that belongs to the targeted topic, the probability that it contains the keyword is higher than most other words. As such, when

[1] The "frequency" is also referred to as semantic "coherence" in Roberts et al. (2019).

searching with the keyword, most targeted documents are included likely. This is akin to the “recall” in classification.

Exclusivity The keyword is used specifically in discussing the targeted topic and less in the other topics. Any document that contains the keyword has a higher probability belonging to the targeted topic others. Hence, when searching with the keyword, most assembled documents are targeted. This is akin to the “precision” in classification.

The trade-off between these two strategies of prioritization (over words within a topic versus over the same word across topics) has been discussed in the statistical literature, such as Mosteller and Wallace (1984); Canny (2004); Airoldi and Bischof (2016); Roberts et al. (2019).

Prior studies of the keyword searching approach are extensive and mostly empirical. For example, O’Mara-Eves et al. (2015) performed a systematic review of literature reviews that used keyword searches. However, few statistical results are available for this technique; the choices of keywords are largely subjective to social scientists. In particular, (i) there is no statistical foundation for including or excluding a certain keyword in terms of the above frequency and exclusivity criteria. (ii) More importantly, the initial set of keywords is usually incomplete and needs expansion or trimming in order to satisfy the above two criteria for tageted document assembling. While the expansion can be done manually or heuristically, there are many reasons to introduce a computational method to assist the process. For example, if efficient, a computational helper allows the modification of keyword sets to be iterative and more robust.

Our contributions

This paper focuses on finding additional keywords for targeted document assembling. To this end, we study a network-based method to prioritize words. In the network of words, two words are connected if and only if they co-occur in at least one document. On the network of words, we propose to rank words using personalized PageRank (PPR) (Berkhin, 2006; Lofgren et al., 2016) and its degree-adjusted version to measure words' frequency and exclusivity to the targeted topic respectively. We call this method WordPPR, for word personalized PageRank. WordPPR is computationally efficient because it only requires the partially assembled sample of documents and does not examined the entire text database. The paper also provides statistical guarantees for WordPPR using the LDA. We argue that under the LDA, word frequency and exclusivity to a topic can be characterized by the per-topic word distribution and the per-word topic distribution respectively. With that,

> *the word PPR vector ranks words consistently by their the frequency in the targeted topic, and the adjusted word PPR vector ranks words consistently by their exclusivity to the targeted topic, provided sufficient sample documents.*

Hence, the two versions of word PPR vectors can be use to evaluate keywords to be added or deleted. We also conducted simulation studies that validate the theoretical results. Finally, through assembling tweets of the "#metoo" movement, we demonstrate that WordPPR's efficacy in modern social science researches.

5.2 WordPPR: a network approach

Personalized PageRank

PPR is a variance of Google's PageRank (Page et al., 1998). It ranks the nodes in a network by their "closeness" to a given seed node. The PPR vector quantifies such closeness for every node and is define to be the stationary distribution of the following personalized random walk. The random walk starts at a given seed node. At any step, the random walker teleports to the seed node with probability τ, or randomly goes to one of the adjacent nodes to the current node with probability $1-\tau$. Here, τ is called the *teleportation constant*. If $\tau = 0$, PPR reduces to a random walk on the graph (Pearson, 1905).

Consider a connected graph of n nodes, $G = (V, E)$, where V is the vertex set, and E is the edge set. Define the *adjacency* matrix $A \in \{0,1\}^{n\times n}$ with

```
Input: Graph adjacency matrix $A$, a set $S$ of $n_0$ seed nodes,
       teleportation constant $\tau$.
Procedure PPR($A, S, \tau$):
  1. Define $\pi \in \mathbb{R}^n$ with $\pi_i = 1/n_0$ if $i \in S$ and $\pi_i = 0$ otherwise.
  2. Initialize $p = \pi$.
  3. while not converged do
       $p \leftarrow \tau\pi + (1-\tau)P^T p$, where $P$ is the transition matrix of the
        graph.
Output: $p$
```

Algorithm 5.1: Compute the PPR vector.

$A_{ij} = 1$ if and only if there is an edge between node i and node j. In addition, define the *transition* matrix $P \in \mathbb{R}^{n \times n}$ with $P_{ij} = A_{ij}/d_i$, where $d_i = \sum_{i=1}^{n} A_{ij}$ is the ith row sum of A. For PPR, we assume without loss of generality that node 1 is the seed. Then, the PPR vector $x \in \mathbb{R}^n$ is the solution to the eigenvalue problem:

$$x^T = \tau\pi^T + (1-\tau)x^T P, \tag{5.1}$$

where $\tau \in (0, 1]$ is the teleportation constant, and $\pi \in \mathbb{R}^n$ is the *preference vector* with $\pi_1 = 1$ and $\pi_i = 0$ for $i = 2, 3, ..., n$. Let $\Pi \in \{0, 1\}^{n \times n}$ with $\Pi_{ij} = 1$ if and only if $j = 1$ and let $Q = \tau\Pi + (1-\tau)P$. From (5.1), the PPR vector is the unique left singular eigenvector of Q and is associated with the simple eigenvalue 1 and can be solved by the iterative power method (or the Richardson method), which enjoys geometric convergence property (Haveliwala, 2003; Jeh and Widom, 2003; Gleich, 2015). Algorithm 5.1 outlines the procedure that computes the PPR vector.

PPR can be defined with multiple seed nodes. The calculation of a PPR vector with multiple seeds is due to the linearity property of PPR. That is, let $p(\pi_1)$ and $p(\pi_2)$ be two PPR vectors corresponding to two preference vectors π_1 and π_2 respectively. Then, for a new preference vector that is a convex combination of π_i, the resulting PPR vector is constructive of $p(\pi_i)$,

$$p(w_1\pi_1 + w_2\pi_2) = w_1 p(\pi_1) + w_2 p(\pi_2),$$

where $w_i \geqslant 0$ and $w_1 + w_2 = 1$. As such, to get the vector, first compute the

PPR vectors for individual seeds separately, then calculate the element-wise average of PPR vectors.

PPR can also be defined in weighted graphs. Suppose the adjacency matrix contains non-negative values, $A_{ij} \geqslant 0$. On a weighted graph, the interpretation of a personalized random walk is slightly generalized. That is, conditional on the random walker continues its walk (this happens with probability $1 - \alpha$), it goes to any adjacent node with probability proportional to P_{ij} (instead of uniformly at random as in an unweighted graph).

The following proposition is useful (the proof is included in Appendix D.1 for completeness).

Proposition 5.1. *Let* p *be the PPR vector of graph* G *with the teleportation constant* $\tau \in (0, 1]$ *and the preference vector* $\pi = (1, 0, \cdots, 0)$. *Denote* $d^* = (d_1, d_2, ..., d_n) / \sum_{i=1}^{n} d_i$ *the distribution of word degrees. Then,*

(a) p *is a continuous function of* τ, *and*

$$p \overset{\tau \to 0}{\longrightarrow} d^* \quad \textit{and} \quad p \overset{\tau \to 1}{\longrightarrow} \pi.$$

(b) p *is the infinite sum of landing probability* $\{(P^s)^T \pi\}_{s=0}^{\infty}$ *with weights* $\{\tau(1 - \tau)^s\}_{s=0}^{\infty}$,

$$p^T = \sum_{s=0}^{\infty} \tau(1 - \tau)^s \pi^T P^s. \tag{5.2}$$

The WordPPR algorithm

The algorithm of WordPPR applies PPR to the co-occurrence graph of words, which is undirected and weighted. Throughout the paper, we assume the graphs are connected. The idea of using word graphs was previously motivated by the fact that the distribution of both nodes in a graph and words in a corpus follow a power law (e.g., Perozzi et al., 2014). In addition, the network of words contains rich information of how words are used in context, which is believed to define the meaning of word (Firth, 1957) In order to rank candidate words for targeted document assembling, we apply the PPR vector to the word graph. PPR ranks the nodes in a network by their "closeness" to the seed. As such, if the initial keywords enjoys frequency and exclusivity properties, we expect the other words that are close to the seed to also share the properties. In addition, the expansion of keywords can be think of as a targeted sampling from some word graphs, for which PPR has been shown to provide a consistent estimate under the stochastic block model (SBM, Karrer and Newman, 2011b; Chen et al., 2020a). Although the SMB does not necessarily characterize a graph of words, we reason the similar results under the LDA model (see Section 5.3).

To construct the word graph, we use a set of m documents that is obtained using an initial keyword search. Let $X \in \mathbb{R}^{m \times n}$ be the document-term matrix where X_{iw} is the number of word w in document i, and n is the number of unique words in all documents. We define the adjacency matrix $A \in \mathbb{R}^{n \times n}$ of word co-occurrence graph as

$$A = X^T X. \tag{5.3}$$

Here, A_{wv} is the number of documents in which word w and v co-occur.

To use PPR, we treat the initial n_0 keywords as the seeds and define the preference vector $\pi \in \mathbb{R}^n$ with $\pi_w = 1/n_0$ if word w is a keyword or 0 otherwise. Let p be the PPR vector as defined in (5.1). To assess the frequency and exclusivity, define the following two quantities $x, y \in \mathbb{R}^n$:

$$x_w = \frac{p_w - \tau \pi_w}{1 - \tau}, \quad \text{and} \quad y_w = \frac{x_w}{d_w}, \tag{5.4}$$

where $d_w = \sum_v A_{wv}$ is the degree of word i. For any word w that is not an initial keyword, x_w is equals to p_w. y simply adjusts it by word degrees.

Algorithm 5.2 summarizes the procedure of WordPPR. The algorithm uses a random sample of documents to construct the co-occurrence graph of words. This is usually a much smaller subset of the corpus and can be repeatedly use for multiple runs of Algorithm 5.2. The output includes two empirical cumulative distribution functions (ECDFs) applied to the values of x and y. To leverage the frequency and exclusivity of words, for example, Airoldi and Bischof (2016) adopted the harmonic mean to of the word's rank in the

```
Input:
    A set $S$ of $n_0$ keywords, teleportation constant $\tau$, and a sample of
    documents.
Procedure:
    1. Construct the word co-occurrence matrix $A$ as defined in (5.3)
    2. Compute the PPR vector $p \leftarrow \texttt{PPR}(A, S, \tau)$.   // Algorithm 5.1
    3. Compute two vectors $x$ and $y$ as defined in (5.4).
Output:
    The frequency and exclusivity measure $x$ and $y$.
```

Algorithm 5.2: WordPPR for finding additional keywords

distribution of x and y:

$$\text{FREX}(w) = \left(\frac{1-a}{\text{ECDF}_x(x_w)} + \frac{a}{\text{ECDF}_y(y_w)} \right)^{-1}, \tag{5.5}$$

where a is the weight for exclusivity (which is default to 0.5).

5.3 The consistency of WordPPR

Topic model

The Latent Dirichlet Allocation (LDA) is a popular generative model for text analysis or natural language processing (Blei et al., 2003; Hand and Adams, 2014). In particular, this model is widely adapted for studying document classification algorithm (Jain, 2010). LDA models the number of times that each word appears in a document (i.e., the document-term matrix), disregarding the order of words. The key idea for LDA is that there are multiple underlying topics. Each document has a probability distribution over topics, and each topic has a probability distribution over word occurrences.

Suppose there are m documents covering $k > 1$ topics and containing n unique words. Each topic t has a word distribution $\phi_t \in \mathbb{R}^n$ (that is, ϕ_t has non-negative elements which sum to one). Under the LDA, ϕ_t is sampled from the Dirichlet distribution with some sparse parameter $\beta \in \mathbb{R}^n$. In this paper, we consider the entirety of $\Phi = (\phi_1, \cdots, \phi_k^T)^T \in \mathbb{R}^{k \times n}$ as the fixed parameter. To sample one word for document i, first sample the topic $z(i)$ that this document belongs to from the multinomial distribution, $z(i) \overset{\text{i.i.d.}}{\sim}$ Multinomial(α), then sample the word from the multinomial distribution, Multinomial$(\phi_{z(i)})$. We denote $Z \in \{0,1\}^{m \times k}$ with $Z_{it} = 1$ if and only if $z(i) = t$ for notation simplicity. Blei et al. (2003) proposes that the number of words in each document could be distributed as Poisson(λ), where λ is the expected length of each document.

Remark 5.2. *In this section, we do not presume the prior $z(i) \overset{i.i.d.}{\sim}$ Dirichlet(α) in LDA, but instead suppose $z(i)$ is sampled from a Multinomial distribution. This will simplify the definition of "true class" that is required for defining targeted document. In addition, Chierichetti et al. (2018) showed the equivalence of identifying latent topics under the two settings. After this initial work, it is of interest to extend to the Dirichlet prior on $z(i)$ and explore alternative definitions of "true class," such as the maximum likelihood or any likelihood larger than some threshold.*

The LDA provides some simple quantification of word popularity and exclusivity. For example, suppose we want to study how well we can capture all documents of topic "regression", which is indexed as the first topic. For simplicity, suppose that we only have one keyword "linear", and it is indexed as the first word. Then, the probability of including document i is simply the probability that document i contains "linear". Due to the Poisson-Multinomial relationship, the number of occurrences of "linear" in the document is Poisson with rate parameter $\lambda\,[Z\Phi]_{i1}$.

Frequency: The probability of failing to include a targeted document,

$$\mathbb{P}(\text{no occurrences of ``linear''} \mid z(i) = 1) = \exp(-\lambda\Phi_{11}).$$

Here, Φ_{11} is the probability mass assigned to the word "linear" in topic "regression." To minimized such error probability, we can rank keywords using the word distribution of the targeted topic, ϕ_1 (i.e., the first row of Φ), and prioritize the most frequent words.

Exclusivity: The probability that other non-targeted documents (i.e., $z(i) \neq 1$) are mistakenly included is

$$\mathbb{P}(\text{“linear” occurs} \mid z(i) \neq 1) = \sum_{t=2}^{k} \frac{\alpha_t}{1-\alpha_1}(1-\exp(-\lambda\Phi_{t1})) \approx \frac{\lambda}{1-\alpha_1}\sum_{t=2}^{k}\alpha_t\Phi_{t1}.$$

Here, we use the approximation $1 - e^{-x} \approx x$. Note that $\lambda\sum_{t=1}^{k}\alpha_t\Phi_{t1}$ is the expected number of occurrences of “linear”, averaged across all topics. Then, minimizing the above error probability is equivalent to maximizing Γ_{11}. Here, $\Gamma \in \mathbb{R}^{k\times n}$ is the per-word topic distribution, in which Γ_{tw} is the probability that word w occurs in a document of topic t,

$$\Gamma_{tw} = \frac{\alpha_t\Phi_{tw}}{\sum_{s=1}^{k}\alpha_s\Phi_{sw}}.$$

For exclusivity, we could prioritize keywords using the first row of Γ.

In WordPPR, we use X_{iw} to indicate whether document i contains at least one word w. In this paper, we assume that X_{iw} follows a Poisson distribution,

$$X_{iw} \mid Z, \Phi \overset{\text{ind.}}{\sim} \text{Poisson}(\lambda\Phi_{z(i)w}). \tag{5.6}$$

In the following subsections, we show the statistical consistency of WordPPR under the LDA. That is, the PPR vector estimates the per-topic word distribution at the targeted topic (ϕ_1), and the aPPR vector estimates the per-word topic distribution at the targeted topic (γ_1). Throughout the discussion, we use three distinct typefaces to denote three classes of objects. Calligraphic typeface is given to the population version of any observable quantities in random graphs (e.g., Equation (5.7)). Normal typeface is given to unobserved model parameters, such as per-topic word distribution (Φ). Bold face is given to all topic-level quantities and parameters like the topic-level adjacency matrix.

A population result

Define the population (expectation) word adjacency matrix

$$\mathscr{A} = \mathbb{E}(A \mid \alpha, \Phi), \tag{5.7}$$

and let the diagonal matrix $\mathscr{D} = \operatorname{diag}(\mathscr{d}_w)$ contain the population word degrees, where $\mathscr{d}_w = \sum_{v=1}^{n} \mathscr{A}_{wv}$. Then, the population word transition matrix is $\mathscr{P} = \mathscr{D}^{-1}\mathscr{A}$. Similarly at the topic-level, define the population adjacency matrix $\mathbf{A} \in \mathbb{R}^{k\times k}$ with $\mathbf{A}_{ts} = \sum_{w=1}^{n} \mathscr{d}_w \Gamma_{tw}\Gamma_{sw}$ being the expected number of common words in the documents of topic t and s. Let the diagonal matrix $\mathbf{D} \in \mathbb{R}^{k\times k}$ contain the row sums of $\mathbf{A}$, and define $\mathbf{P} \in \mathbb{R}^{k\times k}$ to be the population transition matrix, $\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$. The following lemma writes the above population adjacency and transition matrices compactly (the proof is included in Appendix D.1).

Lemma 5.3. *Under the LDA as defined in* (5.6)*,*

(a) *let $\mathscr{A}$ and $\mathscr{P}$ be the population word adjacency and transition matrices, then*

$$\mathscr{A} = \Phi^{T}\mathbf{D}\Phi \quad \textit{and} \quad \mathscr{P} = \Gamma^{T}\Phi.$$

(b) *Let $\mathbf{A}$ and $\mathbf{P}$ be the population topic-level adjacency and transition matrices, then*

$$\mathbf{A} = \Gamma\mathscr{D}\Gamma^{T} \quad \textit{and} \quad \mathbf{P} = \Phi\Gamma^{T}.$$

Lemma 5.3 reveals the simple factorization of $\mathscr{A}$ and $\mathbf{A}$. The population word adjacency matrix is the outer product of per-topic word distribution (Φ) weighted by the expected topic sizes ($\mathbf{D}$), while the population topic adjacency matrix is the outer product of per-word topic distribution (Γ) weighted by the expected word degrees ($\mathscr{D}$). In addition, Lemma 5.3 implies a simple form of powers of the population word transition matrix, that is, $\mathscr{P}^{s} = \Gamma^{T}\mathbf{P}^{s-1}\Phi$ for $s = 1, 2, \cdots$ This leads to the relationship between the word PPR vector and the topic-level PPR vector, which we state in the following proposition (the proof is in Appendix D.1).

Lemma 5.4. *Under the population LDA with* k *topics, topic distribution* α*, and per-topic word distribution* Φ*, let* $\mathscr{p}$ *and* $\mathbf{p}$ *be the PPR vectors of the population word and topic-level adjacency matrices,* $\mathscr{A}$ *and* $\mathbf{A}$*, respectively with the same teleportation constant* $0 < \tau < 1$ *and the first word or topic as the seed. Assume that* $\Gamma_{11} = 1$*, then*

$$\mathscr{p} = \tau\pi + (1-\tau)\Phi^{T}\mathbf{p}.$$

Lemme 5.4 shows that $\mathscr{x} = (\mathscr{p} - \tau\pi)/(1-\tau)$ serves to estimate $\Phi^{T}\mathbf{p}$. Then, by Proposition 5.1, there exists a sufficiently large $\tau \leqslant 1$ such that $\mathbf{p}_1 > 1 - \varepsilon$ for any $\varepsilon > 0$. Hence, $\Phi^{T}\mathbf{p} = \phi_1 + \mathscr{O}(\varepsilon)$ is approximately proportional to the population word distribution of the target topic (i.e., the first row of Φ). If ε is sufficiently small, then ranking words by $\mathscr{x}$ is consistent with the frequency criterion under the population LDA. Next, since $\mathscr{d}_w = m\lambda^2 \sum_{t=1}^{k} \alpha_t \Phi_{tw}$ (see the proof of Lemma 5.3), adjusting $\mathscr{x}$ by the population word degrees yields an approximation of γ_1 (i.e., the first row of Γ),

$$\mathscr{y}_w = \frac{\mathscr{x}_w}{\mathscr{d}_w} = \frac{\Gamma_{1w}}{m\lambda^2\alpha_1} + \mathscr{O}(\varepsilon).$$

Hence, under the population LDA, $\mathscr{y}$ evaluates words precisely by the per-word topic distribution at the first direction (i.e., γ_1 up to a positive constant), provided that ε is sufficiently small.

WordPPR on a random corpus

This section demonstrates the consistency of WordPPR. This result is based on the concentration of the word PPR vectors under the LDA. Specifically, if the corpus is generated from the LDA (more precisely, Equation (5.6)), then WordPPR (Algorithm 5.2) ranks words by the target-topic word distribution and the per-word target-topic distribution with high probability. To this end, we first present a useful tool that controls the entrywise errors of a PPR vector in random graphs. Recall that $\mathscr{p}$ is the stationary distribution of probability transition matrix $\mathscr{Q} = \tau\Pi + (1-\tau)\mathscr{P}$. For any vector $x \in \mathbb{R}^n$, define the vector infinity norm as $\|x\|_\infty = \max_i |x_i|$. The following theorem bounds the entrywise error of the stationary distribution of $\mathscr{Q}$ (the proof is included in

Appendix D.1).

Theorem 5.5 (Concentration of the PPR vectors)**.** *Let* P *be the word transition matrix of* n *unique words in* m *sample documents generated from LDA (Equation (5.6)), where the average expected document length is* λ*. Define* $\mathscr{P}$ *to be the population version of* P*. Let* p *and* $\mathscr{p}$ *be the PPR vector corresponding to* P *and* $\mathscr{P}$ *respectively, with the some sufficiently large teleportation constant* τ*. Assume that* $\rho = \frac{\max_{1 \leqslant w \leqslant n} \mathscr{d}_w}{\min_{1 \leqslant w \leqslant n} \mathscr{d}_w}$ *is bounded by some finite constant and that* $m > c_0 n \log n / \lambda^2$*, for some sufficiently large constant* $c_0 > 0$*. Then, with probability at least* $1 - \mathscr{O}(n^{-5})$*,*

$$\frac{\|p - \mathscr{p}\|_\infty}{\|\mathscr{p}\|_\infty} \leqslant c_1 \sqrt{\frac{n \log n}{m \lambda^2}},$$

for some sufficiently large constant $c_1 > 0$*.*

Theorem 5.5 demonstrates that if the number of sample documents m exceeds $n \log n / \lambda^2$ to some sufficiently large extent, then with high probability, the random PPR vector concentrates around the population PPR vector entrywisely. The theorem also require the ratio between the most word occurrence and the least word occurrence is sufficiently large. This is achievable because in practice, we often pre-process the unique words by filtering out extremely common and extremely rare words. For example, Grinberg et al. (2019) suggest to ignore those words that occur in (i) fewer than 0.02% or (ii) more than 90% of the sample documents.

Next, we give the exact recovery of top frequent words; one can draw the exact recovery result of top exclusive words analogously. Specifically, suppose we select the top w words ($1 < w < n$) with the largest x values defined in 5.4. Then, with high probability, these w words are exactly the top w words with the largest values in $\phi_1 \in \mathbb{R}^n$. Without loss of generality, assume the words are ordered by ϕ_1 in non-increasing order. A key challenge is to distinguish the wth and the $(w+1)$th words, as the difference in the two word characterizes the distance between the first n_1 words and the others. Only if the two words are sufficiently separated (in ϕ_1), can the exact recovery be possible using a finite sample algorithm. For this, we introduce a *separation measure* $\Delta_w \in (0, 1]$

of words $(1 < w < n)$ under the population LDA as

$$\Delta_w = \frac{\mathcal{p}_w - \mathcal{p}_{w+1}}{\mathcal{p}_{\max}}.$$

This turns out to be crucial in determining the sample complexity required to guarantee the exact recovery. With Theorem 5.5 and the separation measure, we then give following corollary that bounds the accuracy of Algorithm 5.2, in terms of graph edge density (the proof is included in Appendix D.1).

Corollary 5.6 (Exact recovery of frequent words)**.** *Suppose the corpus is generated from the LDA* (*or from Equation* (5.6) *of document-term matrix*) *with* m *documents and* n *unique words. Let* $\mathcal{S}$ *be the top* w *words in the targeted-topic word distribution* ϕ_1*. Let* S *be the top* w *words in* x *as computed in Algorithm 5.2 with a sufficiently large teleportation constant* $\tau < 1$ *and a seed keyword* $w_0 \in \mathcal{S}$ *exclusive to the targeted topic. Assume that* $\rho < c_0$ *and that*

$$m > \frac{4c_1 n \log n}{\lambda^2 \Delta_w^2}, \tag{5.8}$$

for some sufficiently large constants $c_0, c_1 > 0$*. Then,* $S = \mathcal{S}$ *with probability at least* $1 - \mathcal{O}(n^{-5})$*.*

5.4 Simulation studies

We assess the accuracy of WordPPR in identifying the top frequent and exclusive words, given a seed word and a sample of documents simulated from the LDA.

We sampled document-term matrix from the LDA with 10 underlying topics and 500 unique words. Specifically, we set up equally distributed topic, $\alpha = (0.1, \cdots, 0.1)$. The per-topic word distribution ϕ_t is generated from the Dirichlet distribution, $\phi_t \overset{\text{ind.}}{\sim} \text{Dirichlet}(\beta)$ for $t = 1, 2, \cdots, 10$, where the elements in $\beta \in \mathbb{R}^{500}$ are sample from the exponential distribution, Exponential(10). To investigate the effects of the sample size and expected document length, we experimented the parameter grids with $m \in \{1000, 2000, 5000, 10000, 20000\}$ and $\lambda \in \{5, 10, 20\}$. In addition, we study the effect of teleportation constant with $\tau \in \{0.25, 0.5, 0.75\}$. For each combination of parameters, we simulated the document-term matrix for 30 times and evaluated the WordPPR algorithm (Algorithm 5.2) independently.

To evaluate the word frequency and exclusivity measurements of WordPPR, we examined the top words ranked by x and y (Equation (5.5)) respectively. Specifically, let $S(x, n_2)$ to be the set of top-n_2 words (or indices) in x. Then, the accuracy of frequency measure over the n_2th most frequent words is defined as

$$\text{Accuracy}(x, \phi_1, n_2) = \frac{1}{n_2} \left| S(x, n_2) \cap S(\phi_1, n_2) \right| .$$

Similarly, the accuracy of exclusivity measure over the n_2th most exclusive words is defined as $\text{Accuracy}(y, \gamma_1, n_2) = \frac{1}{n_2} |S(y, n_2) \cap S(\gamma_1, n_2)|$. In this simulation, we set $n_2 = 25$.

Figure 5.1 displays the accuracy of WordPPR's frequency and exclusivity estimates, for five different document sample sizes, three different expected document lengths, and three different teleportation constants. The results show that WordPPR can identify both frequent and exclusive words. Moreover, both the increase of number of sample documents (m) and the increase of expected document length ("lambda") help the accuracy of WordPPR. In addition, the choice of teleportation constant ("tau") appears to have little effect on the accuracy of WordPPR, except in the top left panel. This agree with the fact that Corollary 5.6 only requires τ to be sufficiently large in a finite graph. Asymptotically (i.e., as n grows), this requirement on τ should reduce.

5.5 Targeted tweet assembling of the #MeToo movement

We collected a random 5% sample of all tweets from the random 1% Twitter archive within the four months since the start of #MeToo on Twitter. This resulted in 4,491,833 tweets that might cover all topics on Twitter during that time frame. After removing punctuations, numbers, stop words, we constructed a word graph, where nodes are either hashtags or bigrams in text and edges represent the co-occurrence of nodes in a tweet. The WordPPR

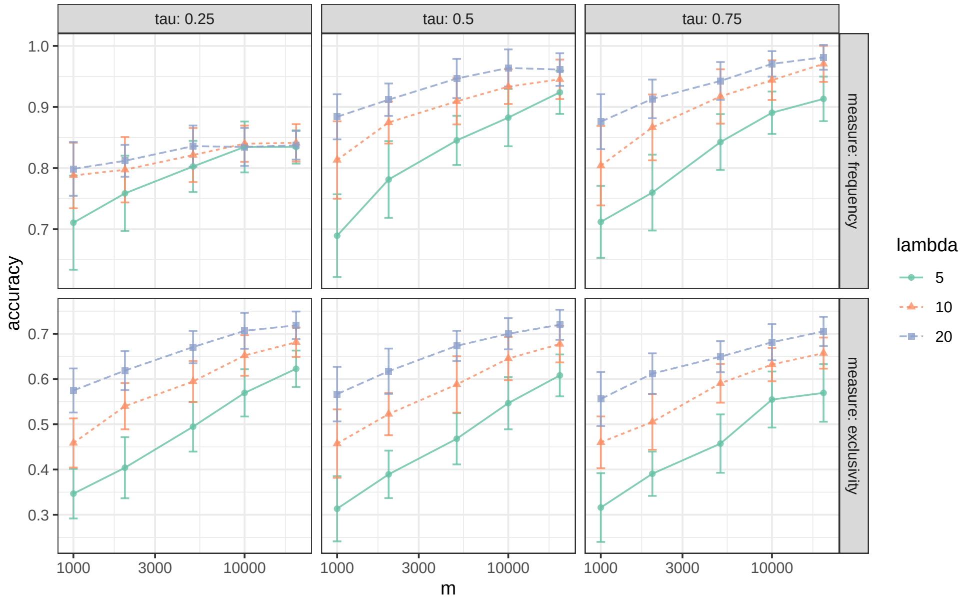


Figure 5.1: Accuracy of frequency and exclusivity measures of WordPPR. Column panels compares the effect of different teleportation constant (tau). Row panels compare different word prioritization (method). The different colors indicate different expected document length (lambda) used for generating the document samples. Each point depicts the average accuracy of including the top 25 words. The error bars indicates two times of the standard deviation of accuracy across 30 repeated experiments.

method applied to this word graph yields two sets of results: a list of terms (i.e., unigrams and bigrams) ranked by frequency and another list ranked by exclusivity. Table 5.1 displays the top 20 terms ranked by frequency and exclusivity.

Both lists seem to suggest highly relevant terms about the #MeToo movement on Twitter. Some terms concern the issues taht this movement seeks to address, like “sexual assault,” “sexual harassment,” “sexual abuse,” “sexually assault,” “sexual predator,” and “sexual misconduct.” Some other terms are related to victims speaking up by sharing their own stories (“share story”) and Times’s Person of the year that honored people who came forward to report sexual violence (“person year”). These terms seem highly promising as the seed terms for targeted data collection. Other terms include more generic terms like “#rt,” “#follow,” “#goldenglobes” and “#women” as well as names of accused public figures, like “harvey weinstein” (the accusations of whom inspired people coming forward to expose similar behaviors), “donald trump,” “roy moore,” and “bill clinton.” Such terms might not be ideal as search terms for targeted document assembling because they might be used in other contexts and thus bring much irrelevant tweets/documents. Going down the lists, the terms become increasingly generic. Therefore, for such results, researchers might want to focus on the top 20-30 results.

Also, it can be seen that “exclusity” tends to uprank terms that are specific to the movement more than “frequency” does, though the top 5 terms on the two lists are exactly the same. On the “frequency” list, generic terms like “#rt,” “#follow,” “#goldenglobes,” and “social medium” are ranked 6th to 10th, while on the “exclusivity” list, the generic terms like “#goldenglobes”(6th), “golden globe”(12th), “#follow”(14th), “social medium”(15th), and “#rt”(16th) are ranked lower.

To compare the results obtained through the random sample, we collected a targeted/“metoo” sample from the same archive using the seed keyword “metoo,” ending up with 38,177 tweets from October 15, 2017 to February 15, 2017. Then we applied the same text procedures as above and WordPPR. Results are shown in Table 5.2.

The difference between the two lists is similar to the difference above. The top 4 terms ranked by frequency are “#metoo,”“#timesup,” “sexual harassment,” and “sexual assault.” These terms are relevant to the #Metoo movement and the #Timesup movement it inspired, as well as to the underlying issues that these movements seek to expose and address. In a similar vein, the other top terms are salient hashtags representing similar movements in other parts of the world, like “#abusefreeindia” (India), “#balancetonporc” (France), “#withyou” (Korea), and terms that concern the substantive issues, like “sexual abuse” and “share story.” These terms can reasonably be used as keywords to collect more tweets in the next round of data collection. However, as the above results, this list also contains noise. The fifth to the seventh most frequent nodes, “#follow,” “#change,” and “#howto,” might reasonably be used by users to

	frequency	exclusivity
1	#metoo	#metoo
2	sexual assault	share story
3	sexual harassment	sexual harassment
4	share story	person year
5	person year	sexual abuse
6	#rt	sexual assault
7	#follow	sexually assault
8	#goldenglobes	#goldenglobes
9	social medium	golden globe
10	sexual abuse	speak truth
11	sexually assault	sexual predator
12	black woman	sexual misconduct
13	year ago	#women
14	donald trump	white woman
15	#women	harvey weinstein
16	sexual predator	black woman
17	sexual misconduct	donald trump's
18	#maga	spend year
19	golden globe	locker room
20	roy moore	bill clinton

Table 5.1: Top 20 word ranked by WordPPR with $\tau = 0.75$ using a random sample of tweets.

support victims of sexual assault and abuse by following them and expressing the determination to change or push for change. However, they lack specificity and can also be widely used in a wide range of other settings. This also applies to terms like "#ifb" (ranked 9th), "pm report" (15th), "#goldenglobes" (16th) that also ranked high among the top 20 terms. Other terms down the list display a similar mixture of specific terms and generic terms.

Among the top 20 terms on the list ranked by exclusivity, the majority of terms are specific to the "#Metoo" discourse. Besides "#metoo,""#timesup," "#withyou" (Korea), "#balancetonporc" (France), they include additional hashtags representing the international "#Metoo" movements, "#wetoo" (Japan), "#yotambi"/"#yotambien" (the Spanish-speaking world), and "#moiaussi"(France). Many other terms are the same with those appearing among the top 20 ranked by frequency, like “sexual harassment” and “sexual assault.” However, terms that might bring noise if used to collect more data are ranked much lower: “#goldenglobes” (12th), “social medium”(14th), “#oscars”(16th), “#svpol”(17th) and “golden globe”(20th).

A comparison of results across the two samples shows that “frequency” and “exclusivity” prioritize different terms. In practice, they can both be helpful for researchers to determine the keywords that can be used for data query. However, understanding their distinct qualities can help researchers make better decision.

It is also noteworthy that the results based on the random sample contain fewer relevant terms than the results based on the “metoo” sample do. Going further down the two lists produced by the “metoo” sample, a lot of terms still seem to be promising candidates for the next round of data collection. For example, “#silencebreakers,” “#fightforshiori” (the Japanese journalist who spoke out about her abuse), “#believewomen,” and “#himthough”(asking men to share the responsibility for fixing the issue) all appear beyond the top 100 terms. In contrast, the terms after the top 50 in the random sample results become largely irrelevant to the #MeToo movement. Those country-specific hashtags like “#withyou” and “#balancetonporc” are not present in such results. This suggests that the targeted/“metoo” sample might yield more relevant terms for further search than the random sample does.

5.6 Discussion

In this paper, we present the WordPPR, a new method to rank words in terms of their popularity (or frequency) and exclusivity to a given targeted topic. We show that under the LDA, the WordPPR estimates is consistent, provided that the sample documents is sufficient. WordPPR is computationally efficient and can be used to find additional keywords. It can also be used to validate and diagnose existing keywords.

While the LDA offers a simple and interpretable test bed for the method, it does not capture several features that are unique to the text data in online social media like Twitter. For example, tweets are short, which means fewer

	frequency	exclusivity
1	#metoo	#withyou
2	#timesup	#metoo
3	sexual harassment	#balancetonporc
4	sexual assault	#wetoo
5	#follow	#timesup
6	#change	sexual harassment
7	#how to	share story
8	#abusefreeindia	#yotambi
9	#ifb	harvey weinstein
10	share story	sexual assault
11	#stopabuse	#yotambien
12	#balancetonporc	#goldenglobes
13	#withyou	sexually assault
14	person year	social medium
15	pm report	sexual abuse
16	#goldenglobes	#oscars
17	sexual abuse	#svpol
18	sexually harass	#moiaussi
19	sexually assault	#sexualharassment
20	#sexualassault	golden globe

Table 5.2: Top 20 word ranked by WordPPR with $\tau = 0.75$ using tweets that contain "metoo."

(sparser) word occurrences (data) than longer documents. In addition, due to "retweeting", many tweets appear multiple times, sometimes thousands of times, in the corpus, which means documents (samples) are not independent. Last but not least, many tweets have extremely simple word distribution, such as "cat cat cat cat" which repeats a single word, which means the occurrence of individual words in a document are not independent. These facts are not hardly represented by the LDA. As such, it is of future interest to investigate WordPPR under alternative, more sophisticated models.

6 SOCIAL MEDIA PUBLIC OPINION AS FLOCKS IN A MURMURATION

6.1 Introduction

As the cornerstone of democracy, public opinion has been predominantly treated as mass opinion–an aggregate of individual opinions gathered by survey-based public opinion polls (Gallup and Rae, 1940). Though a powerful measure of the pulse of the public, this approach tends to yield a snapshot of private preferences that are prompted by pollsters and contingent on the artificial context of polling, thus overlooking the social context and the uneven influence of opinion expression (Blumer, 1948; Lin et al., 2013; Zaller et al., 1992). The drawbacks of survey-based polls are amplified by increasing non-response rates (Groves and Peytcheva, 2008) due to factors like changing patterns of technology use and public distrust in polling.

Social media platforms like Twitter have emerged as one key battleground of public discourse, where people from different backgrounds actively comment on current events and public issues, and strive to exert influence (Conway et al., 2015; Tufekci, 2013; Kim et al., 2015). This leads to naturally occurring, temporally sensitive, and inherently social opinions (Anstead and O'Loughlin, 2014; Boyd, 2010; McGregor, 2019), which are drastically different from those gathered by survey-based opinion polls. Another key aspect of social media public opinion is the various homogeneous networks that it is embedded in, where like-minded individuals interact with each other and reinforce opinions (Colleoni et al., 2014; Conover et al., 2011; Barberá et al., 2015; Sunstein, 2018). Although social media users are not representative of the general public (Barberá and Rivero, 2015; Wojcik and Hughes, 2019), those actively engaged in opinion expression on social media can shape public opinion (Dubois and Gaffney, 2014; Lasorsa et al., 2012). More importantly, blended into individual day-to-day practice and the social world (Becker et al., 2010; Couldry, 2012; McGregor, 2020; Tufekci and Wilson, 2012), social media are an important public opinion domain in and of itself.

Existing studies that leverage social media to studying public opinion have primarily used natural language processing of text to identify patterns of expressions, such as sentiments or topics (Bollen et al., 2011; Cody et al., 2015; Tumasjan et al., 2011), and to compare the results with survey-based opinion polls (O'Connor et al., 2010). However, as the conception of public opinion is deeply intertwined with tools of opinion measurement (Herbst, 2001; Zaller, 1994), a blunt comparison between social media and survey-based opinion polls can be misleading. Furthermore, the text-centric approach fails to take full advantage of social media data to reveal the networks that opinions are embedded in and the social and conversational aspects of public opinion (Anstead and O'Loughlin, 2014), such as "Who talks about which events?" and "How are they talking about these events?" The "who" and "how" questions are especially important to address as a myriad of actors, ranging from social movement activists to propagandists, use social media to influence public opinion (Freelon et al., 2018; Tucker et al., 2018). Some studies have moved beyond text to account for characteristics of social media users, which can be detected with high accuracy (Kosinski et al., 2013; Pennacchiotti and Popescu, 2011). For example, Twitter accounts have been selected as "computational focus groups" based on hashtag use to map shared attention (Lin et al., 2014) or classified into hierarchical groups based on Twitter lists to trace opinion flow (Wu et al., 2011).

Here, we introduce a framework called "murmuration" for the study of public opinion on social media. Given homophily driving friendship formation (McPherson et al., 2001; De Choudhury, 2011) and abundant empirical evidence for the effectiveness of social network structure (i.e., friendship relations) in predicting individual characteristics (Al Zamal et al., 2012; Barberá et al., 2015; Grabowicz et al., 2012; Pan et al., 2019; Pennacchiotti and Popescu, 2011), this framework uses social network structure to computationally identifies focus groups, which we call "flocks" (drawing on the idiom "birds of a feather flock together"). We expect social media public opinion to exhibit homogeneity within a given flock or similar flocks and heterogeneity across different flocks. Therefore, the unfolding of the opinions of various flocks on social media in response to external events is akin to a murmuration of starlings whose formation changes fluidly. A conceptual illustration of the framework is shown in Fig. 6.1, which includes network sampling of targeted accounts, identification of flocks, analysis, and presentation of flock opinions over time.

We apply this framework to the case of political opinion leaders on Twitter, a unique social media platform featuring political commentary and debate and boasting instantaneous response to external events (Hu et al., 2012). As information on Twitter primarily flows from elites to average accounts (Wu et al., 2011), even within echo chambers (Min et al., 2019), opinion leader flocks in the Twitter sphere is essential for understanding general opinion climate on Twitter. By analyzing social network structure and opinion expression, we provide empirical evidence for (1) flocks being homogeneous, interactive, and stable networks on Twitter and (2) flocks predicting opinion expression. Using three distinct news events, we demonstrate the power of our approach in capturing (a) the intensity and (b) temporal dynamics of opinion expression by flocks and (c) the opinion contestation among them. These results demonstrate how the murmuration framework can reveal the social dynamics of public opinion and increase our understanding of social media as the battleground of public discourse.

6.2 Results

The murmuration framework

Our framework for large-scale measurement of social media public opinion contains three analytic modules and one presentation module, the four of which form a cyclic working system (Fig. 6.1). The first two modules are executed infrequently, and the last two modules are performed on a daily basis.

The first module—targeted sampling (Fig. 6.1a)—samples from a targeted population using seed accounts known to be highly influential in the population. We apply personalized PageRank (PPR) sampling to obtain a targeted subset of the massive Twitter friendship network (Materials and Methods).

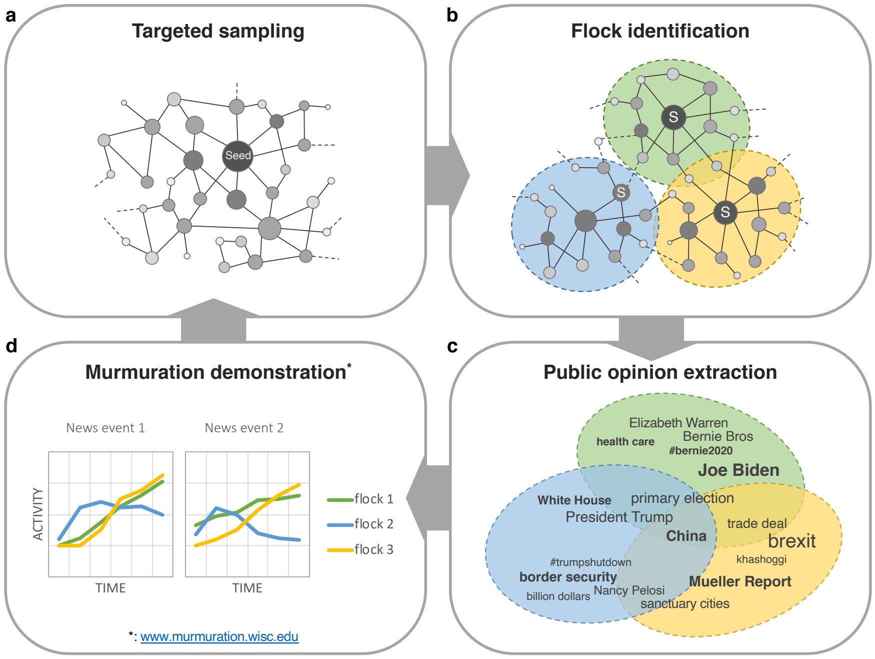


Figure 6.1: An overview of the murmuration framework. (a) Targeted sampling: the first module uses a set of seed nodes, queries the massive Twitter friendship network, and returns a targeted subset of Twitter accounts. (b) Flock identification: based on the friendship network, the second module identifies flocks among sampled Twitter accounts. (c) Public opinion extraction: the third module analyzes public opinion at the flock level. (d) Murmuration demonstration: the fourth module presents analytical results on a website. Over time, we identify newly emerged seed nodes and update the targeted sampling.

Under the degree-corrected stochastic block model (DCSBM), PPR sampling can consistently locate members of a targeted population (Chen et al., 2020a). This excludes low influence accounts such as bots and spammers, establishing a solid foundation for downstream analysis. The second module—flock identification (Fig. 6.1b)—detects the underlying community (or "flock") of Twitter accounts based on their following relationships. We apply a spectral method called vintage sparse principal component analysis (VSP) (Rohe and Zeng, 2020; Chen and Rohe, 2020), which can effectively cluster millions of Twitter accounts in less than an hour (Materials and Methods). Under the DCSBM (a variant of the mixed-membership stochastic block model), VSP provides a consistent estimate of community memberships (Rohe and Zeng, 2020). We interpret each flock based on the profile descriptions of their members and then treat flocks as the unit of analysis in downstream analysis. The third module—opinion extraction (Fig. 6.1c)—analyzes daily opinions from the sample. Given the Twitter flocks identified in the previous module, we collect tweets from each, identify trending news events as observed in tweets, and analyze if and how different flocks respond to the events. This module presents a timely digest of flock opinions, and over time it offers a unique window into public opinion dynamics. The fourth module—murmuration demonstration (Figure 6.1d)—presents social media public opinion by flock in response to major news events. Analytical results are updated daily and made available on a website.

Flocks are homogeneous, interactive and stable networks

In this section, we provide empirical evidence for flocks detected via following relationships as homogeneous, interactive, and stable networks. For the reasons discussed above, we focus on political opinion leaders on Twitter. In August 2018, we performed network sampling of political opinion leaders using a curated list of Twitter accounts including activists, pundits, journalists, and media outlets spanning the whole political spectrum in the United States (supporting Table E.1). We obtained a total of 193,120 Twitter accounts, which followed a total of 1,310,051 accounts (after filtering, Materials and Methods).

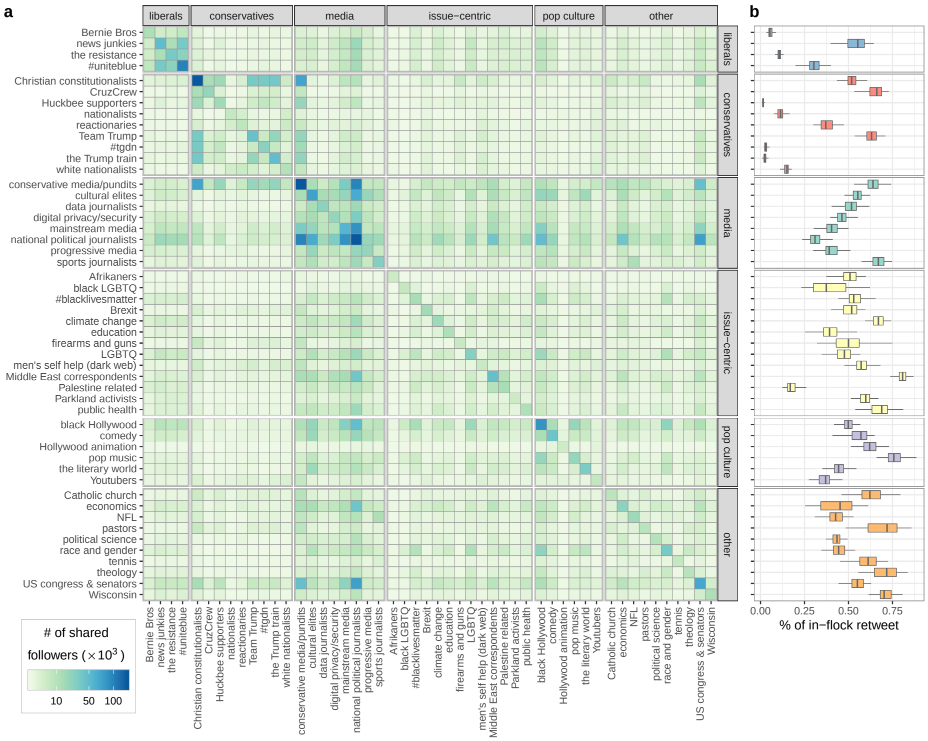


Figure 6.2: Shared followers and retweeting are concentrated within flocks. (a) Heat map of the number of shared followers among flocks. Each row and column corresponds to one flock in the same order (i.e., the shown matrix is symmetric). Rows and columns are grouped into panels by flock category, with strips on the top and right indicating the categories. The shade of color is determined by the number of shared followers between pairs of flocks. (b) Box plots showing the distribution of in-flock retweeting percentages (i.e., for each member of a flock, the percentage of retweeting that he/she initiated of tweets from another flock member was calculated). The box plots align horizontally with the rows in (a).

