

REPORT DOCUMENTATION PAGE				Form Approved OMB NO. 0704-0188	
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA, 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE (DD-MM-YYYY) 08-08-2023		2. REPORT TYPE Final Report		3. DATES COVERED (From - To) 1-Aug-2019 - 31-Jul-2023	
4. TITLE AND SUBTITLE Final Report: Geometric Factorization Tools for Community Mining				5a. CONTRACT NUMBER W911NF-19-1-0407	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER 611102	
6. AUTHORS				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAMES AND ADDRESSES University of Virginia The Rector and Visitors of the University of Virginia 1001 North Emmet Street Charlottesville, VA 22901 -4195				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES) U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211				10. SPONSOR/MONITOR'S ACRONYM(S) ARO	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S) 75544-HC.35	
12. DISTRIBUTION AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.					
13. SUPPLEMENTARY NOTES The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.					
14. ABSTRACT					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Nikolaos Sidiropoulos
a. REPORT UU	b. ABSTRACT UU	c. THIS PAGE UU			19b. TELEPHONE NUMBER 434-924-3977

RPPR Final Report
as of 05-Sep-2023

Agency Code: 21XD

Proposal Number: 75544HC

Agreement Number: W911NF-19-1-0407

INVESTIGATOR(S):

Name: Xiao Fu
Email: fuxia@oregonstate.edu
Phone Number: 5417373617
Principal: N

Name: Nikolaos Sidiropoulos
Email: nikos@virginia.edu
Phone Number: 4349243977
Principal: Y

Name: Zack Almquist
Email: zalmquis@uw.edu
Phone Number: 2065435882
Principal: N

Organization: **University of Virginia**

Address: The Rector and Visitors of the University of Virginia, Charlottesville, VA 229014195

Country: USA

DUNS Number: 065391526

EIN: 546001796

Report Date: 31-Jul-2023

Date Received: 08-Aug-2023

Final Report for Period Beginning 01-Aug-2019 and Ending 31-Jul-2023

Title: Geometric Factorization Tools for Community Mining

Begin Performance Period: 01-Aug-2019

End Performance Period: 31-Jul-2023

Report Term: 0-Other

Submitted By: Nikolaos Sidiropoulos

Email: nikos@virginia.edu

Phone: (434) 924-3977

Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 10

STEM Participants: 10

Major Goals: Clustering nodes (users, entities) in a social network into different communities is an instrumental task in network analytics. Community detection and association is a core problem in machine learning, data mining, and signal processing. However, there are still many challenges remaining, and fundamental aspects that are poorly understood. Whether it is possible to identify the unknown underlying communities and pin down the association of nodes to these communities under realistic conditions (e.g., from a randomly or systematically sampled subset of links) is *a priori* unclear. Theory and scalable algorithms for extracting large quasi-cliques (densely connected communities) in a graph are sorely missing; and how to infer network typologies with limited prior knowledge has also been an open problem. This project developed a suite of principled analytical and computational tools to advance the state-of-art on several fronts in the broad area of network analytics and graph information retrieval. Its primary goal was the development, analysis, and validation of geometric matrix factorization tools for community mining.

Accomplishments: Please see uploaded PDF document containing a detailed account of our accomplishments under this project.

RPPR Final Report as of 05-Sep-2023

Training Opportunities: (The description below is also included in the uploaded PDF report document)

At Oregon State, PhD student Tri Nguyen, was fully supported by this project through a Graduate Research Assistantship (12 months) until August 2022. Mr. Nguyen is close to his final PhD defense.

At the University of Virginia, a postdoctoral research scientist (Mikael Sorensen) was partially supported (50%) through this project until October 2022, and another postdoctoral research scientist (Aritra Konar) was engaged in research under this project. Several PhD students were also engaged in the early stages of the research (N. Kargas, F. Almutairi, C. Kanatsoulis). These students and postdoctoral researchers have been trained in state-of-art theoretical and algorithmic tools for network analysis, which also have many applications in other areas. Konar is now tenure-track assistant professor at KU Leuven in Belgium, one of the oldest and most prestigious research Universities in Europe.

At University of Washington, several students were involved, as detailed below by Co-PI Almquist:

Aja Sutton, a PhD student in Geography Graduating in 2023. Sutton worked on the REACH data set and the book chapter on probabilistic clustering (community detection) with spatial and demographic data. Dissertation focuses on statistical models for small area estimation. First author on book chapter on small area estimation for demography which provides a clear baseline comparison for using the community detection models for improving small area estimation of migration flows. She worked on this project in the summer and fall quarters of 2022.

Elizabeth Nova, a URM PhD student in Sociology. Nova worked on her masters paper that looks at social media and mainstream media impact on mask wearing (as multiple category events). This provides a standard statistical approach to understanding the relationship between mask wearing (none of the time to all the time) in relationship to social media usage and mainstream media usage. Social media and mainstream media represent overlapping sets and good applications for community detection and mining applications. She worked on this project in the summer and fall quarters of 2022.

Ihsan Kahveci, a 4th year PhD student in Sociology working on his comprehensive exam. Kahveci worked on Social Media data and more recently worked on comparing Facebook run survey to representative based survey on pro-social/anti-social activities. He is showing how you can locate similar enough respondents in each survey to use in propensity matching schemes. This matching of respondents is an ideal application of probabilistic community matches over Blau Spaces (or socio-demographic arrays). Ihsan worked on this project during the previous reporting year, as well as summer and fall of 2022.

June Yang, a PhD student in Sociology. 4th year and working on her dissertation prospectus. Yang worked on a paper relating religious identification (multiple and overlapping categories, e.g. Christian and its various sects) and mask wearing (none to all the time). This provides a ground truth for automated community detection methods which can handle overlapping sets. She worked on this project in the Spring and Fall of 2022.

Larisa Ozeryansky, a PhD student in the Interdisciplinary PhD program. First year and is working on her dissertation prospectus on refugees from the Ukraine/Russia conflict. Ozeryansky helped with developing simulation models for realistic networks of people living unsheltered. These models include the ability to add realistic overlapping category variables (e.g. tent communities) for use in testing new community detection methods. Worked on this project in the summer of 2022.

Results Dissemination: Please see the uploaded PDF report document that also details our dissemination activities via refereed journal and conference papers, including many presentations at the most selective data and graph mining conferences.

RPPR Final Report

as of 05-Sep-2023

Honors and Awards: (The information below is also included in the uploaded PDF report)

Both co-PIs, Zack W. Almquist and Xiao Fu received the NSF CAREER Award, from the SBE Sociology division, and the ECCS program under the Engineering Directorate, respectively.

Both co-PIs, Zack W. Almquist and Xiao Fu were promoted to Associate Professor with tenure at their respective institutions.

Co-PI Almquist was selected to be the Editor-in-Chief for the Journal of Mathematical Sociology (the flagship journal for the ASA section on Mathematical Sociology).

PI Sidiropoulos, received:

1. The Technical Achievement Award of the European Association for Signal Processing (EURASIP). The award reflects distinguished and sustained contributions over the recipients' career.
2. The IEEE SPS Claude Shannon - Harry Nyquist Technical Achievement Award for exemplary contributions to tensor decomposition, beamforming, and spectral analysis, Dec. 2022. This is the highest technical career-level award of the IEEE Signal Processing Society.
3. (Joint with co-PI Fu:) The IEEE Signal Processing Society 2023 Best Paper Award, for the paper , "Learning to Optimize: Training Deep Neural Networks for Interference Management," by Haoran Sun, Xiangyi Chen, Qingjiang Shi, Mingyi Hong, Xiao Fu, and Nicholas D. Sidiropoulos, IEEE Trans. on Signal Processing, Oct. 2018.
4. The IEEE Signal Processing Society 2023 Donald G. Fink Overview Paper Award, for the paper "Tensor Decomposition for Signal Processing and Machine Learning," by Nicholas D. Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E. Papalexakis, Christos Faloutsos, IEEE Transactions on Signal Processing, July 2017.

Protocol Activity Status:

Technology Transfer: Interactions with DoD laboratories included:

Collaborated with Dr. Ananthram Swami (ST) ST USARMY DEVCOM ARL on paper:

M. Sorensen, N. D. Sidiropoulos and A. Swami, "Overlapping Community Detection via Semi-binary Matrix Factorization: Identifiability and Algorithms," IEEE Transactions on Signal Processing, vol. 70, pp. 4321-4336, 2022, doi: 10.1109/TSP.2022.3200215.

Co-authored with Dr. Kevin Chan, DEVCOM ARL the paper:

R.S. Pasricha, U.S. Saini, N.D. Sidiropoulos, F. Fang, K. Chan, E.E. Papalexakis, "Harvester: Principled Factorization-based Temporal Tensor Granularity Estimation," in Proc. of the 2023 SIAM International Conference on Data Mining (SDM), Society for Industrial and Applied Mathematics, April 27 - 29, 2023, Minneapolis, MN. doi: 10.1137/1.9781611977653 (acceptance rate: 27.4%)

PARTICIPANTS:

Participant Type: PD/PI

Participant: Nikolaos Dimitrios Sidiropoulos

Person Months Worked: 4.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Co PD/PI

Participant: Xiao Fu

Person Months Worked: 3.00

Project Contribution:

National Academy Member: N

Funding Support:

RPPR Final Report
as of 05-Sep-2023

Participant Type: Co PD/PI
Participant: Zack W Almquist
Person Months Worked: 3.00
Project Contribution:
National Academy Member: N
Funding Support:

Participant Type: Staff Scientist (doctoral level)
Participant: Mikael Sorensen
Person Months Worked: 14.00
Project Contribution:
National Academy Member: N
Funding Support:

Participant Type: Staff Scientist (doctoral level)
Participant: Aritra Konar
Person Months Worked: 3.00
Project Contribution:
National Academy Member: N
Funding Support:

Participant Type: Graduate Student (research assistant)
Participant: Tri Nguyen
Person Months Worked: 15.00
Project Contribution:
National Academy Member: N
Funding Support:

Participant Type: Graduate Student (research assistant)
Participant: Sahana Ibrahim
Person Months Worked: 3.00
Project Contribution:
National Academy Member: N
Funding Support:

Participant Type: Graduate Student (research assistant)
Participant: Nikolaos Kargas
Person Months Worked: 2.00
Project Contribution:
National Academy Member: N
Funding Support:

Participant Type: Graduate Student (research assistant)
Participant: Faisal Almutairi
Person Months Worked: 2.00
Project Contribution:
National Academy Member: N
Funding Support:

RPPR Final Report
as of 05-Sep-2023

Participant Type: Graduate Student (research assistant)
Participant: Charilaos Kanatsoulis
Person Months Worked: 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Aja Sutton
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Elizabeth Nova
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Ihsan Kahveci
Person Months Worked: 8.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: June Yang
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Larisa Ozeryansky
Person Months Worked: 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

ARTICLES:

RPPR Final Report as of 05-Sep-2023

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 4-Under Review

Journal: submitted to IEEE Trans. on Knowledge and Data Engineering (TKDE)

Publication Identifier Type: Publication Identifier:

Volume: Issue: First Page #:

Date Submitted: 5/25/23 12:00AM Date Published:

Publication Location:

Article Title: Mining Triangle-Dense Subgraphs of a Fixed Size:Hardness, Lovasz extension and Applications

Authors: A. Konar and N. D. Sidiropoulos

Keywords: Dense subgraph discovery, triangle-motifs, intractability and approximation, submodularity, convex optimization

Abstract: We introduce the triangle-densest-k-subgraph problem (TDkS) for undirected graphs: given a size parameter k , compute a subset of k vertices that maximizes the number of induced triangles. The problem corresponds to the simplest generalization of the edge based densest-k-subgraph problem (DkS) to the case of higher-order network motifs. We prove that TDkS is NP-hard and is not amenable to efficient approximation, in the worst-case. By judiciously exploiting the structure of the problem, we propose a relaxation algorithm for the purpose of obtaining high-quality, sub-optimal solutions. Our approach utilizes the fact that the cost function of TDkS is submodular to construct a convex relaxation for the problem based on the Lovasz extension for submodular functions. We demonstrate that our approaches attain state-of-the-art performance on real-world graphs and can offer substantially improved exploration of the optimal density-size curve compared to sophisticated approximation baselines ...

Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info
Acknowledged Federal Support: Y

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 4-Under Review

Journal: Field Methods

Publication Identifier Type: Publication Identifier:

Volume: Issue: First Page #:

Date Submitted: 5/25/23 12:00AM Date Published:

Publication Location:

Article Title: Identity and Generosity Norms Among People Experiencing Homelessness in

Authors: MC Anderson, Jessica Perkins, Ashley Hazel, James Holland Jones, Zack W. Almquist

Keywords: Homelessness

Abstract: Although society frequently labels people experiencing homelessness (PEH) as a single, marginalized social group—i.e., “the homeless,” it is unknown whether PEH create social identities around their shared characteristic of homelessness. Using a dictator game, we tested whether PEH display in-group bias, a common feature of social groups, toward other PEH (an in-group) compared with housed people (an out-group). We did not observe in-group bias, suggesting that the shared vulnerability of homelessness may not necessarily foster the formation of a meaningful social group. On average, participants allocated 29% of their game endowment to recipients, suggesting that PEH are not entirely selfish and display behavior consistent with dictator game studies across various cultures. Among men, we found a positive association between game allocations the number of emotional and material supports in one’s social network, and the opposite among women. Finally, we found that PEH were more ...

Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info
Acknowledged Federal Support: Y

RPPR Final Report as of 05-Sep-2023

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 4-Under Review

Journal: Social Forces

Publication Identifier Type:

Publication Identifier:

Volume:

Issue:

First Page #:

Date Submitted: 5/25/23 12:00AM

Date Published:

Publication Location:

Article Title: Political and Educational Dynamics in Religious Group's Mask Noncompliance under COVID-19

Authors: June Yang, Zack W. Almquist, James Holland Jones

Keywords: Political and religious groups

Abstract: This study investigates the mediation effect of conservative political ideology on the relationship between religiosity and the noncompliance of masks under COVID-19, and the moderated effect of education on the mediation relationship. Using nationally representative data from COVID-19 and Social Distancing Survey, a moderated mediation analysis is applied to examine the pathway from religiosity via political conservativeness to the noncompliance of masks, and the interaction effect between years of education and political ideology. A logistic regression model is used to investigate each path in the mediation analysis. Results suggest that controlling for socio-demographic background, religiosity positively drives mask noncompliance. Further, findings confirm that political orientation is not only an established predictor of the polarized public support of masks as found in existing studies, but is also a key mechanism by which religiosity positively predicts mask resistance. ...

Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info
Acknowledged Federal Support: Y

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 4-Under Review

Journal: Social Science and Medicine

Publication Identifier Type:

Publication Identifier:

Volume:

Issue:

First Page #:

Date Submitted: 5/25/23 12:00AM

Date Published:

Publication Location:

Article Title: Barefaced and voting: maskwearing choice and the Trump vote in the United States during the COVID-19 pandemic

Authors: Aja Sutton, Zack W. Almquist, James Holland Jones

Keywords: COVID-19, logistic choice model, mask use, political ecology of health

Abstract: A poorly understood aspect of the choice to wear a mask during the COVID-19 pandemic in the United States is the role of politicization of mask-wearing along political party lines. We consider the role of voting for Donald Trump in mask-wearing choice during the COVID-19 pandemic in the United States using a logistic choice model of indoor mask-wearing choice and the Trump vote in the 2016 United States presidential election, accounting for sociodemographic and ecological characteristics, trust in government, approximate time and variation in relative space, as well as a Pearson's correlation analysis. Data used are from the American Community Survey 2019 5-year estimates and The Behavioral Change COVID-19 (BCC19) Survey, a nationally representative longitudinal survey of adults (over 18 years) in the United States of opinions and behaviors during the COVID-19 pandemic over three waves from September 2020 to June 2021. We find that voting for Donald Trump in the 2016 federal ...

Distribution Statement: 3-Distribution authorized to U.S. Government Agencies and their contractors
Acknowledged Federal Support: Y

CONFERENCE PAPERS:

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published

Conference Name: Advances in Neural Information Processing Systems (NeurIPS 2019).

Date Received: 03-Aug-2020

Conference Date: 09-Dec-2019

Date Published: 04-Aug-2019

Conference Location: Vancouver, BC

Paper Title: Crowdsourcing via Pairwise Co-occurrences: Identifiability and Algorithms

Authors: S. Ibrahim, X. Fu, N. Kargas, and K. Huang

Acknowledged Federal Support: Y

RPPR Final Report
as of 05-Sep-2023

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: IEEE Sensor Array and Multichannel Signal Processing Work- shop (SAM 2020), Hangzhou, China
Date Received: 29-Aug-2021 Conference Date: 09-Jun-2020 Date Published:
Conference Location: Hangzhou, China (conference was virtual)
Paper Title: Multiview Analysis: Identifiability and Neural Network- based Implementation
Authors: Q. Lyu and X. Fu
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: Proc. ACM SIGKDD 2020
Date Received: 29-Aug-2021 Conference Date: 24-Aug-2020 Date Published:
Conference Location: Virtual
Paper Title: Mining Large Quasi-cliques with Quality Guarantees from Vertex Neighborhoods
Authors: A. Konar and N. D. Sidiropoulos
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 5-Submitted
Conference Name: NeurIPS 2020 (under double blind review)
Date Received: 29-Aug-2021 Conference Date: 06-Dec-2020 Date Published:
Conference Location: Vancouver, cA
Paper Title: Graph Query-based Mixed Membership Learning: Identifiability and Algorithm
Authors: S. Ibrahim and X. Fu
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: Asilomar Conference on Signals, Systems and Computers 2020
Date Received: 29-Aug-2021 Conference Date: 01-Nov-2020 Date Published:
Conference Location: Virtual event this year
Paper Title: Recovering Joint PMF from Pairwise Marginals
Authors: S. Ibrahim and X. Fu
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: IEEE Statistical Signal Processing Workshop (SSP), July 11–14, 2021, Rio de Janeiro, Brazil (virtual event).
Date Received: 29-Aug-2021 Conference Date: 11-Jul-2021 Date Published:
Conference Location: Virtual event
Paper Title: Robust Probabilistic Simplex Component Anal- ysis
Authors: Y. Li, W.-K. Ma, N.D. Sidiropoulos
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: SIAM International Conference on Data Mining (SDM) 2021
Date Received: 29-Aug-2021 Conference Date: 29-Apr-2021 Date Published:
Conference Location: Virtual event
Paper Title: TeX-Graph: Coupled tensor-matrix knowledge-graph embedding for COVID-19 drug repurposing
Authors: C. Kanatsoulis and N.D. Sidiropoulos
Acknowledged Federal Support: **Y**

RPPR Final Report as of 05-Sep-2023

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: AAAI 2021, February 2-9, 2021
Date Received: 29-Aug-2021 Conference Date: 02-Feb-2021 Date Published:
Conference Location: Virtual event
Paper Title: STELAR: Spatio- temporal Tensor Factorization with Latent Epidemiological Regularization
Authors: N. Kargas, C. Qian, N.D. Sidiropoulos, C. Xiao, L. Glass, J. Sun
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: AAAI 2021, February 2-9, 2021.
Date Received: 29-Aug-2021 Conference Date: 02-Feb-2021 Date Published:
Conference Location: Virtual event
Paper Title: eTREE: Learning Tree- structured Embeddings
Authors: F. Almutairi, Y. Wang, D. Wang, E. Zhao, N.D. Sidiropoulos
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: WSDM 2021, March 8-12, 2021
Date Received: 29-Aug-2021 Conference Date: 08-Mar-2021 Date Published:
Conference Location: Virtual event
Paper Title: Exploring the Subgraph Density-Size Trade- off via the Lovasz Extension
Authors: A. Konar, N.D. Sidiropoulos
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021
Date Received: 29-Aug-2021 Conference Date: 09-Jun-2021 Date Published:
Conference Location: Virtual event
Paper Title: Learning Mixed Membership from Adjacency Graph Via System- atic Edge Query: Identifiability and Algorithm
Authors: S. Ibrahim and X. Fu
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: Fifteenth ACM International Conference on Web Search and Data Mining (WSDM '22)
Date Received: 25-Aug-2022 Conference Date: 02-Aug-2022 Date Published:
Conference Location: New York, New York
Paper Title: GAGE: Geometry Preserving Attributed Graph Embeddings
Authors: C.I. Kanatsoulis and N.D. Sidiropoulos
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: The Thirty-Sixth AAAI Conference on Artificial Intelligence
Date Received: 25-May-2023 Conference Date: 22-Feb-2022 Date Published: 28-Jun-2022
Conference Location: Virtual only
Paper Title: The Triangle-Densest-K-Subgraph Problem: Hardness, Lovasz Extension, and Application to Document Summarization
Authors: A. Konar, N. D. Sidiropoulos
Acknowledged Federal Support: **Y**

RPPR Final Report as of 05-Sep-2023

Publication Type: Conference Paper or Presentation **Publication Status:** 0-Other
Conference Name: 2023 SIAM International Conference on Data Mining (SDM), Society for Industrial and Applied Mathematics
Date Received: 25-May-2023 Conference Date: 27-Apr-2023 Date Published: 27-Apr-2023
Conference Location: Minneapolis, Minnesota
Paper Title: Harvester: Principled Factorization-based Temporal Tensor Granularity Estimation
Authors: R.S. Pasricha, U.S. Saini, N.D. Sidiropoulos, F. Fang, K. Chan, E.E. Papalexakis
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 0-Other
Conference Name: 56th Asilomar CSSC Conference, Pacific Grove, CA, USA
Date Received: 25-May-2023 Conference Date: 30-Oct-2022 Date Published: 15-Apr-2023
Conference Location: Pacific Grove, CA
Paper Title: Memory-Efficient Convex Optimization for Self-Dictionary Nonnegative Matrix Factorization
Authors: T. Nguyen, X. Fu and R.Wu
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: WSDM '21: The Fourteenth ACM International Conference on Web Search and Data Mining
Date Received: Conference Date: Date Published: 08-Mar-2021
Conference Location: Virtual Event Israel
Paper Title: Exploring the Subgraph Density-Size Trade-off via the Lova?z Extension
Authors: Aritra Konar, Nicholas D. Sidiropoulos,
Acknowledged Federal Support:

Partners

I certify that the information in the report is complete and accurate:
Signature: Nikolaos D. Sidiropoulos
Signature Date: 8/8/23 10:52AM

Geometric Matrix Factorization Tools for Community Mining

Final Report

N.D. Sidiropoulos*, X. Fu[†] and Z. Almquist[‡]

July 19, 2023

United States Army
Army Research Office
W911NF19-1-0407

*Department of Electrical & Computer Engineering
University of Virginia

[†]School of Electrical Engineering & Computer Science
Oregon State University

[‡]Department of Sociology & Statistics
University of Washington

Approved for public release; distribution unlimited. Preprints in the Annex should be treated as confidential until they appear in print.

Contents

1	Abstract	3
2	Motivation and Problem Statement	4
2.1	Relevance	4
3	Year 4 results (NCE period)	4
4	Overview of important earlier results (years 1-3)	7
4.1	Overlapping Community Detection using Semi-Binary Matrix Factorization	7
4.2	Probabilistic Geometric Factorization Tools	8
4.3	Provable Memory-Efficient Convex Geometric Factorization Tools	8
4.4	Maximum Likelihood Based Geometric Matrix Factorization from Partial Observations	9
4.5	Dense Subgraph Mining	10
4.6	Network Topology Inference	11
4.7	Advanced Optimization Algorithm for Geometric Matrix Factorization	12
4.8	Common latent factor analysis of multiple network views	13
4.9	Tensor Decomposition-Based Spatio-Temporal Migration Community Discovery . . .	13
4.10	Cummulative list of publications across the life of the project	14
4.11	Student and postdoctoral researcher development	17
5	Annex: Awards	19
6	Annex: Publications	20

1 Abstract

Clustering nodes (users, entities) in a social network into different communities is an instrumental task in network analytics. Community detection and association is a core problem in machine learning, data mining, and signal processing. However, there are still many challenges remaining, and fundamental aspects that are poorly understood. Whether it is possible to identify the unknown underlying communities and pin down the association of nodes to these communities under realistic conditions (e.g., from a randomly or systematically sampled subset of links) is *a priori* unclear. Theory and scalable algorithms for extracting large quasi-cliques (densely connected communities) in a graph are sorely missing; and how to infer network typologies with limited prior knowledge has also been an open problem. This project developed a suite of principled analytical and computational tools to advance the state-of-art on several fronts in the broad area of network analytics and graph information retrieval.

In year 4 (no-cost extension period) of the project, we worked on the problem of suitably aggregating multiple network “views” to come up with fewer compound views that exhibit low-rank structure. The key problem is that of aggregation granularity estimation for the third mode of a tensor whose other two modes represent nodes in the network. In joint work with UCR and ARL collaborators, we tackled this problem in a systematic way, and published the results in SIAM SDM 2023. We also worked on triangle-dense subgraph mining, specifically on the problem of finding a subgraph comprising k nodes with the highest number of triangles. We showed that the problem is NP-hard and designed highly effective approximation algorithms based on Lov  tz extension. The results were published in AAAI 2022, and an extended journal version is under review for TKDE. Co-PI Almquist worked on computational demography, social science aspects of the Covid-19 pandemic (the interplay between political affiliation and social distancing non-compliance) and homelessness. He published two chapters and submitted three papers under review in social science journals. He also worked on data homeless data collection under the REACH project. Co-PI Fu had concluded his part of the project before the no-cost extension started.

The full report provides a comprehensive summary of the technical and scientific contributions made throughout the course of the project.

Keywords: Network science, information retrieval and management, graph mining, graph clustering, community detection, geometric nonnegative matrix factorization, canonical correlation analysis.

2 Motivation and Problem Statement

Given large but possibly incomplete data regarding entities and their interactions in cyber/physical networks, can we accurately infer what are the latent communities and how they interact with each other? Can we extract densely connected subgraphs (tight communities) from ultra large networks by efficient local search? Can we understand the local connectivity as well as the global topology of the graph/network with partial or incomplete observations? These questions lie at the heart of homeland security, national defense, and battlefield intelligence operations, where it is important to identify extremist groups, detect clandestine activities, and anticipate tactical moves from seemingly inconspicuous observations. The goal of community and graph mining is to reveal the community membership of the various actors and the interaction patterns both within and across communities. The latter are particularly important in understanding how communities coalesce to pursue common objectives, or break apart. Beyond national security and defense applications, community mining and graph analysis also have a wide range of civilian applications, ranging from social network-aware recommender systems to gene analytics and document categorization and summarization, to name a few.

2.1 Relevance

With reference to Part II of BAA WW911NF-17-S-0002-07 for 01 April 2017 – 31 March 2022, this project fits in research program 1.c.iii (network sciences). The research addresses a number of topics under “(2) Social and Cognitive Networks”. The results of the project have advanced the state-of-art in terms of methodology and algorithms for timely problems in “information and knowledge management” and “social network analysis”: graph clustering, community detection, large dense subgraph (quasi-clique) extraction, network topology analysis, and associated factorization and optimization models. Research under this grant addresses fundamental and practical problems in network analysis and community mining. Taking identifiability into consideration when designing network factorization criteria is a crucial innovation that will boost the performance of community mining and our confidence in the results. Building upon identifiability, scalable and robust algorithms are being developed, providing a powerful suite of novel graph mining tools. Multilinear factorization has many more applications in data analytics, signal processing and machine learning. The tools developed in this project will therefore also benefit several ‘nearby’ application areas. The computational tools developed in this research will be released under a creative commons license, and the new datasets to be created will be valuable additions to existing data collections for community mining research.

3 Year 4 results (NCE period)

In year 4, we worked on the following problems.

Principled Factorization-based Temporal Tensor Granularity Estimation. Given a tensor that captures temporal data, such as (user, item, time), the way that we set the granularity of the “time” mode can *make or break* our analysis of the data. If we set the granularity to be extremely fine, we end up with a very sparse and high-rank tensor which is essentially incompatible with what virtually all tensor decomposition models expect, i.e., tensors with low-rank structure, which can be expressed in some form of factorization.

Traditionally, this problem has been *avoided* by setting the “time” granularity to a “reasonable” aggregation level (say hourly or daily intervals), an approach which has certainly served tensor analysis of temporal methods well so far. However, such an approach requires *tedious* trial-and-error experimentation across a number of such fixed aggregations, where typically the one that provides the most sensible results is retained, and furthermore it is *arbitrary*, since the optimal aggregation over time need not necessarily be uniform. In joint work with collaborators from UC Riverside and ARL, we introduced HARVESTER, the first principled factorization-based approach which seeks to identify the best temporal granularity of a given tensor. Unlike existing methods which follow a greedy approach, HARVESTER leverages multiple aggregated views of the tensor and a carefully-designed optimization problem in order to uncover an aggregation of a tensor which has a “good” structure for factor analysis or a downstream task.

We extensively evaluated HARVESTER on synthetic and real data, and demonstrated that it consistently produces tensors of very high quality, compared to the state-of-the-art, across a number of different popular quality measures that have been used by the community.

Outcome:

- R.S. Pasricha, U.S. Saini, N.D. Sidiropoulos, F. Fang, K. Chan, E.E. Papalexakis, “Harvester: Principled Factorization-based Temporal Tensor Granularity Estimation,” in *Proc. of the 2023 SIAM International Conference on Data Mining (SDM)*, Society for Industrial and Applied Mathematics, April 27 - 29, 2023, Minneapolis, MN. doi: 10.1137/1.9781611977653 (acceptance rate: 27.4%)

Mining Triangle-Dense Subgraphs. We introduced the triangle-densest- k -subgraph mining problem (TD k S) for undirected graphs: given a size parameter k , compute a subset of k vertices that maximizes the number of induced triangles. The problem corresponds to the simplest generalization of the edge based densest- k -subgraph problem (D k S) to the case of higher-order network motifs. We proved that TD k S is NP-hard and is not amenable to efficient approximation in the worst-case. By judiciously exploiting the structure of the problem, we proposed a relaxation algorithm for the purpose of obtaining high-quality, sub-optimal solutions. Our approach utilizes the fact that the cost function of TD k S is submodular to construct a convex relaxation for the problem based on the Lovász extension for submodular functions. We demonstrated that our approaches attain state-of-the-art performance on real-world graphs and can offer substantially improved exploration of the optimal density-size curve compared to sophisticated approximation baselines for D k S. We also used document summarization to showcase why TD k S is a useful generalization of D k S.

Outcomes:

- A. Konar and N. D. Sidiropoulos, “The Triangle-Densest-K-Subgraph Problem: Hardness, Lovász Extension, and Application to Document Summarization,” in *Proc. AAAI*, vol. 36, no. 4, pp. 4075-4082, Jun. 2022. (acceptance rate: 15%)
- A. Konar and N. D. Sidiropoulos, “Mining Triangle-Dense Subgraphs of a Fixed Size: Hardness, Lovász extension and Applications,” submitted to *IEEE Trans. on Knowledge and Data Engineering (TKDE)*, Jan. 2023. Under review.

In July 2023, PI Sidiropoulos worked on preparing a tutorial for IEEE CAMSAP 2023 reporting work on canonical correlation analysis with applications in graph mining, partially supported by this grant. This will be presented in Dec. 2023, see <https://camsap23.ig.umons.ac.be/tutorials.php>.

Computational demography and social science. Co-PI Almquist worked on computational demography and social science aspects of the Covid-19 pandemic, focusing on the interplay between political affiliation and social distancing non-compliance, and homelessness.

Sutton and Almquist (forthcoming book chapter) review the current state of the art statistical models for analyzing and discovering spatial aggregates (this can be extended temporally in a straightforward manner, e.g., VAR models). This provides the baseline for any future spatial/temporal clustering models that actively take distance into account. In another book chapter, Almquist, Allen and Kahveci (2023) provide general background in computational demography and health.

Sutton, Almquist and Jones (2023) provide a decision theoretic interpretation of mask wearing during COVID-19 at three points in time. This provides a clear clustering criterion based on political beliefs and actions. This would provide a baseline for any advanced methods on determining mask wearing during the pandemic. There is a clear gradient relationship between sets (Republican or Democrat) and not a perfect prediction of mask wearing. This is a good data set for ground truth of two communities with fuzzy relationships.

Yang, Almquist and Jones (2023) provide a causal explanation for mask wearing and religious adherence during COVID-19. This would provide a baseline model for either predicting religion or mask-wearing due to religious activity.

Anderson et al (2022) provides a game theoretic notion of identifying as a person experiencing homelessness. This provides ground truth for predicting who would identify with the term “homeless.” This is extremely important in a policy setting as identification can be key to use of and engagement with policy interventions and other care activities. This provides a data set with ground truth for overlapping identification.

Outcomes:

- Aja Sutton, Zack Almquist (forthcoming). “Small Area Estimation for Demographic Data” [Book chapter]. In: Recent Trends in Demographic Data. Ed. by Parfait M Eloundou-Enyegue. IntechOpen.
- Zack Almquist, C. Allen, and I. Kahveci (2023). “Computational Demography and Health” [Book chapter]. In: Handbook of Computational Social Science. Ed. by T. Yasseri. Edward Elgar Publishing.
- Aja Sutton, Zack W. Almquist, James Holland Jones. (2023). “Barefaced and voting: mask-wearing choice and the Trump vote in the United States during the COVID-19 pandemic”, submitted to *Social Science and Medicine*. Under review.
- June Yang, Zack W. Almquist, James Holland Jones (2023). “Political and Educational Dynamics in Religious Group’s Mask Noncompliance under COVID-19”, submitted to *Social Forces*. Under review.
- MC Anderson, Jessica Perkins, Ashley Hazel, James Holland Jones, Zack W. Almquist (2023). “Identity and Generosity Norms Among People Experiencing Homelessness in Nashville, TN: A Dictator Game Experiment”, submitted to *Field Methods*. Under review.

Co-PI Almquist made five presentations (without accompanying papers) at *Population Association of America* meetings, and two more at *Pacific Sociological Association* meetings. He also has a number of working papers in progress.

Co-PI Almquist also worked on data collection for the REACH data project, covering movement between tent encampments in Seattle and King County (broadly) over 8-10 years of recorded events. REACH is a homeless care support network that has hundreds of Navigators who cover geographic areas all over Seattle and King County. The Navigators record locations of tent encampments, who is there and other interactions (e.g., Veteran Status or demographics). The Navigators are present when Sweeps (forced migration events) occur. Almquist is also putting together environmental data, including locations of hospitals, needle exchanges, parks, trees, elevation, toxicity of ground layer (e.g., lead), air quality, and temperature. This is an ongoing project. The final dataset will have the movement patterns by event timing of people experiencing homelessness (PEH) between tent locations over a tenure period – including when tent communities fall apart and when they are rebuilt (if ever). This will include characteristics such as Veteran status, gender and age, environmental data and distance to resources, exogenous forced migration events when the city or county sweep away a tent encampment and information on whether the inhabitants were provided information about the incoming sweep or not and if they were directed where to go or not, and finally the tent or other locations the PEH migrated to.

Access to the data from REACH is extremely sensitive and is required to be accessed on a secure computer with no access to the internet on Center for Studies in Demography and Ecology (CSDE) secure servers at UW. Those interested in accessing the data should submit a short proposal for the research goals intended and it will be reviewed by CSDE and REACH for approval. A more detailed description of the REACH data is provided in the form of a .pdf file included with the latest publications in the Annex.

4 Overview of important earlier results (years 1-3)

In this section, we summarize and highlight the most important results of research under this project during years 1-3.

4.1 Overlapping Community Detection using Semi-Binary Matrix Factorization

Whereas geometric factorization techniques such as Volume Minimization (VolMin) constrain the overall membership / social affinity capacity of individuals and aim to estimate community engagement with the different groups, we have been exploring a different approach based on a semi-binary matrix factorization (SBMF) model. In our SBFM model, community membership is either 0 or 1 (i.e., we are interested in solving the *detection* version of the problem, and each individual or entity is free to choose the number of communities in which it participates. Thus there is no sum-to-one constraint when it comes to community membership. This is consistent with the fact that certain people / actors / entities are much more “social” than others, thus constraining everyone to a one-size-fits-all in terms of community activity does not seem appropriate in several application scenarios of interest.

The proposed SBFM model can be understood as a marriage between K -means clustering and non-negative matrix factorization. This leads to an easy-to-interpret factorization that at the same time can handle overlapping communities. In year 2, we had derived relatively easy-to-check uniqueness conditions, which corroborate that meaningful communities can be obtained via SBFM. Computing a (least-squares) optimal SBFM is a nontrivial problem. For this reason, we had also proposed a (coupled) tensor decomposition based method for SBFM that can also be used to initialize an

optimization based method, such as alternating least squares.

In year 3, we further refined our SBMF approach, including new theory, comparison to distinct but relevant work (including overlapping k-means and recent work by Tropp *et al.*), and applications. In collaboration with Dr. A. Swami of the Army Research Laboratory (ARL), we explored the effectiveness of the proposed SBMF approach for community detection from co-authorship data and stock-holding data.

Outcome: M. Sorensen, N. D. Sidiropoulos and A. Swami, “Overlapping Community Detection via Semi-binary Matrix Factorization: Identifiability and Algorithms,” IEEE Transactions on Signal Processing, vol. 70, pp. 4321-4336, 2022, doi: 10.1109/TSP.2022.3200215.

4.2 Probabilistic Geometric Factorization Tools

We finished work on PRISM, a probabilistic simplex component analysis approach to identifying the vertices of a data-circumscribing simplex from data. The problem has a rich variety of applications, including community detection, hyperspectral unmixing in remote sensing and non-negative matrix factorization in machine learning. PRISM uses a simple probabilistic model, namely, uniform simplex data distribution and additive Gaussian noise, and it carries out inference by maximum likelihood. The inference model is sound in the sense that the vertices are provably identifiable under some assumptions, and it suggests that PRISM can be effective in combating noise when the number of data points is large. PRISM has strong, but hidden, relationships with simplex volume minimization, a powerful geometric approach for the same problem. In collaboration with Prof. W.-K. Ma and his group at the Chinese University of Hong Kong, we studied these fundamental aspects, and we also considered algorithmic schemes based on importance sampling and variational inference. In year 3, we prepared a major revision of the paper, including extensive additional experiments and comparisons, and we derived some new insights as a result. Full details can be found in the published paper.

Outcome: R. Wu, W.-K. Ma, Y. Li, A.M.-C. So, and N.D. Sidiropoulos, “Probabilistic Simplex Component Analysis”, IEEE Trans. on Signal Proc., vol. 70, pp. 582-599, 2022, doi: 10.1109/TSP.2021.3133690.

4.3 Provable Memory-Efficient Convex Geometric Factorization Tools

In Year 3, we continued our development of provable memory-efficient geometric factorization tools. Our goal is a global optimality-guaranteed geometric matrix factorization algorithm that can be applied to *very large-scale data*. Such a computational tool is particularly needed for large-scale community detection problems – as overlapped community detection is often formulated as a noisy geometric matrix factorization problem.

In Year 2 we started working on an algorithm leveraging the separability condition that is widely used in the geometric matrix factorization literature. In the context of community mining, separability means that each community has an “anchor node” – i.e., every community has at least one member who is solely associated with this community. This condition, in large networks,

is often not very hard to be even approximately met. Under separability, the geometric matrix factorization-based community detection problem can be formulated as a self-dictionary multiple measurement vector (SD-MMV) problem, which has a convex relaxation form. Convex SD-MMV is provably robust to noise. In addition, convex optimization is polynomial-time solvable. Hence, this formulation has attracted much attention in the past decade. However, this type of approach has prohibitively large memory consumption, which scales quadratically with the number of nodes (i.e. $O(N^2)$) in a graph/network if applied for solving community detection problems. In the past 10 years, there had been several attempts to solve this memory challenge, yet the complexity order remained the same.

To solve this challenge, we developed a Frank-Wolfe (FW) (or, conditional gradient) algorithm for handling the self-dictionary based geometric matrix factorization formulation. The idea was nurtured and tested in Year 2. In Year 3, we have solidified our developments and published this work in the top signal processing journal, i.e., IEEE TRANSACTIONS ON SIGNAL PROCESSING. To be specific, our work in Year 3 includes:

1. Studying an unregularized SD-MMV formulation and analyzing its noise robustness and memory efficiency. This study helps distilling insights on the reason why the FW algorithm is in unique position to solve the long-lingering memory challenge. The result also revealed for the first time the connection between a specific algorithm and the identifiability of the latent factors of the GMF model in some cases – whereas these two aspects had been generally considered to be fully disconnected in the past.
2. Proposing a smoothed row-sparsity regularization for SD-MMV to enhance performance. The smoothing technique is important for applying FW for the regularized formulation. We proved that such smoothed regularization retains the identifiability properties of the original non-smooth version of SD-MMV.
3. Providing theoretical underpinning of the noise robustness and memory-efficiency of our proposed regularized SD-MMV formulation. Our work proved that the FW algorithm will never use more than $O(NK)$ memory (where K is the model rank and often a small number) under reasonable conditions. This approach offers much improved noise robustness over the greedy approaches and also solves the long-standing memory challenge in convex optimization-based SD-MMV.

Outcome: T. Nguyen, X. Fu and R. Wu, “Memory-Efficient Convex Self-Dictionary Nonnegative Matrix Factorization: A Frank-Wolfe Approach,” IEEE Transactions on Signal Processing, vol. 70, pp. 3221–3236, 2022, DOI: 10.1109/TSP.2022.3177845.

4.4 Maximum Likelihood Based Geometric Matrix Factorization from Partial Observations

The overlapped community detection problem is often formulated as a model identification problem under the so-called *mixed membership stochastic blockmodel* (MMSB). Under this model, the observed adjacency matrix consists of binary link indicators (i.e., “1” means there is a link between

two nodes and “0” means the opposite). The indicators are sampled from a low-rank matrix parameterization using the Bernoulli distribution. The community detection or membership estimation problem boils down to estimating the low-rank matrix’s latent factors. The MMSB identification problem has been tackled from a noisy geometric matrix factorization (GMF) viewpoint, but there are a number of challenges: First, the existing GMF methods often rely on eigendecomposition to estimate the range space of the adjacency matrix or the graph statistics as the pre-processing. This step needs the full observation of the adjacency matrix to guarantee good performance (like in spectral clustering). Second, most existing approaches do not have any type of statistical optimality.

In Year 3, we proposed a *maximum likelihood estimation* (MLE)-based formulation for estimating the membership matrix from *incomplete* MMSB link observations. Our formulation treats the membership matrix as a deterministic model parameter and uses the Bernoulli prior. The constraints of the membership matrix and other key parameters are considered. We proposed a gradient projection based algorithm to tackle this problem. So far we have the following results:

Observations: First, the MLE-based formulation can effectively recover the membership matrix under the MMSB model with as few as less than 1% observed entries, if the rank (i.e., the number of communities/clusters) is small and the number of nodes is large. Second, the MLE-based formulation needs the conditions that guarantee GMF identifiability (e.g., the anchor node condition or the “sufficiently scattered condition”) to work properly, which is expected. The reason is that the method implicitly has an GMF step “embedded” in the MLE formulation. Third, using gradient-based algorithms to handle the MLE works to a certain extent, but the algorithm is very slow. The bottleneck lies at the computation of the gradient in each step, which costs $O(N^4)$ flop and thus is not really scalable for large network analysis problems.

4.5 Dense Subgraph Mining

Dense subgraph mining is a key primitive in graph analysis with many applications – ranging from discovering protein complexes (from protein-protein interaction networks), to detecting spam farms in Web graphs, spotting collusion, and making sense of concerted actions in multi-agent networks. It can also be employed to detect events in network streams since abrupt changes tend to create local dense interactions. Dense subgraph mining is also a way to detect communities in larger ‘host’ networks.

A fully connected subset of nodes is called a clique, and there has been a lot of work on clique mining. Most communities are well-connected, but do not constitute cliques – hence the interest in dense subgraph mining. Any subgraph of a clique is a clique, and subgraphs of a dense subgraph are typically dense – hence we seek the largest dense subgraphs, because these correspond to the largest coherent communities we can extract from the global graph. Discovering large dense subgraphs is highly nontrivial. Existing approaches suffer from either high computational complexity or lack of theoretical guarantees (or both). Various approaches have been proposed, based on different definitions of what constitutes a dense subgraph. The classic Bron-Kerbosch algorithm can be used to discover the set of all maximal cliques (i.e., cliques that are not included in other cliques) of a graph, but the complexity is prohibitive for large-scale graphs. Various ideas have been proposed to alleviate the computational burden, including the α -quasi-clique based methods that relax the notion of “strictly” dense subgraph and aim at mining “approximately” dense subgraphs. Such a relaxation makes the dense subgraph mining problem more tractable in concept, but theory and methods for realizing quasi-clique mining are still largely behind. The existing algorithms either lack solution quality guarantees or exhibit very high computational/memory complexity (e.g., due

to using semidefinite programming).

Our approach is refreshingly different: it is very simple, and it comes with performance guarantees. The key is to leverage two recurring characteristics of real-world graphs, namely heavy-tailed degree distributions and large clustering coefficients. We have proven that these two characteristics imply the existence of substantially large vertex neighborhoods with high edge-density. Consequently, we proposed a very simple method for dense subgraph extraction. The method consists of two stages. First, visit every vertex in the graph and compute the edge-density of the subgraph induced by its one-hop neighbors. Second, output the “best” subgraph (according to certain metrics). This simply entails computing the local clustering coefficient of every node in the graph, which can be accomplished by enumerating all triangles in the graph – a task for which there exist several efficient algorithms. Our analysis has some surprising implications. One is that, contrary to what one might be primed to expect from worst-case theoretical analysis, mining cliques and quasi-cliques of non-trivial sizes from real-world graphs is often not difficult – which is quite remarkable given the combinatorial nature of the problem considered.

Outcome: A. Konar and N. D. Sidiropoulos, “Mining Large Quasi-cliques with Quality Guarantees from Vertex Neighborhoods,” in *Proc. ACM SIGKDD 2020*. (This is the prime conference in graph mining, with acceptance rate 17% that year).

We also considered a different approach for mining dense subgraphs. Given an undirected graph, the Densest-k-Subgraph problem (DkS) seeks to find a subset of k vertices such that the sum of the edge weights in the corresponding subgraph is maximized. The problem is known to be NP-hard, and is also very difficult to approximate, in the worst-case. We derived a new convex relaxation for the problem. Our key idea is to reformulate DkS as minimizing a submodular function subject to a cardinality constraint. Exploiting the fact that submodular functions possess a convex, continuous extension (known as the Lovász extension), we proposed to minimize the Lovász extension over the convex hull of the cardinality constraints. Although the Lovász extension of a submodular function does not admit an analytical form in general, for DkS we showed that it does. We leveraged this result to develop a highly scalable algorithm based on the Alternating Direction Method of Multipliers (ADMM) for solving the relaxed problem. Coupled with a pair of fortuitously simple rounding schemes, we demonstrated that our approach outperforms existing baselines on real-world graphs and can yield high quality sub-optimal solutions which typically are a posteriori no worse than 65-80% of the optimal density.

Outcome: A. Konar, and N.D. Sidiropoulos, “Exploring the Subgraph Density-Size Trade-off via the Lovasz Extension”, in *Proc. WSDM 2021*, pp. 743751, Jerusalem, March 8-12, 2021 (acceptance rate: 18.6%). doi: 10.1145/3437963.3441756.

4.6 Network Topology Inference

Estimating network topology from incomplete measurements is an essential step towards understanding complex systems such as brain-, social-, and financial networks. Existing methods approach this challenging problem using so-called *structural equation models* (SEMs). Essentially, these methods assume that some exogenous inputs “injected” to the network are available. Then, using the “excited” graph responses and the known inputs, the topology inference problem can be recast as a least squares problem. The idea is elegant, but many challenges remain. The biggest

hurdle is that assuming that we know the exogenous inputs is unrealistic for many networks. A recent method using factor analysis tools (in particular, low-rank tensor decomposition) to handle the same problem relaxes this assumption to some extent, but still assumes that the second-order statistics of the network inputs are known— even these are hard to acquire in practice.

Our approach puts forth an input-agnostic (‘blind’) topology inference method based on a classical factor analysis tool, namely, joint diagonalization (JD). JD can be viewed as a variant of the canonical polyadic tensor decomposition model with structural constraints on its latent factors. There are several advantages for employing JD, which include more efficient algorithms, and the convenience of readily accounting for known network structure, such as sparsity. Unlike existing alternatives, we established identifiability guarantees without resorting to the exogenous inputs or their statistics. We proposed three new JD algorithms specifically designed for network topology inference, and we illustrated their performance using stock market network inference as an application example where we seek to identifying connections among publicly traded companies. These can be corroborated using stakeholder and market sector / supply chain / commodity data. Our approach can be applied to other types of networks where we have little knowledge of the exogenous inputs, incentives and motives that drive the behavior of different actors in the network.

Outcome: Y. Shen, X. Fu, G. B. Giannakis and N. D. Sidiropoulos, “Topology Identification of Directed Graphs via Joint Diagonalization of Correlation Matrices,” in *IEEE Transactions on Signal and Information Processing over Networks*, vol. 6, pp. 271-283, 2020.

4.7 Advanced Optimization Algorithm for Geometric Matrix Factorization

Geometric matrix factorization is an NP-hard problem. The origin of this problem can be traced back to the 1990s in the remote sensing community, but only heuristic algorithms were known during that period. In recent years, many more principled continuous nonconvex optimization algorithms were proposed to handle this problem. However, due to its challenging nature, many compromises have to be made when handling geometric factorization. There are two major lines of work for computing geometric matrix factorization. One is to convert this matrix factorization problem to an “inverse filtering” problem. Instead of directly looking for the two latent factor matrices of the factorization model, this line of work seeks the inverse matrix of the basis factor. This reformulation significantly reduces the number of parameters in the optimization problem, and often leads to stable and fast algorithms. However, since the algorithms operate in an ‘inverse’ domain, prior knowledge on the original latent factors cannot be utilized to enhance performance. The other line of work employs fitting criteria with geometric volume regularization, leading to regularized NMF-type formulations. The upshot is that the physical meaning of the latent factors can be exploited to fend against heavy noise or outliers. But this type of formulation has a large number of parameters, and often results in algorithms with lower efficiency relative to the first type.

We proposed a new primal-dual algorithm to handle the geometric matrix decomposition problem. Instead of dealing with ‘inverse filtering’ or data fitting formulations, the idea is to introduce a dual variable for relaxing the hard factorization constraint and tackle the augmented Lagrangian problem. This way, the latent factors are kept in the formulation (so that their physical meaning and associated constraints can be incorporated, if needed), but the dual variable can be used to significantly improve the convergence efficiency. This algorithm outperforms a number of state-of-the-art geometric matrix factorization algorithms in terms of offering a balanced solution between

efficiency and numerical stability. The new augmented Lagrangian framework is also backed by rigorous convergence proof.

Outcome: Q. Shi, M. Hong, X. Fu, and T.-H. Chang, “Penalty Dual Decomposition Method For Nonsmooth Nonconvex Optimization—Part II: Applications,” *IEEE Transactions on Signal Processing*, June 16, 2020.

4.8 Common latent factor analysis of multiple network views

In many network science applications, we observe multiple instances (“snapshots” or “views”) of a given network, taken over time or in different contexts (e.g., Linkedin / professional coauthorship, and facebook connections). In such cases, we are often interested in what is *common* across these multiple network instances – we do not care about personal connections that are not related to business activities, for example. The typical way in which such multi-view networks are handled in the literature is through coupled latent factorization – often called *data fusion* – which only works if the common latent components are also the dominant ones – which is usually not true when we are dealing with seemingly inconspicuous low-intensity activities. We have recently re-discovered a tool that is particularly useful in this regard. (Generalized) Canonical Correlation Analysis (GCCA) has found numerous applications in data mining, machine learning, and artificial intelligence, where it is used to reveal ‘common’ random variables that are strongly correlated across multiple feature representations (views) of the same set of entities. CCA and to a lesser extent GCCA have been studied from the statistical and algorithmic points of view (we have been working on developing advanced GCCA algorithms for several years), but not as much from the standpoint of linear algebra. Our research has made an important connection that offers a fresh algebraic perspective on GCCA, based on a (bi-)linear generative model that naturally captures its essence. We have shown that from a linear algebra point of view, GCCA is tantamount to subspace intersection; and we have provided tight conditions under which the common subspace of the different views is identifiable. We also proposed a novel GCCA algorithm based on subspace intersection, which scales up to handle large GCCA tasks.

Outcome: M. Sorensen, C. Kanatsoulis and N. D. Sidiropoulos, “Generalized Canonical Correlation Analysis: A Subspace Intersection Approach,” submitted to *IEEE Transactions on Signal Processing*, vol. 69, pp. 2452-2467, 2021, doi: 10.1109/TSP.2021.3061218.

4.9 Tensor Decomposition-Based Spatio-Temporal Migration Community Discovery

In Year 3, we continued advancing our study on spatio-temporal tensor decomposition based migration system analysis. Understanding migration systems is a core task in understanding human mobility. In Year 2, the team has discovered that low-rank tensor decomposition is ideally suited for describing “stable” spatial interaction clusters across time – which is an intuitively pleasing way to define and understand a migration system. Consequently, we proposed a novel *spatio-temporal (ST) tensor co-clustering* method for migration system mining. In the first phase of the research, we focused on domestic migration between counties in the US from 1990-2018, a period for which pertinent data has been made available through the US Internal Revenue Service. Specifically, we concentrated on three illustrative case studies: (i) US Metropolitan Areas, (ii) the state of

California, and (ii) Louisiana with focus on detecting exogenous events such as Hurricane Katrina in 2005. Our findings have three important implications: (i) using the proposed method we can extract data-driven migration systems that are stable over time and space; (ii) we can correctly derive expected migration systems across the US (e.g. Northern and Southern California), and (iii) we can detect exogenous shocks to the migration system (e.g. Hurricane Katrina). Beyond the purely descriptive, our approach and results have the potential to improve population forecasting, as Andris et al. demonstrated the usefulness of migration clusters in predicting future migration flows. In Year 3, we applied the ST tensor co-clustering method to discover larger-scale migration systems at the globe level. The identification of international migration systems continues to be an open problem in the field. A number of works have posited that recurrent international migration systems exist, and provided a set of general principles underpinning such systems. On the other hand, there is relative scarcity of analytical and data-centric methods in this area. Most existing works treat the migration system discovery problem as a year by year process and do not exploit the strong temporal dependence across subsequent years. This is very different from our work, whose goal is to find stable co-evolving migration patterns over time. In Year 3, we applied the ST tensor co-clustering method to a dataset that contains international migration records comprising 200 countries from 1990 to 2015. Our algorithm leads to important differences in output and understanding of the migration systems. The ST tensor co-clustering produces a set of migration systems which exist over the entire observation period. Each system is characterized by proportionally co-evolving origin to destination migration patterns, whose intensity varies over time in the same way for all origin-destination pairs within the given system. The results offer a clear and explainable association from year to year, unlike e.g., the year-by-year community detection methods of Abel et al. which cannot guarantee any persistence of the extracted systems. Our ST co-clustering method, on the other hand, captures persistent migration systems and their dynamics, in line with what Kritz and Zlotnik called “dynamic stability”.

Outcome: Z. Almquist, T.-D. Nguen, M. Sorensen, X. Fu, and N.D. Sidiropoulos, “Uncovering migration systems through spatio-temporal tensor co-clustering,” revised version to be submitted to *Sociological Science* or *Socius*. Preprint: <https://arxiv.org/abs/2112.15296>.

4.10 Cumulative list of publications across the life of the project

Full details on our results are provided in the paper reprints and preprints, which are included in the Annex.

- **Archival journal papers:**

1. Y. Shen, X. Fu, G. B. Giannakis and N. D. Sidiropoulos, “Topology Identification of Directed Graphs via Joint Diagonalization of Correlation Matrices,” in *IEEE Transactions on Signal and Information Processing over Networks*, vol. 6, pp. 271-283, 2020, doi: 10.1109/TSIPN.2020.2984131.
2. Q. Shi, M. Hong, X. Fu and T. Chang, “Penalty Dual Decomposition Method For Nonsmooth Nonconvex Optimization—Part II: Applications,” in *IEEE Transactions on Signal Processing*, doi: 10.1109/TSP.2020.3001397.

3. X. Fu, E. Seo, J. Clarke and R. A. Hutchinson, "Link prediction under imperfect detection: Collaborative filtering for ecological networks," in *IEEE Transactions on Knowledge and Data Engineering*, doi: 10.1109/TKDE.2019.2962031.
4. Q. Lyu and X. Fu, "Nonlinear multiview analysis: Identifiability and neural network-assisted implementation," in *IEEE Transactions on Signal Processing*, vol. 68, pp. 2697-2712, 2020, doi: 10.1109/TSP.2020.2986351.
5. X. Fu, N. Vervliet, L. De Lathauwer, K. Huang and N. Gillis, "Computing large-scale matrix and tensor decomposition with structured factors: A unified nonconvex optimization perspective", *IEEE Signal Processing Magazine*, vol. 37, no. 5, pp. 78-94, Sept. 2020, doi: 10.1109/MSP.2020.3003544.
6. R. Wu, W.-K. Ma, Y. Li, A.M.-C. So, and N.D. Sidiropoulos, "Probabilistic Simplex Component Analysis", *IEEE Trans. on Signal Processing*, vol. 70, pp. 582-599, 2022, doi: 10.1109/TSP.2021.3133690.
7. M. Sorensen, C. Kanatsoulis, and N.D. Sidiropoulos, "Generalized Canonical Correlation Analysis: A Subspace Intersection Approach", *IEEE Trans. on Signal Processing*, vol. 69, pp. 2452-2467, 2021, doi: 10.1109/TSP.2021.3061218.
8. S. Ibrahim and X. Fu, "Recovering Joint Probability of Discrete Random Variables From Pairwise Marginals," *IEEE Transactions on Signal Processing*, vol. 69, pp. 4116-4131, 2021, doi: 10.1109/TSP.2021.3090960.
9. S. Ibrahim and X. Fu, "Mixed Membership Graph Clustering via Systematic Edge Queries", *IEEE Transactions on Signal Processing*, vol. 69, pp. 5189-5205, 2021, doi: 10.1109/TSP.2021.3109380.
10. T. Nguyen, X. Fu and R. Wu, "Memory-Efficient Convex Self-Dictionary Nonnegative Matrix Factorization: A Frank-Wolfe Approach," *IEEE Transactions on Signal Processing*, vol. 70, pp. 3221-3236, 2022, doi: 10.1109/TSP.2022.3177845.
11. M. Sorensen, N. D. Sidiropoulos and A. Swami, "Overlapping Community Detection via Semi-binary Matrix Factorization: Identifiability and Algorithms," *IEEE Transactions on Signal Processing*, vol. 70, pp. 4321-4336, 2022, doi: 10.1109/TSP.2022.3200215.
12. A. Konar and N. D. Sidiropoulos, "Mining Triangle-Dense Subgraphs of a Fixed Size: Hardness, Lovász extension and Applications," submitted to *IEEE Trans. on Knowledge and Data Engineering (TKDE)*, Jan. 2023. Under review.
13. Z. Almquist, T.-D. Nguen, M. Sorensen, X. Fu, and N.D. Sidiropoulos, "Uncovering migration systems through spatio-temporal tensor co-clustering," revised version to be submitted to *Sociological Science* or *Socius*. Preprint: <https://arxiv.org/abs/2112.15296>.
14. Aja Sutton, Zack W. Almquist, James Holland Jones. (2023). "Barefaced and voting: mask-wearing choice and the Trump vote in the United States during the COVID-19 pandemic", submitted to *Social Science and Medicine*. Under review.
15. June Yang, Zack W. Almquist, James Holland Jones (2023). "Political and Educational Dynamics in Religious Group's Mask Noncompliance under COVID-19", submitted to *Social Forces*. Under review.
16. MC Anderson, Jessica Perkins, Ashley Hazel, James Holland Jones, Zack W. Almquist (2023). "Identity and Generosity Norms Among People Experiencing Homelessness in

Nashville, TN: A Dictator Game Experiment”, submitted to *Field Methods*. Under review.

- **Archival highly selective conference papers:**

1. A. Konar and N. D. Sidiropoulos, “Mining Large Quasi-cliques with Quality Guarantees from Vertex Neighborhoods,” in *Proc. ACM SIGKDD* 2020. (acceptance rate 17%)
2. S. Ibrahim, X. Fu, N. Kargas, and K. Huang, “Crowdsourcing via Pairwise Co-occurrences: Identifiability and Algorithms”, In *Advances in Neural Information Processing Systems* (NeurIPS 2019). (acceptance rate: 21.6%)
3. Charilaos Kanatsoulis, N.D. Sidiropoulos, “TeX-Graph: Coupled tensor-matrix knowledge-graph embedding for COVID-19 drug repurposing”, SIAM International Conference on Data Mining (SDM) 2021, April 29 – May 1, 2021. (acceptance rate: 21.25%)
4. N. Kargas, C. Qian, N.D. Sidiropoulos, C. Xiao, L. Glass, J. Sun, “STELAR: Spatio-temporal Tensor Factorization with Latent Epidemiological Regularization”, AAAI 2021, February 2-9, 2021. (acceptance rate: 21%)
5. F. Almutairi, Y. Wang, D. Wang, E. Zhao, N.D. Sidiropoulos, “eTREE: Learning Tree-structured Embeddings”, AAAI 2021, February 2-9, 2021. (acceptance rate: 21%)
6. C.I. Kanatsoulis and N.D. Sidiropoulos, “GAGE: Geometry Preserving Attributed Graph Embeddings,” in *Proc. Fifteenth ACM International Conference on Web Search and Data Mining (WSDM ’22)*. Association for Computing Machinery, New York, NY, USA, 439–448. (acceptance rate: 20.23%) <https://doi.org/10.1145/3488560.3498467>
7. A. Konar and N. D. Sidiropoulos, “The Triangle-Densest-K-Subgraph Problem: Hardness, Lovász Extension, and Application to Document Summarization,” in *Proc. AAAI*, vol. 36, no. 4, pp. 4075-4082, Jun. 2022. (acceptance rate: 15%)
8. R.S. Pasricha, U.S. Saini, N.D. Sidiropoulos, F. Fang, K. Chan, E.E. Papalexakis, “Harvester: Principled Factorization-based Temporal Tensor Granularity Estimation,” in *Proc. of the 2023 SIAM International Conference on Data Mining (SDM)*, Society for Industrial and Applied Mathematics, April 27 - 29, 2023, Minneapolis, MN. doi: 10.1137/1.9781611977653 (acceptance rate: 27.4%)

- **Archival conference papers:**

1. Q. Lyu and X. Fu, “Nonlinear Multiview Analysis: Identifiability and Neural Network-based Implementation”, *IEEE Sensor Array and Multichannel Signal Processing Workshop* (SAM 2020), June 8-11, Hangzhou, China. (**Best Student Paper Finalist**)
2. S. Ibrahim and X. Fu, “Learning Mixed Membership from Adjacency Graph Via Systematic Edge Query: Identifiability and Algorithm,” ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 5370-5374, doi: 10.1109/ICASSP39728.2021.9413541.
3. S. Ibrahim and X. Fu, “Recovering Joint PMF from Pairwise Marginals,” 2020 54th Asilomar Conference on Signals, Systems, and Computers, 2020, pp. 356-360, doi: 10.1109/IEEECONF51394.2020.9443425.

4. T. Nguyen, X. Fu and R. Wu, "Memory-Efficient Convex Optimization for Self-Dictionary Nonnegative Matrix Factorization," 2022 56th Asilomar CSSC Conference, Pacific Grove, CA, USA, 2022, pp. 1372-1376, doi: 10.1109/IEEECONF56349.2022.10051825.

- **Book chapters:**

1. Aja Sutton, Zack Almquist (forthcoming). "Small Area Estimation for Demographic Data" [Book chapter]. In: Recent Trends in Demographic Data. Ed. by Parfait M Eloundou-Enyegue. IntechOpen.
2. Zack Almquist, C. Allen, and I. Kahveci (2023). "Computational Demography and Health" [Book chapter]. In: Handbook of Computational Social Science. Ed. by T. Yasseri. Edward Elgar Publishing.

- **Non-archival abstract-only conference presentations:**

1. Wang, S, Gilroy, C., Spiro, E. and Almquist, ZW. (2022) "Online Information Sharing of Spanish Speaking HUD Sponsored Homeless Care Organizations Response to COVID-19 Pandemic." Poster presented at INSNA Sunbelt 2022, Cairns, Australia (virtually).
2. Almquist, Z. W., T. D. Nguyen, M. Sorensen, X. Fu, and N. D. Sidiropoulos (2022). "Uncovering migration systems through spatio-temporal tensor co-clustering." The 2022 Annual Meeting of the Population Association of America, Atlanta, GA.
3. Kahveci, I., Nova, E, Z. Hsiao, K. Koltai, E. S. Spiro and Z. W. Almquist (2022). "Talking about the Jab: Medical Professionals' Expressions of Vaccine Hesitancy Online." The 2022 Annual Meeting of the American Sociological Association paper session on Harnessing Internet Data for Sociological Research, Los Angeles CA.
4. Sutton, A., Z. W. Almquist, A. Hazel, M. Kline, P. Smaldino, C. Moya, and J. H. Jones (2022) "Re-evaluating spatial multi-level regression with poststratification for small area estimation." The 2022 Annual Meeting of the American Sociological Association paper session on Quantitative Methodology, Los Angeles CA.
5. Kahveci, I., Z. W. Almquist, Z. Hsiao, K. Koltai, and E. S. Spiro (2022). "Talking about the Jab: Medical Professionals' Expressions of Vaccine Hesitancy Online." The 2022 Annual Meeting of the Population Association of America, Atlanta, GA.
6. Sutton, A., Z. W. Almquist, A. Hazel, M. Kline, P. Smaldino, C. Moya, and J. H. Jones (2022). Adherence to maskwearing and trust in government: a Bayesian hierarchical analysis of COVID-19 social behavior in the United States. The 2022 Annual Meeting of the Population Association of America, Atlanta, GA.

4.11 Student and postdoctoral researcher development

At Oregon State, PhD student Tri Nguyen, was fully supported by this project through a Graduate Research Assistantship (12 months) until August 2022. Mr. Nguyen is close to his final PhD defense.

At the University of Virginia, a postdoctoral research scientist (Mikael Sorensen) was partially supported (50%) through this project until October 2022, and another postdoctoral research scientist (Aritra Konar) was engaged in research under this project. Several PhD students were also

engaged in the early stages of the research (N. Kargas, F. Almutairi, C. Kanatsoulis). These students and postdoctoral researchers have been trained in state-of-art theoretical and algorithmic tools for network analysis, which also have many applications in other areas. Konar is now tenure-track assistant professor at KU Leuven in Belgium, one of the oldest and most prestigious research Universities in Europe.

At University of Washington, several students were involved, as detailed below by Co-PI Almquist:

Aja Sutton, a PhD student in Geography Graduating in 2023. Sutton worked on the REACH data set and the book chapter on probabilistic clustering (community detection) with spatial and demographic data. Dissertation focuses on statistical models for small area estimation. First author on book chapter on small area estimation for demography which provides a clear baseline comparison for using the community detection models for improving small area estimation of migration flows. She worked on this project in the summer and fall quarters of 2022.

Elizabeth Nova, a URM PhD student in Sociology. Nova worked on her masters paper that looks at social media and mainstream media impact on mask wearing (as multiple category events). This provides a standard statistical approach to understanding the relationship between mask wearing (none of the time to all the time) in relationship to social media usage and mainstream media usage. Social media and mainstream media represent overlapping sets and good applications for community detection and mining applications. She worked on this project in the summer and fall quarters of 2022.

Ihsan Kahveci, a 4th year PhD student in Sociology working on his comprehensive exam. Kahveci worked on Social Media data and more recently worked on comparing Facebook run survey to representative based survey on pro-social/anti-social activities. He is showing how you can locate similar enough respondents in each survey to use in propensity matching schemes. This matching of respondents is an ideal application of probabilistic community matches over Blau Spaces (or socio-demographic arrays). Ihsan worked on this project during the previous reporting year, as well as summer and fall of 2022.

June Yang, a PhD student in Sociology. 4th year and working on her dissertation prospectus. Yang worked on a paper relating religious identification (multiple and overlapping categories, e.g. Christian and its various sects) and mask wearing (none to all the time). This provides a ground truth for automated community detection methods which can handle overlapping sets. She worked on this project in the Spring and Fall of 2022.

Larisa Ozeryansky, a PhD student in the Interdisciplinary PhD program. First year and is working on her dissertation prospectus on refugees from the Ukraine/Russia conflict. Ozeryansky helped with developing simulation models for realistic networks of people living unsheltered. These models include the ability to add realistic overlapping category variables (e.g. tent communities) for use in testing new community detection methods. Worked on this project in the summer of 2022.

5 Annex: Awards

- Both co-PIs, Zack W Almquist and Xiao Fu received the NSF CAREER Award, from the SBE Sociology division, and the ECCS program under the Engineering Directorate, respectively.
- Both co-PIs, Zack W Almquist and Xiao Fu were promoted to Associate Professor with tenure at their respective institutions.
- Co-PI Almquist was selected to be the Editor-in-Chief for the *Journal of Mathematical Sociology* (the flagship journal for the ASA section on Mathematical Sociology).
- PI Sidiropoulos, received
 1. The Technical Achievement Award of the European Association for Signal Processing (EURASIP). The award reflects distinguished and sustained contributions over the recipients' career.
 2. The IEEE SPS Claude Shannon - Harry Nyquist Technical Achievement Award *for exemplary contributions to tensor decomposition, beamforming, and spectral analysis*, Dec. 2022. this is the highest technical career-level award of the IEEE Signal Processing Society.
 3. (Joint with co-PI Fu:) The IEEE Signal Processing Society 2023 Best Paper Award, for the paper , "Learning to Optimize: Training Deep Neural Networks for Interference Management," by Haoran Sun, Xiangyi Chen, Qingjiang Shi, Mingyi Hong, Xiao Fu, and Nicholas D. Sidiropoulos, IEEE Trans. on Signal Processing, Oct. 2018.
 4. The IEEE Signal Processing Society 2023 Donald G. Fink Overview Paper Award, for the paper "Tensor Decomposition for Signal Processing and Machine Learning," by Nicholas D. Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E. Papalexakis, Christos Faloutsos, IEEE Transactions on Signal Processing, July 2017.

6 Annex: Publications

Paper reprints/preprints are included in the following pages. Please treat preprints as confidential until they appear in print or electronic form.

HARVESTER: Principled Factorization-based Temporal Tensor Granularity Estimation

Ravdeep S Pasricha* Uday Singh Saini[†] Nicholas D. Sidiropoulos[‡] Fei Fang[§]
Kevin Chan[¶] Evangelos E. Papalexakis^{||}

Abstract

Given a tensor that captures temporal data, such as (user, item, time), the way that we set the granularity of the “time” mode can *make or break* our analysis of the data. If we set the granularity to be extremely fine, we end up with a very sparse and high-rank tensor which is essentially incompatible with what virtually all tensor decomposition models expect, i.e., tensors with low-rank structure, which can be expressed in some form of factorization.

Traditionally, this problem has been *avoided* by setting the granularity of the “time” to a “reasonable” aggregation (say hourly or daily intervals), an approach which has certainly served tensor analysis of temporal methods well so far. However, such an approach requires *tedious* trial-and-error experimentation across a number of such fixed aggregations, where typically the one that provides the most sensible results is retained, and furthermore it is *arbitrary*, since the optimal aggregation over time need not necessarily be uniform. In our work, we directly tackle this problem.

We introduce HARVESTER, the first principled factorization-based approach which seeks to identify the best temporal granularity of a given tensor. Unlike existing methods which follow a greedy approach, HARVESTER leverages multiple aggregated views of the tensor, and a carefully-designed optimization problem, in order to uncover an aggregation of a tensor which has a “good” structure for factor analysis or a downstream task.

We extensively evaluate HARVESTER on synthetic and real data, and demonstrate that it consistently produces tensors of very high quality, compared to the state-of-the-art, across the board for a number of different popular quality measures that have been used by the community.

1 Introduction

Tensor decomposition methods have been used to find latent structures in multi-modal data in a wide variety of applications like web link analysis [13], social network analysis [4], health care data analysis [11] and many more. When dealing with data that is temporal in nature, the granularity of data plays a crucial role in determining its usability in any data mining or machine

learning algorithms. In applications like web mining, social networks or computer network analysis, if the data is collected at very fine temporal granularity the resulting tensor can be extremely sparse and noisy, in addition to being very high-dimensional. Unfortunately, tensors of that form are typically not amenable to popular and widely-used tensor decompositions¹, which seek to identify low-rank latent structure in the data.

There have been recent works which modify the decomposition algorithm in order to accommodate temporal irregularities, such as imposing smoothness [8] or burstiness [9] in the temporal factor or, especially in the streaming case, defining a “continuous tensor model” which accommodates incoming data in an “any-time” fashion by implicitly aggregating locally over time [15]. Instead of customizing, and thus restricting, the decomposition algorithm, in this work we seek to pre-process the tensor data in order to allow for flexible downstream analysis. Note that extreme aggregation hides away the temporal information (and the model identifiability boost that comes with it), while no aggregation leaves us with very sparse and possibly noisy data, not amenable to low-rank modeling. Hence the optimal level of aggregation lies somewhere in between, and our goal is to find it.

To circumvent the problem of having a long tensor in temporal mode and no exploitable structure, the most intuitive and widely-used approach is to aggregate the temporal mode using certain *fixed window sizes*. For example converting milliseconds to seconds, seconds to minutes, minutes to hours, hours to days, and so on. This solution can work, and has been serving tensor analysis of temporal data well for many years; however, there are a few issues with it, which we seek to mitigate in our work. First, it requires a copious amount of trial-and-error experimentation to evaluate what aggregation level works the best for the given data and tensor decomposition method. Second, and more importantly,

*University of California Riverside. Email: rpassr001@ucr.edu

[†]University of California Riverside. Email: usain001@ucr.edu

[‡]University of Virginia. Email: nikos@virginia.edu

[§]Carnegie Mellon University. Email: feifang@cmu.edu

[¶]DEVCOM Army Research Laboratory Email: kevin.s.chan.civ@army.mil

^{||}University of California Riverside. Email: epalex@cs.ucr.edu

¹We use the terms “decomposition” and “factorization” interchangeably.

tensor decomposition on these fixed size aggregated tensors may result in finding “good” latent structures however there might exist more “natural” aggregation which does not follow any fixed or arbitrary window size aggregations and might provide latent structures that are missed because of fixed window aggregation. To tackle this problem, [19, 20] introduced the problem *Trapped Under Ice* and provided greedy solutions ICEBREAKER and ICEBREAKER++, which iterate over the temporal mode and decide to aggregate two timestamps together or not based on certain utility functions. However, the greedy nature of those methods is prone to suboptimal solutions and requires a lot of recursive iterations, which hurts scalability, as we demonstrate in our experiments.

In this work, we introduce a principled way of finding an optimal aggregation of the temporal mode. We take inspiration from [2, 3] which essentially tackles the inverse problem to ours: given multiple views of an aggregated tensor in different modes (temporal and non-temporal), estimate the disaggregated tensor. Motivated by that approach, we propose HARVESTER where we leverage multiple aggregated views of a tensor, derived from the tensor in the original granularity in the temporal mode, and define a decomposition-based approach for recovering the best granularity.

Our contributions in this paper are as follows:

- **Novel Problem Formulation:** We provide a new and novel way to formulate the problem in terms of a factorization problem which can be solved using optimization techniques.
- **Efficient & Scalable Algorithm:** We propose HARVESTER, which solves the formulated problem in alternating fashion using non-negative multiplicative updates.
- **Experimental Evaluation:** We demonstrate the effectiveness of HARVESTER by performing extensive experiments on synthetic and real datasets.

2 Problem Formulation

2.1 Preliminary Definitions Tensors are multidimensional arrays (of order higher than two, i.e., indexed by three or more indices) which are used to model data that are multi-aspect in nature.

Matricization of a tensor is defined as unfolding or flattening a n -mode tensor in one of its modes to form a matrix. Consider a 3-mode tensor $\mathcal{X}$ of dimension $I \times J \times K$, which can be unfolded in three ways along each mode. Unfolding in first mode is given by $\mathbf{X}_1$ of dimension $I \times JK$ and similarly unfolding in other two modes is given by $\mathbf{X}_2$ and of $\mathbf{X}_3$ of dimension $J \times IK$ and $K \times IJ$ respectively.

The **n -mode product** of a tensor $\mathcal{X}$ of dimension $I_1 \times I_2 \cdots \times I_k \cdots \times I_n$ with a matrix $\mathbf{W}$ of dimension

$I_k \times F$ is given by $\mathcal{Y} = \mathcal{X} \times_k \mathbf{W}$ where dimension of tensor $\mathcal{Y}$ is $I_1 \times I_2 \cdots \times I_{k-1} \times F \times I_{k+1} \cdots \times I_n$. This can also be extended to the matricization format as $\mathbf{Y}_k = \mathbf{W} \mathbf{X}_k$. The **CANDECOMP/PARAFAC decomposition** (henceforth referred to as CP) [7, 10] of a tensor $\mathcal{X}$ of size $I \times J \times K$ and (tensor) rank F is a sum of F rank-1 tensors (outer products of three vectors in the 3-D case) $\mathcal{X} \approx \sum_{r=1}^F a_r \circ b_r \circ c_r$, which is irreducible, in that an alternative decomposition cannot be found for $F' < F$. Here $\circ$ denotes the outer product and a_r, b_r, c_r are vectors of dimension I, J, K respectively. Another way of writing the above decomposition is $\mathcal{X} \approx [\mathbf{A}, \mathbf{B}, \mathbf{C}]$ where $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ are factor matrices of dimension $I \times F, J \times F$ and $K \times F$ respectively such that each outer product of the r^{th} column of $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ forms one of the rank one tensors involved in the CP decomposition.

For more in-depth details on tensors and tensor decompositions we refer readers to the popular surveys in the literature [12, 22, 18].

2.2 Problem Definition Consider a tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ with extremely fine raw temporal granularity, and which is very long and sparse/noisy in the temporal mode (in this example we consider the third mode as the temporal mode) which makes it ill-suited for compact modeling via CP decomposition. Our problem is:

Given a tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ and multiple views of the tensor $\mathcal{X}, \mathcal{Y}^1 \in \mathbb{R}^{I \times J \times K_1}$ and $\mathcal{Y}^2 \in \mathbb{R}^{I \times J \times K_2}$, where $K_2 < K_1 < K$. Find an aggregated tensor $\mathcal{Y} \in \mathbb{R}^{I \times J \times K^*}$ where $K^* < K_2 < K_1 < K$, such that its temporal mode latent factor matrix is an optimal low-rank compression of the original temporal mode latent factor, while respecting the temporal sequence of indices in that mode.

Creating Views: To give a concrete example, let's assume $\mathcal{X}$ is of size $I \times J \times 9$ and if we want to aggregate the third mode on some fixed window size, say 3, so $\mathbf{W}$ is of size 3×9 and takes the form:

$$\mathbf{W} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

The resulting tensor will be of size $I \times J \times 3$, where the first 3 slices (1 to 3) are aggregated into one slice, the next 3 slices (4 to 6) into a second slice and last three slices (7 to 9) into the third slice.