Based on the observed social network, we then identified 100 flocks (i.e., communities of accounts followed by the political-related accounts), which cover various social, cultural, political and geographical entities (supporting Table E.2). We excluded most regional flocks and selected 50 flocks of interest for downstream analysis, including media flocks, partisan flocks, issue flocks and non-political flocks (supporting Table E.2). In addition, we considered the 1000 most central accounts from each flock (Materials and Methods) to control for the effect sizes of individual flocks. We evaluated the effectiveness of flock identification through (i) shared followers, (ii) retweeting network, (iii) stability and fidelity of flock membership. Taken together, evidence indicates that our approach to flock identification discovers meaningful networks with high accuracy and resolution.

First, member accounts of a flock demonstrate homogeneity because they have more shared followers than do accounts from different flocks. As followers of a Twitter account constitute its imagined audience with whom in mind it crafts messages (Litt, 2012), similar accounts should attract similar audiences. Aggregating the number of shared followers between any pairs of accounts, as observed in our sample, we found marked more followers were shared by members of the same flock (Fig. 6.2a). In addition, flocks of the same category (supporting Table E.2) also shared more followers (e.g., “mainstream media” and “national political journalists” under the “media” category), revealing inter-flock structure. To quantify this pattern, we calculated an “in-and-out ratio” to measure the average number of shared followers by two accounts within a flock over the average number of shared followers by one from the flock and one outside it (Materials and Methods). Overall, we observed an average in-and-out ratio of 15.628 across 50 selected flocks, with a minimum of 5.52. Notably, an account of the “#uniteblue” flock shared on average 35.2 fold more followers with accounts within the flock than with accounts outside the flock; similar results hold for the “Christian constitutionalists” and “national political journalists” flocks with 31.4 and 20.2 folds respectively.

Second, interaction in the form of retweeting is concentrated among member accounts of a flock, showing the similarity between flocks and offline social networks where interactions are localized (Grabowicz et al., 2012). We constructed a random sample of tweets, quantified the proportion of retweeting that occurred between accounts within each flock (Materials and Methods) and found that the retweeting network was consistent with the flock structure. Among the 7,379,555 retweeting relationships between all accounts in the 50 flocks, on average 44.1% were between accounts within a flock, with the "Brexit" flock having as high as 80.8% of within-flock retweeting (Fig. 6.2b). In fact, we found strong statistical evidence in the correlation between an account retweeting other accounts in the 50 flocks and the retweeted post originating from other accounts within its own flock: p-value $< 2.2 \times 10^{-16}$ in χ^2 test, after multiplicity correction with Benjamini-Hochberg (BH) procedure. For the flocks with low levels of within-flock retweeting, we found that they retweeted a large number of tweets from flocks of the same category. For example, 50.6% of retweeting by "#uniteblue" was of accounts in similar flocks, i.e., flocks under the "liberals" category; and 52.1% of retweeting by "Christian constitutionalists" was of accounts under the "conservatives" category (supporting Figure E.1). Given that retweeting reflects existing ties or is conducive to new tie formation (Golder and Yardi, 2010), such evidence might further suggest redundant friendship ties between flock members.

Third, the flock structure we identified is stable and consistent even after one year. Unlike fluid networks organized by communication (Bennett and Segerberg, 2013), flocks, based on the following relationships, should be relatively stable. This means that despite Twitter users' ability to freely follow additional accounts or unfollow existing ones, flock members should exhibit relative consistency in accounts they follow, which we investigate here. To this end, we ran murmuration modules 1 and 2 (Fig. 6.1ab) in August of 2018 and 2019 separately; then we compared the flock identification results, based on the 100 flocks from 2018 and 2019. Specifically, we evaluated each of the 100 flocks on its (i) stability: the percentage of flock members that remained in the sample after one year and (ii) fidelity: the percentage of recurring flock members that fell into a similar flock. We first observed that flocks exhibit stability. On average, 60.3% (median 71.6%) member accounts across the 100 flocks of 2018 recurred among the 100 flocks of 2019 (Fig. 6.3a). Particularly, 68 flocks in 2018 saw more than half of their members reappear after one

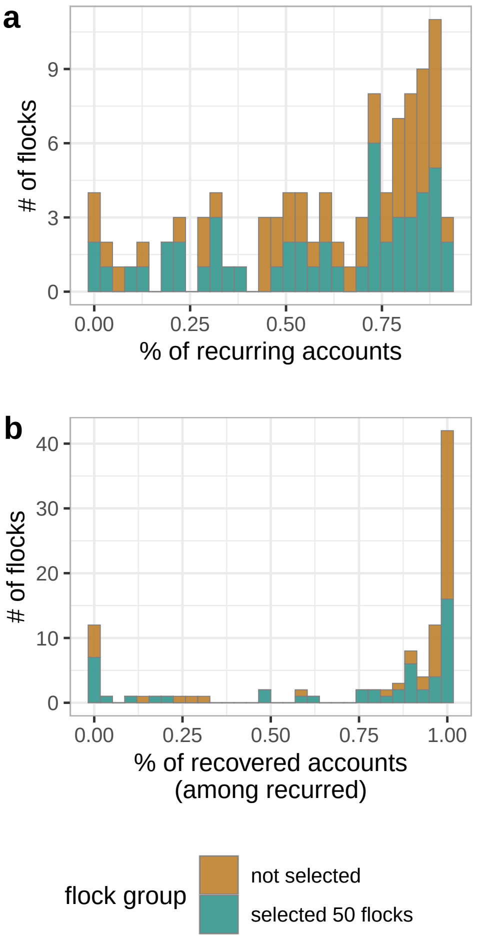


Figure 6.3: The percentage of flock members that recurred and recovered after one year. (a) Histogram of the percentages of flock members in the 2018 flocks that remained in the new sample in 2019. (b) Histogram of the percentages of recovered accounts in the 2018 flocks, i.e., accounts reappearing in a similar flock in 2019. In both panels, the 100 flocks are stratified by whether they belong to the 50 flocks that we selected for downstream analysis in 2018.

year and only 18 flocks less than 30%. The fidelity of flocks provides further assurance for flock stability. Among the 2018 accounts that reappeared in 2019, on average 75.9% fell into a similar flock (Materials and Methods). In particular, as many as 60 flocks in 2018 matched a new flock of 2019 (with more than 90% shared account) (Fig. 6.3b).

Flocks predict opinion expression

As shown above, member accounts of a flock are similar, interactive, and embedded in a stable social network structure, suggesting that a flock is a meaningful network organized on social media. Situated in such a group context, accounts within a flock are expected to share topical emphases in opinion expression, which we examine in this section. For this analysis, we again relied on our random sample of tweets. Given that hashtags are semantic markers of full tweets, we focused on the pattern of hashtags used across the 50 flocks. The most frequently used hashtags were grouped into 6 categories and their occurrences in tweets were computed by flock (Materials and Methods). For illustration, we present the use of selected hashtags in Fig. 6.4. Overall, we found a high level of correspondence between hashtags and flocks that used them, suggesting the predictability of opinion expression by flock membership. Hashtags presumably used by liberals appeared most frequently in liberal flocks' tweets. Similarly, hashtags often used by conservatives to indicate conservative values, or by Trump supporters to show their allegiance, or by conspiracy believers, appeared most frequently in conservative flocks' tweets. So were the issue- and topic-specific hashtags: #syria and #iran were overwhelmingly used by "Middle East correspondents." Hashtags even validated the distinction between similar flocks: #bernie2020 and #notmeus were nearly exclusively used by the "Bernie Bros" flock on the liberal side. These results are consistent with previous research showing the similarity between friendship ties and tweets (Aiello et al., 2012). However, some seemingly discriminative hashtags failed to neatly align with their corresponding flocks. For example, the use of #maga, a hashtag presumably indicating support of Trump's presidential campaign, was split among liberal, conservative and Trump supporter

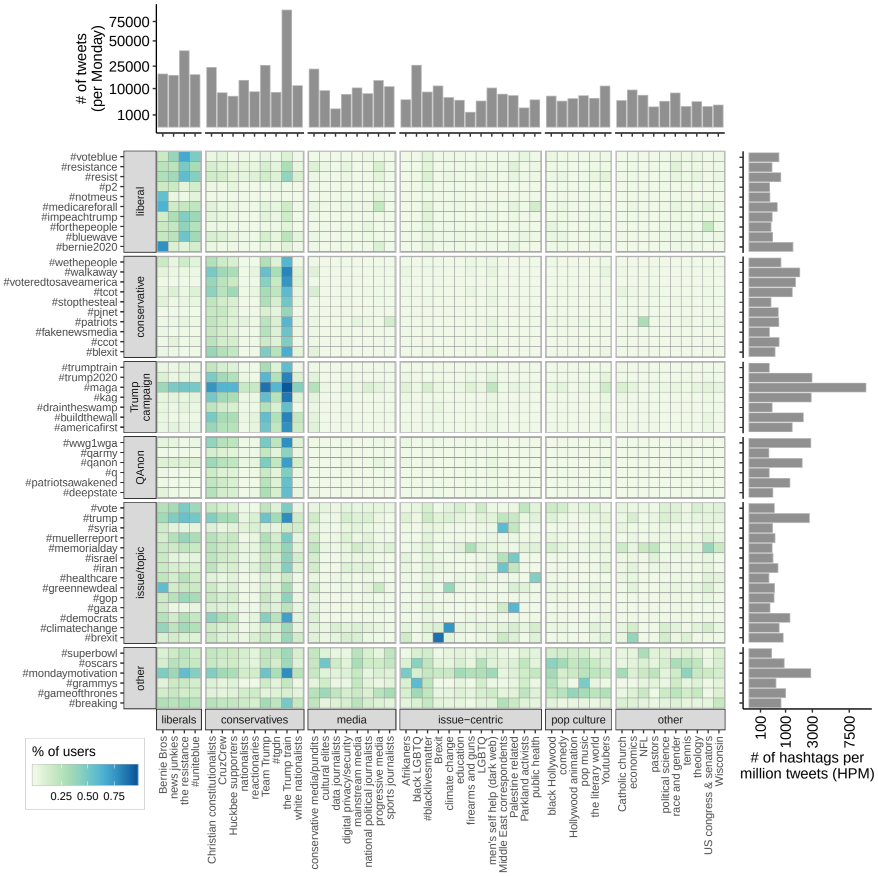


Figure 6.4: Heat map of 53 hashtags frequently used by 50 flocks. Each column corresponds to one flock, with column panels indicating flock category and column strips on the bottom indicating the category name. Each row corresponds to one hashtag, with row panels indicating the hashtag category and row strips on the left indicating the category name. The shade of color indicates the percentage of active accounts in the flock that utilized the hashtag. The bar plot above the heat map reports the number of daily tweets from each flock; the bar plot on the right reports the number of hashtags observed per million tweets collected.

flocks. Similarly, #resistance and #resist, used to express opposition toward the Trump presidency, also appeared saliently in tweets from "the Trump train" flock. Such idiosyncratic hashtags being used by heterogeneous flocks might be explained by hashjacking, a practice of infiltrating into opponents' networks (Bode et al., 2015). This suggests that the nuance that cannot be picked up by patterns of hashtag use can be revealed through the social network structure encoded in flocks. In addition, we fit a topic model, treating all of an account's tweets as a single document. A similar pattern was observed between the actual topics of tweets and flocks: topics were generally consistent with flock membership (Appendix E.2).

Measuring social media public opinion with flocks

This section presents an example of social media public opinion in response to news events with flocks, focusing on the intensity, temporal dynamics, and difference of opinion expression by flocks. To illustrate, we chose 10 flocks from different rungs in the influence hierarchy, including more influential media flocks and less influential activist flocks (Table 6.1), and collected tweets corresponding to three news events for analysis (Materials and Methods). Our website `www.murmuration.wisc.edu` updates every day to summarize how flocks discussed yesterday's events.

The three news events were selected to balance liberal and conservative political issues: (1) the concluding phase of the Mueller investigation (March 1, 2019 to July 31, 2019), (2) the passing of anti-abortion laws by several states (March 1, 2019 to May 31, 2019) and (3) the killing of the Washington Post journalist Jamal Khashoggi (November 1, 2018 to December 31, 2018).

We first investigated the intensity of opinion expression by all 10 flocks in response to the three events (Fig. 6.5ab). Though in general "the Trump train" and "Christian conservatives" on the conservative side and the "#unite-blue" and "the resistance" on the liberal side were the most active, the pattern of the 10 flocks' expression intensity varied across events. For the Mueller investigation, the three conservative flocks and the three liberal flocks were nearly equally engaged in talking about the investigation, accounting for 45.6%

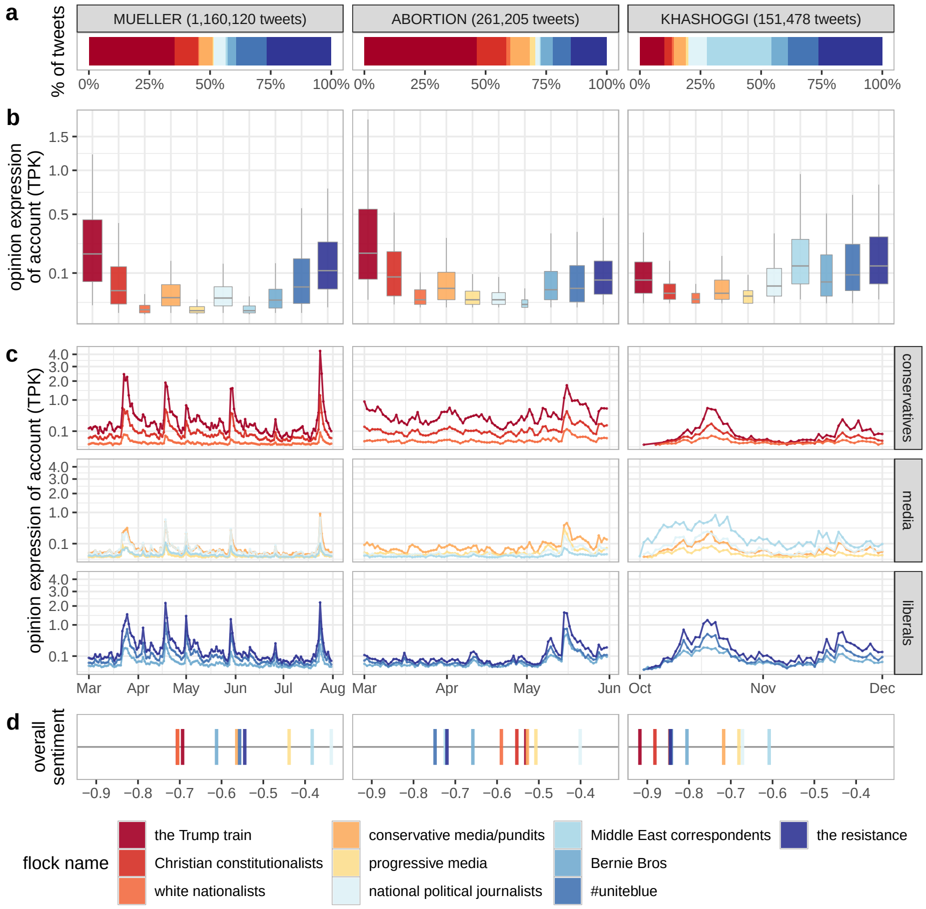


Figure 6.5: Opinions of 10 flocks about three news events. Throughout, each column panel corresponds to one event, and each color corresponds to one flock. (a) The percentage of tweets contributed by each flock. The top strips show the names of events and the brackets the total number of tweets collected. (b) Box plots of the opinion expression of account, as measured by the number of tweets per thousand event tweets (TPK, Materials and Methods). Each account's TPK is averaged across all days in the event period. The widths of boxes indicate the number of accounts that expressed opinions. (c) The opinion expression of account (in TPK) averaged across individual flocks as a function of days during the three events, stratified by flock category. Right-hand side strips indicate the flock category. (d) Weighted average sentiment observed in opinions on each event by each flock. Each vertical line represents one flock.

Table 6.1: Ten exemplary flocks of political relevance, as of August 2018.

Category	**Name**	**Description**
conservatives	“the Trump train”	Vowing clear and strong support for Donald Trump
	“Christian constitutionalists”	Showing firm conservative beliefs yet do not explicitly express solidarity with a specific political figure
	“white nationalists”	Espousing beliefs in ethnocentrism and nationalism
liberals	“Bernie Bros”	Alleging support for Bernie Sanders
	“#uniteblue”	Promoting progressive causes and values
	“the resistance”	Opposed to the Trump presidency
media	“conservative media/pundits”	Appealing to conservative partisan audience, e.g., Stephen Miller, Ben Shapiro, and National Review
	“progressive media”	Appealing to progressive partisan audience, e.g., Jacobin and Splinter News
	“national political journalists”	Covering US national politics
	“Middle East correspondents”	Covering or commenting on Middle East affairs

and 42.8% respectively of the conversations. In “the Trump train,” member accounts have on average 97.1 relevant tweets per month (TPM) and in “the resistance,” about 70.4 TPM. However, the passing of abortion laws was primarily a conservative issue: “the Trump train” alone accounted for 46.5% of total tweets (with 50.3 TPM), and the three conservative flocks combined tweeted 60.2% of relevant content. This pattern of activity contrasts with that surrounding the killing of Khashoggi, which caught mostly the attention of “Middle East correspondents” and the three liberal flocks.

Besides the level of expression intensity, the temporal pattern of expression diverges across events (Fig. 6.5c). For the Mueller investigation, all 10 flocks were relatively in sync in terms of tweets per day, suggesting that opinion expression about the Mueller investigation was driven by key moments. However, the passing of anti-abortion laws witnessed a completely different temporal pattern. The conservative flocks, spearheaded by “the Trump train,” had remained agitated on the abortion ban, as evidenced in their constant hyperactivity. However, the liberal flocks did not join the conversations en masse much later, when the Alabama governor signed the most extreme abortion ban. A different pattern can be observed in the killing of Khashoggi. His disappearance first and foremost concerned “Middle East correspondents,” spreading next to “national political journalists” and liberal flocks. Conservatives flocks, unlike their response in the other two events, reacted to this event later than other flocks.

The drastically different words used by flocks in their opinions toward each event (supporting Table E.4, Materials and Methods) demonstrate how opinion expression was tied to the flock context. For the Mueller report, conservative flocks saw it as a vindication of Trump (suggested by keywords like “#maga,” “trump2020”) and shifted the target to Democrats (“democrats” “witch,” “hunt,” “obama,” “hillary”). However, liberals saw it as evidence for “obstruction” of “justice” and reason for “impeachment” of Trump. They also called upon the public to “read” the “report” and the Department of Justice to release the full report. Responding to the anti-abortion laws, conservative flocks emphasized the sanctity of life and liberal flocks women’s rights. Conservatives and media invoked pro-life tropes, characterized by terms like “babies,” "heartbeat," "life," "murder" and "infanticide," whereas liberals and other media couched their language in legal and activism terms, like "access," "rights," "ban," "#stopthebans." For the Khashoggi event, while "Middle east correspondents" and "national political journalists" mainly focused on the event itself, the three liberal networks and the three conservative networks politicized this event. The liberal flocks tied it to Trump's and Kushner's relationships with the Saudis, while the conservatives focused on Khashoggi's alleged tie with Muslim brotherhood and on Obama's "mistreatment" of western journalists, and tried to channel the attention back to the Benghazi attack.

6.3 Methods

Targeted sampling from Twitter friendship network

In August 2018, we sampled elite Twitter accounts who actively expressed political opinions in the Twitter friendship network using personalized PageRank (PPR) sampling (Chen et al., 2020a). The PPR sampling evaluates nodes in the network with an approximate PPR vector and samples those nodes with the highest scores. The PPR vector is defined as the stationary probability distribution of which we call a personalized random walk (Page et al., 1998). At each step of the random walk, the walker returns to the seed node with probability α, and, with probability $1 - \alpha$, the random walker goes to an adjacent node chosen uniformly at random. The details of the algorithm and implementation are described in Appendix E.1. We chose 59 Twitter accounts as seed nodes (supporting Table E.1) and implemented the method (`https://github.com/RoheLab/aPPR`) to collect following network data. We obtained a total of 267,117 Twitter accounts, with a total of 10,174,291 friends that they followed. Given that an account who follows or is followed by few accounts is difficult to classify, we removed any accounts who follow fewer than 2 friends and those followed by fewer than 5 accounts. This resulted in the reported sample of the following network in the main text. In August 2019, a year after we first performed the targeted sampling, we updated the seed nodes by removing inactive seeds (such as @RealAlexJones and @RichardB-

Spencer), and added new seeds that emerged in the 2018 data. This resulted in a total of 75 seed nodes for the PPR sampling in 2019 (supporting Table E.1).

Flock identification

Through PPR sampling, we obtained a bipartite network consisting of followers and accounts followed by them, from the Twitter friendship network. To identify flocks, we employed an spectral method called *vintage sparse principal component analysis* (VSP) to detect community structure in the observed friendship network. VSP is a simple algorithm for sparse principal component analysis (SPCA), where the loadings of principal components (PCs) are sparse (Zou and Xue, 2018; Rohe and Zeng, 2020; Chen and Rohe, 2020). The coefficients of sparse PCs (also known as loadings) estimate the (mixed) community membership for each account (Rohe and Zeng, 2020). In particular, we applied the two-way version of VSP to detect 100 communities among the followers as well as the followed accounts. The estimated communities of follower and followed accounts are matched (Rohe et al., 2016), that is, the k-th follower community tends to follow members in the k-th community of followed accounts (supporting Figure E.2). Additional details about VSP and a schema of the algorithm are provided in Appendix E.1. For downstream analysis, we focused on communities of followed accounts, which we refer to as *flocks*. Such choice is based on the assumption that these communities are prominent Twitter opinion leaders as they are followed by the political-related accounts that we sampled. In our analysis of the 2018 Twitter sample, the size (number of member accounts) of 100 flocks is 13,101 on average, with only four flocks smaller than 1,000 and the largest being 56,943. We compared the 100 flocks identified in the 2018 sampling and 2019 sampling to assess the stability and fidelity of flocks. For this, we defined a matching between two sets of 100 flocks by maximizing the total number of shared accounts between pairs of matched flocks, whose solution was computed with the Hungarian algorithm (Kuhn, 1955). Such matching was then used to calculate the percentage of recovered accounts in each flock.

Computation-assisted interpretation of flocks

The flocks were interpreted based on the profile descriptions of flock members. Since each flock has 1000 members, this process was assisted by a computational approach that identifies the keywords of each flock using the best feature function (BFF), a feature selection method. A detailed description of BFF is provided in Appendix E.1. BFF takes tokenized unigrams present in the profile descriptions of all accounts and extracts unigrams that are most unique to one flock as compared to all other flocks. Based on the best unigrams associated with each flock, validated by the authors based on the actual profile descriptions, we interpreted and named each flock. The full list of 100 flocks and their inter-relationship are provided in supporting Table E.3. We selected 50 flocks of interest for the downstream analysis and demonstrate 24 on our website.

Event detection and tweet classification

Our extraction of public opinion is event-based. Basically, we collect tweets on a daily basis and perform a two-stage text analysis: (i) identify news events across the whole corpus of tweets and (ii) designate the relevant tweets to individual news events. This pipeline is a data-driven mechanism informed by human input. For example, while the labeling of each news event relies on the cluster of words and short phrases, it was validated with news reports from mainstream outlets. Additional details about text pre-processing, news event identification, and tweet classification are provided in Appendix E.1

Evaluation of flocks by shared followers

To estimate the pattern of shared followers among flocks, we utilized the friendship information of the accounts who followed accounts in at least one flock. Specifically, we counted the number of shared followers between any pair of accounts in the 50 selected flocks, that is, a total of 50×1000 accounts. We then aggregated these individual counts into flock's shared follower counts:

for each follower, if it follows n_i member accounts of flock i and n_j of flock j, then we add $n_i n_j \times 10^{-6}$ to the shared follower counter between flock i and j.

Sampling of tweets and three news event tweets

To validate flocks with retweeting relationships and tweet text, we constructed a random sample containing all tweets on Mondays from October 1, 2018 to October 1, 2019. Span the sampling across a whole year strategy serves to prevent the peculiarity of a certain time period from skewing general patterns. This yielded 30,028,074 tweets with 15,846,255 being retweets, which were used to analyze hashtag usage and perform topic modeling of text contents (Appendix E.2). Among our sample of retweets, 7,379,555 (46.6%) were originally posted by accounts in the 50 flocks, which is then used to examine the retweeting relationship among flocks. To obtain a low-noise set of tweets about each news event, we applied restrictive search strings to retrieve content. For the concluding phase of the Mueller investigation, any tweet containing "mueller" or"russia probe" (case insensitive) or any tweet quoting another tweet containing the same terms was included, resulting in a total of 1,160,120 tweets. For the passing of anti-abortion laws, we collected a total of 261,205 tweets using the search term "abortion." Lastly, for the killing of Khashoggi, "khashoggi" yielded 151,478 tweets.

Frequently used hashtags

For the analysis of hashtag usage, we included a total of 129 hashtags that appeared over 4000 times in our tweet sample. These hashtags were grouped into 6 categories. For example, hashtags presumably used by progressive accounts, like #voteblue and #bluewave, were labeled "liberal." Likewise, hashtags often used by conservative accounts, such as #tcot and #votered, were categorized as "conservative." The "Trump campaign" category included hashtags like #maga and #trumptrain presumably used by Trump supporters to rally around Trump. "QAnon" hashtags, like #qanon and #thegreatawakening, were presumably used by people holding conspiracy beliefs that "deepstate" traitors were scheming to thwart the Trump presidency. The "issue/topic" category included miscellaneous hashtags concerning political issues or topics. The remaining hashtags, mainly pop culture related, fell under the “other” category. The distribution of each hashtag use across all 50 flocks is presented in supporting Figure E.3. A subset of hashtags was selected to represent the range of patterns observed in all hashtags in Fig. E.3.

Activity, keywords, and sentiment of flocks in response to news events

We define a measure to assess the daily activity level of opinion expression of a Twitter account in response to a news event: the number of tweets per thousand event tweets (TPK). The measure of TPK normalizes for the total of event tweets thus is comparable across different news events. Given a set of Twitter accounts and their event tweets, TPK is computed in two steps. First, calculate the “per thousand event tweet scaling factor”, which is defined as the average number of daily event tweets divided by 1,000. Second, divide the event tweet counts of individual accounts by the event scaling factor. This quantity is averaged over all days across news event period in Fig. E.5b and is averaged over individual flocks in Fig. E.5c. We identified the keywords in each flock’s tweets using BFF, the same procedure used to find keywords in each flock’s profile descriptions. We conducted sentiment analysis using the AFINN lexicon (Nielsen, 2011). Specifically, given a set of tweets (e.g., tweets grouped by flock), the average of all words’ sentiment scores, weighted by the square root of their frequency, is treated as the overall sentiment. Here, the square root was taken for variance stabilization under the Poisson rate model (Bartlett, 1947).

6.4 Discussion

In this paper, we introduce “murmuration,” a framework for the large-scale measurement of opinions on social media. It treats flocks, which encode social network structure, as the unit of analysis of social media public opinion. Overall, our results speak to the effectiveness of the murmuration framework in capturing the temporal and social dynamics of public opinion on social media. In particular, we demonstrate that flocks are a homogeneous and stable structure that predicts opinion expression. We further show that the murmuration framework identifies public opinion with distinct patterns of opinion intensity, temporality, and contestation. The patterns from our case study suggest that networks on Twitter who talk politics exhibit shared attention, though at varying levels of intensity, to events that might not align with their views; and they attempt to frame those events from different angles in line with their own values and identities. Flocks might engage in such practice not so much to convince the outside world as to invoke their core beliefs or "ideological priors" to defend their egos against any ideologically disruptive evidence (Katz, 1960). Alternatively, they might seize the opportunities that those high-profile events afford them to jostle for power by advancing ideological definitions of issues and shaping the corresponding public response (Entman, 1993; Jungherr et al., 2019).

Methodically, this study offers one way to study public opinion that is different than survey-based public opinion polls and the text-centric approach to mining social media opinions. Our results suggest that to analyze social media opinions, researchers should combine the dominant text-as-data approach and the social network approach. This synthetic approach helps discover patterns of expression and interaction that can be traced back to social actors and the networks they are part of. As a result, we can better take advantage of social media data to understand public opinion as a form of social interaction and to reveal underlying social dynamics.

We must note that the opinions that we measure in this paper belong to the elite layer of public opinion, though the murmuration framework can be applied in various contexts and for different purposes. However, we see this more as a feature than a limitation. Given previous studies showing the two-step flow of opinions, understanding this stratum of opinion leaders is essential. Moreover, since these opinion leaders on social media might interact with mass media, this project in its next phase will examine how social media public opinion, in terms of both intensity and content, interacts with news media attention and coverage.

A.1 Technical proofs

Proof of Proposition 2.1

Proof. We apply Perron-Frobenius theorem for the first part (Perron, 1907; Frobenius et al., 1912), and complete the proof by construction.

(a) First, notice that Q is a Markov transition matrix by modifying $G = (V, E)$ a little. To this end, (i) shrink the weights of every existing edge by factor $1 - \alpha$, and (ii) add an edge weighted α between seed node v_0 and all nodes in the graph. Then Q represents the new graph $G'(V, E')$, which is strongly connected by construction. Hence Q is irreducible.

The PPR vector p is all-positive. To see this, note that the equation $p^T = p^T Q$ implies that p is a stationary distribution for the standard random walk on G'. Since G' is strongly connected, it follows that the stationary distribution must be all-positive.

From the Perron-Frobenius theorem, the only all-positive eigenvector of a non-negative irreducible matrix is associated with the leading eigenvalue, which is 1 in our case. Since the leading eigenvalue of non-negative irreducible matrix is simple, we conclude that p is unique.

(b) We finish the proof by constructing an explicit form of the PPR vector. Let $R_\alpha = \alpha \sum_{s=0}^{\infty} (1-\alpha)^s P^s$. The infinite sum converges for $\alpha \in (0, 1]$. Then, $p = R_\alpha^T \pi$ satisfies the definition of personalized PageRank vector,

$$\begin{aligned} \alpha\pi^T + (1-\alpha)\pi^T R_\alpha P &= \alpha\pi^T + (1-\alpha)\pi^T \left(\alpha \sum_{s=0}^{\infty} (1-\alpha)^s P^s \right) P \\ &= \alpha\pi^T + \alpha \sum_{s=1}^{\infty} (1-\alpha)^s \pi^T P^s \\ &= \pi^T R_\alpha. \end{aligned}$$

Since the solution is unique, we have $p = R_\alpha^T \pi$. □

Proof of Proposition 2.2

Proof. Algorithm 2.1 maintains two vectors, p^{ε} and r, by transporting probability mass from r to p^{ε} at each updating step. Note that the termination criterion implies that $r_u < \varepsilon d_u$ for any u sampled, thus it suffices to prove that

$$|p_u - p_u^{\varepsilon}| \leqslant r_u.$$

For a fixed α, let $p(x)$ be the PPR vector with preference vector $x \in \mathbb{R}^N$ satisfying $x_i \geqslant 0$ and $\|x\|_1 \leqslant 1$. Then $p(\pi)$ is the exact PPR vector as in Equation (2.2). Since $p(x)^T P = p(x^T P)$, we have (Jeh and Widom, 2003)

$$p(x) = \alpha x + (1-\alpha) p(P^T x). \tag{A.1}$$

We argue that $p^{\varepsilon} + p(r)$ is invariant in updating steps. To see this, suppose $(p^{\varepsilon})'$ and r' are the results of performing one update on p^{ε} and r after sampling node u. We have

$$\begin{aligned} (p^{\varepsilon})' &= p^{\varepsilon} + \alpha r_u e_u, \\ r' &= r - r_u e_u + (1-\alpha) r_u P^T e_u. \end{aligned}$$

where e_u is the unit vector on the direction of u. Then,

$$\begin{aligned} p(r) &= p(r - r_u e_u) + p(r_u e_u) \\ &\overset{\text{(i)}}{=} p(r - r_u e_u) + \alpha r_u e_u + (1-\alpha) p\left(r_u P^T e_u\right) \\ &\overset{\text{(ii)}}{=} p\left(r - r_u e_u + (1-\alpha) r_u P^T e_u\right) + \alpha r_u e_u \\ &= p(r') + (p^{\varepsilon})' - p^{\varepsilon}, \end{aligned}$$

where (i) is applying Equation (A.1) at $x = r_u e_u$ and (ii) comes from the linearity of PPR vector in the preference vector.

The desired result follows from recognizing that $p^{\varepsilon} + p(r)$ is initially $\vec{0} + p(\pi)$ and that when the algorithm terminates, $[p(r)]_u \leqslant r_u$ for any sampled u. □

Remark. If $\varepsilon d_1 > 1$, Algorithm 2.1 terminates after the first round and simply output $p = \vec{0}$. Under this circumstance, Proposition 2.2 still holds, because $|p_u - p_u^\varepsilon| \leqslant |p_u| + |p_u^\varepsilon| \leqslant 1$.

Lemmas for the DC-SBM

Lemma A.1 (Properties of the DC-SBM)**.** *Under the population directed DC-SBM with* K *blocks and parameters* $\{\mathbf{B}, Z, \Theta^{\text{in}}, \Theta^{\text{out}}\}$*,*

(a) $\mathbf{D}^{\text{in}} = Z^T \mathscr{D}^{\text{in}} Z$*, and* $\mathbf{D}^{\text{out}} = Z^T \mathscr{D}^{\text{out}} Z$ *, and*

(b) $\mathscr{d}_v^{\text{in}} = \theta_v^{\text{in}} \mathbf{d}_{z(v)}^{\text{in}}$*, and* $\mathscr{d}_v^{\text{out}} = \theta_v^{\text{out}} \mathbf{d}_{z(v)}^{\text{out}}$*.*

Proof. a is an alternative way of writing the definition. For b, we prove the first equation. Recall that for any i, $\sum_{u:z(u)=i} \theta_u^{\text{out}} = 1$, then by definition,

$$\mathscr{d}_v^{\text{in}} = \sum_u \theta_u^{\text{out}} \theta_v^{\text{in}} B_{z(u)z(v)} = \theta_v^{\text{in}} \sum_{j=1}^K \left(\mathbf{B}_{jz(v)} \sum_{u:z(u)=j} \theta_u^{\text{out}} \right) = \theta_v^{\text{in}} \mathbf{d}_{z(v)}^{\text{in}}.$$

□

Remark. Since $Z^T \Theta^{\text{in}} Z = I_K$, a implies $\left[\mathscr{D}^{\text{in}}\right]^{-1} \Theta^{\text{in}} Z = Z \left[\mathbf{D}^{\text{in}}\right]^{-1}$.

Lemma A.2 (Explicit form of $\mathscr{P}$ and its powers)**.** *Under the population directed DC-SBM with* K *blocks and parameters* $\{\mathbf{B}, Z, \Theta^{\text{in}}, \Theta^{\text{out}}\}$*, the population graph transition is the product*

$$\mathscr{P} = Z\mathbf{P}Z^T\Theta^{\text{in}}.$$

and its matrix powers are

$$\mathscr{P}^k = Z\mathbf{P}^k Z^T \Theta^{\text{in}}.$$

Proof. By definition and Lemma A.1b, for any $u, v \in V$,

$$\mathscr{P}_{uv} = \left(\theta_u^{\text{out}} \mathbf{d}_{z(u)}^{\text{out}}\right)^{-1} \theta_u^{\text{out}} \theta_v^{\text{in}} \mathbf{B}_{z(u)z(v)} = \theta_v^{\text{in}} \mathbf{B}_{z(u)z(v)} / \mathbf{d}_{z(u)}^{\text{out}} = \theta_v^{\text{in}} \mathbf{P}_{z(u)z(v)}.$$

For the powers of $\mathscr{P}$, noticing that $Z^T \Theta^{\text{in}} Z = I_K$,

$$\mathscr{P}^2 = Z\mathbf{P}Z^T\Theta^{\text{in}} Z\mathbf{P}Z^T\Theta^{\text{in}} = Z\mathbf{P}^2 Z^T \Theta^{\text{in}}.$$

The desired result follows from the principle of induction on k-th power. □

Proof of Theorem 2.4

Proof. By Proposition 2.1 and Lemma A.2, we have

$$\begin{aligned} \mathscr{p} &= \alpha \sum_{s=0}^{\infty} (1-\alpha)^s \left(\mathscr{P}^s\right)^{\mathsf{T}} \pi \\ &= \alpha \sum_{s=0}^{\infty} (1-\alpha)^s \Theta^{\text{in}} Z \left(\mathbf{P}^s\right)^{\mathsf{T}} Z^{\mathsf{T}} \pi \\ &= \Theta^{\text{in}} Z \left(\alpha \sum_{s=0}^{\infty} (1-\alpha)^s \left(\mathbf{P}^s\right)^{\mathsf{T}} \boldsymbol{\pi} \right) \\ &= \Theta^{\text{in}} Z \mathbf{p}. \end{aligned}$$

In addition, it follows from Lemma A.1a that

$$\mathscr{p}^* = \left[\mathscr{D}^{\text{in}}\right]^{-1} \mathscr{p} = \left[\mathscr{D}^{\text{in}}\right]^{-1} \Theta^{\text{in}} Z \mathbf{p} = Z \left[\mathbf{D}^{\text{in}}\right]^{-1} \mathbf{p} = Z \mathbf{p}^*.$$

This completes the proof. □

Proof of Lemma 2.5

Proof. For any $\alpha > 0$, the PPR vector with seed node $v_0 = 1$ is the solution to the equation $\mathscr{p}^{\mathsf{T}} = \mathscr{p}^{\mathsf{T}} \mathscr{Q}$, where $\mathscr{Q} = \alpha \Pi + (1-\alpha)\mathscr{P}$. Define a sequence of probability distribution $\mathscr{p}^s \in \mathbb{R}^N$ such that $\mathscr{p}^s = \left(\mathscr{Q}^s\right)^{\mathsf{T}} \mathscr{p}^0$, where $\mathscr{p}^0$ is an arbitrary initial probability distribution. Then, $\lim_{s \to \infty} \mathscr{p}^s = \mathscr{p}$. For simplicity, we assume $\mathscr{p}^0$ is close to $\mathscr{p}$, that is, for any $\varepsilon > 0$ and $s \geqslant 0$,

$$\|\mathscr{p}^s - \mathscr{p}\|_{\infty} < \varepsilon/2. \tag{A.2}$$

This can be achieved by finding an integer $S(\varepsilon)$ large enough and setting $\mathscr{p}^0 = \mathscr{p}^S$.

We first claim that

$$\max_{u \neq 1} \frac{p_u^{s+1}}{d_u} \leqslant (1-\alpha) \max_{u \in V} \frac{p_u^s}{d_u}. \tag{A.3}$$

In fact, for any $u \neq 1$,

$$\begin{aligned} p_u^{s+1} &= \alpha \mathbb{1}_{\{u=1\}} + (1-\alpha) \sum_{v \in V} \frac{\mathscr{A}_{vu}}{d_v} p_v^s \\ &\leqslant (1-\alpha) \left(\sum_{v \in V} \mathscr{A}_{vu} \right) \max_{v \in V} \frac{p_v^s}{d_v} \\ &= (1-\alpha) d_u \max_{v \in V} \frac{p_v^s}{d_v}. \end{aligned}$$

We then show $\frac{p_1^s}{d_1} > \frac{p_v^s}{d_v}$ for any $v \neq 1$ by contradiction. Suppose otherwise that $\frac{p_1^s}{d_1} \leqslant \max_{u \neq 1} \frac{p_u^s}{d_u}$, then Equation (A.2) implies for any s',

$$\frac{p_1^{s'}}{d_1} \leqslant \frac{p_1^s + \varepsilon}{d_1} \leqslant \max_{u \neq 1} \frac{p_u^s}{d_u} + \frac{\varepsilon}{d_1} \leqslant \max_{u \neq 1} \frac{p_u^{s'} + \varepsilon}{d_u} + \frac{\varepsilon}{d_1} \leqslant \max_{u \neq 1} \frac{p_u^{s'}}{d_u} + \frac{2\varepsilon}{d_{\min}},$$

where $d_{\min} = \min_{v \in V} d_v$. Hence, $\max_{u \in V} \frac{p_u^{s'}}{d_u} \leqslant \max_{u \neq 1} \frac{p_u^{s'}}{d_u} + \frac{2\varepsilon}{d_{\min}}$. In addition, applying Equation (A.3) recursively we have

$$\begin{aligned} \max_{u \in V} \frac{p_u^s}{d_u} &= \max_{u \neq 1} \frac{p_u^s}{d_u} \\ &\leqslant (1-\alpha) \max_{u \in V} \frac{p_u^{s-1}}{d_u} \\ &\leqslant (1-\alpha) \left(\max_{u \neq 1} \frac{p_u^{s-1}}{d_u} + \frac{2\varepsilon}{d_{\min}} \right) \\ &\leqslant (1-\alpha)^s \max_{u \in V} \frac{p_u^0}{d_u} + \frac{2\varepsilon}{d_{\min}} \sum_{t=1}^{s-1} (1-\alpha)^t. \end{aligned}$$

The inequality means that if $d_{\min} > 0$ is fixed, p_u^s can be arbitrarily small when $s \to \infty$, which contradicts the fact that p is a probability distribution. This completes the proof.

Remark. When the teleportation constant is zero, the PPR vector becomes the stationary probability distribution of a standard random walk,

$$\left(\frac{d_1}{\sum_i d_i}, \frac{d_2}{\sum_i d_i}, ..., \frac{d_N}{\sum_i d_i}\right).$$

After adjusting by node degrees, every entry becomes identical $(1/\sum_i d_i)$. The lemma is intuitive, recognizing that the teleportation introduces a particular favor of the seed node.

Remark. When the edges are weighted (non-negative), the stationary distribution of a random walk is still proportional to node degrees, if one defines the degree as sum of edge weights incident to the node (Lovász, 1993). Note also that the stationary distribution of a random walk in a directed graph is characterized by the in-degree of nodes (Ghoshal and Barabási, 2011; Lu et al., 2013). The conclusion and a modified proof apply to directed or weighted graphs. □

Proof of Theorem 2.6

We start with a few lemmas to prepare for the proof of Theorem 2.6. Specifically, we introduce a few notations used in Lemma A.5 and list a few properties of vector norm and matrix norm (Brémaud, 2013). For completeness, Section A.1 lists a few inequalities that are used throughout the proofs.

For any strictly positive probability distribution vector $p \in \mathbb{R}^N$, the inner product space indexed by p is a real vector space $\mathbb{R}^N$ endowed with the inner product

$$\langle x, y \rangle_p = \sum_{\nu=1}^{N} p_\nu x_\nu y_\nu.$$

The corresponding vector norm and the induced matrix norm are defined respectively as

$$\|x\|_p = \sqrt{\langle x, x \rangle_p} \text{ and } \|A\|_p = \sup_{\|x\|_p = 1} \|A^T x\|_p.$$

Lemma A.3. *If* $0 \leqslant p_{\min} \leqslant p_\nu \leqslant p_{\max}$ *for all* $\nu = 1, 2, ..., N$, *then the following* *inequalities hold*

$$\sqrt{p_{\min}}\|x\|_2 \leqslant \|x\|_p \leqslant \sqrt{p_{\max}}\|x\|_2 \ \textit{and} \ \sqrt{\frac{p_{\min}}{p_{\max}}}\|A\|_2 \leqslant \|A\|_p \leqslant \sqrt{\frac{p_{\max}}{p_{\min}}}\|A\|_2.$$

The following lemma provides concentration of the node degrees in a graph generated from the DC-SBM.

Lemma A.4 (Degree concentration)**.** *Let* $G = (V, E)$ *be a graph of* N *vertices generated from the DC-SBM with* K *blocks and parameters* $\{\mathbf{B}, Z, \Theta\}$*. Let* $d_{\min}$ *and* $d_{\max}$ *be the smallest and the largest node degree observed. Let* δ *be the average expected node degree, and define* $\rho = \mathscr{d}_{\max}/\mathscr{d}_{\min}$*. If* $\delta \geqslant c_0(1-\alpha)\log N$ *for some sufficiently large constant* $c_0 > 0$*, then with probability at least* $1 - \mathcal{O}(N^{-10})$*, it holds that*

$$\frac{\delta}{2\rho} \leqslant d_{\min} \leqslant d_{\max} \leqslant \frac{3\rho\delta}{2}. \tag{A.4}$$

Proof. Note that the definition of ρ immediately implies that

$$\frac{\delta}{\rho} \leqslant \mathscr{d}_{\min} \leqslant \mathscr{d}_{\max} \leqslant \delta\rho.$$

The lemma follows from the standard Chernoff's bound, hence is omitted. □

The following useful lemma concerns the eigenvector perturbation for probability transition matrices, promoted from the celebrated Davis-Kahan $\sin\Theta$ Theorem (Davis and Kahan, 1970).

Lemma A.5 (Eigenvector perturbation)**.** *Suppose that* Q*,* $\hat{Q}$*, and* $\mathcal{Q}$ *are probability transition matrices with stationary distributions* p*,* $\hat{p}$*, and* $\mathscr{p}$ *respectively. Assume that* $\mathcal{Q}$ *represents a reversible Markov chain. Then,*

$$\|p - \hat{p}\|_{\mathscr{p}} \leqslant \frac{\|(Q - \hat{Q})^T p\|_{\mathscr{p}}}{1 - \max\{\lambda_2(\mathcal{Q}), -\lambda_N(\mathcal{Q})\} - \|\hat{Q} - \mathcal{Q}\|_{\mathscr{p}}}.$$

The proof the Lemma A.5 can be found in Chen et al. (2019) Section 3, thus omitted.

Proof. The proof processes as follows. We first bound the entrywise error rate of p,

$$\frac{\|p - \mathscr{p}\|_\infty}{\|\mathscr{p}\|_\infty} \leqslant c_0 \sqrt{\frac{\log N}{\delta}},$$

by invoking the novel leave-one-out techniques (Chen et al., 2019), The entrywise error bounds of p^* follows immediately.

Recall that both p and $\mathscr{p}$ are stationary distribution, which means

$$p = Q^T p \text{ and } \mathscr{p} = \mathscr{Q}^T \mathscr{p}.$$

Due to this, for any $w = 1, 2, ..., N$, we can decompose

$$\begin{aligned} p_w - \mathscr{p}_w &= Q^T_{\cdot w} p - \mathscr{Q}^T_{\cdot w} \mathscr{p} \\ &= \underbrace{(Q_{\cdot w} - \mathscr{Q}_{\cdot w})^T \mathscr{p}}_{:= I_1^w} + \underbrace{Q^T_{\cdot w}(p - \mathscr{p})}_{:= I_2^w}, \end{aligned}$$

where $Q_{\cdot w}$ denotes the w-th column of Q.

(a) We start with the first term I_1^w. Note that

$$\begin{aligned} I_1^w &= (1-\alpha) \sum_{v=1}^N \left[\frac{A_{vw}}{d_v} - \frac{\mathscr{A}_{vw}}{\mathscr{d}_v} \right] \mathscr{p}_v \\ &= \underbrace{(1-\alpha) \sum_{v=1}^N \left[(A_{vw} - \mathscr{A}_{vw}) \frac{1}{\mathscr{d}_v} \right] \mathscr{p}_v}_{:= I_{11}^w} + \underbrace{(1-\alpha) \sum_{v=1}^N A_{vw} \left(\frac{1}{d_v} - \frac{1}{\mathscr{d}_v} \right) \mathscr{p}_v}_{:= I_{12}^w} . \end{aligned}$$

Recall that A_{vw}'s correspond to independent Bernoulli random variables, we can easily bound the first term using Bernstein's inequality (Lemma

A.8), with probability at least $1 - \mathscr{O}(N^{-8})$,

$$\begin{aligned} |I_{11}^{w}| &\leqslant (1-\alpha)\left|\sum_{v=1}^{N} (A_{vw} - \mathscr{A}_{vw})\right| \frac{\|p\|_\infty}{\delta} \\ &\leqslant (1-\alpha)\left(\sqrt{16\log N \sum_{v=1}^{N} \mathscr{A}_{vw}} + \frac{16\log N}{3}\right)\frac{\|p\|_\infty}{\delta} \\ &\overset{(i)}{\leqslant} (1-\alpha)\left(4\sqrt{\frac{\rho\log N}{\delta}} + \frac{16\log N}{3\delta}\right)\|p\|_\infty, \end{aligned}$$

where (i) follows from the fact that $\rho\delta \leqslant \mathscr{d}_{\max}$.