3 Proposed Method

Consider a three mode tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ where I, J denote the dimensions of the first two modes and K that of the third mode which is temporal in nature and

has fine granularity, making tensor $\mathcal{X}$ extremely sparse and having no exploitable latent structure. The CP decomposition of tensor $\mathcal{X}$ for a rank F is given by:

$$(3.1) \quad \mathcal{X} \approx [\mathbf{A}, \mathbf{B}, \mathbf{C}]$$

Where $\mathbf{A} \in \mathbb{R}^{I \times F}$, $\mathbf{B} \in \mathbb{R}^{J \times F}$ and $\mathbf{C} \in \mathbb{R}^{K \times F}$.

Let us assume there exists an oracle tensor $\mathcal{Y}$, which has perfect latent structure, which is created by the third mode-product of the tensor $\mathcal{X}$ with a matrix $\mathbf{W}_{opt}$ which has special structure as specified in Section 2.2

$$(3.2) \quad \mathcal{Y} = \mathcal{X} \times_3 \mathbf{W}_{opt}$$

$\mathbf{W}_{opt}$ is of size $K_{opt} \times K$ so the dimension of $\mathcal{Y}$ is $I \times J \times K_{opt}$, such that $K_{opt} \ll K$. The CP decomposition of oracle tensor $\mathcal{Y}$ is given by

$$(3.3) \quad \mathcal{Y} \approx [\mathbf{A}, \mathbf{B}, \mathbf{C}_{opt}]$$

where matrix $\mathbf{C}_{opt}$ is of size $K_{opt} \times F$

The key assumption we make is that $\mathbf{C}$ can be obtained by some dimensionality-expanding linear transformation of the unknown (latent) $\mathbf{C}_{opt}$. This is tantamount to assuming that there exists low rank matrix structure in the temporal mode.

$$(3.4) \quad \mathbf{C} = \mathbf{M}\mathbf{C}_{opt}$$

where $\mathbf{M} \in \mathbb{R}^{K \times K_{opt}}$ and $\mathbf{M}$ is a tall matrix.

We create $m \geq 2$ views of tensor $\mathcal{X}$ as follows:

$$(3.5) \quad \mathcal{Y}^1 = \mathcal{X} \times_3 \mathbf{W}_1 = [\mathbf{A}, \mathbf{B}, \mathbf{C}_1]; \mathbf{C}_1 = \mathbf{W}_1 \mathbf{C}$$

$$(3.6) \quad \mathcal{Y}^2 = \mathcal{X} \times_3 \mathbf{W}_2 = [\mathbf{A}, \mathbf{B}, \mathbf{C}_2]; \mathbf{C}_2 = \mathbf{W}_2 \mathbf{C}$$

⋮

$$(3.7) \quad \mathcal{Y}^m = \mathcal{X} \times_3 \mathbf{W}_m = [\mathbf{A}, \mathbf{B}, \mathbf{C}_m]; \mathbf{C}_m = \mathbf{W}_m \mathbf{C}$$

Before we formalize the problem, we need to address some issues regarding the problem

1. Since we don't have access to the oracle tensor $\mathcal{Y}$ we cannot directly estimate $\mathbf{C}_{opt}$.
2. Most importantly, unlike in the case of [2, 3], where the "right" K_{opt} is known beforehand, here we don't know the size K_{opt} for the matrix $\mathbf{C}_{opt}$.

Thus, we seek to estimate $\tilde{\mathbf{C}}$ as shown in equation 3.9, which is supposed to be good approximation of $\mathbf{C}_{opt}$. Using equation 3.4 and the fact that $\mathbf{C}_i = \mathbf{W}_i \mathbf{C}$ per equations 3.5-3.7, we have,

$$(3.8) \quad \mathbf{C}_i = \mathbf{W}_i \mathbf{M} \mathbf{C}_{opt}$$

In the above equation, we know $\mathbf{C}_i$ and $\mathbf{W}_i$, but we do not know $\mathbf{M}$ and $\mathbf{C}_{opt}$ – not even K_{opt} . As a result, we do not know the product $\mathbf{W}_i \mathbf{M}$. How can we bypass this seemingly insurmountable challenge? We propose to bring in ideas from sparse regression. The idea is to over-parameterize the product $\mathbf{W}_i \mathbf{M}$ as a new matrix variable $\mathbf{P}_i$, and use a sparse diagonal row-selection matrix to pick up the essential (reduced-dimension) row span of $\mathbf{C}_{opt}$. In more detail, we approximate

$$(3.9) \quad \mathbf{C}_i \approx \mathbf{P}_i \mathbf{\Lambda} \tilde{\mathbf{C}}$$

where $\mathbf{P}_i$ is of size $K_i \times K$, $\tilde{\mathbf{C}}$ is of size $K \times F$, and $\mathbf{\Lambda}$ is a sparse diagonal matrix of size $K \times K$, whose zero entries are meant to strike out columns of $\mathbf{P}_i$ and corresponding rows of $\tilde{\mathbf{C}}$. The product $\mathbf{P}_i \mathbf{\Lambda}$ is a proxy for $\mathbf{W}_i \mathbf{M}$, with the diagonal $\mathbf{\Lambda}$ effectively providing us with a tall matrix of *a priori* unknown number of columns K_{opt} .

Using equation 3.9, our optimization problem is:

$$(3.10) \quad \mathcal{L} = \min_{\tilde{\mathbf{C}}, \mathbf{P}_i, \mathbf{\Lambda}} \sum_{i=1}^m \|\mathbf{C}_i - \mathbf{P}_i \mathbf{\Lambda} \tilde{\mathbf{C}}\|_F^2 + \alpha \|\tilde{\mathbf{C}}\|_F^2 + \beta \|\mathbf{\Lambda}\|_1 + \gamma \|\mathbf{P}_i\|_F^2$$

We optimize equation 3.10 for matrices $\tilde{\mathbf{C}}$, $\mathbf{\Lambda}$ and $\mathbf{P}_i$. the sparsity or L-1 regularization constraint on $\mathbf{\Lambda}$ is there to help reduce the granularity of the the $\tilde{\mathbf{C}}$. The non-zero entries in the diagonal of $\mathbf{\Lambda}$ control the final granularity of the resulting tensor, thus we impose a sparsity penalty on the diagonal.

Discussion: Since we are trying to find the best aggregation in the third mode and we assume that there exists a low dimensional structure in that mode, one may ask why not just perform singular value decomposition (SVD) on the third mode matricization of the tensor and use the top-k right singular vectors as the basis for the aggregation, which would yield the best compression in the least squares sense.

$$[\mathbf{U}, \mathbf{S}, \mathbf{V}] = \text{svd}(\mathbf{X}_3)$$

$$\mathcal{Y} = \mathcal{X} \times_3 \mathbf{V}_k$$

In our case, this does not solve the problem we set out to tackle. Even though there may be low-rank structure in the temporal mode, captured by the aggregation matrix $\mathbf{V}_k$, this aggregation does not respect the temporal sequence of slices, therefore is not appropriate for recovering the type of aggregation we are interested in.

3.1 Proposed method: Harvester In this section we present our method HARVESTER to solve the problem in equation 3.10. We solve it using non negative multiplicative update (NMU) [14] in an alternating fashion. Here, we derive the update steps only using two views but this can easily be extended to more views.

Using tensor $\mathbf{X}$ we create two views for that tensor namely $\mathbf{Y}^1$ and $\mathbf{Y}^2$ which have lower resolution than the original tensor. To create these views we multiply the third mode of the tensor with a $\mathbf{W}_1$ and $\mathbf{W}_2$ matrix.

$$(3.11) \quad \mathbf{Y}^1 = \mathbf{X} \times_3 \mathbf{W}_1$$

$$(3.12) \quad \mathbf{Y}^2 = \mathbf{X} \times_3 \mathbf{W}_2$$

We create a tensor $\mathbf{Z}$ by stacking tensor $\mathbf{Y}^1$ and $\mathbf{Y}^2$ onto one another hence making them coupled in the first two modes. The dimension of tensor $\mathbf{Z}$ is $I \times J \times (K_1 + K_2)$. Then we perform CP decomposition on tensor $\mathbf{Z}$ which yields factor matrices needed for our optimization algorithm. We split factor matrix $\mathbf{D}$ into $\mathbf{C}_1$ and $\mathbf{C}_2$ based on their respective dimensions as shown below:

$$(3.13) \quad \begin{aligned} \mathbf{Z} &= [\mathbf{Y}^1; \mathbf{Y}^2] \\ \mathbf{Z} &\approx [\mathbf{A}, \mathbf{B}, \mathbf{D}] \\ \mathbf{C}_1 &= \mathbf{D}(1 : K_1, :) \\ \mathbf{C}_2 &= \mathbf{D}(K_1 + 1 : K_2, :) \end{aligned}$$

We use equations 3.9 and 3.13 to create our optimization problem that follows from equation 3.10. The optimization steps to solve equation 3.14 are derived along the lines of [21].

$$(3.14) \quad \mathcal{L} = \min_{\tilde{\mathbf{C}}, \mathbf{P}_1, \mathbf{P}_2, \Lambda} \underbrace{\|\mathbf{C}_1 - \mathbf{P}_1 \Lambda \tilde{\mathbf{C}}\|_F^2}_f + \underbrace{\|\mathbf{C}_2 - \mathbf{P}_2 \Lambda \tilde{\mathbf{C}}\|_F^2}_g + \underbrace{\alpha \|\tilde{\mathbf{C}}\|_F^2 + \beta \|\Lambda\|_1 + \gamma \|\mathbf{P}_1\|_F^2 + \gamma \|\mathbf{P}_2\|_F^2}_h$$

The gradient of $\mathcal{L}$ in eq. 3.14 w.r.t. $\tilde{\mathbf{C}}$:

$$(3.15) \quad \begin{aligned} \frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{C}}} &= \frac{\partial f}{\partial \tilde{\mathbf{C}}} + \frac{\partial g}{\partial \tilde{\mathbf{C}}} + \frac{\partial h}{\partial \tilde{\mathbf{C}}} \\ \frac{\partial f}{\partial \tilde{\mathbf{C}}} &= \frac{\partial}{\partial \tilde{\mathbf{C}}} (\|\mathbf{C}_1 - \mathbf{P}_1 \Lambda \tilde{\mathbf{C}}\|_F^2) \\ \frac{\partial f}{\partial \tilde{\mathbf{C}}} &= \frac{\partial}{\partial \tilde{\mathbf{C}}} \text{Tr}((\mathbf{C}_1 - \mathbf{P}_1 \Lambda \tilde{\mathbf{C}})(\mathbf{C}_1 - \mathbf{P}_1 \Lambda \tilde{\mathbf{C}})^T) \\ \frac{\partial f}{\partial \tilde{\mathbf{C}}} &= -\Lambda^T \mathbf{P}_1^T \mathbf{C}_1 - \Lambda^T \mathbf{P}_1^T \mathbf{C}_1 + \Lambda^T \mathbf{P}_1^T \mathbf{P}_1 \Lambda \tilde{\mathbf{C}} + \Lambda^T \mathbf{P}_1^T \mathbf{P}_1 \Lambda \tilde{\mathbf{C}} \\ \frac{\partial f}{\partial \tilde{\mathbf{C}}} &= 2\Lambda^T \mathbf{P}_1^T \mathbf{P}_1 \Lambda \tilde{\mathbf{C}} - 2\Lambda^T \mathbf{P}_1^T \mathbf{C}_1 \end{aligned}$$

We can similarly derive $\frac{\partial g}{\partial \tilde{\mathbf{C}}}$,

$$(3.16) \quad \frac{\partial g}{\partial \tilde{\mathbf{C}}} = 2\Lambda^T \mathbf{P}_2^T \mathbf{P}_2 \Lambda \tilde{\mathbf{C}} - 2\Lambda^T \mathbf{P}_2^T \mathbf{C}_2$$

Gradient of h w.r.t $\tilde{\mathbf{C}}$

$$(3.17) \quad \begin{aligned} \frac{\partial h}{\partial \tilde{\mathbf{C}}} &= \frac{\partial}{\partial \tilde{\mathbf{C}}} \text{Tr}(\alpha \tilde{\mathbf{C}}^T \tilde{\mathbf{C}}) \\ &= 2\alpha \tilde{\mathbf{C}} \end{aligned}$$

Combining equation 3.15, 3.16 and 3.17

$$(3.18) \quad \frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{C}}} = 2(\Lambda^T \mathbf{P}_1^T \mathbf{P}_1 \Lambda \tilde{\mathbf{C}} + \Lambda^T \mathbf{P}_2^T \mathbf{P}_2 \Lambda \tilde{\mathbf{C}} - \Lambda^T \mathbf{P}_1^T \mathbf{C}_1 - \Lambda^T \mathbf{P}_2^T \mathbf{C}_2 + \alpha \tilde{\mathbf{C}})$$

Setting $\frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{C}}}$ to zero and using non-negative multiplicative update method to compute update step. $\otimes$ here represents element wise product and fraction equation below represents element wise division.

$$(3.19) \quad \tilde{\mathbf{C}} \leftarrow \tilde{\mathbf{C}} \otimes \frac{\Lambda^T \mathbf{P}_1^T \mathbf{C}_1 + \Lambda^T \mathbf{P}_2^T \mathbf{C}_2}{(\Lambda^T \mathbf{P}_1^T \mathbf{P}_1 \Lambda + \Lambda^T \mathbf{P}_2^T \mathbf{P}_2 \Lambda + \alpha I) \tilde{\mathbf{C}}}$$

The gradient of $\mathcal{L}$ in eq. 3.14 w.r.t $\mathbf{P}_1$ (or $\mathbf{P}_2$):

$$(3.20) \quad \begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{P}_1} &= \frac{\partial f}{\partial \mathbf{P}_1} + \frac{\partial g}{\partial \mathbf{P}_1} + \frac{\partial h}{\partial \mathbf{P}_1} \\ \frac{\partial f}{\partial \mathbf{P}_1} &= \frac{\partial}{\partial \mathbf{P}_1} (\|\mathbf{C}_1 - \mathbf{P}_1 \Lambda \tilde{\mathbf{C}}\|_F^2) \\ \frac{\partial f}{\partial \mathbf{P}_1} &= \frac{\partial}{\partial \mathbf{P}_1} \text{Tr}((\mathbf{C}_1 - \mathbf{P}_1 \Lambda \tilde{\mathbf{C}})(\mathbf{C}_1 - \mathbf{P}_1 \Lambda \tilde{\mathbf{C}})^T) \\ &= 2(\mathbf{P}_1 \Lambda \tilde{\mathbf{C}} \tilde{\mathbf{C}}^T \Lambda^T - \mathbf{C}_1 \tilde{\mathbf{C}}^T \Lambda^T) \end{aligned}$$

Gradient of g with respect to $\mathbf{P}_1$ is zero. Now computing gradient of h with respect to $\mathbf{P}_1$

$$(3.21) \quad \begin{aligned} \frac{\partial h}{\partial \mathbf{P}_1} &= \frac{\partial}{\partial \mathbf{P}_1} \text{Tr}(\gamma \mathbf{P}_1 \mathbf{P}_1^T) \\ &= 2\gamma \mathbf{P}_1 \end{aligned}$$

Combining equation 3.20 and 3.21

$$(3.22) \quad \frac{\partial \mathcal{L}}{\partial \mathbf{P}_1} = 2(\mathbf{P}_1 \Lambda \tilde{\mathbf{C}} \tilde{\mathbf{C}}^T \Lambda^T - 2\mathbf{C}_1 \tilde{\mathbf{C}}^T \Lambda^T + 2\gamma \mathbf{P}_1)$$

Setting $\frac{\partial \mathcal{L}}{\partial \mathbf{P}_1}$ to zero and using non-negative multiplicative update method to compute update step. Again $\otimes$ here represents element wise product and fraction equation below represents element wise division.

$$(3.23) \quad \mathbf{P}_1 \leftarrow \mathbf{P}_1 \otimes \frac{\mathbf{C}_1 \tilde{\mathbf{C}}^T \Lambda^T}{\mathbf{P}_1 (\Lambda \tilde{\mathbf{C}} \tilde{\mathbf{C}}^T \Lambda^T + \gamma \mathbf{I})}$$

We can compute update step for $\mathbf{P}_2$ similarly,

$$(3.24) \quad \mathbf{P}_2 \leftarrow \mathbf{P}_2 \otimes \frac{\mathbf{C}_2 \tilde{\mathbf{C}}^T \Lambda^T}{\mathbf{P}_2 (\Lambda \tilde{\mathbf{C}} \tilde{\mathbf{C}}^T \Lambda^T + \gamma \mathbf{I})}$$

The gradient of $\mathcal{L}$ in eq. 3.14 with respect to Λ :

$$(3.25) \quad \begin{aligned} \frac{\partial \mathcal{L}}{\partial \Lambda} &= \frac{\partial f}{\partial \Lambda} + \frac{\partial g}{\partial \Lambda} + \frac{\partial h}{\partial \Lambda} \\ \frac{\partial f}{\partial \Lambda} &= \frac{\partial}{\partial \Lambda} (\|\mathbf{C}_1 - \mathbf{P}_1 \Lambda \tilde{\mathbf{C}}\|_F^2) \\ \frac{\partial f}{\partial \Lambda} &= \frac{\partial}{\partial \Lambda} \text{Tr}((\mathbf{C}_1 - \mathbf{P}_1 \Lambda \tilde{\mathbf{C}})(\mathbf{C}_1 - \mathbf{P}_1 \Lambda \tilde{\mathbf{C}})^T) \\ &= 2(\mathbf{P}_1^T \mathbf{P}_1 \Lambda \tilde{\mathbf{C}} \tilde{\mathbf{C}}^T - \mathbf{P}_1^T \mathbf{C}_1 \tilde{\mathbf{C}}^T) \end{aligned}$$

We can similarly derive the derivative of g w.r.t. Λ ,

$$(3.26) \quad \frac{\partial g}{\partial \Lambda} = 2(\mathbf{P}_2^T \mathbf{P}_2 \Lambda \tilde{\mathbf{C}} \tilde{\mathbf{C}}^T - \mathbf{P}_2^T \mathbf{C}_2 \tilde{\mathbf{C}}^T)$$

Next, based on the update steps in Section 2 of [14], we compute the Gradient of h w.r.t. Λ

$$(3.27) \quad \frac{\partial h}{\partial \Lambda} = \frac{\partial}{\partial \Lambda} \|\beta \Lambda\|_1 = \beta \mathbf{E}$$

where $\mathbf{E}$ is a matrix of all ones. We would like to reiterate that, although L-1 Norm is not a differentiable function in the entirety of its domain and thus we must instead compute its sub-gradient, the constraint of non-negativity on Λ allows us to circumvent that issue and compute the gradient as long as values of Λ are not negative (which we guarantee by non-negative initialization and subsequent iterations).

Combining equations 3.25, 3.26 and 3.27

$$(3.28) \quad \frac{\partial \mathcal{L}}{\partial \Lambda} = 2\mathbf{P}_1^T \mathbf{P}_1 \Lambda \tilde{\mathbf{C}} \tilde{\mathbf{C}}^T - 2\mathbf{P}_1^T \mathbf{C}_1 \tilde{\mathbf{C}}^T + 2\mathbf{P}_2^T \mathbf{P}_2 \Lambda \tilde{\mathbf{C}} \tilde{\mathbf{C}}^T - 2\mathbf{P}_2^T \mathbf{C}_2 \tilde{\mathbf{C}}^T + \beta \mathbf{E}$$

Setting $\frac{\partial \mathcal{L}}{\partial \Lambda}$ to zero, we use the non-negative multiplicative update method to compute the update step. $\circledast$ represents element wise product and fraction equation below represents element wise division.

$$(3.29) \quad \Lambda \leftarrow \Lambda \circledast \frac{2\mathbf{P}_1^T \mathbf{C}_1 \tilde{\mathbf{C}}^T + 2\mathbf{P}_2^T \mathbf{C}_2 \tilde{\mathbf{C}}^T}{2\mathbf{P}_1^T \mathbf{P}_1 \Lambda \tilde{\mathbf{C}} \tilde{\mathbf{C}}^T + 2\mathbf{P}_2^T \mathbf{P}_2 \Lambda \tilde{\mathbf{C}} \tilde{\mathbf{C}}^T + \beta \mathbf{E}}$$

Aggregation strategies So far, we have computed Λ , $\mathbf{C}$, $\mathbf{P}_1$ and $\mathbf{P}_2$ which are needed to construct the aggregated tensor. We can construct the aggregated tensor in the following ways, given $\mathbf{C}_o = \Lambda \tilde{\mathbf{C}}$

-aggregateOnNorm: We use one of the techniques from [20], namely norm aggregation and apply it to matrix instead of tensor. Essentially we iterate over the rows of $\mathbf{C}_o$, add a candidate row to previous row, if the rate of change of norm between the previous row and sum of previous and candidate row is more than a certain threshold. If not, the candidate row becomes the previous row and process continues until we reach the end. We use that norm aggregated matrix $\mathbf{C}_{norm}$ to construct a tensor.

$$\begin{aligned} \mathbf{W}_{norm} &= \text{aggregateOnNorm}(\mathbf{C}_o, \text{normThreshold}) \\ \mathbf{C}_{norm} &= \mathbf{W}_{norm} \mathbf{C}_o \\ \mathbf{y}^{norm} &\approx [\mathbf{A}, \mathbf{B}, \mathbf{C}_{norm}] \end{aligned}$$

-aggregateZeroRows: Since Λ is not only a diagonal matrix but it's also a diagonally sparse matrix, which is used in update steps to turn off (zero out) the rows of $\tilde{\mathbf{C}}$. We can just use those non zero rows as anchor

points to create a matrix $\mathbf{C}_{spar}$ which contains only the non-zero rows of $\mathbf{C}_o$.

$$\begin{aligned} \mathbf{W}_{spar} &= \text{aggregateZeroRows}(\mathbf{C}_o) \\ \mathbf{C}_{spar} &= \mathbf{W}_{spar} \mathbf{C}_o \\ \mathbf{y}^{spar} &\approx [\mathbf{A}, \mathbf{B}, \mathbf{C}_{spar}] \end{aligned}$$

We use tensors $\mathbf{y}^{norm}$ and $\mathbf{y}^{spar}$ for evaluation.

4 Experimental Evaluation

We implemented our method² in Matlab using tensor toolbox [5] and for using CP-WOPT [1] algorithm in tensor toolbox for tensor completion we make use of L-BFGS-B Matlab wrapper³.

Hyperparameter Tuning: Our problem formulation has three hyperparameters which need to be tuned namely α , β and γ as in equation 3.14. We use different ranges for the hyperparameters search space used in equation 3.14 based on the dataset, we specify those ranges while discussing the results for those datasets in their respective sub sections. We perform grid search over these values and choose two points from this grid search results based on following conditions:

Sparsity threshold: Number of non zero entries in diagonal of lambda (Λ) matrix is less than some threshold. For synthetic dataset we set that threshold to be 1/10 of the third dimension of $\mathbf{X}$, e.g., for $\mathbf{X}$ being $100 \times 100 \times 500$, the threshold is set at $500/10 = 50$.

Error threshold: From all the points that meet the above threshold condition, we find the median error and only consider points with error less than that threshold.

After applying the above two thresholds, we choose two points, one which has smallest error on the loss function (min error point) and other one which have the lowest number of non-zero entries on the diagonal of the lambda (Λ) matrix (min sparsity point). We use these two points for our analysis. For each of these points we create two tensors namely $\mathbf{y}_{norm}$ and $\mathbf{y}_{spar}$ as mentioned in previous section. In total, HARVESTER produces 4 candidate tensors for evaluation namely min error $\mathbf{y}_{norm}$, min error $\mathbf{y}_{spar}$, min sparsity $\mathbf{y}_{norm}$, and min sparsity $\mathbf{y}_{spar}$. The result presented in this work for our method HARVESTER will be of one or more of those points unless specified otherwise.

4.1 Evaluation Metrics and Baselines We evaluate our method HARVESTER on two model-based measures namely relative fit of the model and CORCONDIA [6, 17] and also on a task-based measure on tensor completion [1]. In order to choose the “best” rank for the

²<https://github.com/ravdeep003/harvester-sdm>

³<https://github.com/stephenbeckr/L-BFGS-B-C>

tensor at hand, we use AutoTen [17] (modified with non-negativity constraints). Given the rank identified by AutoTen, we compute the following quality measures:

Relative Fit of the model: We evaluate the fit of non-negative CP using the AutoTen rank:

$$(4.30) \text{ Relative Fit} = 1 - \left(\frac{\|\mathbf{X}_{Input} - \mathbf{X}_{computed}\|_F}{\|\mathbf{X}_{Input}\|_F} \right)$$

CORCONDIA: The Core Consistency Diagnostic (CORCONDIA) [6] returns a score for how well the given factors obey the CP model, given a tensor and a decomposition at a certain rank. AutoTen, uses CORCONDIA to arrive to the selected rank, so we are essentially using the value reported by AutoTen.

Tensor Completion: This can be viewed as downstream task oriented evaluation, in which we hide 20% of non-zero entries and use CP_WOPT [1] for predicting the missing values. We report RMSE between the predicted and actual hidden values. The lower the RMSE the better the performance on the tensor completion task. For consistency with the previous two measures, for which higher is better, we compute $1/RMSE$.

Baselines: We compare against the following:

IceBreaker++: We compare our method HARVESTER against the greedy based method ICEBREAKER++ [20], which introduced the problem of *Trapped Under Ice* and provided a greedy approach to tackle the problem.

Fixed aggregations: We also compare our method against fixed aggregations views generated from the input tensor, which just aggregates the slices based on certain fixed window intervals.

4.2 Synthetic data generation We created the synthetic datasets in a similar fashion as mentioned in the [20]. We first create a sparse tensor of certain sparsity and then we distribute the non-zero entries in each frontal slice $\mathbf{X}(:, :, k)$ over certain fixed number of slices (B). We repeat this for every frontal slice in original tensor, every time distributing the non-zero entries to new fixed bucket of slices(B) and concatenate all the buckets to create the dataset. Table 1 shows the different synthetic datasets used in the experiments. For example take SD1 (synthetic dataset 1), the original data is of size $100 \times 100 \times 10$, all the non-zero entries in the k -th slice $\mathbf{X}(i, j, k)$ are randomly distributed over 50 slices in the third mode with the same i, j indices.

We create 4 types of dataset as shown in table 1, for each type of dataset we create 10 datasets and we report our findings for these 40 datasets. Hyperparameter ranges used are : for α we use the values $[10^{-4}, 10^{-3}, 10^{-2}]$, for β we use $[10^{-2}, 10^{-1}, 10^0, 10]$ and for γ we use $[10^{-4}, 10^{-3}, 10^{-2}]$.

4.3 Pareto-based evaluation for synthetic data

We evaluate the quality of our method HARVESTER and baseline methods on the three evaluation criteria as mentioned in section 4.1. We are seeking a tensor which maximizes CP_FIT, CORCONDIA and minimizes the RMSE for tensor completion (or maximize $1/RMSE$). This is essentially a multi-objective optimization problem, and a widely accepted means of measuring optimality in this case is to identify points in that 3D space which lie in the so-called *Pareto frontier* [16], i.e., points which dominate the rest of the points with respect to the three dimensions/objectives. To that effect, we find the tensors which lie on the Pareto frontier of these evaluation metrics, and we count how frequently a given method produces such tensors. The more frequently a method yields tensors lying on the Pareto frontier, the better the quality of those tensors, thus, the higher performing the method. We perform Pareto finding operation on all the data points obtained using method HARVESTER, ICEBREAKER++, fixed aggregation views of the tensor which we used with our method. In finding Pareto-optimal points we excluded the original tensor.

To measure the effectiveness of our method HARVESTER against baselines we count the number of times each method is on the Pareto frontier for all the synthetic datasets we evaluated on and we present that information in table 2. We observe that points from our method HARVESTER are always on the Pareto frontier with variety of the synthetic datasets. As shown in table 2, HARVESTER performs far better than the baseline methods. In all of the synthetic datasets, we see that tensor generated by HARVESTER was found on the Pareto frontier. In SD4 ICEBREAKER++ did reasonably well, as it was on the Pareto frontier six times as opposed to HARVESTER which was on the Pareto frontier for all ten datasets.

4.3.1 Rank Sensitivity In this section, we explore the sensitivity of our algorithm HARVESTER with respect to different ranks. We run this experiment using a dataset of size $100 \times 100 \times 10$ generated using rank 10, which was then used to generate a fine grained tensor of size $100 \times 100 \times 500$, using the same method as described in section 4.3. In Figure 1(a), 1(b) and 1(c) we report mean and standard deviation of Corcondia, Relative fit and RMSE for the dataset over 5 runs. We observe that HARVESTER has a steady performance for the initial rank increase but further increase in the rank deteriorate some of the evaluation metrics.

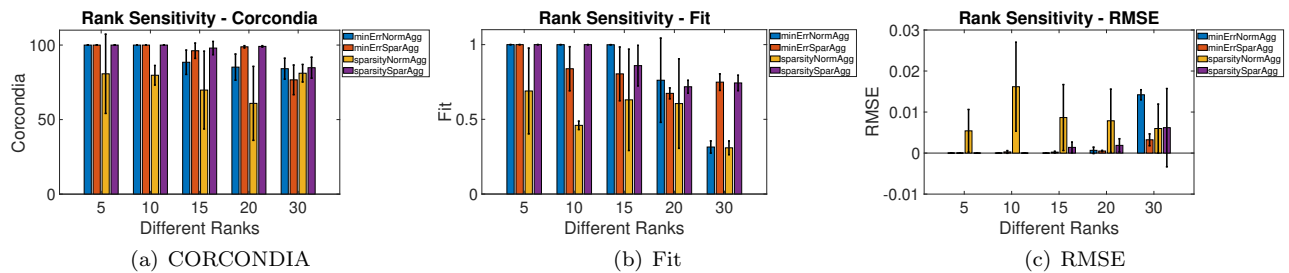
4.4 Scalability Analysis In this section we show the scalability analysis for HARVESTER as the third mode grows, since HARVESTER depends on the views

Dataset	Original Dimension ($\mathcal{X}_{\text{og}}$)	Rank (R)	Bucket size (B)	Approximate Final Dimension ($\mathcal{X}$)	View 1 ($\mathcal{Y}^1$)	View 2 ($\mathcal{Y}^2$)	Number of datasets
SD1	$100 \times 100 \times 10$	5	50	$100 \times 100 \times 500$	$100 \times 100 \times 100$	$100 \times 100 \times 50$	10
SD2	$100 \times 100 \times 10$	10	50	$100 \times 100 \times 500$	$100 \times 100 \times 100$	$100 \times 100 \times 50$	10
SD3	$100 \times 100 \times 10$	30	50	$100 \times 100 \times 500$	$100 \times 100 \times 100$	$100 \times 100 \times 50$	10
SD4	$100 \times 100 \times 100$	20	20	$100 \times 100 \times 2000$	$100 \times 100 \times 400$	$100 \times 100 \times 200$	10

Table 1: Table of Synthetic Datasets analyzed

Dataset	Original Dimension ($\mathcal{X}_{\text{og}}$)	Approximate Final Dimension ($\mathcal{X}$)	Harvester Count	IceBreaker Count	Views Count
SD1	$100 \times 100 \times 10$	$100 \times 100 \times 500$	10	0	0
SD2	$100 \times 100 \times 10$	$100 \times 100 \times 500$	10	0	0
SD3	$100 \times 100 \times 10$	$100 \times 100 \times 500$	10	1	0
SD4	$100 \times 100 \times 100$	$100 \times 100 \times 2000$	10	6	0

Table 2: How many times each method was on the Pareto frontier

Figure 1: HARVESTER sensitivity as a function of the rank on a $100 \times 100 \times 10$ dataset with true rank 10

of the original tensor. We run HARVESTER on the different settings one with coarser views and other one with finer views. In Figure 2(a), we report time taken by HARVESTER for both of these inputs. When the third mode is smaller both aggregations roughly take the same time, but as the dimension grows we see that HARVESTER with coarser aggregation runs faster.

We also evaluated the scalability of HARVESTER against ICEBREAKER++. Figure 2(b) shows the time taken by each method. Since ICEBREAKER++ depends on the performance of the utility functions, we ran ICEBREAKER++ without the missing value prediction utility function (the most time consuming utility function used in that work). We also ran ICEBREAKER++ with one iteration of missing value prediction utility function as opposed to 10 times reported in the paper [20]. We compared it with HARVESTER with both coarser and fine grained views. As shown in Figure 2(b), HARVESTER is *significantly faster* than ICEBREAKER++.

Finally, we evaluate the behavior of HARVESTER as the rank which is used to compute the coupled tensor factorization to set up the problem with $\mathbf{C}_i$'s increases. For an input dataset to HARVESTER of size $100 \times 100 \times 500$ we ran it 10 times for each rank and report runtimes in Figure 2(c). We observe that

HARVESTER scales linearly with respect to rank.

4.5 Real-world case study: Foursquare Dataset

We use the foursquare dataset⁴ [25] for check-ins in New York City collected over 10 months from 12 April, 2012 to 16 February, 2013. We constructed a 3-mode tensor with user, venue categories and time as the three modes where each entry in tensor is number of check-in for a user for a particular venue category for a given day. We aggregated temporal mode on the daily basis to create an input tensor and we used top 126 venue categories (venue categories which had more check-ins than the median check-in value of each categories). We, thus, end up with a tensor of size $1083 \times 126 \times 250$ (user, venue categories, days). We create two views from the input tensor of size $1083 \times 126 \times 125$ and $1083 \times 126 \times 50$ respectively. These two views were used by HARVESTER to find tensor of optimal granularity. By conducting the same Pareto-based evaluation, we observed that *only HARVESTER tensors are on the Pareto frontier*.

In the interest of space, we present a small indicative set of the latent factors extracted from one of those

⁴<https://www.kaggle.com/datasets/chetanism/foursquare-nyc-and-tokyo-checkin-dataset>

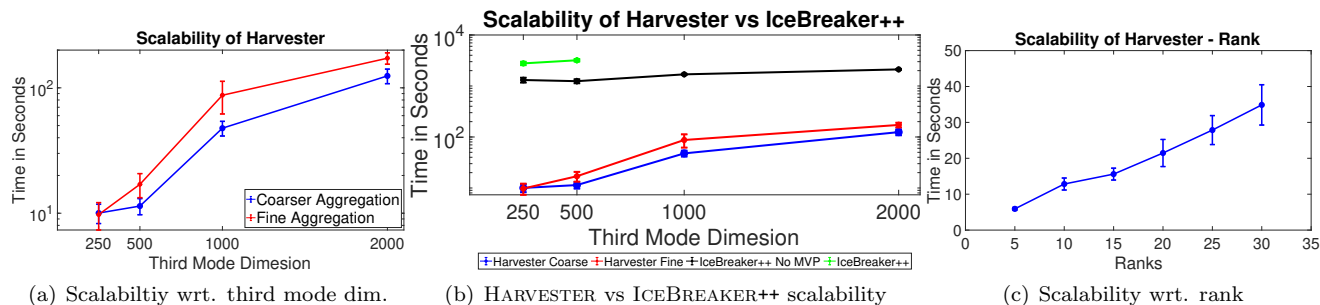


Figure 2: Scalability Analysis

HARVESTER tensors. Figure 3 shows three indicative latent factors, where we highlight the venues mostly associated with that factor with high intensity in a word cloud, alongside the temporal profile of that latent factor. In Figure 3(a) we observe venues which are mostly associated with workplaces, and the corresponding temporal profile shows activity which dips during vacation periods such as the summer. Fig 3(b) appears to have clustered various parks and venues related to outdoor activity, with an overwhelming spike during one of the Springs in our dataset, which is a popular season for outdoor activities in NYC. Finally, the last component in Figure 3(c) has identified what appears to be areas where college students congregate and study, with activity that aligns with times during the academic year.

It is important to note here without any external information about different events that coincide with the activity observed, we can only make speculations about the spikes and fluctuations we observe in the temporal activity of different components. The patterns we show in Figure 3 are intuitive and lend themselves to simple explanation by non experts, a fact which lends more credence to the quality of the components extracted. A platform like Foursquare, or different vendors, however, can use such extracted components to understand different spatio-temporal trends over the City, and potentially integrate those components with other information (such as deals, specials, etc) which is not available to us and would provide more insight for time points identified by the decomposition.

5 Related Work

In addition to existing tensor works which we reviewed in the Introduction [8, 15], a neighboring problem to ours is the aggregation of temporal edges into a mature graph [23, 24]. However, those works are specific to graphs and do not generalize to arbitrary tensor data. Closest to our work is [19, 20], which provides a greedy solution to our problem. However to the best of our knowledge we are the first work to introduce a principled



Figure 3: Indicative latent factors for Foursquare way to finding an optimal aggregation in the temporal mode using aggregated views of the input data and formalizing it in terms of an optimization problem.

6 Conclusions

We developed HARVESTER, a factorization-based tensor temporal granularity estimation, which uses multiple aggregated views of an input tensor and identifies the best aggregation. We extensively evaluate HARVESTER and demonstrate that it generates tensors of superior quality to the current state of the art.

7 Acknowledgements

Research was supported by the National Science Foundation under CAREER grant IIS 2046086 and grant IIS-1908070, by the Army Research Office W911NF19-1-0407, and by the U.S. Army Combat Capabilities Development Command Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-13-2-0045 (ARL Cyber Security CRA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Combat Capabilities Development Command Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

- [1] Evrim Acar, Daniel M Dunlavy, Tamara G Kolda, and Morten Mørup. Scalable tensor factorizations for incomplete data. *Chemometrics and Intelligent Laboratory Systems*, 106(1):41–56, 2011.
- [2] Faisal M Almutairi, Charilaos I Kanatsoulis, and Nicholas D Sidiropoulos. Tendi: Tensor disaggregation from multiple coarse views. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 867–880. Springer, 2020.
- [3] Faisal M Almutairi, Charilaos I Kanatsoulis, and Nicholas D Sidiropoulos. Prema: Principled tensor data recovery from multiple aggregated views. *IEEE Journal of Selected Topics in Signal Processing*, 15(3):535–549, 2021.
- [4] Miguel Araujo, Spiros Papadimitriou, Stephan Günnemann, Christos Faloutsos, Prithwish Basu, Ananthram Swami, Evangelos E Papalexakis, and Danaï Koutra. Com2: fast automatic discovery of temporal (‘comet’) communities. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 271–283. Springer, 2014.
- [5] Brett W. Bader, Tamara G. Kolda, et al. Matlab tensor toolbox version 3.1. Available online, June 2019.
- [6] Rasmus Bro and Henk AL Kiers. A new efficient method for determining the number of components in parafac models. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 17(5):274–286, 2003.
- [7] J Douglas Carroll and Jih-Jie Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of “eckart-young” decomposition. *Psychometrika*, 35(3):283–319, 1970.
- [8] Alexander Gorovits, Ekta Gujral, Evangelos E Papalexakis, and Petko Bogdanov. Larc: Learning activity-regularized overlapping communities across time. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1465–1474, 2018.
- [9] Alexander Gorovits, Lin Zhang, Ekta Gujral, Evangelos Papalexakis, and Petko Bogdanov. Mining bursty groups from interaction data. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 596–605, 2021.
- [10] Richard A Harshman et al. Foundations of the PARAFAC procedure: models and conditions for an explanatory multimodal factor analysis. *UCLA Working Papers in Phonetics*, 16:1–84, 1970.
- [11] Joyce C Ho, Joydeep Ghosh, and Jimeng Sun. Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 115–124, 2014.
- [12] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.
- [13] Tamara G Kolda, Brett W Bader, and Joseph P Kenny. Higher-order web link analysis using multilinear algebra. In *Fifth IEEE International Conference on Data Mining (ICDM’05)*, pages 8–pp. IEEE, 2005.
- [14] Denis Krompaß, Maximilian Nickel, Xueyan Jiang, and Volker Tresp. Non-negative tensor factorization with rescal. In *Tensor Methods for Machine Learning, ECML workshop*, pages 1–10, 2013.
- [15] Taehyung Kwon, Inkyu Park, Dongjin Lee, and Kijung Shin. Slicenstitch: Continuous cp decomposition of sparse tensor streams. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 816–827. IEEE, 2021.
- [16] Alexander V Lotov and Kaisa Miettinen. Visualizing the pareto frontier. In *Multiobjective optimization*, pages 213–243. Springer, 2008.
- [17] Evangelos E Papalexakis. Automatic unsupervised tensor mining with quality assessment. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 711–719. SIAM, 2016.
- [18] Evangelos E Papalexakis, Christos Faloutsos, and Nicholas D Sidiropoulos. Tensors for data mining and data fusion: Models, applications, and scalable algorithms. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(2):1–44, 2016.
- [19] Ravdeep Pasricha, Ekta Gujral, and Evangelos Papalexakis. Adaptive granularity in time evolving graphs as tensors. In *Proceedings of the 16th International Workshop on Mining and Learning with Graphs (MLG)*, 2020.
- [20] Ravdeep Pasricha, Ekta Gujral, and Evangelos E Papalexakis. Adaptive granularity in tensors: A quest for interpretable structure. *arXiv preprint arXiv:1912.09009*, 2019.
- [21] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.
- [22] Nicholas D Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582, 2017.
- [23] Sucheta Soundarajan, Acar Tamersoy, Elias B Khalil, Tina Eliassi-Rad, Duen Horng Chau, Brian Gallagher, and Kevin Roundy. Generating graph snapshots from streaming edge data. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 109–110, 2016.
- [24] Rajmonda Sulo, Tanya Berger-Wolf, and Robert Grossman. Meaningful selection of temporal resolution for dynamic networks. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, pages 127–136, 2010.
- [25] Dingqi Yang, Daqing Zhang, Vincent W Zheng, and Zhiyong Yu. Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(1):129–142, 2014.