Note that the second term is

$$I_{12}^{w} = (1-\alpha)\sum_{v=1}^{N} \mathbb{1}_{(v,w)\in E}\left(\frac{1}{d_v} - \frac{1}{\mathscr{d}_v}\right)\mathscr{p}_v,$$

to which we can apply the Hoeffding's inequality (Lemma A.6) and obtain

$$\mathbb{P}\left(|I_{12}^{w}| \leqslant \rho(1-\alpha)\sqrt{\frac{\rho\log N}{\delta}}\|\mathscr{p}\|_\infty\right) \geqslant 1 - 2N^{-8}.$$

In sums, we have high probability event

$$|I_{1}^{w}| \leqslant (1-\alpha)\left((4+\rho)\sqrt{\rho} + 3\sqrt{\frac{\log N}{\delta}}\right)\sqrt{\frac{\log N}{\delta}}\|p\|_\infty. \tag{A.5}$$

(b) The statistical dependency between p and Q introduces difficulty in sharply bounding I_{2}^{w}. Nevertheless, we can invoke the leave-one-out techniques to decouple the dependency. To this end, we define, for each $w = 1, 2, ..., N$, a new transition matrix $Q^{(w)} = \alpha\Pi + (1-\alpha)P^{(w)}$ that bridges between Q and $\mathscr{Q}$. $P^{(w)}$ has almost the same entries as P except for replacing those in w-th row or column by their expectations; that is, for any $u \neq v$,

$$P^{(w)}_{uv} = \begin{cases} P_{uv}, & u \neq w \text{ and } v \neq w, \\ \mathscr{P}_{uv}, & u = w \text{ or } v = w, \end{cases}$$

and for any $u = 1, 2, ..., N$,

$$P^{(w)}_{uu} = 1 - \sum_{v : v \neq u} P^{(w)}_{uv},$$

in order to ensure that $P^{(w)}$ and $Q^{(w)}$ are transition matrices. In addition, define $p^{(w)}$ to be the stationary distribution corresponding to $Q^{(w)}$. As demonstrated in Chen et al. (2019), $p^{(w)}$ helps us well approximate p, yet it is statistically independent of $Q_{\cdot w}$.

Now we decompose I^w_2 as follows:

$$\begin{aligned} I^w_2 &= \sum_{v=1}^{N} Q_{vw}(p_v - \mathscr{p}_v) \\ &= \underbrace{\sum_{v=1}^{N} Q_{vw}\left(p_v - p^{(w)}_v\right)}_{:=I^w_{21}} + \underbrace{\sum_{v=1}^{N} Q_{vw}\left(p^{(w)}_v - \mathscr{p}_v\right)}_{:=I^w_{22}}. \end{aligned}$$

(c) In this part, we focus on the first term I^w_{21}, where we would need another intermediate quantity to facilitate our estimation. To be specific, consider the leave-one-out version of Q conditioning on the graph $G = (V, E)$, $Q^{(w,G)} = \alpha\Pi + (1 - \alpha)P^{(w,G)}$, which is almost the same as Q except for replacing the non-zero entries in w-th row or column by their expectations. Concretely, for and $u \neq v$,

$$P^{(w,G)}_{uv} = \begin{cases} P_{uv}, & u \neq w \text{ and } v \neq w, \\ \mathbb{1}_{(u,v) \in E}\mathscr{P}_{uv}, & u = w \text{ or } v = w, \end{cases}$$

and for any $u = 1, 2, ..., N$, define

$$P_{uu}^{(w,G)} = 1 - \sum_{v:v\neq u} P_{uv}^{(w,G)},$$

so that $P^{(w,G)}$ is a probability transition matrix.

With $Q^{(w,G)}$ in mind, we now apply Cauchy-Schwarz inequality on I_{21}^{w} to reach

$$\begin{aligned}
|I_{21}^{w}| &= \left|\sum_{v=1}^{N} Q_{vw}\left(p_v - p_v^{(w)}\right)\right| \\
&\leqslant \left(\sum_{v=1}^{N} Q_{vw}^2\right)^{\frac{1}{2}} \left\|p - p^{(w)}\right\|_2 \\
&\overset{(i)}{\leqslant} \sqrt{\alpha + \frac{1}{d_{\min}}}\sqrt{\frac{\mathscr{p}_{\max}}{\mathscr{p}_{\min}}} \left\|p - p^{(w)}\right\|_{\mathscr{p}} \\
&\overset{(ii)}{\leqslant} \sqrt{\alpha + \frac{1}{d_{\min}}}\sqrt{\frac{\mathscr{p}_{\max}}{\mathscr{p}_{\min}}}\frac{1}{\gamma} \left\|\left(Q - Q^{(w)}\right)^{T} p^{(w)}\right\|_2 \\
&\overset{\substack{(iii)\\ \text{w.h.p.}}}{\leqslant} \sqrt{\alpha + \frac{2\rho}{\delta}}\frac{\sqrt{\kappa}}{\gamma}\left(\underbrace{\left\|(Q - Q^{(w,G)})^{T} p^{(w)}\right\|_2}_{:=I_{211}^{w}} + \underbrace{\left\|(Q^{(w,G)} - Q^{(w)})^{T} p^{(w)}\right\|_2}_{:=I_{212}^{w}}\right).
\end{aligned}$$

where (i) follows from Lemma A.3 and the fact that $P_{vw} \leqslant \frac{1}{d_{\min}}$, (ii) comes from Lemma A.5, and (iii) results from Lemma A.4 and the triangle inequality, and recognizing $\kappa = \mathscr{p}_{\max}/\mathscr{p}_{\min}$ (from the proof of Proposition 2.1, it is bounded), and "w.h.p." is short for "with high probability". Note that Π adds at most 1 to the rank of $\mathcal{Q}$, and because we presume B is positive definite $\mathscr{P}$ has exactly K positive eigenvalues among other zeros (Section A.2). Here, $\gamma = 1 - \max\{\lambda_2(\mathcal{Q}), -\lambda_N(\mathcal{Q})\} - \|Q^{(w,G)} - \mathcal{Q}\|_{\mathscr{p}}$ is the spectral gap and is lower bounded by some positive constant (due to Khanna et al. (2017)). Then, it boils down to controlling I_{211}^{w} and I_{212}^{w}.

For I^{w}_{211}, the w-th entry inside the vector norm is

$$\begin{aligned}\left[\left(Q - Q^{(w,G)}\right)^{T} p^{(w)}\right]_{w} &= \left[(Q - \mathscr{Q})^{T} p^{(w)}\right]_{w} \\ &= (1-\alpha)\sum_{v=1}^{N} (P_{vw} - \mathscr{P}_{vw}) p^{(w)}_{v}.\end{aligned}$$

Note that $p^{(w)}_{v}$ is statistically independent of $P_{\cdot w}$. Then, by Hoeffding's inequality (Lemma A.6) and Lemma A.4, we have with probability at least $1 - 2N^{-8}$,

$$\left[\left(Q - Q^{(w,G)}\right)^{T} p^{(w)}\right]_{w} \leqslant 4\rho(1-\alpha)\sqrt{\frac{\rho \log N}{\delta}} \left\|p^{(w)}\right\|_{\infty}. \tag{A.6}$$

As for any $u \neq w$, applying Hoeffding's inequality again yields

$$\begin{aligned}\left[\left(Q - Q^{(w,G)}\right) p^{(w)}\right]_{u} &= (1-\alpha)\sum_{v=1}^{N} (P_{vu} - \mathscr{P}_{vu})\, p^{(w)}_{v} \\ &= (1-\alpha)\left(P_{uu} - P^{(w,G)}_{uu}\right) p^{(w)}_{u} \\ &\quad + (1-\alpha)\left(P_{uw} - P^{(w,G)}_{uw}\right) p^{(w)}_{w} \\ &= -(1-\alpha)\left(P_{uw} - P^{(w,G)}_{uw}\right) p^{(w)}_{u} \\ &\quad + (1-\alpha)\left(P_{uw} - P^{(w,G)}_{uw}\right) p^{(w)}_{w}.\end{aligned}$$

Recognizing that

$$\left|P_{uw} - P^{(w,G)}_{uw}\right| = \begin{cases} A_{uw} d^{-1}_{uu} - \mathscr{A}_{uw} \mathscr{d}^{-1}_{uu}, & (u,w) \in E, \\ 0, & (u,w) \notin E, \end{cases}$$

we apply again the Hoeffding's inequality (Lemma A.6) together with

(A.4), and obtain with probability at least $1 - \mathcal{O}\left(N^{-8}\right)$,

$$\left|\left[\left(Q - Q^{(w,G)}\right)^{T} p^{(w)}\right]_{u}\right| \leqslant \begin{cases} 4\rho(1-\alpha)\frac{\sqrt{\log N}}{\delta}\left\|p^{(w)}\right\|_{\infty}, & (u,w) \in E, \\ 0, & (u,w) \notin E. \end{cases} \tag{A.7}$$

Combining (A.6) and (A.7) yields

$$\begin{aligned} I_{211}^{w} &\leqslant 4\rho(1-\alpha)\left(1 + \sqrt{\sum_{u:u\neq w} \mathbb{1}_{(u,w)\in E}}\right)\sqrt{\frac{\rho \log N}{\delta}}\left\|p^{(w)}\right\|_{\infty} \\ &\overset{\substack{(i) \\ \text{w.h.p.}}}{\leqslant} 8\rho^{2}\sqrt{\rho}(1-\alpha)\sqrt{\frac{\log N}{\delta}}\left\|p^{(w)}\right\|_{\infty}, \end{aligned}$$

where (i) follows from the high probability event that $d_{\max} \leqslant 3\rho\delta/2$.

Regarding I_{212}^{w}, since $(Q^{(w,G)} - Q^{(w)})\mathscr{p} = \vec{0}$, we can rewrite this as

$$I_{212}^{w} = \left\|\left(Q^{(w,G)} - Q^{(w)}\right)^{T}\left(p^{(w)} - \mathscr{p}\right)\right\|_{2}.$$

Similarly, note that $P_{vw}^{(w)} - P_{vw}^{(w,G)} = \frac{\mathscr{A}_{vw}}{\mathscr{d}_{v}}\mathbb{1}_{(w,v)\notin E}$, we apply Bernstein's inequality on w-th term inside the vector norm to obtain that with probability at least $1 - 2N^{-8}$,

$$\begin{aligned} &\left[\left(Q^{(w,G)} - Q^{(w)}\right)\left(p^{(w)} - \mathscr{p}\right)\right]_{w} \\ = \quad &(1-\alpha)\sum_{v=1}^{N}\left(P_{vw}^{(w,G)} - P_{vw}^{(w)}\right)\left(p_{v}^{(w)} - \mathscr{p}_{v}\right) \\ = \quad &(1-\alpha)\sum_{v=1}^{N}\frac{1}{\mathscr{d}_{v}}\left(p_{v}^{(w)} - \mathscr{p}_{v}\right)\mathbb{1}_{(w,v)\notin E} \\ \overset{\text{w.h.p.}}{\leqslant} \quad &(1-\alpha)\left(4\rho\sqrt{\frac{\rho \log N}{\delta}} + \frac{16}{3}\frac{\log N}{\delta}\right)\left\|p^{(w)} - \mathscr{p}\right\|_{\infty}. \end{aligned}$$

For any $u \neq w$, the u-th term inside vector norm is

$$\begin{aligned}
& \left[\left(Q^{(w,G)} - Q^{(w)}\right)\left(p^{(w)} - \mathscr{p}\right)\right]_u \\
= \; & (1-\alpha)\sum_{v=1}^{N}\left(P_{vu}^{(w,G)} - P_{vu}^{(w)}\right)\left(p_v^{(w)} - \mathscr{p}_v\right) \\
= \; & (1-\alpha)\left(P_{vv}^{(w,G)} - P_{uu}^{(w)}\right)\left(p_u^{(w)} - \mathscr{p}_u\right) \\
& +(1-\alpha)\left(P_{vw}^{(w,G)} - P_{uw}^{(w)}\right)\left(p_w^{(w)} - \mathscr{p}_w\right) \\
= \; & -(1-\alpha)\left(P_{vw}^{(w,G)} - P_{uw}^{(w)}\right)\left(p_u^{(w)} - \mathscr{p}_u\right) \\
& +(1-\alpha)\left(P_{vw}^{(w,G)} - P_{uw}^{(w)}\right)\left(p_w^{(w)} - \mathscr{p}_w\right).
\end{aligned}$$

Recognizing that

$$P_{uw}^{(w)} - P_{uw}^{(w,G)} = \mathscr{A}_{uw}\mathscr{d}_u^{-1}\mathbb{1}_{(u,w)\notin E},$$

we have from (A.4) that

$$\left|\left[\left(Q^{(w,G)} - Q^{(w)}\right)^{T}\left(p^{(w)} - \mathscr{p}\right)\right]_u\right| \leqslant 2\mathscr{A}_{uw}\mathscr{d}_u^{-1}\mathbb{1}_{(u,w)\notin E}(1-\alpha)\left\|p^{(w)} - \mathscr{p}\right\|_\infty.$$

Hence, we have with probability at least $1 - \mathscr{O}\left(N^{-8}\right)$,

$$\begin{aligned}
I_{212}^{w} \;&\leqslant\; (1-\alpha)\left(4\rho\sqrt{\rho}\sqrt{\frac{\log N}{\delta}} + \frac{16}{3}\frac{\log N}{\delta} + 2\sqrt{\sum_{u:u\neq w}\frac{\mathbb{1}_{(u,w)\notin E}}{\mathscr{D}_{uu}^2}}\right)\left\|p^{(w)} - \mathscr{p}\right\|_\infty \\
&\overset{(i)}{\leqslant}\; (1-\alpha)\left(4\rho\sqrt{\rho}\sqrt{\frac{\log N}{\delta}} + \frac{16}{3}\frac{\log N}{\delta} + 2\rho\sqrt{\frac{\rho}{\delta}}\right)\left\|p^{(w)} - \mathscr{p}\right\|_\infty,
\end{aligned}$$

where (i) follows from the high probability event that $d_{\max} \leqslant 3\rho\delta/2$. Combining the above two bounds, we have with probability at least $1 -$

$\mathcal{O}\left(N^{-8}\right)$ that

$$\begin{aligned} I_{21}^{w} &\leqslant \sqrt{\alpha + \frac{2\rho}{\delta}\frac{\sqrt{\kappa}}{\gamma}}\,(I_{211}^{w} + I_{212}^{w}) \\ &\overset{(i)}{\leqslant} 8c\rho^2(1-\alpha)\sqrt{\frac{\rho \log N}{\delta}}\|\mathscr{p}\|_\infty \\ &\quad + c(1-\alpha)\left(8\rho^2\sqrt{\frac{\rho\log N}{\delta}} + 2\rho\sqrt{\frac{\rho}{\delta}} + 4\sqrt{\frac{\rho\log N}{\delta}} + \frac{16}{3}\frac{\log N}{\delta}\right)\|p^{(w)} - \mathscr{p}\|_\infty \\ &\overset{(ii)}{\leqslant} 8c\rho^2(1-\alpha)\sqrt{\frac{\rho\log N}{\delta}}\|\mathscr{p}\|_\infty + \frac{c}{2}\|p^{(w)} - \mathscr{p}\|_\infty \\ &\overset{(iii)}{\leqslant} 16c\rho^2(1-\alpha)\sqrt{\frac{\rho\log N}{\delta}}\|\mathscr{p}\|_\infty + c\|p - \mathscr{p}\|_\infty. \end{aligned}$$

where $c = \sqrt{\alpha + \frac{2\rho}{\delta}\frac{\sqrt{\kappa}}{\gamma}}$, and (i) follows from the triangle inequality $\left\|p^{(w)}\right\|_\infty \leqslant \left\|p^{(w)} - \mathscr{p}\right\|_\infty + \|\mathscr{p}\|_\infty$, and (ii) holds as long as $\delta > c_0(1-\alpha)^2 \log N$ for some $c_0 > 0$ sufficiently large, and (iii) comes from the triangle inequality $\left\|p^{(w)} - \mathscr{p}\right\|_\infty \leqslant \left\|p^{(w)} - p\right\|_2 + \|p - \mathscr{p}\|_\infty$.

(d) Now it is left to estimate the last item I_{22}^{w}. Note that

$$\begin{aligned} I_{22}^{w} &= \sum_{v=1}^{N} \mathbb{1}_{(v,w)\in E} Q_{vw}\left(p_v^{(w)} - \mathscr{p}_v\right) \\ &= \sum_{v=1}^{N}\left[\alpha \mathbb{1}_{\{w=1\}} + (1-\alpha)\frac{1}{d_v}\mathbb{1}_{(v,w)\in E}\right]\left(p_v^{(w)} - \mathscr{p}_v\right) \\ &= \underbrace{\alpha\sum_{v=1}^{N}\mathbb{1}_{\{w=1\}}\left(p_v^{(w)} - \mathscr{p}_v\right)}_{:=I_{221}^{w}} + \underbrace{(1-\alpha)\sum_{v=1}^{N}\frac{\mathbb{1}_{(v,w)\in E}}{\mathscr{d}_v}\left(p_v^{(w)} - \mathscr{p}_v\right)}_{:=I_{222}^{w}} \\ &\quad \underbrace{+ (1-\alpha)\sum_{v=1}^{N}\left(\frac{1}{d_v} - \frac{1}{\mathscr{d}_v}\right)\mathbb{1}_{(v,w)\in E}\left(p_v^{(w)} - \mathscr{p}_v\right)}_{:=I_{223}^{w}}. \end{aligned}$$

Since both $p^{(w)}$ and $\mathscr{p}$ are distribution vector, $I_{221}^{w} = 0$. Then, due to

Hoeffding's inequality (Lemma A.6),

$$\begin{aligned}
|I_{222}^{w}| &\leqslant 4\rho(1-\alpha)\sqrt{\frac{\rho\log N}{\delta}}\left\|p^{(w)}-\mathscr{p}\right\|_{\infty},\\
|I_{223}^{w}| &\leqslant 2\rho(1-\alpha)\sqrt{\frac{\rho\log N}{\delta}}\left\|p^{(w)}-\mathscr{p}\right\|_{\infty},
\end{aligned}$$

with probability at least $1-\mathcal{O}\left(N^{-8}\right)$. Thus, we reach the high probability event

$$|I_{22}^{w}| \leqslant 6\rho(1-\alpha)\sqrt{\frac{\rho\log N}{\delta}}\left\|p^{(w)}-\mathscr{p}\right\|_{\infty}.$$

In sums, we reach with probability at least $1-\mathcal{O}\left(N^{-8}\right)$,

$$\begin{aligned}
|I_{2}^{w}| \leqslant\ & 16\rho^2(1-\alpha)\frac{\sqrt{\kappa\rho}}{\gamma}\sqrt{\alpha+\frac{2\rho}{\delta}}\sqrt{\frac{\log N}{\delta}}\|\mathscr{p}\|_{\infty}\\
&+\left(\frac{\sqrt{\kappa}}{\gamma}\sqrt{\alpha+\frac{2\rho}{\delta}}+6\rho(1-\alpha)\sqrt{\frac{\rho\log N}{\delta}}\right)\|p-\mathscr{p}\|_{\infty}. \quad \text{(A.8)}
\end{aligned}$$

(e) Collecting the preceding bounds (A.5) and (A.8) together, we conclude that with high probability

$$\begin{aligned}
\|p-\mathscr{p}\|_{\infty} &= \max_{w}|p_w-\mathscr{p}_w|\\
&\leqslant c_2(1-\alpha)\sqrt{\frac{\log N}{\delta}}\|\mathscr{p}\|_{\infty}+c_3\|p-\mathscr{p}\|_{\infty},
\end{aligned}$$

as long as $\delta/[(1-\alpha)\log N]$ is sufficiently large, which controls the entry-wise error of p,

$$\frac{\|p-\mathscr{p}\|_{\infty}}{\|\mathscr{p}\|_{\infty}} \leqslant c_1(1-\alpha)\sqrt{\frac{\log N}{\delta}}, \tag{A.9}$$

for some sufficiently large constant $c_1, c_2, c_3 > 0$.

Remark. c_2 and c_3 are controlled by constants ρ, κ, γ, which are thereby driven from the model parameters $\mathbf{B}$, Θ, K, and Z.

(f) Finally, we accomplish the proof by observing that

$$\begin{aligned}\frac{\|p^* - \mathscr{p}^*\|_\infty}{\|\mathscr{p}^*\|_\infty} &\leqslant \frac{2\max\left(d_{\min}^{-1}, \mathscr{d}_{\min}^{-1}\right)}{\mathscr{d}_{\min}^{-1}} \frac{\|p - \mathscr{p}\|_\infty}{\|\mathscr{p}\|_\infty} \\ &\leqslant \frac{4\|p - \mathscr{p}\|_\infty}{\|\mathscr{p}\|_\infty}.\end{aligned}$$

Above observation together with the inequality (A.9) allow us to control the entrywise error of p^* as claimed, with probability at least $1 - \mathscr{O}\left(N^{-5}\right)$,

$$\frac{\|p^* - \mathscr{p}^*\|_\infty}{\|\mathscr{p}^*\|_\infty} \leqslant c_2(1-\alpha)\sqrt{\frac{\log N}{\delta}},$$

for some sufficiently large constant $c_2 > 0$. □

Proof of Corollary 2.7

Proof. The algorithm ranks all vertices according to $p^{\varepsilon *}$, and the population local cluster can be explicitly written as

$$\mathscr{C} = \{v \in V : \mathscr{p}_v^* = \mathbf{p}_1^*\}.$$

It suffices to show that

$$p_v^{\varepsilon *} > p_u^{\varepsilon *}, \text{ for } \forall v \in \mathscr{C}, u \in V \backslash \mathscr{C},$$

where $p^{\varepsilon *}_v = p_v^\varepsilon / d_v$. To this end, we apply triangle inequality and get

$$\begin{aligned}\frac{p_v^{\varepsilon *} - p_u^{\varepsilon *}}{\|\mathscr{p}^*\|_\infty} &\geqslant \frac{\mathscr{p}_v^* - \mathscr{p}_u^*}{\|\mathscr{p}^*\|_\infty} - \frac{|p_v^* - \mathscr{p}_v^*|}{\|\mathscr{p}^*\|_\infty} - \frac{|p_u^* - \mathscr{p}_u^*|}{\|\mathscr{p}^*\|_\infty} - \frac{|p_u^{\varepsilon *} - p_u^*|}{\|\mathscr{p}^*\|_\infty} - \frac{|p_v^{\varepsilon *} - p_v^*|}{\|\mathscr{p}^*\|_\infty} \\ &\geqslant \Delta - \frac{2\|p^* - \mathscr{p}^*\|_\infty}{\|\mathscr{p}^*\|_\infty} - \frac{2\|p^{\varepsilon *} - p^*\|_\infty}{\|\mathscr{p}^*\|_\infty}.\end{aligned}$$

Since $\Delta_\alpha \leqslant 1$, assumption (2.8) contains condition (2.7) in Theorem 2.6, which together with Proposition 2.2 implies that

$$\frac{\|p^* - \mathscr{p}^*\|_\infty}{\|\mathscr{p}^*\|_\infty} < \frac{1}{4}\Delta, \qquad \frac{\|p^{\varepsilon *} - p^*\|_\infty}{\|\mathscr{p}^*\|_\infty} < \frac{1}{4}\Delta,$$

if $\Delta^2\delta/\log N$ is large enough. These collectively imply $p_v^* > p_u^*$ as desired. □

Concentration inequalities

The following is a standard concentration inequality used throughout the paper.

Lemma A.6 (Hoeffding's inequality)**.** *Let* $\{X_i\}_{1\leqslant i\leqslant n}$ *be a sequence of independent random variables where* $X_i \in [a_i, b_i]$ *for each* $1 \leqslant i \leqslant n$*, and* $S_n = \sum_{i=1}^n X_i$*. Then,*

$$\mathbb{P}(|S_n - \mathbb{E}\, S_n| \geqslant t) \leqslant 2\exp\left\{-\frac{t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}.$$

The next lemma is a special case of Chernoff's bound.

Lemma A.7 (Chernoff's bounds)**.** *Let* $\{X_i\}_{1\leqslant i\leqslant n}$ *be a sequence of independent random variables, whose sum is* S_n*, each having probability* p_i *of being equal to* a_i*, otherwise 0. Define* $\mu = \sum_i p_i a_i$*. Then, for any* $\varepsilon > 0$*,*

$$\mathbb{P}\left(X_i \geqslant (1+\varepsilon)\mu\right) \leqslant (1+\varepsilon)^{-\varepsilon\mu},$$

$$\mathbb{P}\left(X_i \leqslant (1-\varepsilon)\mu\right) \leqslant (1-\varepsilon)^{\varepsilon\mu}.$$

For the use of this paper, we only invoke a simpler version of Bernstein inequality.

Lemma A.8 (Bernstein's inequality)**.** *Let* $\{X_i\}_{1\leqslant i\leqslant n}$ *be a sequence of independent random variables with* $|X_i| \leqslant B$ *for each* $1 \leqslant i \leqslant n$*, and* $S_n = \sum_{i=1}^n X_i$ *and* $T_n = \sum_{i=1}^n X_i^2$*. Then, with probability at least* $1 - 2n^{-a}$*,*

$$|S_n - \mathbb{E}[S_n]| \leqslant \sqrt{2a\log n\,\mathbb{E}[T_n]} + \frac{2a}{3}B\log n$$

for any $a \geqslant 2$*.*

The proofs of Lemma A.6, A.7, and A.8 can be found in Boucheron et al. (2013b), hence are omitted.

A.2 Additional information on model parameters

Block connectivity matrix B

In the paper, we assume that the block connectivity matrix **B** corresponds to a strongly connected graph at block level and is positive definite. These assumptions asserts the efficacy of PPR clustering and is primarily a technical assumption sufficient for our theoretical results. In fact, we require **B** to represent to a strongly connected graph because this enables the block-wise PPR vector to have the largest value corresponding to the block of seed(s) (Lemma 2.5 in the paper). On the other hand, we impose the positive definiteness on **B** because this allows us to intuitively define the notion of local cluster, yet our statistical theory (i.e., the entrywise control of sample PPR vector) does not explicitly rely on such positive-definiteness per se. It is not clear yet whether these constraints are *necessary* in order for PPR clustering to function; possible generalizations of them are of research interest.

We list a few concrete examples showing that (i) if we break the strongly-connectivity assumption, the PPR clustering can fail, despite a reasonable teleportation constant, $\alpha = 0.15$, but (ii) PPR clustering often works as hoped even when **B** is not positive-definite. Throughout, we assume that the first block is targeted and consider directed graphs with three underlying blocks $(K = 3)$. The first two instances of **B** demonstrates the necessity of the strongly-connectivity constraint, which ensures the block-wise aPPR vector to possess the largest first element. The third and forth instances, on the other hand, indicate that **B** need not to be positive definite.

Violating the strongly-connective assumption

Hierarchy case. Let the block connectivity matrix

$$\mathbf{B} = \begin{bmatrix} p & p & p \\ 0 & p & p \\ 0 & 0 & p \end{bmatrix}$$

for some constant $p > 0$. $\mathbf{B}_{ij}$ is the number (or the probability) of edges from the i-th block to the j-th block in population. Then, the directed graph represented by $\mathbf{B}$ is not strongly connected, as block 3 has no path to the first block. In fact, this graph (specified by upper triangular $\mathbf{B}$) has a hierarchical structure, where the third block is in the center (or the highest hierarchy) of the graph, and the member of first block are essentially satellite from outside. Particularly, edges only come from outsiders to insiders.

We now perform the PPR clustering on the first cluster. The block-wise transition matrix is

$$\mathbf{P} = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \end{bmatrix}.$$

Then, both $\mathbf{B}$ and $\mathbf{P}$ are positive definite, with eigenvalues of (p, p, p) and $(1, 1/2, 1/3)$ respectively. To ease the calculation, we set $p = 3$. Then the block-wise PPR vector is approximately

$$\mathbf{p} = (0.209, 0.103, 0.688),$$

and the block-wise aPPR vector is approximately (after adjusting by column sums of $\mathbf{B}$)

$$\mathbf{p}^* = (0.0698, 0.0172, 0.0764).$$

As shown, neither block-wise PPR vector nor aPPR vector properly recognize the local cluster 1.

Adding a small amount of circulation. If we add a small quantity to the left bottom element of above $\mathbf{B}$ matrix, then the block connectivity matrix corresponds to a connected graph. To illustrate, we assign a small value to it, $\mathbf{B}_{31} = 0.1$, then the new block connectivity matrix becomes

$$\mathbf{B}' = \begin{bmatrix} p & p & p \\ 0 & p & p \\ 0.1 & 0 & p \end{bmatrix}.$$

To explore the PPR vector, we set $p = 3$ once again. In this case, $\mathbf{B}'$ has one real eigenvalue (≈ 4.069) and two imaginary eigenvalues. The block-wise PPR vector is approximately

$$\mathbf{p} = (0.235, 0.115, 0.650),$$

and the block-wise aPPR vector is approximately (after adjusting by column sums of $\mathbf{B}'$)

$$\mathbf{p}^* = (0.0755, 0.0192, 0.0723).$$

In this case, the PPR clustering works like a charm.

Violating the positive-definite assumption

Consider again the $K = 3$ design with equally distributed block size. We present two examples breaking the positive-definite assumption on $\mathbf{B}$, where the PPR cluster still operates properly.

Indefinite case. Given some constants $r > p > 0$, define

$$\mathbf{B} = \begin{bmatrix} p & r & r \\ r & p & r \\ r & r & p \end{bmatrix}.$$

In this case, the random graphs generated from such configuration of $\mathbf{B}$ have a unique characteristic: two vertices with different block memberships are more likely to connect than those pairs belonging to the same block. Note that the three eigenvalues of $\mathbf{B}$ are $p + 2r$, $p - r$, and $p - r$. Hence, B is an indefinite matrix (so does the block-wise transition matrix $\mathbf{P}$).

Interestingly, the PPR clustering continues working under this circumstance. For simplicity, setting $p = 3$ and $r = 9$, and we articulate the block-wise PPR vector and aPPR vector. In fact, the block-wise PPR vector is approximately

$$\mathbf{p} = (0.386, 0.306, 0.306).$$

Since $\mathbf{B}$ has homogeneous column sums, it follows that the first element in the block-wise aPPR vector is also the largest, suggesting the effectiveness of PPR clustering. The same conclusion hold when we set $p = 3$ and $r = 99$ (or 999).

Singular case. Suppose $p > 0$ and let

$$\mathbf{B} = \begin{bmatrix} 0 & p & 0 \\ p & 0 & p \\ 0 & p & 0 \end{bmatrix}.$$

In this case, nodes in block 1 only connect with those nodes in block 2, and the nodes in block 3 only have edges with block 2's members. Note that $\mathbf{B}$ is singular because three of its eigenvalues are 1, -1, and 0. So does the block-wise transition matrix. However, the PPR clustering remain effective. In fact, the block-wise PPR vector and aPPR vector are

$$\mathbf{p} = (0.345, 0.459, 0.195) \quad \text{and} \quad \mathbf{p}^* = \frac{1}{p}(0.345, 0.230, 0.195).$$

In both cases (when $\mathbf{B}$ is not positive-definite), the block-wise aPPR vector correctly assigns the largest value to the first element and thus is still effective for targeted sampling. These examples suggest a potentially greater applicability of the PPR clustering under the block model graph.

Comments

Putting together above demonstrations, we briefly comment on $\mathbf{B}$ and the PPR clustering. (i) The strongly-connectivity assumption is essential for the PPR clustering to be consistent. (ii) The efficacy of PPR clustering is conditioning on the fact that teleportation constant is sufficiently large. If we assign an extremely small to it, e.g. $\alpha = 0.001$, the PPR clustering collapses. (iii) Beyond community-like graphs (where $\mathbf{B}$ is positive-definite), the PPR clustering has potential for working on a more general block model graphs.

Spectral analysis on graph transition $\mathscr{P}$

In this section, we present a spectral analysis of graph transition matrix, which demonstrates that (1) under the population DC-SBM, a graph transition matrix $\mathscr{P}$ has exactly K positive eigenvalues, and $N - K$ zero eigenvalues, and (2) in a random graph generated from the DC-SBM, the graph transition matrix P is close to its population, with respect to spectral norm.

Lemma A.9 (Eigen-decomposition for $\mathscr{P}$ and $\mathbf{P}$)**.** *Under the population DC-SBM with K blocks and parameters $\{\mathbf{B}, Z, \Theta\}$, let $\mathscr{P} \in \mathbb{R}^{N\times N}$ be the population graph transition matrix and $\mathbf{P} \in \mathbb{R}^{K\times K}$ be the block-wise transition matrix. Then, $\mathscr{P}$ and $\mathbf{P}$ have the same K positive eigenvalues. The remaining $N - K$ eigenvalues of $\mathscr{P}$ are all zeros. Denote the K positive eigenvalues of both matrices as $\lambda_1 \geqslant \lambda_2 \geqslant \cdots ...\lambda_K \geqslant 0$, and let $\mathscr{X} \in \mathbb{R}^{N\times K}$ and $\mathscr{Y} \in \mathbb{R}^{K\times K}$ contain the left eigenvector of $\mathscr{P}$ and $\mathbf{P}$ respectively, corresponding to λ_i in their i-th column. Then, there exists a orthogonal matrix $U \in \mathbb{R}^{K\times K}$, such that*

(a) $\mathscr{X}^T = \mathscr{D}^{-1/2}\Theta^{1/2}ZU$; *and*

(b) $\mathscr{Y}^T = \mathbf{D}^{-1/2}U$.

Proof. We follow the proof of Lemma 3.3 in Qin and Rohe (2013). Define $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{B}\mathbf{D}^{-1/2}$, then $\mathbf{P} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{1/2}$. By model assumption, $\mathbf{P} \succ 0$.

Define the graph Laplacian $\mathscr{L} = \mathscr{D}^{-1/2}\mathscr{A}\mathscr{D}^{-1/2}$, then by Lemma A.1b,

$$\mathscr{L}_{uv} = \frac{\mathscr{A}_{uv}}{\sqrt{\mathscr{d}_u \mathscr{d}_v}} = \frac{\theta_u\theta_v\mathbf{B}_{z(u)z(v)}}{\sqrt{\mathscr{d}_u \mathscr{d}_v}} = \frac{\mathbf{B}_{z(u)z(v)}\sqrt{\theta_u\theta_v}}{\sqrt{\mathbf{d}_{z(u)}\mathbf{d}_{z(v)}}} = [\mathbf{L}]_{z(u)z(v)}\sqrt{\theta_u\theta_v},$$

or equivalently,

$$\mathscr{L} = \Theta^{1/2}Z\mathbf{L}Z^T\Theta^{1/2}.$$

Then

$$\mathscr{X}^T\Lambda\mathscr{X}' = \mathscr{D}^{-1/2}\Theta^{1/2}ZU\Lambda U^T Z^T\Theta^{1/2}\mathscr{D}^{1/2} = \mathscr{D}^{-1/2}\mathscr{L}\mathscr{D}^{1/2} = \mathscr{D}^{-1}\mathscr{A} = \mathscr{P},$$

and

$$\mathscr{Y}^T\Lambda\mathscr{Y}' = \mathbf{D}^{-1/2}U\Lambda U^T\mathbf{D}^{1/2} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{1/2} = \mathbf{P},$$

where $\mathscr{X}' = \mathrm{U}^{\mathrm{T}}\mathrm{Z}^{\mathrm{T}}\Theta^{1/2}\mathscr{D}^{1/2}$ and $\mathscr{Y}' = \mathrm{U}^{\mathrm{T}}\mathbf{D}^{1/2}$ are right eigenvectors if $\mathscr{P}$ and $\mathbf{P}$ respectively. Recognizing that $\mathscr{X}^{\mathrm{T}}\mathscr{X}' = \mathscr{Y}^{\mathrm{T}}\mathscr{Y}' = \mathrm{I}_{\mathrm{K}}$ completes the proof. □

Lemma A.10. *Let* L *be a symmetric matrix, let* D *be a diagonal matrix, and let* $\mathrm{P} = \mathrm{D}^{-1/2}\mathrm{L}\mathrm{D}^{1/2}$. *If* x *is an eigenvector of* L *corresponding to eigenvalue* λ, *then*

(a) $\mathrm{D}^{-1/2}x$ *is a right eigenvector of* P *with eigenvalue* λ,

(b) $\|\mathrm{P}^{\mathrm{T}}\mathrm{P}\| = \|\mathrm{L}\|^2$.

Proof. Let $y = \mathrm{D}^{-1/2}x$, then

$$\mathrm{P}y = \mathrm{D}^{-1/2}\mathrm{L}\mathrm{D}^{1/2}y = \mathrm{D}^{-1/2}\mathrm{L}x = \lambda\mathrm{D}^{-1/2}x = \lambda y.$$

Part a of the lemma follows. To see b, observe that y is also an eigenvector of $\mathrm{P}^{\mathrm{T}}\mathrm{P}$ with eigenvalue λ^2. □

Lemma A.10 implies that P has the same spectral norm of graph Laplacian L. Since L concentrates to $\mathscr{L}$ (see for example Qin and Rohe (2013) for a proof), we have under a random graph generated from the DC-SBM, the graph transition matrix P concentrates to its population $\mathscr{P}$ with respect to spectral norm.

Teleportation constant α

In the paper, we state that a sufficiently large teleportation constant α enables the entrywise control of sample PPR vector, thus facilitating the PPR clustering in a random graph. Here, from a practical perspective, we further illustrate the sensitivity of PPR clustering to α, with the Twitter friendship network. To this end, we investigate the targeted sampling returned by four configurations of the teleportation constant, $\alpha \in \{0.1, 0.15, 0.25, 1/3\}$, where NBC Politics (@NBCPolitics) is the seed. The tolerance parameter is fixed, $\varepsilon = 10^{-7}$, in four targeted sampling.

Table A.1 lists the number of Twitter users we examined and the total number of users we “reached” (as of August 2019) in four attempts. Here, we examine a user by retrieving its friend list (after which it gets a positive p_u

Table A.1: Number of nodes examined/reached by Algorithm 2.3 with seed node @NBCPolitics and different teleportation constants, and a fixed tolerance parameter $\varepsilon = 10^{-7}$, as in August 2019.

α	Examined	Reached
0.1	7,445	342,454
0.15	5,919	272,985
0.25	4,860	228,561
1/3	3,984	193,848

value in Algorithm 2.3), and reach a user once it appears in a user's friend list (at which point, it possesses a positive r_v value in Algorithm 2.3). Given the same tolerance parameter, varying the teleportation constant largely affects the number of nodes examined/reached. This demonstrates the role of teleportation constant in leveraging between the seed preference and the standard random walk.

Despite the fact that different α's result in substantial difference in network coverage, when the algorithm stops, the estimated local clusters appear to share a vast majority in common. To demonstrate this immediately, we inspect the local clusters of size $n = 300$ returned by Algorithm 2.4 with four α's and quantify to what degree do they overlap each other. Table A.2 shows the percentage of common members between each pair of four returned local clusters. As shown, most pairs have about 90% overlapping members, indicating that PPR clustering is fairly robust against the teleportation constant.

The stability of PPR clustering continues to show when we vary the cluster size, $n = 100, 150, ..., 700$. Figure A.1 shows the proportion of common members across *all* four local clusters, returned by PPR, aPPR, and rPPR (with the regularizer $\tau = 10$). Overall, the PPR clustering produces a fairly consistent local cluster, with around 80% of members overlapping across four different strengths of teleportation (see Supplementary Materials).

We conclude that in practice, PPR clustering (i) is mainly influential to the number of nodes examined in the targeted sampling and (ii) has fairly robust performance with respect to the choice of teleportation constant.

Table A.2: Percentage of pairwise overlapping among three local clusters around @NBCPolitics with different teleportation constants, $\alpha \in \{0.1, 0.15, 0.25, 1/3\}$, as in August 2019.

α	0.1	0.15	0.25	1/3
0.1	100%	92.7%	89.3%	87.7%
0.15		100%	93.3%	90%
0.25			100%	92%
1/3				100%

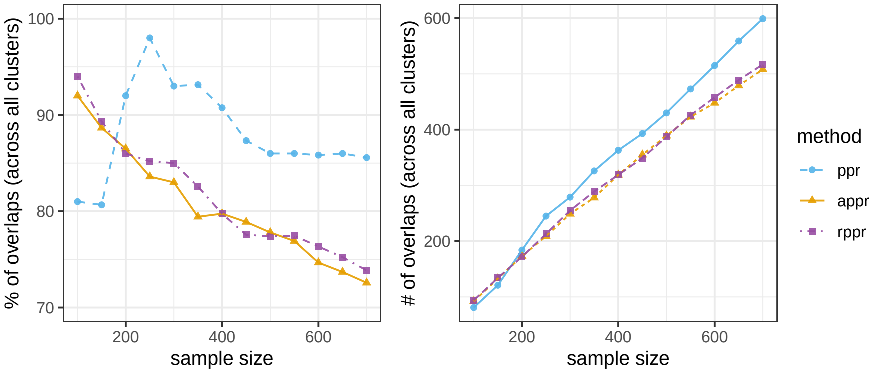


Figure A.1: Sensitivity to the teleportation constant, $\alpha = \{0.1, 0.15, 0.25, 1/3\}$. Shown are the percentage (left) and number (right) of common members across all four local clusters returned by three PPR clustering methods. The targeted sample size increases from 100 to 700 with the increment of 50.

The graph size N

In the paper, we provide an entrywise error control for the PPR vector and the aPPR vector (Theorem 2.6), assuming the edge density is sufficiently large (i.e., inequality (2.7) in the main paper). Simulation 3 in Section 4.4 demonstrates the relationship between the expected degree (δ) and the error rate of PPR clustering, as promised by the theorem. Here, we provide another simulation to illustrate Theorem 2.6. Specifically, we further investigate the affect of graph size (N) on the relative entrywise error (REE) of the PPR vector ($\frac{\|p-\mathcal{p}\|_\infty}{\|\mathcal{p}\|_\infty}$), given some edge density (δ).

We generate 30 replicates of networks of size $N = e^x$, where $x \in \{6.5, 7, 7.5, 8, 8.5, 9\}$,

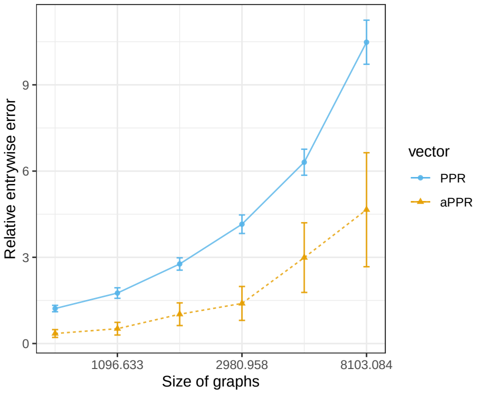


Figure A.2: Entrywise error rate versus the graph sizes. Shown are relative entrywise error (REE) corresponding to different underlying graph sizes, averaged over 30 replicates. For each dot, an error bar indicates the standard error. The RER for aPPR vector is scaled down by a factor of 240 to improve visualization. The ticks in x-axis are transformed through logarithm with the natural base.

from the four-parameter stochastic block model, SBM$(K = 3, N, b_1 = 9, b_2 = 3)$. The average expect degree is set to $\delta = 125$. Both PPR vectors and aPPR vectors are calculated for every network, with teleportation constant $\alpha = 0.15$ and 10 seeds randomly selected from the first block. Figure A.2 depict the REE with respect to different graph sizes (scaled by a logarithm transformation with the natural base). As shown, with δ fixed (not growing at the rate of $\log N$), the REE increases as the graph expands, so does the variance of REE for both PPR and aPPR vectors, matching the results in Theorem 2.6.

A.3 Connection to linear discriminant analysis

In this section, we give another representation of PPR vectors in the landing probability space, which builds upon Kloumann et al. (2017). This assorts PPR to a greater functional regime. Then, we extend the previous result that links the PPR vectors with linear discriminant functions under the DC-SBM. In particular, when every block has the same degree (volume), where $\mathbf{D}$ becomes a scalar matrix, the PPR vector is asymptotically equivalent to the optimal linear discriminant function.

First, we briefly introduce linear discriminant analysis in landing probability space, which the PPR vector also lives in. Consider a random walk on the graph starting from a seed node. Define the *landing probability* r_s^{ν} to be the probability that the random walk ends up at $\nu \in V$ after exactly s steps. The *landing probability space* is the space of landing probability of any nodes.

A *linear discriminant* (LD) analysis keeps the first S landing probability on each node, $r^{\nu} = (r_0^{\nu}, r_1^{\nu}, ..., r_S^{\nu}) \in \mathbb{R}^S$, and divides vertices into two sets by thresholding on the linear discriminant score vector $l \in \mathbb{R}^N$, whose ν-th entry is defined to be inner product

$$l_{\nu} = \langle \omega, r^{\nu} \rangle$$

with some weights $\omega \in \mathbb{R}^S$. For example, let $\omega = r^{\nu_1} - r^{\nu_2}$, where ν_1, ν_2 are empirical centroids of two node sets. Then l_{ν} increases as ν slides from ν_1 to ν_2, and thresholding $(\|\nu_1\|^2 - \|\nu_2\|^2)/2$ allocates vertices to nearest centroid.

Remark. The landing probability of the s-th step, $r_s = \left(r_s^1, r_s^2, ..., r_s^N\right) \in \mathbb{R}^N$, is defined as $\left(P^s\right)^T \pi$. It follows from proposition 2.1 that PPR vector $p = \sum_{s=0}^{\infty} \phi_s r_s$ with $\phi_s = \alpha(1-\alpha)^s$. Keeping the first S terms yields an LD score vector with the weights $\omega_{PPR} = (\phi_0, \phi_1, ..., \phi_{S-1})$.

We then perform population (expectation) analysis for PPR in the landing probability space. Define the population *block landing probability* $\mathbf{W}_s^k$ to be the probability that a random walk from ν_0 ends up in block k after exactly s steps, where $k = 1, 2, ..., K$ and $s = 0, 1, ..., S-1$. Given that ν_0 is in block 1, $\mathbf{W}_0^{\cdot} = (1, 0, ..., 0)^T$. Using the first S steps block landing probabilities, the next lemma gives an explicit form of LD vectors.

Lemma A.11 (Explicit form of LD vectors)**.** *Under the population DC-SBM with* K *blocks and parameters* $\{\mathbf{B}, Z, \Theta\}$*, assume all blocks have the same degrees. Let* $\ell(k)$ *be the linear discriminant score vector between block 1 and block* k*. Then,*

(a) $\mathbf{W}_s^{\cdot} = \mathbf{P}^T \mathbf{W}_{s-1}^{\cdot}$, $s = 1, 2, ..., S-1$*; and*

(b) $\ell(k) = \Theta Z \boldsymbol{l}(k)$, $k = 2, ..., K$, *where* $\boldsymbol{l}(k) = \mathbf{W}\mathbf{W}^T(e_1 - e_k)$.

Here, e_k *is the elementary unit vector on the direction of* k*-th block.*

Proof. We prove a using following quantities. Let E_s^k be the number of paths from v_0 to block k with exact length s, and let $\mathscr{E}_s^k$ be the expected number of paths from v_0 to block k with exact length s. Recall from 2.3 that $\mathbf{B}_{ij}$ represents the expected number of edges between block i and j if $i \neq j$, or twice of that if $i = j$. Then,

$$\mathscr{E}_s^k = \sum_{j=1}^{K} \mathbf{B}_{kj} \mathscr{E}_{s-1}^j.$$

To see $\mathbf{W}_s^{\cdot} = \mathbf{P}^T \mathbf{W}_{s-1}^{\cdot}$, observe that

$$\mathbf{W}_s^k = \frac{\mathscr{E}_s^k}{\sum_{i=1}^K \mathscr{E}_s^i} = \frac{\sum_{j=1}^K \mathbf{B}_{kj} \mathscr{E}_{s-1}^j}{\sum_{i=1}^K \sum_{j=1}^K \mathbf{B}_{ij} \mathscr{E}_{s-1}^j} = \frac{\sum_{j=1}^K \mathbf{B}_{kj} \mathscr{E}_{s-1}^j}{\sum_{j=1}^K \mathbf{d}_j \mathscr{E}_{s-1}^j} = \sum_{j=1}^{K} \mathbf{P}_{kj} \mathbf{W}_{s-1}^k.$$

The last equality comes from the assumption that all blocks have the same degrees, which means $\mathbf{d}_i$ is constant.