The Triangle-Densest- k -Subgraph Problem: Hardness, Lovász Extension, and Application to Document Summarization

Aritra Konar, and Nicholas D. Sidiropoulos

Department of Electrical and Computer Engineering, University of Virginia, VA, USA
(aritra,nikos)@virginia.edu

Abstract

We introduce the triangle-densest- k -subgraph problem (TD k S) for undirected graphs: given a size parameter k , compute a subset of k vertices that maximizes the number of induced triangles. The problem corresponds to the simplest generalization of the edge based densest- k -subgraph problem (D k S) to the case of higher-order network motifs. We prove that TD k S is NP-hard and is not amenable to efficient approximation, in the worst-case. By judiciously exploiting the structure of the problem, we propose a relaxation algorithm for the purpose of obtaining high-quality, sub-optimal solutions. Our approach utilizes the fact that the cost function of TD k S is submodular to construct a convex relaxation for the problem based on the Lovász extension for submodular functions. We demonstrate that our approaches attain state-of-the-art performance on real-world graphs and can offer substantially improved exploration of the optimal density-size curve compared to sophisticated approximation baselines for D k S. We use document summarization to showcase why TD k S is a useful generalization of D k S.

Introduction

The task of extracting dense subgraphs from a given graph has diverse applications in graph mining ranging from fraud detection (Hooi et al. 2016; Zhang et al. 2017), chemical informatics (Podolyan and Karypis 2009), computational biology (Saha et al. 2010) and knowledge discovery (Angel et al. 2012; Tixier, Malliaros, and Vazirgiannis 2016). Owing to its practical relevance, the problem has received extensive attention (see (Cadena, Chen, and Vullikanti 2018) and references therein) - we briefly highlight some prominent formulations.

Given an undirected graph, the classic densest-subgraph (DS) problem (Goldberg 1984) aims to detect the subgraph with the maximum average induced degree. The problem is known to be polynomial-time solvable, and admits a simple linear-time $1/2$ approximation via a greedy algorithm (Charikar 2000). However, real-world examples are known (Tsourakakis et al. 2013) where the greedy algorithm returns the trivial solution corresponding to the graph itself as the densest subgraph. This undesirable behavior can be attributed to the fact that the approach does not allow explicit

specification of the desired subgraph size. Adding a simple size constraint to the DS problem results in the densest- k -subgraph (D k S) problem (Feige, Peleg, and Kortsarz 2001), which, for a specified node-size k , aims to find the subgraph with the maximum number of induced edges. Unfortunately, the constraint also renders D k S NP-hard. Moreover, the problem is notorious for being very difficult to approximate, in the worst-case sense (Khot 2006; Bhaskara et al. 2012; Manurangsi 2017). Notwithstanding such pessimistic results, polynomial-time algorithms which work well in practice for D k S are known – these include low (constant) rank matrix approximation techniques (Papailiopoulos et al. 2014) and a recent work (Konar and Sidiropoulos 2021), which uses tools from submodular optimization to construct a convex relaxation for D k S.

A salient feature of the aforementioned formulations is that they quantify subgraph density in terms of induced edges, which represent pair-wise relationships between vertices. However, real-world graphs are often rich in higher-order motifs, which signify stronger associations among vertices as compared to pair-wise relationships alone (Watts and Strogatz 1998). This suggests that leveraging higher-order motif structure for dense subgraph discovery can detect subgraphs which are more clique-like compared to those obtained via the edge-based formulations. For example, prior work (Tsourakakis 2015) has introduced the ℓ -clique densest subgraph problem to extract the subgraph with the largest average number of induced ℓ -cliques. This is a generalization of the DS problem (the latter corresponds to choosing $\ell = 2$) which remains polynomial-time solvable and also admits effective approximation via a greedy algorithm (Tsourakakis 2015). More importantly, applying this formulation with $\ell = 3$ (triangles, the simplest example of a higher-order motif) to real-world graphs yields subgraphs of higher *edge density* compared to using the DS with $\ell = 2$. However, like its edge-based counterpart, the ℓ -clique densest subgraph problem formulation does not provide a means of explicitly controlling the desired subgraph size. We argue that this is a restrictive feature, since it does not allow the end-user the flexibility in picking a desired solution. By varying the size explicitly, one can obtain small subsets of vertices which are tightly knit, to larger subsets which exhibit smaller density, and everything in-between.

For this purpose, in this paper, we introduce the triangle-

densest- k -subgraph problem (TD k S). Given an undirected graph $\mathcal{G}$ on n vertices and a desired subgraph size k , we aim to compute the subgraph with the maximum number of induced triangles over all possible $\binom{n}{k}$ subgraphs. Clearly, TD k S is the simplest higher-order generalization of its edge based counterpart D k S. To the best of our knowledge, however, this is the first time that the problem has been studied. Can adopting such a formulation enable us to discover denser subgraphs compared to D k S, and thereby do a better job at exploring the optimal density-size curve on real-world graphs? Can TD k S extract more meaningful subsets than D k S in real-world applications? These are the main questions considered in our paper. Given this context, our contributions can be summarized as follows.

- **Hardness:** We prove that TD k S is NP-hard in the worst-case. Additionally, we show that it is difficult to obtain a favorable approximation of the optimal objective value of TD k S in polynomial-time.

- **Submodular relaxation and algorithm:** Notwithstanding such pessimistic worst-case results, we focus on developing an approximation algorithm which can work well on real-world instances. We show and leverage the fact that the discrete cost function of TD k S is endowed with a specific type of combinatorial structure - namely, it is a submodular function. As such functions possess a unique, continuous, convex extension (i.e., the Lovász extension), we devise a convex relaxation for TD k S that minimizes the Lovász extension over the convex hull of the cardinality constraints. Additionally, a key technical contribution of our paper is to show that for TD k S, the Lovász extension admits an analytical functional form, which is difficult to determine for general submodular functions. We exploit this structure to develop a scalable Mirror Descent algorithm for solving the problem, which, combined with a simple rounding procedure, can be employed for extracting candidate triangle-dense subgraphs.

- **Experiments:** Our experiments reveal that the proposed approach is very effective in mining triangle-dense subgraphs on real-world datasets. Interestingly, it can also extract subgraphs of higher edge density than state-of-the-art D k S baselines, which is a bonus. Our experiments further indicate that when TD k S is used for unsupervised document summarization it yields more meaningful and interpretable summaries than D k S does. A sneak preview of our results can be found in Table 1, which depicts a summary of 30 words extracted from a text document representing a collection of reviews for the movie Joker, which we obtained from the review aggregator website <https://metacritic.com>. In order to apply our method to this text dataset, we constructed a graph-based representation where unique words in the document are vertices, with triplets of vertices being connected via triangles if the words they correspond to co-occur together in a tri-gram (i.e., a sub-sentence of length 3). Note that the TD k S represents a document summary composed of a subset of k words which co-occur most frequently in triples. An alternative graph model of text can be constructed from bi-grams - here, an edge connects a pair of vertices if the words they represent appear together in a bi-gram. In this context, the D k S is the subset of k words which appear

Approach	Extracted Summary
TD k S	movie, joker, comic, book, character, reimagined, movies, one, anchored, performance, joaquin, phoenix, oscar, worthy, allows, would, study, depressing, engrossing, masterful, iconic, well, made, great, fine, almost, created, downside, killer, work.
D k S	film, movie, joker, comic, book, character, dc, much, work, performance, joaquin, phoenix, arthur, villain, todd, phillips, study, one, bad, social, last, well, point, also, go, enough, anything, specific, like, movies.

Table 1: Thirty-word summaries extracted from text reviews of the 2019 movie Joker by applying TD k S (resp. D k S) on a tri-gram (resp. bi-gram) graph-of-words model (Mihalcea and Tarau 2004).

most frequently in pairs. Since TD k S exploits higher-order co-occurrences (modeled using triangle motifs) relative to D k S, we intuitively expect that for a fixed summary size k , the former approach will yield more cohesive summaries compared to the latter. The results in Table 1 confirm this intuition.

We point out that, at a high level, our use of the Lovász relaxation for TD k S is in the spirit of (Konar and Sidiropoulos 2021) which introduced the Lovász relaxation for D k S. That being said, there are also important differences (apart from the fact that the two problems are distinct), the one key being that computing an analytical functional form for the Lovász extension of TD k S is substantially more challenging compared to the classical edge based case. Additionally, the form that the Lovász extension of TD k S takes is more complicated than that for D k S, which necessitates an entirely different algorithmic approach. Finally, to put our contributions into broader context, several recent works (Tsourakakis 2015; Tsourakakis, Pachocki, and Mitzenmacher 2017; Benson, Gleich, and Leskovec 2016; Zhang and Parthasarathy 2012) have considered generalizing classical edge-based graph mining tasks to account for higher-order network motifs. Our present work seeks to contribute to this thread of research by developing new tools for tackling a challenging problem in this area.

Primer on Submodularity

We provide a brief overview of basic concepts regarding submodular functions (Lovász 1983; Bach et al. 2013; Fujishige 2005). For a set of n objects $\mathcal{V} = \{1, \dots, n\}$, a *set function* $F : 2^{\mathcal{V}} \rightarrow \mathbb{R}$ assigns a real value to any subset $\mathcal{S} \subseteq \mathcal{V}$. A set function F is said to be *submodular* if and only if $F(\mathcal{A} \cup \mathcal{B}) + F(\mathcal{A} \cap \mathcal{B}) \leq F(\mathcal{A}) + F(\mathcal{B})$ for all subsets $\mathcal{A}, \mathcal{B} \subseteq \mathcal{V}$. For the special case where $n = 2$ and $\mathcal{V} = \{a, b\}$, the above condition simplifies to $F(\emptyset) + F(\mathcal{V}) \leq F(\{a\}) + F(\{b\})$. A notable feature of submodular functions is that they possess a continuous, *convex* extension known as the Lovász extension, which extends their domain from $2^{\mathcal{V}}$ to the unit interval $[0, 1]^n$ (recall

$n = |\mathcal{V}|$). Formally, the Lovász extension $f_L : [0, 1]^n \rightarrow \mathbb{R}$ of a submodular function F is given by

$$f_L(\mathbf{x}) := \max_{\mathbf{g} \in \mathcal{B}_F} \mathbf{g}^T \mathbf{x}, \quad (1)$$

where the set $\mathcal{B}_F$ is the *base polytope* associated with F and is defined as

$$\mathcal{B}_F := \{\mathbf{g} \in \mathbb{R}^n : \mathbf{g}^T \mathbf{1}_{\mathcal{V}} = F(\mathcal{V}); \mathbf{g}^T \mathbf{1}_S \leq F(S), \forall S \subseteq \mathcal{V}\}. \quad (2)$$

It can be seen that the Lovász extension is the support function of the base polytope $\mathcal{B}_F$, and is thus a convex function. In fact, f_L is convex if and only if F is submodular. Furthermore, when evaluated at a binary vector $\mathbf{x} \in \{0, 1\}^n$, the Lovász extension equals the value of the submodular function F .

Problem Statement

Given an undirected graph $\mathcal{G} := (\mathcal{V}, \mathcal{E})$ on n vertices and a parameter $4 \leq k < n$, we consider the problem of determining the subgraph of size k that exhibits the maximum weighted sum of induced triangles. Let $\mathcal{X} := \{\mathbf{x} \in \{0, 1\}^n : \mathbf{1}^T \mathbf{x} = k\}$ be the set of all binary vectors with k non-zero entries. Formally, the triangle-densest- k -subgraph (TDkS) problem can be expressed as

$$\max_{\mathbf{x} \in \mathcal{X}} \left\{ f(\mathbf{x}) := \sum_{(u,v,w) \in \Delta} w_t x_u x_v x_w \right\}, \quad (3)$$

where Δ denotes the set of triangles in the graph (each counted once), and w_t is a positive weight associated with triangle $t := \{u, v, w\} \in \Delta$ ¹. We denote the optimal value of (3) as f^* (which represents the optimal weighted sum of induced triangles) and the optimal triangle density as $\rho_3^*(\mathcal{G}, k) := f^* / \binom{k}{3}$. Evidently, the above problem corresponds to maximizing a discrete third-order polynomial, subject to a cardinality constraint. This suggests that it is no easier to solve compared to the discrete quadratic maximization form associated with its classic edge-based counterpart, which is known to be NP-hard (Feige, Peleg, and Kortsarz 2001), and also very difficult to approximate. We make these notions concrete by proving the following pair of negative results regarding TDkS.

Theorem 1. *TDkS is NP-hard.*

In light of the above result, it is unlikely that the problem admits an efficient solution in polynomial-time. Consequently, we focus on developing effective approximation algorithms for problem (3) that run in polynomial-time. However, we first show that TDkS is fundamentally not amenable to favorable approximation in the worst-case sense; in fact it is more difficult to approximate compared to DkS.

More precisely, given an instance of DkS, let $\rho_2^*(\mathcal{G}, k)$ denote the optimal edge-density, defined as the ratio of the number of induced edges in the optimal size- k subgraph and the maximum possible number of induced edges $\binom{k}{2}$. Regarding the hardness of approximation of DkS, the following result is known (Manurangsi 2017).

¹If $\mathcal{G}$ is unweighted, each triangle $t \in \Delta$ has weight $w_t = 1$.

Lemma 1. *Assuming that the Exponential Time Hypothesis (ETH) is valid, there is no polynomial-time algorithm that can approximate the optimal value of DkS better than a multiplicative factor $\alpha(n) := n^{1/(\log \log n)^c}$, where $c > 0$ is a universal constant.*

Note that the quantity $\alpha(n) > 1$. Hence, given an arbitrary instance of DkS, there is no polynomial-time algorithm which can output a size- k subgraph whose edge-density is guaranteed to be no worse than a fraction $1/(\alpha(n))^{1-\epsilon}$ of the optimal edge-density $\rho_2^*(\mathcal{G}, k)$, for any $\epsilon > 0$. In other words - if $\bar{\rho}_2(\mathcal{G}, k)$ denotes the edge density achieved by any polynomial-time approximation algorithm applied on a fixed instance of DkS, it must hold that

$$\rho_2^*(\mathcal{G}, k) \geq \bar{\rho}_2(\mathcal{G}, k) \geq O\left(\frac{1}{\alpha(n)}\right) \rho_2^*(\mathcal{G}, k). \quad (4)$$

We now demonstrate that the above hardness result for DkS can be utilized to derive an analogous hardness of approximation result for TDkS as well.

Theorem 2. *Assuming ETH is true, there is no polynomial-time algorithm that can approximate the optimal value of TDkS better than a multiplicative factor $\beta(n) := (\alpha(n))^{3/2}$.*

The above result implies that TDkS is more difficult to approximate compared to DkS, which is already known to be a challenging problem. Roughly speaking, Theorem 2 asserts that even the best possible polynomial-time approximation algorithm for TDkS must exhibit an approximation gap that grows as a polynomial in the size of the problem input n , which is a very pessimistic result.

That being said, the results of Theorem 1 and 2 are based on viewing the problem from the perspective of the worst-case scenario, which may not always arise in practice. With this in mind, we propose a convex relaxation for TDkS with the aim of obtaining high-quality, sub-optimal solutions on real-world instances.

The Lovász Relaxation

In order to explain our approach, we first reformulate TDkS in combinatorial form as follows. Let $\mathcal{C} := \{S \subseteq \mathcal{V} : |S| = k\}$ denote the collection of subsets of vertices of size k . Note that there is a one-to-one correspondence between the elements of $\mathcal{X}$ and $\mathcal{C}$; every vector $\mathbf{x} \in \mathcal{X}$ is precisely the indicator function of a subset of vertices $S \in \mathcal{C}$, i.e., given a vector $\mathbf{x} \in \mathcal{X}$ and a set $S \in \mathcal{C}$, we have the equivalence

$$x_u = \begin{cases} 1 & \Leftrightarrow u \in S \\ 0 & \Leftrightarrow u \notin S. \end{cases} \quad (5)$$

This observation allows us to equivalently express problem (3) in minimization form as

$$\min_{S \in \mathcal{C}} \left\{ F(S) := \sum_{(u,v,w) \in \Delta} F_{uvw}(S) \right\}, \quad (6)$$

where for each triangle $(u, v, w) \in \Delta$, we have defined the function

$$F_{uvw}(S) := F(S \cap \{u, v, w\}) = \begin{cases} -w_t, & \text{if } (u, v, w) \in S, \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

Hence, the cost function $F(\mathcal{S})$ linearly decomposes over the set of triangles of $\mathcal{G}$, with each component function $F_{uvw}(\mathcal{S})$ contributing to the overall cost if and only if all three vertices constituting a triangle are included in the subgraph induced by $\mathcal{S} \in \mathcal{C}$. Our starting point is the following observation regarding the cost function $F(\mathcal{S})$.

Theorem 3. $F(\mathcal{S})$ is a submodular function.

Note that Theorem 3 does not change the fact that problem (3) is difficult to solve in the worst-case. However, it does allow us to adopt the following relaxation strategy. Let $\mathcal{P} := \{\mathbf{x} \in [0, 1]^n; \mathbf{1}^T \mathbf{x} = k\}$ denote the convex hull of the combinatorial sum-to- k constraints. The key idea underpinning our approach is the following. Since the cost function of (6) is submodular, we can replace it by its Lovász extension to obtain the following equivalent problem

$$\begin{aligned} \min \quad & f_L(\mathbf{x}) \\ \text{s.to} \quad & \mathbf{x} \in \{0, 1\}^n \cap \mathcal{P}. \end{aligned} \quad (8)$$

Note that the equivalence stems from the fact that the Lovász extension equals the value of $F(\cdot)$ at all binary $\{0, 1\}^n$ vectors. Upon dropping the discrete constraints, we obtain the relaxed problem

$$\min_{\mathbf{x} \in \mathcal{P}} f_L(\mathbf{x}) \quad (9)$$

which corresponds to minimizing the Lovász extension of F over the convex hull of the combinatorial set $\mathcal{C}$. Our rationale for employing the Lovász extension as a convex surrogate of F stems from the fact that it corresponds to the convex closure of F on the domain $[0, 1]^n$. In other words, in a certain sense, the Lovász extension is the tightest convex under-estimator of F .

It is evident that problem (9) is convex, and hence can be optimally solved in polynomial-time to obtain a lower bound on the optimal value of (6). However, from an algorithmic standpoint, a major issue in solving the above problem is that the Lovász extension of a submodular function does not admit an analytical functional form in general. This can be attributed to the fact that the base polytope $\mathcal{B}_F$ is characterized by (potentially) an exponential number of inequalities in the problem dimension n . In a seminal paper, Edmonds (Edmonds 1970) established that a greedy algorithm based on sorting and querying F on n specific subsets suffices to compute a subgradient of the Lovász extension at any point $\mathbf{x} \in [0, 1]^n$ without requiring explicit specification of the base polytope $\mathcal{B}_F$. While this result can be utilized within a projected subgradient framework for solving (9), for our present problem, we elect not to do so. This is due to the fact that the greedy algorithm is generic, i.e., it is not tailored to exploit the form of the submodular cost function of (6), which, in addition to its incremental nature, can result in a heavy computational footprint on large graphs.

We now demonstrate that it is possible to circumvent the aforementioned challenges related to solving (9) efficiently, and the main reason is that the Lovász extension for TDkS *does* admit an analytical form. In order to formally establish the result, we exploit the fact that F is linearly decomposable over the triangle-set Δ , which in turn implies that its base polytope can be expressed as the Minkowski sum of the

base polytopes of the constituent functions F_{uvw} (Schrijver 2003, Theorem 44.6), i.e., we have

$$\mathcal{B}_F = \sum_{(u,v,w) \in \Delta} \mathcal{B}_{F_{uvw}}, \quad (10)$$

where $\mathcal{B}_{F_{uvw}}$ is the base polytope associated with the component F_{uvw} , and we have overloaded notation to represent set addition using the standard addition operator. Our next result shows that each such “sub”-polytope $\mathcal{B}_{F_{uvw}}$ admits a simple characterization.

Lemma 2. The base polytope of F_{uvw} is given by

$$\mathcal{B}_{F_{uvw}} = -w_t \text{conv}(\mathbf{e}_u, \mathbf{e}_v, \mathbf{e}_w). \quad (11)$$

Hence, the base polytope of F_{uvw} is the probability simplex in the space spanned by the coordinates indexed via (u, v, w) reflected about the origin and scaled by the weight w_t . Next, we exploit this result to derive an analytical form for the Lovász extension of F .

Theorem 4. The Lovász extension of F is given by

$$f_L(\mathbf{x}) = - \sum_{(u,v,w) \in \Delta} w_t \min\{x_u, x_v, x_w\}$$

The above result allows us to express problem (9) as

$$\min_{\substack{\mathbf{x} \in [0,1]^n, \\ \sum_{u=1}^n x_u = k}} \sum_{(u,v,w) \in \Delta} w_t \max\{-x_u, -x_v, -x_w\}, \quad (12)$$

which we designate as the Lovász relaxation. On inspecting the problem, however, it offers little in terms of an intuitive explanation as to why it can serve as a useful approximation for TDkS. To this end, our next result shows that the Lovász extension can be cast in an alternate form, which provides additional insight regarding (12).

Theorem 5. The Lovász extension of F can be expressed as

$$f_L(\mathbf{x}) = -\mathbf{t}^T \mathbf{x} + \sum_{(u,v,w) \in \Delta} w_t \cdot \phi_t(x_u, x_v, x_w) \quad (13)$$

where $\mathbf{t} \in \mathbb{R}^n$ denotes the vector of triangle degrees and $\phi_t(x_u, x_v, x_w) := \max\{x_u + x_v - 2x_w, x_v + x_w - 2x_u, x_u + x_w - 2x_v\}$.

Using the above derived form, we now provide an intuitive explanation for the Lovász relaxation. Given any subset of vertices $\mathcal{S} \subseteq \mathcal{V}$, define the triangle “volume” $\text{vol}_\Delta(\mathcal{S}) := \sum_{v \in \mathcal{S}} t_v$ of $\mathcal{S}$ to be the sum of the weighted triangle counts of the vertices that constitute $\mathcal{S}$. Using a double counting argument, the triangle volume of any subset $\mathcal{S} \subseteq \mathcal{V}$ can be equivalently expressed as

$$\text{vol}_\Delta(\mathcal{S}) = t_1(\mathcal{S}) + 2t_2(\mathcal{S}) + 3t_3(\mathcal{S}), \quad (14)$$

where $t_1(\mathcal{S})$, $t_2(\mathcal{S})$ and $t_3(\mathcal{S})$ denote the weighted sum of triangles with one, two and three endpoints in $\mathcal{S}$, respectively. The above identity can be re-written as

$$t_3(\mathcal{S}) = (\text{vol}_\Delta(\mathcal{S}) - [2t_2(\mathcal{S}) + t_1(\mathcal{S})])/3, \forall \mathcal{S} \subseteq \mathcal{V}. \quad (15)$$

Note that the term on the left hand side corresponds to the objective function of TDkS. Hence, among subgraphs of a

given size, those containing a large number of induced triangles must exhibit a large triangle volume (the first term on the right hand side) while simultaneously having few triangles being cut as a result of crossing the boundary of $\mathcal{S}$ (measured by the sum of the two terms subtracted from the volume). To be precise, for any given subset, a severed triangle with two endpoints $\{u, v\} \in \mathcal{S}$ affects the triangle counts (t_u, t_v) of both respective vertices (and hence the -2 factor), whereas a cut triangle with a single endpoint $u \in \mathcal{S}$ affects the triangle count t_u of only that vertex (and hence the -1 factor). The above equation asserts that for subgraphs with high triangle density, these losses stemming from severed triangles should be small compared to the triangle volume.

In order to establish the link with the form of the Lovász extension established in Theorem 5, we re-write (15) as

$$-t_3(\mathcal{S}) = -(\text{vol}_\Delta(\mathcal{S}) + [2t_2(\mathcal{S}) + t_1(\mathcal{S})])/3. \quad (16)$$

Since we have already established that $-t_3(\mathcal{S})$ is a submodular function and $\text{vol}_\Delta(\mathcal{S})$ is a modular (and thus submodular) function, the remainder on the right hand side must also be submodular, as submodularity is preserved under addition. Furthermore, as the Lovász extension of the sum of submodular functions equals the sum of the Lovász extensions of the component functions, inspecting the result of Theorem 5 reveals that it corresponds to the sum of the Lovász extensions of the terms on the right hand side of (16). Hence, the extension preserves the first term, corresponding to the triangle volume, whereas it uses a convex surrogate for the second term to approximate the losses in the volume stemming from severed triangles. In particular, when solving the Lovász relaxation, each vertex is assigned a soft score that indicates how likely it is to belong to the triangle-densest- k subgraph. The formulation then assigns the highest emphasis on those vertices which have large triangle counts, but also exhibit small variation in scores across triangles.

Algorithm: Mirror Descent

In this section, we describe our algorithm for efficiently solving the Lovász relaxation (12), which is a convex problem. Since the Lovász extension is non-differentiable, this suggests employing a Euclidean projected subgradient algorithm for solving (12). The algorithm starts from an initial feasible point $\mathbf{x}^0 \in \mathcal{P}$ and then proceeds in the following iterative fashion

$$\mathbf{x}^{r+1} = \arg \min_{\mathbf{x} \in \mathcal{P}} \left\{ (\mathbf{g}^r)^T \mathbf{x} + \frac{1}{\beta^r} \|\mathbf{x} - \mathbf{x}^r\|_2^2 \right\}, \forall r \in \mathbb{N} \quad (17)$$

where $\mathbf{g}^r \in \partial f_L(\mathbf{x}^r)$ denotes a subgradient of the Lovász extension $f_L(\mathbf{x})$ at the current iterate $\mathbf{x} = \mathbf{x}^r$ and $\beta^r > 0$ is the learning rate. A standard result in convex optimization states that if the subgradients of f_L are bounded in the Euclidean sense, i.e., there exists a constant $G > 0$ such that

$$\|\mathbf{g}\|_2 \leq G, \forall \mathbf{g} \in \partial f_L(\mathbf{x}), \forall \mathbf{x} \in \mathcal{P}, \quad (18)$$

then using the learning rate schedule $\beta_r = O(1/(\sqrt{r}))$ is sufficient to guarantee convergence to the optimal cost of (12) at a sublinear-rate of $O(G/\sqrt{r})$ (Bubeck 2015, Theorem 3.2). From this result, one can hope that the iteration

complexity of the Euclidean subgradient algorithm is independent of the problem dimension n , which is a desirable trait for scaling up to large problem instances. However, the above claim is true provided that the Lipschitz constant G of the Lovász extension is independent of n . Unfortunately, this is not the case for our problem, as it can be shown that G varies like $O(n)$. Consequently, the Euclidean subgradient method (17) applied to solve (12) attains a dimension-dependent convergence rate of $O(\sqrt{\frac{n}{r}})$, which has undesirable implications for large-scale instances.

Thus, the non-Euclidean geometry of the problem renders the standard subgradient method (which measures distances in the ℓ_2 -sense) a poor fit. In order to correct for this “mismatch” in geometry, we propose to employ the Mirror Descent algorithm (MDA) (Beck and Teboulle 2003), which can be viewed as a generalization of the subgradient algorithm to non-Euclidean spaces. To be specific, MDA is an iterative first-order algorithm that starts from a point $\mathbf{x}^0 \in \mathcal{P}$ and performs the following updates

$$\mathbf{x}^{r+1} = \arg \min_{\mathbf{x} \in \mathcal{P}} \left\{ (\mathbf{g}^r)^T \mathbf{x} + \frac{1}{\beta^r} D(\mathbf{x}, \mathbf{x}^r) \right\}, \forall r \in \mathbb{N} \quad (19)$$

where $D(\cdot, \cdot)$ is an appropriate “proximity”-measuring function. For example, on choosing $D(\mathbf{x}, \mathbf{x}^r) = \|\mathbf{x} - \mathbf{x}^r\|_2^2$, we obtain the standard subgradient algorithm. This proximal term can be viewed as the Bregman divergence associated with the function $\|\mathbf{x}\|_2^2$, which is strongly convex w.r.t. the ℓ_2 norm.

For our problem, the subgradients of the Lovász extension have constant size in the ℓ_∞ sense (see Section H of supplement), which motivates measuring distances using the ℓ_1 norm (which is the dual norm of the ℓ_∞ norm). This observation also suggests the choice of $D(\cdot, \cdot)$ in MDA to be the un-normalized Kullback-Leibler (KL) divergence between the points $\mathbf{x}$ and $\mathbf{x}^r$, which is defined as $D_{\text{KL}}(\mathbf{x}, \mathbf{x}^r) = \sum_{i=1}^n x_i \left(\log \frac{x_i}{x_i^r} - 1 \right) + x_i^r$. Such a choice is based on the fact that the KL divergence is the Bregman divergence associated with the negative entropy function, which is strongly convex w.r.t. ℓ_1 norm on the feasible set $\mathcal{P}$. On performing the MDA updates (19) using KL divergence with the learning rate schedule $\beta_r = O(\sqrt{\log n/r})$, invoking a standard result in convex optimization (Bubeck 2015, Theorem 4.2) guarantees a convergence rate of $O(G_\infty \sqrt{\log n/r})$, where G_∞ denotes the Lipschitz constant of f_L w.r.t. the ℓ_1 norm. Since this quantity is a constant, we obtain a convergence rate that exhibits a significantly improved dependence on the problem dimension n compared to that of the standard subgradient algorithm. Hence, fixing the geometry mismatch by employing the ℓ_1 norm to measure distances in MDA pays substantial dividends in this case. Owing to space constraints, the details of deriving the MDA updates are omitted. A full description of MDA is provided in Algorithm 1.

Since the computed solution $\mathbf{x}_L$ is not guaranteed to be integral in general, we perform a simple post-processing rounding step in order to obtain a binary indicator vector corresponding to a candidate subgraph. This is accomplished by simply projecting $\mathbf{x}_L$ onto the discrete sum-to- k constraints,

Algorithm 1: MIRROR DESCENT

Input: Triangle list Δ , triangle weights $\{w_t\}_{t \in \Delta}$, subgraph size k , bisection tolerance $\epsilon > 0$.

Initialize: $\mathbf{x}^1 = (k/n)\mathbf{1}$, $r = 1$.

1: **while** Convergence criterion is not met **do**

2: Obtain $\mathbf{g}^r \in \partial f_L(\mathbf{x}^r)$.

3: Update step-size $\beta^r = c/\sqrt{r}$.

4: $\mathbf{y}^r := \mathbf{x}^r \otimes \exp(-\beta^r \mathbf{g}^r)$.

5: $\mathbf{x}^{r+1} = \text{BISECTION}(\mathbf{y}^r, k, \epsilon)$.

6: Update $r = r + 1$.

7: **end while**

8: **return** $\mathbf{x}_L = (1/r) \sum_{i=1}^r \mathbf{x}^i$

Graph	n	m	$ \Delta $
PPI-HUMAN	21,557	342K	2.39M
FACEBOOK-B	63,731	817K	3.51M
CAIDA	192K	609K	455K
WEB-STANFORD	281K	2.31M	11.33M
WEB-GOOGLE	875K	5.10M	13.39M
WIKI-TOPCATS	1.8M	28.51M	52.11M

Table 2: Summary of graph statistics: the number of vertices (n), the number of edges (m), and the number of triangles ($|\Delta|$).

which is equivalent to identifying the support of the k -largest entries in $\mathbf{x}_L$, and can be performed in $O(n \log k)$ time using heaps.

Experiments

In this section, we test the effectiveness of our proposed method in exploring the triangle density-size trade-off across a collection of real-world graphs. Additionally, we consider an application of TDkS to document summarization. Our results indicate that contrary to the worst-case scenario, real-world instances of TDkS can be far from adversarial, with the Lovász relaxation being effective at identifying high-quality, sub-optimal solutions.

Baselines

To the best of our knowledge, we are unaware of any pre-existing algorithms for the TDkS problem. Hence, we employ two state-of-the-art baselines for the (edge) densest- k -subgraph DkS problem, and test their efficacy at discovering triangle-dense subgraphs. These methods are described in brief below - additional details can be found in Section K of the supplement.

Lovász Relaxation for DkS (Konar and Sidiropoulos 2021): Similar to the approach considered herein (at a high level), but applied to the edge-density based formulation. Utilizes a variant of the Alternating Direction Method of Multipliers (ADMM) (Condat 2013) to solve the relaxed problem.

Low-rank Binary Matrix Principal Component (Papailiopoulos et al. 2014): Employs a rank-1 decomposition of the graph adjacency matrix $\mathbf{A}$, followed by solving the DkS

problem with the rank-1 approximation in place of $\mathbf{A}$. The resulting problem admits a simple solution in $O(n)$ time which also provides an instance-specific upper bound on the optimal edge density for a given subgraph size. While not attainable in general, this bound can serve as a useful performance benchmark.

Triangle-density upper bound: The edge-density upper bound for DkS obtained via the above approach can also be converted into an upper bound on the optimal triangle density for TDkS via the Kruskal-Katona theorem (Kruskal 1963; Katona 1972). However, such a bound is not attainable in general for every choice of k as it is more loose compared to the bound on the optimal edge-density. In spite of this, we observed that on real-world graphs the Lovász relaxation for TDkS can attain this upper bound, or capture a significant fraction of it.

Since the first two baselines do not aim to directly detect triangle dense subgraphs, for fair comparison, we also compare the efficacy of our proposed methods for TDkS at detecting edge-dense subgraphs against the above baselines.

Datasets

We used a collection of graph datasets (summarized in Table 2) from standard repositories (Leskovec and Krevl 2014; Kunegis 2013) to test the performance of all methods. For TDkS, we used the well-known NODEITERATOR++ algorithm (Suri and Vassilvitskii 2011, Algorithm 2) to obtain a list of triangles in the graph, which incurs a run-time complexity of $O(|\mathcal{E}|^{3/2})$.

Results and Discussion

The outcomes of our experiments on the considered datasets are depicted in Figure 1. Our main findings are:

- With regard to subgraph triangle density (left column in Figure 1), solving the Lovász relaxation for TDkS via Mirror Descent followed by rounding consistently yields the best results across all considered graphs. In fact, for small subgraph sizes (≤ 100), it is the only method that attains, or comes close to attaining the upper bound on the optimal triangle density. Our results demonstrate that although TDkS is NP-hard and difficult to approximate in the worst-case, the Lovász relaxation can still prove to be an effective tool for detecting triangle-dense subgraphs in real-world graphs.
- Although the Mirror Descent algorithm aims to detect subgraphs with high triangle density, it turns in a commendable performance in terms of edge density as well (right column in Figure 1). In fact, for subgraph sizes ≤ 200 , it outperforms the dedicated edge based formulations, often by a significant margin and comes closest to attaining the edge density upper bound. This can be viewed as a consequence of the Kruskal-Katona theorem which formalizes the following intuitive notion: if a subgraph has high triangle density, then it must possess high edge density as well. Looking at Figure 1 (right column) confirms this observation.
- For large subgraph sizes, the edge density obtained by Mirror Descent / TDkS is often second (although by a small margin) to that obtained by applying the Lovász relaxation for DkS. Empirically, we note that this occurs (i.e., the blue

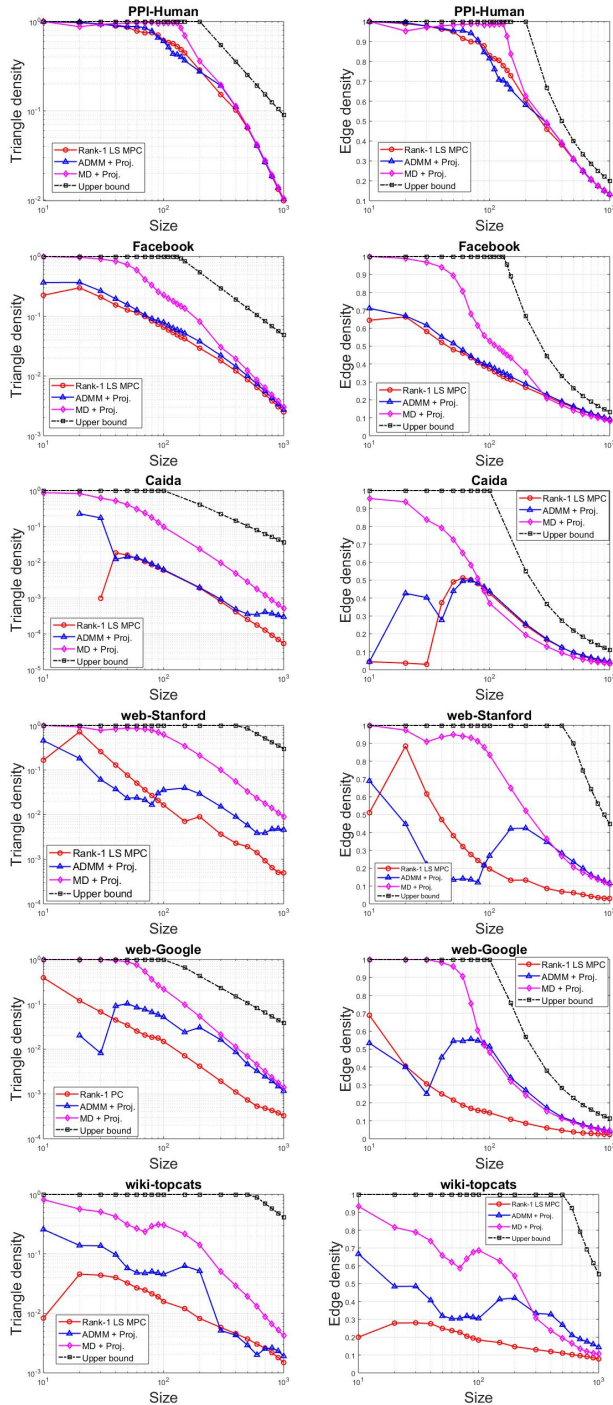


Figure 1: Left column: Triangle density (on a log-scale) vs size. Right column: Edge density vs size. Red (Rank-1 least-squares matrix principal component) and blue (ADMM) curves are methods for DkS while the magenta curve (MD) is for $TDkS$. The black curve in the right column depicts the edge density upper bound for DkS . This bound combined with the Kruskal-Katona theorem also yields an upper bound on $TDkS$ (black curve in the left column).

curve “overtakes” the magenta curve) when the edge density falls below the 50% threshold. A possible explanation is as follows. Turán’s theorem (Turán 1941) states that a graph can exhibit an edge density at most 0.5 without harboring any triangles. In other words, below this threshold, there do exist graphs with edge density up to 50% while containing very few triangles. Consequently, in the regime where the densest subgraph of a given size has edge density upper bounded by 0.5, employing a density measure based on edges may prove to be more beneficial as opposed to using triangles, if one cares more about edge density.

Conclusions

We considered the triangle-densest- k -subgraph problem ($TDkS$) which aims to compute the size k subgraph with the largest number of induced triangles. Unfortunately, not only is the problem NP-hard, but it is also difficult to approximate in polynomial-time, in the worst-case sense. With the aim of computing high-quality, sub-optimal solutions on real-world instances, we exploited the fact that the cost function of $TDkS$ is submodular to construct a convex relaxation of the problem based on the Lovász extension of submodular functions. As we derived an analytical functional form for the extension, this enabled us to devise a Mirror Descent algorithm for efficiently solving the problem at scale. Our results on real-world graphs showcased that our approach can effectively exploit triangle motifs to attain state-of-the-art performance, and can provide a more effective means of exploring the density-size trade-off compared to baselines that only use edges for density maximization. Additionally, we utilized the problem of document summarization to showcase that $TDkS$ can generate more informative word summaries compared to DkS .

Acknowledgements

Supported by the National Science Foundation and the Army Research Office under Grants No. IIS-1908070 and ARO-W911NF1910407 respectively. Aritra Konar acknowledges the assistance of Soumyajit Gupta in pre-processing the movie-review text datasets.

References

- Angel, A.; Sarkas, N.; Koudas, N.; and Srivastava, D. 2012. Dense subgraph maintenance under streaming edge weight updates for real-time story identification. In *Proceedings of the 38th International Conference on Very Large Data Bases*, 574–585. VLDB Endowment.
- Bach, F.; et al. 2013. Learning with Submodular Functions: A Convex Optimization Perspective. *Foundations and Trends® in Machine Learning*, 6(2-3): 145–373.
- Beck, A.; and Teboulle, M. 2003. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3): 167–175.
- Benson, A. R.; Gleich, D. F.; and Leskovec, J. 2016. Higher-order organization of complex networks. *Science*, 353(6295): 163–166.

- Bhaskara, A.; Charikar, M.; Guruswami, V.; Vijayaraghavan, A.; and Zhou, Y. 2012. Polynomial integrality gaps for strong sdg relaxations of densest k-subgraph. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, 388–405. SIAM.
- Bubeck, S. 2015. Convex Optimization: Algorithms and Complexity. *Foundations and Trends® in Machine Learning*, 8(3-4): 231–357.
- Cadena, J.; Chen, F.; and Vullikanti, A. 2018. Graph anomaly detection based on Steiner connectivity and density. *Proceedings of the IEEE*, 106(5): 829–845.
- Charikar, M. 2000. Greedy approximation algorithms for finding dense components in a graph. In *International Workshop on Approximation Algorithms for Combinatorial Optimization*, 84–95. Springer.
- Condat, L. 2013. A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 158(2): 460–479.
- Edmonds, J. 1970. Submodular functions, matroids, and certain polyhedra. *Edited by G. Goos, J. Hartmanis, and J. van Leeuwen*, 11.
- Feige, U.; Peleg, D.; and Kortsarz, G. 2001. The dense k-subgraph problem. *Algorithmica*, 29(3): 410–421.
- Fujishige, S. 2005. *Submodular functions and optimization*. Elsevier.
- Goldberg, A. V. 1984. *Finding a maximum density subgraph*. Technical report, University of California Berkeley, CA.
- Hooi, B.; Song, H. A.; Beutel, A.; Shah, N.; Shin, K.; and Faloutsos, C. 2016. Fraudar: Bounding graph fraud in the face of camouflage. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 895–904. ACM.
- Katona, G. O. 1972. A simple proof of the Erdős-Chao Ko-Rado theorem. *Journal of Combinatorial Theory, Series B*, 13(2): 183–184.
- Khot, S. 2006. Ruling out PTAS for graph min-bisection, dense k-subgraph, and bipartite clique. *SIAM Journal on Computing*, 36(4): 1025–1071.
- Konar, A.; and Sidiropoulos, N. D. 2021. Exploring the Subgraph Density-Size Trade-off via the Lovász Extension. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 743–751.
- Kruskal, J. B. 1963. The number of simplices in a complex. *Mathematical optimization techniques*, 10: 251–278.
- Kunegis, J. 2013. KONECT – The Koblenz Network Collection. In *Proc. Int. Conf. on World Wide Web Companion*, 1343–1350.
- Leskovec, J.; and Krevl, A. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. <https://snap.stanford.edu/data>. Accessed: 2021-08-15.
- Lovász, L. 1983. Submodular functions and convexity. In *Mathematical programming the state of the art*, 235–257. Springer.
- Manurangsi, P. 2017. Almost-polynomial ratio ETH-hardness of approximating densest k-subgraph. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, 954–961.
- Mihalcea, R.; and Tarau, P. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, 404–411.
- Papailiopoulos, D.; Mitliagkas, I.; Dimakis, A.; and Caramanis, C. 2014. Finding dense subgraphs via low-rank bilinear optimization. In *International Conference on Machine Learning*, 1890–1898.
- Podolyan, Y.; and Karypis, G. 2009. Common pharmacophore identification using frequent clique detection algorithm. *Journal of chemical information and modeling*, 49(1): 13–21.
- Saha, B.; Hoch, A.; Khuller, S.; Raschid, L.; and Zhang, X.-N. 2010. Dense subgraphs with restrictions and applications to gene annotation graphs. In *Annual International Conference on Research in Computational Molecular Biology*, 456–472. Springer.
- Schrijver, A. 2003. *Combinatorial optimization: polyhedra and efficiency*, volume 24. Springer Science & Business Media.
- Suri, S.; and Vassilvitskii, S. 2011. Counting triangles and the curse of the last reducer. In *Proceedings of the 20th international conference on World wide web*, 607–614.
- Tixier, A.; Malliaros, F.; and Vazirgiannis, M. 2016. A graph degeneracy-based approach to keyword extraction. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, 1860–1870.
- Tsourakakis, C. 2015. The k-clique densest subgraph problem. In *Proceedings of the 24th International Conference on World Wide Web*, 1122–1132. International World Wide Web Conferences Steering Committee.
- Tsourakakis, C.; Bonchi, F.; Gionis, A.; Gullo, F.; and Tsiarli, M. 2013. Denser than the densest subgraph: extracting optimal quasi-cliques with quality guarantees. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 104–112.
- Tsourakakis, C. E.; Pachocki, J.; and Mitzenmacher, M. 2017. Scalable motif-aware graph clustering. In *Proceedings of the 26th International Conference on World Wide Web*, 1451–1460.
- Turán, P. 1941. On an external problem in graph theory. *Mat. Fiz. Lapok*, 48: 436–452.
- Watts, D. J.; and Strogatz, S. H. 1998. Collective dynamics of ‘small-world’ networks. *Nature*, 393(6684): 440.
- Zhang, S.; Zhou, D.; Yildirim, M. Y.; Alcorn, S.; He, J.; Davulcu, H.; and Tong, H. 2017. Hidden: hierarchical dense subgraph detection with application to financial fraud detection. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, 570–578. SIAM.
- Zhang, Y.; and Parthasarathy, S. 2012. Extracting analyzing and visualizing triangle k-core motifs within networks. In *2012 IEEE 28th international conference on data engineering*, 1049–1060. IEEE.

Mining Triangle-Dense Subgraphs of a Fixed Size: Hardness, Lovász extension and Applications

Aritra Konar, and Nicholas D. Sidiropoulos, *Fellow, IEEE*

Abstract—We introduce the triangle-densest- k -subgraph problem (TD k S) for undirected graphs: given a size parameter k , compute a subset of k vertices that maximizes the number of induced triangles. The problem corresponds to the simplest generalization of the edge based densest- k -subgraph problem (D k S) to the case of higher-order network motifs. We prove that TD k S is NP-hard and is not amenable to efficient approximation, in the worst-case. By judiciously exploiting the structure of the problem, we propose a relaxation algorithm for the purpose of obtaining high-quality, sub-optimal solutions. Our approach utilizes the fact that the cost function of TD k S is submodular to construct a convex relaxation for the problem based on the Lovász extension for submodular functions. We demonstrate that our approaches attain state-of-the-art performance on real-world graphs and can offer substantially improved exploration of the optimal density-size curve compared to sophisticated approximation baselines for D k S. We use document summarization to showcase why TD k S is a useful generalization of D k S.

Index Terms—Dense subgraph discovery, triangle-motifs, intractability and approximation, submodularity, convex optimization.



1 INTRODUCTION

The task of extracting dense subgraphs from a given graph has diverse applications in graph mining ranging from fraud detection [1], [2], chemical informatics [3], computational biology [4] and knowledge discovery [5], [6]. Owing to its practical relevance, the problem has received extensive attention (see [7] and references therein) - we briefly highlight some prominent formulations.

Given an undirected graph, the classic densest-subgraph (DS) problem [8] aims to detect the subgraph with the maximum average induced degree. The problem is known to be polynomial-time solvable, and admits a simple linear-time $1/2$ approximation via a greedy algorithm [9]. However, real-world examples are known [10] where the greedy algorithm returns the trivial solution corresponding to the graph itself as the densest subgraph. This undesirable behavior can be attributed to the fact that the approach does not allow explicit specification of the desired subgraph size. Adding a simple size constraint to the DS problem results in the densest- k -subgraph (D k S) problem [11], which, for a specified node-size k , aims to find the subgraph with the maximum number of induced edges. Unfortunately, the constraint also renders D k S NP-hard. Moreover, the problem is notorious for being very difficult to approximate, in the worst-case sense [12]–[14]. Notwithstanding such pessimistic results, polynomial-time algorithms which work well in practice for D k S are known – these include low (constant) rank matrix approximation techniques [15] and a recent work [16], which uses tools from submodular optimization to construct a convex relaxation for D k S.

A salient feature of the aforementioned formulations is

that they quantify subgraph density in terms of induced edges, which represent pair-wise relationships between vertices. However, real-world graphs are often rich in higher-order motifs, which signify stronger associations among vertices as compared to pair-wise relationships alone [17]. This suggests that leveraging higher-order motif structure for dense subgraph discovery can detect subgraphs which are more clique-like compared to those obtained via the edge-based formulations. For example, prior work [18] has introduced the ℓ -clique densest subgraph problem to extract the subgraph with the largest average number of induced ℓ -cliques. This is a generalization of the DS problem (the latter corresponds to choosing $\ell = 2$) which remains polynomial-time solvable and also admits effective approximation via a greedy algorithm [18]. More importantly, applying this formulation with $\ell = 3$ (triangles, the simplest example of a higher-order motif) to real-world graphs yields subgraphs of higher *edge density* compared to using the DS with $\ell = 2$. However, like its edge-based counterpart, the ℓ -clique densest subgraph problem formulation does not provide a means of explicitly controlling the desired subgraph size. We argue that this is a restrictive feature, since it does not allow the end-user the flexibility in picking a desired solution. By varying the size explicitly, one can obtain small subsets of vertices which are tightly knit, to larger subsets which exhibit smaller density, and everything in-between.

For this purpose, in this paper, we introduce the triangle-densest- k -subgraph problem (TD k S). Given an undirected graph $\mathcal{G}$ on n vertices and a desired subgraph size k , we aim to compute the subgraph with the maximum number of induced triangles over all possible $\binom{n}{k}$ subgraphs. Clearly, TD k S is the simplest higher-order generalization of its edge based counterpart D k S. To the best of our knowledge, however, this is the first time that the problem has been studied. Can adopting such a formulation enable us to

- A. Konar is with the Dept. of Electrical Engineering (ESAT), KU Leuven, Belgium. E-mail: aritra.konar@kuleuven.be
- Nicholas D. Sidiropoulos is with the Dept. of Electrical and Computer Engineering, University of Virginia, USA. E-mail: nikos@virginia.edu

discover denser subgraphs compared to DkS , and thereby do a better job at exploring the optimal density-size curve on real-world graphs? Can $TDkS$ extract more meaningful subsets than DkS in real-world applications? These are the main questions considered in our paper. Given this context, our contributions can be summarized as follows.

- **Hardness:** We prove that $TDkS$ is NP-hard in the worst-case. Additionally, we show that it is difficult to obtain a favorable approximation of the optimal objective value of $TDkS$ in polynomial-time.

- **Submodular relaxation and algorithm:** Notwithstanding such pessimistic worst-case results, we focus on developing an approximation algorithm which can work well on real-world instances. We show and leverage the fact that the discrete cost function of $TDkS$ is endowed with a specific type of combinatorial structure - namely, it is a submodular function. As such functions possess a unique, continuous, convex extension (i.e., the Lovász extension), we devise a convex relaxation for $TDkS$ that minimizes the Lovász extension over the convex hull of the cardinality constraints. Additionally, a key technical contribution of our paper is to show that for $TDkS$, the Lovász extension admits an analytical functional form, which is difficult to determine for general submodular functions. We exploit this structure to develop a scalable Mirror Descent algorithm for solving the problem, which, combined with a simple rounding procedure, can be employed for extracting candidate triangle-dense subgraphs.

- **Experiments:** Our experiments reveal that the proposed approach is very effective in mining triangle-dense subgraphs on real-world datasets. Interestingly, it can also extract subgraphs of higher edge density than state-of-the-art DkS baselines, which is a bonus. Our experiments further indicate that when $TDkS$ is used for unsupervised document summarization it yields more meaningful and interpretable summaries than DkS does.

We point out that, at a high level, our use of the Lovász relaxation for $TDkS$ is in the spirit of [16] which introduced the Lovász relaxation for DkS . That being said, there are also important differences (apart from the fact that the two problems are distinct), the one key being that computing an analytical functional form for the Lovász extension of $TDkS$ is substantially more challenging compared to the classical edge based case. Additionally, the form that the Lovász extension of $TDkS$ takes is more complicated than that for DkS , which necessitates an entirely different algorithmic approach. Finally, to put our contributions into broader context, several recent works [18]–[21] have considered generalizing classical edge-based graph mining tasks to account for higher-order network motifs. Our present work seeks to contribute to this thread of research by developing new tools for tackling a challenging problem in this area.

A conference version of this work has previously appeared in [22]. Relative to that, the present journal version adds a new section on problem motivation to highlight the conceptual appeal of our formulation, additional experiments including a comprehensive and insightful case study on document summarization, and fully fleshed out technical proofs.

2 PRIMER ON SUBMODULARITY

We provide a brief overview of basic concepts regarding submodular functions [23]–[25]. For a set of n objects $\mathcal{V} = \{1, \dots, n\}$, a *set function* $F : 2^{\mathcal{V}} \rightarrow \mathbb{R}$ assigns a real value to any subset $S \subseteq \mathcal{V}$. A set function F is said to be *submodular* if and only if $F(\mathcal{A} \cup \mathcal{B}) + F(\mathcal{A} \cap \mathcal{B}) \leq F(\mathcal{A}) + F(\mathcal{B})$ for all subsets $\mathcal{A}, \mathcal{B} \subseteq \mathcal{V}$. For the special case where $n = 2$ and $\mathcal{V} = \{a, b\}$, the above condition simplifies to $F(\emptyset) + F(\mathcal{V}) \leq F(\{a\}) + F(\{b\})$. A notable feature of submodular functions is that they possess a continuous, *convex* extension known as the Lovász extension, which extends their domain from $2^{\mathcal{V}}$ to the unit interval $[0, 1]^n$ (recall $n = |\mathcal{V}|$). Formally, the Lovász extension $f_L : [0, 1]^n \rightarrow \mathbb{R}$ of a submodular function F is given by

$$f_L(\mathbf{x}) := \max_{\mathbf{g} \in \mathcal{B}_F} \mathbf{g}^T \mathbf{x}, \quad (1)$$

where the set $\mathcal{B}_F$ is the *base polytope* associated with F and is defined as

$$\mathcal{B}_F := \{\mathbf{g} \in \mathbb{R}^n : \mathbf{g}^T \mathbf{1}_{\mathcal{V}} = F(\mathcal{V}); \mathbf{g}^T \mathbf{1}_S \leq F(S), \forall S \subseteq \mathcal{V}\}. \quad (2)$$

It can be seen that the Lovász extension is the support function of the base polytope $\mathcal{B}_F$, and is thus a convex function. In fact, f_L is convex if and only if F is submodular. Furthermore, when evaluated at a binary vector $\mathbf{x} \in \{0, 1\}^n$, the Lovász extension equals the value of the submodular function F .

3 PROBLEM STATEMENT

Consider an undirected graph $\mathcal{G} := (\mathcal{V}, \mathcal{E})$ on n vertices. Given a size parameter $k \in \{4, \dots, n\}$ and a subset of k vertices $S \subseteq \mathcal{V}$, let $\rho_2(S, k)$ denote the edge density of the subgraph $\mathcal{G}_S$ induced by S . This quantity equals the ratio of the sum of the induced edges and the maximum possible number of induced edges $\binom{k}{2}$. In an analogous fashion, let $\rho_3(S, k)$ denote the triangle density of S . In this paper, we consider the problem of extracting the subgraph of size k that exhibits the maximum sum of induced triangles. Let $\mathcal{X}_k := \{\mathbf{x} \in \{0, 1\}^n : \mathbf{1}^T \mathbf{x} = k\}$ be the set of all binary vectors with k non-zero entries. Formally, the triangle-densest- k -subgraph ($TDkS$) problem can be expressed as

$$\max_{\mathbf{x} \in \mathcal{X}_k} \left\{ f(\mathbf{x}) := \sum_{(u,v,w) \in \Delta} w_t x_u x_v x_w \right\}, \quad (3)$$

where Δ denotes the set of triangles in the graph (each counted once), and w_t is a positive weight associated with triangle $t := \{u, v, w\} \in \Delta$.

Let $\mathbf{x}^*$ denote an optimal solution of (3) and $f(\mathbf{x}^*)$ represent the optimal sum of induced triangles. If we denote the optimal vertex subset as $\mathcal{S}^* = \{i \in \mathcal{V} : x_i^* = 1\}$, then the maximum triangle density is $\rho_3(\mathcal{S}^*, k) := f(\mathbf{x}^*) / \binom{k}{3}$. Note that by varying the size parameter k , $TDkS$ outputs a spectrum of dense subgraphs. We designate this spectrum as the triangle density versus size curve - each point on this curve corresponds to a pair $(k, \rho_3(\mathcal{S}^*, k))$.

1. If $\mathcal{G}$ is unweighted, each triangle $t \in \Delta$ has weight $w_t = 1$.

3.1 Motivation

It is evident that $TDkS$ is a higher-order extension of DkS , i.e., it quantifies subgraph density with respect to triangles, as opposed to edges. Since the triangle motif is a higher-order clique compared to an edge, it is then natural to consider what additional benefits the $TDkS$ formulation (3) can offer relative to DkS .

To this end, consider a subgraph $\mathcal{S} \subseteq \mathcal{V}$ on k vertices, with edge and triangle densities $\rho_2(\mathcal{S}, k)$ and $\rho_3(\mathcal{S}, k)$ respectively. It is known that these two quantities obey the following relationship.

Fact 1. The Kruskal-Katona Theorem [26], [27]: For any subgraph, it always holds that

$$\rho_2(\mathcal{S}, k) \geq \rho_3(\mathcal{S}, k)^{\frac{2}{3}}. \quad (4)$$

Hence, maximizing triangle density for a fixed subgraph size serves as a surrogate for maximizing edge density as well. The above fact formalizes the simple notion that maximizing the number of induced triangles in a subgraph also has the effect of increasing the number of induced edges.

Additionally, for a subgraph $\mathcal{S} \subseteq \mathcal{V}$, let $\tau(\mathcal{S})$ denotes its transitivity (i.e., global clustering coefficient). Recall that it can be expressed as

$$\tau(\mathcal{S}) = 3 \cdot \frac{\text{No. of triangles induced by } \mathcal{S}}{\text{No. of paths of length 2 in } \mathcal{S}}. \quad (5)$$

The value of subgraph transitivity always lies in the unit interval, and can be interpreted as the probability of sampling (uniformly at random) a path of length 2 from $\mathcal{G}_{\mathcal{S}}$, and having its endpoints connected to form a triangle. Simply put, it determines how globally “triangle-like” the subgraph is. Hence, transitivity can serve as an alternate means of quantifying density with regard to triangle motifs. However, it turns out that transitivity obeys the following relationship with triangle density.

Lemma 1. Given a subgraph $\mathcal{S}$ on $k := |\mathcal{S}|$ vertices, its transitivity $\tau(\mathcal{S})$ is no lesser than its triangle density $\rho_3(\mathcal{S}, k)$, with equality if and only if $\mathcal{S}$ is a k -clique.

Proof. The denominator of $\tau(\mathcal{S})$ can be expressed as $\sum_{u \in \mathcal{S}} \binom{d_u}{2}$, where d_u is the induced degree of vertex $u \in \mathcal{S}$. It then holds that

$$\frac{1}{3} \sum_{u \in \mathcal{S}} \binom{d_u}{2} \leq \frac{k}{3} \binom{d_{\max}}{2} \leq \frac{k}{3} \binom{k-1}{2} = \binom{k}{3}, \quad (6)$$

where $d_{\max} \leq k-1$ is the largest induced degree in $\mathcal{S}$. Consequently, we have

$$\tau(\mathcal{S}) \geq \frac{\text{No. of triangles induced by } \mathcal{S}}{\binom{k}{3}} = \rho_3(\mathcal{S}, k). \quad (7)$$

The above inequality is satisfied with equality if and only if all inequalities in (6) hold with equality, which is true if and only if the subgraph induced by $\mathcal{S}$ is a k -clique. $\square$

The result is intuitive, since increasing the number of induced triangles in a subgraph cannot decrease its transitivity. Hence, for a fixed subgraph size, maximizing triangle density additionally serves as a surrogate for maximizing transitivity.

Table 2: Statistics of the triangle-densest-subgraph computed using the method of [28] on representative datasets.

Graph	Size	Edge density	Triangle density
PPI-HUMAN	361	0.42	0.14
FACEBOOK-B	198	0.36	0.08
CAIDA	75	0.55	0.20
WEB-STANFORD	684	0.17	0.02
WEB-GOOGLE	66	0.85	0.64

Thus far, we have demonstrated that employing triangle motifs for maximizing subgraph density via $TDkS$ offers the twin advantages of additionally improving the edge density and transitivity. At this point, it is instructive to compare the following two problems - $TDkS$ and its closest counterpart in the literature, the triangle-densest-subgraph (TDS) formulation [18], [28]. The TDS problem can be expressed as

$$\max_{\mathbf{z} \in \{0,1\}^n} \left\{ g(\mathbf{z}) := \frac{f(\mathbf{z})}{\mathbf{1}^T \mathbf{z}} \right\}. \quad (8)$$

The objective function of TDS assigns a greater reward to those subgraphs for which the average (induced) triangle degree of the vertices comprising the subgraph is large. In contrast to $TDkS$, the problem does not explicitly constrain the size of the desired subgraph. Interestingly, it turns out that we can establish a link between the solution of the two problems. Let $\mathbf{z}^*$ denote an optimal solution of (8) and $\mathcal{Z}^* = \{i \in \mathcal{V} : z_i^* = 1\}$ denote the optimal vertex subset. Then, we have the following claim.

Lemma 2. For the choice of the size parameter $k = |\mathcal{Z}^*|$, $\mathcal{Z}^*$ is also an optimal solution for $TDkS$.

Proof. We proceed via contradiction. Suppose that $\mathcal{Z}^*$ is not optimal for $TDkS$ with $k = |\mathcal{Z}^*|$. Then, there exists a subgraph $\bar{\mathcal{Z}} \neq \mathcal{Z}^*$ on k vertices such that the weighted sum of triangles it induces $f(\bar{\mathbf{z}}) > f(\mathbf{z}^*)$. Since both subgraphs are of size k , this further implies that $g(\bar{\mathbf{z}}) > g(\mathbf{z}^*)$, which contradicts the optimality of $\mathcal{Z}^*$ for TDS. $\square$

Thus, the solution of TDS corresponds to a point on the triangle-density versus size curve. The caveat of TDS, however, is that one cannot control the subgraph size explicitly. Hence, given a graph, one cannot predict *a priori* how large or small the size of the solution will be; i.e., pinpoint the exact location on the triangle-density versus size curve where the solution of TDS will land.

In order to empirically assess the quality of the solution of TDS, we tested the method of [28]² on several publicly available graph datasets³. The results are summarized in Table 1, from which it can be seen that in general, the solution computed by TDS is large in size with low density (see Table 1). These observations motivate the need to consider the additional flexibility afforded by $TDkS$ in explicitly specifying the size parameter k , as TDS may not yield a satisfactory solution.

² The code is available at <https://github.com/tsourolampis/Scalable-Large-Near-Clique-Detection>.

³ For details about the datasets, see Section V.

3.2 Hardness

Problem (3) corresponds to maximizing a discrete third-order polynomial subject to a cardinality constraint. This suggests that it is no easier to solve compared to the discrete quadratic maximization form associated with its classic edge-based counterpart, which is known to be NP-hard [11], and also very difficult to approximate. We make these notions concrete by proving the following pair of negative results regarding TDkS.

Theorem 1. *TDkS is NP-hard.*

Proof. See Appendix A in the supplement. $\square$

In light of the above result, it is unlikely that the problem admits an efficient solution in polynomial-time. Consequently, we focus on developing effective approximation algorithms for problem (3) that run in polynomial-time. However, we first show that TDkS is fundamentally not amenable to favorable approximation in the worst-case sense; in fact it is more difficult to approximate compared to DkS.

More precisely, given an instance of DkS, let $\mathcal{M}^*$ denote the optimal solution and $\rho_2(\mathcal{M}^*, k)$ denote the optimal edge-density. Regarding the hardness of approximation of DkS, the following result is known [14].

Fact 2. *Assuming that the Exponential Time Hypothesis (ETH) is valid, there is no polynomial-time algorithm that can approximate the optimal value of DkS better than a multiplicative factor $\alpha(n) := n^{1/(\log \log n)^c}$, where $c > 0$ is a universal constant.*

Note that the quantity $\alpha(n) > 1$. Hence, given an arbitrary instance of DkS, there is no polynomial-time algorithm which can output a size- k subgraph whose edge-density is guaranteed to be no worse than a fraction $1/(\alpha(n))^{1-\epsilon}$ of the optimal edge-density $\rho_2(\mathcal{M}^*, k)$, for any $\epsilon > 0$. In other words, if $\mathcal{M}$ denotes the output of *any* polynomial-time approximation algorithm applied on a fixed instance of DkS and $\rho_2(\mathcal{M}, k)$ is the achieved edge density, it must hold that

$$\rho_2(\mathcal{M}^*, k) \geq \rho_2(\mathcal{M}, k) \geq O\left(\frac{1}{\alpha(n)}\right) \rho_2(\mathcal{M}^*, k). \quad (9)$$

We now demonstrate that the above hardness result for DkS can be utilized to derive an analogous hardness of approximation result for TDkS as well.

Theorem 2. *Assuming ETH is true, there is no polynomial-time algorithm that can approximate the optimal value of TDkS better than a multiplicative factor $\beta(n) := (\alpha(n))^{3/2}$.*

Proof. See Appendix B in the supplement. $\square$

The above result implies that TDkS is more difficult to approximate compared to DkS, which is already known to be a challenging problem. Roughly speaking, Theorem 2 asserts that even the best possible polynomial-time approximation algorithm for TDkS must exhibit an approximation gap that grows as a polynomial in the size of the problem input n , which is a very pessimistic result.

That being said, the results of Theorem 1 and 2 are based on viewing the problem from the perspective of the worst-case scenario, which may not always arise in practice. With this in mind, we propose a convex relaxation for TDkS with

the aim of obtaining high-quality, sub-optimal solutions on real-world instances.

4 THE LOVÁSZ RELAXATION

In order to explain our approach, we first reformulate TDkS in combinatorial form as follows. Let $\mathcal{C} := \{S \subset \mathcal{V} : |S| = k\}$ denote the collection of subsets of vertices of size k . Note that there is a one-to-one correspondence between the elements of $\mathcal{X}_k$ and $\mathcal{C}$; every vector $\mathbf{x} \in \mathcal{X}_k$ is precisely the indicator function of a subset of vertices $S \in \mathcal{C}$, i.e., given a vector $\mathbf{x} \in \mathcal{X}_k$ and a set $S \in \mathcal{C}$, we have the equivalence

$$x_u = \begin{cases} 1 & \Leftrightarrow u \in S \\ 0 & \Leftrightarrow u \notin S. \end{cases} \quad (10)$$

This observation allows us to equivalently express problem (3) in minimization form as

$$\min_{S \in \mathcal{C}} \left\{ F(S) := \sum_{(u,v,w) \in \Delta} F_{uvw}(S) \right\}, \quad (11)$$

where for each triangle $(u, v, w) \in \Delta$, we have defined the function

$$F_{uvw}(S) := F(S \cap \{u, v, w\}) = \begin{cases} -w_t, & \text{if } (u, v, w) \in S, \\ 0, & \text{otherwise.} \end{cases} \quad (12)$$

Hence, the cost function $F(S)$ linearly decomposes over the set of triangles of $\mathcal{G}$, with each component function $F_{uvw}(S)$ contributing to the overall cost if and only if all three vertices constituting a triangle are included in the subgraph induced by $S \in \mathcal{C}$. Our starting point is the following observation regarding the cost function $F(S)$.

Theorem 3. *$F(S)$ is a submodular function.*

Proof. See Appendix C in the supplement. $\square$

Note that Theorem 3 does not change the fact that problem (3) is difficult to solve in the worst-case. However, it does allow us to adopt the following relaxation strategy. Let $\mathcal{P}_k := \{\mathbf{x} \in [0, 1]^n; \mathbf{1}^T \mathbf{x} = k\}$ denote the convex hull of the combinatorial sum-to- k constraints. The key idea underpinning our approach is the following. Since the cost function of (11) is submodular, we can replace it by its Lovász extension to obtain the following equivalent problem

$$\begin{aligned} \min \quad & f_L(\mathbf{x}) \\ \text{s.to} \quad & \mathbf{x} \in \{0, 1\}^n \cap \mathcal{P}_k. \end{aligned} \quad (13)$$

Note that the equivalence stems from the fact that the Lovász extension equals the value of $F(\cdot)$ at all binary $\{0, 1\}^n$ vectors. Upon dropping the discrete constraints, we obtain the relaxed problem

$$\min_{\mathbf{x} \in \mathcal{P}_k} f_L(\mathbf{x}) \quad (14)$$

which corresponds to minimizing the Lovász extension of F over the convex hull of the combinatorial set $\mathcal{C}$. Our rationale for employing the Lovász extension as a convex surrogate of F stems from the fact that it corresponds to the convex closure of F on the domain $[0, 1]^n$. In other words, in a certain sense, the Lovász extension is the tightest convex under-estimator of F .

It is evident that problem (14) is convex, and hence can be optimally solved in polynomial-time to obtain a lower bound on the optimal value of (11). However, from an algorithmic standpoint, a major issue in solving the above problem is that the Lovász extension of a submodular function does not admit an analytical functional form in general. This can be attributed to the fact that the base polytope $\mathcal{B}_F$ is characterized by (potentially) an exponential number of inequalities in the problem dimension n . In a seminal paper, Edmonds [29] established that a greedy algorithm based on sorting and querying F on n specific subsets suffices to compute a subgradient of the Lovász extension at any point $\mathbf{x} \in [0, 1]^n$ without requiring explicit specification of the base polytope $\mathcal{B}_F$. While this result can be utilized within a projected subgradient framework for solving (14), for our present problem, we elect not to do so. This is due to the fact that the greedy algorithm is generic, i.e., it is not tailored to exploit the form of the submodular cost function of (11), which, in addition to its incremental nature, can result in a heavy computational footprint on large graphs.

We now demonstrate that it is possible to circumvent the aforementioned challenges related to solving (14) efficiently, and the main reason is that the Lovász extension for TDkS *does* admit an analytical form. In order to formally establish the result, we exploit the fact that F is linearly decomposable over the triangle-set Δ , which in turn implies that its base polytope can be expressed as the Minkowski sum of the base polytopes of the constituent functions F_{uvw} [30, Theorem 44.6], i.e., we have

$$\mathcal{B}_F = \sum_{(u,v,w) \in \Delta} \mathcal{B}_{F_{uvw}}, \quad (15)$$

where $\mathcal{B}_{F_{uvw}}$ is the base polytope associated with the component F_{uvw} , and we have overloaded notation to represent set addition using the standard addition operator. Our next result shows that each such “sub”-polytope $\mathcal{B}_{F_{uvw}}$ admits a simple characterization.

Lemma 3. *The base polytope of F_{uvw} is given by*

$$\mathcal{B}_{F_{uvw}} = -w_t \text{conv}(\mathbf{e}_u, \mathbf{e}_v, \mathbf{e}_w). \quad (16)$$

Proof. See Appendix E in the supplement. $\square$

Hence, the base polytope of F_{uvw} is the probability simplex in the space spanned by the coordinates indexed via (u, v, w) reflected about the origin and scaled by the weight w_t . Next, we exploit this result to derive an analytical form for the Lovász extension of F .

Theorem 4. *The Lovász extension of F is given by*

$$f_L(\mathbf{x}) = - \sum_{(u,v,w) \in \Delta} w_t \min\{x_u, x_v, x_w\}$$

Proof. See Appendix F in the supplement. $\square$

The above result allows us to explicitly express problem (14) as

$$\min_{\substack{\mathbf{x} \in [0,1]^n \\ \sum_{u=1}^n x_u = k}} \sum_{(u,v,w) \in \Delta} w_t \max\{-x_u, -x_v, -x_w\}, \quad (17)$$

which we designate as the Lovász relaxation. On inspecting the problem, however, it offers little in terms of an intuitive

explanation as to why it can serve as a useful approximation for TDkS. To this end, our next result shows that the Lovász extension can be cast in an alternate form, which provides additional insight regarding (17).

Theorem 5. *The Lovász extension of F can be expressed as*

$$f_L(\mathbf{x}) = -\mathbf{t}^T \mathbf{x} + \sum_{(u,v,w) \in \Delta} w_t \cdot \phi_t(x_u, x_v, x_w) \quad (18)$$

where $\phi_t(x_u, x_v, x_w) := \max\{x_u + x_v - 2x_w, x_v + x_w - 2x_u, x_u + x_w - 2x_v\}$.

Proof. See Appendix G in the supplement. $\square$

Using the above derived form, we now provide an intuitive explanation for the Lovász relaxation. Given any subset of vertices $\mathcal{S} \subseteq \mathcal{V}$, define the triangle “volume” $\text{vol}_\Delta(\mathcal{S}) := \sum_{v \in \mathcal{S}} t_v$ of $\mathcal{S}$ to be the sum of the weighted triangle counts of the vertices that constitute $\mathcal{S}$. Using a double counting argument, the triangle volume of any subset $\mathcal{S} \subseteq \mathcal{V}$ can be equivalently expressed as

$$\text{vol}_\Delta(\mathcal{S}) = t_1(\mathcal{S}) + 2t_2(\mathcal{S}) + 3t_3(\mathcal{S}), \quad (19)$$

where $t_1(\mathcal{S})$, $t_2(\mathcal{S})$ and $t_3(\mathcal{S})$ denote the weighted sum of triangles with one, two and three endpoints in $\mathcal{S}$, respectively. The above identity can be re-written as

$$t_3(\mathcal{S}) = (\text{vol}_\Delta(\mathcal{S}) - [2t_2(\mathcal{S}) + t_1(\mathcal{S})])/3, \forall \mathcal{S} \subseteq \mathcal{V}. \quad (20)$$

Note that the term on the left hand side corresponds to the objective function of TDkS. Hence, among subgraphs of a given size, those containing a large number of induced triangles must exhibit a large triangle volume (the first term on the right hand side) while simultaneously having few triangles being cut as a result of crossing the boundary of $\mathcal{S}$ (measured by the sum of the two terms subtracted from the volume). To be precise, for any given subset, a severed triangle with two endpoints $\{u, v\} \in \mathcal{S}$ affects the triangle counts (t_u, t_v) of both respective vertices (and hence the -2 factor), whereas a cut triangle with a single endpoint $u \in \mathcal{S}$ affects the triangle count t_u of only that vertex (and hence the -1 factor). The above equation asserts that for subgraphs with high triangle density, these losses stemming from severed triangles should be small compared to the triangle volume.

In order to establish the link with the form of the Lovász extension established in Theorem 5, we re-write (20) as

$$-t_3(\mathcal{S}) = -(\text{vol}_\Delta(\mathcal{S}) + [2t_2(\mathcal{S}) + t_1(\mathcal{S})])/3. \quad (21)$$

Since we have already established that $-t_3(\mathcal{S})$ is a submodular function and $\text{vol}_\Delta(\mathcal{S})$ is a modular (and thus submodular) function, the remainder on the right hand side must also be submodular, as submodularity is preserved under addition. Furthermore, as the Lovász extension of the sum of submodular functions equals the sum of the Lovász extensions of the component functions, inspecting the result of Theorem 5 reveals that it corresponds to the sum of the Lovász extensions of the terms on the right hand side of (21). Hence, the extension preserves the first term, corresponding to the triangle volume, whereas it uses a convex surrogate for the second term to approximate the losses in the volume stemming from severed triangles. In particular, when solving the Lovász relaxation, each vertex

is assigned a soft score that indicates how likely it is to belong to the triangle-densest- k subgraph. The formulation then assigns the highest emphasis on those vertices which have large triangle counts, but also exhibit small variation in scores across triangles.

5 ALGORITHM: MIRROR DESCENT

In this section, we describe our algorithm for efficiently solving the Lovász relaxation (17), which is a convex problem. Since the Lovász extension is non-differentiable, this suggests employing a Euclidean projected subgradient algorithm for solving (17). The algorithm starts from an initial feasible point $\mathbf{x}^0 \in \mathcal{P}_k$ and then proceeds in the following iterative fashion

$$\mathbf{x}^{r+1} = \arg \min_{\mathbf{x} \in \mathcal{P}_k} \left\{ (\mathbf{g}^r)^T \mathbf{x} + \frac{1}{\beta^r} \|\mathbf{x} - \mathbf{x}^r\|_2^2 \right\}, \forall r \in \mathbb{N} \quad (22)$$

where $\mathbf{g}^r \in \partial f_L(\mathbf{x}^r)$ denotes a subgradient of the Lovász extension $f_L(\mathbf{x})$ at the current iterate $\mathbf{x} = \mathbf{x}^r$ and $\beta^r > 0$ is the learning rate. A standard result in convex optimization states that if the subgradients of f_L are bounded in the Euclidean sense; i.e., there exists a constant $G > 0$ such that

$$\|\mathbf{g}\|_2 \leq G, \forall \mathbf{g} \in \partial f_L(\mathbf{x}), \forall \mathbf{x} \in \mathcal{P}_k, \quad (23)$$

then using the learning rate schedule $\beta_r = O(1/(\sqrt{r}))$ is sufficient to guarantee convergence to the optimal cost of (17) at a sublinear-rate of $O(G/\sqrt{r})$ [31, Theorem 3.2]. From this result, one can hope that the iteration complexity of the Euclidean subgradient algorithm is independent of the problem dimension n , which is a desirable trait for scaling up to large problem instances. However, the above claim is true provided that the Lipschitz constant G of the Lovász extension is independent of n . Unfortunately, this is not the case for our problem, as we now demonstrate.