Now, we prove part b of the lemma. Let $R \in \mathbb{R}^{N \times S}$ collect all landing probabilities r_s^v of the first S steps, where $v = 1, 2, ..., N$ and $s = 0, 1, ..., S-1$. Without loss of generality, assume the seed node corresponds to the first row. Define $\mathscr{R} = \mathbb{E}(R) \in [0,1]^{N \times S}$ to be the population version of R. Then the population landing probability is explicitly

$$\mathscr{R}_s^v = \frac{\mathscr{d}_v}{\mathbf{d}_{z(v)}} \mathbf{W}_s^{z(v)} = \theta_v \mathbf{W}_s^{z(v)},$$

or compactly,

$$\mathscr{R} = \Theta Z \mathbf{W}.$$

In linear discriminant, the weights vector ω is the geometric difference between centroid of block 1 and k, which can be written as

$$\left(\sum_{v:z(v)=1} \mathscr{R}_1^v - \sum_{v:z(v)=k} \mathscr{R}_1^v, \sum_{v:z(v)=1} \mathscr{R}_2^v - \sum_{v:z(v)=k} \mathscr{R}_2^v, ..., \sum_{v:z(v)=1} \mathscr{R}_S^v - \sum_{v:z(v)=k} \mathscr{R}_S^v \right),$$

or compactly

$$\omega = \mathscr{R}^{\mathrm{T}} Z(e_1 - e_k).$$

By Lemma A.1, the linear discriminant score vector reads

$$\begin{aligned}\langle \mathscr{R} \cdot \omega \rangle &= \mathscr{R}\mathscr{R}^{\mathrm{T}} Z(e_1 - e_k)\\ &= \Theta Z \mathbf{W}\mathbf{W}^{\mathrm{T}}(e_1 - e_k),\end{aligned}$$

for $k = 2, ..., K$. Setting $\mathbf{l}(k) = \mathbf{W}\mathbf{W}^{\mathrm{T}}(e_1 - e_k)$ completes the proof. □

Recall from Theorem 2.4 that $\mathscr{p} = \Theta Z \mathbf{p}$. The LD score vector ℓ has a similarly simple form that separates the block-related information ($\mathbf{W}$) and the node specific information (Θ and Z). Lemma A.11 provides a population (expectation) representation of PPR in the landing probability space. To facilitate its application in random graphs, the next lemma provides a control of the landing probabilities on a random block model graph.

Lemma A.12 (Concentration of landing probabilities)**.** *Let* $G = (V, E)$ *be a graph of* N *vertices generated from the DC-SBM with* K *blocks and parameters* $\{B, Z, \Theta\}$. *Let* $R_s^{\cdot} \in [0, 1]^N$ *be the landing probabilities of the* k*-th step, and* $\mathscr{R}_s^{\cdot} = \mathbb{E}(R)$ *be its expectation. Then, for any* $\varepsilon > 0$ *and any vertex* $u = 1, 2, ..., N$,

$$\mathbb{P}\left(R_s^u \geqslant (1 + \varepsilon)\mathscr{R}_s^u\right) \leqslant (1 + \varepsilon)^{-\varepsilon N r},$$

$$\mathbb{P}\left(R_s^u \geqslant (1 - \varepsilon)\mathscr{R}_s^u\right) \leqslant (1 - \varepsilon)^{\varepsilon N r},$$

where $r = \min_{v \in V} \theta_u \theta_v \mathbf{P}_{z(u)z(v)} \mathbf{W}_{s-1}^{z(v)}$.

Proof. Note that $R_s^u = \sum_{v \in V} X_{uv}$, where

$$X_{uv} = \frac{\mathbf{W}_{s-1}^{z(v)}}{\mathbf{d}_{z(v)}} \mathbb{1}_{\{A_{uv} = 1\}}$$

are independent random variables having probability $\theta_u \theta_v \mathbf{B}_{z(u)z(v)}$ of being equal to $\mathbf{W}_{s-1}^{z(\nu)}/\mathbf{D}_{z(\nu)z(\nu)}$. Then,

$$\mathbb{E}\left[R_s^u\right] = \sum_{\nu \in V} \frac{\mathbf{W}_{s-1}^{z(\nu)}}{\mathbf{d}_{z(\nu)}} \theta_u \theta_\nu \mathbf{B}_{z(u)z(\nu)} = \theta_u \sum_{k=1}^{K} \mathbf{P}_{z(u)k} \mathbf{W}_{s-1}^{k} = \mathscr{R}_s^u.$$

We can apply Chernoff's bounds (Lemma A.7) on R_s^u and obtain bounds for any fixed u,

$$\mathbb{P}\left(R_s^u \geqslant (1+\varepsilon)\mathscr{R}_s^u\right) \leqslant (1+\varepsilon)^{-\varepsilon \mathscr{R}_s^u},$$

and

$$\mathbb{P}\left(R_s^u \leqslant (1-\varepsilon)\mathscr{R}_s^u\right) \leqslant (1-\varepsilon)^{\varepsilon \mathscr{R}_s^u}.$$

Recognizing that $\mathscr{R}_s^u \geqslant Nr$ completes the proof. □

Lemma A.12 provides an entrywise concentration bound for landing probabilities. The next theorem equates PPR and LD vectors when blocks are equally distributed. Together, they asserts the asymptotically equivalence between PPR and LD vectors, in symmetric block model graphs.

Theorem A.13 (Equivalence between PPR and LD vectors)**.** *Under the population DC-SBM with* K *blocks and parameters* $\{\mathbf{B}, Z, \Theta\}$*, assume* $B_{ii} = b_1$ *for all* i*, and* $\mathbf{B}_{ij} = b_2$ *for* $i \neq j$ $(b_1 > b_2 > 0)$*. Let* λ_2 *the second largest eigenvalue of* $\mathscr{P}$*. Let* $\not{p}$ *be the personalized PageRank vector, and let* $\ell(k)$ *be the linear discriminant score vector between block 1 and block* k*,* $k = 2, ..., K$*. If the teleportation constant* $\alpha = 1 - \lambda_2$*, then*

$$\not{p} \propto \ell(k).$$

Proof. From Section A.2 and Lemma A.11a, the block landing probability is precisely

$$\mathbf{W}_s^k = \sum_{j=1}^{K} \lambda_j^s U_{kj} U_{1j},$$

where λ_k is the k-th eigenvalues of $\mathscr{P}$ and U is the orthogonal matrix used in Lemma A.2.

Note that $\mathbf{B}$ has eigenvalues $\lambda_1 = 1$ and $\lambda_2 = \frac{b_1 - b_2}{b_1 + b_2}$, with complexity indices 1 and $k - 1$ respectively. In addition, we know the orthogonal matrix above precisely as well,

$$U = \begin{bmatrix} \frac{1}{\sqrt{N}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \cdots & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{N}} & -\frac{1}{\sqrt{2}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{N}} & 0 & -\frac{1}{\sqrt{2}} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \frac{1}{\sqrt{N}} & 0 & 0 & \cdots & -\frac{1}{\sqrt{2}} \end{bmatrix}.$$

Then it follows from Lemma A.11b that the LD weight vector is

$$\begin{aligned} \omega_{LD} &= \mathscr{R}^T Z(e_1 - e_k) \\ &= \mathbf{W}^T(e_1 - e_k) \\ &= \mathbf{W}_{\cdot}^1 - \mathbf{W}_{\cdot}^k \\ &= \sum_{j=1}^{K} \lambda_j^s (U_{1j} - U_{kj}) U_{1j} \\ &= \frac{K}{2} \begin{bmatrix} 1 \\ \lambda_2 \\ \lambda_2^2 \\ \vdots \\ \lambda_2^{S-1} \end{bmatrix}. \end{aligned}$$

On the other hand, the weight vector of PPR on landing probability space is $\omega_{PPR} = (\phi_0, \phi_1, ...)$, where $\phi_s = \alpha(1-\alpha)^s$. Hence, setting the teleportation constant $\alpha = 1 - \lambda_2$ asymptotically equates approximate PPR and LD vectors, up to a scalar factor. □

Remark. First, a positive factor that differentiates PPR and LD vectors does not change the relative ranking of the nodes, because the ranking via $\mathscr{p}$ or $c\mathscr{p}$ is equivalent. Hence, Theorem A.13 shows that the PPR vector is equivalent to an optimal LD score vector under described population DC-SBM. Second, Theorem A.13 is an extension of Kloumann et al. (2017). Combining Theorem A.13 and Lemma A.12 gives the asymptotic equivalence between PPR and LD vectors under the particular DC-SBM stated.

A.4 Lists of top 200 handles

In this section, we supply three lists of handles resulting from sampling using PPR, aPPR, and rPPR vectors with @NBCPolitics as the seed, as of December 2018. We conceal handles with followers count fewer than 200 for privacy considerations. The biographical descriptions are trimmed for unifying displays. In addition, we annotate handles with whether or not they are followed ("Followed") by the seed node.

A PPR's sample of 200

Table A.3: Top (selected) handles returned by PPR.

	Name	Followed	Followers	Description
1	Melania Trump	Yes	11242283	This account is run by the Office of First Lady Melania Trum...
2	The White House	Yes	17625630	Welcome to @WhiteHouse! Follow for the latest from President...
3	Chuck Todd	Yes	2032038	Moderator of @meetthepress and @nbcnews political director; ...
4	NBC News	Yes	6280551	The leading source of global news and info for more than 75 ...
5	NBC Nightly News	Yes	962290	Breaking news, in-depth reporting, context on news from arou...
6	Andrea Mitchell	Yes	1737764	NBC News Chief Foreign Affairs Correspondent/anchor, Andrea ...
7	Savannah Guthrie	Yes	881669	Mom to Vale & Charley, TODAY Co-Anchor, Georgetown Law...
8	Joe Scarborough	Yes	2521215	With Malice Toward None
9	MSNBC	Yes	2261911	The place for in-depth analysis, political commentary and in...
10	Rachel Maddow MSNBC	Yes	9498076	I see political people... (Retweets do not imply endorsement...
11	Breaking News	Yes	9223158	
12	NBC News First Read	Yes	53847	The first place for news and analysis from the @NBCNews Poli...
13	TODAY	Yes	4276453	America's favorite morning show | Snapchat: todayshow
14	Meet the Press	Yes	566713	Meet the Press is the longest-running television show in his...
15	The Wall Street Journal	Yes	16188842	Breaking news and features from the WSJ.
16	Pete Williams	Yes	70062	NBC News Justice Correspondent. Covers US Supreme Court, ...
17	Mark Murray	Yes	97571	Mark Murray is the senior political editor for NBC News, as ...
18	POLITICO	Yes	3695835	Nobody knows politics like POLITICO. Got a news tip for us? ...
19	Katy Tur	Yes	587474	MSNBC anchor @2pm, NBC News correspondent, author of NYT ...
20	Bill Clinton	Yes	10697521	Founder, Clinton Foundation and 42nd President of the United ...
21	Kasie Hunt	Yes	381704	@NBCNews Capitol Hill Correspondent. Host, @KasieDC, Sundays...
22	TIME	Yes	15584815	Breaking news and current events from around the globe. Host...
23	Kelly O'Donnell	Yes	195765	White House Correspondent @NBCNews Veteran of Cap Hill ...
24	John McCain	Yes	3181773	Memorial account for U.S. Senator John McCain, 1936-2018. To...
25	Peter Alexander	Yes	283522	@NBCNews White House Correspondent / Weekend @TODAYshow ...
26	Hallie Jackson	Yes	359099	Chief White House Correspondent / @NBCNews / @MSNBC ...
27	Kristen Welker	Yes	182244	@NBCNews White House Correspondent. Links and retweets ...
28	Carrie Dann	Yes	37119	.@NBCNews / @NBCPolitics. RTs not endorsements.

...continued

	Name	Followed	Followers	Description
29	Willie Geist	Yes	807536	Host @NBC #SundayTODAY, Co-Host @Morning_Joe, “Sunday ...
30	Morning Joe	Yes	563650	Live tweet during the show! Links to must-read op-eds and ...
31	Frank Thorp V	Yes	58152	Producer & Off-Air Reporter covering Congress at @NBCNews ...
32	Mark Knoller	Yes	318923	CBS News White House Correspondent
33	Tom Brokaw	Yes	308276	Special correspondent, @NBCNews
34	Mika Brzezinski	Yes	868124	"Bipartisanship helps to avoid extremes and imbalances. It ...
35	Chris Jansing	Yes	72375	@msnbc Senior National Correspondent, intrepid traveler and ...
36	John Harwood	Yes	251246	a Dad who covers Washington, the economy and national politi...
37	Nicolle Wallace	Yes	413153	Author of 18 Acres series, mom, dog walker, wife, gardener. ...
38	NBC News Signal	Yes	83715	A new streaming news channel from @NBCNews. Catch us Thursda...
39	Sam Stein	Yes	392003	Daily Beast/MSNBC newsletter: https://t.co/DVURxntWdL Emai...
40	Chris Matthews	Yes	882434	Host of @hardball M-F at 7PM ET on @MSNBC and author of "Bob...
41	Carol Lee	Yes	51240	Reporter for NBC News, former WSJ & POLITICO, Hudson’s mom, ...
42	Ali Vitali	Yes	78839	@NBCnews Political Reporter. Covered Trump campaign, WH + ...
43	Ken Dilanian	Yes	124635	Intelligence and national security reporter for the NBC News...
44	Jim Miklaszewski	Yes	14196	Jim Miklaszewski is Chief Pentagon Correspondent for NBC New...
45	John Heilemann	Yes	247616	@SHO_TheCircus host/ep; NBCNews/@MSNBC natl affairs analyst;...
46	Stephanie Ruhle	Yes	352895	Mom, MSNBC LIVE Anchor 9AM M-F, VELSHI & RUHLE 1 PM ...
47	Nick Confessore	Yes	172359	Reporter for @NYTimes, writer-at-large for @NYTmag, MSNBC ...
48	Talking Points Memo	Yes	275692	Breaking news and analysis from the TPM team. “I’ll leave ...
49	Tom Costello	Yes	17268	NBC News Correspondent covering Aviation, Transportation, Ec...
50	Post Politics	Yes	384611	The latest political news and analysis from The Washington P...
51	Alex Moe	Yes	28245	@NBCNews Capitol Hill Producer + Off-Air Reporter; ’12 & ’16...
52	Benjy Sarlin	Yes	100896	Political reporter for @NBCNews. I cover elections and their...
53	Preet Bharara	Yes	945030	Patriotic American & proud immigrant. Movie buff. @Springste...
54	Matthew Miller	Yes	229867	Partner at Vianovo. MSNBC Justice & Security Analyst. Recove...
55	Leigh Ann Caldwell	Yes	20714	NBC Capitol Hill reporter. Formerly at CNN and public radio....
56	Ken Strickland	Yes	2693	NBC News Washington Bureau Chief
57	Ron Fournier	Yes	64356	President: Truscott Rossman. Best-seller https://t.co/09CdTN...
58	Mike Memoli	Yes	39693	National Political Reporter @nbcnews; @latimes alum mike dot...
59	Miguel Almaguer	Yes	14082	Prolific coffee drinker. Chronic under sleeper. Raging road ...
60	Courtney Kube	Yes	9494	NBC News National Security & Military Reporter. Links and ...
61	NBC News World	Yes	279165	A dynamic look at world events from @NBCNews.
62	Jonathan Martin	Yes	241690	Nat’l Political Correspondent, NY Times. Husband of the ...
63	Steve Schmidt	Yes	498812	"Patriotism means to stand by the country. It does not mean ...
64	Jenna Bush Hager	Yes	207106	Mama to M and P, NBC News correspondent, Editor-at-Large ...
65	Sean Spicer	Yes	406957	President of RigWil, Sr Advisor @AmericaFirstPAC check out ...
66	Roll Call	Yes	356374	Breaking news, reporter tweets and analysis from the Source ...
67	POLITICO 45	Yes	88470	A daily diary of the 45th president of the United States.
68	Scott Foster	Yes	3464	Senior Producer, Washington @NBCNEWS @TODAYshow
69	Domenico Montanaro	Yes	83999	"Congress shall make no law respecting an est. of religion, ...
70	Tom Winter	Yes	40777	NBC News Investigations reporter based in New York focusing ...
71	Kailani Koenig	Yes	11416	Producer with @MSNBC & @NBCNews. Team @MeetThePress ...
72	Capital Journal	Yes	131212	WSJ’s home for politics, policy and national security news. ...

...continued

	Name	Followed	Followers	Description
73	NBC News Videos	Yes	7838	The latest video from http://t.co/xPyvMOTEF6
74	Diane Sawyer	Yes	876906	I like my news 24/7, my food spicy, my drinks caffeinated, ...
75	Jane C. Timm	Yes	6478	@nbcnews political reporter and fact checker. More fun than ...
76	Elyse PG	Yes	2697	White House producer @nbcnews |@USCAnnenberg alum | LA kid ...
77	Libby Leist	Yes	7946	Executive Producer @todayshow
78	Mike Barnicle	Yes	116588	Mike Barnicle is an award-winning print and broadcast journa...
79	Reuters Politics	Yes	259106	U.S. political coverage, breaking news and special investiga...
80	Beth Fouhy	Yes	13684	Senior editor, politics, NBC News and MSNBC
81	HuffPost	Yes	11401771	Know what's real.
82	Joey Scarborough	Yes	6277	NBC News Social Media Editor. New York Daily News Alum. RTs ...
83	Marianna Sotomayor	Yes	11965	Running around Capitol Hill for @NBCNews. Covers politics ...
84	Shaquille Brewster	Yes	5362	@NBCNews Producer/Politics | @HowardU Alum| Journalist ...
85	Joyce Alene	Yes	185116	U of Alabama Law Professor|@MSNBC Contributor|Obama US ...
86	Garrett Haake	Yes	40714	Correspondent @msnbc • Taller than I look on TV • Long-suffe...
87	Andrew Rafferty	Yes	16567	Senior political editor for @newsy Before that @NBCNews ...
88	Jacob Soboroff	Yes	144153	@MSNBC correspondent. Instagram & Snapchat: jacobsoboroff
89	Perry Bacon Jr.	Yes	26853	I write about government (mostly federal, often state, ...
90	Alex Witt	Yes	28126	Weekend host on @MSNBC (9am, noon & 1pm). Tigger's mom ...
91	Mark Halperin	Yes	332564	New York, New York
92	Heidi Przybyla	Yes	66489	NBC News, n'tl political reporter "Prezbella" Heidi.Przyb...
93	Morgan Radford	Yes	20967	@NBCnews Correspondent: @TODAYShow/@NBCNightlyNews .
94	Savannah Sellers	Yes	4637	News junkie. Host of NBC's "Stay Tuned" on Snapchat. Storyte...
95	Marist Poll	Yes	16030	Founded in 1978, MIPO is home to the Marist Poll and regular...
96	Jill Wine-Banks	Yes	158753	@NBCNews & @MSNBC Contributor. Speaker. Watergate prosecutor...
97	NBC Field Notes	Yes	1390	NBC News correspondents and http://t.co/1eSopOQt8s reporters...
98	Olivia Nuzzi	Yes	190919	Washington Correspondent, New York Magazine
99	NBC News THINK	Yes	12017	THINK is NBC News' home for fresh opinion, sharp analysis ...
100	Making a Difference	Yes	670	@NBCNightlyNews' popular feature profiles ordinary people do...
101	adam nagourney	Yes	25307	LA Bureau Chief for The New York Times. Story ideas welcome ...
102	Phil McCausland	Yes	2519	@NBCNews Digital reporter focused on the rural-urban divide....
103	Katie Couric	Yes	1746116	Journalist, podcaster, @SU2C founder, doc filmmaker of @FedU...
104	Monica Alba	Yes	30034	@NBCNews White House team. Covered Hillary Clinton on the ...
105	Vicente Fox Quesada	Yes	1244017	Presidente de México de 2000 a 2006 y ahora trabajando po...
106	Alex Johnson	Yes	4371	News, data and analysis for @NBCNews; data geek; non-celebri...
108	Alex Seitz-Wald	Yes	50168	Political reporter for @NBCNews covering Democrats | Tips, ...
109	Anthony Terrell	Yes	6827	Emmy Award winning journalist. Political observer. Covered ...
110	Sam Petulla	Yes	2588	Editor @cnnpolitics • Usually looking for datasets. You can ...
111	Debra Messing	Yes	532941	Actor. Mama. Global Ambassador for HIV/AIDS for PSI. Activis...
112	Corky Siemaszko	Yes	2538	Senior Writer at NBC News Digital (former NY Daily News rewr...
114	Zach Haberman	Yes	3693	Lead Breaking News Editor, @NBCNews. Previously had other jobs ...
115	NBC Latino	Yes	67920	Elevating the conversation around Latino news in the United ...
116	Vivian Salama	Yes	16020	White House reporter for @WSJ. Formerly AP Baghdad bureau ...
117	Zeke Miller	Yes	215054	White House Reporter @AP. Email: zekejmiller@gmail.com Links...
118	Vaughn Hillyard	Yes	31464	On the Road, Meeting Good Folk | NBC News | Arizonan | IG: @...

...continued

	Name	Followed	Followers	Description
119	Jonathan Allen	Yes	44477	political reporter, @NBCNews Digital | co-author, NYT bestse...
121	HuffPost Politics	Yes	1428870	The latest political news from HuffPost's politics team.
122	Nick Akerman	Yes	14949	Partner in the AmLaw 100 law firm of Dorsey & Whitney, Water...
123	CSPAN	Yes	1915821	Capitol Hill. The White House. National Politics.
124	John McCormack	Yes	30688	Senior writer at The Weekly Standard.
125	Jo Ling Kent	Yes	32957	NBC News Correspondent @NBCNightlyNews, @TODAYshow ...
126	PolitiFact	Yes	628659	Home of the Truth-O-Meter and independent fact-checking ...
127	Bob Corker	Yes	10042	Serving Tennesseans in the U.S. Senate
128	Elise Jordan	Yes	58884	Co-host of @WMM_podcast podcast. @MSNBC/@NBCNews political...
129	Greg Martin	Yes	1161	Political Booking Producer at @nbcnews @todayshow
130	Education Nation	Yes	276468	Hosted by @NBCNews. Creator of Parent Toolkit & moderator of...
131	Micah Grimes	Yes	25948	Head of Social, @NBCNews & @MSNBC – Foreign and domestic ...
132	Jill Lawrence	Yes	17282	Commentary editor and columnist @USATODAY. Author of The Art...
133	McKay Coppins	Yes	131623	Staff writer at @TheAtlantic. Author of THE WILDERNESS. 'Sor...
134	Emmanuelle Saliba	Yes	4004	Head of Social Media Strategy @Euronews | Launched #THECUBE ...
135	Hasani Gittens	Yes	3002	Level 29 Mage. Senior News Ed. @NBCNews. Sheriff of Nattahna...
136	Rebecca Sinderbrand	Yes	18691	Now: @NBCNews Senior Washington Editor, visiting lecturer @Y...
137	BuzzFeed Politics	Yes	121646	News and updates from the politics team @BuzzFeedNews.
138	Adam Edelman	Yes	2341	Political reporter @nbcnews. Wisconsin native, Bestchester r...
139	Ethan Klapper	Yes	18292	Journalist (@YahooNews) and #avgeek.
140	President Trump	No	24593638	45th President of the United States of America, @realDonaldT...
141	Vice President Mike ...	No	6795022	Vice President Mike Pence. Husband, father, & honored to ...
142	Donald J. Trump	No	56050499	45th President of the United States of America
143	Karen Pence	No	403315	Educator, mom, wife of @VP Pence. Passionate about art thera...
144	Sarah Sanders	No	3522219	@WhiteHouse Press Secretary. Proudly representing @POTUS ...
145	Kellyanne Conway	No	2506546	Mom. Patriot. Catholic. Counselor.
146	DRUDGE REPORT	No	1408129	The DRUDGE REPORT is a U.S. based news aggregation website ...
147	White House History	No	104010	The White House Historical Association is a non-profit organ...
148	The New York Times	No	42412491	Where the conversation begins. Follow for breaking news, ...
149	White House Archived	No	13379715	This is an archive of an Obama Administration account mainta...
150	Dan Scavino Jr.	No	324561	Assistant to President @realDonaldTrump, Director of Social ...
151	Drudge Buzz	No	104111	Tracking the buzz made by Americas #1 newsmaker Matt Drudge....
152	David Gregory	No	1749373	CNN, Georgetown U
153	Hillary Clinton	No	23643522	2016 Democratic Nominee, SecState, Senator, hair icon. Mom, ...
154	CNN Breaking News	No	54476034	Breaking news from CNN Digital. Now 54M strong. Check @cnn ...
155	The Cabinet	No	123597	The @WhiteHouse Office of Cabinet Affairs. Tweets may be arc...
156	Lester Holt	No	501427	Anchor @NBCNightlyNews and @datelinenbc, reporting on the to...
157	John Dickerson	No	48122	Co-host CBS This Morning. This account @johndickerson is mos...
158	CNN	No	40854429	It's our job to #GoThere & tell the most difficult stories. ...
159	J Earnest (Archived)	No	1182091	WH Press Secretary. This is an archive of an Obama Administr...
160	The Washington Post	No	13117609	Breaking news, analysis, and opinion. Founded in 1877. Our ...
161	Adam Liptak	No	61589	Supreme Court reporter for The New York Times
162	NSC	No	35905	National Security Council | Tweets may be archived ...
163	MSNBC video	No	40669	Favorite video highlights from @msnbc.

...continued

	Name	Followed	Followers	Description
164	Gorsuch Facts	No	39143	Judge Gorsuch will be fair to all regardless of their backgr...
165	Greg Stohr	No	11651	Supreme Court reporter for Bloomberg News. Baseball dad ...
166	OMB Press	No	11182	Office of Management and Budget | Tweets may be archived: ...
167	Richard Engel	No	288066	@NBCNews Chief Foreign Correspondent
168	Norah O'Donnell	No	195549	Wife, mother of 3, Co-Host @cbsthismorning, #1 fan of @chefg...
169	Robert Barnes	No	37361	Robert Barnes covers the Supreme Court for The Washington Po...
170	Luke Russert	No	253495	Sometimes nothing can be a real cool hand. STA'04/BC'08
171	Stephen Colbert	No	18269222	the guy on CBS
172	Mark Sherman	No	6336	
173	U.S. Attorney EDVA	No	5709	Led by U.S. Attorney G. Zachary Terwilliger. 130+ attorneys ...
174	The Associated Press	No	13051963	News from The Associated Press, and a taste of the great jou...
175	Joe Palazzolo	No	10938	WSJ reporter covering legal issues. joe.palazzolo@wsj.com. ...
176	Natalie Morales	No	443991	@TODAYshow Anchor and @AccessOnline Anchor, Author, mom ...
177	Brent Kendall	No	5451	WSJ legal affairs reporter in Washington. Native Tar Heel, ...
178	Joan Biskupic	No	11021	CNN legal analyst & Supreme Court biographer; Chicago native...
179	Keith Olbermann	No	1097676	Dogs. And sports. And whales (Tom Jumbo-Grumbo on BoJack ...
180	Brian Williams	No	230947	
181	Pope Francis	No	17791867	Welcome to the official Twitter page of His Holiness Pope Fr...
182	Ezra Klein	No	2500383	Founder and editor-at-large, https://t.co/5gESirESRH. Why ...
183	Anderson Cooper	No	9967099	tweets by Anderson Cooper. Anchor @AC360 and correspondent...
184	BBC News (World)	No	24153838	News, features and analysis from the World's newsroom. Break...
185	Reince Priebus	No	935431	President @MichaelBestLaw; Exclusive Speaker @WashSpeakers; ...
186	Joe Biden	No	3111675	Represented Delaware in the Senate for 36 years, 47th Vice P...
187	Department of State	No	5149607	Welcome to the official U.S. Department of State Twitter acc...
188	Jim Miklaszewski	No	1956	Chief Pentagon Correspondent for NBC News
189	Tony Mauro	No	20310	Supreme Court correspondent, https://t.co/571ZdQnzo2 and The...
190	David Axelrod	No	1113850	Director, UChicago Institute of Politics. Senior Political ...
191	Nate Silver	No	3176243	Editor-in-Chief, @FiveThirtyEight. Author, The Signal and ...
192	George Bush	No	356042	A tribute site to the 41st President of the United States of...
193	CBS News	No	6537991	Your source for original reporting and trusted news.
194	Jonathan Karl	No	206986	ABC News Chief White House Correspondent. insta @jonkarl ...
195	BBC Breaking News	No	38539186	Breaking news alerts and updates from the BBC. For news, ...
196	Mitt Romney	No	1977201	Senator-elect from Utah.
197	ABC News	No	13985606	All the news and information you need to see, curated by the...
198	Deborah Turness	No	10389	President of NBC News International
199	The Hill	No	3162118	The Hill is the premier source for policy and political news...
200	Ann Curry	No	1536122	Journalism is an act of faith in the future.

An aPPR’s sample of 200

Table A.4: Top (selected) handles returned by aPPR. The handles with fewer than 200 followers are hidden for privacy considerations.

	Name	Followed	Followers	Description
1		Yes	198	Enroll America National Regional Director http://t.co/X6jJIE...
2	Jennifer Sizemore	Yes	386	
3	Alissa Swango	Yes	441	Director of Digital Programming at @natgeo. All things food....
4	Making a Difference	Yes	670	@NBCNightlyNews’ popular feature profiles ordinary people do...
5		No	1	
6		No	3	
7	Greg Martin	Yes	1161	Political Booking Producer at @nbcnews @todayshow
8		No	1	I am Area Man. I pwn your news feed.
9		No	2	
10	NBC Field Notes	Yes	1390	NBC News correspondents and http://t.co/1eSopOQt8s reporters...
11		No	2	
12		No	2	
13		No	1	
14		No	1	
15		No	1	
16		No	1	
17		No	3	yet another activist twitter, fighting all those fun -isms ...
18		No	4	
19		No	7	Dianne Kube is an Author with a passion, for family, holiday...
20		No	7	
21	Adam Edelman	Yes	2341	Political reporter @nbcnews. Wisconsin native, Bestchester ...
22	Phil McCausland	Yes	2519	@NBCNews Digital reporter focused on the rural-urban divide....
23	Corky Siemaszko	Yes	2538	Senior Writer at NBC News Digital (former NY Daily News rewr...
24	Sam Petulla	Yes	2588	Editor @cnnpolitics • Usually looking for datasets. You can ...
25	Ken Strickland	Yes	2693	NBC News Washington Bureau Chief
26		No	7	
27	Elyse PG	Yes	2697	White House producer @nbcnews |@USCAnnenberg alum | LA kid ...
28		No	2	Change your thoughts & you change your world. -Norman Vincen...
29		No	4	
30		No	13	
31		No	6	
32		No	154	We distribute new, never-worn clothing and merchandise to ...
33		No	10	
34	Hasani Gittens	Yes	3002	Level 29 Mage. Senior News Ed. @NBCNews. Sheriff of Nattahna...
35		No	1	
36	Scott Foster	Yes	3464	Senior Producer, Washington @NBCNEWS @TODAYshow
37		No	2	
38		No	13	
39		No	5	
40	Zach Haberman	Yes	3693	Lead Breaking News Editor, @NBCNews. Previously had other jobs...

...continued

	Name	Followed	Followers	Description
41		No	3	just like to stay in the know :) just like to stay in the ...
42		No	2	
43		No	5	
44		No	7	
45		No	1	
46	Emmanuelle Saliba	Yes	4004	Head of Social Media Strategy @Euronews | Launched #THECUBE ...
47		No	2	
48	Alex Johnson	Yes	4371	News, data and analysis for @NBCNews; data geek; non-celebri...
49		No	8	
50	Savannah Sellers	Yes	4637	News junkie. Host of NBC's "Stay Tuned" on Snapchat. Storyte...
51		No	21	
52		No	6	Anti-money laundering professional with federal law enforcem...
53		No	15	
54	Shaquille Brewster	Yes	5362	@NBCNews Producer/Politics | @HowardU Alum| Journalist | Pol...
55		No	2	Just another DIY, punk kid from the black land dirt of NEPA'...
56		No	18	Cdr Bob Mehal, Public Affairs Office, Office of the Secretar...
57		No	5	
58		No	4	
59		No	8	
60		No	10	
61		No	2	
62	Joey Scarborough	Yes	6277	NBC News Social Media Editor. New York Daily News Alum. RTs ...
63		No	5	
64		No	1	
65	Voices United	No	310	Voices United is a non profit educational organization for ...
66	Jane C. Timm	Yes	6478	@nbcnews political reporter and fact checker. More fun than ...
67	Social Headlines	No	344	Daily roundup of top social media and networking stories.
68	James Miklaszewski	No	337	Writer, Photographer, Editor, Director, Producer, Newshound ...
69		No	12	
70	Anthony Terrell	Yes	6827	Emmy Award winning journalist. Political observer. Covered ...
71		No	10	
72		No	8	
73		No	8	I'm the real Charlie Sheen. If you are a Winner, stick aroun...
74		No	9	Quotes from a nice jewish mom who's just tryna get some nice...
75		No	2	
76		No	4	
77		No	6	"Rawr!"
78	NBC News Videos	Yes	7838	The latest video from http://t.co/xPyvMOTEF6
79		No	9	
80		No	4	
81	Libby Leist	Yes	7946	Executive Producer @todayshow
82		No	8	
83		No	2	I'm running for President of the United States of America.
84		No	35	

...continued

	Name	Followed	Followers	Description
85		No	8	
86		No	2	
87		No	2	
88		No	16	
89		No	4	
90		No	5	Happy princess
91		No	1	
92		No	4	
93	Courtney Kube	Yes	9494	NBC News National Security & Military Reporter. Links and ...
94		No	5	
95		No	5	
96		No	169	
97		No	5	
98		No	2	
99	Vets Helping Heroes	No	449	Raising funds to sponsor the training of assistance dogs for...
100		No	12	
101		No	4	
102		No	8	
103	Bob Corker	Yes	10042	Serving Tennesseans in the U.S. Senate
104		No	4	
105		No	2	
106		No	11	Spécialiste développement produit et marketing des produits ...
107		No	4	
108		No	8	Not your average Grandma
109		No	29	
110		No	2	
111		No	6	
112	Kailani Koenig	Yes	11416	Producer with @MSNBC & @NBCNews. Team @MeetThePress alum. 20...
113		No	13	
114		No	14	
115	Gloria Turkin	No	204	I am honest and straight to the point. Retired Civilian Fed...
116		No	7	
117		No	28	An unconventional appreciation account for @DeadlineWH host,...
118		No	6	
119		No	10	Live like Bones
120		No	2	
121	Marianna Sotomayor	Yes	11965	Running around Capitol Hill for @NBCNews. Covers politics ...
122	NBC News THINK	Yes	12017	THINK is NBC News' home for fresh opinion, sharp analysis ...
123		No	1	
124		No	15	
125		No	2	
126		No	3	Photographer, artist, newsletter editor, designer, writer...
127		No	18	
128		No	5	

...continued

	Name	Followed	Followers	Description
129		No	5	The Quest for the Denim Jacket
130		No	9	
131		No	15	
132		No	16	Author of A Traumatic History: A Unique Look at PTSD and ...
133		No	5	
134		No	7	
135		No	5	
136		No	7	
137		No	7	
138	Beth Fouhy	Yes	13684	Senior editor, politics, NBC News and MSNBC
139	Jim Miklaszewski	Yes	14196	Jim Miklaszewski is Chief Pentagon Correspondent for NBC New...
140	Miguel Almaguer	Yes	14082	Prolific coffee drinker. Chronic under sleeper. Raging road ...
141		No	16	
142		No	4	
143		No	3	
144		No	19	The Northeast Tennessee Victory program will create a grassr...
145		No	17	
146		No	14	Just a dude with a crappy job.
147		No	5	
148	Nick Akerman	Yes	14949	Partner in the AmLaw 100 law firm of Dorsey & Whitney, Water...
149		No	5	
150		No	59	
151		No	8	
152		No	8	
153		No	4	Grad student at JHU
154		No	6	
155	Marist Poll	Yes	16030	Founded in 1978, MIPO is home to the Marist Poll and regular...
156		No	10	Sharing the best news from the e-Discovery world. Tweets by ...
157		No	7	
158		No	4	
159		No	11	Workforce and Economic Development Consultant; Employment ...
160		No	7	We're the workers of the @villagevoice, trying to get a fair...
161	Vivian Salama	Yes	16020	White House reporter for @WSJ. Formerly AP Baghdad bureau ...
162		No	8	
163		No	24	
164		No	19	I should be the real trix rabbit
165		No	4	
166		No	24	Curious food lover always looking for the best food everywhe...
167	Andrew Rafferty	Yes	16567	Senior political editor for @newsy Before that @NBCNews. And...
168		No	5	
169		No	36	
170	Tom Costello	Yes	17268	NBC News Correspondent covering Aviation, Transportation, ...
171		No	68	Wanderlust journalist ... A man is but the product of ...
172		No	6	Bibliophile, Animal lover, Realtor, Volunteer,

...continued

	Name	Followed	Followers	Description
173		No	25	
174		No	70	Director or Product Marketing @ Microsoft. My tweets. My li...
175		No	3	Experienced (and successful) grantwriter, author, wife, moth...
176		No	5	
177	Jill Lawrence	Yes	17282	Commentary editor and columnist @USATODAY. Author of The Art...
178		No	8	Howard McKinnon is Town Manager of Havana, Florida.
179		No	136	
180		No	59	
181		No	8	
182		No	12	
183		No	7	
184		No	8	
185		No	8	
186		No	15	Old and getting older.
187		No	15	
188		No	15	Married
189		No	4	
190		No	2	Director of the Essex, Connecticut Public Library aka "Your ...
191	Ethan Klapper	Yes	18292	Journalist (@YahooNews) and #avgeek.
192		No	38	
193		No	5	
194	Rebecca Sinderbrand	Yes	18691	Now: @NBCNews Senior Washington Editor, visiting lecturer ...
195		No	3	
196		No	11	Tireless trend researcher.
197		No	5	
198		No	5	
199		No	11	
200		No	3	

An rPPR's sample of 200

Table A.5: Top (selected) handles returned by rPPR. The handles with fewer than 200 followers are hidden for privacy considerations.

	Name	Followed	Followers	Description
1		Yes	198	Enroll America National Regional Director http://t.co/X6jJIE...
2	Jennifer Sizemore	Yes	386	
3	Alissa Swango	Yes	441	Director of Digital Programming at @natgeo. All things food....
4	Making a Difference	Yes	670	@NBCNightlyNews' popular feature profiles ordinary people do...
5	Greg Martin	Yes	1161	Political Booking Producer at @nbcnews @todayshow
6	NBC Field Notes	Yes	1390	NBC News correspondents and http://t.co/1eSopOQt8s reporters...
7	Adam Edelman	Yes	2341	Political reporter @nbcnews. Wisconsin native, Bestchester ...
8	Phil McCausland	Yes	2519	@NBCNews Digital reporter focused on the rural-urban divide....
9	Corky Siemaszko	Yes	2538	Senior Writer at NBC News Digital (former NY Daily News ...
10	Sam Petulla	Yes	2588	Editor @cnnpolitics • Usually looking for datasets. You can ...
11	Ken Strickland	Yes	2693	NBC News Washington Bureau Chief
12	Elyse PG	Yes	2697	White House producer @nbcnews |@USCAnnenberg alum | LA kid ...
13	Hasani Gittens	Yes	3002	Level 29 Mage. Senior News Ed. @NBCNews. Sheriff of Nattahna...
14	Scott Foster	Yes	3464	Senior Producer, Washington @NBCNEWS @TODAYshow
15	Zach Haberman	Yes	3693	Lead Breaking News Editor, @NBCNews. Previously had other jobs ...
16	Emmanuelle Saliba	Yes	4004	Head of Social Media Strategy @Euronews | Launched #THECUBE ...
17	Alex Johnson	Yes	4371	News, data and analysis for @NBCNews; data geek; non-celebri...
18	Savannah Sellers	Yes	4637	News junkie. Host of NBC's "Stay Tuned" on Snapchat. Storyte...
19		No	154	We distribute new, never-worn clothing and merchandise to ...
20	Shaquille Brewster	Yes	5362	@NBCNews Producer/Politics | @HowardU Alum| Journalist | Pol...
21	Joey Scarborough	Yes	6277	NBC News Social Media Editor. New York Daily News Alum. RTs ...
22	Jane C. Timm	Yes	6478	@nbcnews political reporter and fact checker. More fun than ...
23	Anthony Terrell	Yes	6827	Emmy Award winning journalist. Political observer. Covered ...
24	NBC News Videos	Yes	7838	The latest video from http://t.co/xPyvMOTEF6
25	Libby Leist	Yes	7946	Executive Producer @todayshow
26	Voices United	No	310	Voices United is a non profit educational organization for ...
27	Social Headlines	No	344	Daily roundup of top social media and networking stories.
28	James Miklaszewski	No	337	Writer, Photographer, Editor, Director, Producer, Newshound ...
29	Courtney Kube	Yes	9494	NBC News National Security & Military Reporter. Links and ...
30	Bob Corker	Yes	10042	Serving Tennesseans in the U.S. Senate
31	Kailani Koenig	Yes	11416	Producer with @MSNBC & @NBCNews. Team @MeetThePress alum...
32	Vets Helping Heroes	No	449	Raising funds to sponsor the training of assistance dogs for...
33	Marianna Sotomayor	Yes	11965	Running around Capitol Hill for @NBCNews. Covers politics ...
34	NBC News THINK	Yes	12017	THINK is NBC News' home for fresh opinion, sharp analysis ...
35	Beth Fouhy	Yes	13684	Senior editor, politics, NBC News and MSNBC
36	Jim Miklaszewski	Yes	14196	Jim Miklaszewski is Chief Pentagon Correspondent for NBC New...
37	Miguel Almaguer	Yes	14082	Prolific coffee drinker. Chronic under sleeper. Raging road ...
38		No	169	
39	Nick Akerman	Yes	14949	Partner in the AmLaw 100 law firm of Dorsey & Whitney, Water...
40	Marist Poll	Yes	16030	Founded in 1978, MIPO is home to the Marist Poll and regular...

...continued

	Name	Followed	Followers	Description
41	Vivian Salama	Yes	16020	White House reporter for @WSJ. Formerly AP Baghdad bureau ...
42	Andrew Rafferty	Yes	16567	Senior political editor for @newsy Before that @NBCNews. And...
43	Tom Costello	Yes	17268	NBC News Correspondent covering Aviation, Transportation, ...
44	Gloria Turkin	No	204	I am honest and straight to the point. Retired Civilian Fed...
45	Jill Lawrence	Yes	17282	Commentary editor and columnist @USATODAY. Author of The Art...
46	Ethan Klapper	Yes	18292	Journalist (@YahooNews) and #avgeek.
47	Rebecca Sinderbrand	Yes	18691	Now: @NBCNews Senior Washington Editor, visiting lecturer ...
48	Leigh Ann Caldwell	Yes	20714	NBC Capitol Hill reporter. Formerly at CNN and public radio....
49	Morgan Radford	Yes	20967	@NBCnews Correspondent: @TODAYShow/@NBCNightlyNews/...
50	GuardAnglSolPet	No	927	Supporting the Military, our Veterans and their Beloved Pets...
51	adam nagourney	Yes	25307	LA Bureau Chief for The New York Times. Story ideas welcome ...
52		No	13	
53	Micah Grimes	Yes	25948	Head of Social, @NBCNews & @MSNBC – Foreign and domestic ...
54	Perry Bacon Jr.	Yes	26853	I write about government (mostly federal, often state, occas...
55		No	21	
56	Alex Moe	Yes	28245	@NBCNews Capitol Hill Producer + Off-Air Reporter; '12 & '16...
57	Ray Farmer	No	603	NBC News staff photographer. Colorado based
58	Alex Witt	Yes	28126	Weekend host on @MSNBC (9am, noon & 1pm). Tigger's mom + ...
59	Monica Alba	Yes	30034	@NBCNews White House team. Covered Hillary Clinton on the ...
60	Jim Miklaszewski	No	1956	Chief Pentagon Correspondent for NBC News
61		No	13	
62	John McCormack	Yes	30688	Senior writer at The Weekly Standard.
63		No	136	
64	Vaughn Hillyard	Yes	31464	On the Road, Meeting Good Folk | NBC News | Arizonan | IG...
65		No	35	
66	Madelyn Monteath	No	257	NFL Network, wife, mother.. not necessarily in that order.
67	Thomas DeFrank	No	593	Veteran White House correspondent (every prez since LBJ) and...
68	Jo Ling Kent	Yes	32957	NBC News Correspondent @NBCNightlyNews, @TODAYshow...
69		No	10	
70	Carrie Dann	Yes	37119	.@NBCNews / @NBCPolitics. RTs not endorsements.
71		No	3	
72		No	7	Dianne Kube is an Author with a passion, for family, holiday...
73		No	18	Cdr Bob Mehal, Public Affairs Office, Office of the Secretar...
74		No	7	
75	Mike Memoli	Yes	39693	National Political Reporter @nbcnews; @latimes alum mike dot...
76	John Boxley	No	1201	NBC News Producer...Living life one day at a time.
77		No	15	
78	Tom Winter	Yes	40777	NBC News Investigations reporter based in New York focusing ...
79		No	7	
80	Garrett Haake	Yes	40714	Correspondent @msnbc • Taller than I look on TV • Long-suffe...
81		No	59	
82		No	70	Director or Product Marketing @ Microsoft. My tweets. My li...
83		No	68	Wanderlust journalist ... A man is but the product of ...
84		No	158	Marketing nerd at Cornerstone OnDemand.

...continued

	Name	Followed	Followers	Description
85	Jonathan Allen	Yes	44477	political reporter, @NBCNews Digital | co-author, NYT bestse...
86	NBC News First Read	Yes	53847	The first place for news and analysis from the @NBCNews Poli...
87		No	92	Smokin Meat & Raising Kids That Raise Hell. Live Every Day ...
88	Sam Singal	No	1016	Executive Producer, @nbcnightlynews
89		No	29	
90		No	59	
91	Carol Lee	Yes	51240	Reporter for NBC News, former WSJ & POLITICO, Hudson’s mom, ...
92	Alex Seitz-Wald	Yes	50168	Political reporter for @NBCNews covering Democrats | Tips, ...
93		No	28	An unconventional appreciation account for @DeadlineWH host,...
94		No	188	I am a Senior Video Producer at NBCNews.com, as well a few ...
95	HailYeah63	No	483	#RedskinsTweetTeam #HTTR
96	Eva’s Heroes	No	2067	To enrich the lives of individuals with intellectual special...
97		No	6	
98	Chi Omega	No	278	Chi Omega Chapter at CU Boulder
99	Aarne Heikkila	No	1210	Coordinating Producer for @JacobSoboroff @MSNBC & @NBCNews, ...
100	Dani	No	447	only here to talk shit & complain
101	Frank Thorp V	Yes	58152	Producer & Off-Air Reporter covering Congress at @NBCNews. ...
102	Youcef	No	228	
103		No	76	Pentagon correspondent http://t.co/Qo0w3AnYOb
104	project c.u.r.e.	No	2260	delivering donated medical supplies and equipment to develop...
105		No	117	
106		No	4	
107	Elise Jordan	Yes	58884	Co-host of @WMM_podcast podcast. @MSNBC/@NBCNews political ...
108	Patrick Burkey	No	2313	Executive Producer, @NBCNews, @MSNBC. Former EP, @NBCNightly...
109	bill hartnett	No	2500	Stripmining the internets for remarkable ephemera Social Mus...
110		No	7	
111		No	8	
112		No	16	
113		No	36	
114	Ron Fournier	Yes	64356	President: Truscott Rossman. Best-seller https://t.co/09CdTN...
115		No	12	
116	Pete Williams	Yes	70062	NBC News Justice Correspondent. Covers US Supreme Court, ...
117		No	65	Wife, Mother. Litigation Specialist. Designer. Activist for ...
118		No	10	
119	Heidi Przybyla	Yes	66489	NBC News, n’tl political reporter "Prezbella" Heidi.Przyb...
120	NBC Latino	Yes	67920	Elevating the conversation around Latino news in the United ...
121		No	189	
122		No	38	
123	Chris Jansing	Yes	72375	@msnbc Senior National Correspondent, intrepid traveler and ...
124		No	1	
125	Brent Kendall	No	5451	WSJ legal affairs reporter in Washington. Native Tar Heel, ...
126		No	2	
127	U.S. Attorney EDVA	No	5709	Led by U.S. Attorney G. Zachary Terwilliger. 130+ attorneys ...
128		No	74	Life long learner Paralegal Arts & Culture Black Community ...

...continued

	Name	Followed	Followers	Description
129		No	2	
130		No	2	
131	Tammy Fine	No	1584	Corporate Communications by day. Teen Negotiator by night...
132	Bonnie Optekman	No	2242	Digital media strategist. Voice over artist. News junkie, ...
133		No	3	yet another activist twitter, fighting all those fun -isms ...
134		No	109	Communicator through an eclectic lens of #healthcare #hospit...
135		No	88	Earth and Physical Science Teacher, Mom of 2, Self-declared ...
136	Amy Lynn-Cramer	No	1590	Mommy to 2 amazing kiddos, Wife to @tecramer AND Corporate ...
137		No	5	
138	prodjay	No	304	NBC News producer
139		No	109	
140		No	10	
141	Meghann Ludemann	No	216	Stay Tuned Associate Producer @NBCNews on @Snapchat
142		No	4	
143	Ali Vitali	Yes	78839	@NBCnews Political Reporter. Covered Trump campaign, WH + ...
144	Doug Adams	No	1902	NBC Sr. Political desk editor; Father; Baseball fan; Lover ...
145		No	99	
146	Mark Sherman	No	6336	
147	Robin Gradison	No	272	NBC News DC Deputy Bureau Chief, politics junkie, road run...
148	NBC News Signal	Yes	83715	A new streaming news channel from @NBCNews. Catch us Thursda...
149		No	45	Professor at Columbia Journaism School.
150		No	8	
151		No	18	
152	Stacey Klein	No	914	@NBCNews White House Producer, Born and raised in BalDimore ...
153		No	97	
154	Rich Latour	No	1883	From Broadcast News to Digital Storytelling. Dad of 3 Boys ...
155	Domenico Montanaro	Yes	83999	"Congress shall make no law respecting an est. of religion, ...
156		No	5	
157		No	6	Anti-money laundering professional with federal law enforcem...
158		No	24	
159		No	24	Curious food lover always looking for the best food everywhe...
160		No	25	
161		No	161	1 of 12 U.S.-led PRTs. Improving Panjshir's stability, incre...
162	Anna Matthews	No	230	
163		No	46	
164	POLITICO 45	Yes	88470	A daily diary of the 45th president of the United States.
165		No	9	Quotes from a nice jewish mom who's just tryna get some nice...
166		No	19	The Northeast Tennessee Victory program will create a grassr...
167		No	130	@NBCNews Producer in London, Links & retweets aren't endorse...
168	samgo	No	1161	Executive Producer, @MSNBC Digital
169	Megan Stark	No	263	over served Coloradan
170		No	70	
171	Katie Yu	No	484	NBC News Senior Producer / formerly @Nightline, @NBCNightlyN...
172	Mark Murray	Yes	97571	Mark Murray is the senior political editor for NBC News, as ...

...continued

	Name	Followed	Followers	Description
173	Kevin Thurm	No	1946	Chief Executive Officer @ClintonFdn. Dad, sports fan & trivi...
174		No	122	Mom, wife, grandma, Airedale Terrier lover
175		No	173	Providing conservatives with breaking news, opinion, blogs ...
176		No	14	
177		No	12	
178		No	137	Celebrate the simple loveliness of every day things, scarves...
179		No	15	
180		No	8	
181		No	16	Author of A Traumatic History: A Unique Look at PTSD and ...
182		No	9	
183		No	8	I'm the real Charlie Sheen. If you are a Winner, stick aroun...
184	David Espo	No	1308	Dad, AP Special Correspondent, Dad, Red Sox fan, Dad.
185		No	40	
186	matt toder	No	253	supervising producer, documentaries/verticals at NBC News ...
187		No	13	
188	Benjy Sarlin	Yes	100896	Political reporter for @NBCNews. I cover elections and their...
189		No	15	
190		No	29	
191		No	17	
192		No	28	Director of the Marist Poll, poll obsessed, epistemophilic,...
193		No	16	
194		No	144	Vice President, Standards @NBCNews
195		No	108	trey.daly@gmail.com
196	Daniella Mayer	No	314	DON'T forget the A. I think everything about North Korea is ...
197	Bill Hatfield	No	635	Washington news producer for NBC News TODAY; politics/histor...
198		No	19	I should be the real trix rabbit
199		No	50	
200	Phil Griffin	No	231	

B APPENDIX FOR CHAPTER 3

B.1 Technical proofs

Proof of central limit theorem and other results from Section 3.2

Proof of Proposition 3.2

Proof. Using the symmetry of P and the cyclic property of the trace, we obtain

$$\begin{aligned}\|P - \hat{X}\Gamma\hat{X}^T\|_F^2 &= \operatorname{tr}(P^2) + \operatorname{tr}(\hat{X}\Gamma^2\hat{X}^T) - 2\operatorname{tr}(P\hat{X}\Gamma\hat{X}^T) \\ &= \operatorname{tr}(P^2) + \operatorname{tr}(\Gamma^2) - 2\operatorname{tr}(\hat{X}^T P\hat{X}\Gamma).\end{aligned}$$

Taking a derivative with respect to the diagonal of Γ and setting equal to zero gives

$$\Gamma = \operatorname{diag}(\hat{X}^T P\hat{X})$$

which contains $\lambda_P(\hat{x}_1), \dots, \lambda_P(\hat{x}_q)$ down the diagonal. □

Proof of Proposition 3.3

Proof. Suppose that $\mathbb{E}(A) = U\Lambda U^T$. It follows that $\mathbb{E}(\tilde{A}) = (1-\varepsilon)U\Lambda U^T$ and $\mathbb{E}(\tilde{A}_{\text{test}}) = \varepsilon U\Lambda U^T$. This shows that they have the same eigenvectors and the simple relationship between their eigenvalues in the statement. The independence of $\tilde{A}$ and $\tilde{A}_{\text{test}}$ follows from the next lemma, often referred to as thinning (see, e.g., Durrett (2019, Section 3.7.2)).

Lemma B.1. *Define* $X \sim \operatorname{Poi}(\lambda)$ *and conditionally on* X, *define* $Y \sim \operatorname{Bin}(X, p)$ *and* $Z = X - Y$. *Unconditionally on* X, *the random variables* Y *and* Z *are independent Poisson random variables and, further,* $Y \sim \operatorname{Poi}(p\lambda)$, $Z \sim \operatorname{Poi}((1-p)\lambda)$.

To apply the lemma, take $p = \varepsilon$, define A_{ij} as X, and let Y, Z be the (i, j)-th elements of $\tilde{A}$ and $\tilde{A}_{\text{test}}$ respectively. □

Proof of Theorem 3.4

Proof. We will use Lyapunov's CLT for triangular arrays with fourth moment condition (see, e.g., Durrett (2019, Exercise 3.4.12)). Recall that B_{ij} is Poisson with mean Q_{ij}. Its mean and variance are Q_{ij} while its central fourth moment is $Q_{ij}(1+3Q_{ij}) \leqslant 4Q_{ij}$ under the assumption $Q_{ij} \leqslant 1$. Note that $\sigma^2 = \sum_{ij}(x_i x_j)^2 Q_{ij}$. To use Lyapunov's CLT, we show that the following ratio converges to zero:

$$\begin{aligned} \frac{\sum_{ij} \mathbb{E}\,|x_i x_j B_{ij} - x_i x_j Q_{ij}|^4}{\sigma^4} &\leqslant \frac{\sum_{ij}(x_i x_j)^4 (4Q_{ij})}{\sigma^4} \\ &\leqslant \frac{4\|x\|_\infty^4 \sum_{ij}(x_i x_j)^2 Q_{ij}}{\sigma^4} \\ &= \frac{4\|x\|_\infty^4}{\sigma^2} \\ &= o(1), \end{aligned} \tag{B.1}$$

where we used the bound on the fourth moment on the first line and the delocalization condition on the last line. This shows that

$$\frac{\lambda_B(x) - \lambda_Q(x)}{\sigma^2} \Rightarrow N(0,1). \tag{B.2}$$

Via Slutsky's Lemma, we can multiply the ratio in Equation (B.2) by any sequence that converges to one in probability and the result still holds. The proof is then concluded by showing that $\sigma/\hat{\sigma}$ converges to one in probability. Indeed, we have

$$\begin{aligned} \mathrm{Var}\left(\frac{\hat{\sigma}^2}{\sigma^2}\right) &= \frac{\mathrm{Var}[(x^2)^T B x^2]}{\sigma^4} \\ &= \frac{\sum_{ij}(x_i x_j)^4 Q_{ij}}{\sigma^4}, \end{aligned}$$

which is Equation (B.1) up to a factor of 4 and thus $o(1)$. So, by Chebyshev's inequality, $\hat{\sigma}^2/\sigma^2$ converges in probability to its expectation. Note that $\mathbb{E}(\hat{\sigma}^2/\sigma^2) = 1$ and that taking the inverse and the square root is continuous transformation. So, the ratio $\sigma/\hat{\sigma}$ converges in probability to one. □

Corollary B.2

The following corollary gives a sufficient condition for $\|x\|_\infty^2 = o(\sigma)$ to hold in terms of m and the expected number of edges in B.

Corollary B.2. *Using the setting of Theorem 3.4, let $\pi \in \mathbb{R}^n$ be a probability distribution on the nodes with π_i proportional to a node's expected degree. Define $\langle \pi, x^2 \rangle$ be the expected value of x_I^2 for I drawn from π and define $m = 2^{-1} \sum_i d_i$ as the expected total number of edges. If Q is positive semi-definite and*

$$\frac{\|x\|_\infty^2}{\langle \pi, x^2 \rangle} = o\left(\sqrt{m}\right),$$

then the CLT in Equation (3.5) *holds.*

Proof. The proof of Corollary B.2 follows directly from the next lemma.

Lemma B.3. *Suppose $Q \in \mathbb{R}^{n\times n}$ is positive semi-definite. Define $d = Q\mathbb{1}_n \in \mathbb{R}^n$ to be the expected degrees of the nodes $1, \dots, n$, where $\mathbb{1}_n \in \mathbb{R}^n$ is a vector of 1's. Then,*

$$\sigma^2 = (x^2)^T Q x^2 \geqslant \frac{\langle d, x^2 \rangle^2}{\sum_i d_i}.$$

Proof. Define $y = x^2, \theta = d^{1/2}, \Theta = \mathrm{diag}(\theta) \in \mathbb{R}^{n\times n}, y_\theta = \Theta y$, and $\mathscr{L} = \Theta^{-1} Q \Theta^{-1}$. Because the elements of θ are non-negative, $\mathscr{L}$ is non-negative definite.

The first part of the proof is to show that $\mathscr{L}\theta = \theta$. This is because $\Theta^{-2}Q$ is a Markov transition matrix. So,

$$\Theta^{-2} Q \mathbb{1}_n = \mathbb{1}_n \implies \Theta^{-1} Q \Theta^{-1} \Theta \mathbb{1}_n = \Theta \mathbb{1}_n$$

and this implies that $\mathscr{L}\theta = \theta$. So, by the Perron-Frobenius Theorem, θ is the leading eigenvector of $\mathscr{L}$ with eigenvalue 1.

Let $\mathscr{L}$ have eigenvectors and eigenvalues $(\phi_1, \lambda_1), \dots, (\phi_n, \lambda_n)$, where $\phi_1 = \theta/\|\theta\|_2$, $\lambda_1 = 1$ and $0 \leqslant \lambda_j \leqslant 1$ for $j \neq 1$. Then,

$$y^T Q y = y_\theta^T \mathscr{L} y_\theta = \sum_{\ell=1}^{n} \lambda_\ell \langle \phi_\ell, y_\theta \rangle^2.$$

Keeping only the first order term on the right-hand side, we have

$$y^T Q y \geqslant \lambda_1 \langle \phi_1, y_\theta \rangle^2 = \frac{\langle d, x^2 \rangle^2}{\sum_i d_i}.$$

The desired result follows. □

Applying the bound in the lemma to the delocalization condition and rearranging gives the claim. □

Proof of consistency

This section details the proof of Theorem 3.12.

Notation We use the notation $[n]$ to refer to $\{1, 2, ..., n\}$. For any real numbers $a, b \in \mathbb{R}$, we denote $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. For non-negative a_n and b_n that depend on n, we write $a_n \lesssim b_n$ to mean $a_n \leqslant C b_n$ for some constant $C > 0$, and similarly for $a_n \gtrsim b_n$. The matrix spectral norm is $\|M\| = \max_{\|x\|_2=1} \|Mx\|_2$, the matrix max-norm is $\|M\|_{\max} = \max_{i,j} |M_{ij}|$, and the matrix $2 \to \infty$ norm is $\|M\|_{2,\infty} = \max_i \|M_{i,\cdot}\|_2$.

Modified algorithm

Algorithm B.1 is used in the consistency result.

Useful concentration bounds

We will need several concentration bounds for Poisson random variables. We derive them from standard results.

We begin with a simple moment growth bound.

Lemma B.4 (Poisson moment growth)**.** *Let Z be a Poisson random variable with mean $\mu \leqslant 1$. There exists a universal constant $C > 0$ such that, for all integers $p \geqslant 2$,*

$$\mathbb{E}[|Z - \mu|^p] \leqslant C\mu \frac{p!}{2} \left(\frac{e}{2}\right)^{p-2}.$$

```
Input: Adjacency matrix $A \in \mathbb{N}^{n \times n}$, edge splitting probability
       $\varepsilon \in (0,1)$
Procedure EigCV'$(A, \varepsilon, k_{\max})$:
  1. Obtain $\tilde{A}$, $\tilde{A}_{\text{test}} \leftarrow \mathtt{ES}(A, \varepsilon)$ from splitting $A$ and set $S = \emptyset$.
   // Algorithm 3.1
  2. for $k = 2, \dots, k_{\max}$ do
     a - compute $\tilde{\lambda}_{\text{test}}(\tilde{x}_k) = \tilde{x}_k^T \tilde{A}_{\text{test}} \tilde{x}_k$ and $\tilde{\sigma}_k = \sqrt{\frac{\varepsilon}{1-\varepsilon}(\tilde{x}_k^2)^T \tilde{A} \tilde{x}_k^2}$
     b - if
           $\|\tilde{x}_k\|_\infty^2 \leqslant \min\left\{\frac{\tilde{\sigma}_k^2}{\log^2 n}, \frac{\log n}{n}\right\},$
       add $k$ to $S$ and compute
           $T_k = \frac{\tilde{\lambda}_{\text{test}}(\tilde{x}_k)}{\tilde{\sigma}_k}.$
Output: The graph dimensionality estimate:
        $\hat{K} = |\{T_k \geqslant \sqrt{n \log n} : k \in S\}|.$
```

Algorithm B.1: Modified eigenvalue cross-validation

Proof. We show that

$$\mathbb{E}[|Z - \mu|^p] \leqslant C'\mu \left(\frac{p}{2}\right)^p. \tag{B.3}$$

for some constant $C' > 0$. The claim then follows from Stirling's formula in the form

$$\sqrt{2\pi} p^{p+1/2} e^{-p} \leqslant p!, \qquad \forall p \geqslant 1.$$

By the definition of the Poisson distribution and using the fact that $0 \leqslant \mu \leqslant 1$

by assumption, we have

$$\begin{aligned}
\mathbb{E}[|Z-\mu|^{p}] &= \sum_{z\geqslant 0}|z-\mu|^{p}e^{-\mu}\frac{\mu^{z}}{z!}\\
&= |\mu|^{p}e^{-\mu}+|1-\mu|^{p}e^{-\mu}\mu+\sum_{z\geqslant 2}|z-\mu|^{p}e^{-\mu}\frac{\mu^{z}}{z!}\\
&\leqslant 2\mu+\mu^{2}e\left\{\sum_{z\geqslant 0}z^{p}\frac{e^{-1}}{z!}\right\}.
\end{aligned}$$

The term in curly brackets on the last line is the p-th moment of a Poisson random variable with mean 1, which is $\leqslant C''\left(\frac{p}{2}\right)^{p}$ for some constant $C''>0$ by Ahle (2021, Theorem 1). Equation (B.3) follows. □

The moment growth bound implies concentration for linear combinations of independent Poisson random variables.

Lemma B.5 (General Bernstein for Poisson variables)**.** *Let* $Z_1,\dots,Z_m$ *be independent Poisson random variables with respective means* $\mu_1,\dots,\mu_m\leqslant 1$*. For any* $\boldsymbol{\alpha}=(\alpha_1,\dots,\alpha_m)\in\mathbb{R}^m$ *and* $t>0$*,*

$$\mathbb{P}\left[\sum_{i=1}^{m}\alpha_i(Z_i-\mu_i)\geqslant t\right]\leqslant\exp\left(-\frac{t^2}{C'\mu_{\max}\|\boldsymbol{\alpha}\|_2^2+C''\|\boldsymbol{\alpha}\|_\infty t}\right)$$

where $\mu_{\max}=\max_i\mu_i$ *and* $C',C''>0$ *are universal constants.*

Proof. We use Boucheron et al. (2013a, Corollary 2.11). Observe that

$$\sum_{i=1}^{m}\mathbb{E}[\alpha_i(Z_i-\mu_i)^2]=\sum_{i=1}^{m}\alpha_i^2\mu_i\leqslant\mu_{\max}\|\boldsymbol{\alpha}\|_2^2.$$

Moreover, by Lemma B.4 and Stirling's formula,

$$\begin{aligned}\sum_{i=1}^{m} \mathbb{E}[\alpha_i^p (Z_i - \mu_i)_+^p] &\leqslant \sum_{i=1}^{m} \alpha_i^p C \mu_i \frac{p!}{2}\left(\frac{e}{2}\right)^{p-2} \\ &\leqslant C\mu_{\max} \|\boldsymbol{\alpha}\|_2^2 \frac{p!}{2}\left(\frac{e}{2}\|\boldsymbol{\alpha}\|_\infty\right)^{p-2} \\ &\leqslant \frac{p!}{2} \nu \left(\frac{e}{2}\|\boldsymbol{\alpha}\|_\infty\right)^{p-2},\end{aligned}$$

where we define

$$\nu := \max\{1, C\}\, \mu_{\max} \|\boldsymbol{\alpha}\|_2^2.$$

The claim then follows from Boucheron et al. (2013a, Corollary 2.11). □

The moment growth bound also implies spectral norm concentration.

Lemma B.6 (Spectral norm of Poisson graph). *Suppose $B \in \mathbb{R}^{n \times n}$ is the adjacency matrix of a Poisson graph with mean matrix Q satisfying $Q_{ij} \leqslant 1$ for all i, j. Let $q_{\max} = \max_{ij} Q_{ij}$ and assume that $nq_{\max} \geqslant c_0 \log^{\xi_0} n$ for some $\xi_0 > 2$. Then, for any $\delta > 0$, there exists a constant $C''' > 0$ such that*

$$\|B - Q\| \leqslant C''' \sqrt{nq_{\max} \log n}$$

with probability at least $1 - n^{-\delta}$.

Proof. We use Tropp (2012, Theorem 6.2). We first rewrite the matrix as a finite sum of independent symmetric random matrices,

$$B - Q = \sum_{i=1}^{n} \sum_{j=i}^{n} (B_{ij} - Q_{ij}) E^{i,j}$$

where $E^{i,j} \in \mathbb{R}^{n \times n}$ with $E^{i,j}_{ij} = E^{i,j}_{ji} = 1$ and 0 elsewhere.

Observe that, for $i \neq j$,

$$(E^{i,j})^p = \begin{cases} E^{i,i} + E^{j,j} & \text{if } p = 2, 4, \ldots \\ E^{i,j} & \text{if } p = 3, 5, \ldots \end{cases}$$

while, if $i = j$,

$$(E^{i,i})^p = E^{i,i}, \qquad p \geqslant 2.$$

Let $X^{i,j} := (B_{ij} - Q_{ij})E^{i,j}$. Then $\mathbb{E}X^{i,j} = 0$. Moreover, for $i \neq j$ and $p = 2, 4, \ldots$, we have

$$\mathbb{E}(X^{i,j})^p = \mathbb{E}(B_{ij} - Q_{ij})^p \, (E^{i,i} + E^{j,j}) \preceq Cq_{\max} \frac{p!}{2} \left(\frac{e}{2}\right)^{p-2} (E^{i,i} + E^{j,j}),$$

by Lemma B.4. Similarly, for $i \neq j$ and $p = 3, 5, \ldots$,

$$\mathbb{E}(X^{i,j})^p = \mathbb{E}(B_{ij} - Q_{ij})^p \, E^{i,j} \preceq Cq_{\max} \frac{p!}{2} \left(\frac{e}{2}\right)^{p-2} (E^{i,i} + E^{j,j}),$$

where we used the fact that the matrix $\left(\begin{smallmatrix} 1 & \alpha \\ \alpha & 1 \end{smallmatrix}\right)$ has eigenvalues $1 + \alpha, 1 - \alpha \geqslant 0$ when $|\alpha| \leqslant 1$. When $i = j$,

$$\mathbb{E}(X^{i,i})^p = \mathbb{E}(B_{ii} - Q_{ii})^p \, E^{i,i} \preceq Cq_{\max} \frac{p!}{2} \left(\frac{e}{2}\right)^{p-2} (2E^{i,i}).$$

Define

$$(\Sigma^2)^{i,j} := Cq_{\max}(E^{i,i} + E^{j,j}).$$

and

$$\sigma^2 \;=\; \left\| \sum_{i=1}^{n} \sum_{j=i}^{n} (\Sigma^2)^{i,j} \right\| = \left\| Cq_{\max} \sum_{i=1}^{n} \sum_{j=i}^{n} (E^{i,i} + E^{j,j}) \right\| \leqslant 2Cq_{\max} n,$$

where the inequality holds since $\sum_{i=1}^{n} \sum_{j=i}^{n} (E^{i,i} + E^{j,j})$ is a diagonal matrix with maximum entry $2n$. Then, by Tropp (2012, Theorem 6.2),

$$\begin{aligned} \mathbb{P}\left[\|B - Q\| \geqslant t\right] \;&=\; \mathbb{P}\left[\left\| \sum_{i=1}^{n} \sum_{j=i}^{n} X^{i,j} \right\| \geqslant t\right] \\ &\leqslant\; n \exp\left(\frac{-t^2/2}{\sigma^2 + (e/2)t}\right) \\ &\leqslant\; n \exp\left(\frac{-t^2/2}{2Cq_{\max}n + (e/2)t}\right). \end{aligned}$$

Taking $t = C''' \sqrt{nq_{\max} \log n}$ and using the fact that $nq_{\max} \geqslant c_0 \log^{\xi_0} n$, $\xi_0 > 2$, gives the result. □

Key properties of sample eigenvectors

Consider the adjacency matrix A of a Poisson graph satisfying Assumptions 3.10 and 3.11. Fixing $\varepsilon \in (0, 1)$, let $\tilde{A}$ and $\tilde{A}_{\text{test}}$ be as in Section 3.2. Let $P = \rho_n P^0 = \mathbb{E}A = \sum_{j=1}^{K} \lambda_j x_j^T x_j$ with $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_K > 0$. Let $\{\tilde{x}_l\}_{l=1}^{k_{\max}}$ be the collection of eigenvectors associated with eigenvalues $\{\tilde{\lambda}_l\}_{l=1}^{k_{\max}}$ of $\tilde{A}$. Without loss of generality, we assume $\tilde{\lambda}_1 \geqslant \tilde{\lambda}_2 \geqslant \cdots \geqslant \tilde{\lambda}_{k_{\max}}$. Define $\hat{U} = (\tilde{x}_1, \cdots, \tilde{x}_K)$ and $U = (x_1, \cdots, x_K) \in \mathbb{R}^{n \times K}$. We will need the following event:

$$\mathscr{E}^0 = \left\{ \left\| \tilde{A} - (1 - \varepsilon) P \right\| \leqslant C''' \sqrt{n\rho_n \log n} \right\}.$$

Applying Lemma B.6 with $B := \tilde{A}$ and $Q := (1 - \varepsilon)P$ shows that $\mathscr{E}^0$ holds with high probability.

Concentration of signal eigenspace First, we use a version of the Davis-Kahan theorem to show that the signal sample eigenvectors are close to the signal population eigenspace.

Lemma B.7 (Signal eigenspace)**.** *Under event $\mathscr{E}^0$, there exists an orthonormal matrix $O \in \mathbb{R}^{K \times K}$ such that, for all $k \in [K]$,*

$$\|\tilde{y}_k - x_k\|_2 = O\left(\sqrt{\frac{\log n}{n\rho_n}}\right), \qquad \|\tilde{x}_k - y_k\|_2 = O\left(\sqrt{\frac{\log n}{n\rho_n}}\right),$$

where

$$\tilde{y}_l = \left(\hat{U}O\right)_{\cdot l} = \left(\sum_{i=1}^{K} (\tilde{x}_i)_j O_{il}\right)_{j=1}^{n}, \quad y_l = (UO^T)_{\cdot l} = \left(\sum_{i=1}^{K} (x_i)_j O_{li}\right)_{j=1}^{n}.$$

Moreover, for all $k \in [K]$, $s \in [K]$, and $t \in [k_{\max}] \setminus [K]$,

$$\langle x_s, y_k \rangle = O_{ks}, \qquad \langle \tilde{x}_t, \tilde{y}_k \rangle = 0.$$

Proof. We use the variant of the Davis-Kahan theorem in Yu et al. (2015, Theorem 2). Under $\mathscr{E}^0$, $\left\|\tilde{A} - (1-\varepsilon)P\right\| = O(\sqrt{n\rho_n \log n})$. By Yu et al. (2015, Theorem 2), there exists an orthonormal matrix $O \in \mathbb{R}^{K \times K}$ such that, for all $l \in [K]$,

$$\|\tilde{y}_l - x_l\|_2 \leqslant \|\hat{U}O - U\|_F = O\left(\frac{\|\tilde{A} - (1-\varepsilon)P\|}{\lambda_K}\right) = O\left(\sqrt{\frac{\log n}{n\rho_n}}\right),$$

and

$$\|\tilde{x}_l - y_l\|_2 \leqslant \|\hat{U} - UO^T\|_F = \|(\hat{U}O - U)O^T\|_F = \|\hat{U}O - U\|_F = O\left(\sqrt{\frac{\log n}{n\rho_n}}\right),$$

where we used $\lambda_K \geqslant \psi_1^{-1}\psi_1' n\rho_n$, which holds under Assumption 3.10.

By the orthonormality of $\{x_l\}_l$ and $\{\tilde{x}_l\}_l$, we have for $s \in [K]$,

$$\langle x_s, y_k \rangle = \sum_{l=1}^{n} (x_s)_l \left(\sum_{i=1}^{K} (x_i)_l O_{ki} \right) = \sum_{i=1}^{K} \left(\sum_{l=1}^{n} (x_s)_l (x_i)_l \right) O_{ki} = O_{ks},$$

and for $t \in [k_{\max}] \setminus [K]$,

$$\langle \tilde{x}_t, \tilde{y}_k \rangle = \sum_{l=1}^{n} (\tilde{x}_t)_l \left(\sum_{i=1}^{K} (\tilde{x}_i)_l O_{ik} \right) = \sum_{i=1}^{K} O_{ik} \left(\sum_{l=1}^{n} (\tilde{x}_t)_l (\tilde{x}_i)_l \right) = \sum_{i=1}^{K} O_{ik} \mathbf{1}_{\{i=t\}} = 0.$$

□

Bounds on population quantities The previous lemma implies bounds on the population quantity of interest, $\lambda_P(\tilde{x}_l)$.

Lemma B.8 (Bounding $\lambda_P(\tilde{x}_l)$)**.** *Under event* $\mathscr{E}^0$*,*

$$\begin{aligned} \tilde{x}_l^T P \tilde{x}_l &= \Omega(n\rho_n), && \forall l \in [K], \\ \tilde{x}_l^T P \tilde{x}_l &= O(\log n), && \forall l \in [k_{\max}] \setminus [K]. \end{aligned}$$

Proof. For $s \in [K]$, expanding $\tilde{x}_s$ over an orthonormal basis including $\{x_l\}_{l \in K}$, we get

$$
\begin{aligned}
\tilde{x}_s^T P \tilde{x}_s &= \sum_{k=1}^{K} \lambda_k \langle \tilde{x}_s, x_k \rangle^2 \\
&= \sum_{k=1}^{K} \lambda_k \left[\langle x_k, y_s \rangle^2 - \langle x_k, y_s - \tilde{x}_s \rangle \langle x_k, \tilde{x}_s + y_s \rangle \right] \\
&\geqslant \sum_{k=1}^{K} \lambda_k O_{sk}^2 - \sum_{k=1}^{K} \lambda_k \| y_s - \tilde{x}_s \|_2 \| x_k \|_2^2 \left(\| \tilde{x}_s \|_2 + \| y_s \|_2 \right) \qquad \text{(B.4)} \\
&\geqslant \psi_1^{-1} \psi_1' n\rho_n - O\left(2Kn\rho_n \sqrt{\frac{\log n}{n\rho_n}} \right) \qquad \text{(B.5)} \\
&= \Omega\left(n\rho_n \right)
\end{aligned}
$$

where inequality (B.4) follows from Cauchy–Schwarz, the triangle inequality and $\langle x_k, y_s \rangle^2 = O_{sk}$ by Lemma B.7. Inequality (B.5) holds since $\sum_{k=1}^{K} O_{sk}^2 = 1$, $\psi_1^{-1}\psi_1' n\rho_n \leqslant \lambda_k \leqslant n\rho_n$ by Assumption 3.10, $\|\tilde{x}_s - y_s\|_2 = O\left(\sqrt{\frac{\log n}{n\rho_n}}\right)$ by Lemma B.7 and $\|\tilde{x}_k\|_2 = \|x_k\|_2 = \|y_s\|_2 = 1$.

For $t \in [k_{\max}] \setminus [K]$,

$$
\begin{aligned}
\tilde{x}_t^T P \tilde{x}_t &= \sum_{k=1}^{K} \lambda_k \langle \tilde{x}_t, x_k \rangle^2 \\
&= \sum_{k=1}^{K} \lambda_k \langle \tilde{x}_t, x_k - \tilde{y}_k + \tilde{y}_k \rangle^2 \\
&= \sum_{k=1}^{K} \lambda_k \left[\langle \tilde{x}_t, x_k - \tilde{y}_k \rangle + \langle \tilde{x}_t, \tilde{y}_k \rangle \right]^2 \\
&= \sum_{k=1}^{K} \lambda_k \langle \tilde{x}_t, x_k - \tilde{y}_k \rangle^2 \qquad \text{(B.6)} \\
&\leqslant K\lambda_1 \max_{k \in [K]} \| x_k - \tilde{y}_k \|_2^2 = O(\log n) \qquad \text{(B.7)}
\end{aligned}
$$

where equality (B.6) follows from $\langle \tilde{x}_t, \tilde{y}_k \rangle = 0$ by Lemma B.7. Equation (B.7) holds since $\|\tilde{y}_k - x_k\|_2 = O\left(\sqrt{\log n / n\rho_n}\right)$ by Lemma B.7 and $\lambda_k \leqslant \lambda_1 \leqslant n\rho_n$

by Assumption 3.10. □

Delocalization of signal eigenvectors To establish concentration of the estimate $\tilde{\lambda}_{\text{test}}(\tilde{x}_l)$ around $\varepsilon\lambda_P(\tilde{x}_l)$ for $l \in [K]$, we first need to show that $\tilde{x}_l$ is delocalized. That result essentially follows from an entrywise version of Lemma B.7.

Lemma B.9 (Delocalization of signal sample eigenvectors)**.** *There exist constants* $\delta_1 > 0$, $C_1 > 0$ *such that the event*

$$\mathscr{E}^1 = \left\{ \|\tilde{x}_l\|_\infty \leqslant C_1\sqrt{\frac{\mu_0}{n}}, \forall l \in [K] \right\}$$

holds with probability at least $1 - 3n^{-\delta_1}$.

Proof. We use Abbe et al. (2020, Theorem 2.1) on $\tilde{A}$, which requires four conditions. We check these conditions next. First, let $\tilde{A}^* = (1-\varepsilon)P$, $\Delta^* = \lambda_K$,

$$\kappa = \frac{\lambda_1}{\lambda_K} \leqslant \psi_1, \tag{B.8}$$

where the inequality follows from Assumption 3.10,

$$\varphi(x) = \frac{1}{32\psi_1}\min\{\sqrt{n}x, 1\},$$

and

$$\gamma = C'''\psi_1(\psi_1')^{-1}\sqrt{\frac{\log n}{n\rho_n}} \gtrsim \sqrt{\frac{\log n}{n^{1-\xi_1}}}, \tag{B.9}$$

where C''' is the constant in Lemma B.6 and $\psi_1, \psi_1' > 0$, $\xi_1 \in (0,1)$ are the constants in Assumption 3.10.

(A1) (*Incoherence*) By Abbe et al. (2020, Equation (2.4)) and the remarks that follow it, the incoherence condition is satisfied provided

$$\mu(U) := \frac{n}{K}\|U\|_{2,\infty}^2 \leqslant \frac{n\gamma^2}{K\kappa^2}.$$

Under Assumption 3.11, $\mu(U) \leqslant \mu_0$ while (B.9) implies $n\gamma^2 = \Omega(\log n)$ and (B.8) implies $\kappa = O(1)$. Hence the condition is satisfied.

(A2) (*Row and columnwise independence*) By Proposition 3.3, $\tilde{A}$ is the adjacency matrix of a Poisson graph with independent entries. In particular, $\{\tilde{A}_{ij} : i = m \text{ or } j = m\}$ are independent of $\{\tilde{A}_{ij}; i \neq m, j \neq m\}$.

(A3) (*Spectral norm concentration*) As observed previously, applying Lemma B.6 with $B := \tilde{A}$, $Q := (1-\varepsilon)P$ and $\delta > 0$ shows that the event

$$\mathscr{E}^0 = \left\{ \left\| \tilde{A} - (1-\varepsilon)P \right\| \leqslant C''' \sqrt{n\rho_n \log n} \right\},$$

holds with probability $1 - n^{-\delta}$. Moreover, by the remark after Assumption 3.10,

$$\gamma\Delta^* = C'''\psi_1(\psi_1')^{-1}\sqrt{\frac{\log n}{n\rho_n}}\lambda_K \geqslant C'''\sqrt{n\rho_n \log n}.$$

Hence,

$$\mathbb{P}\left[\left\| \tilde{A} - \tilde{A}^* \right\| \leqslant \gamma\Delta^* \right] \geqslant 1 - n^{-\delta}.$$

Note further that, under Assumption 3.10, $\gamma = o(1)$, which implies

$$32\kappa \max\{\gamma, \varphi(\gamma)\} \leqslant 32\kappa \max\left\{\gamma, \frac{1}{32\psi_1}\right\} \leqslant 1,$$

for n large enough, as required in Abbe et al. (2020, Assumption (A3)).

(A4) (*Row concentration*) As required in Abbe et al. (2020, Assumption (A4)), the function φ is continuous and non-decreasing on $\mathbb{R}_+$ with $\varphi(0) = 0$ and $\varphi(x)/x$ nonincreasing on $\mathbb{R}_+$. Let $W \in \mathbb{R}^{n\times K}$. By standard norm bounds

$$\frac{1}{\sqrt{n}} \leqslant \frac{\|W\|_F}{\sqrt{n}\|W\|_{2,\infty}} \leqslant 1.$$

As result, by definition of φ,

$$\varphi\left(\frac{\|W\|_F}{\sqrt{n}\|W\|_{2,\infty}} \right) = \frac{1}{32\psi_1}.$$

Let

$$g = \Delta^* \|W\|_{2,\infty} \varphi \left(\frac{\|W\|_F}{\sqrt{n}\|W\|_{2,\infty}} \right) = \frac{1}{32\psi_1} \lambda_K \|W\|_{2,\infty}.$$

Fix $m \in [n]$ and $r \in [K]$. Applying Lemma B.5 on $\tilde{A}_{m\cdot}$ with $\max_{ij} \mathbb{E}\tilde{A}_{ij} \leqslant (1-\varepsilon)\rho_n$, there exist $c_2 > 0$, $c_2' > 1$ such that

$$\begin{aligned} & \mathbb{P}\left(\left| \sum_{i\in[n]} (\tilde{A}_{mi} - \tilde{Q}_{mi}) W_{ir} \right| \geqslant g/\sqrt{K} \right) \\ \leqslant\ & 2\exp\left(-\frac{g^2/K}{C'(1-\varepsilon)\rho_n \|W_{\cdot r}\|_2^2 + C''\|W_{\cdot r}\|_\infty g/\sqrt{K}} \right) \\ =\ & 2\exp\left(-\frac{\lambda_K^2 \|W\|_{2,\infty}^2}{32^2\psi_1^2 K C'(1-\varepsilon)\rho_n \|W_{\cdot r}\|_2^2 + 32\psi_1\sqrt{K}C''\|W_{\cdot r}\|_\infty \lambda_K \|W\|_{2,\infty}} \right) \\ \leqslant\ & 2\exp\left(-\frac{\lambda_K^2}{32^2\psi_1^2 K C'(1-\varepsilon) n\rho_n + 32\psi_1\sqrt{K}C''\lambda_K} \right) \\ \leqslant\ & 2\exp(-c_2 n\rho_n) \\ \leqslant\ & n^{-c_2'}, \end{aligned}$$

where C' and C'' are the constants in Lemma B.5 and we used again that, by the remark after Assumption 3.10, $\lambda_K \geqslant \psi_1^{-1}\psi_1' n\rho_n$. In the final inequality, we use that $n\rho_n \geqslant c_0 \log^{\xi_0} n$, $\xi_0 > 2$ under Assumption 3.10. Since

$$\left\|(\tilde{A} - \tilde{Q})_{m\cdot} W\right\|_2 \leqslant \sqrt{K} \sup_r \left| \sum_{i\in[n]} (\tilde{A}_{mi} - \tilde{Q}_{mi}) W_{ir} \right|,$$

a union bound over r implies

$$\mathbb{P}\left[\left\|(\tilde{A} - \tilde{Q})_{m\cdot} W\right\|_2 \leqslant g \right] \geqslant 1 - Kn^{-c_2'}.$$

Applying Abbe et al. (2020, Theorem 2.1) and using Abbe et al. (2020, Equation (2.4)) again, there exists $\tilde{C} > 0$ such that

$$\begin{aligned}\max_{l\in[K]} \|\tilde{x}_l\|_\infty &\leqslant \left\|\hat{U}\right\|_{2,\infty} \\ &\leqslant \tilde{C}(2\kappa + \varphi(1))\|U\|_{2,\infty} \\ &\leqslant \tilde{C}\left(2\psi_1 + \frac{1}{32\psi_1}\right)\sqrt{K}\sqrt{\frac{\mu_0}{n}},\end{aligned}$$

with probability $1 - n^{-\delta} - 2n^{-(c_2'-1)}$, where we used Assumption 3.11 on the last line. Taking $C_1 = \tilde{C}(2\psi_1 + \frac{1}{32\psi_1})\sqrt{K}$ and $\delta_1 = \min\{\delta, c_2' - 1\} > 0$ gives the claim. □

Concentration of quadratic forms Next, we show that $\tilde{\lambda}_{\text{test}}(\tilde{x}_l)$ is concentrated around $\varepsilon\lambda_P(\tilde{x}_l)$.

Lemma B.10 (Concentration of $\tilde{\lambda}_{\text{test}}(x)$). *Let $x \in \mathbb{R}^n$ be a unit vector such that*

$$\|x\|_\infty^2 \leqslant \frac{\log n}{n}, \tag{B.10}$$

for some constant $C_2 > 0$. Then there exists $\delta_2 > 0$ such that

$$\mathbb{P}\left[\left|\sum_{i,j} x_i x_j (\tilde{A}_{\text{test}} - \varepsilon P)_{ij}\right| \leqslant \sqrt{\rho_n \log n}\right] \geqslant 1 - n^{-\delta_2}.$$

Proof. We use Lemma B.5. Using that $\|x\|_2 = 1$, we get

$$\begin{aligned}&\mathbb{P}\left[\left|\sum_{i,j} x_i x_j (\tilde{A}_{\text{test}} - \varepsilon P)_{ij}\right| \geqslant \sqrt{\rho_n \log n}\right] \\ \leqslant\ & 2\exp\left(-\frac{(\sqrt{\rho_n \log n})^2/2}{C'\varepsilon\rho_n \sum_{i,j}(x_i x_j)^2 + C'' \max_{ij}|x_i x_j|\sqrt{\rho_n \log n}}\right) \\ \leqslant\ & 2\exp\left(-\frac{\rho_n \log n/2}{C'\varepsilon\rho_n + C''\|x\|_\infty^2\sqrt{\rho_n \log n}}\right).\end{aligned}$$

By Assumption 3.10, $\rho_n \gg \frac{\log n}{n}$ while $\sqrt{\rho_n \log n} = o(1)$. By (B.10), the denominator on the last line is $\lesssim \rho_n$ and the claim follows. □

We also bound the variance estimate for the signal eigenvectors.

Lemma B.11 (Bound on the variance estimate). *Under event $\mathscr{E}^0 \cap \mathscr{E}^1$, for all $l \in [K]$*

$$\tilde{\sigma}_l^2 := \frac{\varepsilon}{1-\varepsilon}(\tilde{x}_l^2)^T \tilde{A} \tilde{x}_l^2 = \Theta(\rho_n).$$

Proof. Let $\tilde{Q} = (1-\varepsilon)P$. Under $\mathscr{E}^0$, $\tilde{\sigma}_l^2$ can be controlled through $(\tilde{x}_l^2)^T \tilde{Q} \tilde{x}_l^2$. Indeed observe that for each $l \in [K]$

$$\begin{aligned}
\left|(\tilde{x}_l^2)^T \tilde{A} \tilde{x}_l^2 - (\tilde{x}_l^2)^T \tilde{Q} \tilde{x}_l^2\right| &= |(\tilde{x}_l^2)^T (\tilde{A} - \tilde{Q}) \tilde{x}_l^2| \\
&\leqslant \left\|\tilde{A} - \tilde{Q}\right\| \|\tilde{x}_l^2\|_2^2 \\
&\leqslant \left\|\tilde{A} - \tilde{Q}\right\| \|\tilde{x}_l\|_\infty^2 \|\tilde{x}_l\|_2^2 \\
&= O\left(\sqrt{n\rho_n \log n} \cdot \frac{1}{n}\right) \\
&= O\left(\sqrt{\frac{\rho_n \log n}{n}}\right)
\end{aligned}$$

where we used that $\left\|\tilde{A} - \tilde{Q}\right\| = O(\sqrt{n\rho_n \log n})$ under event $\mathscr{E}^0$ and $\left\|\tilde{x}_l^2\right\|_\infty = \|\tilde{x}_l\|_\infty^2 = O(\frac{1}{n})$ under $\mathscr{E}^1$. Moreover, observe that $\sqrt{\rho_n \log n / n} \ll \rho_n$ since $n\rho_n \geqslant c_0 \log^{\xi_0} n$ under Assumption 3.10. So

$$\left|(\tilde{x}_l^2)^T \tilde{A} \tilde{x}_l^2 - (\tilde{x}_l^2)^T \tilde{Q} \tilde{x}_l^2\right| \ll \rho_n. \tag{B.11}$$

To get an upper bound on $\tilde{\sigma}_l^2$, note that

$$\begin{aligned}
(\tilde{x}_l^2)^T P \tilde{x}_l^2 &\leqslant \lambda_1 \|\tilde{x}_l^2\|_2^2 \\
&\leqslant \lambda_1 \cdot \|\tilde{x}_l\|_\infty^2 \cdot \|\tilde{x}_l\|_2^2 \\
&= O\left(n\rho_n \cdot \frac{1}{n} \cdot 1\right) \\
&= O(\rho_n),
\end{aligned}$$

where we used $\lambda_1 \leqslant n\rho_n$ by Assumption 3.10. Hence, we get

$$\begin{aligned}
\tilde{\sigma}_l^2 &= \frac{\varepsilon}{1-\varepsilon}(\tilde{x}_l^2)^T \tilde{A} \tilde{x}_l^2 \\
&\leqslant \frac{\varepsilon}{1-\varepsilon}|(\tilde{x}_l^2)^T \tilde{A} \tilde{x}_l^2 - (\tilde{x}_l^2)^T \tilde{Q} \tilde{x}_l^2| + \frac{\varepsilon}{1-\varepsilon}(\tilde{x}_l^2)^T \tilde{Q} \tilde{x}_l^2 \\
&\leqslant \frac{\varepsilon}{1-\varepsilon}|(\tilde{x}_l^2)^T \tilde{A} \tilde{x}_l^2 - (\tilde{x}_l^2)^T \tilde{Q} \tilde{x}_l^2| + \varepsilon(\tilde{x}_l^2)^T P \tilde{x}_l^2 \\
&= O(\rho_n),
\end{aligned}$$

by (B.11).

In the other direction, by Cauchy-Schwarz,

$$(\tilde{x}_l^2)^T P \tilde{x}_l^2 \geqslant \frac{\left(\tilde{x}_l^T P \tilde{x}_l\right)^2}{\sum_{ij} P_{ij}} \gtrsim \frac{(n\rho_n)^2}{n^2 \rho_n} \gtrsim \rho_n,$$

where the middle inequality follows from Lemma B.8. Combining with (B.11), we have

$$\begin{aligned}
\tilde{\sigma}_l^2 &= \frac{\varepsilon}{1-\varepsilon}(\tilde{x}_l^2)^T \tilde{A} \tilde{x}_l^2 \\
&\geqslant \frac{\varepsilon}{1-\varepsilon}(\tilde{x}_l^2)^T \tilde{Q} \tilde{x}_l^2 - \frac{\varepsilon}{1-\varepsilon}|(\tilde{x}_l^2)^T \tilde{A} \tilde{x}_l^2 - (\tilde{x}_l^2)^T \tilde{Q} \tilde{x}_l^2| \\
&\geqslant \varepsilon(\tilde{x}_l^2)^T P \tilde{x}_l^2 - \frac{\varepsilon}{1-\varepsilon}|(\tilde{x}_l^2)^T \tilde{A} \tilde{x}_l^2 - (\tilde{x}_l^2)^T \tilde{Q} \tilde{x}_l^2| \\
&\gtrsim \rho_n.
\end{aligned}$$

That concludes the proof. □

Proof of Theorem 3.12

Now, we are ready to prove Theorem 3.12.

Proof of Theorem 3.12. By Lemmas B.6 and B.9, the event $\mathscr{E}^0 \cap \mathscr{E}^1$ holds with probability $1 - 4n^{-\delta_1}$. Under $\mathscr{E}^0 \cap \mathscr{E}^1$, which depends only on $\tilde{A}$, the claims in Lemmas B.7, B.8 and B.11 also hold. For the rest of the proof, we condition on $\mathscr{E}^0 \cap \mathscr{E}^1$ and use the fact that $\tilde{A}_{\text{test}}$ is independent of $\tilde{A}$ by Proposition 3.3.

Let $\tilde{x}_l$, $l \in [k_{\max}]$, be the top $k_{\max}$ unit eigenvectors of $\tilde{A}$ and let

$$\tilde{\sigma}_l^2 = \frac{\varepsilon}{1-\varepsilon}(\tilde{x}_l^2)^T \tilde{A} \tilde{x}_l^2.$$

Define

$$S = \left\{ l \in [k_{\max}] : \|\tilde{x}_l\|_\infty^2 \leqslant \min\left\{ \frac{\tilde{\sigma}_l^2}{\log^2 n}, \frac{\log n}{n} \right\} \right\},$$

to be the subset of $[k_{\max}]$ corresponding to sufficiently delocalized eigenvectors. Recall that the test statistic associated to $\tilde{x}_l$ is

$$T_l = \frac{\tilde{x}_l^T \tilde{A}_{\text{test}} \tilde{x}_l}{\tilde{\sigma}_l}.$$

We say that l is rejected if

$$l \in S \quad \text{and} \quad |T_l| \geqslant \sqrt{n \log n} =: \tau_n.$$

No under-estimation We show that the test statistic associated with the K leading eigenvectors of $\tilde{A}$ will reject the null hypothesis with high probability, that is,

- $[K] \subset S$; and
- $|T_l| \geqslant \tau_n, \forall l \in [K]$.

Fix $s \in [K]$. First, we check that $s \in S$. Under $\mathscr{E}^1$, $\left\|\tilde{x}_s^2\right\|_\infty = O(1/n) \ll \log n/n$. We need to check that $\left\|\underline{\tilde{x}}_s^2\right\|_\infty \leqslant \tilde{\sigma}_s^2/\log^2 n$, for n sufficiently large. This follows from the fact that $\tilde{\sigma}_s^2 = \Theta(\rho_n)$ by Lemma B.11 and $\rho_n \geqslant c_0 n^{-1} \log^{\xi_0} n$ with $\xi_0 > 2$ under Assumption 3.10.