Since f_L is linearly separable over the set of triangles in the graph, a standard result in convex analysis [32] asserts that a subgradient of the Lovász extension $\mathbf{g} \in \partial f_L(\mathbf{x})$ at a feasible point $\mathbf{x} \in \mathcal{P}_k$ can be expressed as

$$\mathbf{g} = \sum_{(u,v,w) \in \Delta} \mathbf{g}_{uvw}, \quad (24)$$

where $\mathbf{g}_{uvw} \in \partial f_{uvw}(\mathbf{x})$ denotes a subgradient of f_{uvw} at the point $\mathbf{x} \in \mathcal{P}_k$, and we have overloaded notation to represent set addition using the standard addition operator. The subdifferential set of each component function f_{uvw} is characterized by the equation

$$\partial f_{uvw}(\mathbf{x}) \in \arg \max_{\mathbf{y} \in \mathcal{B}_{F_{uvw}}} \mathbf{y}^T \mathbf{x}. \quad (25)$$

From the form of the base polytope $\mathcal{B}_{F_{uvw}}$ given by Lemma 3, it follows that any subgradient $\mathbf{g}_{uvw} \in \partial f_{uvw}(\mathbf{x})$ obeys

$$\|\mathbf{g}_{uvw}\|_\infty \leq w_t, \forall (u, v, w) \in \Delta. \quad (26)$$

Combining this observation with equation (24), we obtain that a subgradient of the Lovász extension $\mathbf{g} \in \partial f_L(\mathbf{x})$ is bounded in the ℓ_∞ sense as

$$\|\mathbf{g}\|_\infty \leq \sum_{(u,v,w) \in \Delta} \|\mathbf{g}_{uvw}\|_\infty \leq \sum_{(u,v,w) \in \Delta} w_t. \quad (27)$$

Since the sum of all triangle weights is a constant and $\|\mathbf{g}\|_2 \leq \sqrt{n} \|\mathbf{g}\|_\infty$, this implies that the Lipschitz constant G of f_L as defined in (23) in terms of ℓ_2 -distances is $O(\sqrt{n})$, which is dependent on the dimension. Consequently, the Euclidean subgradient method (22) applied to solve (17) attains a dimension-dependent convergence rate of $O(\sqrt{\frac{n}{r}})$, which has undesirable implications for large-scale instances.

Thus, the non-Euclidean geometry of the problem renders the standard subgradient method (which measures distances in the ℓ_2 -sense) a poor fit. In order to correct for this “mismatch” in geometry, we propose to employ the Mirror Descent algorithm (MDA) [33], which can be viewed as a generalization of the subgradient algorithm to non-Euclidean spaces. To be specific, MDA is an iterative first-order algorithm that starts from a point $\mathbf{x}^0 \in \mathcal{P}_k$ and performs the following updates

$$\mathbf{x}^{r+1} = \arg \min_{\mathbf{x} \in \mathcal{P}_k} \left\{ (\mathbf{g}^r)^T \mathbf{x} + \frac{1}{\beta^r} D(\mathbf{x}, \mathbf{x}^r) \right\}, \forall r \in \mathbb{N} \quad (28)$$

where $D(\cdot, \cdot)$ is an appropriate “proximity”-measuring function. For example, on choosing $D(\mathbf{x}, \mathbf{x}^r) = \|\mathbf{x} - \mathbf{x}^r\|_2^2$, we obtain the standard subgradient algorithm. This proximal term can be viewed as the Bregman divergence associated with the function $\|\mathbf{x}\|_2^2$, which is strongly convex w.r.t. the ℓ_2 norm.

Since the subgradients of the Lovász extension have constant size when measured using the ℓ_∞ norm, this motivates measuring distances using the ℓ_1 norm (which is the dual norm of the ℓ_∞ norm). This observation also suggests the choice of $D(\cdot, \cdot)$ in MDA to be the un-normalized Kullback-Leibler (KL) divergence between the points $\mathbf{x}$ and $\mathbf{x}^r$, which is defined as $D_{\text{KL}}(\mathbf{x}, \mathbf{x}^r) = \sum_{i=1}^n x_i \left(\log \frac{x_i}{x_i^r} - 1 \right) + x_i^r$. Such a choice is based on the fact that the KL divergence is the Bregman divergence associated with the negative entropy function, which is strongly convex w.r.t. ℓ_1 norm on the feasible set $\mathcal{P}_k$. On performing the MDA updates (28) using KL divergence with the learning rate schedule $\beta_r = O(\sqrt{\log n/r})$, invoking a standard result in convex optimization [31, Theorem 4.2] guarantees a convergence rate of $O(G_\infty \sqrt{\log n/r})$, where G_∞ denotes the Lipschitz constant of f_L w.r.t. the ℓ_1 norm. From (27), since this quantity is a constant, we obtain a convergence rate that exhibits a significantly improved dependence on the problem dimension n compared to that of the standard subgradient algorithm. Hence, fixing the geometry mismatch by employing the ℓ_1 norm to measure distances in MDA pays substantial dividends in this case.

With the above choice of KL divergence, the MDA updates (28) can be equivalently expressed as

$$\mathbf{x}^{r+1} = \arg \min_{\mathbf{x} \in \mathcal{P}_k} \left\{ D_{\text{KL}}(\mathbf{x}, \mathbf{x}^r \otimes \exp(-\beta^r \mathbf{g}^r)) \right\}, \quad (29)$$

where the $\otimes$ operator denotes element-wise multiplication. The update reveals that the algorithm utilizes a subgradient of the Lovász extension f_L to perform a multiplicative update on the present iterate $\mathbf{x}^r$ followed by computing the KL projection of the result onto $\mathcal{P}_k$ in order to ensure iterate feasibility. Hence, the potentially intensive tasks that have to be performed at each step are: (a) computing a subgradient of f_L , and (b) computing the KL projection onto the feasible

set $\mathcal{P}_k$. As it turns out, these operations can be computed efficiently - with task (a) requiring $O(|\Delta|)$ time and task (b) being solvable via bisection search.

(A) Computing a subgradient: In order to compute a subgradient of f_L at a given point $\mathbf{x} \in \mathcal{P}_k$, it suffices to compute a subgradient of each component function f_{uvw} , and then sum the results up. From equation (25), a subgradient $\mathbf{g}_{uvw} \in \partial f_{uvw}(\mathbf{x})$ is given by

$$\mathbf{g}_{uvw} = -w_t \mathbf{e}_{s^*}, s^* \in \arg \min_{s \in \{u,v,w\}} \{x_s\}. \quad (30)$$

Then, the full subgradient $\mathbf{g} \in \partial f_L(\mathbf{x})$ is $\mathbf{g} = \sum_{(u,v,w) \in \Delta} \mathbf{g}_{uvw}$. We implement this procedure as follows: we visit each component function f_{uvw} and extract the index of the sub-vector $[x_u, x_v, x_w]^T$ that attains the minimum (with ties broken arbitrarily). This operation incurs $O(|\Delta|)$ time, and can be trivially parallelized. For a vertex $u \in \mathcal{V}$, let Δ_u denote the sub-collection of component functions f_{uvw} (and corresponding triangles) in which u participates in, and, additionally, is the index that attains the minimum in (30). Since each subgradient $\mathbf{g}_{uvw}$ corresponds to updating a single coordinate of the full subgradient $\mathbf{g}$, its entries are given by $g_u = -\sum_{t \in \Delta_u} w_t, \forall u \in \mathcal{V}$. In the case where the graph is unweighted, i.e., $w_t = 1, \forall t \in \Delta$, the update can be further simplified to $g_u = -|\Delta_u|, \forall u \in \mathcal{V}$. Implementing this update requires counting the number of times x_u attains the minimum across all the component functions that it participates in (which correspond to the triangles that vertex $u \in \mathcal{V}$ belongs to).

(B) Computing KL projections: Let $\mathbf{y}^r := \mathbf{x}^r \otimes \exp(-\beta^r \mathbf{g}^r)$ denote the vector obtained by performing the multiplicative subgradient update. The MDA update (29) can then be expressed as

$$\mathbf{x}^{r+1} = \arg \min_{\mathbf{x} \in \mathcal{P}_k} \left\{ D_{\text{KL}}(\mathbf{x}, \mathbf{y}^r) \right\}. \quad (31)$$

Our next result provides an explicit characterization of the optimal solution of the above problem.

Lemma 4. *The solution $\mathbf{x}^{r+1}$ satisfies the conditions*

$$x_i^{r+1} = \min \left\{ 1, \exp(-\nu^*) y_i^r \right\}, \forall i \in [n], \sum_{i=1}^n x_i^{r+1} = k,$$

where $\nu^* \in \mathbb{R}$ is the optimal dual variable associated with the sum-to- k constraint.

Proof. See Appendix H in the supplement. $\square$

The above result can be exploited to solve (31) via a simple procedure. Define the positive variable $\alpha^* := \exp(-\nu^*)$. Note that in order to solve for $\mathbf{x}^{r+1}$, it suffices to solve for α^* . This in turn, can be accomplished by finding the root of the non-linear, continuous equation

$$\phi(\alpha) := \left[\sum_{i=1}^n \min \{1, \alpha y_i^r\} \right] - k. \quad (32)$$

To this end, note that $\phi(\alpha)$ is a continuous function that is monotone non-decreasing in α , which suggests a simple bisection search procedure. The lower and upper limits of the initial bisection interval can be set to be $\alpha_l := 1/\max_{i \in [n]} \{y_i^r\}$ and $\alpha_u := 1/(n \max_{i \in [n]} \{y_i^r\})$ respectively, for which value of

Algorithm 1 MIRROR DESCENT

Input: Triangle list Δ , triangle weights $\{w_t\}_{t \in \Delta}$, subgraph size k , bisection tolerance $\epsilon > 0$.
Initialize: $\mathbf{x}^1 = (k/n)\mathbf{1}, r = 1$.
1: **while** Convergence criterion is not met **do**
2: Obtain $\mathbf{g}^r \in \partial f_L(\mathbf{x}^r)$.
3: Update step-size $\beta^r = c/\sqrt{r}$.
4: $\mathbf{y}^r := \mathbf{x}^r \otimes \exp(-\beta^r \mathbf{g}^r)$.
5: $\mathbf{x}^{r+1} = \text{BISECTION}(\mathbf{y}^r, k, \epsilon)$.
6: Update $r = r + 1$.
7: **end while**
8: **return** $\mathbf{x}_L = (1/r) \sum_{i=1}^r \mathbf{x}^i$

TABLE 1: Summary of graph statistics: the number of vertices (n), the number of edges (m), the number of triangles ($|\Delta|$), and network type.

Graph	n	m	$ \Delta $	Network Type
PPI-HUMAN	21,557	342K	2.39M	Biological
FACEBOOK-B	63,731	817K	3.51M	Social
CAIDA	192K	609K	455K	Router
WEB-STANFORD	281K	2.31M	11.33M	Web graph
WEB-GOOGLE	875K	5.10M	13.39M	Web graph
WIKI-TOPCATS	1.8M	28.51M	52.11M	Hyperlinks

the lower interval is $\phi(\alpha_l) \leq 1 - k < 0$ and that of the upper interval is $\phi(\alpha_u) = n - k > 0$. For a prescribed exit tolerance $\epsilon > 0$, the algorithm requires $O(\log[\phi(\alpha_l) - \phi(\alpha_u)/\epsilon])$ iterations to exit.

A full description of MDA is provided in Algorithm 1. Since the computed solution $\mathbf{x}_L$ is not guaranteed to be integral in general, we perform a simple post-processing rounding step in order to obtain a binary indicator vector corresponding to a candidate subgraph. This is accomplished by simply projecting $\mathbf{x}_L$ onto the discrete sum-to- k constraints, which is equivalent to identifying the support of the k -largest entries in $\mathbf{x}_L$, and can be performed in $O(n \log k)$ time using heaps.

6 EXPERIMENTS

In this section, we test the effectiveness of our proposed method in exploring the triangle density-size trade-off across a collection of real-world graphs. Our results indicate that contrary to the worst-case scenario, real-world instances of TD k S can be far from adversarial, with the Lovász relaxation being effective at identifying high-quality, sub-optimal solutions.

6.1 Baselines

To the best of our knowledge, we are unaware of any pre-existing algorithms for the TD k S problem. Hence, we employ two state-of-the-art baselines for the (edge) densest- k -subgraph D k S problem, and test their efficacy at discovering triangle-dense subgraphs. These methods are described in brief below.

Lovász Relaxation for D k S [16]: The same approach considered herein, but applied to the edge-density based formulation, i.e., minimizing the Lovász extension for induced edges over the convex hull of the sum-to- k constraints. Utilizes a variant of the Alternating Direction Method of Multipliers (ADMM) [34] to solve the relaxed problem. As the solution is not guaranteed to be integral, a rounding

post-processing step is used to obtain the candidate subgraph.

Low-rank Binary Matrix Principal Component [15]: Employs a low-rank decomposition of the graph adjacency matrix A , followed by solving the DkS problem with the low-rank approximation in place of A . For the rank-1 approximation scenario, the resulting problem admits a simple solution in $O(n)$ time, whereas for constant ranks (i.e., $r = O(1)$), the problem can be surprisingly solved in polynomial-time $O(n^{r+1})$. Furthermore, the resulting solution can be utilized to construct an instance-specific upper bound on the optimal edge density for a given subgraph size, which, while not attainable in general, can serve as a useful performance benchmark. In practice, the algorithm is run using ranks $r \leq 5$, owing to its high complexity. In our experiments, we ran the algorithm with rank-1 approximation for all our datasets to generate a candidate subgraph and the edge-density upper bound, as even the rank-2 case proved too expensive to compute.

Triangle density upper bound: We demonstrate that the edge density upper bound for DkS obtained via the above approach can also be converted into an upper bound on the optimal triangle density for $TDkS$ via the Kruskal-Katona theorem, which asserts that for any unweighted, undirected (sub)graph $\mathcal{G}$, its edge density $\rho_2(\mathcal{G})$ and triangle density $\rho_3(\mathcal{G})$ must obey the relationship $\rho_3(\mathcal{G}) \leq (\rho_2(\mathcal{G}))^{3/2}$. Maximizing both sides of the above inequality w.r.t. all subgraphs of a fixed size k then yields the following relationship between the optimal triangle density of $TDkS$ $\rho_3^*(\mathcal{G}, k)$ and the optimal edge density of DkS $\rho_2^*(\mathcal{G}, k)$: we must have $\rho_3^*(\mathcal{G}, k) \leq (\rho_2^*(\mathcal{G}, k))^{3/2}$. Hence, an upper bound on $\rho_2^*(\mathcal{G}, k)$ translates into an upper bound on $\rho_3^*(\mathcal{G}, k)$ as well. However, such a bound is not attainable in general for every choice of k as it is more loose compared to the bound on $\rho_2^*(\mathcal{G}, k)$. In spite of this, we observed that on real-world graphs the Lovász relaxation for $TDkS$ can attain this upper bound, or capture a significant fraction of it.

Since the first two baselines do not aim to directly detect triangle dense subgraphs, for fair comparison, we also compare the efficacy of our proposed methods for $TDkS$ at detecting edge-dense subgraphs against the above baselines.

6.2 Datasets, pre-processing and implementation

We used a collection of graph datasets (summarized in Table 1) from standard repositories [35], [36] to test the performance of all methods. Each dataset is unweighted, and pre-processed by symmetrizing any directed arcs, removing self-loops, and extracting the largest connected component.

For $TDkS$, we used the well-known NODEITERATOR++ algorithm [37, Algorithm 2] to obtain a list of triangles in the graph, which incurs a run-time complexity of $O(m^{3/2})$. A beneficial byproduct of this step is that we can eliminate all vertices which do not participate in forming triangles from belonging to a triangle-dense graph. Since this step does not change the number of triangles in the graph, it does not affect the input to $TDkS$. This, in turn, reduces the problem dimension n and results in a substantial improvement in the practical run-time of the Mirror Descent Algorithm. Note that this step, however, does change the edge set of the graph, and hence, the input to DkS . Consequently, the edge-

based baselines outlined in the previous section are applied on the full graph.

All our experiments were performed in Matlab on a Windows workstation equipped with 16GB RAM and an Intel i7 processor. The Matlab code for the low-rank principal component approximation approach [15] was obtained via personal communication with the respective authors. Regarding our Mirror Descent algorithm for solving the Lovász relaxation for $TDkS$, we employed a diminishing step-size schedule $(c/t_{\max})/\sqrt{\log n/r}$, where $t_{\max}$ denotes the largest triangle “degree” in the graph, and $c > 0$ is a constant that was empirically chosen for each dataset. The maximum number of iterations applied was no more than 500 across all datasets.

6.3 Results and Discussion

The outcomes of our experiments on the considered datasets are depicted in Figure 1. Our main findings are:

- With regard to subgraph triangle density (left column in Figure 1), solving the Lovász relaxation for $TDkS$ via Mirror Descent followed by rounding consistently yields the best results across all considered graphs. In fact, for small subgraph sizes (≤ 100), it is the only method that attains, or comes close to attaining the upper bound on the optimal triangle density. Our results demonstrate that although $TDkS$ is NP-hard and difficult to approximate in the worst-case, the Lovász relaxation can still prove to be an effective tool for detecting triangle-dense subgraphs in real-world graphs.
- Although the Mirror Descent algorithm aims to detect subgraphs with high triangle density, it turns in a commendable performance in terms of edge density as well (right column in Figure 1). In fact, for subgraph sizes ≤ 200 , it outperforms the dedicated edge based formulations, often by a significant margin and comes closest to attaining the edge density upper bound. This can be viewed as a consequence of the Kruskal-Katona theorem which formalizes the following intuitive notion: if a subgraph has high triangle density, then it must possess high edge density as well. Looking at Figure 1 (right column) confirms this observation.
- For large subgraph sizes, the edge density obtained by Mirror Descent / $TDkS$ is often second (although by a small margin) to that obtained by applying the Lovász relaxation for DkS . Empirically, we note that this occurs (i.e., the blue curve “overtakes” the magenta curve) when the edge density falls below the 50% threshold. A possible explanation is as follows. Turán’s theorem [38] states that a graph can exhibit an edge density at most 0.5 without harboring any triangles. In other words, below this threshold, there do exist graphs with edge density up to 50% while containing very few triangles. Consequently, in the regime where the densest subgraph of a given size has edge density upper bounded by 0.5, employing a density measure based on edges may prove to be more beneficial as opposed to using triangles, if one cares more about edge density.

7 CASE STUDY – SINGLE DOCUMENT KEYWORD EXTRACTION

Overview: We consider an application of $TDkS$ to the real-world problem of unsupervised single document keyword

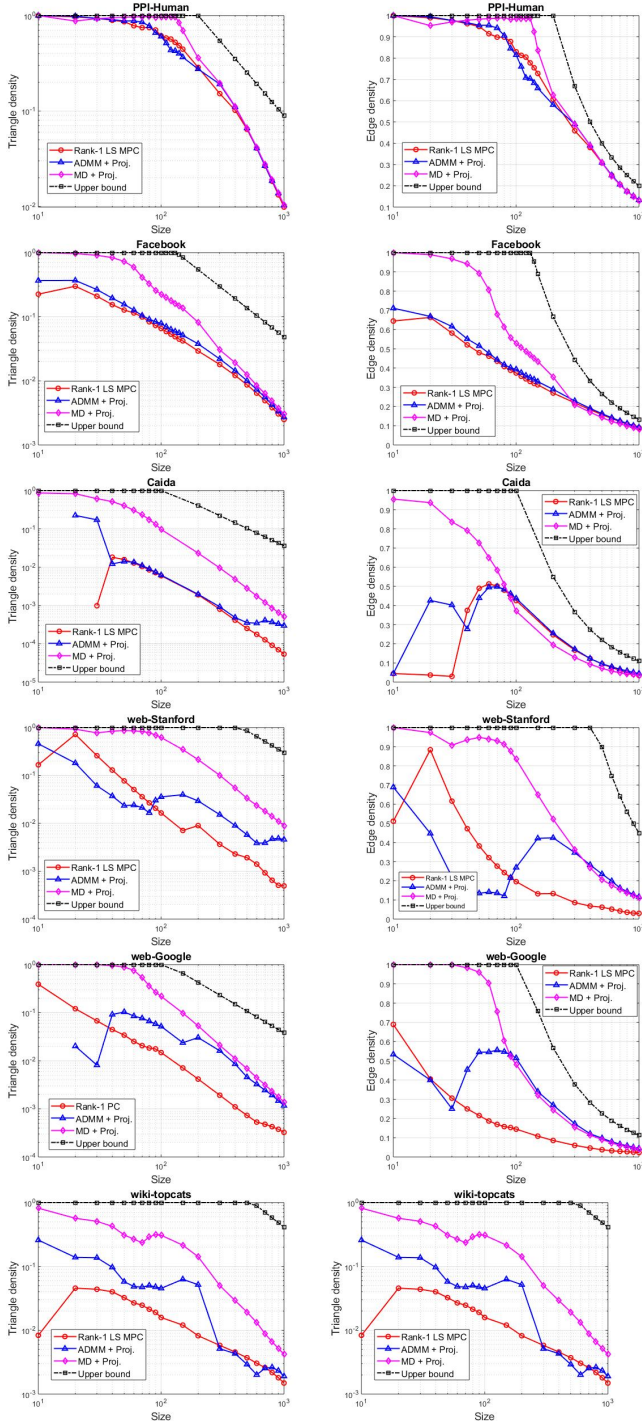


Fig. 1: **Left column:** Triangle density (on a log-scale) vs. size. **Right column:** Edge density vs. size. Red (Rank-1 least-squares matrix principal component) and blue (ADMM) curves are methods for D_kS while the magenta curve (MD) is for TD_kS . The black curve in the right column depicts the edge density upper bound for D_kS . This bound combined with the Kruskal-Katona theorem also yields an upper bound on TD_kS (black curve in the left column).

extraction. For this purpose, we adopt the popular Graph-of-Words (GoW) model of [39]. In this model, a given text document is represented as an undirected graph, where vertices correspond to unique words occurring in the document and an edge connects a pair of vertices if the words they represent co-occur within a window of pre-specified length L which is slid from the start to the end of the document, spanning across sentences. For example, the choice $L = 2$ corresponds to connecting pairs of words appearing in bi-grams, whereas $L = 3$ connects triplets of words forming tri-grams via triangles. In this context, extracting a dense subgraph from a GoW-representation of a text document corresponds to determining a highly cohesive subset of keywords, which can form an informative summary of a document’s content.

Prior Work: Such an approach has been adopted previously in [5] to extract summaries of trending news stories from streams of Twitter data, based on a GoW model with window length $L = 2$. Later work of [6] considered GoW representations of general window length $L \geq 2$, and proposed the following two-step approach. First, a k -core decomposition [40] is performed on the graph followed by a k -truss⁴ decomposition [41] refinement step in order to extract informative keywords. We point out that such density based approaches to keyword extraction, which quantify importance of keywords based on how frequently they co-occur with other groups of words (i.e., cohesiveness), are different from random-walk based approaches such as TextRank [39], which assign importance to words based on their degree alone. It is known that the former approach based on density generally outperforms the latter on real-world data [6], [42].

Limitations: That being said, the majority of density based approaches to keyword extraction still suffer from two main limitations: (i) they do not feature a natural means of explicitly controlling the size of the extracted summary, and (ii), few methods can effectively deal with higher-order word co-occurrences (corresponding to GoW models with $L \geq 3$). An exception to the second limitation is the method of [6], which employs k -truss decomposition as a proxy for detecting a triangle-dense subgraph. However, the authors of [6] noted that their approach was effective in extracting informative summaries from real text data only for GoW models constructed from higher-order n -grams (with window length L at least 4, 5). This observation was attributed to the fact that the number of triangles in the constructed graph increases with L , which in turn facilitates the detection of a dense k -truss. Note that constructing such higher-order GoW representations comes at the expense of computing all higher-order n -grams, which can prove to be expensive for large documents.

We argue that TD_kS is well positioned to address these shortcomings. Since our problem allows pre-specification of the desired subgraph size, given a GoW representation of a document, it enables direct control of the size of the extracted summary. Additionally, the objective function of TD_kS is directly geared towards maximizing the weighted sum of induced triangles in the subgraph, which we intu-

4. A k -truss is a subgraph where every edge participates in at least $k - 2$ triangles.

itively expect will allow our formulation to extract meaningful summaries from simple GoW representations constructed using tri-grams alone (with window length $L = 3$). **Experiments:** In order to provide empirical corroboration of our hypothesis, we applied TD k S to the problem of generating high-level descriptions of recently released Hollywood movies based on publicly available text reviews submitted by movie critics. We used the popular review aggregator website <https://www.metacritic.com/> to obtain reviews for the following movies, which are briefly described below.

- **John Wick 3: Parabellum**⁵ - an action movie in the John Wick franchise directed by Chad Stahelski and starring Keanu Reeves as the titular character.
- **Baby Driver**⁶ - a heist movie directed by Edgar Wright.
- **Arrival**⁷ - a science fiction movie that doubles as a drama; directed by Denis Villeneuve and starring Amy Adams in the lead role.
- **Hereditary**⁸ - a horror movie centered around the evil that befalls a family; written and directed by Ari Aster (in his directorial debut) and starring Toni Collette in the lead role.
- **Mad Max: Fury Road**⁹ - an action movie set in a wasteland that serves as a soft reboot of the Mad Max franchise; directed by George Miller.
- **Joker**¹⁰ - an origin story about a popular comic book villain, The Joker; directed by Todd Phillips and starring Joaquin Phoenix in the lead role.

For each movie, we collected all the available reviews submitted by movie critics into a single text document. On average, a movie had 45 reviews, each of which represents a summary/opinion of a critic. We performed a simple pre-processing step on each text document where we filtered out short, commonly occurring words using the list of stop-words provided in Python’s Natural Language Processing Toolkit (NLTK)¹¹. The resulting text document comprised, on average, 740 unique words used to describe a movie by multiple critics. Then, from each document, we enumerated all bi-grams and tri-grams in order to construct two different GoW representations. In the first model, triplets of vertices are connected by a triangle if their corresponding words co-occur together in a tri-gram; each triangle has a weight denoting the co-occurrence probability of the word triplet in the text document. We apply the Mirror Descent algorithm developed for the Lovász relaxation of TD k S in order to detect a triangle-dense subgraph of a pre-specified size k . Note that the vertices comprising such a subgraph correspond to words which frequently co-occur together in sub-sentences of length 3. As a baseline, we construct a second GoW model, where a pair of vertices are connected by an edge if their respective words co-occur in a bi-gram, with each edge being weighted by the pair-wise

co-occurrence probability of the word pair. We then apply the ADMM algorithm developed in [16] for the Lovász relaxation of D k S to extract a dense subgraph of size- k based on pair-wise cohesiveness. For a fixed summary size k , we expect the summary generated using TD k S to be more cohesive and informative as compared to D k S, since the former approach exploits higher-order co-occurrences in text.

Results and Discussion: The outcomes of our experiments are depicted in Tables 4 and 5, which display the 30-word summaries generated for each movie using TD k S and D k S respectively. At a high level, it is evident that the movie summaries generated from the tri-gram based GoW model using TD k S are substantially more cohesive and descriptive compared to that obtained from the bi-gram GoW model using D k S. A minor drawback, however, is that the depth of coverage obtained from the tri-gram model sometimes comes at the expense of breadth. For example, from Table 5 it can be seen that the summary of each movie contains the name of the director, but this is true for 4/6 movies in Table 4. The similarities and differences in the two kinds of summaries for each movie are elucidated below.

- **John Wick 3: Parabellum** - both summaries extract the name of the movie and the fact that it belongs in a franchise. However, the TD k S summary additionally identifies that the movie is the third in the franchise. While the D k S summary contains the director, it does not contain the full name of the movie’s main star Keanu Reeves, who is present in the TD k S summary. The D k S summary reveals the movie’s content as being violent and also mentions the choreography, although few additional details are present. In contrast, the other summary reveals that the critics found the level of violence excessive, but enjoyed the choreography of the fight scenes.
- **Baby Driver** - the twin summaries contain the movie’s name and that of the director. The TD k S summary reveals that the critics enjoyed the story and pacing of the movie (**tale, perfect, enjoy, fun, fast, furious, inventive**), and also the fact that the movie plot simultaneously combines elements of action, romance and musicals (**heist, romance, love, musicals**). The soundtrack of the movie is also mentioned (**jukebox**). In contrast, the D k S summary is sparser in details regarding the plot, and beyond the action elements does not pick up on the other aspects of the movie’s plot. The soundtrack is also mentioned (**music, beat**), and the fact the movie was released in the summer.
- **Arrival** - common to both summaries are the names of the movie, director and the lead actress. However, the TD k S summary identifies the genre of the film as being both science fiction *and* a drama – the D k S summary does not pick up the latter. This serves as key context for description of the movie’s content which was praised by the critics for the weight of its intellectual themes, as well as the moving story. In contrast, the D k S summary, while picking up on some of the emotional themes in the story, misses out this key intellectual aspect.
- **Hereditary** - both summaries highlight the movie’s

5. <https://www.metacritic.com/movie/john-wick-chapter-3---parabellum>

6. <https://www.metacritic.com/movie/baby-driver>

7. <https://www.metacritic.com/movie/arrival>

8. <https://www.metacritic.com/movie/hereditary>

9. <https://www.metacritic.com/movie/mad-max-fury-road>

10. <https://www.metacritic.com/movie/joker>

11. <https://www.nltk.org/>

name, the genre being horror, and the fact that the first-time director was also the writer. The TD k S summary additionally reveals that the critics praised the direction (**excels**, **smugly**, **promising**) and found the movie impressive, and even designated it as a masterpiece. They also commented on the fact that the movie is more in the vein of old school horror movies - we note that these details are missing from the D k S summary. However, the latter summary does mention the performance of the lead actress, Toni Collette.

- **Mad Max: Fury Road** - while having the movie's and director's name in common with the D k S summary, the triangle-based summary contains substantially more information about the content - the fact the reviewers praised the style and energy of the action sequences (**delivers**, **effective**, **loco**), and concurred that the movie marked a successful return to the Mad Max universe, and even surpassed previous installments.
- **Joker** - the TD k S summary reveals that the movie is a re-imagining of a comic book character (the D k S summary adds the detail that it is a DC comic), and is anchored by a powerful performance from the actor Joaquin Phoenix, which is described as being Oscar worthy (Joaquin Phoenix indeed won the Oscar for Best Male Actor for his portrayal of the Joker). The critics also described the movie as being depressing, but at the same time also offered praise (**engrossing**, **masterful**, **iconic**, **well**, **made**, **great**, **fine**). Praise was not uniform, however - downsides are mentioned, but more specific details are not captured. By comparison, the D k S summary is spartan in its descriptive quality - the main addition being the director's name.

8 CONCLUSIONS

We considered the triangle-densest- k -subgraph problem (TD k S) which aims to compute the size k subgraph with the largest number of induced triangles. Unfortunately, not only is the problem NP-hard, but it is also difficult to approximate in polynomial-time, in the worst-case sense. With the aim of computing high-quality, sub-optimal solutions on real-world instances, we exploited the fact that the cost function of TD k S is submodular to construct a convex relaxation of the problem based on the Lovász extension of submodular functions. As we derived an analytical functional form for the extension, this enabled us to devise a Mirror Descent algorithm for efficiently solving the problem at scale. Our results on real-world graphs showcased that our approach can effectively exploit triangle motifs to attain state-of-the-art performance, and can provide a more effective means of exploring the density-size trade-off compared to baselines that only use edges for density maximization. Additionally, we utilized the problem of document summarization to showcase that TD k S can generate more informative word summaries compared to D k S.

TABLE 2: Top-30 words which most frequently co-occur in trigrams obtained from movie review text data; detected using the Lovasz relaxation for TD k S (with $k = 30$). Key terms are highlighted in bold.

John Wick 3	Baby Driver	Arrival	Hereditary	Mad Max: Fury Road	Joker
movies	movie	movie	movie	films	movie
john	movies	film	film	mad	joker
wick	baby	arrival	hereditary	max	comic
3	driver	makes	also	fury	book
parabellum	almost	villeneuve	director	road	character
still	feels	take	writer	made	reimagined
starring	edgar	almost	ari	director	movies
keanu	wright	science	aster	george	one
reeves	literalizes	fiction	first	millar	anchored
may	like	sci	time	also	performance
another	always	fi	feature	action	joaquin
one	watching	drama	debut	sequences	phoenix
chapter	tale	used	us	scenes	oscar
series	perfect	adams	buy	delivers	worthy
third	enjoy	head	gives	effective	allows
anything	fun	like	excels	loco	would
action	fast	moving	smugly	style	study
fight	furious	thoughtful	promising	orgy	depressing
scenes	inventive	great	impressively	stands	engrossing
gratuitously	nothing	intimate	masterpiece	return	masterful
violent	much	poignant	horror	revived	iconic
choreographed	going	intellectual	agitating	universe	well
gorgeously	musical	intelligent	unnerving	wasteland	made
grander	romance	beguiling	old	surpassed	great
baring	heist	leaves	school	previous	fine
fluid	jukebox	questions	makes	interested	almost
story	love	truly	making	indebted	created
closer	plays	best	family	stone	downside
draw	sorts	something	takes	cold	killer
parabolically	mashed	conundrum	much	warrior	work

TABLE 3: Top-30 words which most frequently co-occur in bigrams obtained from movie review text data; detected using the Lovasz relaxation for D k S (with $k = 30$). Key terms are highlighted in bold.

John Wick 3	Baby Driver	Arrival	Hereditary	Mad Max: Fury Road	Joker
film	film	film	film	movie	film
movies	movie	movie	movie	film	movie
movie	movies	arrival	hereditary	mad	joker
john	baby	much	see	max	comic
wick	driver	less	director	fury	book
3	director	denis	writer	road	character
parabellum	edgar	villeneuve	ari	movies	dc
though	wright	like	aster	films	much
get	something	science	first	george	work
first	also	fiction	time	millar	performance
like	much	sci	mind	never	joaquin
director	story	fi	debut	much	phoenix
chad	fun	one	feature	things	arthur
stahelski	nothing	adams	every	us	villain
reeves	even	performance	gets	mayhem	todd
chapter	heist	moving	toni	action	phillips
series	action	emotional	collette	chase	study
franchise	crime	human	performance	makes	one
action	work	great	takes	also	bad
still	us	story	even	get	social
violence	one	best	long	work	last
violent	high	life	genre	feels	well
good	filmmaking	truly	horror	first	point
far	love	made	family	even	also
point	part	open	supernatural	muscle	go
scenes	like	time	atmosphere	new	enough
two	beat	ideas	dread	like	anything
choreographed	music	first	much	post	specific
set	car	makes	new	story	like
new	summer	questions	one	one	movies

REFERENCES

- [1] B. Hooi, H. A. Song, A. Beutel, N. Shah, K. Shin, and C. Faloutsos, "Fraudar: Bounding graph fraud in the face of camouflage," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 895–904, ACM, 2016.
- [2] S. Zhang, D. Zhou, M. Y. Yildirim, S. Alcorn, J. He, H. Davulcu, and H. Tong, "Hidden: hierarchical dense subgraph detection with application to financial fraud detection," in *Proceedings of the 2017 SIAM International Conference on Data Mining*, pp. 570–578, SIAM, 2017.
- [3] Y. Podolyan and G. Karypis, "Common pharmacophore identification using frequent clique detection algorithm," *Journal of chemical information and modeling*, vol. 49, no. 1, pp. 13–21, 2009.
- [4] B. Saha, A. Hoch, S. Khuller, L. Raschid, and X.-N. Zhang, "Dense subgraphs with restrictions and applications to gene annotation graphs," in *Annual International Conference on Research in Computational Molecular Biology*, pp. 456–472, Springer, 2010.
- [5] A. Angel, N. Sarkas, N. Koudas, and D. Srivastava, "Dense subgraph maintenance under streaming edge weight updates for real-time story identification," in *Proceedings of the 38th International Conference on Very Large Data Bases*, pp. 574–585, VLDB Endowment, 2012.