Next, we bound $|T_s|$ from below. We have, with probability $1 - n^{-\delta_1}$,

$$\begin{aligned}
|T_s| &= \left| \frac{\tilde{x}_s^T \tilde{A}_{\text{test}} \tilde{x}_s}{\tilde{\sigma}_s} \right| \\
&\geqslant \frac{\varepsilon \left| \tilde{x}_s^T P \tilde{x}_s \right| - \left| \tilde{x}_s^T (\tilde{A}_{\text{test}} - \varepsilon P) \tilde{x}_s \right|}{\tilde{\sigma}_s} \\
&\gtrsim \frac{\varepsilon n \rho_n - \sqrt{\rho_n \log n}}{\sqrt{\rho_n}} \\
&\gtrsim n \sqrt{\rho_n} \\
&\gg \sqrt{n \log n}, \qquad \text{(B.12)}
\end{aligned}$$

where the dominating term is controlled through $|\tilde{x}_s^T P \tilde{x}_s| \gtrsim n\rho_n \gg \log n$ by Lemma B.8, the term $|\tilde{x}_s^T (\tilde{A}_{\text{test}} - \varepsilon P) \tilde{x}_s|$ is bounded above by $\sqrt{\rho_n \log n} \ll \log n$ from Lemma B.10 and the denominator satisfies $\tilde{\sigma}_s^2 = \Theta(\rho_n)$ by Lemma B.11. The final bound follows from Assumption 3.10. By a union bound, (B.12) holds simultaneously for $s \in [K]$ with probability $1 - Kn^{-\delta_1}$.

No over-estimation Then, we show that the noise eigenvectors of $\tilde{A}$ will either be too localized or the test statistic associated with them will fail to reject the null hypothesis. In other words, we show that for any $s \in S \setminus [K]$, it holds that $|T_s| < \tau_n$ with high probability.

Let $t \in S \setminus [K]$. We bound $|T_t|$ from above as follows

$$\begin{aligned}
|T_t| &= \left| \frac{\tilde{x}_t^T \tilde{A}_{\text{test}} \tilde{x}_t}{\tilde{\sigma}_t} \right| \\
&\leqslant \frac{\varepsilon \left| \tilde{x}_t^T P \tilde{x}_t \right| + \left| \tilde{x}_t^T (\tilde{A}_{\text{test}} - \varepsilon P) \tilde{x}_t \right|}{\tilde{\sigma}_t} \qquad \text{(B.13)} \\
&= O\left(\sqrt{\frac{n}{\log^2 n}} \cdot (\log n + \sqrt{\rho_n \log n}) \right) \\
&= O\left(\sqrt{n}\right) \qquad \text{(B.14)}
\end{aligned}$$

The first term in the numerator of (B.13) satisfies $\left| x_t^T P x_t \right| = O(\log n)$ by Lemma B.8 while the term $|\tilde{x}_t^T (\tilde{A}_{\text{test}} - \varepsilon P) \tilde{x}_t|$ in (B.13) is bounded above by $\sqrt{\rho_n \log n} \ll \log n$ from Lemma B.10. For the denominator $\tilde{\sigma}_t$, $t \in S$ implies that $\|\tilde{x}_t^2\|_\infty \leqslant \tilde{\sigma}_t^2/\log^2 n$, thus

$$\tilde{\sigma}_t^2 \geqslant \log^2 n \cdot \|\tilde{x}_t^2\|_\infty \geqslant \frac{\log^2 n}{n} \cdot n\|\tilde{x}_t^2\|_\infty \geqslant \frac{\log^2 n}{n} \cdot \|\tilde{x}_t\|_2^2 = \frac{\log^2 n}{n}.$$

Consistency Therefore, it follows that the algorithm outputs $\hat{K} = K$ with probability tending to 1. □

B.2 Choosing the splitting probability

This section discusses how to select ε, the probability that an edge is placed in the testing graph. This quantity controls the number of edges used to compute the test statistics T_j. For simplicity, this section assumes the true graph dimensionality is $k = 2$ and focuses on the second test statistics T_2.

Upper bound: Reconstruction threshold

The upper bound ensures that the eigenvectors of $\tilde{A}$, $\tilde{x}_2$, could possibly estimate the eigenvector of $P = \mathbb{E}(A)$, x_2. Note that the eigenvalues of $\mathbb{E}(\tilde{A}) = (1-\varepsilon)P$ are proportional to those of P by a factor of $1-\varepsilon$. Hence, we need ε to be sufficiently small.

Two-block graphs

This subsection considers a simplified case where the graph is generated from the Stochastic Block Model (SBM). Suppose that A is generated from an SBM with two blocks, where the in-block probabilities are a/n and the out-of-block probabilities are b/n, for $a, b > 0$. Its expectation P has two non-zero eigenvalues of $\lambda_1 = (a+b)/2$ (with eigenvector $\vec{1}_n/\sqrt{n}$) and $\lambda_2 = (a-b)/2$ (with eigenvector $x_2 \in \mathbb{R}^n$ taking the values $1/\sqrt{n}$ on the nodes in the first block and $-1/\sqrt{n}$ on the nodes in the second).

Define the critical value for P as $c(P) = \frac{\lambda_2^2}{\lambda_1} = \frac{(a-b)^2}{2(a+b)}$. In fact, $c = 1$ characterizes the boundary of the reconstruction threshold of two-block graph (Mossel et al., 2015). Only above or on this reconstruction threshold (i.e. $c > 1$), $\tilde{x}_2$ could possibly recovered a signal of block membership.

Let $\tilde{P} = \mathbb{E}(\tilde{A})$ be the expectation of the splitting graph $\tilde{A}$. Since $\tilde{P} = (1-\varepsilon)P$, we have $c(\tilde{P}) = (1-\varepsilon)c(P)$. In order for the splitting graph to be above or on the reconstruction threshold, $c(\tilde{P})$ must be greater or equal to 1. That yields an upper bound on ε,

$$\varepsilon \leqslant 1 - \frac{1}{c(P)}.$$

Lower bound: Power calculation

The lower bound ensures that we put sufficiently many edges into the test graph $\tilde{A}_{test}$ in order to control the type I and type II error. For this, we perform the following power calculation. The power calculation relates μ_2, the desired level of the test α (i.e., type I error rate), and the desired power of the test $1-\beta$ (here, β is the type II error rate).

Under the alternative hypothesis, $T_2 \sim N(\mu_2, 1)$, where

$$\mu_2 = \frac{\tilde{x}_2^T(\varepsilon P)\tilde{x}_2}{\sqrt{(\tilde{x}_2^2)^T(\varepsilon P)\tilde{x}_2^2}}, \tag{B.15}$$

and $\tilde{x}_2$ is the second eigenvector of $\tilde{A}$. Note that the ε in the numerator and denominator in (B.15) follows from Proposition 3.3.

Type I error. To control the type I error rates across the k_{max} tests, we will use the Bonferroni correction. So, we will choose a cut off value t such that under the null hypothesis where $T_2 \sim N(0,1)$, $P(|T_2| > t) \leqslant \alpha/k_{max}$. To compute t, use the tail bound $P(|T_2| > t) \leqslant 2\exp(-t^2/2)$ and set the right hand side equal to α/k_{max}. This yields

$$t = \sqrt{2\log(2k_{max}/\alpha)}. \tag{B.16}$$

Type II error. Next, we will find a lower bound on μ_2 so that $P(|T_2| < t) \leqslant \beta$ under the alternative hypothesis (i.e., $T_2 \sim N(\mu_2, 1)$). We imagine that $\mu_2 > t$, it follows that

$$P(|T_2| < t) \leqslant P(T_2 < t) = P(T_2 - \mu_2 < t - \mu_2) < \exp(-(t-\mu_2)^2/2).$$

Setting the right hand side to β and solve for μ_2, we have a lower bound on μ_2,

$$\mu_2 \geqslant t + \sqrt{2\log(1/\beta)}.$$

Square both sides and substitute t by (B.16), then we want to ensure that

$$\mu_2 \geqslant \sqrt{2\log(2k_{\max}/\alpha)} + \sqrt{2\log(1/\beta)} \tag{B.17}$$

Finally, combining (B.15) and (B.17), we find a lower bound for ε,

$$\frac{\varepsilon\left(\tilde{x}_2^T P\tilde{x}_2\right)^2}{(\tilde{x}_2^2)^T P\tilde{x}_2^2} \geqslant \left(\sqrt{2\log(2k_{\max}/\alpha)} + \sqrt{2\log(1/\beta)}\right)^2.$$

Note that such lower bound is implicit because $\tilde{x}_2$ depends on ε.

Two-block graphs

We again consider the two-block graph case. Suppose that $\tilde{x}_2$ is sufficiently close to x_2. Then, $\tilde{x}_2$ is approximately orthogonal to $\vec{1}$, and the numerator in (B.15) is $\sqrt{\varepsilon}\tilde{x}_2^T P\tilde{x}_2 \gtrsim \sqrt{\varepsilon}\lambda_2(x_2^T\tilde{x}_2)^2$. In addition, the denominator is

$$\sqrt{(\tilde{x}_2^2)^T P(\tilde{x}_2^2)^T} \leqslant \sqrt{\|\tilde{x}_2^2\|_2^2\lambda_1} \approx \sqrt{\lambda_1/n}.$$

[1] Putting together, we have

$$\mu_2 \gtrsim \sqrt{c\varepsilon n}(x_2^T\tilde{x}_2)^2,$$

where $c = \lambda_2^2/\lambda_1$, and λ_1 and λ_2 are the leading two eigenvalues of P. Finally, combined with (B.17), we find a implicit lower bound on ε,

$$\sqrt{cn\varepsilon}(x_2^T\tilde{x}_2)^2 \geqslant \sqrt{2\log(2k_{\max}/\alpha)} + \sqrt{2\log(1/\beta)}.$$

[1] Since x_2 is not localized under this two-block SBM (with elements being either $1/\sqrt{n}$ or $-1/\sqrt{n}$), the approximation $\|\tilde{x}_2^2\|_2^2 \approx 1/n$ introduces a $1/n$ overhead in the denominator. For other (more general) graphs, such overhead will be different.

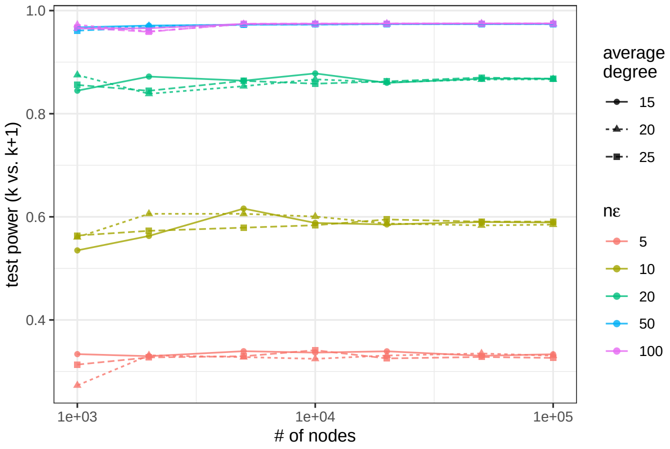


Figure B.1: Statistical power of the test on graphs simulated from the SBM with two equally sized blocks. Each graph contains n nodes (x-axis). The in-block edge probability is a/n, and the out-block edge probability is b/n, with the critical quantity $c = (a-b)^2/[2(a+b)]$ set to 2. Each dot depicts the statistical power of T_2 subtracted by that of T_3, averaged across 5 repeated experiments (y-axis). The colors indicate the value of $n\varepsilon$. The dot/line shapes indicate the value of expected average node degrees.

If $\tilde{x}_2$ is sufficiently close to x_2 (i.e., $(x_2^T \tilde{x}_2)^2 \approx 1$), then the lower bound on ε suggests that for a target statistical power and c, it suffices to ensure that $n\varepsilon$ is above some threshold. Figure B.1 shows that indeed, the testing power is determined by $n\varepsilon$ and is invariant in average node degrees, when we fix c.

Remark B.12 (Simplify Equation (B.17))**.** *Note that*

$$\begin{aligned}
&\left(\sqrt{2\log(2k_{\max}/\alpha)} + \sqrt{2\log(1/\beta)}\right)^2 \\
=\ & 2\log(2k_{\max}/\beta) + 2\log(1/\beta) + 2\sqrt{4\log(2k_{\max}/\beta)\log(1/\beta)} \\
=\ & 2\log(2k_{\max}) + 4\log(1/\beta) + 4\log(1/\beta)\sqrt{1+\log(2k_{\max})/\log(1/\beta)} \\
=\ & 2\log(2k_{\max}) + 4\log(1/\beta)(1+\sqrt{1+\log(2k_{\max})/\log(1/\beta)}).
\end{aligned}$$

We set α *and* β *to* $1/\log^2 n$. *So long as* $n \geqslant 1000$ *and* $k_{\max} \leqslant 100$, *it holds that* $1 + \sqrt{1 + \log(2k_{\max})/\log(1/\beta)} \leqslant 2.5$. *So, if*

$$\mu_2^2 \geqslant 2\log(2k_{\max}/\alpha) + 10\log(1/\beta), \tag{B.18}$$

then μ_2 *will satisfy* (B.17) *and thus be large enough for the test to have level* α *and power* $1 - \beta$. *Furthermore, substituting* $\mu_2 = \sqrt{c\varepsilon n}$ *into Equation* (B.18) *yields*

$$\varepsilon \geqslant \frac{1}{cn}\left(2\log(2k_{\max}) + 20\log(\log n)\right).$$

To have a sense for the scale of this ε, *define* $\bar{d}$ *to be the average node degree. If* ε *is set to the lower bound on the reconstruction threshold boundary (i.e.,* $c = 1$*), then the expected number of edges placed into the the test set is* $n\,\varepsilon\,\bar{d} = \bar{d}\,(2\log(2k_{\max}) + 20\log(\log n))$. *This is a very small quantity, thus revealing the power of the test statistics* T_2. *For a fixed* $\bar{d}$, *it grows with* n *at the rate* $\log\log n$ *and it grows with* $k_{\max}$ *at the rate* $\log k_{\max}$. *With* $\bar{d}$, *it grows linearly.*

B.3 Supporting figures and tables

In Figure 3.1, we evaluated the accuracy of each method when requiring the exact recovery of k. In order to illustrate how each method either under-estimates or over-estimates k, Figure B.2 displays the results in Figure 3.1 by the relative error for each estimate $\hat{k}$, which is defined as

$$\text{relative error} = \frac{\hat{k} - k^*}{k^*},$$

where $k^* = 10$ is the true k. From the simulation results, we observed that most methods under-estimate k when the average degree of the graph is smaller (i.e., sparser), except for StGoF which over-estimates it. In addition, from the standard deviation of the relative error, we observe that EigCV provides a more accurate and less variable estimation of k as the graph sparsity varies.

In Section 3.4, we removed the 14 small departments that consist of less than 10 members. Among these, two departments have only one members,

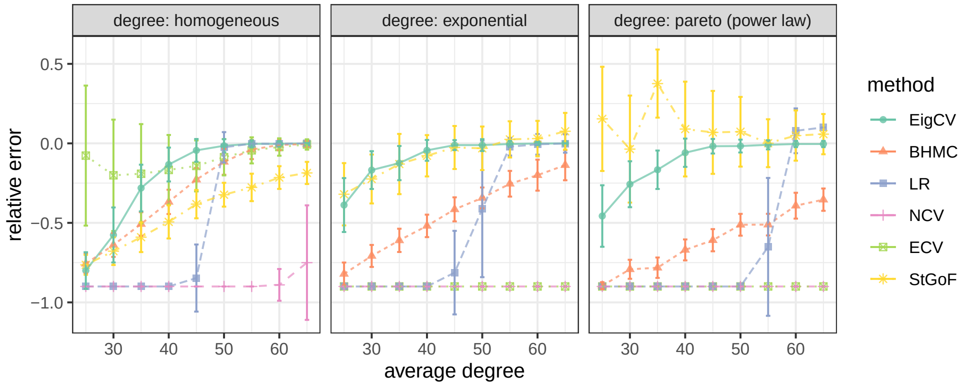


Figure B.2: Comparison of relative error for different graph dimensionality estimators under the DCSBM. The panel strips on the top indicate the node degree distribution used. Within each panel, each colored line depicts the relative error of each estimation method as the average node degree increases. Each point on the lines are averaged across 100 repeated experiments. For each point, an error bar indicates the sample standard deviation of relative errors.

Table B.1: Comparison of graph dimensionality estimates using the email network among members in a large European research institution. Each members belongs to one of 42 departments.

Method	Estimate (mean)	Runtime (second)
EigCV	30.56	0.81
BHMC	14.00	0.04
LR	13.00	128.17
NCV	6.96	271.15
ECV	20.08	60.13
StGoF	> 50	544.66

and eight departments have less than five members. Table B.1 compares six methods using this email network without filtering. We observed similarly that EigCV provided a closer estimate of k than other methods.

C.1 Technical proofs

Proof. **of Proposition 4.2** We show that for any fixed Z and Y, the inequality holds for the minimization over B on the left-hand-side and the diagonal D on the right-hand-side,

$$\min_{B} \left\|X - ZBY^T\right\|_F^2 \leqslant \min_{D} \left\|X - ZDY^T\right\|_F^2.$$

In fact, the maximizer of the left-hand-side is $B^* = \left(Z^TZ\right)^{-1} Z^TXY \left(Y^TY\right)^{-1}$ if Z and Y are full-rank, or $B^* = \left(Z^TZ\right)^{+} Z^TXY \left(Y^TY\right)^{+}$ if either Z or Y is singular, where A^+ is the Moore–Penrose inverse of matrix A. Since B^* is not diagonal in general, the inequality follows. □

Proof. **of Lemma 4.4** We rewrite the objective function:

$$\begin{aligned} \left\|X - ZBY^T\right\|_F^2 &= \operatorname{tr}\left[\left(X - ZBY^T\right)^T \left(X - ZBY^T\right)\right] \\ &= \|X\|_F^2 - 2\operatorname{tr}\left(X^TZBY^T\right) + \operatorname{tr}\left(B^TB\right) \\ &= \|X\|_F^2 - \operatorname{tr}\left[B^T\left(2Z^TXY - B\right)\right]. \end{aligned}$$

For fixed Z and Y, take the derivative of B and set it to zero. We have the optimizer $B^* = Z^TXY$ and the squared optimal value is $\|X\|_F^2 - \left\|Z^TXY\right\|_F^2$. Recognizing that $\|X\|_F^2$ is determined, the desired formulation (4.13) follows. □

Remark C.1 (Minimal matrix reconstruction error of PMD)**.** *If* B *is constrained to a diagonal matrix in* (4.12)*, then the squared minimal value is* $\|X\|_F^2 - \sum_{i=1}^k d_i^2$*, where* $d_i = \left[Z^TXY\right]_{ii}$ *for* $i = 1, 2, ..., k$.

Proof. From the proof of Lemma 4.4, we have

$$\left\|X - ZDY^T\right\|_F^2 = \|X\|_F^2 - \operatorname{tr}\left[D^T\left(2Z^TXY - D\right)\right].$$

Then, take the derivative of D and set it to zero. This yields the solution $\hat{D} = \mathrm{diag}(d_i)$, where $d_i = \left[U^T X V\right]_{ii}$. Finally, plugging-in the maximizer $\hat{D}$ gives the claimed optimal value. Note that $\sum_{i=1}^{k} d_i^2 \leqslant \left\|U^T X V\right\|_F^2$. □

Proof. **of Lemma 4.5** Suppose the low-rank SVD of $C \in \mathbb{R}^{p \times k}$ is UDV^T, where $U \in \mathscr{V}(p, k)$, $V \in \mathscr{U}(k)$, and $D \in \mathbb{R}^{k \times k}$ is diagonal. Then,

$$\left\|C^T X\right\|_F^2 = \mathrm{tr}\left(X^T C C^T X\right) = \mathrm{tr}\left(X^T U D^2 U^T X\right).$$

The trace quadratic form is maximized at $X^* = UR$, for any orthogonal matrix $R \in \mathscr{U}(k)$. In particular, when $R = V$, $X^* = \mathrm{polar}(C)$. □

C.2 Choosing the sparsity parameter

The sparsity controlling parameters in SCA and SMA—γ, γ_y, and γ_z—are meaningful if they take values from a certain range, depending on the choice of ℓ_p-norm constraint. In this section, we discuss the sparsity constraint on Y; the constraint on Z is similar. First, consider the ℓ_1-norm constraint $\|Y\|_1 \leqslant \gamma$. The sparsity parameter should satisfy $k \leqslant \gamma \leqslant k\sqrt{p}$. This is for the set $\{Y \in \mathbb{R}^{p \times k} \mid \|Y\|_1 = \gamma\}$ to intersect with $\mathscr{V}(p, k)$. On the right hand side, if $\gamma > k\sqrt{p}$, any element in $\mathscr{V}(p, k)$ satisfies $\|Y\|_1 < \gamma$, so the sparsity constraint is ineffective (Figure C.1: left panel). On the left hand side, if $\gamma < k$, none of the elements in $\mathscr{V}(p, k)$ satisfies $\|Y\|_1 \leqslant \gamma$, so the solution to (4.4) does not fall on $\mathscr{V}(p, k)$. Similarly, for the $\ell_{4/3}$-norm sparsity constraint $\|Y\|_{4/3} \leqslant \gamma$, the sparsity controlling parameter should take value within $k^{3/4} \leqslant \gamma \leqslant p^{1/4} k^{3/4}$ (Figure C.1: right panel).

In Algorithm 4.2, the sparsity parameter is optional. If absent, the algorithm uses a default value of $\gamma = \sqrt{pk}$ (or $\gamma_z = \sqrt{nk}$ and $\gamma_y = \sqrt{pk}$ in SMA). This is supported by our simulation results showing that the SCA algorithm is robust against various choices of γ (Section 4.4). In addition, we observed that the default settings generally yielded meaningful estimates in real data applications.

The sparsity parameter can also be tuned based on the data. We provide a schema for cross-validate the parameters of SCA and SMA (e.g., the approxi-

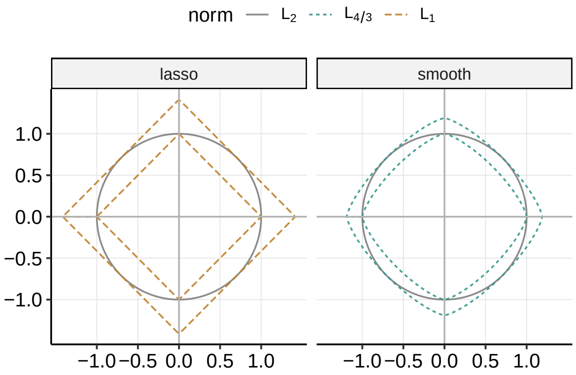


Figure C.1: Comparison of the ℓ_p norms. Left (lasso): Two ℓ_1-norm contours (brown) of 1 and $\sqrt{2}$ and the ℓ_2-norm contour (grey) of 1. Right (smooth): $\ell_{4/3}$-norm contours (green) of 1 and $2^{1/4}$ and the ℓ_2-norm contour (grey) of 1.

mation rank k and the sparsity parameter γ). To assess a candidate parameter, we adapt a K-fold cross-validation framework (K often takes the value 10) as previously introduced by Wold (1978):

Step 1) Given the input data $X \in \mathbb{R}^{n \times p}$, we first construct K leave-out data matrices $X^{(1)}, X^{(2)}, ..., X^{(K)} \in \mathbb{R}^{n \times p}$, each of which has one-Kth disjoint portion of elements being randomly sampled and removed (i.e., set to zero). Let $C^{(k)}$ collects the indices of those left-out elements in $X^{(k)}$, for $k = 1, 2, ..., K$.

Step 2) Next, we apply SCA (or the SMA) to every new matrix $X^{(k)}$ with the candidate tuning parameters and obtain its low-rank approximation $\hat{X}^{(k)}$. That is, for SCA, $\hat{X}^{(k)} = X^{(K)}\hat{Y}^{(k)}[\hat{Y}^{(k)}]^T$, and for SMA, $\hat{X}^{(k)} = \hat{Z}^{(k)}\hat{B}^{(k)}[\hat{Y}^{(k)}]^T$

Step 3) Finally, calculate the mean square error (MSE) of $\hat{X}^{(k)}$ over those left-out elements $C^{(k)}$, defined as

$$\mathrm{MSE}(k) = \sum_{(i,j) \in C^{(k)}} \left(\hat{X}_{ij}^{(k)} - X_{ij}\right)^2, k = 1, 2, ..., K.$$

> We then evaluate the “goodness” of a candidate parameter by the average MSE across K leave-out data matrices.

Upon the construction of leave-out data matrices, the left-out elements are randomly sampled; this typically removes scattered entries of X, rather than trunks of adjacent ones. For example, if X is the adjacency matrix of a graph, then this procedure is akin to the edge cross-validation studied by Li et al. (2020). Setting the left-out elements to zero eliminates all terms in $\left\|Z^TXY\right\|_F$ that related to them. Our low-rank estimation for the missing entries is closely related to the SVD-based methods in data imputation literature (Troyanskaya et al., 2001).

C.3 Properties of soft-thresholding

In the PRS update, the last step uses a shrinkage operator to project the rotated matrices onto the feasible set. Shrinkage operators are widely used for creating sparse structure, as it is easy to implement. The threshold value t can be found in $\mathcal{O}(\log_2(1/\varepsilon))$ time through a binary search, where ε is the convergence tolerance.

For the ℓ_1-norm constraint (or penalty), we show that a soft-thresholding shrinkage is “appropriate.” Let $Y \in \mathscr{V}(p,k)$ and $\hat{Y} \in \mathscr{B}(p,k)$ be the two matrices before and after a shrinkage operation respectively. A direct calculation shows that given a constraint $\|\hat{Y}\|_1 \leqslant \gamma$, the soft-thresholding shrinkage, $\hat{Y} = T_\gamma(Y)$, minimizes $\|\hat{Y} - Y\|_F$. After the shrinkage, the objective value in (4.4) (i.e., explained variance) decreases by at most $\|Z^TX\|_F\|\hat{Y} - Y\|_F$. Note that we update Y fixing Z (and X).

We provide theoretical properties for the soft-thresholding, regarding preservation of orthogonality and the explained variance. Let $Y \in \mathscr{V}(p,k)$ and let $\hat{Y} = T_\gamma(Y)$ be the result of soft-thresholding Y as defined in (4.9).

First, we denote the included angles between any two columns of $\hat{Y}$ and Y as θ_{ij}, for $i,j = 1,2,...,k$. When it is clear, we also write θ_{ii} as θ_i for simplicity. We define the *deviation* between $\hat{Y}$ and Y as $\sum_{i=1}^k \sin^2(\theta_i)$. The following proposition bounds the sum of deviations.

Proposition C.2 (Deviation due to soft-thresholding)**.** *If* t *is sufficiently small, then*

$$\sum_{j=1}^{k} \sin^2(\theta_j) \leqslant \left\|\hat{Y} - Y\right\|_F^2.$$

Proof. Let $\hat{y}_i$ and y_i be the ith column of $\hat{Y}$ and Y respectively. For the included angle θ_i,

$$\begin{aligned} \cos(\theta_i) &= \hat{y}_i^T y_i / \|\hat{y}\|_2 \\ &= \|\hat{y}_i\|_2 + \hat{y}_i^T(y_i - \hat{y}_i)/\|\hat{y}\|_2 \\ &> \|\hat{y}_i\|_2. \end{aligned}$$

The last inequality results from the definition of soft-thresholding. Then, by the Pythagorean trigonometric identity, we have

$$\begin{aligned} \sin^2(\theta_i) &= 1 - \cos^2(\theta_i) \\ &< 1 - \|\hat{y}_i\|_2^2 \\ &\leqslant \|\hat{y}_i - y_i\|_2^2. \end{aligned}$$

The last inequality is due to the triangular inequality. Finally, summing over the columns yields the desired result. □

Proposition C.2 controls the deviation with the Frobenius norm of $Y - \hat{Y}$. Since the columns of Y are mutually orthogonal, for any two columns of $\hat{Y}$, we have

$$\left|\hat{y}_i^T \hat{y}_j\right| \leqslant \sin\left(\theta_j + \theta_l\right) \|\hat{y}_i\|_2 \|\hat{y}_j\|_2$$

assuming $\theta_i + \theta_j \leqslant \pi/2$. Hence, a small deviation indicates that the orthogonality of $\hat{Y}$ is conserved after soft-thresholding.

Next, we investigate the change in explained variation due to soft-thresholding. Define the explained variance (EV) of a data matrix X by the loading matrix Y as $\mathrm{EV}(Y) = \|XY\|_F$. The following proposition bounds the EV for $\hat{Y}$ and is due to the Theorem 13 in Hu et al. (2016).

Proposition C.3 (Explained variance after soft-thresholding)**.** *If for all* $1 \leqslant i \leqslant k$, $\theta_i = \theta$ *and* $\sum_{j=1}^{k} \cos(\theta_{ij}) \leqslant 1$, *then*

$$\left(\cos^2\theta - \sqrt{k-1}\sin 2\theta\right) \mathrm{EV}(Y) \leqslant \mathrm{EV}(\hat{Y})$$

for any data matrix X.

Proposition C.3 implies that if the deviation between Y and $\hat{Y}$ is small, then the EV of $\hat{Y}$ is close to that of Y,

$$\left(\cos^2\theta - \mathcal{O}(\theta)\right) \mathrm{EV}(Y) \leqslant \mathrm{EV}(\hat{Y}).$$

C.4 Independent component analysis

In this section, we demonstrate the connection between sparse PCA (specifically, our SCA formulation) and independent component analysis (ICA).

ICA is motivated by blind-source (or blind-signal) separation in signal processing (see, e.g., Georgiev et al., 2005; Comon and Jutten, 2010), where we observe a series of multivariate signals $X_{i\cdot} \in \mathbb{R}^p$ for $i = 1, 2, ..., n$, where n is the number of observations. In ICA, there exist k independent, non-Gaussian and unobserved *source* signals underlying each observation, $Z_{i\cdot} \in \mathbb{R}^k$ for $i = 1, 2, ..., n$, and each observation is a linear mixture of these source signals, this is, $X = ZM^T$ (or $X_{i\cdot} = Z_{i\cdot}M$ for $i = 1, 2, ..., n$), where $M \in \mathbb{R}^{p \times k}$ is the *mixing* matrix. ICA aims to "un-mix" the observed X and extract Z from it. In particular, since the k source signals are independent, it is often assumed that Z's columns have unit length and are orthogonal to each other (i.e., $Z \in \mathscr{V}(n, k)$). The ICA literature is rich in theoretical results (Hyvärinen and Oja, 2000; Chen and Bickel, 2006; Samworth and Yuan, 2012; Miettinen et al., 2015), and most methods for ICA (e.g. fastICA) identifies both platykurtic- and leptokurtic-sourced signals.

We consider a sparse version of ICA, sparse ICA, where Z is sparse (or the columns of Z follow leptokurtic distributions). We show that sparse ICA and sparse PCA are unified by the SMA. To see this, recall from Section 4.2 that the SMA of a data matrix is ZBY^T, where Z and Y are both sparse but B. We interpret the SMA for the two modern multivariate data analysis:

Sparse PCA For sparse PCA, we treat Y as the sparse loadings, and ZB together as the component scores.

Sparse ICA For sparse ICA, the sparse source signals (or the independent components) are the columns of Z, the mixing matrix is BY^T.

It can be seen that both sparse PCA and sparse ICA seek a sparse component in the data: sparse PCA extracts them for the column space (Y), while ICA the row space (Z). Hence, performing sparse PCA to the transposed input data matrix actually accomplishes sparse ICA to the original data. This highlights the similarities between sparse PCA and sparse ICA.

Example: Blind source separation with SCA

We apply SCA to the blind source separation of image data (Comon and Jutten, 2010). For example, suppose the source signals are individual images, and a sensor senses several mixed images, each an linear mixture of the sources. The objective is then to identify the source images from the observed ones (i.e., to decipher the linear coefficients).

We selected three 512 × 512-pixels pictures of diverse genres from the internet (Figure C.2, the first row). The sample excess kurtosis of the images are 1.53, 3.32, and -0.45 respectively. Next, we generated three $(n = 3)$ mixtures of the original images, with the linear coefficients randomly drawn from the uniform distribution, Unif(0,1). The three mixed images are displayed in the second row of Figure C.2. For sparse PCA, we vectorize the mixed images (that is 512^2-pixels) and put them in a shallow matrix $X \in \mathbb{R}^{n \times p}$, where $p = 262,144$. This matrix is then input to SCA (Algorithm 4.2) for three sparse PCs $(k = 3)$, with the sparsity parameter γ set to $\sqrt{nk}$. The resulting sparse loadings $Y \in \mathbb{R}^{p \times k}$ contains the three separated source images and the scores $S \in \mathbb{R}^{n \times k}$ decodes the mixing coefficients. The third row in Figure C.2 displays the three separated images (i.e., the three rows of Y.) The clean-cut

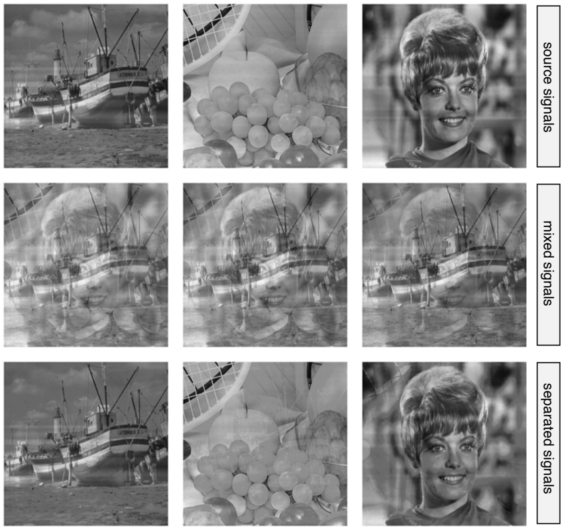


Figure C.2: Blind image signal separation using SCA. The three panel rows display three source images, three linear mixtures of the source images, and the three separated images using SCA.

identification of the source images suggests that sparse PCA is capable of extracting sparse and independent components from the data.

Algorithmic comparisons

Another insight for sparse PCA and sparse ICA can be gleaned from their algorithms. In this section, we demonstrate that the fastICA algorithm (Hyvarinen, 1999) and our SCA algorithm are both closely related to kurtosis (Mardia, 1970).

The fastICA algorithm finds Z in two steps. The first step is to pre-process X. The pre-processing of centering and whitening (see, e.g., Comon (1994)) results in the leading k left singular vectors $\hat{U} \in \mathscr{V}(n, k)$. The second steps searches for an orthogonal rotation that maximize the non-gaussianity of $\hat{U}R$, as measured by the approximation of negentropy,

$$\underset{R}{\text{maximize}} \quad \textstyle\sum_{j=1}^{k} \left\{ G([\hat{U}R]_{\cdot j}) - G(\nu) \right\}^2 \quad \text{subject to } R \in \mathscr{U}(k), \quad \text{(C.1)}$$

where $G(x)$ is a non-quadratic function for $x \in \mathbb{R}^n$, and $\nu \sim N(0, I_n)$ is the multivariate standard Gaussian vector. Finally, $\hat{U}\hat{R}$ is the fastICA estimate for Z, where $\hat{R}$ is the solution to (C.1). Hyvarinen (1999) noted that setting $G(x) = \|x\|_4^4/n$, the optimization in (C.1) takes the form[1]

$$\underset{R}{\text{maximize}} \quad \textstyle\sum_{j=1}^{k} \text{kurt}^2([UR]_{\cdot j}) \quad \text{subject to } R \in \mathscr{U}(k), \tag{C.2}$$

where kurt(x) is the sample excess kurtosis of $x \in \mathbb{R}^n$ and is defined as kurt(x) $= n \sum_{i=1}^{n} (x_i - \bar{x})^4 / \left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)^2 - 3$, where $\bar{x} = \sum_{i=1}^{n} x_i/n$ is the mean. It can be seen from (C.2) that fastICA produces either leptokurtic (kurt(x) > 0) or platykurtic (kurt(x) < 0) estimation for the columns of Z, because of the squared kurtosis in the objective function. This primarily explains that fastICA allows both platykurtic- and leptokurtic-sourced signals.

As for SCA, the algorithm uses the varimax rotation to find the orthogonal rotation. Suppose $Y \in \mathscr{V}(n, k)$. Since the sum of squares of Y's columns are constant, $\sum_{j=1}^{k} Y_{ij}^2 = 1$, maximizing the varimax rotation is equivalent to maximizing the sum of sample kurtosis of Y's columns,

$$C_{\text{varimax}}(Y) = \sum_{j=1}^{k} \text{kurt}(Y_{\cdot j}) + \text{constant}.$$

This suggests that the varimax rotation in SCA promotes some leptokurtic columns in the loading Y of sparse PCs. Note that any sparse distribution is leptokurtic (see Theorem 2.1 of Rohe and Zeng (2020)). Hence, SCA generates specifically sparse PCs.

In many applications of ICA, the number of independent components and the number of observed variables are the same (i.e., $p = k$), in which case, the mixing matrix is square. The $p = k$ regime is generally challenging. As such, many theoretical results presume no or very little noise in X, in order for estimating guarantees. By contrast, sparse PCA typically presumes the data to comprise noise and the statistical model usually contain a noise term. In addition, it is showed that sparse PCA is consistent even when the observed

[1] The authors also suggested different forms of $G(x)$.

data is high-dimensional (i.e., p grows at the same rate as n) or sparse by itself (i.e. contains many zeros) (Rohe and Zeng, 2020), while it is unclear yet whether ICA is consistent or not under these settings.

D APPENDIX FOR CHAPTER 5

D.1 Technical proofs

Proof of Proposition 5.1

Proof. For (a), when the teleportation constant is zero, the PPR vector becomes the stationary probability distribution of a standard random walk (Lovász, 1993), which is proportional to the node degree distribution d^*. The second part of (a) is intuitive, recognizing that the teleportation introduces a particular favor of the seed node.

We prove (b) by constructing an explicit form of the PPR vector. Let $R_\tau = \tau \sum_{s=0}^{\infty} (1-\tau)^s P^s$. The infinite sum converges for $\tau \in (0, 1]$. Then,

$$\begin{aligned} \tau\pi^T + (1-\tau)\pi^T R_\tau P &= \tau\pi^T + (1-\tau)\pi^T \left(\tau \sum_{s=0}^{\infty} (1-\tau)^s P^s \right) P \\ &= \tau\pi^T + \tau \sum_{s=1}^{\infty} (1-\tau)^s \pi^T P^s \\ &= \pi^T R_\tau. \end{aligned}$$

Hence, $p = R_\tau^T \pi$ satisfies the definition of personalized PageRank vector. □

Proof of Lemma 5.3

Proof. For part (a), since $X_{iw} \mid Z, \Phi \overset{\text{ind.}}{\sim} \text{Bernoulli}(\lambda \Phi_{z(i)w})$, it follows from the law of expectation that

$$\mathscr{A} = \mathbb{E}\left(\mathbb{E}\left(X^T X \mid Z\right) \mid \alpha, \lambda, \Phi\right) = \Phi^T \left(\lambda^2\, \mathbb{E}\left(Z^T Z \mid \alpha\right)\right) \Phi = \Phi^T \mathbf{D} \Phi,$$

where $\mathbf{D} = m\lambda^2 \operatorname{diag}(\alpha)$. Then,

$$\mathscr{d}_w = \sum_{v=1}^{n} \mathscr{A}_{wv} = \sum_{t=1}^{k} \mathbf{d}_t \Phi_{tw} = m\lambda^2 \sum_{t=1}^{k} \alpha_t \Phi_{tw},$$

and

$$\mathscr{P} = \mathscr{D}^{-1}\Phi^{\mathrm{T}}\mathbf{D}\Phi = \Gamma^{\mathrm{T}}\Phi,$$

where $\Gamma = \mathbf{D}\Phi\mathscr{D}^{-1}$. For part (b), since $\mathbf{A} = \Gamma\mathscr{D}\Gamma^{\mathrm{T}}$ by definition,

$$\mathbf{P} = \mathbf{D}^{-1}\mathbf{A} = \mathbf{D}^{-1}\Gamma\mathscr{D}\Gamma^{\mathrm{T}} = \Phi\Gamma^{\mathrm{T}}.$$

Remark. It can be seen that $\mathscr{P}^2 = \Gamma^{\mathrm{T}}(\Phi\Gamma^{\mathrm{T}})\Phi = \Gamma^{\mathrm{T}}\mathbf{P}\Phi$. □

Proof of Lemma 5.4

Proof. From Lemma 5.3, we have $\mathscr{P}^s = \Gamma^{\mathrm{T}}\mathbf{P}^{s-1}\Phi$ for any integer $s > 1$. Then, from Proposition 5.1(b), we have

$$\begin{aligned} \mathscr{p}^{\mathrm{T}} &= \sum_{s=0}^{\infty} \tau(1-\tau)^s \pi^{\mathrm{T}} \mathscr{P}^s \\ &= \tau\pi^{\mathrm{T}} + (1-\tau)\left(\sum_{s=0}^{\infty} \tau(1-\tau)^s (\Gamma\pi)^{\mathrm{T}} \mathbf{P}^s\right)\Phi \\ &= \tau\pi^{\mathrm{T}} + (1-\tau)\mathbf{p}^{\mathrm{T}}\Phi. \end{aligned}$$

In the second equation, $\Gamma\pi = (1, 0, \cdots, 0)$ by the assumption. □

Proof of Theorem 5.5

The proof of Theorem 5.5 follows directly from Theorem 1 of Chen et al. (2020a), recognizing that the average expected node degrees δ is

$$\frac{1}{n}\sum_{w=1}^{n} \mathscr{d}_w = \frac{m\lambda^2}{n}.$$

Proof of Corollary 5.6

Proof. For frequency, the algorithm ranks words by $x = p - \tau\pi$. It suffices to show that with high probability, $x_w > x_v$ for two words $w \in \mathcal{S}$ and $v \notin \mathcal{S}$ that are not seed keywords. To this end, we apply triangle inequality and get

$$\begin{aligned}\frac{x_w - x_v}{\|\mathscr{x}\|_\infty} &\geqslant \frac{p_w - p_v}{\|\mathscr{p}\|_\infty} \\ &\geqslant \frac{\mathscr{p}_w - \mathscr{p}_v}{\|\mathscr{p}\|_\infty} - \frac{|p_w - \mathscr{p}_w|}{\|\mathscr{p}\|_\infty} - \frac{|p_v - \mathscr{p}_v|}{\|\mathscr{p}\|_\infty} \\ &\geqslant \Delta_w - \frac{2\|p - \mathscr{p}\|_\infty}{\|\mathscr{p}\|_\infty}.\end{aligned}$$

Since $\Delta_w \leqslant 1$, assumption (5.8) contains the condition $m\lambda^2/n > c_0 \log n$ in Theorem 5.5, which implies that

$$\frac{\|p - \mathscr{p}\|_\infty}{\|\mathscr{p}\|_\infty} < \frac{\Delta_w}{2},$$

if $m\lambda^2\Delta_w^2/n \log n$ is large enough. These collectively imply $x_w > x_v$ as desired.

□

E.1 Supporting materials and methods

Targeted sampling with Personalized PageRank

We performed a target sampling from the Twitter friendship network with personalized PageRank (PPR) (Chen et al., 2020a). Consider the Twitter friendship network (or graph) of n accounts, $G = (V, E)$, where $V = \{1, 2, ..., n\}$ is the list of accounts, and E is the edge list representing friendships (or followings). For example, if account u follows account v, then $(u, v) \in E$. Note that G is a directed graph, that is, $(u, v) \in E$ does not imply that $(v, u) \in E$. Graph G can be represented by an adjacency matrix, $A \in \{0, 1\}^{n \times n}$, where $A_{uv} = 1$ if $(u, v) \in E$ and 0 otherwise. For account u, define $d_u = \sum_{v \in V} A_{uv}$ to be the number of u's friends (or followings). We consider those accounts that have at least one friend (i.e., $d_u > 0$).

The PPR vector $x \in \mathbb{R}^n$ is defined as the stationary distribution of which we call a *personalized random walk* (Page et al., 1998) on G. The random walk starts at a seed account $u \in V$ of interest. At each step, the walker returns to u with probability τ, called the *teleportation constant*, and with probability $1 - \tau$, the random walker randomly goes to a friend of the current account. For any account v, its corresponding value in the PPR vector, x_v, is the probability that the random walker lands on it. x_v quantifies the "closeness" of account v to the seed account (the larger the closer). We used the PPR vector to determine the inclusion of Twitter accounts to our sample. The PPR vector has a simple and useful algebraic representation. Define the preference vector $\pi \in \mathbb{R}^n$ with $\pi_u = 1$ and $\pi_v = 0$ for all $v \neq u$. Then, the PPR vector x is the solution to a linear system:

$$x = \tau\pi + (1 - \tau)Px,$$

where $P \in \mathbb{R}^{n \times n}$ is the graph transition matrix with $P_{vu} = A_{uv}/d_u$.

For a massive graph like the Twitter friendship network (large n), it is computationally intractable to calculate the exact PPR vector. We utilized an approximate algorithm to estimate the PPR vector, which examines only the friendship information of accounts nearby the seed account. Algorithm E.1, which extends the algorithm in Andersen et al. (2006), outlines this procedure. The PPR sampling takes as input the preference vector, the teleportation constant, and a tolerance controlling parameter $\varepsilon \in \mathbb{R}$. The algorithm maintains two vectors: the approximate PPR vector $p \in \mathbb{R}^n$ and the probability mass residual $r \in \mathbb{R}^n$, where p is initialized to be a vector of zeros and r is initialized as π. Once initialized, Algorithm 2.3 is an iterative procedure. At each iteration, the algorithm randomly samples one account that is "sufficiently close" to the seed using the criterion $r_u/d_u \geqslant \varepsilon$. Here, we used a heap to store the non-zero elements of r, which allows fast query of an eligible u. Then, the algorithm sends a Twitter application programming interface (API) request for u's friend list. Twitter API returns the basic information of d_u accounts that u follows, including the number d_v of their friends. Next, the probability mass at r_u is divided into three parts proportional to (i) τ, (ii) $(1-\tau)/2$, and (iii) $(1-\tau)/2$ and updated as follows. Part (i) is moved to p_u. Part (ii) is evenly distributed to r_v's, where v's are the accounts followed by u. Finally, part (iii) stays at r_u. When the iteration terminates, p becomes an approximate PPR vector. Algorithm 2.3 then includes any account v that has sufficiently large p_v ($> \varepsilon$) to the output sample. In some rare occasions where an account's friendship information is unavailable (e.g. private account), we excluded the node and put any non-zero values in p and r to r_u (the seed account).

In the murmuration 2018, we performed the PPR sampling (`https://github.com/RoheLab/aPPR`) for 59 seed accounts (Appendix Table E.1) with the teleportation constant $\tau = 0.1$ and tolerance $\varepsilon = 10^{-8}$. Such parameter settings leveraged satisfactory approximation rate within reasonable computational time. The sampling procedure was terminated when the sample size reached roughly 2×10^4. We recorded the friends (followings) and followers of all examined Twitter accounts along the PPR sampling. Collectively, the PPR sampling retried friendship information of 267,117 Twitter accounts who follow a total of 10,174,291 accounts. We then removed those accounts who follow fewer than 2 friends or are followed by fewer than 5 accounts, which resulted in a total of $n' = 193,120$ Twitter accounts who follow a total of $n'' = 1,310,051$ accounts. Finally, we represented the PPR sampling results

```
Input: Preference vector $\pi$, teleportation constant $\tau$, and tolerance $\varepsilon$
Procedure:
    Initialize $p \leftarrow 0$, $r \leftarrow \pi$, $\alpha' \leftarrow \alpha/(2-\alpha)$.
    while $\exists u \in V$ such that $r_u \geqslant \epsilon d_u$ do
        Send a Twitter API request for account $u$'s friend list
        Add $\tau r_u$ to $p_u$                                    // part (i)
        for all nodes $v$ that $u$ follows do
            Add $(1-\tau)r_u/(2d_u)$ to $r_v$                      // part (ii)
        Replace $r_u$ by $(1-\tau)r_u/2$                           // part (iii)
Output: A set of accounts that satisfy $p_v > \varepsilon$.
```

Algorithm E.1: PPR sampling using the Twitter API

with matrix $A' \in \{0,1\}^{n'' \times n''}$ with $A'_{uv} = 1$ if and only if account u follows account v and performed community detection on A'.

Vintage sparse PCA for flock identification

In this section, we explain the algorithm of vintage sparse PCA (VSP) in the context of Twitter friendship network. VSP is analogous to a simple form of factor analysis and estimates sparse components of the input data matrix (Chen and Rohe, 2020). To identify flocks with the sparse components, we applied VSP to the Twitter friendship network returned by PPR sampling. Let $A \in \{0,1\}^{n \times m}$ be the observed friendship matrix with $A_{ij} = 1$ if account i follows account j and 0 otherwise, where $n = 193,120$ is the number of accounts examined in PPR sampling, and $m = 1,310,051$ is the number of account followed by some of the n accounts. We focus on identifying flocks among the 1,310,051 Twitter accounts (columns of A), on which we performed the downstream analysis.