- [6] A. Tixier, F. Malliaros, and M. Vazirgiannis, "A graph degeneracy-based approach to keyword extraction," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1860–1870, 2016.
- [7] J. Cadena, F. Chen, and A. Vullikanti, "Graph anomaly detection based on steiner connectivity and density," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 829–845, 2018.
- [8] A. V. Goldberg, *Finding a maximum density subgraph*. Technical report, University of California Berkeley, CA, 1984.
- [9] M. Charikar, "Greedy approximation algorithms for finding dense components in a graph," in *International Workshop on Approximation Algorithms for Combinatorial Optimization*, pp. 84–95, Springer, 2000.
- [10] C. Tsourakakis, F. Bonchi, A. Gionis, F. Gullo, and M. Tsiarli, "Denser than the densest subgraph: extracting optimal quasi-cliques with quality guarantees," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 104–112, 2013.
- [11] U. Feige, D. Peleg, and G. Kortsarz, "The dense k-subgraph problem," *Algorithmica*, vol. 29, no. 3, pp. 410–421, 2001.
- [12] S. Khot, "Ruling out ptas for graph min-bisection, dense k-subgraph, and bipartite clique," *SIAM Journal on Computing*, vol. 36, no. 4, pp. 1025–1071, 2006.
- [13] A. Bhaskara, M. Charikar, V. Guruswami, A. Vijayaraghavan, and Y. Zhou, "Polynomial integrality gaps for strong sdp relaxations of densest k-subgraph," in *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pp. 388–405, SIAM, 2012.
- [14] P. Manurangsi, "Almost-polynomial ratio hardness of approximating densest k-subgraph," in *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 954–961, 2017.
- [15] D. Papailiopoulos, I. Mitliagkas, A. Dimakis, and C. Caramanis, "Finding dense subgraphs via low-rank bilinear optimization," in *International Conference on Machine Learning*, pp. 1890–1898, 2014.
- [16] A. Konar and N. D. Sidiropoulos, "Exploring the subgraph density-size trade-off via the lovasz extension," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 743–751, 2021.
- [17] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, p. 440, 1998.
- [18] C. Tsourakakis, "The k-clique densest subgraph problem," in *Proceedings of the 24th International Conference on World Wide Web*, pp. 1122–1132, International World Wide Web Conferences Steering Committee, 2015.
- [19] C. E. Tsourakakis, J. Pachocki, and M. Mitzenmacher, "Scalable motif-aware graph clustering," in *Proceedings of the 26th International Conference on World Wide Web*, pp. 1451–1460, 2017.
- [20] A. R. Benson, D. F. Gleich, and J. Leskovec, "Higher-order organization of complex networks," *Science*, vol. 353, no. 6295, pp. 163–166, 2016.
- [21] Y. Zhang and S. Parthasarathy, "Extracting analyzing and visualizing triangle k-core motifs within networks," in *2012 IEEE 28th international conference on data engineering*, pp. 1049–1060, IEEE, 2012.
- [22] A. Konar and N. D. Sidiropoulos, "The triangle-densest-k-subgraph problem: Hardness, lovasz extension, and application to document summarization," in *Proceedings of AAAI*, 2022.
- [23] L. Lovász, "Submodular functions and convexity," in *Mathematical programming the state of the art*, pp. 235–257, Springer, 1983.
- [24] F. Bach et al., "Learning with submodular functions: A convex optimization perspective," *Foundations and Trends® in Machine Learning*, vol. 6, no. 2-3, pp. 145–373, 2013.
- [25] S. Fujishige, *Submodular functions and optimization*. Elsevier, 2005.
- [26] J. B. Kruskal, "The number of simplices in a complex," *Mathematical optimization techniques*, vol. 10, pp. 251–278, 1963.
- [27] G. O. Katona, "A simple proof of the erdős-chao ko-rado theorem," *Journal of Combinatorial Theory, Series B*, vol. 13, no. 2, pp. 183–184, 1972.
- [28] M. Mitzenmacher, J. Pachocki, R. Peng, C. Tsourakakis, and S. C. Xu, "Scalable large near-clique detection in large-scale networks via sampling," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 815–824, 2015.
- [29] J. Edmonds, "Submodular functions, matroids, and certain polyhedra," *Edited by G. Goos, J. Hartmanis, and J. van Leeuwen*, vol. 11, 1970.
- [30] A. Schrijver, *Combinatorial optimization: polyhedra and efficiency*, vol. 24. Springer Science & Business Media, 2003.
- [31] S. Bubeck, "Convex optimization: Algorithms and complexity," *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.
- [32] R. T. Rockafellar, *Convex analysis*. No. 28, Princeton university press, 1970.
- [33] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," *Operations Research Letters*, vol. 31, no. 3, pp. 167–175, 2003.
- [34] L. Condat, "A primal-dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms," *Journal of Optimization Theory and Applications*, vol. 158, no. 2, pp. 460–479, 2013.
- [35] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection." <http://snap.stanford.edu/data>, 2014.
- [36] J. Kunegis, "KONECT – The Koblenz Network Collection," in *Proc. Int. Conf. on World Wide Web Companion*, pp. 1343–1350, 2013.
- [37] S. Suri and S. Vassilvitskii, "Counting triangles and the curse of the last reducer," in *Proceedings of the 20th international conference on World wide web*, pp. 607–614, 2011.
- [38] P. Turán, "On an external problem in graph theory," *Mat. Fiz. Lapok*, vol. 48, pp. 436–452, 1941.
- [39] R. Mihalcea and P. Tarau, "TextRank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, pp. 404–411, 2004.
- [40] S. B. Seidman, "Network structure and minimum degree," *Social networks*, vol. 5, no. 3, pp. 269–287, 1983.
- [41] J. Cohen, "Trusses: Cohesive subgraphs for social network analysis," *National security agency technical report*, vol. 16, no. 3.1, 2008.
- [42] F. Rousseau and M. Vazirgiannis, "Main core retention on graph-of-words for single-document keyword extraction," in *European Conference on Information Retrieval*, pp. 382–393, Springer, 2015.

Aritra Konar received the BTech degree in electronics and communications engineering from the West Bengal University of Technology, West Bengal, India, in 2011, and the MS and PhD degrees in electrical engineering from the University of Minnesota, Minneapolis, Minnesota, in 2014 and 2017, respectively. He is currently an assistant professor in the Department of Electrical Engineering at KU Leuven, Belgium. His research interests include network science, graph mining, nonlinear optimization, and signal processing.

Nicholas D. Sidiropoulos (Fellow, IEEE) received the Diploma in electrical engineering from the Aristotle University of Thessaloniki, Greece, and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland at College Park, College Park, MD, USA, in 1988, 1990, and 1992, respectively. He is the Louis T. Rader Professor at the University of Virginia. He has previously served on the faculty of the University of Minnesota, and the Technical University of Crete, Greece. His research interests are in signal processing, communications, optimization, and tensor decomposition with applications in machine learning and communications. He received the NSF/CAREER award in 1998, the IEEE Signal Processing Society (SPS) Best Paper Award in 2001, 2007, 2011, and 2023, and the IEEE SPS Donald G. Fink Overview Paper Award in 2023. He served as IEEE SPS Distinguished Lecturer (2008–2009), as Vice President - Membership of IEEE SPS (2017–2019), and as chair of the SPS IEEE Fellow Evaluation Committee (2020–2021). He received the IEEE Signal Processing Society Meritorious Service Award (2010), the Distinguished Alumni Award of the Department of ECE, University of Maryland (2013), the EURASIP Technical Achievement Award (2022), and the IEEE SPS Claude Shannon - Harry Nyquist Technical Achievement Award (2023). He is a Fellow of EURASIP (2014).

Supplement

Aritra Konar, and Nicholas D. Sidiropoulos, *Fellow, IEEE*



APPENDIX A: PROOF OF THEOREM 1

We provide a simple reduction from the decision version of the maximum clique problem, which is known to be NP-complete [1]. For an unweighted, undirected graph $\mathcal{G}$, let $\omega(\mathcal{G})$ denote the size of its largest clique. Given $\mathcal{G}$ and a positive integer $\alpha \geq 3$, the decision variant of the maximum clique problem asks whether the maximum clique size $\omega(\mathcal{G})$ is at least as large as α . For such an instance, let $f_\alpha(\mathbf{x}^*)$ denote the optimal value of problem (1) with $k = \alpha$. Note that $f_\alpha(\mathbf{x}^*)$ cannot exceed $\binom{\alpha}{3}$, with equality attained if and only if $\mathcal{G}$ contains an α -clique, i.e., the answer to $\omega(\mathcal{G}) \geq \alpha$ is affirmative. Hence, solving problem (1) is at least as hard as solving an arbitrary decision instance of the maximum clique problem.

APPENDIX B: PROOF OF THEOREM 2

Our starting point is the Kruskal-Katona Theorem [2], [3], which states that for any graph $\mathcal{G}$ of size k , its edge-density $\rho_2(\mathcal{G}, k)$ and triangle-density $\rho_3(\mathcal{G}, k)$ must obey the relationship

$$\rho_2(\mathcal{G}, k) \geq \rho_3(\mathcal{G}, k)^{2/3}. \quad (1)$$

As an immediate consequence of this result, we obtain the following relationship between the optimal values of TDkS and DkS, for a fixed input.

$$\rho_2(\mathcal{M}^*, k) \geq \rho_3(\mathcal{S}^*, k)^{2/3}, \quad (2)$$

where the size- k vertex subsets $\mathcal{M}^*$ and $\mathcal{S}^*$ represent optimal solutions of DkS and TDkS respectively. Combining the above result with the statement of Fact 2, we obtain the following relationship

$$\rho_2(\mathcal{M}^*, k) \geq \rho_2(\mathcal{M}, k) \geq O\left(\frac{1}{\alpha(n)}\right) \rho_3(\mathcal{S}^*, k)^{2/3}, \quad (3)$$

which must hold for any joint instance of TDkS and DkS, and for any polynomial-time approximation algorithm for DkS, which produces a subgraph $\mathcal{M}$ of size k as output.

We are now ready to establish the proof of our main result, which relies on an argument based on contradiction. Assume that there exists a polynomial-time approximation algorithm $\mathcal{A}$ for TDkS which outputs a size- k subgraph $\mathcal{S}$ whose triangle density $\rho_3(\mathcal{S}, k)$ is guaranteed to be no worse than a fraction $O(1/(\alpha(n))^{3/2-\epsilon})$ of the optimal triangle density $\rho_3(\mathcal{S}^*, k)$, for some $\epsilon > 0$. In other words, the triangle density achieved by the output of the algorithm $\mathcal{A}$ obeys the relationship

$$\rho_3(\mathcal{S}^*, k) \geq \rho_3(\mathcal{S}, k) \geq \frac{C_1}{(\alpha(n))^{3/2-\epsilon}} \cdot \rho_3(\mathcal{S}^*, k) \quad (4)$$

for every instance of TDkS (here $C_1 > 0$ is a universal constant).

Let $\rho_2(\mathcal{S}, k)$ denote the edge density of the subgraph induced by $\mathcal{S}$. Applying the Kruskal-Katona theorem and the second inequality in (4), we obtain the following lower bound on $\rho_2(\mathcal{S}, k)$.

$$\rho_2(\mathcal{S}, k) \geq \rho_3(\mathcal{S}, k)^{2/3} \geq \frac{C_2}{(\alpha(n))^{1-\delta}} \cdot \rho_3(\mathcal{S}^*, k)^{2/3}, \quad (5)$$

where $C_2 := C_1^{2/3} > 0$ and $\delta := \frac{2\epsilon}{3} > 0$. Furthermore, we point out that $\rho_2(\mathcal{S}, k)$ cannot exceed the optimal edge density $\rho_2(\mathcal{M}^*, k)$ for the same problem instance. Combining this fact with (5), we obtain

$$\rho_2(\mathcal{M}^*, k) \geq \rho_2(\mathcal{S}, k) \geq \frac{C_2}{(\alpha(n))^{1-\delta}} \cdot \rho_3(\mathcal{S}^*, k)^{2/3}, \quad (6)$$

which is a clear contradiction of (3).

APPENDIX C: PROOF OF THEOREM 3

Recall that $F(\mathcal{S})$ can be linearly decomposed over the set of triangles Δ as follows

$$F(\mathcal{S}) = \sum_{(u,v,w) \in \Delta} F_{uvw}(\mathcal{S}). \quad (7)$$

Since submodularity is preserved under addition and restriction, in order to establish the desired result, it suffices to show that each component function $F_{uvw}(\mathcal{S})$ is submodular. To this end, note that by construction, the domain of F_{uvw} is the power-set of the reduced ground set comprising vertices $\{u, v, w\}$. In order to demonstrate that F_{uvw} is submodular, we invoke the following fact about submodular functions - a function is submodular if and only if all its projections onto 2 variables are submodular [4]. Considering the discrete polynomial form of F_{uvw}

$$f_{uvw}(x_u, x_v, x_w) = -w_t x_u x_v x_w,$$

the above test for submodularity is tantamount to fixing any one of the variables and testing whether the function is submodular in the remaining pair of free variables, for all such possible configurations. Since the function is symmetric with respect to (w.r.t.) its arguments, it follows without loss of generality that we can fix x_u and check whether the projection of f_{uvw} onto (x_v, x_w) is submodular. This in turn simplifies to testing whether $g_t(x_v, x_w) := f(1, x_v, x_w) = -w_t x_v x_w$ is submodular (note that the case

$x_u = 0$ is trivial). Recalling the definition of submodularity for functions of two variables, we note that the condition

$$g_t(1, 0) + g_t(0, 1) \geq g_t(0, 0) + g_t(1, 1)$$

is always satisfied since w_t is non-negative. Hence, each component function $F_{uvw}(S)$ is submodular, from which the claim follows.

Remark: We point out that when the graph is unweighted and each triangle has unit weight, the above theorem recovers the result of [5, Theorem 3] as a special case. Furthermore, as we now explain, the above line of reasoning can be extended to establish a considerably more general result regarding submodularity of the weighted sum of induced higher-order cliques in a subgraph. Consider an undirected graph on n vertices represented as $\mathcal{G} = (\mathcal{V}, \mathcal{C}_k, w)$, where $\mathcal{V}$ denotes the vertex set, $\mathcal{C}_k$ denotes the set of all k -cliques in the graph (with $k \geq 3$), and each clique $c \in \mathcal{C}_k$ comprising k vertices $\{i_1, \dots, i_k\} \subset \mathcal{V}$ is associated with a non-negative weight w_c . Given a subset of vertices $\mathcal{S} \subset \mathcal{V}$, define the function

$$h(\mathbf{x}) := \sum_{(i_1, i_2, \dots, i_k) \in \mathcal{C}_k} w_c \prod_{j=1}^k x_{i_j}, \quad (8)$$

where $\mathbf{x}$ is the binary indicator vector of $\mathcal{S}$. Note that for a given subset of vertices $\mathcal{S} \subset \mathcal{V}$, h returns the weighted sum of k -cliques induced by $\mathcal{S}$. Then, we have the following result.

Corollary 1. $-h$ is a submodular function.

Proof. See Appendix D. $\square$

APPENDIX D: PROOF OF COROLLARY 1

It suffices to show that each component function

$$h_c(x_{i_1}, \dots, x_{i_k}) := -w_c \prod_{j=1}^k x_{i_j} \quad (9)$$

corresponding to a given clique $c \in \mathcal{C}_k$ is submodular. To do so, we employ the same sequence of observations as in the previous result. Namely, that each function is defined on a reduced ground set comprising vertices $\{i_1, \dots, i_k\}$ and is symmetric w.r.t. to its arguments. Hence, w.l.o.g., we can fix the first $k-2$ variables $(x_{i_1}, \dots, x_{i_{k-2}})$ to one and test whether the projection of the function

$$h_c(\underbrace{1, \dots, 1}_{k-2}, x_{i_{k-1}}, x_{i_k}) = -w_c x_{i_{k-1}} x_{i_k} \quad (10)$$

onto the pair of variables $(x_{i_{k-1}}, x_{i_k})$ is submodular. Since w_c is non-negative, as shown previously, this is indeed the case.

APPENDIX E: PROOF OF LEMMA 3

From the definitions of F_{uvw} and the base polytope of a submodular function, we obtain the following characterization of $\mathcal{B}_{F_{uvw}}$.

$$\begin{aligned} \mathcal{B}_{F_{uvw}} = & \left\{ \mathbf{y} \in \mathbb{R}^n : y_u \leq 0, y_v \leq 0, y_w \leq 0, \right. \\ & y_u + y_v \leq 0, y_v + y_w \leq 0, y_u + y_w \leq 0, \\ & \left. y_u + y_v + y_w \leq -w_t, y_u + y_v + y_w = -w_t \right\}. \end{aligned} \quad (11)$$

Observe that the first three inequalities imply the next three, whereas the final inequality is made redundant by the lone equality. The result then follows after eliminating all redundant inequalities.

APPENDIX F: PROOF OF THEOREM 4

Let f_{uvw} denote the Lovász extension of F_{uvw} . Invoking the result of Lemma 3, we obtain

$$f_{uvw}(\mathbf{x}) = \max_{\mathbf{y} \in \mathcal{B}_{F_{uvw}}} \mathbf{y}^T \mathbf{x} = -w_t \min\{x_u, x_v, x_w\}. \quad (12)$$

The result then follows since $f_L = \sum_{(u,v,w) \in \Delta} f_{uvw}$.

APPENDIX G: PROOF OF THEOREM 5

Our starting point is the result of Lemma 3, which states that the base polytope $\mathcal{B}_{F_{uvw}}$ is the probability simplex appropriately scaled and reflected about the origin. While every vector lying in this set can be expressed as the convex combination of the three extreme points of the scaled simplex, it possesses only two degrees of freedom (due to the summation constraint). Hence, it follows that $\mathcal{B}_{F_{uvw}}$ can be equivalently expressed as

$$\begin{aligned} \mathcal{B}_{F_{uvw}} = & \left\{ \alpha_t \geq 0, \beta_t \geq 0, \right. \\ & \alpha_t + \beta_t \leq 1 : -w_t(\alpha_t \mathbf{e}_u + \beta_t \mathbf{e}_v + (1 - \alpha_t - \beta_t) \mathbf{e}_w) \Big\}, \\ & = \left\{ \alpha_t \geq 0, \beta_t \geq 0, \right. \\ & \left. \alpha_t + \beta_t \leq 1 : -w_t(\alpha_t(\mathbf{e}_u - \mathbf{e}_w) + \beta_t(\mathbf{e}_v - \mathbf{e}_w) + \mathbf{e}_w) \right\}. \end{aligned} \quad (13)$$

Define the variables $\gamma_t := 1 - 3\alpha_t$ and $\delta_t := 1 - 3\beta_t$. By construction, these variables obey the relationships

$$\gamma_t \leq 1, \delta_t \leq 1, \gamma_t + \delta_t \geq -1. \quad (14)$$

On performing a change of variables, we can express $\mathcal{B}_{F_{uvw}}$ as

$$\begin{aligned} \mathcal{B}_{F_{uvw}} = & -w_t \left[\frac{(1 - \gamma_t)(\mathbf{e}_u - \mathbf{e}_w) + (1 - \delta_t)(\mathbf{e}_v - \mathbf{e}_w)}{3} + \mathbf{e}_w \right], \\ & = -w_t \left[\frac{(\mathbf{e}_u + \mathbf{e}_v + \mathbf{e}_w)}{3} - \frac{\gamma_t(\mathbf{e}_u - \mathbf{e}_w) + \delta_t(\mathbf{e}_v - \mathbf{e}_w)}{3} \right]. \end{aligned} \quad (15)$$

Since the polytope $\mathcal{B}_F$ is the Minkowski sum of the polytopes $\mathcal{B}_{F_{uvw}}$, we obtain the following expression

$$\begin{aligned} \mathcal{B}_F = & - \sum_{(u,v,w) \in \Delta} w_t(\mathbf{e}_u + \mathbf{e}_v + \mathbf{e}_w) \\ & + \sum_{(u,v,w) \in \Delta} w_t \left\{ \gamma_t(\mathbf{e}_u - \mathbf{e}_w) + \delta_t(\mathbf{e}_v - \mathbf{e}_w) \right\} \\ & = -\mathbf{t} + \sum_{(u,v,w) \in \Delta} w_t \left\{ \gamma_t(\mathbf{e}_u - \mathbf{e}_w) + \delta_t(\mathbf{e}_v - \mathbf{e}_w) \right\}, \end{aligned} \quad (16)$$

where for every triangle $t \in (u, v, w) \in \Delta$, the variables $\{\gamma_t, \delta_t\}$ obey (14). The second equality stems from the fact that the contribution of each vertex $u \in \mathcal{V}$ to the first

summand is precisely t_u . From the definition of the Lovász extension, we obtain

$$f_L(\mathbf{x}) = -\mathbf{t}^T \mathbf{x} + \sum_{(u,v,w) \in \Delta} w_t \max_{\substack{\gamma_t \leq 1, \delta_t \leq 1, \\ \gamma_t + \delta_t \geq -1}} \left\{ \gamma_t(\mathbf{e}_u - \mathbf{e}_w) + \delta_t(\mathbf{e}_v - \mathbf{e}_w) \right\}. \quad (17)$$

We focus on each maximization subproblem within the summand, which is a linear program in the variables $\{\gamma_t, \delta_t\}$. We now make the following observation regarding the constraint set defined by (14).

$$\text{conv}((1, 1), (-2, 1), (1, -2)) \Leftrightarrow \begin{cases} \gamma_t \leq 1, \delta_t \leq 1, \\ \gamma_t + \delta_t \geq -1. \end{cases} \quad (18)$$

With the convex-hull representation of the inequality constraints in hand, and exploiting the fact that the optimal solution of a linear program is always attained at an extreme point of its feasible set, we conclude that the solution of each subproblem is given by

$$\max\{x_u + x_v - 2x_w, x_v + x_w - 2x_u, x_u + x_w - 2x_v\},$$

which concludes the proof.

APPENDIX H: PROOF OF LEMMA 4

Define the function

$$h(\mathbf{x}) := \begin{cases} D_{\text{KL}}(\mathbf{x}, \mathbf{y}^r), & \mathbf{0} \leq \mathbf{x} \leq \mathbf{1}, \\ +\infty, & \text{otherwise.} \end{cases} \quad (19)$$

Then, the MDA update at each step (i.e., problem (31)) can be equivalently written as

$$\mathbf{x}^{r+1} = \arg \min_{\mathbf{1}^T \mathbf{x} = k} h(\mathbf{x}). \quad (20)$$

Let $\nu \in \mathbb{R}$ denote the dual variable associated with the equality constraint. Then, the Lagrangian of the above problem is

$$L(\mathbf{x}, \nu) := \begin{cases} \sum_{i=1}^n x_i (\log(x_i/y_i^r) + \nu - 1), & \mathbf{0} \leq \mathbf{x} \leq \mathbf{1}, \\ +\infty, & \text{otherwise.} \end{cases} \quad (21)$$

Let $(\mathbf{x}^*, \nu^*)$ denote a pair of primal-dual optimal solutions for (20). Applying the Karush-Kuhn-Tucker (KKT) conditions (which are necessary and sufficient for optimality in this case) yields the conditions

$$\mathbf{x}^* = \arg \min_{\mathbf{0} \leq \mathbf{x} \leq \mathbf{1}} L(\mathbf{x}, \nu^*), \quad \mathbf{1}^T \mathbf{x}^* = k. \quad (22a)$$

Since the Lagrangian is linearly separable in $\mathbf{x}$, the first condition can be further simplified as

$$x_i^* = \arg \min_{0 \leq x_i \leq 1} \left\{ x_i (\log(x_i/y_i^r) + \nu^* - 1) \right\}, \quad \forall i \in [n]. \quad (23)$$

Each sub-problem admits an analytical solution of the form

$$x_i^* = \min\{1, \exp(-\nu^*) y_i^r\}, \quad \forall i \in [n], \quad (24)$$

which completes the proof.

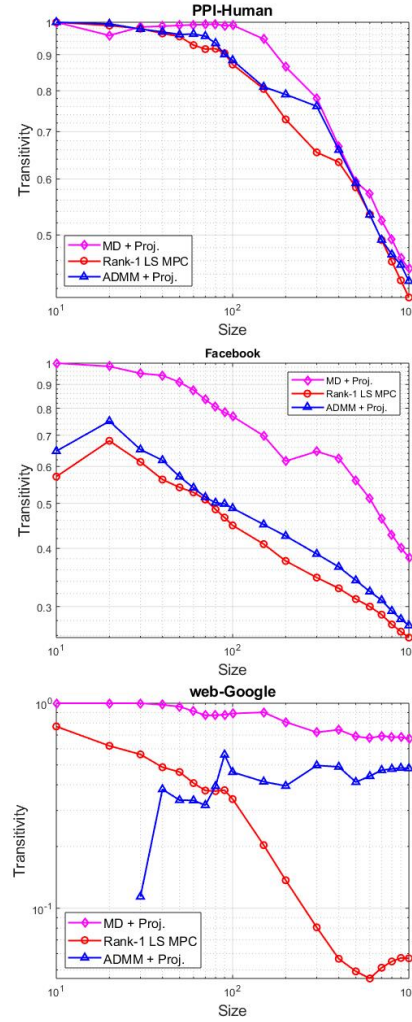


Fig. 1: Subgraph transitivity versus size.

APPENDIX I. TRANSITIVITY RESULTS

Figure 1 depicts the transitivity of the detected subgraphs as a function of the size on representative datasets. From Lemma 2, we know that for a fixed subgraph size, maximizing triangle density is a surrogate for maximizing transitivity as well. Since we developed a dedicated algorithm for TDkS, we expect it to perform better with regard to transitivity compared to the edge-based baselines, which is indeed the case.

APPENDIX J. TIMING RESULTS

The timing results of the methods (in seconds) across different datasets are presented in Figure 2. It can be seen that the rank-1 principal component approximation approach of [6] is consistently the fastest, as it simply requires computing the top-eigenvector of the graph adjacency matrix once followed by extracting the support of the top- k entries to determine the candidate subgraph of size k . In contrast, the Lovász relaxation schemes for both DkS and TDkS (blue and magenta curves respectively) require more time in general as they entail solving convex optimization problems via iterative methods. For the larger datasets (bottom row of Figure 2), we remark that the Mirror Descent algorithm for

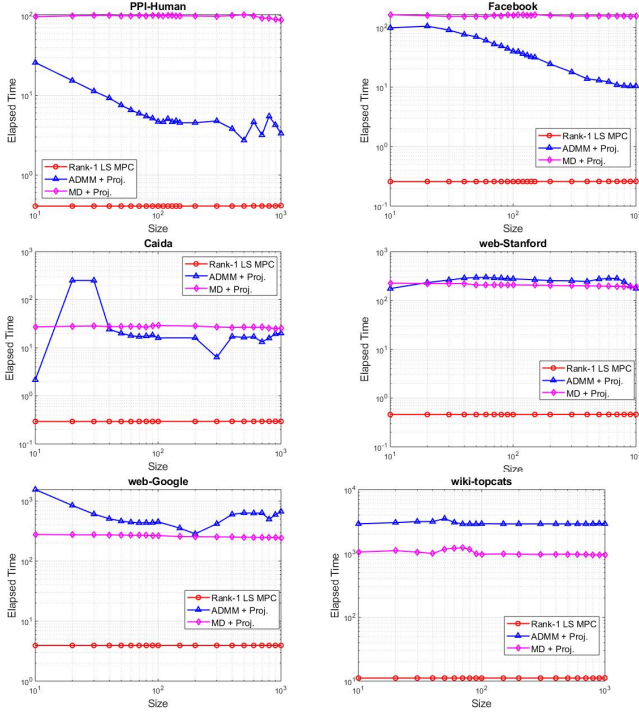


Fig. 2: Run-time of the methods across datasets. Red and blue curves are methods for DkS while the magenta curve is the MD algorithm for $TDkS$.

$TDkS$ is faster compared to the ADMM algorithm described in [7] for DkS .

APPENDIX K. MIRROR DESCENT PARAMETERS

TABLE 1: Step-size constants (c) and number of iterations (N)

Graph	c	N
PPI-HUMAN	40	500
FACEBOOK-B	100	500
CAIDA	5	400
WEB-STANFORD	1000	200
WEB-GOOGLE	100	200
WIKI-TOPCATS	1000	200

We run Algorithm 1 with the learning-rate $\beta^r = (c/t_{\max})\sqrt{\log n/r}$, where the choice of constant c and the number of total iterations is listed in Table 1 for each dataset.

REFERENCES

- [1] R. M. Karp, “Reducibility among combinatorial problems,” in *Complexity of computer computations*, pp. 85–103, Springer, 1972.
- [2] J. B. Kruskal, “The number of simplices in a complex,” *Mathematical optimization techniques*, vol. 10, pp. 251–278, 1963.
- [3] G. O. Katona, “A simple proof of the erdős-chao ko-rado theorem,” *Journal of Combinatorial Theory, Series B*, vol. 13, no. 2, pp. 183–184, 1972.
- [4] V. Kolmogorov and R. Zabih, “What energy functions can be minimized via graph cuts?,” in *European conference on computer vision*, pp. 65–81, Springer, 2002.

- [5] C. Tsourakakis, “The k -clique densest subgraph problem,” in *Proceedings of the 24th International Conference on World Wide Web*, pp. 1122–1132, International World Wide Web Conferences Steering Committee, 2015.
- [6] D. Papailiopoulos, I. Mitliagkas, A. Dimakis, and C. Caramanis, “Finding dense subgraphs via low-rank bilinear optimization,” in *International Conference on Machine Learning*, pp. 1890–1898, 2014.
- [7] A. Konar and N. D. Sidiropoulos, “Exploring the subgraph density-size trade-off via the lovasz extension,” in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 743–751, 2021.

Memory-Efficient Convex Optimization for Self-Dictionary Nonnegative Matrix Factorization

Tri Nguyen and Xiao Fu

School of Electrical Engineering and Computer Science
Oregon State University
(nguyetr9,xiao.fu)@oregonstate.edu

Ruiyuan Wu

Department of Electronic Engineering
The Chinese University of Hong Kong
rywu@cuhk.edu.hk

Abstract—The separable nonnegative matrix factorization (NMF) model has been adopted in many applications (e.g., topic modeling and community detection), as it ensures identifiability of the latent factors with polynomial time algorithms under reasonable assumptions. Separable NMF is often formulated as a self-dictionary (SD) sparse regression problem, and algorithms under this formulation can be categorized into two classes, namely, greedy methods and all-at-once convex optimization. The latter has been shown to be more resilient to noise, but has serious scalability issues. Particularly, existing convex optimization algorithms for the SD sparse regression problem cost a memory that grows *quadratically* with respect to the data size—making it hardly useful in practice. This work puts forth a Frank-Wolfe algorithm to tackle this challenge. The proposed method has a memory cost that grows *linearly* in the data size—with provable guarantees. To our best knowledge, this is the first linear memory algorithm for this long-existing problem. Synthetic and real data experiments are used to showcase the memory efficiency of the proposed algorithm.

Index Terms—Nonnegative matrix factorization, separability, identifiability, memory-efficient algorithm, Frank-Wolfe

I. INTRODUCTION

Given a matrix $\mathbf{X} \in \mathbb{R}^{M \times N}$, the *nonnegative matrix factorization* (NMF) techniques [1] seek two latent factor matrices $\mathbf{W}$, and $\mathbf{H}$ such that

$$\mathbf{X} \approx \mathbf{W}\mathbf{H}, \quad (1)$$

where $\mathbf{W} \in \mathbb{R}_+^{M \times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K \times N}$, and $K \leq \min(M, N)$ is the model order. It is noted that the latent factors $\mathbf{W}$ and $\mathbf{H}$ are identifiable up to some often inconsequential ambiguities under mild conditions [1], [2]. This property can be used to design latent parameter estimation and unsupervised model learning methods, which are cornerstones of a wide spectrum of applications, e.g., hyperspectral unmixing [3], topic modeling [4] and community detection [5]–[7].

In terms of computation, NMF is an NP-hard problem in the worst case [8]. Nonetheless, a special sub-class of NMF problem is found to be polynomial-time solvable. In particular, if the so-called *separability* condition [9] is satisfied, then the NMF problem can be solved efficiently [10]–[15]. Interestingly, separability often has nice physical interpretations in applications (e.g., the existence of anchor words and pure

pixels in topic modeling [16] and hyperspectral unmixing [3], respectively).

An important perspective to tackle separable NMF is the so-called *self-dictionary multiple measurement vector* (SD-MMV) sparse regression [13], [15], [17]–[19]. There are mainly two types of algorithms for solving the SD-MMV problem. First, a number of greedy pursuit algorithms exist [11]–[13], [16], [20], [21]. The second type uses convex relaxations of SD-MMV; see, e.g., [14], [15], [19], [22]. Since these methods find the solution factors all-at-once (instead of part-by-part as in the greedy methods), they are often more robust to noise.

However, existing convex SD-MMV algorithms often encounters serious memory issues. The reason is that the SD-MMV formulation involves an optimization variable that has a size of $N \times N$, where N often stands for the number of data samples (e.g., the number of pixels in hyperspectral unmixing). Hence, algorithms relying on general-purpose convex optimization tools, e.g., projected gradient [15], [19] and ADMM [14], do not scale well in terms of memory cost. For example, when $N = 100,000$, the algorithms in [14], [15], [19] need 74.5GB memory to instantiate and store the optimization variable—which makes the convex SD-MMV approach hardly useful for big data analytics.

In this work, we develop a memory-efficient Frank-Wolfe (FW) algorithm for convex SD-MMV¹. In the noiseless case, we show that the FW algorithm ensures the identifiability of the latent factors using $O(KN)$ memory (where $K \ll N$ often holds) under a simple SD-MMV formulation. When noise is present, we show that using a carefully designed regularization term can assist attaining the same memory complexity without losing identifiability guarantees. Our result presents the first algorithm that costs a memory linear in N (i.e., the data size) for convex SD-MMV. FW has been well-known for offering memory-efficient solutions to find *unconstrained* low-rank matrix models [24], and our work expands this idea to cover an important constrained low-rank model, i.e., separable NMF. We applied the algorithm onto synthetic data and real-world community analysis data to showcase the effectiveness.

¹This work is a shortened version of a submitted journal article, whose arXiv version is online [23]. Permissions for submitting this conference version have been given by the associate editor and editor-in-chief.

This work is supported in part by the Army Research Office (ARO) under Project ARO W911NF19-1-0407.

II. BACKGROUND AND PROBLEM STATEMENT

Consider the noisy NMF model as follows:

$$\mathbf{X} = \mathbf{W}\mathbf{H} + \mathbf{V}, \quad (2)$$

where $\mathbf{X}$, $\mathbf{W}$ and $\mathbf{H}$ are defined as before in (1), and $\mathbf{V}$ represents a noise term. We further assume that

$$\mathbf{1}^\top \mathbf{H} = \mathbf{1}^\top, \quad \mathbf{H} \geq \mathbf{0}; \quad (3)$$

i.e., the columns of $\mathbf{H}$ live in the probability simplex. In many applications like topic modeling, hyperspectral unmixing, and community detection, this assumption naturally holds [1]. For other application, this assumption can be “enforced” by simple pre-processing [12].

The NMF techniques aim to find the data generating ground-truth latent factors $\mathbf{W}$ and $\mathbf{H}$ from the ambient data $\mathbf{X}$. Note that this is a very challenging task in general—even if there is no noise nonnegative $\mathbf{W}$ and $\mathbf{H}$ that can synthesize $\mathbf{X}$ may not be unique [1]. The separability condition was first introduced in [9] to address the model identification issue:

Assumption 1 (Separability) *There exists an index set $\mathcal{K} \subseteq [N] = \{1, \dots, N\}$ such that $\mathbf{H}(:, \mathcal{K}) = \mathbf{I}_K$.*

Assumption 1 makes uniquely identifying $\mathbf{W}$ and $\mathbf{H}$ from $\mathbf{X}$ (up to a permutation ambiguity) a tractable problem. To see this, under noiseless case, if the index set $\mathcal{K}$ can be identified, then $\mathbf{W} = \mathbf{X}(:, \mathcal{K})$ can be easily recovered. Separability was connected to physically meaningful conditions in various applications—e.g., the anchor word assumption in topic modeling [4], [16], the pure node assumption in community detection [5], [6], [25], the pure pixel assumption in hyperspectral imaging [3], and the specialized annotator condition in crowdsourcing [26].

The set $\mathcal{K}$ can be found via locating columns of $\mathbf{X}$ that can represent other columns of $\mathbf{X}$ as convex combinations (since $\mathbf{x}_\ell = \mathbf{W}\mathbf{h}_\ell = \mathbf{X}(:, \mathcal{K})\mathbf{h}_\ell$ and $\mathbf{h}_\ell$ resides in the probability simplex) [13]. If $\text{rank}(\mathbf{W}) = K$, $\mathbf{X}(:, \mathcal{K})$ represents a *smallest* set of columns that achieves this goal. Therefore, identifying $\mathcal{K}$ can be done by solving a row-sparsity minimization problem [13], [22]:

$$\begin{aligned} & \underset{\mathbf{C}}{\text{minimize}} \quad \|\mathbf{C}\|_{\text{row}-0} \\ & \text{subject to} \quad \mathbf{1}^\top \mathbf{C} = \mathbf{1}^\top, \mathbf{C} \geq \mathbf{0}, \mathbf{X} = \mathbf{X}\mathbf{C}, \end{aligned} \quad (4)$$

where $\|\mathbf{C}\|_{\text{row}-0}$ counts number of nonzero rows in $\mathbf{C}$. Under the noiseless model, an optimal solution of (4) is

$$\mathbf{C}^*(\mathcal{K}, :) = \mathbf{H}, \quad \mathbf{C}^*(\mathcal{K}^c, :) = \mathbf{0}. \quad (5)$$

By inspecting row of $\mathbf{C}^*$, $\mathcal{K}$ is obtained.

Developments of solving (4) are mostly divided into two types. First, recognizing (4) as a combinatorial problem, a number of greedy pursuit approaches that pick the indices in $\mathcal{K}$ one by one were proposed; see, e.g., [11]–[13], [16], [20], [21]. However, when the noise is present, the greedy algorithms sometimes suffer from error propagation. The second type of approaches works in an all-at-once manner. These methods

use convex relaxations as surrogate functions for $\|\mathbf{C}\|_{\text{row}-0}$. For instance, the works in [14], [22] used the mixed-norm $\|\mathbf{C}\|_{\infty,1}$, defined as $\|\mathbf{C}\|_{\infty,1} = \sum_{n=1}^N \|\mathbf{C}(n, :)\|_\infty$ to replace $\|\mathbf{C}\|_{\text{row}-0}$. The working objective was recast as follows:

$$\begin{aligned} & \underset{\mathbf{C}}{\text{minimize}} \quad \frac{1}{2} \|\mathbf{X} - \mathbf{X}\mathbf{C}\|_F^2 + \lambda \|\mathbf{C}\|_{\infty,1} \\ & \text{subject to} \quad \mathbf{1}^\top \mathbf{C} = \mathbf{1}^\top, \mathbf{C} \geq \mathbf{0}. \end{aligned} \quad (6)$$

The formulation (6) is shown in [14] to admit identifiability of $\mathcal{K}$ in the presence of noise. Since convex SD-MMV approaches do not pick indices in $\mathcal{K}$ in a sequential manner as in greedy pursuit, it is able to circumvent the error propagation problem—and enhanced separable NMF performance is often observed [14], [18], [19]. However, the variable $\mathbf{C}$ in the optimization problem (6) has a size of $N \times N$, which poses a memory challenge when N is large.

III. PROPOSED APPROACH

In this section, we propose a FW algorithm based method for convex SD-MMV.

A. Preliminaries of Frank-Wolfe

The FW algorithm is also called conditional gradient, which was proposed in the 1950s [27]. The FW algorithm deals with optimization problems with the following form:

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad f(\boldsymbol{\theta}) \quad \text{subject to} \quad \boldsymbol{\theta} \in \mathcal{C} \quad (7)$$

where $\mathcal{C}$ is a compact convex set and f is a convex continuously differentiable function. FW solves the optimization problem by the following updates:

$$\mathbf{u}^t = \arg \min_{\mathbf{u} \in \mathcal{C}} \nabla f(\boldsymbol{\theta}^t)^\top \mathbf{u}, \quad (8a)$$

$$\boldsymbol{\theta}^{t+1} = (1 - \alpha)\boldsymbol{\theta}^t + \alpha \mathbf{u}^t, \quad (8b)$$

where α is step size and is determined as $\alpha = 2/(2 + t)$. For convex problems, FW converges to the optimal solution with at least a sublinear rate [24].

The FW algorithm’s update rule can sometimes help come up with memory-efficient solutions for various problems—e.g., large-scale low-rank matrix completion where $\mathcal{C}$ is the set of low-rank matrices with small nuclear norms [24]. In this work, we will use the following property of FW. To be specific, if $\mathcal{C} = \{\mathbf{z} \mid \mathbf{z} \geq \mathbf{0}, \mathbf{1}^\top \mathbf{z} = 1\}$, then the constrained optimization in (8a) reduces to a linear program, whose solution is as follows:

$$\mathbf{u}^t = \mathbf{e}_{n^*}, \quad n^* = \arg \min_n [\nabla f(\boldsymbol{\theta}^t)]_n, \quad (9)$$

where $\mathbf{e}_i \in \mathbb{R}^N$ denotes the i th unit vector.