Define the *regularized Laplacian* $L \in \mathbb{R}^{n \times m}$ as

$$L_{ij} = \frac{A_{ij}}{\sqrt{r_i + \tau_r}\sqrt{c_j + \tau_c}}, \quad \text{for} \quad i = 1,2,\ldots,n, j = 1,2,\ldots,m,$$

where $r_i = \sum_{j=1}^{m} A_{ij}$ and $c_j = \sum_{i=1}^{n} A_{ij}$ are the row and column sums of A, and $\tau_r = \sum_i r_i/n$ and $\tau_c = \sum_j c_j/m$ are row- and column-regularizer respectively. Given L, Algorithm E.2 implements VSP and computes k sparse components in two main steps. The first step is to calculate a low-rank approximation using the singular value decomposition (SVD) of L,

$$L \approx \hat{U}\hat{\Lambda}\hat{V}^T,$$

where $\hat{U} \in \mathbb{R}^{n \times k}$ and $\hat{V} \in \mathbb{R}^{m \times k}$ have orthogonal columns and contain the leading left and right singular vectors, and $\hat{\Lambda} \in \mathbb{R}^{k \times k}$ is a diagonal matrix $\text{diag}(\lambda_1, \lambda_2, ..., \lambda_k)$ where $\lambda_i \geqslant \lambda_{i+1}$ for all $i = 1, 2, ..., k-1$. The second step applies a varimax rotation to $\hat{V}$ from its right. The varimax rotation maximizes the varimax criterion (Kaiser, 1958),

$$\begin{aligned} \text{maximize} \quad & \sum_{j=1}^{k} \left[\frac{1}{n} \sum_{i=1}^{n} Y_{ij}^4 - \left(\frac{1}{n} \sum_{i=1}^{n} Y_{ij}^2 \right)^2 \right], \\ \text{subject to} \quad & Y = \hat{V}O,\ O^T O = I_k,\ O \in \mathbb{R}^{k \times k}. \end{aligned} \tag{E.1}$$

Varimax maximizes the fourth sample moment of the rotated components $\hat{U}O$.[1] By maximizing this "kurtosis," varimax promotes sparsity in the rotated components. Finally, Algorithm E.2 outputs the rotated components $Y = \hat{V}\hat{O}$, where $\hat{O}$ is the orthogonal matrix that maximizes the objective in (E.1). Each of the k columns corresponds to a different community.[2] VSP is computationally feasible and suitable for Twitter friendship network. (Rohe and Zeng, 2020) showed that under the degree corrected mixed-membership stochastic block model, VSP offers a consistent estimate of the underlying block membership, provided the network is sufficiently large and dense.[3]

In the murmuration 2018, we identified $k = 100$ communities among the 1,310,051 Twitter accounts. We assigned each account a community member-

[1] Equation (E.1) is a simplified statement of varimax; because we do not use the row normalization step proposed in Kaiser (1958), the additional term $\left(\sum \sum Z_{ij}^2\right)^2$ is constant in O.

[2] In our implementation of Algorithm E.2, we configure the signs of O's columns such that every rotated singular vector is positively skewed (i.e., the sums of third moments of elements are positive, $\sum_i Y_{ij}^3 > 0$). This is valid because the optimal value in Equation (E.1) is invariant in sign flips of O's columns.

[3] This result requires an additional centering step which we find is unnecessary for our analysis.

```
Input: The regularized Laplacian $L$ and number of communities $k$.
Procedure VSP($L, k$):
    SVD. Calculate the SVD of $L$ and let $\hat{V}$ contain the top-$k$ right
      singular vectors.
    Varimax. Find orthogonal matrices $\hat{O}$ that maximizes the varimax
      criterion in Equation (E.1).
Output: $k$ sparse components in the columns of $Y = \hat{V}\hat{O}$
```

Algorithm E.2: Vintage Sparse PCA (VSP)

ship according to the rows of Y as returned from Algorithm E.2. Specifically, the i-th account was assigned to the l-th flock if Y_{il} is greater than other flock loadings of account i,

$$Y_{il} \geqslant Y_{il'}, \quad \text{for} \quad l' = 1, 2, ..., l-1, l+1, ..., k.$$

For the downstream analysis, we investigated the "leading members" of each flock, as measured by size of the elements in Y. In particular, for flock j with n_j member accounts, we focused on the top 1,000 member accounts that have the largest flock loadings, $Y_{i'j}$, where i' is the member accounts of flock j. The selected 50 flocks presented in the murmuration 2018 all have more than 1,000 members ($n_j > 1,000$, Appendix Table E.2). For those flocks with no more than 1,000 members ($n_j \leqslant 1,000$), we included their entirety for the downstream analysis.

Best feature function for keywords extraction in profile descriptions

Provided the member accounts of flocks, we analyzed the profile descriptions of Twitter accounts in order to label individual flocks. Consider the profile descriptions of m accounts that belong to the same flock. We treated each description as a text document and tokenized the words (terms) in it, followed by stop word removal. Then, we represent the m descriptions with a document-term matrix $X \in \mathbb{R}^{m \times w}$, where $X_i j$ is the number of times that word j occurs in document i, and w is the total number of distinct words (terms).

We applied a "best feature function" (BFF) technique that helped us extract keywords from profile descriptions. BFF is complementary to VSP. For each flock (i.e., each component in VSP), BFF assigns each word (or term) a feature score. The feature score $F_{ij} \in \mathbb{R}$ of word i for flock j is the average of word frequency in each profile, weighted by flock loadings $Y_{\cdot j}$,

$$F_{ij} = \sum_{l=1}^{m} X_{li} Y_{lj}.$$

For variance stabilization, we applied a square root transformation on F_{ij}. The feature score F_{ij} measures the importance of term j to the i-th flock. To interpret each flock, we inspected the most representative 15 keywords suggested by BFF while also relying on skimming through the actual profile descriptions. BFF effectively reduces the work load from domain experts; once the top terms are cross-validated by examining several representative accounts, a clear interpretation would surface. For example, the top scored terms for the "national political journalists" flock, such as "reporter," "political", "correspondent,", "white house", and "political reporter,", suggest that this flock consists of reporters/correspondents stationed in Washington DC covering the White House (Appendix Table E.3). An examination of profile descriptions of the top accounts validates this interpretation.

Text analysis for news event identification

This section describes the first stage of our daily workflow that identifies ongoing news events from tweets. After pre-processing of text (e.g., stop word removal, white space removal, emoji removal, tokenization), we transform all the tweets into a document-term matrix, where a document is a tweet and a term is a word. Given a set of current tweets (from one day), we measure the distinctive usage for each terms by contrasting its frequency to that in a historical corpus. Here, we define the historical corpus as the collection of previous tweets spanning the preceding two months. For any terms i in a corpus of w words, let $\tilde{c}_i$ be the count of its occurrence in the historical corpus, and let c_i be its frequency in the current corpus. We define the contrast of word i as

$$\hat{c}_i = \sqrt{\frac{wc_i}{\sum_i c_i}} - \sqrt{\frac{w\tilde{c}_i}{\sum_i \tilde{c}_i}}, \tag{E.2}$$

where a larger value suggests a more distinctive usage. In particular, we add a smoothing constant to $\tilde{c}_i$'s, which is set to 10 (Manning et al., 2008). We collect terms with big contrast of the day for downstream analysis. In our experience, proceeding with top 50 terms (in contrast measure, $\hat{c}$) generally offers a satisfactory coverage of news events reported by the mainstream media.

We cluster the selected terms based on their co-occurrence in tweets. Specifically, the similarity of two terms, i and j is measured by the number of co-occurrences, normalized by the total number of times they appear in the current day's corpus. This is calculated with the document-term matrix X of the current day's corpus, where X_{ki} is the number of times that term i appears in document j. The similarity between i and j is defined as

$$S_{ij} = \frac{\sum_k X_{ki} X_{kj}}{\sqrt{\sum_k X_{ki}}\sqrt{\sum_k X_{kj}}}.$$

The similarity measure does not normalize for popularity of individual terms, that is, some terms co-occur broadly with many other terms and some terms only co-occur with a few specific terms. To account for this, we further normalize similarity matrix by its row sums, $\Sigma_{ij} = S_{ij}/(\tau \cdot \sum_j S_{ij})$, where $\tau = 1.02$ is a regularization constant (assuming 50 terms are selected previously). Next, we apply Ward's hierarchical clustering (Murtagh and Legendre, 2014) to the matrix $\Sigma\Sigma \circ \Sigma$, where $\circ$ is element-wise matrix multiplication and obtain clusters of terms. Each cluster of terms are related to a news event and the terms are used to initiate the downstream searching for related tweets about the event.

Forward searching for tweet classification

In the second stage of the daily text analysis, we assign relevant tweets into events. To this end, we define inclusion terms for each event so that any tweets that contain at least one of the inclusion words are treated as pertaining to the

```
Input: An initial set of inclusion terms I and the desired number of
       inclusion terms $n$
Procedure FS-Keywords($I, n$):
    Initialize the set of searched inclusion terms $S = I$.
    while set $S$ contains less than $n$ terms do
        1. Update $X$ with the document-term matrix of tweets currently
           included by the terms in $S$.
        2. Calculate normalized word count $\hat{c}$ as defined in Equation
           (E.2) with $X$.
        3. Find the term $i$ outside $S$ with the greatest word contrast
           measure, i.e., $\hat{c}_i \geqslant \hat{c}_j$ for $\forall j \notin S$.
        4. Add term $i$ to the set of inclusion terms, $S \leftarrow S \cup \{i\}$.
Output: The final set of inclusion terms $S$
```

Algorithm E.3: Forward searching for inclusion keywords

event. The initial set of inclusion terms for an event are the terms identified in the previous step. To expand the set of inclusion terms, we devised a forward searching algorithm (Algorithm E.3) that adds one inclusion term at a time. Algorithm E.3 includes each term that is the most "representative" in the already included tweets, as measured by Equation (E.2). In particular, we have a specialist who adjudicates whether the suggested inclusion terms are valid and exclude those that are inappropriate based on their understanding of the news events. Finally, given the lists of inclusion terms for all events, we identify event tweets that include any one of the inclusion terms.

E.2 Supporting discussions

Comparing social-network-based and text-based contextualization

Apart from identifying flocks by social network structure, an alternative is to reply on text to do so. Here we present a comparison between text-based contextualization and social-network-based approach.

The social-network-based approach was elaborated in previous sections (Appendix Table E.2). For text-based approach, we first aggregated tweets by account and performed topic modeling by adapting Latent Dirichlet allocation

(LDA), which yielded estimation of the probability of each account's tweets belonging to a certain topic. Specifically, we fit the LDA model with 50 topics by invoking the Gibbs Sampling (Phan et al., 2008) (with the default settings implemented in R package `topicmodels`). Appendix Table E.5 lists the top 20 terms for each topic estimated by LDA. Then, we aggregated the probabilities by flock. Appendix Figure E.4 displays the relations between flocks and the 50 topics. The overall results show that each flock tended to be uniquely associated with a certain topic, suggesting similar effectiveness between a social-network-based approach and a text-based approach.

However, we did find subtle differences. We computed entropy of the two estimations at individual accounts level. Specifically, for LDA, we computed the entropy of topical distribution for each aggregated document (i.e. each account's tweets), and for VSP, we calculated the entropy of row-normalized loadings (Y) for each account. Comparing the two approaches, we observed that the social-network-based approach yields a lower entropy for a vast majority of clusters, indicating less uncertainty in classification (Appendix Figure E.5ab). This could potentially be attributed to the fact that a flock may express opinion on a wide range of topics over time, whereas their social context may stay relatively stable. As a result, finding flocks based on the topics they discussed in their tweets introduces a greater level of uncertainty. This suggests that that our method is better suited for finding finer-grained communities than the approach based only on text.

E.3 Supporting figures

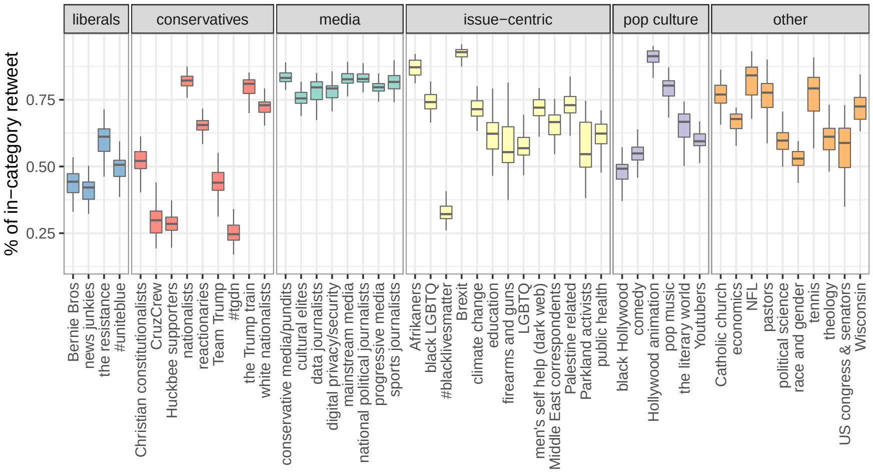


Figure E.1: Box plot of proportions of same-category retweeting. For each account, the percentage of its retweets that originated from other accounts of the same flock category was computed. Each box corresponds to one flock, with the panels indicating flock category.

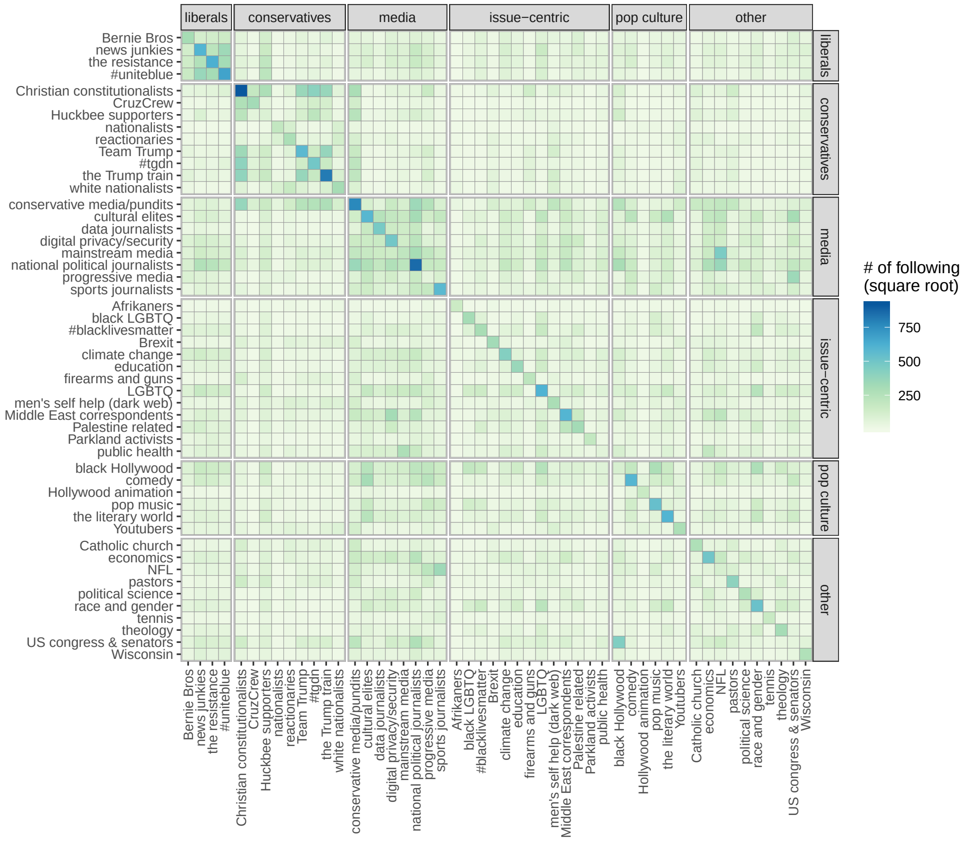


Figure E.2: Heat map of friend counts among flocks. Each row and column corresponds to a flock, in the same order. The row and column panels indicate flock categories, with the category names shown in the top and right strips. The color shades indicate the number of followings from the row flock to the column flock, with the square root transformation.

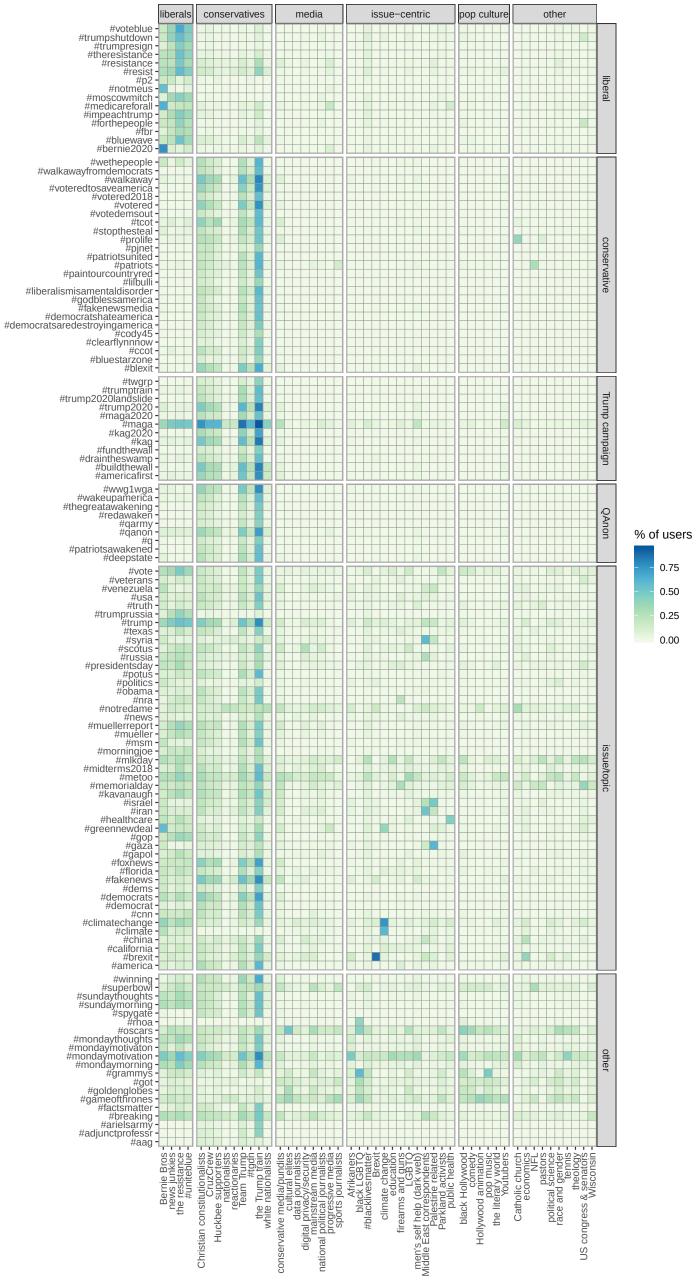


Figure E.3: The usage of 129 frequently used hashtags by 50 flocks. Each row corresponds to one hashtag, with row panels on the right indicating the topical category. Each column represents one flocks with column panels on the top indicating the flock category. The darkness in color shows the percentage of active accounts in the flock who used the hashtag.

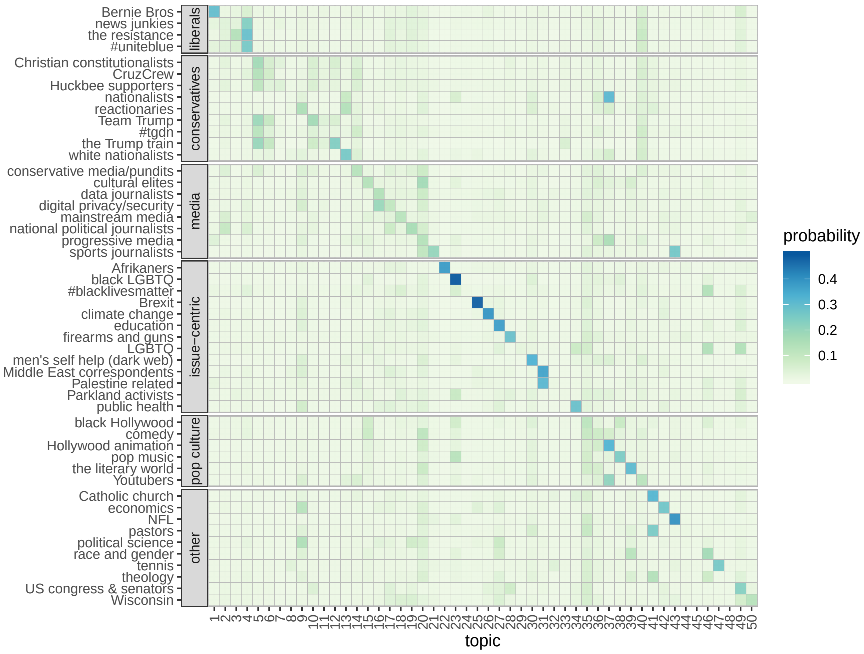


Figure E.4: The topical distribution of each account's tweets aggregated by flock. Each row corresponds to one flock, with column panels indicating flock category. Each column represents to one topic identified by adapting the Latent Dirichlet allocation with each account's tweets. The color darkness indicates the average topical probability distribution of individual account's tweets across each flock, weighted by the corresponding loadings from VSP.

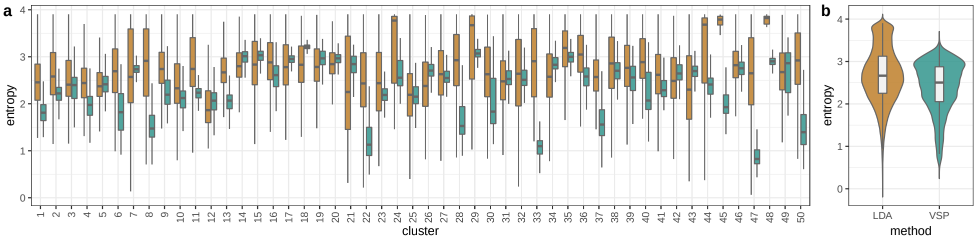


Figure E.5: Comparison of the entropy in LDA and VSP estimations. For each account in the 50 flocks, the entropy of its social context estimation by LDA (brown) and VSP (green) is shown, (a) stratified by 50 matching clusters, (b) across all accounts examined.

E.4 Supporting tables

Table E.1: Seed nodes used for PPR sampling in 2018 and 2019. Handles belongs to one of two types – "individuals" or "media". The category of seed nodes is also reported.

	Twitter handle	Type	Category	In 2018	In 2019
1	@anamariecox	individuals	liberal	Yes	Yes
2	@andersoncooper	individuals	liberal	Yes	Yes
3	@GStephanopoulos	individuals	liberal	Yes	Yes
4	@chucktodd	individuals	liberal	Yes	Yes
5	@maddow	individuals	liberal	Yes	Yes
6	@ezraklein	individuals	liberal	Yes	Yes
7	@NateSilver538	individuals	liberal	Yes	Yes
8	@ggreenwald	individuals	liberal	Yes	No
9	@deray	individuals	blacktwitter	Yes	Yes
10	@nhannahjones	individuals	blacktwitter	Yes	Yes
11	@MHarrisPerry	individuals	liberal	Yes	No
12	@Moore_Darnell	individuals	blacktwitter	Yes	Yes
13	@WesleyLowery	individuals	blacktwitter	Yes	Yes
14	@FeministaJones	individuals	blacktwitter	Yes	Yes
15	@glennbeck	individuals	conservative	Yes	Yes
16	@GovMikeHuckabee	individuals	conservative	Yes	No
17	@seanhannity	individuals	conservative	Yes	No
18	@benshapiro	individuals	conservative	Yes	Yes
19	@DineshDSouza	individuals	conservative	Yes	Yes
20	@AnnCoulter	individuals	conservative	Yes	Yes
21	@RealAlexJones	individuals	alt-light	Yes	No
22	@StefanMolyneux	individuals	alt-light	Yes	Yes
23	@PrisonPlanet	individuals	alt-light	Yes	Yes
24	@Cernovich	individuals	alt-light	Yes	Yes
25	@gatewaypundit	individuals	alt-light	Yes	Yes
26	@RichardBSpencer	individuals	alt-right	Yes	Yes
27	@dailykos	media	liberal	Yes	Yes
28	@politicususa	media	liberal	Yes	Yes
29	@thinkprogress	media	liberal	Yes	Yes
30	@voxdotcom	media	liberal	Yes	Yes
31	@mmfa	media	liberal	Yes	Yes
32	@PolitiFact	media	liberal	Yes	Yes
33	@Salon	media	liberal	Yes	Yes
34	@thenation	media	liberal	Yes	Yes
35	@HuffPost	media	liberal	Yes	Yes
36	@MSNBC	media	liberal	Yes	Yes
37	@washingtonpost	media	mainstream	Yes	Yes
38	@nytimes	media	mainstream	Yes	Yes

Table E.1: Seed nodes used for PPR sampling in 2018 and 2019 (continued).

	Twitter handle	Type	Category	In 2018	In 2019
39	@CNN	media	mainstream	Yes	Yes
40	@AP_Politics	media	mainstream	Yes	Yes
41	@CBSPolitics	media	mainstream	Yes	Yes
42	@NBCPolitics	media	mainstream	Yes	Yes
43	@ABCPolitics	media	mainstream	Yes	Yes
44	@politico	media	mainstream	Yes	Yes
45	@USATODAY	media	mainstream	Yes	Yes
46	@WSJ	media	mainstream	Yes	Yes
47	@FoxNews	media	mainstream	Yes	Yes
48	@theblaze	media	conservative	Yes	Yes
49	@dcexaminer	media	conservative	Yes	Yes
50	@DailyCaller	media	conservative	Yes	Yes
51	@conserv_tribune	media	conservative	Yes	Yes
52	@BreitbartNews	media	conservative	Yes	Yes
53	@infowars	media	conservative	Yes	No
54	@instapundit	media	conservative	Yes	No
55	@townhallcom	media	conservative	Yes	Yes
56	@RedState	media	conservative	Yes	Yes
57	@NRO	media	conservative	Yes	Yes
58	@TheRoot	media	blackmedia	Yes	Yes
59	@EBONYMag	media	blackmedia	Yes	Yes
60	@davidaxelrod	individuals	liberal	No	Yes
61	@maggieNYT	individuals	liberal	No	Yes
62	@jonfavs	individuals	liberal	No	Yes
63	@TuckerCarlson	individuals	conservative	No	Yes
64	@BretBaier	individuals	conservative	No	Yes
65	@brithume	individuals	conservative	No	Yes
66	@greggutfeld	individuals	conservative	No	Yes
67	@IngrahamAngle	individuals	conservative	No	Yes
68	@RealJamesWoods	individuals	conservative	No	Yes
69	@DrDavidDuke	individuals	alt-light	No	Yes
70	@JackPosobiec	individuals	alt-right	No	Yes
71	@Blklivesmatter	individuals	blacktwitter	No	Yes
72	@Nettaaaaaaaa	individuals	blacktwitter	No	Yes
73	@BreeNewsome	individuals	blacktwitter	No	Yes
74	@paulkrugman	individuals	liberal	No	Yes
75	@BuzzFeedBen	individuals	liberal	No	Yes
76	@TPM	media	liberal	No	Yes
77	@WestJournalism	media	conservative	No	Yes
78	@FreeBeacon	media	conservative	No	Yes
79	@Colorlines	media	blackmedia	No	Yes
80	@thehill	media	liberal	No	Yes
81	@theintercept	media	liberal	No	Yes
82	@jacobinmag	media	liberal	No	Yes

Table E.2: The 100 flocks of 2018. For each flock, the total number of its member accounts and the median ratio of members' followers over followings ratios (FFR) are reported. The 50 selected flocks are marked as "Yes" in the last column.

	Category	Name	# of members	FFR	Selected
1	conservatives	Christian constitutionalists	27835	1.07	Yes
2	conservatives	CruzCrew	7524	0.82	Yes
3	conservatives	Huckbee supporters	20023	0.98	Yes
4	conservatives	nationalists	2861	2.11	Yes
5	conservatives	reactionaries	6815	2.00	Yes
6	conservatives	Team Trump	20770	1.23	Yes
7	conservatives	#tgdn	16561	0.98	Yes
8	conservatives	the Trump train	46918	1.05	Yes
9	conservatives	white nationalists	6626	1.53	Yes
10	foreign	Australia	8393	8.32	No
11	foreign	Brazil	4925	150.59	No
12	foreign	Canada	17308	7.71	No
13	foreign	France	5712	10.01	No
14	foreign	Germany	4311	7.66	No
15	foreign	Israel	5779	16.49	No
16	foreign	South Africa	15509	30.90	No
17	foreign	Sweden	2625	5.21	No
18	foreign	UK	26853	20.42	No
19	issue-centric	Afrikaners	2237	6.74	Yes
20	issue-centric	black LGBTQ	12929	2.68	Yes
21	issue-centric	#blacklivesmatter	10706	1.52	Yes
22	issue-centric	Brexit	9832	7.73	Yes
23	issue-centric	climate change	20259	4.70	Yes
24	issue-centric	education	14813	3.89	Yes
25	issue-centric	firearms and guns	4880	12.38	Yes
26	issue-centric	LGBTQ	27263	4.84	Yes
27	issue-centric	men's self help (dark web)	7052	11.13	Yes
28	issue-centric	Middle East correspondents	18549	15.17	Yes
29	issue-centric	Palestine related	11265	7.30	Yes
30	issue-centric	Parkland activists	2846	1.75	Yes
31	issue-centric	public health	14882	6.54	Yes
32	liberals	Bernie Bros	12173	1.30	Yes
33	liberals	news junkies	19035	0.98	Yes
34	liberals	the resistance	32604	1.07	Yes
35	liberals	#uniteblue	13708	1.00	Yes
36	media	conservative media/pundits	17828	12.68	Yes
37	media	cultural elites	17041	9.53	Yes
38	media	data journalists	13910	2.93	Yes
39	media	digital privacy/security	19384	10.43	Yes
40	media	mainstream media	11151	9.51	Yes

Table E.2: The 100 flocks of 2018 (continued).

	Category	Name	# of members	FFR	Selected
41	media	national political journalists	9619	11.78	Yes
42	media	progressive media	8217	8.71	Yes
43	media	sports journalists	17658	35.22	Yes
44	other	arts and culture	16006	14.04	No
45	other	bitcoin	6820	28.81	No
46	other	black women	12906	3.18	No
47	other	chess	1431	10.09	No
48	other	culinary circle	14409	13.09	No
49	other	fashion	20122	262.37	No
50	other	Kardashians	994	29.40	No
51	other	life (entertainment)	177	33.88	No
52	other	life (entertainment)	864	242.04	No
53	other	NASA	9072	51.89	No
54	other	social media marketers	41841	1.09	No
55	other	social media marketers	56944	1.22	No
56	other	tech & vc	28714	23.90	No
57	other	US law enforcement agencies	586	10.81	No
58	other	video games	5703	0.94	No
59	other	video games	18356	65.68	No
60	other	Catholic church	8240	7.51	Yes
61	other	economics	15054	15.44	Yes
62	other	NFL	18117	70.95	Yes
63	other	pastors	13323	42.78	Yes
64	other	political science	7564	1.50	Yes
65	other	race and gender	19048	2.28	Yes
66	other	tennis	2881	36.04	Yes
67	other	theology	9908	2.19	Yes
68	other	US congress & senators	9466	21.85	Yes
69	other	Wisconsin	7277	2.25	Yes
70	other	black Hollywood	24585	81.59	Yes
71	other	comedy	17718	25.51	Yes
72	other	Hollywood animation	3670	23.22	Yes
73	other	pop music	25133	93.76	Yes
74	other	the literary world	23117	4.54	Yes
75	other	Youtubers	6784	5.50	Yes
76	regional	Arkansas	3421	1.57	No
77	regional	Baltimore	8646	3.03	No
78	regional	Boise	3125	1.72	No
79	regional	Boston	15224	4.70	No
80	regional	Chicago	11752	2.89	No
81	regional	Colorado	17168	3.84	No
82	regional	Florida	6740	2.08	No
83	regional	Florida	17134	3.27	No
84	regional	Iowa	6685	2.06	No

Table E.2: The 100 flocks of 2018 (continued).

	Category	Name	# of members	FFR	Selected
85	regional	Louisiana	8932	2.70	No
86	regional	Michigan	9881	2.42	No
87	regional	Minnesota	8975	3.35	No
88	regional	Nashville	26254	280.59	No
89	regional	New Jersey	7948	2.09	No
90	regional	North Carolina	9583	2.32	No
91	regional	NYC	14543	4.83	No
92	regional	Ohio	12148	2.91	No
93	regional	Philly	12084	2.99	No
94	regional	Portland	11985	2.68	No
95	regional	South Carolina	6456	2.16	No
96	regional	St. Louis	10887	2.88	No
97	regional	Texas	14937	2.87	No
98	regional	Utah	4155	3.12	No
99	regional	Wake Forest University	2127	1.39	No
100	regional	Washington, DC	15242	3.50	No

Table E.3: Keywords in member accounts' profiles of the 50 selected flocks as determined by BFF.

	Name	Keywords in member's profile
1	Christian constitutionalists	conservative, maga, christian, patriot, tcot, god, prolife, nra, constitution, pjnet, cruzcrew, constitutional, ccot, deplorable, usa
2	CruzCrew	cruzcrew, conservative, cruz, christian, ted, constitutional, constitution, nevertrump, prolife, pjnet, tedcruz, god, country, ccot, liberty
3	Huckbee supporters	conservative, tcot, huckabee, prolife, mike, teaparty, christian, sgp,republican, grassroots, liberty, constitution, american, freedom, president
4	nationalists	frogtwitter, nice, respecter, nationalism, trad, racc, orthodox, priv, paulhead, illuminati, screws, iq, normantwitter, loose, anime
5	reactionaries	nrx, reactionary, catholic, traditionalist, neoreactionary, patriarchy, srx, evolution, neoreaction, philosophy, monarchist, civilization, enemies, hitler, menciian
6	Team Trump	trump, maga, donald, makeamericagreatagain, america, president, americafirst, supporter, trumptrain, support, american, kag, buildthewall, god, potus
7	#tgdn	tgdn, conservative, tcot, nra, libertarian, maga, christian, lgod, liberty, country, patriot, constitution, vet
8	the Trump train	maga, kag, trump, fb, nra, military, buildthewall, americafirst, patriot,vets, wwgwga, trumptrain, prolife, god, conservative
9	white nationalists	nationalist, identitarian, white, european, altright, traditionalist, identity, american, nationalism, proeuropean, liftwaffe, altmedia, racist, whitegenocide, antiwhite
10	Afrikaners	en, south, die, afrikaans, van, vir, sa, afriforum, african, op, africa, wat,nuus, ons, pretoria
11	black LGBTQ	ig, gay, black, i_—Èm, actor, instagram, lgbt, hiv, snapchat, bitch, model,gaymer, atl, morehouse, hivaids
12	#blacklivesmatter	louis, st, ferguson, stl, justice, black, blacklivesmatter, activist,organizer, liberation, postdispatch, people, mo, freedom, fighter
13	Brexit	mp, parliament, brexit, email, minister, conservative, casework, bbc, editor, uk,secretary, labour, mep, queries, constituency
14	climate change	climate, energy, environment, change, environmental, science, clean, global, transition, renewable, solutions, defense, earth, sustainable, solar
15	education	education, schools, ed, school, students, policy, teacher, educators, public, educational, astronaut, teachers, nonprofit, charter, college
16	firearms and guns	firearms, shooting, hunting, gun, gear, outdoor, guns, manufacturer, accessories, tactical, shooters, worlds, industry, rifles, rifle

Table E.3: Keywords in member accounts' profiles of the 50 selected flocks (continued).

	Name	Keywords in member's profile
17	LGBTQ	trans, lgbtq, reproductive, rights, justice, lgbt, queer, gender, lesbian, women, people, transgender, gay, feminist, community
18	men's self help (dark web)	entrepreneur, emails, daily, masculinity, money, philosophy, online, fitness, coach, teach, sign, dating, psychology, copywriter, author
19	Middle East correspondents	middle, east, fellow, syria, correspondent, senior, endorsement, iraq, egypt, mena, foreign, security, analyst, beirut, views
20	Palestine related	palestine, palestinian, rights, bds, middle, east, gaza, boycott, human, israeli, israel, journalist, occupation, solidarity, apartheid
21	Parkland activists	msd, msdstrong, neveragain, march, douglasstrong, eagle, marchforourlives, lives, douglas, activist, oliver, joaquin, ucf, change, guac
22	public health	health, policy, care, economist, professor, medical, economics, healthcare, medicine, medicaid, kaiser, senior, researcher, medicare, reporter
23	Bernie Bros	bernie, feelthebern, sanders, revolution, ourrevolution, berniesanders, progressive, grassroots, presidential, vote, progressives, medicareforall, campaign, support, affiliated
24	news junkies	liberal, progressive, resist, theresistance, democrat, notmypresident, obama, junkie, connecttheleft, atheist, retired, rwnj, left, stillwithher, lgbt
25	the resistance	theresistance, fbr, resist, bluewave, resistance, trumprussia, notmypresident, blm, geeksresist, trump, blocked, impeach, impeachtrump, fbpe, wearethepatriots
26	#uniteblue	uniteblue, resist, liberal, theresistance, progressive, fbr, democrat, notmypresident, obama, resistance, lgbt, impeachtrump, equality, rwnjs, blue
27	conservative media/pundits	contributor, host, columnist, editor, author, conservative, bestselling, review, examiner, fox, cohost, contributing, fellow, syndicated, senior
28	cultural elites	editor, writer, critic, senior, culture, dot, book, york, times, magazine, reporter, staff, cohost, buzzfeed, film
29	data journalists	data, graphics, editor, design, journalism, visualization, previously, code, york, visual, product, visuals, times, designer, computational
30	digital privacy/security	security, privacy, law, pgp, technology, liberties, tech, hacker, aclu, civil, cto, surveillance, author, professor, lawyer
31	mainstream media	news, cnn, correspondent, breaking, anchor, emergencies, weather, monitored, official, police, dial, cbs, nbc, twitter, department
32	national political journalists	correspondent, political, reporter, cnn, washington, white, politics, chief, politico, national, senior, covering, alum, bureau, analyst

Table E.3: Keywords in member accounts' profiles of the 50 selected flocks (continued).

	Name	Keywords in member's profile
33	progressive media	bylines, writer, cohost, hire, deadspin, gmail, dot, jokes, genious, podcast, dsa, boy, polygon, album, stone
34	sports journalists	nba, espn, writer, sports, baseball, mlb, nfl, senior, insider, basketball, network, podcast, athletic, analyst, columnist
35	Catholic church	catholic, prolife, diocese, faith, archdiocese, vatican, catholics, abortion, life, priest, bishop, church, archbishop, roman, religion
36	economics	economics, economist, economic, policy, bloomberg, professor, health, reserve,markets, financial, macro, ft, wall, research, bank
37	NFL	nfl, football, miami, espn, sports, college, hurricanes, analyst, coach, network, insider, ig, university, national, draft
38	pastors	pastor, church, jesus, husband, author, christ, hillsong, god, christian, apologetics, father, baptist, faith, president, gospel
39	political science	political, professor, science, scientist, university, prof, politics, assistant, associate, study, behavior, elections, american, methods, data
40	race and gender	black, professor, historian, studies, author, african, race, prof, feminist, american, scholar, phd, history, sociologist, justice
41	tennis	tennis, player, professional, instagram, atp, tour, pro, wta, play, champion, cup, grand, slam, world, official
42	theology	religion, author, pastor, faith, church, preacher, professor, justice, theological, theologian, historian, theology, black, seminary, speaker
43	US congress & senators	district, congressional, proudly, representing, congressman, represent, congress, house, serving, chairman, committee, subcommittee, representatives, serve, honored
44	Wisconsin	wisconsin, milwaukee, assembly, wisconsins, wi, journal, madison, representative, district, sentinel, news, senator, racine, senate, counties
45	black Hollywood	black, producer, actress, actor, bookings, booking, grammy, ig, host, instagram, award, infocom, entertainment, tv, activist
46	comedy	comedian, comedy, netflix, writer, standup, podcast, itunes, special, watch, snl,central, streaming, actor, late, album
47	Hollywood animation	animator, artist, animation, storyboard, icon, draw, creator, nsfw, voice, contact, actor, illustrator, yotta, game, cartoon
48	pop music	booking, bookings, grammy, music, album, infocom, mgmt, dj, inquiries, ig, contact, beats, hop, tde, bookingscom
49	the literary world	literary, fiction, books, poetry, nonfiction, book, publisher, literature, bookstore, magazine, publishing, writing, independent, author, writers
50	Youtubers	youtube, youtuber, videos, egalitarian, channel, im, creator, shit, gamer, patreon, internet, atheist, poisoning, twitch, merch

Table E.4: Top 30 keywords in tweets of 10 flocks. The keywords are extracted using the BFF technique.

Flock	MUELLER	ABORTION	KHASHOGGI
the Trump train	mueller, collusion, democrats, dems, report, investigation, #maga, fbi, obama,#patriotsawakened, trump, russia, president, time, #muellerreport, robert, years, #trump2020, fake, #mueller, hillary, dossier, media, #wwg1wga, hunt, witch, #factsmatter, flynn, fisa, clinton	abortion, babies, life, baby, murder, #abortionismurder, abortions, heartbeat, matters, choose, democrats, planned, #prolife, parenthood, dems, killing, #chooselife, bill, born, infanticide, kill, unborn, birth, god, alive, term, democrat, support, #abortionisnothealthcare, innocent	brotherhood, muslim, msm, laden, #patriotsunited, osama, obama, benghazi, #benghazi, terrorist, media, outrage, citizen, #votered, connected, #fakenews, #fakenewsmedia, visa, russiagate, alqaeda, #voteredtosaveamerica, illegal, american, #maga, libya, holder, brennan, card, truthleaks, #osamabinladen
Christian constitutionalists	#aag, #maga, democrats, collusion, #tcot, obama, hillary, #tlot, strzok, clinton, #the200, media, hoax, dossier, fbi, #trump2020, #uniteblue, flatlined, comey, jw, #pjnet, brennan, mass, weissmann, #news, #nahbabynah, #teaparty, #bbc, hunt, witch	#aag, lifenews, #ccot, #homeposts, #abortion, #tcot, babies, #pjnet, #state, parenthood, #national, planned, democrats, #unbornlivesmatter, baby, proabortion, democrat, born, abortions, #maga, prolife, alive, unborn, survive, #praytoendabortion, abortion, theblaze, infanticide, black, newsbusters	#keepamericagreat, thinker, hebdo, charlie, #kag, contempt, foley, #saudikillsjamalkhashoggi, treated, james, paris, slain, obama, #maga, #charliehebdo, #jamesfoley, #aag,muslim, brotherhood, #tcot, disguised, tw403, democrat, tw395, tw520, breathless, american, #yahoo, soldiers, #ccot

white nationalists	#aag, mcfeels, #ftn, ftn, pilpul, israel, #golanheights, blrompf, miga,prn, changedand, workrelease, deflating, boomer, gay, boomers, #collegecheatingscandal, asimov, #editorial, repudiates, dylan, #wallst, avenattis, sambei, chicanery, promueller, hunger, ralliers, songbook, probation	#aag, whites, kushner, jewish, casual, wypipo, invaded, soulsucking, childrearing, $22t, monogamy, multiculturalism, mcfeels, expression, miscegenation, endless, tiddies, hoes, cancelled, degeneracy, antiwhite, nationalists, logically, prostitutes, atheist, cartel, mutual, subs, riddance, certificates	confusing, raghead, israel, mindless, posturing, overhyped,abortion, #aag, poison, surprising, carlson, myths, buchanan, privilege, countless, snowden, routine, introducing, class, frenzy, starved, tucker, injected, bombed, nails, undercut, symptom, pitiful, acted, desired
conservative media/pundits	#aag, gp, #ampfw, flatlined, #wallst, podcast, milking, episode, #faultlines, #bft, journos, themccarthyreport, #twtfrontpage, brackets, #allyourdreamsaredead, pimps, #leftunhinged, madnesscom, #frontpage, #a1, pouncing, journo, melanin, rowling, joins, cohenprague, peters, column, oversold, #teamtrump	#aag, #homeposts, gp, conservatism, gosnell, opinion, #state, filmed, abortionrights, argument, #national, bulwark, editor, debate, #ncpol, episode, media, croatia, unrepresentative, uncivil, boli, #ncga, proabortion, jedi, modified,civilly, pounce, arguments, rareness, noting	islamist, #aag, #keepamericagreat, thinker, charlie, hebdo, proiran, chamber, echo, #saudikillsjamalkhashoggi, caper, contempt, treated, irresponsible, #kag, abattoir, usiran, chicoms, foley, cudgel, disaffected, floats, axis, #frontpage, #a1, #twtfrontpage, assad, iranbacked, turkishqatari, rick
progressive media	#aag, maturity, flatlined, #wallst, cursor, dms, tracey, astute, habit, milking, krassensteins, bonuses, pimps, blinking, tattoo, laptop, tl, krassenstein, neocons, hipster, staring, pornographic, salads, denialists, #bbc, gaming, undisguised, vra, amounting, dipshits	dsa, donate, donation, bowlathon, #bowl19, receipt, raising, upton, fund, doggo, scoping, swag, comrades, lipinski, fundraising, teams, #abortionsolidarity, #homeposts, #pissedoffpeaches, comics, dccc, raised, funds, hotdogs, #aag, matched, otto, fred, fundraiser, #fundabortionbuildpower	podesta, iming, qanon, strainer, bulldozing, pulverizing, towns, wears, nonsensical, wictor,friedman, retweets, subscribers, shia, careless, impression, village, swipe, province, federalist, unbelievably, endorsing, censored, davis, blockbuster, heaven, bernie, junior, fuckin, lol

national political journalists	#mtpdaily, #mtp, #nhpolitics, aides, hse, #ifitssunday, cmte, ap, cillizza, stakeout, hrng, spox, arrives, #amrstaff, collinson, #fitn, sxm, negotiations,spotted, nbcwsj, nonmueller, gillibrand, caphill, nh, mbrs, palm, #cnnstakeout, longerterm, adds, #ap	#mtp, shifting, #mtpdaily, postalabama, lifenews, #nhpolitics, mccarthy, #lagov, dga,nh, #lalege, primary, #fitn, suburban, alito, mostall, recennservative, scrapped, localities, weighs, emerging, spate, phenomenon, competitive, erupt, foreshadows, #valeg, broader, breyer, reports	#mtp, ryan, fred, publisher, scoop, tells, #powerup, pence, #amr, pres, editorial, statement, sen, adds, #mtpdaily, ap, asked, column, deception, longestablished, answers, hill, colleague, disparaged, #podsavetheworld, capitol, tonight, recommending, reporters, antiterror
Middle East correspondents	#trumpwatch, kayla, kotkin, #mog, rapport, captors, yazidi, #world, sayyaf, jailer, baghdadi, foley, peskov, #isis, quil, congr, tre, #trumprussia, stasi, headfirst, graf, bibi, lawenforcement, husbands, #wallst, iracontrolled, gulf, parliament, gg, #factchecker	arab, societies, islamophobia, gp, koreas, objecting, excerpt, philippines, conspired, commentators, summarize, resolution, #globalgagrule, #trumps, #mog, stoning, lesbian, sultans, discomfort, amputation, deprived, extramarital, greenhouse, romania, bomb, charities, briefings, uk, brunei, #us	saudi, khashoggi, consulate, #saudi, #khashoggi, jamal, #jamalkhashoggi, istanbul, turkish, disappearance, mbs, case, turkey, #turkey, killing, erdogan, riyadh, investigation,murder, officials, official, affair, arabia, prince, latest, #mbs, arab, authorities, piece, #saudiarabia
Bernie Bros	#readthemuellerrepor, bribed, russiagate, steadfastly, corp, #bernie2020, gopnazi, #trumpresign, kleptocrats, bribing, #dailykos, fellow, hacks, jokes, resignation, bernie, dt, puppets, #tulsi2020, corporate, vips, #tulsiforpresident, #russiagate, oligarchs, #bernie, theory, #biden, assange, #tulsigabbard, pass	#bernie2020, biden, bernie, birmingham, constitutional, upton, joe, iraq, womans, lipinski, extremist, medicare, sanders, dccc, fred, rights, comprehensive, anita, campaigned, sole, donate, #pissedoffpeaches, guaranteeing, access, #medicareforall, pathological, desegregation, berniesanders, #dailykos, compromise	yemen, #wikileaks, #freejulian, #yemen, yemeni, saudi, rt, arabia, saudiled, united, bernie, julian, children, states, bombing, reevaluate, humanitarian, support, sanders, corporate, starving, genocide, #jamalkhashoggi, theyoungturks, assange, largest, theintercept, regime, dictatorship, unequivocal

#uniteblue	mueller, trump, report, barr, congress, read, president, public, house, justice, american, obstruction, full, trumps, evidence, people, attorney, release, general, impeachment, muellers, clear, letter, summary, donald, robert, republicans, mcconnell, criminal, #muellerreport	women, alabama, ban, bans, republicans, georgia, gop, pregnant, trump, ohio, republican, abortion, politicians, health, conservatives, #alabamaabortionban, rape, law, safe, #abortionisawomansright, access, rights, state, lawmakers, raped, fight, antiabortion, states, legal, missouri	trump, khashoggi, saudi, jamal, murder, prince, crown, kushner, arabia, murdered, saudis, journalist, cia, trumps, house, jared, president, cover, dismembered, #p2, bone, breaking, america, knew, white, money, lied, mohammed, tortured, salman
the resistance	report, mueller, trump, barr, congress, #releasethefullmuellerreport, house, public, read, full, justice, trumps, impeachment, people, obstruction, president, release, #muellerreport, #releasethereport, summary, attorney, muellers, american, gop, mcconnell, donald, letter, clear, democracy, call	women, #stopthebans, gop, access, ban, bans, alabama, rights, antiabortion, safe, rape, #womensrightsarehumanrights, #abortionisawomansright, passed, state, republicans, fight, missouri, reproductive, laws, pregnant, health, #waronwomen, antichoice, legal, alabamas, mortality, roe, unconstitutional, #abortionishealthcare	trump, khashoggi, saudi, murder, jamal, prince, saudis, crown, arabia, journalist, kushner, cia, mbs, khashoggis, house, cover, trumps, ordered, president, jared, murdered, #khashoggi, resident, white, #justiceforkhashoggi, money, donald, killing, coverup, body

Table E.5: Top terms of 50 topics by LDA. For each topic, 20 terms with the highest term-topic probability are listed. The sum of topical probability over all documents (Volumn) is also reported. We fitted 50 topics in order to match the number of flocks for comparative purpose.

Topic	Volumn	Top terms
1	1.45%	bernie, sanders, people, #bernie2020, trump, biden, support, vote, campaign, warren, candidate, party, democratic, time, medicare, progressive, corporate, money, hillary, 2020
2	1.66%	trump, president, house, news, trumps, report, white, 2020, democrats, donald, mueller, fox, border, watch, cnn, calls, rep, biden, gop, democratic
3	1.26%	vote, trump, time, make, people, follow, election, house, day, call, love, #voteblue, voting, support, women, #trumpresign, gop, today, state, senate
4	3.02%	trump, president, white, people, mueller, donald, trumps, america, house, gop, republicans, country, time, russia, american, racist, report, russian, vote, republican
5	2.57%	trump, democrats, president, obama, america, illegal, people, democrat, border, media, american, left, americans, dems, country, hillary, clinton, party, black, time
6	1.57%	vote, trump, democrats, border, people, caravan, #maga, president, america, kavanaugh, red, florida, democrat, election, voting, republican, country, time, american, make
7	0.71%	#maga, #tcot, #trump, latest, daily, #pjnet, #p2, #news, american, #gop, #foxnews, #resist, #ccot, #cnn, #trump2020, news, #kag, trump, #democrats, #uniteblue
8	0.90%	canada, france, #cdnpoli, paris, trudeau, 2019, canadian, #flowerreport, french, pr, st, german, germany, years, europe, 2018, 15, cest, 20, fran
9	3.31%	people, good, point, political, thread, problem, policy, things, work, interesting, social, question, read, time, politics, great, evidence, wrong, lot, thing
10	1.86%	president, trump, great, people, america, country, democrats, border, news, years, american, media, fake, united, time, good, china, back, today, states
11	1.07%	trump, #qanon, fbi, media, clinton, state, epstein, news, #wwg1wga, mueller, people, obama, video, watch, time, truth, russia, deep, hillary, cia
12	1.52%	trump, president, america, god, people, #maga, good, country, patriots, great, love, time, day, back, american, bless, dems, democrats, vote, #trump2020
13	1.95%	white, people, world, america, country, american, jews, jewish, left, black, hate, whites, immigration, children, women, media, israel, europe, race, years
14	2.22%	media, people, twitter, left, news, trump, story, video, political, tweet, racist, jews, american, kids, conservative, white, women, hate, antisemitism, speech
15	2.09%	movie, film, #gameofthrones, show, episode, season, #oscars, watch, tv, series, movies, star, night, love, thrones, king, years, oscar, films, john
16	2.05%	data, news, facebook, tech, work, media, google, internet, security, digital, online, privacy, social, journalism, companies, company, content, technology, information, users
17	2.34%	court, law, house, case, report, public, federal, state, today, government, president, investigation, story, judge, legal, told, justice, committee, officials, office

Table E.5: Top terms of 50 topics by LDA (continued).

Topic	Proportion	Top terms
18	2.18%	police, people, killed, shooting, woman, city, fire, years, school, found, live,california, home, shot, death, arrested, video, family, died, dead
19	2.14%	vote, voters, election, trump, 2020, state, campaign, house, democrats, senate, democratic, gop, president, candidates, republicans, candidate, party, voting, republican, dems
20	4.72%	good, people, time, thing, ve, lot, great, things, love, bad, make, tweet, feel,story, work, yeah, pretty, years, ll, back
21	1.36%	game, baseball, season, team, year, games, sox, today, series, red, back, league, time, mlb, top, players, home, day, yankees, hit, teams
22	1.08%	south, nie, vir, africa, anc, cape, #sabcnews, #statecaptureinquiry, ek, jou, jy, african, oor, minister, 2019, sy, land, zuma, ramaphosa, gauteng
23	3.14%	lol, love, shit, good, ass, lmao, time, people, black, ve, fuck, back, yall, bitch, girl, ain, gonna, wanna, damn, day
24	0.75%	political, media, rich, people, corporate, american, white, nation, americans, democracy, corp, owned, support, racist, economic, real, wealthy, progressive, power, time
25	2.11%	brexit, uk, deal, party, vote, labour, people, leave, mps, government, #brexit,british, parliament, referendum, boris, pm, britain, tory, remain, election
26	1.97%	climate, change, energy, world, global, water, power, #climatechange, action, emissions, oil, gas, report, carbon, solar, future, coal, science, environmental, clean
27	2.66%	students, school, schools, education, student, teachers, learn, research, college, work, public, learning, read, teacher, program, kids, great, high, year, community
28	1.55%	gun, military, veterans, service, army, day, national, guns, honor, veteran, training, air, force, american, great, navy, shooting, today, 2019, defense
29	0.82%	published, news, 2019, gay, march, december, january, meghan, february, fresh, october, entertainment, star, #politics, april, harry, today, #hollywood, ||, reveals
30	2.78%	people, life, make, time, things, women, work, good, day, world, feel, money, love, learn, find, give, start, mind, real, change
31	2.33%	israel, israeli, iran, saudi, palestinian, syria, military, gaza, palestinians, forces, rights, killed, regime, state, world, attack, foreign, international, government, security
32	0.77%	support, pa, share, retweet, ser, za, gracias, great, good, ve, visit, music, #gofundme, hoy, espa, #tenisxespn, books, ideas, #crowdfund, #crowdfunding
33	0.92%	follow, rt, patriots, retweet, back, ride, great, train, #maga, patriot, followers, trump, dm, twitter, love, retweeted, awesome, god, fb, ifb
34	1.81%	health, abortion, patients, medical, cancer, drug, women, research, study, people, learn, risk, treatment, disease, healthcare, dr, hiv, patient, doctors, hospital
35	4.14%	today, week, great, day, live, join, watch, tonight, love, happy, tomorrow, show, morning, time, check, monday, listen, good, amazing, year
36	3.03%	day, good, love, time, happy, home, morning, life, back, food, today, beautiful,night, baby, eat, years, make, house, sweet, cat

Table E.5: Top terms of 50 topics by LDA (continued).

Topic	Proportion	Top terms
37	3.54%	shit, fuck, good, lol, fucking, yeah, game, time, gonna, people, make, video, guy, bad, made, dude, guys, thing, lmao, cool
38	1.90%	music, album, video, song, watch, listen, love, show, back, live, songs, years, rap, 2019, tour, time, #grammys, playlist, favorite, birthday
39	2.56%	book, read, books, reading, writing, story, history, review, work, world, author, love, american, poetry, writers, art, piece, write, life, 2019
40	3.22%	people, good, time, make, stop, back, twitter, thing, lol, shit, hate, wrong, stupid, love, money, give, person, bad, tweet, true
41	2.48%	god, church, jesus, life, lord, catholic, love, christ, pray, faith, st, christian, world, today, pope, prayer, holy, day, bible, people
42	1.91%	china, trade, market, tax, year, economy, economic, growth, global, business, billion, chinese, fed, bank, markets, financial, deal, years, world, oil
43	2.57%	game, season, team, nfl, week, football, play, coach, win, nba, players, year, games, back, teams, player, big, bowl, top, time
44	0.97%	number, dm, hear, message, confirmation, team, flight, link, happy, share, direct, twitter, travel, check, email, send, assist, assistance, apologize, time
45	0.61%	thread, unroll, read, gp, find, good, day, asked, support, share, talk, interesting, enjoy, #democrats, #demswork4usa, #winblue, #progressives, #healthcare, #yeswecan, tweets
46	2.49%	black, people, white, women, folks, trans, racism, work, racist, woman, history,community, violence, police, support, color, justice, years, rights, today
47	1.23%	match, tennis, set, win, round, open, 64, top, 63, title, world, beat, 1st, final, back, cup, court, 62, play, 2019
48	0.69%	read, click, history, black, story, school, people, free, learned, happened, militia, mind, book, made, knew, world, artists, written, 2a, painted
49	2.95%	people, today, rights, women, work, families, make, workers, support, health, country, state, children, violence, act, fight, time, congress, join, working
50	1.08%	wisconsin, road, traffic, county, weather, morning, vehicle, milwaukee, lane, crash, closed, drive, today, area, north, left, snow, rain, channel, update

Abadir, Karim M., Walter Distaso, and Filip Zikes. 2014. Design-free estimation of variance matrices. *Journal of Econometrics* 181(2):165–180.

Abbe, E., A. S. Bandeira, and G. Hall. 2016. Exact Recovery in the Stochastic Block Model. *IEEE Transactions on Information Theory* 62(1):471–487.

Abbe, Emmanuel. 2017. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research* 18(1):6446–6531.

Abbe, Emmanuel, Jianqing Fan, Kaizheng Wang, Yiqiao Zhong, et al. 2020. Entrywise eigenvector analysis of random matrices with low expected rank. *The Annals of Statistics* 48(3):1452–1474.

Abbe, Emmanuel, and Colin Sandon. 2015. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. *arXiv:1503.00609 [cs, math]*.

Achlioptas, Dimitris, and Frank McSherry. 2007. Fast computation of low-rank matrix approximations. *Journal of the ACM (JACM)* 54(2):9–es.

Ahle, Thomas D. 2021. Sharp and Simple Bounds for the raw Moments of the Binomial and Poisson Distributions. *arXiv:2103.17027 [math, stat]*. ArXiv: 2103.17027.

Aiello, Luca Maria, Alain Barrat, Rossano Schifanella, Ciro Cattuto, Benjamin Markines, and Filippo Menczer. 2012. Friendship prediction and homophily in social media. *ACM Transactions on the Web (TWEB)* 6(2):1–33.

Airoldi, Edoardo M, and Jonathan M Bischof. 2016. Improving and evaluating topic models and other models of text. *Journal of the American Statistical Association* 111(516):1381–1403.

Airoldi, Edoardo M, David M Blei, Stephen E Fienberg, and Eric P Xing. 2008. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research* 9(Sep): 1981–2014.

Ajanki, Oskari H., László Erdős, and Torben Krüger. 2017. Universality for general Wigner-type matrices. *Probability Theory and Related Fields* 169(3):667–727.

Al Zamal, Faiyaz, Wendy Liu, and Derek Ruths. 2012. Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 6.

Alamgir, M., and U. von Luxburg. 2010. Multi-agent random walks for local clustering. 18–27. Max-Planck-Gesellschaft, Piscataway, NJ, USA: IEEE.

Alex, Bloemendal, László Erdős, Antti Knowles, Horng-Tzer Yau, and Jun Yin. 2014. Isotropic local laws for sample covariance and generalized Wigner matrices. *Electronic Journal of Probability* 19.

Amini, Arash A, Aiyou Chen, Peter J Bickel, Elizaveta Levina, et al. 2013. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics* 41(4):2097–2122.

Amini, Arash A, and Martin J Wainwright. 2009. High-dimensional analysis of semidefinite relaxations for sparse principal components. *The Annals of Statistics* 37(5B):2877–2921.

Andersen, Reid, Fan Chung, and Kevin Lang. 2006. Local graph partitioning using PageRank vectors. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, 475–486. FOCS '06, Washington, DC, USA: IEEE Computer Society.

Andersen, Reid, and Kevin J. Lang. 2006. Communities from seed sets. In *Proceedings of the 15th International Conference on World Wide Web*, 223–232. WWW '06, New York, NY, USA: ACM.

Andersen, Reid, and Yuval Peres. 2009. Finding sparse cuts locally using evolving sets. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, 235–244. STOC '09, New York, NY, USA: ACM.

Anstead, Nick, and Ben O'Loughlin. 2014. Social media analysis and public opinion: The 2010 UK general election. *Journal of Computer-Mediated Communication* 20(2): 204–220.

Arlot, Sylvain, and Alain Celisse. 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys* 4(none):40–79.

Aslam, Salman. 2021. Twitter by the numbers: Stats, demographics & fun facts.

Athreya, Avanti, Donniell E Fishkind, Minh Tang, Carey E Priebe, Youngser Park, Joshua T Vogelstein, Keith Levin, Vince Lyzinski, and Yichen Qin. 2017. Statistical inference on random dot product graphs: a survey. *The Journal of Machine Learning Research* 18(1):8393–8484.

Baglama, James, and Lothar Reichel. 2005. Augmented implicitly restarted lanczos bidiagonalization methods. *SIAM Journal on Scientific Computing* 27(1):19–42.

Barabási, Albert-László, and Réka Albert. 1999. Emergence of scaling in random networks. *Science* 286(5439):509–512.

Barberá, Pablo, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science* 26(10):1531–1542.

Barberá, Pablo, and Gonzalo Rivero. 2015. Understanding the political representativeness of Twitter users. *Social Science Computer Review* 33(6):712–729.

Baron, Maayan, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, et al. 2016. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Systems* 3(4):346–360.

Bartlett, Maurice S. 1947. The use of transformations. *Biometrics* 3(1):39–52.

Becker, Hila, Mor Naaman, and Luis Gravano. 2010. Learning similarity metrics for event identification in social media. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, 291–300.

Belabbas, Mohamed-Ali, and Patrick J. Wolfe. 2009. Spectral methods in machine learning and new strategies for very large datasets. *Proceedings of the National Academy of Sciences* 106(2):369–374. `https://www.pnas.org/content/106/2/369.full.pdf`.

Belkin, Mikhail, and Partha Niyogi. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation* 15(6):1373–1396.

Bell, Anthony J, and Terrence J Sejnowski. 1997. The "independent components" of natural scenes are edge filters. *Vision Research* 37(23):3327–3338.

Benaych-Georges, Florent, Charles Bordenave, and Antti Knowles. 2019. Largest eigenvalues of sparse inhomogeneous Erdős–Rényi graphs. *Annals of Probability* 47(3):1653–1676.

———. 2020. Spectral radii of sparse random matrices. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* 56(3):2141–2161.

Benjamini, Yoav, and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57(1):289–300.

Bennett, W Lance, and Alexandra Segerberg. 2013. *The logic of connective action: Digital media and the personalization of contentious politics*. Cambridge University Press.

Berkhin, Pavel. 2006. Bookmark-coloring algorithm for personalized PageRank computing. *Internet Mathematics* 3(1):41–62.

Bernaards, Coen A, and Robert I Jennrich. 2005. Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. *Educational and Psychological Measurement* 65(5):676–696.

Berthet, Quentin, and Philippe Rigollet. 2013. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics* 41(4):1780–1815.

Bhattacharyya, Sharmodeep, and Peter J. Bickel. 2015. Subsampling bootstrap of count features of networks. *The Annals of Statistics* 43(6):2384–2411.

Bickel, Peter J., and Purnamrita Sarkar. 2016. Hypothesis testing for automated community detection in networks. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 78(1):253–273.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3(Jan):993–1022.

Blumer, Herbert. 1948. Public opinion and public opinion polling. *American Sociological Review* 13(5):542–549.

Bode, Leticia, Alexander Hanna, Junghwan Yang, and Dhavan V Shah. 2015. Candidate networks, citizen clusters, and political expression: Strategic hashtag use in the 2010 midterms. *The Annals of the American Academy of Political and Social Science* 659(1):149–165.

Bollen, Johan, Huina Mao, and Alberto Pepe. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Fifth International AAAI Conference on Weblogs and Social Media*.

Boot, Arnout B, Erik Tjong Kim Sang, Katinka Dijkstra, and Rolf A Zwaan. 2019. How character limit affects language usage in tweets. *Palgrave Communications* 5(1): 1–13.

Bordenave, C., M. Lelarge, and L. Massoulié. 2015. Non-backtracking Spectrum of Random Graphs: Community Detection and Non-regular Ramanujan Graphs. In *56th Annual Symposium on Foundations of Computer Science*, 1347–1357.

Boucheron, S., G. Lugosi, and P. Massart. 2013a. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford.

Boucheron, Stéphane, Gábor Lugosi, and Pascal Massart. 2013b. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Boyd, Danah. 2010. Social network sites as networked publics: Affordances, dynamics, and implications. In *A Networked Self*, 47–66. Routledge.

Brémaud, Pierre. 2013. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, vol. 31. Springer Science & Business Media.

Brin, Sergey, and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems* 30(1-7):107–117.

Cai, Diana, Trevor Campbell, and Tamara Broderick. 2016. Edge-exchangeable graphs and sparsity. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 4249–4257.

Cai, T Tony, Zongming Ma, and Yihong Wu. 2013. Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics* 41(6):3074–3110.

Candès, Emmanuel J, and Benjamin Recht. 2009. Exact matrix completion via convex optimization. *Foundations of Computational mathematics* 9(6):717.

Canny, John. 2004. GaP: a factor model for discrete data. In *Proceedings of the 27th annual international acm sigir conference on research and development in information retrieval*, 122–129.

Carroll, John B. 1953. An analytical solution for approximating simple structure in factor analysis. *Psychometrika* 18(1):23–38.

Chakrabarty, Arijit, Sukrit Chakraborty, and Rajat Subhra Hazra. 2020. Eigenvalues Outside the Bulk of Inhomogeneous Erdős–Rényi Random Graphs. *Journal of Statistical Physics* 181(5):1746–1780.

Chatterjee, Sourav. 2015. Matrix estimation by Universal Singular Value Thresholding. *The Annals of Statistics* 43(1):177–214.

Chen, Aiyou, and Peter J Bickel. 2006. Efficient independent component analysis. *The Annals of Statistics* 34(6):2825–2855.

Chen, Fan, and Karl Rohe. 2020. A new basis for sparse PCA. `2007.00596`.

Chen, Fan, Yini Zhang, and Karl Rohe. 2020a. Targeted sampling from massive block model graphs with personalized PageRank. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82(1):99–126.

Chen, Kehui, and Jing Lei. 2018. Network Cross-Validation for Determining the Number of Communities in Network Data. *Journal of the American Statistical Association* 113(521):241–251.

Chen, Yuxin, Yuejie Chi, Jianqing Fan, and Cong Ma. 2020b. Spectral methods for data science: A statistical perspective. *arXiv preprint arXiv:2012.08496*.

Chen, Yuxin, Jianqing Fan, Cong Ma, and Kaizheng Wang. 2019. Spectral method and regularized mle are both optimal for top-k ranking. *The Annals of Statistics* 47(4): 2204–2235.

Chierichetti, Flavio, Alessandro Panconesi, and Andrea Vattani. 2018. The equivalence of single-topic and lda topic reconstruction.

Chung, Fan, and Linyuan Lu. 2006. *Complex graphs and networks*. CBMS Regional Conference Series in Mathematics 92, American Mathematical Society.

Cody, Emily M, Andrew J Reagan, Lewis Mitchell, Peter Sheridan Dodds, and Christopher M Danforth. 2015. Climate change sentiment on Twitter: An unsolicited public opinion poll. *PloS ONE* 10(8).

Colleoni, Elanor, Alessandro Rozza, and Adam Arvidsson. 2014. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication* 64(2):317–332.

Comon, Pierre. 1994. Independent component analysis, a new concept? *Signal Processing* 36(3):287–314.

Comon, Pierre, and Christian Jutten. 2010. *Handbook of Blind Source Separation: Independent component analysis and applications*. Oxford, UK: Academic Press.

Conover, Michael D, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on Twitter. In *Fifth International AAAI Conference on Weblogs and Social Media*.

Conway, Bethany A, Kate Kenski, and Di Wang. 2015. The rise of Twitter in the political campaign: Searching for intermedia agenda-setting effects in the presidential primary. *Journal of Computer-Mediated Communication* 20(4):363–380.

Couldry, Nick. 2012. *Media, society, world: Social theory and digital media practice*. Polity.

d'Aspremont, Alexandre, Laurent El Ghaoui, Michael I Jordan, and Gert R G Lanckriet. 2007. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review* 49(3):434–448.

Davis, Chandler, and William Morton Kahan. 1970. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis* 7(1):1–46.

De Choudhury, Munmun. 2011. Tie formation on Twitter: Homophily and structure of egocentric networks. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 465–470. IEEE.

Dobriban, Edgar. 2020. Permutation methods for factor analysis and pca. *The Annals of Statistics* 48(5):2824–2847.

Dobriban, Edgar, and Art B Owen. 2019. Deterministic parallel analysis: an improved method for selecting factors and principal components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81(1):163–183.

Donoho, David L. 1995. De-noising by soft-thresholding. *IEEE Transactions on Information Theory* 41(3):613–627.

Donoho, David L, Matan Gavish, and Iain M Johnstone. 2018. Optimal shrinkage of eigenvalues in the spiked covariance model. *The Annals of Statistics* 46(4):1742.

Dubois, Elizabeth, and Devin Gaffney. 2014. The multiple facets of influence: Identifying political influentials and opinion leaders on Twitter. *American Behavioral Scientist* 58(10):1260–1277.

Dumitriu, Ioana, and Yizhe Zhu. 2019. Sparse general Wigner-type matrices: Local law and eigenvector delocalization. *Journal of Mathematical Physics* 60(2):023301.

Durrett, Rick. 2019. *Probability: Theory and examples*. 5th ed. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Entman, Robert M. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication* 43(4):51–58.

Erdős, Paul, and Alfréd Rényi. 1960. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci* 5(1):17–60.

Erdős, László, Antti Knowles, Horng-Tzer Yau, and Jun Yin. 2012. Spectral Statistics of Erdős-Rényi Graphs II: Eigenvalue Spacing and the Extreme Eigenvalues. *Communications in Mathematical Physics* 314(3):587–640.

———. 2013. The local semicircle law for a general class of random matrices. *Electronic Journal of Probability* 18.

Field, David J. 1987. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A* 4(12): 2379–2394.

———. 1994. What is the goal of sensory coding? *Neural Computation* 6(4):559–601.

Firth, John R. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*.

Fisher, Ronald A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7(2):179–188.

Fortunato, Santo. 2010. Community detection in graphs. *Physics Reports* 486(3-5): 75–174.

Freelon, Deen, Charlton McIlwain, and Meredith Clark. 2018. Quantifying the power and consequences of social media protest. *New Media & Society* 20(3):990–1011.

Frobenius, Georg Ferdinand, Ferdinand Georg Frobenius, Ferdinand Georg Frobenius, Ferdinand Georg Frobenius, and Germany Mathematician. 1912. *Über matrizen aus nicht negativen elementen*. Königliche Akademie der Wissenschaften.

Gallivan, Kyle A, and PA Absil. 2010. Note on the convex hull of the Stiefel manifold. *Florida State University*.

Gallup, George, and Saul Forbes Rae. 1940. The pulse of democracy: The public-opinion poll and how it works.

Gao, Chao, Zongming Ma, Anderson Y Zhang, Harrison H Zhou, et al. 2018. Community detection in degree-corrected block models. *The Annals of Statistics* 46(5): 2153–2185.

Gataric, Milana, Tengyao Wang, and Richard J. Samworth. 2020. Sparse principal component analysis via axis-aligned random projections. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82(2):329–359.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. Text as data. *Journal of Economic Literature* 57(3):535–74.

Georgiev, Pando, Fabian Theis, and Andrzej Cichocki. 2005. Sparse component analysis and blind source separation of underdetermined mixtures. *IEEE Transactions on Neural Networks* 16(4):992–996.

Gharan, Shayan Oveis, and Luca Trevisan. 2012. Approximating the expansion profile and almost optimal local graph clustering. In *Proceedings of the 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, 187–196. FOCS '12, Washington, DC, USA: IEEE Computer Society.

Ghoshal, Gourab, and Albert-László Barabási. 2011. Ranking stability and super-stable nodes in complex networks. *Nature Communications* 2:394.

Gleich, David F. 2015. PageRank beyond the web. *SIAM Review* 57(3):321–363.

Golder, Scott A, and Michael W Macy. 2014. Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology* 40:129–152.

Golder, Scott A, and Sarita Yardi. 2010. Structural predictors of tie formation in twitter: Transitivity and mutuality. In *Second International Conference on Social Computing*, 88–95. IEEE.

Grabowicz, Przemyslaw A, José J Ramasco, Esteban Moro, Josep M Pujol, and Victor M Eguiluz. 2012. Social features of online networks: The strength of intermediary ties in online social media. *PloS ONE* 7(1).

Green, Alden, and Cosma Rohilla Shalizi. 2017. Bootstrapping Exchangeable Random Graphs. *arXiv:1711.00813 [stat]*.

Gregor, Karol, and Yann LeCun. 2010. Learning fast approximations of sparse coding. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 399–406. ICML'10, Madison, WI, USA: Omnipress.

Grimmer, Justin, and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21(3):267–297.

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on twitter during the 2016 us presidential election. *Science* 363(6425):374–378.

Groves, Robert M, and Emilia Peytcheva. 2008. The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly* 72(2):167–189.

Gupta, Pankaj, Ashish Goel, Jimmy Lin, Aneesh Sharma, Dong Wang, and Reza Zadeh. 2013. WTF: The who to follow service at twitter. In *Proceedings of the 22nd International Conference on World Wide Web*, 505–514. ACM.

Hand, David J, and Niall M Adams. 2014. Data mining. *Wiley StatsRef: Statistics Reference Online* 1–7.

Haveliwala, Taher H. 2003. Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering* 15(4): 784–796.

Heckathorn, Douglas D, and Christopher J Cameron. 2017. Network sampling: From snowball and multiplicity to respondent-driven sampling. *Annual Review of Sociology* 43:101–119.

Herbst, Susan. 2001. Public opinion infrastructures: Meanings, measures, media. *Political Communication* 18(4):451–464.

Hoff, Peter. 2008. Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems*, ed. J. Platt, D. Koller, Y. Singer, and S. Roweis, vol. 20. Curran Associates, Inc.

Holland, Paul W, Kathryn Blackmond Laskey, and Samuel Leinhardt. 1983a. Stochastic blockmodels: First steps. *Social Networks* 5(2):109–137.

Holland, Paul W., Kathryn Blackmond Laskey, and Samuel Leinhardt. 1983b. Stochastic blockmodels: First steps. *Social Networks* 5(2):109–137.

Hong, David, Yue Sheng, and Edgar Dobriban. 2020. Selecting the number of components in pca via random signflips. *arXiv preprint arXiv:2012.02985*.

Horn, John L. 1965. A rationale and test for the number of factors in factor analysis. *Psychometrika* 30(2):179–185.

Horn, Roger A, and Charles R Johnson. 1985. *Matrix analysis*. Cambridge, UK: Cambridge University Press.

Hotelling, Harold. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24(6):417.

Hu, Yuheng, Ajita John, Dorée Duncan Seligmann, and Fei Wang. 2012. What were the tweets about? Topical associations between public events and Twitter feeds. In *Sixth International AAAI Conference on Weblogs and Social Media*.

Hu, Zhenfang, Gang Pan, Yueming Wang, and Zhaohui Wu. 2016. Sparse principal component analysis via rotation and truncation. *IEEE Transactions on Neural Networks and Learning Systems* 27(4):875–890.

Hwang, Jong Yun, Ji Oon Lee, and Wooseok Yang. 2020. Local law and Tracy–Widom limit for sparse stochastic block models. *Bernoulli* 26(3):2400–2435.

Hyvarinen, Aapo. 1999. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks* 10(3):626–634.

Hyvärinen, Aapo, and Erkki Oja. 2000. Independent component analysis: algorithms and applications. *Neural Networks* 13(4-5):411–430.

Jain, Anil K. 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* 31(8):651–666.

Jeffers, JNR. 1967. Two case studies in the application of principal component analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 16(3):225–236.

Jeh, Glen, and Jennifer Widom. 2003. Scaling personalized web search. In *Proceedings of the 12th International Conference on World Wide Web*, 271–279. WWW '03, New York, NY, USA: ACM.

Jennrich, Robert I. 2001. A simple general procedure for orthogonal rotation. *Psychometrika* 66(2):289–306.

Jentsch, Carsten, Eun Ryung Lee, and Enno Mammen. 2020. Poisson reduced-rank models with an application to political text data. *Biometrika*.

Jin, Jiashun, Zheng Tracy Ke, Shengming Luo, and Minzhe Wang. 2020. Estimating the number of communities by Stepwise Goodness-of-fit. *arXiv:2009.09177 [math, stat]*.

Johnstone, Iain M, and Arthur Yu Lu. 2009. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association* 104(486):682–693.

Jolliffe, Ian T. 1995. Rotation of principal components: choice of normalization constraints. *Journal of Applied Statistics* 22(1):29–35.

Jolliffe, Ian T, Nickolay T Trendafilov, and Mudassir Uddin. 2003. A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics* 12(3):531–547.

Joseph, Antony, and Bin Yu. 2016. Impact of regularization on spectral clustering. *The Annals of Statistics* 44(4):1765–1791.

Journée, Michel, Yurii Nesterov, Peter Richtárik, and Rodolphe Sepulchre. 2010. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research* 11:517–553.

Jungherr, Andreas, Oliver Posegga, and Jisun An. 2019. Discursive power in contemporary media systems: A comparative framework. *The International Journal of Press/Politics* 24(4):404–425.

Kaiser, Henry F. 1958. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* 23(3):187–200.

———. 1960. The application of electronic computers to factor analysis. *Educational and Psychological Measurement* 20(1):141–151.

Karrer, Brian, and M. E. J. Newman. 2011a. Stochastic blockmodels and community structure in networks. *Physical Review E* 83(1):016107.

Karrer, Brian, and Mark EJ Newman. 2011b. Stochastic blockmodels and community structure in networks. *Physical Review E* 83(1):016107.

Karypis, George, and Vipin Kumar. 1998. Multilevelk-way partitioning scheme for irregular graphs. *Journal of Parallel and Distributed computing* 48(1):96–129.

Katz, Daniel. 1960. The functional approach to the study of attitudes. *Public Opinion Quarterly* 24(2):163–204.

Keshavan, Raghunandan H, Andrea Montanari, and Sewoong Oh. 2010. Matrix completion from noisy entries. *The Journal of Machine Learning Research* 11:2057–2078.

Khanna, Rajiv, Ethan Elenberg, Alexandros G Dimakis, and Sahand Negahban. 2017. On approximation guarantees for greedy low rank optimization. *arXiv preprint arXiv:1703.02721*.

Kim, Yonghwan, Youngju Kim, Joong Suk Lee, Jeyoung Oh, and Na Yeon Lee. 2015. Tweeting the public: journalists' Twitter use, attitudes toward the public's tweets, and the relationship with the public. *Information, Communication & Society* 18(4): 443–458.

Kim, Yoonsang, Jidong Huang, and Sherry Emery. 2016. Garbage in, garbage out: data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection. *Journal of Medical Internet Research* 18(2):e41.

Kloumann, Isabel M., Johan Ugander, and Jon Kleinberg. 2017. Block models and personalized PageRank. *Proceedings of the National Academy of Sciences* 114(1):33–38.

Kosinski, Michal, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110(15):5802–5805.

Kruskal, Joseph B. 1964a. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29(1):1–27.

———. 1964b. Nonmetric multidimensional scaling: A numerical method. *Psychometrika* 29(2):115–129.

Krzakala, Florent, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. 2013. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences* 110(52):20935–20940.

Kuhn, Harold W. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* 2(1-2):83–97.

Lam, Clifford. 2016. Nonparametric eigenvalue-regularized precision or covariance matrix estimator. *The Annals of Statistics* 44(3):928–953.

Lasorsa, Dominic L, Seth C Lewis, and Avery E Holton. 2012. Normalizing Twitter: Journalism practice in an emerging communication space. *Journalism Studies* 13(1): 19–36.

Le, Can M., and Elizaveta Levina. 2019. Estimating the number of communities in networks by spectral methods. *arXiv:1507.00827 [cs, math, stat]*.

Le, Can M, Elizaveta Levina, and Roman Vershynin. 2016. Optimization via low-rank approximation for community detection in networks. *The Annals of Statistics* 44(1): 373–400.

Le, Can M., Elizaveta Levina, and Roman Vershynin. 2017. Concentration and regularization of random graphs. *Random Structures & Algorithms* 51(3):538–561.

Lee, Honglak, Alexis Battle, Rajat Raina, and Andrew Y Ng. 2006. Efficient sparse coding algorithms. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, 801–808. NIPS'06, Cambridge, MA, USA: MIT Press.

Lei, Jing. 2016. A goodness-of-fit test for stochastic block models. *The Annals of Statistics* 44(1):401–424.

Leskovec, Jure, Jon Kleinberg, and Christos Faloutsos. 2007. Graph evolution: Densification and shrinking diameters. *ACM transactions on Knowledge Discovery from Data* (*TKDD*) 1(1):2–es.

Levin, Keith, and Elizaveta Levina. 2019. Bootstrapping Networks with Latent Space Structure. *arXiv:1907.10821* [*math, stat*].

Li, Tianxi, Elizaveta Levina, and Ji Zhu. 2020. Network cross-validation by edge sampling. *Biometrika* 107(2):257–276.

Liao, Chung-Shou, Kanghao Lu, Michael Baym, Rohit Singh, and Bonnie Berger. 2009. IsoRankN: Spectral methods for global alignment of multiple protein networks. *Bioinformatics* 25(12):i253–i258.

Lin, Qiaohui, Robert Lunde, and Purnamrita Sarkar. 2020a. Higher-Order Correct Multiplier Bootstraps for Count Functionals of Networks. *arXiv:2009.06170* [*math, stat*].

———. 2020b. On the Theoretical Properties of the Network Jackknife. In *International Conference on Machine Learning*, 6105–6115. PMLR.

Lin, Yu-Ru, Brian Keegan, Drew Margolin, and David Lazer. 2014. Rising tides or rising stars?: Dynamics of shared attention on Twitter during media events. *PloS ONE* 9(5):e94093.

Lin, Yu-Ru, Drew Margolin, Brian Keegan, and David Lazer. 2013. Voices of victory: A computational focus group framework for tracking opinion shift in real time. In *Proceedings of the 22nd International Conference on World Wide Web*, 737–748. ACM.

Litt, Eden. 2012. Knock, knock. Who's there? The imagined audience. *Journal of Broadcasting & Electronic Media* 56(3):330–345.

Liu, Yan, Zhiqiang Hou, Zhigang Yao, Zhidong Bai, Jiang Hu, and Shurong Zheng. 2019. Community Detection Based on the L_∞ convergence of eigenvectors in DCBM. *arXiv:1906.06713* [*math, stat*].

Lofgren, Peter, Siddhartha Banerjee, and Ashish Goel. 2016. Personalized PageRank estimation and search: A bidirectional approach. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, 163–172. ACM.

Lovász, László. 1993. Random walks on graphs: A survey. *Combinatorics, Paul erdos is eighty* 2(1):1–46.

Lu, Xin, Jens Malmros, Fredrik Liljeros, and Tom Britton. 2013. Respondent-driven sampling on directed networks. *Electronic Journal of Statistics* 7:292–322.

Lunde, Robert, and Purnamrita Sarkar. 2019. Subsampling Sparse Graphons Under Minimal Assumptions. *arXiv:1907.12528* [*math, stat*].

Ma, Shujie, Liangjun Su, and Yichong Zhang. 2019. Determining the Number of Communities in Degree-corrected Stochastic Block Models. *arXiv:1809.01028* [*stat*].

Mackey, Lester. 2008. Deflation methods for sparse PCA. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, 1017–1024. NIPS'08, Red Hook, NY, USA: Curran Associates Inc.

Macropol, Kathy, Tolga Can, and Ambuj K. Singh. 2009. RRW: Repeated random walks on genome-scale protein networks for local cluster discovery. *BMC Bioinformatics* 10(1):283.

Manning, Christopher D, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge university press.

Mardia, Kanti V. 1970. Measures of multivariate skewness and kurtosis with applications. *Biometrika* 57(3):519–530.

Mcauliffe, Jon D, and David M Blei. 2008. Supervised topic models. In *Advances in neural information processing systems*, 121–128.

McGregor, Shannon C. 2019. Social media as public opinion: How journalists use social media to represent public opinion. *Journalism* 20(8):1070–1086.

———. 2020. "taking the temperature of the room" How political campaigns use social media to understand and represent public opinion. *Public Opinion Quarterly* 84(S1):236–256.

McPherson, Miller, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27(1):415–444.

McSherry, Frank. 2001. Spectral partitioning of random graphs. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, 529–537. IEEE.

Miettinen, Jari, Sara Taskinen, Klaus Nordhausen, and Hannu Oja. 2015. Fourth moments and independent component analysis. *Statistical Science* 30(3):372–390.

Min, Yong, Tingjun Jiang, Cheng Jin, Qu Li, and Xiaogang Jin. 2019. Endogenetic structure of filter bubble in social networks. *Royal Society Open Science* 6(11):190868.

Moghaddam, Baback, Yair Weiss, and Shai Avidan. 2006. Generalized spectral bounds for sparse LDA. In *Proceedings of the 23rd International Conference on Machine Learning*, 641–648. ICML '06, New York, NY, USA: Association for Computing Machinery.

Montoro, Daniel T, Adam L Haber, Moshe Biton, Vladimir Vinarsky, Brian Lin, Susan E Birket, Feng Yuan, Sijia Chen, Hui Min Leung, Jorge Villoria, et al. 2018. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* 560(7718):319–324.

Mossel, Elchanan, Joe Neeman, and Allan Sly. 2015. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields* 162(3):431–461.

Mosteller, Frederick, and David L Wallace. 1984. A robust hand-calculated bayesian analysis. In *Applied bayesian and classical inference*, 215–228. Springer.

Murtagh, Fionn, and Pierre Legendre. 2014. Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *Journal of Classification* 31(3):274–295.

Naulet, Zacharie, Daniel M. Roy, Ekansh Sharma, and Victor Veitch. 2021. Bootstrap estimators for the tail-index and for the count statistics of graphex processes. *Electronic Journal of Statistics* 15(1):282–325.

Newman, Mark, Albert-Laszlo Barabasi, and Duncan J. Watts, eds. 2006. *The structure and dynamics of networks*. Princeton, NJ, USA: Princeton University Press.

Ng, Andrew, Michael Jordan, and Yair Weiss. 2001. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems* 14:849–856.

Nielsen, Finn Årup. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Nocedal, Jorge, and Stephen Wright. 2006. *Numerical optimization*. 2nd ed. New York, NY, USA: Springer Science & Business Media.

O'Connor, Brendan, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Fourth International AAAI Conference on Weblogs and Social Media*.

Olshausen, Bruno A, and David J Field. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381(6583):607.

O'Mara-Eves, Alison, James Thomas, John McNaught, Makoto Miwa, and Sophia Ananiadou. 2015. Using text mining for study identification in systematic reviews: A systematic review of current approaches. *Systematic Reviews* 4(1):5.

Page, L., S. Brin, R. Motwani, and T. Winograd. 1998. The PageRank citation ranking: Bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference*, 161–172. Brisbane, Australia.

Pan, Jiaqi, Rishabh Bhardwaj, Wei Lu, Hai Leong Chieu, Xinghao Pan, and Ni Yi Puay. 2019. Twitter homophily: Network based prediction of user's occupation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2633–2638.

Pearson, Karl. 1901. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2(11): 559–572.

———. 1905. The problem of the random walk. *Nature* 72(1867):342.

Pennacchiotti, Marco, and Ana-Maria Popescu. 2011. Democrats, republicans and starbucks afficionados: User classification in Twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 430–438.

Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th acm sigkdd international conference on knowledge discovery and data mining*, 701–710.

Perron, Oskar. 1907. Zur theorie der matrices. *Mathematische Annalen* 64(2):248–263.

Phan, Xuan-Hieu, Le-Minh Nguyen, and Susumu Horiguchi. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th International Conference on World Wide Web*, 91–100.

Picard, Richard R., and R. Dennis Cook. 1984. Cross-Validation of Regression Models. *Journal of the American Statistical Association* 79:575–583.

Plasschaert, Lindsey W, Rapolas Žilionis, Rayman Choo-Wing, Virginia Savova, Judith Knehr, Guglielmo Roma, Allon M Klein, and Aron B Jaffe. 2018. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* 560(7718):377–381.

Qin, Tai, and Karl Rohe. 2013. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, 3120–3128. NIPS'13, USA: Curran Associates Inc.

Roberts, Margaret E, Brandon M Stewart, and Dustin Tingley. 2019. `stm`: An R package for structural topic models. *Journal of Statistical Software* 91(1):1–40.

Rohe, Karl, Sourav Chatterjee, and Bin Yu. 2011. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics* 39(4):1878–1915.

Rohe, Karl, Tai Qin, and Bin Yu. 2016. Co-clustering directed graphs to discover asymmetries and directional communities. *Proceedings of the National Academy of Sciences* 113(45):12679–12684.

Rohe, Karl, Jun Tao, Xintian Han, and Norbert Binkiewicz. 2018. A note on quickly sampling a sparse matrix with low rank expectation. *The Journal of Machine Learning Research* 19(1):3040–3052.

Rohe, Karl, and Muzhe Zeng. 2020. Vintage factor analysis with varimax performs statistical inference. *arXiv preprint arXiv:2004.05387*.

Rohe, Karl, et al. 2019. A critical threshold for design effects in network sampling. *The Annals of Statistics* 47(1):556–582.

Rosen, Aliza. 2017. Tweeting made easier.

Salganik, Matthew J. 2019. *Bit by bit: Social research in the digital age*. Princeton University Press.

Samworth, Richard J, and Ming Yuan. 2012. Independent component analysis via nonparametric maximum likelihood estimation. *The Annals of Statistics* 40(6):2973–3002.

Sengupta, Srijan, and Yuguo Chen. 2018. A block model for node popularity in networks with community structure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80(2):365–386.

Shen, Dan, Haipeng Shen, and James Stephen Marron. 2013. Consistency of sparse PCA in high dimension, low sample size contexts. *Journal of Multivariate Analysis* 115:317–333.

Shen, Haipeng, and Jianhua Z Huang. 2008. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis* 99(6): 1015–1034.

Snijders, Tom AB, and Stephen P Borgatti. 1999. Non-parametric standard errors and tests for network statistics. *Connections* 22(2):161–170.

Soshnikov, Alexander. 1999. Universality at the Edge of the Spectrum¶in Wigner Random Matrices. *Communications in Mathematical Physics* 207(3):697–733.

Spielman, Daniel A, and Shang-Hua Teng. 1996. Spectral partitioning works: Planar graphs and finite element meshes. In *Proceedings of 37th Conference on Foundations of Computer Science*, 96–105. IEEE.

———. 2004. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, 81–90. ACM.

Su, Liangjun, Wuyi Wang, and Yichong Zhang. 2019. Strong Consistency of Spectral Clustering for Stochastic Block Models. *arXiv:1710.06191 [stat]*.

Sunstein, Cass R. 2018. *# republic: Divided democracy in the age of social media*. Princeton University Press.

Thompson, Mary E., Lilia L. Ramirez Ramirez, Vyacheslav Lyubchich, and Yulia R. Gel. 2016. Using the bootstrap for statistical inference on random graphs. *Canadian Journal of Statistics* 44(1):3–24.

Thurstone, Louis Leon. 1931. Multiple factor analysis. *Psychological Review* 38(5): 406.

Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Tillmann, Andreas M, and Marc E Pfetsch. 2014. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Transactions on Information Theory* 60(2):1248–1259.

Tracy, Craig A., and Harold Widom. 1994. Level-spacing distributions and the Airy kernel. *Communications in Mathematical Physics* 159(1):151–174.

Tropp, Joel A. 2012. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics* 12(4):389–434.

Troyanskaya, Olga, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics* 17(6):520–525.

Tucker, Joshua A, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: A review of the scientific literature (March 19, 2018)*.

Tufekci, Zeynep. 2013. "not this one" social movements, the attention economy, and microcelebrity networked activism. *American Behavioral Scientist* 57(7):848–870.

Tufekci, Zeynep, and Christopher Wilson. 2012. Social media and the decision to participate in political protest: Observations from Tahrir Square. *Journal of Communication* 62(2):363–379.

Tumasjan, Andranik, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. 2011. Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social Science Computer Review* 29(4):402–418.

Von Luxburg, Ulrike. 2007. A tutorial on spectral clustering. *Statistics and Computing* 17(4):395–416.

Vu, Vincent Q, Juhee Cho, Jing Lei, and Karl Rohe. 2013. Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, 2670–2678. NIPS'13, Red Hook, NY, USA: Curran Associates Inc.

Vu, Vincent Q, and Jing Lei. 2013. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics* 41(6):2905–2947.

Wang, Tengyao, Quentin Berthet, and Richard J Samworth. 2016. Statistical and computational trade-offs in estimation of sparse principal components. *The Annals of Statistics* 44(5):1896–1930.

Wang, Y. X. Rachel, and Peter J. Bickel. 2017. Likelihood-based model selection for stochastic block models. *The Annals of Statistics* 45(2):500–528.

Watts, Duncan J, and Steven H Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature* 393(6684):440.

Witten, Daniela M, Robert Tibshirani, and Trevor Hastie. 2009. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 10(3):515–534.

Wojcik, Stepan, and Adam Hughes. 2019. Sizing up Twitter users. *www.pewresearch.org*.

Wold, Svante. 1978. Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics* 20(4):397–405.

Wu, Shaomei, Jake M Hofman, Winter A Mason, and Duncan J Watts. 2011. Who says what to whom on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, 705–714. ACM.

Xu, Min, Varun Jog, Po-Ling Loh, et al. 2020. Optimal rates for community estimation in the weighted stochastic block model. *The Annals of Statistics* 48(1):183–204.

Yu, Yi, Tengyao Wang, and Richard J Samworth. 2015. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika* 102(2):315–323.

Yuan, Xiao-Tong, and Tong Zhang. 2013. Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research* 14(1):899–925.

Zaller, John. 1994. Positive constructs of public opinion. *Critical Studies in Mass Communication* 11(3):276–87.

Zaller, John R, et al. 1992. *The nature and origins of mass opinion*. Cambridge University Press.

Zhang, Yilin, and Karl Rohe. 2018. Understanding regularized spectral clustering via graph conductance. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 10654–10663. NIPS'18, Red Hook, NY, USA: Curran Associates Inc.

Zhang, Yini, Fan Chen, and Karl Rohe. 2021. Social media public opinion as flocks in a murmuration. In preparation.

Zhao, Yunpeng, Elizaveta Levina, and Ji Zhu. 2011. Community extraction for social networks. *Proceedings of the National Academy of Sciences* 108(18):7321–7326.

Zhu, Mu, and Ali Ghodsi. 2006. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis* 51(2): 918–930.

Zhu, Yaojia, Xiaoran Yan, and Cristopher Moore. 2013. Oriented and degree-generated block models: Generating and inferring communities with inhomogeneous degree distributions. *Journal of Complex Networks* 2(1):1–18.

Zou, Hui, and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2):301–320.

Zou, Hui, Trevor Hastie, and Robert Tibshirani. 2006. Sparse principal component analysis. *Journal of Computational and Graphical Statistics* 15(2):265–286.

Zou, Hui, and Lingzhou Xue. 2018. A selective overview of sparse principal component analysis. *Proceedings of the IEEE* 106(8):1311–1320.