B. FW-based SD-MMV

Our idea is to use an FW algorithm for solving (6). Note that the classic FW algorithm was proposed for differentiable cost functions, yet $\|\mathbf{C}\|_{\infty,1}$ is not smooth. Although FW can work with non-differential objectives (see, e.g., [28]), the algorithmic framework is more convoluted. In this work, we

propose to employ a smoothed approximation for the $\ell_{\infty,1}$ norm from [29]:

$$\Phi_{\mu}(C) := \mu \sum_{n=1}^N \log \left(\frac{1}{N} \sum_{\ell=1}^N \exp(c_{n,\ell}/\mu) \right), \quad \mu > 0.$$

As a result, we consider the following optimization formula

$$\begin{aligned} & \underset{C}{\text{minimize}} && \frac{1}{2} \|\mathbf{X} - \mathbf{X}C\|_F^2 + \lambda \Phi_{\mu}(C) \\ & \text{subject to} && \mathbf{1}^\top C = \mathbf{1}^\top, \quad C \geq 0. \end{aligned} \quad (10)$$

Our first result shows that this smoothed version of (6) still admits identifiability of $\mathcal{K}$, even in the presence of noise:

Theorem 1 (Identifiability) *Assume that Assumption 1 holds, and that $\text{rank}(\mathbf{W}) = K$. Also assume that $\|\mathbf{v}_i\|^2 \leq (\rho/N) \|\mathbf{V}\|_F^2$ for some ρ and $\|\mathbf{v}_i\|_2 \leq \delta$. Then, any optimal solution $\mathbf{C}_{\text{opt}}$ of Problem (10) satisfies:*

$$\|\mathbf{C}_{\text{opt}}(n, :)\|_{\infty} > 1 - \beta, \quad \forall n \in \mathcal{K}, \quad (11a)$$

$$\begin{aligned} \|\mathbf{C}_{\text{opt}}(n, :)\|_{\infty} &\leq 2\rho \frac{N-K}{\lambda N} \|\mathbf{V}\|_F^2 \\ &\quad + \mu N \log(N) + \beta K, \quad \forall n \in \mathcal{K}^c \end{aligned} \quad (11b)$$

where

$$\beta := \frac{\sqrt{4\rho(1-K/N) \|\mathbf{V}\|_F^2 + 2\lambda K} + 2\delta}{\kappa(\mathbf{W})(1-d(\mathbf{H}))},$$

$$\kappa(\mathbf{W}) := \min_{\substack{1 \leq k \leq K, \\ \mathbf{1}^\top \boldsymbol{\theta} = 1, \boldsymbol{\theta} \geq 0}} \|\mathbf{w}_k - \mathbf{W}(:, -k)\boldsymbol{\theta}\|_2, \quad d(\mathbf{H}) := \max_{\substack{n \in \mathcal{K}, \\ \ell \in \mathcal{K}^c}} h_{n,\ell}$$

Due to space limitations, the proof is relegated to a long version of this work [23]. The result in (11) provides insights into how the identification of $\mathcal{K}$ is affected by the key problem instance parameters (e.g., $\mathbf{H}, \mathbf{W}$) and hyperparameters (e.g., λ, μ). For example, when the noise level is high, one hopes to use a large λ to “suppress” the effect of the increased $\|\mathbf{V}\|_F^2$ in (11b). In the meanwhile, due to the presence of $2\lambda K$ in the expression of β , one does not hope to increase λ too much as ideally one hopes $\beta \approx 0$. Such a nuanced balance is indeed often reflected in the parameter tuning process in practice.

Note that identifiability is a characterization of the optimal solution of (10), regardless of which algorithm to take. As we mentioned, many convex optimization algorithms (e.g., projected gradient) would require $O(N^2)$ for instantiating and storing $\mathbf{C}$, which is not a viable solution when N is large. We propose to employ the FW algorithm, as detailed in Algorithm 1. We show that, under reasonable assumptions, the algorithm can identify $\mathcal{K}$ using only $O(NK)$ memory:

Theorem 2 (Memory Efficiency) *Assume that Assumption 1 holds and that $\text{rank}(\mathbf{W}) = K$. Also assume that (i) in iteration t , $\mathbf{C}^t$ satisfies $\text{supp}(\mathbf{c}_{\ell}^t) \subseteq \mathcal{K}$ for all $\ell \in [N]$ and (ii) there exists at least an $n^* \in \mathcal{K}$ such that $\mathbf{C}^t(n^*, :)$ is not a constant row vector. If the noise bound satisfies*

$$\delta \leq \sqrt{\gamma^2 + v} - \gamma, \quad (12)$$

Algorithm 1 MERIT’s algorithm

```

1: procedure MERIT( $\mathbf{X}, \lambda, t_{\text{init}}, \mathbf{C}^{\text{init}}$ )
2:   if ‘Warm Start’ = True then
3:      $\mathbf{C} = \mathbf{C}^{\text{init}}$ 
4:   else
5:      $\mathbf{C} = \mathbf{0}$ 
6:      $t_{\text{init}} = 0$ 
7:   end if
8:   for  $t \leftarrow t_{\text{init}}, t_{\text{init}} + 1, t_{\text{init}} + 2, \dots$  do
9:      $\alpha \leftarrow 2/(t + 2)$ 
10:     $\mathbf{r} \leftarrow \mathbf{0}$ 
11:    for  $n \leftarrow 1, \dots, N$  do
12:       $r_n \leftarrow \sum_{i=1}^N \exp(\mathbf{C}(n, i)/\mu)$ 
13:    end for
14:    for  $\ell \leftarrow 1, \dots, N$  do
15:       $\mathbf{p}_{\ell} \leftarrow \mathbf{X}^\top(\mathbf{X}\mathbf{c}_{\ell} - \mathbf{x}_{\ell})$ 
16:       $\mathbf{y}_{\ell} \leftarrow \exp(\mathbf{c}_{\ell}/\mu) \cdot \mathbf{r}$   $\triangleright \cdot /$  denotes an
element-wise division between 2 vectors
17:       $\mathbf{g}_{\ell} \leftarrow \mathbf{p}_{\ell} + \lambda \mathbf{y}_{\ell}$ 
18:       $\mathbf{u}_{\ell} \leftarrow \mathbf{e}_i$  such that  $i = \arg \min_j [\mathbf{g}_{\ell}]_j$ 
19:       $\mathbf{c}_{\ell} \leftarrow (1 - \alpha)\mathbf{c}_{\ell} + \alpha \mathbf{u}_{\ell}$ 
20:    end for
21:  end for
22: end procedure

```

where

$$\begin{aligned} \gamma &= \max_{1 \leq k \leq K} \|\mathbf{w}_k\|_2, \\ v &= \frac{1}{8} \left[2\lambda(1/N - \exp(-\psi/\mu)) \right. \\ &\quad \left. - (d'(\mathbf{H})^2 + 2d'(\mathbf{H}) + 5)\lambda_{\max}(\mathbf{W}^\top \mathbf{W}) \right], \end{aligned}$$

in which,

$$\begin{aligned} d'(\mathbf{H}) &= \sqrt{\max(2(d(\mathbf{H}) - 1/2)^2 + 1/2, 2(1/2 - 1/K)^2 + 1/2)}, \\ \psi &:= \min_{\substack{1 \leq i, j \leq N, \\ n \in \mathcal{K}, c_{n,i}^t \neq c_{n,j}^t}} |c_{n,i}^t - c_{n,j}^t|, \end{aligned}$$

then, to attain $\mathbf{c}_{\ell}^{t+1}$ from $\mathbf{c}_{\ell}^t$, FW will update the j_{ℓ} th element of $\mathbf{c}_{\ell}^t$ such that $j_{\ell} \in \mathcal{K}$ for all $\ell \in [N]$. This means that only $O(KN)$ memory is used in all the iterations.

Theorem 2 asserts that if $\mathbf{C}$ is initialized properly such that $\text{supp}(\mathbf{C}) \in \mathcal{K}$ (which can be attained by running an iteration of existing greedy approaches like [12]), then the updates of $\mathbf{C}$ will never go out of $\mathcal{K}$ under reasonable conditions. Note that ψ is defined using $\mathbf{C}^t$ for all t , thereby imposes a seemingly stringent condition. Fortunately, if the initialization of $\mathbf{C}$ satisfies a certain regularity condition, $\psi > 0$ is always guaranteed when the number of iteration T is finite. Specifically, we have the following proposition:

Proposition 1 (Initialization Condition) *Let $\mathbf{C}^{\text{init}}$ be a feasible initial solution. Define*

$$D_{ij}^n(\mathbf{C}) = \frac{t_{\text{init}}(t_{\text{init}} + 1)}{2} |c_{n,i} - c_{n,j}|$$

for some $t_{\text{init}} \in \mathbb{N}, t_{\text{init}} \geq 1$. Suppose that for some $n^* \in \mathcal{K}$, there exists a pair of i^*, j^* such that the following regularity condition is satisfied:

$$D_{i^*, j^*}^{n^*}(\mathbf{C}^{\text{init}}) \notin \mathbb{N}. \quad (13)$$

Running FW with $\mathbf{C}^t = \mathbf{C}^{\text{init}}$ starting with $t = t_{\text{init}}$ for T iterations, the produced solution sequence $\{\mathbf{C}\}^t$ for $t \geq t_{\text{init}} + 1$ satisfies $\mathbf{C}^t(n^*, :)$ is not a constant row vector and

$$\min_{\substack{n \in \mathcal{K}, i, j \\ c_{n,i}^t \neq c_{n,j}^t}} |c_{n,i}^t - c_{n,j}^t| \geq \frac{2\xi}{T(T+1)}$$

where $\xi := \min_{\substack{n \in \mathcal{K}, i, j, z \in \mathbb{N} \\ D_{i,j}^n(\mathbf{C}^{\text{init}}) \neq z}} |D_{i,j}^n(\mathbf{C}^{\text{init}}) - z|$.

Note that regularity condition has zero measure in the continuous space, and thus is believed to be hard to violate in practice. The proofs of Theorem 2 and Proposition 1 can both be found in [23].

IV. EXPERIMENTS

A. Synthetic data

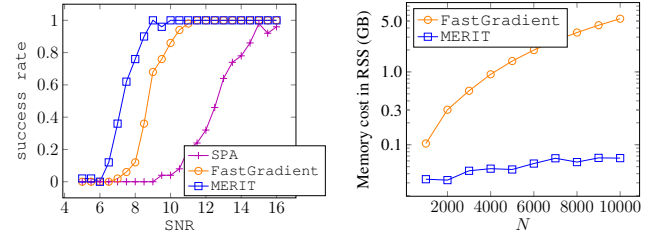
1) *Baselines*: We use SPA [12], [13], [20], a well-known greedy pursuit method for SD-MMV, as a baseline. In addition, FastGradient [19] is included as it is the state-of-the-art convex SD-MMV solver.

2) *Data Generation*: We generate $\mathbf{X}$ as follows: The elements of $\mathbf{W} \in \mathbb{R}^{M \times K}$ are drawn from a uniform distribution $\mathcal{U}(0, 1)$. The first K columns in $\mathbf{H}$ are set to the identity matrix (so that $\mathcal{K} = [K]$), and the remaining $N - K$ columns are drawn from the Dirichlet distribution with $\mathbf{1}$ being the model parameter vector. The noise $\mathbf{V}$ is drawn from the zero-mean Gaussian distribution with a tunable variance σ^2 . The signal-to-noise ratio (SNR) is defined as $\text{SNR} = 10 \log_{10}(\sum_{\ell=1}^N \|\mathbf{W}\mathbf{h}_\ell\|_2^2 / (MN\sigma^2))$ dB. After performing a random shuffle on columns of $\mathbf{H}$, the data $\mathbf{X}$ is obtained as $\mathbf{X} = \mathbf{W}\mathbf{H} + \mathbf{V}$.

3) *Metric*: Given K , all methods are asked to produce an index set $\hat{\mathcal{K}}$ that is an estimate of $\mathcal{K}$. For MERIT, $\hat{K}$ is chosen by picking up the first K indices corresponding to the largest $\|\mathbf{C}(n, :)\|_\infty$'s. For both FastGradient and MERIT, we use the heuristic in [19] to select the regularization coefficient λ . The performance is measured by the success rate defined as $\text{success rate} = \Pr(\hat{\mathcal{K}} = \mathcal{K})$.

4) *Results*: Fig. 1(a) shows the performance of the algorithms under various SNRs. One can see that SPA is more sensitive to noise compared to the convex self-dictionary approaches. At SNR = 10dB, SPA's success rate is still lower than 0.1. However, FastGradient and MERIT attain success rates that are higher than 0.6 and 0.9, respectively.

Fig. 1(b) shows the memory cost of FastGradient and MERIT under various N 's with $(M, K) = (50, 40)$. The memory cost is measured by the so-called maximum resident set size (RSS), recorded by a Linux command named



(a) success rate under different SNRs; $N = 200$. (b) Memory consumption under different N 's; SNR = 10dB.

Fig. 1: Performance of MERIT compared to baselines; $M = 50, K = 40$.

time². RSS measures the amount of memory allocated in RAM for a process, which is mostly aligned with total memory requirement for that process. As shown in Fig. 1(b), FastGradient requires almost 5GB when $N = 10,000$, while MERIT's memory cost is under 0.1GB.

B. Community detection

In this subsection, we evaluate the algorithms on a community detection problem. Under the well-known *mixed membership stochastic blockmodels* (MMSB), estimating the node membership amounts to a noisy matrix factorization [6], [7], [25]. In particular, $\mathbf{X} \approx \mathbf{W}\mathbf{H}$ where rows of $\mathbf{X} \in \mathbb{R}^{F \times N}$ are F leading eigenvectors of the adjacency matrix. The separability condition holds if there is at least one pure node for each community, where a “pure node” is a node that only belongs to a single community.

1) *Baselines*: We compare MERIT with three baselines, i.e., two greedy pursuit methods that are tailored for the community detection problem (namely, GeoNMF [6], SPOC [25]) and FastGradient.

2) *Datasets*: We use two datasets which are the Data Base systems and Logic Programming (DBLP) data and the Microsoft Academic Graph (MAG) data. Both datasets contains community label. Particularly, in DBLP, a group of conferences forms a community, while in MAG, it is the “field of study”; see details in [6]. In pre-processing step, we keep the most significant nodes such that 99% of the energy (in terms of squared Euclidean norm of the corresponding row in the adjacency matrix) are retained. Table I shows detailed statistics of the preprocessed data.

3) *Metric*: We measure the performance using the averaged Spearman's rank correlation (SRC) which is commonly used in community detection papers [6], [7], [25]. SRC ranges from -1 to 1 , and a higher value corresponds to a better estimation of the mixed membership of the nodes.

4) *Results*: Table I presents the SRC performance of different methods over seven different datasets. One can see that the convex optimization-based approach produces competitive results compared to the greedy class. In particular, MERIT are in top 2 highest scores in 5 out of 7 cases.

²<https://man7.org/linux/man-pages/man1/time.1.html>

TABLE I: Statistics and SRC performance on DBLP and MAG.

Dataset	Statistics		SRC			
	# nodes	# communities	GeoNMF	SPOC	FastGradient	MERIT
DBLP1	6437	6	0.2974	0.2996	0.3145	0.2937
DBLP2	3248	3	0.2948	0.2126	0.3237	0.3257
DBLP3	3129	3	0.2629	0.2972	0.1933	0.2763
DBLP4	1032	3	0.2661	0.3479	0.1601	0.3559
DBLP5	3509	4	0.1977	0.1720	0.0912	0.1983
MAG1	1447	3	0.1349	0.1173	0.0441	0.1149
MAG2	4974	3	0.1451	0.1531	0.2426	0.2414

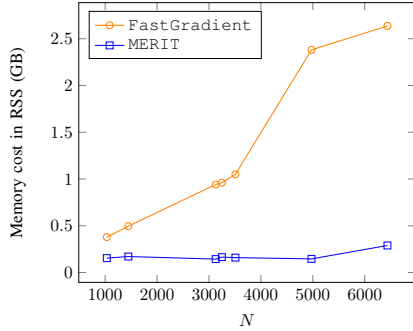


Fig. 2: Memory consumed by FastGradient and MERIT.

Similar to the synthetic experiment, Fig. 2 shows memory consumed by FastGradient and MERIT under different problem sizes. We see that MERIT's memory never exceeds 0.5GB while FastGradient requires more than 2.5GB when $N = 6000$.

V. CONCLUSION

Nonnegative matrix factorization with separability condition provides a tractable framework for identification. Compare to greedy algorithm class, the convex relaxation SD-MMV formulation allows a more noise tolerant approach, but also presents issue on memory consumption of $O(N^2)$ when dealing with big data. In this work, we tackle the memory issue while still offering identifiability. We show that using FW to optimize a smooth objective function can guarantee exact recovery of $\mathcal{K}$ while only using a $O(KN)$ memory under noise case. The experimental result on synthetic settings aligns with our analysis.

REFERENCES

- [1] N. Gillis, *Nonnegative Matrix Factorization*. SIAM, 2020.
- [2] X. Fu, K. Huang, and N. D. Sidiropoulos, "On identifiability of nonnegative matrix factorization," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 328–332, 2018.
- [3] W.-K. Ma, J. M. Bioucas-Dias, T. H. Chan, N. Gillis, P. Gader, A. J. Plaza, A. Ambikapathi, and C. Y. Chi, "A signal processing perspective on hyperspectral unmixing: Insights from remote sensing," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 67–81, 2014.
- [4] K. Huang, X. Fu, and N. D. Sidiropoulos, "Anchor-free correlated topic modeling: Identifiability and algorithm," in *Advances in Neural Information Processing Systems*, 2016, pp. 1786–1794.
- [5] K. Huang and X. Fu, "Detecting overlapping and correlated communities without pure nodes: Identifiability and Algorithm," in *International Conference on Machine Learning*, 09–15 Jun 2019, pp. 2859–2868.
- [6] X. Mao, P. Sarkar, and D. Chakrabarti, "On mixed memberships and symmetric nonnegative matrix factorizations," in *International Conference on Machine Learning*, 2017, pp. 2324–2333.
- [7] S. Ibrahim and X. Fu, "Mixed membership graph clustering via systematic edge query," *IEEE Trans. Signal Process.*, vol. 69, pp. 5189–5205, 2021.
- [8] S. A. Vavasis, "On the complexity of nonnegative matrix factorization," *SIAM Journal on Optimization*, vol. 20, no. 3, pp. 1364–1377, 2009.
- [9] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?" in *Advances in Neural Information Processing Systems*, vol. 16, 2003, pp. 1141–1148.
- [10] S. Arora, R. Ge, and A. Moitra, "Learning topic models—going beyond SVD," in *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, 2012, pp. 1–10.
- [11] T.-H. Chan, W.-K. Ma, A. Ambikapathi, and C.-Y. Chi, "A simplex volume maximization framework for hyperspectral endmember extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 11, pp. 4177–4193, 2011.
- [12] N. Gillis and S. A. Vavasis, "Fast and robust recursive algorithms for separable nonnegative matrix factorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 698–714, 2014.
- [13] X. Fu, W.-K. Ma, T.-H. Chan, and J. M. Bioucas-Dias, "Self-dictionary sparse regression for hyperspectral unmixing: Greedy pursuit and pure pixel search are related," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 6, pp. 1128–1141, 2015.
- [14] X. Fu and W.-K. Ma, "Robustness analysis of structured matrix factorization via self-dictionary mixed-norm optimization," *IEEE Signal Process. Lett.*, vol. 23, no. 1, pp. 60–64, 2015.
- [15] B. Recht, C. Re, J. Tropp, and V. Bittorf, "Factoring nonnegative matrices with linear programs," in *Advances in Neural Information Processing Systems*, 2012, pp. 1214–1222.
- [16] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu, "A practical algorithm for topic modeling with provable guarantees," in *International Conference on Machine Learning*, 2013, pp. 280–288.
- [17] X. Fu, W.-K. Ma, T.-H. Chan, J. M. Bioucas-Dias, and M.-D. Iordache, "Greedy algorithms for pure pixels identification in hyperspectral unmixing: A multiple-measurement vector viewpoint," in *21st European Signal Processing Conference*, 2013, pp. 1–5.
- [18] N. Gillis, "Robustness analysis of hottopix, a linear programming model for factoring nonnegative matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 34, no. 3, pp. 1189–1212, 2013.
- [19] N. Gillis and R. Luce, "A fast gradient method for nonnegative sparse regression with self-dictionary," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 24–37, 2018.
- [20] U. Araújo, B. Saldanha, R. Galvão, T. Yoneyama, H. Chame, and V. Visani, "The successive projections algorithm for variable selection in spectroscopic multicomponent analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 57, no. 2, pp. 65–73, 2001.
- [21] J. M. Nascimento and J. M. Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 4, pp. 898–910, 2005.
- [22] E. Esser, M. Moller, S. Osher, G. Sapiro, and J. Xin, "A convex model for nonnegative matrix factorization and dimensionality reduction on physical space," *IEEE Trans. Image Process.*, vol. 21, no. 7, pp. 3239–3252, 2012.
- [23] T. Nguyen, X. Fu, and R. Wu, "Memory-efficient convex optimization for self-dictionary separable nonnegative matrix factorization: A Frank-Wolfe approach," *arXiv preprint arXiv:2109.11135*, 2021.
- [24] M. Jaggi, "Revisiting Frank-Wolfe: Projection-free sparse convex optimization," in *International Conference on Machine Learning*, 2013, pp. 427–435.
- [25] M. Panov, K. Slavov, and R. Ushakov, "Consistent estimation of mixed memberships with successive projections," *International Workshop on Complex Networks and their Applications*, pp. 53–64, 2017.
- [26] S. Ibrahim, X. Fu, N. Kargas, and K. Huang, "Crowdsourcing via pairwise co-occurrences: Identifiability and Algorithms," in *Advances in Neural Information Processing Systems*, 2019, pp. 7847–7857.
- [27] M. Frank, P. Wolfe *et al.*, "An algorithm for quadratic programming," *Naval Research Logistics Quarterly*, vol. 3, no. 1-2, pp. 95–110, 1956.
- [28] K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh, "Projection efficient subgradient method and optimal nonsmooth Frank-Wolfe method," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 211–12 224, 2020.
- [29] Y. Nesterov, "Smooth minimization of non-smooth functions," *Mathematical programming*, vol. 103, no. 1, pp. 127–152, 2005.

Computational Demography and Health

Zack W. Almquist ^a

Courtney Allen ^b

Ihsan Kahveci ^b

^aDepartment of Sociology, Department of Statistics, Center for Studies in Demography and Ecology, University of Washington, Seattle, WA, USA

^bDepartment of Sociology, University of Washington, Seattle, WA, USA

ARTICLE HISTORY

Compiled January 8, 2023

ABSTRACT

Recent developments in computing, data entry and generation, and analytic tools have changed the landscape of modern demography and health research. These changes have come to be known as computational demography, big data, and precision health in the field. This is an emerging interdisciplinary research comprised of social scientists, physical scientists, engineers, data scientists, and disease experts. This work has changed how we use administrative data, conduct surveys, and allow for complex behavioral studies via big data (electronic trace data from mobile phones, apps, et cetera). This chapter reviews the new data sources, methods, and applications of this emerging field.

KEYWORDS

computational demography, computational health, big data, data science, synthetic data, social media data, online data, migration

Introduction

Recent developments in computing, data entry, and analytic tools have changed the landscape of modern demography and health research. These changes have come to be known as computational demography, big data, and precision health in their respective fields. This is an emerging interdisciplinary research area comprised of social scientists, physical scientists, engineers, data scientists, and disease experts. Major developments include automated data collection such as social media or wearable health trackers; disease tracking such as Google Flu Trends or the Facebook COVID-19 Symptoms Survey; novel methods in machine learning and statistics; complex systems analysis; and so forth. This has led to prominent growth in demographic measurement, public health research, and even areas such as precision medicine (e.g., personal medicine), self-tracking, and public health initiatives. These systems have recently been used in large-scale public health outreach through COVID-19 reporting and contact tracing. Altogether, this collection of methods, data, and applications is changing how we understand demography and health.

In this book Chapter, we are going to review major data developments, applications

of machine learning, statistics and simulation models to demographic and health data, and conclude with a summary and note about COVID-19. This Chapter is not a comprehensive review but a general gateway to computational demography and health — an emerging field of study that has the potential to impact much of the social sciences.

Big data

Big data has a rich and complicated history in the Social Sciences and has come to refer to a variety of data sources, including but not limited to administrative data, behavioral trace data from on and offline sources, synthetic data, and geolocation data. Big data and data science are increasingly becoming a key tool in demographic and public health research. In the following section, we will review a brief background and suitability of big data and conclude with some notable applications and research findings.

Suitability of big data for population processes and health research

The term Big Data once described data that was too large to be stored by conventional means but has since evolved to have a more fluid definition that tries to capture the “bigness” of the information used by scientists across many disciplines (Mooney, Westreich, & El-Sayed, 2015). This data is characterized by the *three V’s*: high variety, volume, and velocity. Big data are often comprised of information from multiple sources with varying original purposes (high-variety), have more cases than previously used data (high-volume), and are generated in real-time (high-velocity) (Chiolero & Buckeridge, 2020; Mooney et al., 2015). The data tend to be organic in nature, accidentally produced, for purposes other than research (Bohon, 2018) and reflect the increasing digitization of health and other areas of life through evolving technologies (Chiolero & Buckeridge, 2020). Big data also describes not only the bigness of the data but data that is so large, varied, or novel that it requires new technologies and methods to understand it (D. Lazer & Radford, 2017). Take, for example, modern computer systems like Google or Facebook, which can range in petabytes per day. This is in comparison to “classic” big data such as US Census data, tax records, and other administrative data (Ruggles, 2014). Some have described a *fourth V* in big data: veracity. This can be thought of similar to face validity and allows to ask the basic question can this data answer the questions we are posing to it.

Big data can be noisy, with poor quality, due to its rapid production (Chiolero & Buckeridge, 2020), and may be overwhelming to manage (Cesare, Lee, McCormick, Spiro, & Zagheni, 2018; Raghupathi & Raghupathi, 2014). Because big data sources are generally byproducts of digitized interactions not designed for research, they possess no defined universe and, therefore, no representativeness of any population (Ruggles, 2014). This creates a shallowness that has come to typify these sources (Bohon, 2018) as they cannot be considered neutral, objective, or complete (Delpierre & Kelly-Irving, 2018) (also cite Ruggles). Many big data sources have clear population biases, such as social media platforms heavily used by women or younger populations (Ruths & Pfeffer, 2014). Using data not designed for statistical inference or representative of any population presents a challenge when considering big data’s use for understanding population health.

Researchers must take steps to understand the selection bias in nonrepresentative

sources that, if not calibrated, would lead to biased and invalid conclusions (Yildiz, Munson, Vitali, Tinati, & Holland, 2017; Zagheni & Weber, 2015). These sources have biases that go beyond nonrepresentative population bias. For example, standard statistical approaches cannot be used due to the network and relational aspect of data when using social media sources (Hunter et al., 2018). In addition, big data can be gleaned from platforms where individual behaviors are affected by proprietary algorithms, of which researchers have no way of correcting for these influences. Furthermore, embedded researchers - those connected to such platforms - may analyze this data while correcting for algorithm influence but not be able to be transparent about their methods (Hunter et al., 2018; Ruths & Pfeffer, 2014). Researchers encounter barriers when using big data that require advanced or novel methods to correct bias while wielding such large data sets. These barriers are overcome with specific and advanced skills largely missing from many graduate training programs in the social sciences and demography (Cesare et al., 2018). This lack of training creates a slower production of robust, cutting-edge methodologies and slows down the the popularity and growth of big data use in demography and other social sciences. This area of research is quickly evolving, and this description is not exhaustive of all of the challenges and biases that exist when using big data. Many of these issues have existed in the social sciences, statistics, epidemiology, and other domains. Utilizing big data, collaboration, and knowledge sharing will push the field further.

Big data often include personal information gleaned from the many digitized areas of people’s everyday lives, creating ethical implications in its use. Internet data is an increasingly used source of big data that has great potential to provide novel insights into demographic behaviors (Zagheni & Weber, 2015). Often referred to as trace data, this data are derived from digitized interactions from social media platforms or other digital activity and may be collected unknowingly from platform users, may contain sensitive information, and their ownership remains unclear as data are sold or shared with third parties (Cesare et al., 2018). Individuals’ privacy is one element of this data that remains unclear. The right to privacy would ensure individuals’ freedom from intrusion, though many digital interactions lie within an ambiguous public-private continuum (Hunter et al., 2018). Some data points may be collected in a manner out of the control of any researcher or individual to collect or provide consent (Hunter et al., 2018). Confidentiality and anonymity are also vital pieces of ethical research. While many researchers have taken steps to remove conventional identifying information, digital data are complex, and even anonymized data may expose individuals through social network ties (Hunter et al., 2018). Data that reflect individual actions, behaviors, or other personal attributes, may not offer safeguard options to those who would not wish to contribute their data for larger research aims. This has created tensions between data protection for individuals in a community and maximizing researchers’ access to big data (Salerno, Knoppers, Lee, Hlaing, & Goodman, 2017).

Comparability of big data to other sources

Google’s Flu Trends (GFT) is the first major big data implementation for predicting health outcomes. It was developed by Ginsberg et al. (2009) to predict CDC’s seasonal estimates using search queries. However, in 2013 Google stopped the service, as it infamously made the headlines after falsely predicting flu rates almost double the CDC estimates. In their work, D. Lazer, Kennedy, King, and Vespignani (2014) explain the potential sources of biases embedded in the model. In addition, they show that if GFT

systematically calibrated its predictions with the CDC estimates, it would result in much more accurate estimates. Following D. Lazer and Radford’s (2017) terminology, we underlined the various ways of “methodological integration” of online data with ground truth data in this chapter.

The proponents of the Big Data Revolution often claim that as the size of data increases, the quality of information gained will also increase. However, one needs to remember that big data is not collected for research purposes. It contains measurement errors and selection bias (Zagheni & Weber, 2015). Recent work has questioned whether big data methods are superior to classic survey sampling for general population representativeness. (Bradley et al., 2021). In addition, the commercial goals behind the services, such as Google’s ad services, significantly alter the data generation process over time, causing problems for replicability (D. Lazer et al., 2014). Two methodological approaches are suggested for reducing the selection bias: (i) First, where ground data exist (such as administrative data), online data can be calibrated to match demographic variables. (ii) Second, where no ground data exist, difference-in-differences can be used to reduce bias when estimating trends from online data (Zagheni & Weber, 2015). An example of this approach includes two studies using Facebook’s advertising platform to predict regional public health outcomes. Gittelman et al. (2015) combine the Behavioral Risk Factor Surveillance System (BRFSS), vital statistics, and Facebook data to predict county-level health outcomes and health-related behaviors. The second, Mejova, Weber, and Fernandez-Luque (2018) employ a quasi-instrumental design to associate disease-specific marker user interests with state-level health conditions. Both results appear promising for employing online data sources for public health monitoring.

Recent statistical developments such as Multilevel Regression with Poststratification (MRP) provide a method to mitigate selection bias in analyzing large-scale online surveys (Cesare et al., 2018; Wang, Rothschild, Goel, & Gelman, 2015). This method is highly related to Bayesian hierarchical models Loux, Nelson, Arnold, Shacham, and Schootman (2019), which can also be used to adjust the data from a biased sample. This includes important use cases such as respondents recruited via a Facebook survey. For example, this method has been applied to the estimation of physical activity in Brazil on a large-scale cross-sectional national survey where they demonstrated that MRP consistently outperforms single-level regression models (Chiolero & Buckeridge, 2020).

Another approach is computational methods such as machine learning (ML) or artificial intelligence (AI) tools that have been used to reduce bias when predicting health outcomes. Modern methods of regularization (e.g., LASSO or Ridge regression) have proved useful for model selection in public health contexts (Young, Torrone, Urata, & Aral, 2018). Work focusing on real-time monitoring using social media data has leveraged ML methods. For example, Joshi et al. (2020) utilized Twitter data for the early detection of thunderstorm asthma. Using a novel monitoring algorithm based on “weighted moving average of time-between-event values.” One of the more important uses of these procedures is that it allows for time-critical application. *Time-criticality* means that the monitoring algorithm cannot wait to alert until it counts all the events in a time interval because, at that time, the outbreak will already have occurred. The frequency of events should be taken into consideration too.

Application of big data to demography and health

Facebook

Facebook data and access to its users has become an important resource for demographic and health data worldwide. Recently, Facebook in conjunction with academic researchers, has been providing micro-level data (e.g., US Zip-code level data) on people’s experience with COVID-19 and self-reported symptoms (Astley et al., 2021; Reinhart et al., 2021). This work has appeared in a special issue in the *Proceedings of the National Academy of Sciences* and has already produced dozens of peer-reviewed articles. Other important contributions include a US county-by-county social connectedness index (Bailey, Cao, Kuchler, Stroebel, & Wong, 2018) and experimental evidence on social contagions via social media posts and their effect on happiness (Kramer, Guillory, & Hancock, 2014).

Twitter

Twitter data, including geolocated tweets, has been used to measure place and space, as well as provide textual content to understand opinions, trending topics, and interaction patterns. Researchers have utilized Twitter to measure migration using geolocated tweets (Alexander, Polimis, & Zagheni, 2019), estimate population totals using face and age recognition on profile photos (Yildiz et al., 2017), inform on trending health topics (Aramburu, Berlanga, & Lanza, 2020; Batbaatar & Ryu, 2019), understand housing quality (Chisholm & O’Sullivan, 2017), explore substance abuse (Stevens et al., 2020), identify food poisoning outbreaks and other acute disease events (Harris et al., 2017; Joshi et al., 2020), and even examine social distancing trends during the COVID-19 pandemic (Younis et al., 2020). Other data sources have even included smart meters to understand electricity use (Newing, Anderson, Bahaj, & James, 2016), Instagram photos that give insight into codeine misuse (Cherian, Westbrook, Ramo, & Sarkar, 2018), and Google Street View to examine the built environment (Keralis et al., 2020).

Geolocation data

One of the most important trace data is that geolocated data. It can come in many forms, for example, from cell phones, IP location information, and mobile apps. Major marketing vendors have built large-scale data sets such as the Safegraph (Andersen, 2020).

Of this data, maybe the most important is geolocation data from mobile phones. Mobile phone data has been used to understand the efficacy of lock-down orders during an epidemic (Peak et al., 2018), understanding mobility patterns (J. E. Blumenstock, 2012; Deville et al., 2014; Palmer et al., 2012), predicting poverty and wealth (J. Blumenstock, Cadamuro, & On, 2015), and estimating de-facto versus permanent populations (Morrison et al., 2020). During the COVID-19 Pandemic, anonymized Google Maps data was used to understand the impact of mobility restrictions on excess deaths (Basellini et al., 2021) and hospital admissions (Davies et al., 2021).

In demography, geolocation data has been particularly important for answering migration problems, especially short-range movements. For example, people displaced due to disaster or moving between neighborhoods.

Geolocation data and migration

Demographers and population health researchers have long been interested in large-scale population movement and their effects on regional populations (Cas, Frankenberg, Suriastini, & Thomas, 2014). This topic has only increased in importance as climate change and rising population densities increased the frequency of natural disasters impacting human settlements (Bouwer, 2011).

Researchers have used mobility data from cell phones and other electronic devices to characterize both short-term and long-term displacement after a disaster (Gething & Tatem, 2011); however, this technique has often been limited by an inability to pair mobility data with key additional information. In particular, many potential hypotheses could be explored if these methods also gave information about who is displaced, the household structure of displaced persons and the nature of locations to which displaced persons migrate (e.g., identification as a home of relatives) (Cas et al., 2014).

Migration decisions are complex and depend on not only the severity of a given disaster but also the demographic composition of those in the area (Cas et al., 2014). A displacement event can last days to years, with many choosing to relocate permanently. Prior work has compared mobility strategies among individuals living in communities that sustained different degrees of damage due to the 2004 Indian Ocean Tsunami (Cas et al., 2014).

Natural Disasters and Displaced Populations

There has been much work around population movement, big data, and computational methods. For example, post-disaster population movements have traditionally been very difficult to measure (e.g., early evacuation). Both voluntary and involuntary displacement can affect demographic change in a local area (Cas et al., 2014). These demographic changes can have material consequences for affected areas (Frankenberg, Sikoki, Sumantri, Suriastini, & Thomas, 2013). Recent work includes Facebook’s Gender-Stratified Displacement Map, which aims to quantify the magnitude of displacement following disasters, describe where the displaced population has migrated and enable the study of these population trends by gender (Maas, 2019). Facebook’s Displacement Map dataset estimates how many people were displaced by a given disaster and where those people are in the period following it at an aggregate level. Specifically, the models identify a person as displaced if their typical nighttime location patterns are disrupted after the event compared to that person’s pre-disaster patterns. These patterns are obtained from a user’s location history — an optional setting on the Facebook app users can enable that provides precise locations (Maas, 2019). The individual data is aggregated into a city-level transition matrix showing how many people are displaced from one city to another for all source cities in the disaster-affected region for each day for a period following the event (Maas, 2019). Facebook’s Data for Good program (Maas, 2019) makes available displacement maps for use by humanitarian partners, and this project is a larger part of that effort.

Recent work in the “big data” area includes following longer-term displacement and evacuation has often used administrative data, such as the Internal Revenue Service (IRS) county migration data (Fussell, Curtis, & DeWaard, 2014) or, for shorter-term estimates, the Federal Reserve Bank of New York/Equifax Consumer Credit Panel (CCP) (DeWaard, Johnson, & Whitaker, 2020). The IRS data has been used to look at the 3–4 year recovery period after Hurricane Katrina for return migration (Fussell et al., 2014), and the CCP data for continued estimation of displacement due to Hurricane

Maria two years after (DeWaard et al., 2020). Data of this nature tend to be limited either in time (e.g., the IRS data is only available yearly) or income (e.g., CCP requires the user has a credit card).

Twitter, a popular social media application, has been used to estimate the evacuation and displacement of people following Hurricane Maria in Puerto Rico. For example, Martín, Cutter, Li, Emrich, and Mitchell (2020) demonstrated that geotagged tweets provide a representative sample of individuals in the age range of 18–54 years and complemented data collected through a survey with a sample bias toward older respondents. Following Hurricane Maria, it was estimated that in the Twitter sample, 8.3% of residents relocated, and nearly 4% continued to be displaced nine months later (Martín et al., 2020). Other work has used Twitter to characterize aspects of displacement such as destination, differential impact, and local severity (Yum, 2020).

Facebook has been used in explorations of demographic topics such as immigrant assimilation (Stewart, Flores, Riffe, Weber, & Zagheni, 2019), world fertility (Ribeiro, Benevenuto, & Zagheni, 2020), and world migration stocks (Zagheni, Weber, & Gum-madi, 2017). Recently, Facebook’s Marketing API has been used to estimate displacement and evacuation from Hurricane Maria in Puerto Rico (Alexander, Polimis, & Zagheni, 2020). Further, the Facebook ads system has been used to recruit survey participants for mobility research Blondel, Decuyper, and Krings (2015). Facebook’s advertising platform has been shown to provide a good sampling frame for providing online surveys and helps with challenges of timeliness, coverage, and cost-effectiveness (Grow et al., 2020). Facebook users are a good cross-section of the overall population with internet access; Grow et al. (2020) advises stratifying by demographic characteristics, which are known to correlate with the outcome of interest and to post-stratify in the final analysis. There is growing literature on how to adjust surveys administered on the Facebook ads (Zagheni et al., 2017) platform or survey system (Feehan & Cobb, 2019) for the estimation of both online and offline populations. Finally, recent work has demonstrated that it is possible to survey evacuation behavior post-disaster via Facebook’s survey infrastructure (Maas, Almquist, Giraudy, & Schneider, 2020).

Activity-based data and health research

Physical trace data and movements like the quantified self-movement (Lee, 2014) have begun to provide real-time analytics on health-related activities (e.g., running or biking) as well as constant health monitoring such as heart rate or blood pressure (Lupton, 2016). Recent work using activity-based data has demonstrated that group size can affect exercise quality (Zeng, Almquist, & Spiro, 2017, 2018, 2019). Other work has been done to show how you can get population estimates of health outcomes from these activity-based data (Almquist, Arya, Zeng, & Spiro, 2019). There is also rising interest in using self-tracking data for personal health and healthcare management (Ajana, 2017). This is an exciting area of study with the potential to have big impacts on individual health and estimates of population-level health measurements (e.g., sleep studies) and our understanding of health mechanisms (e.g., time and quality of running on personal health).

Machine Learning, Statistics and Simulation

A major challenge with big data is that much of it is not developed for research purposes but is organically created by human action patterns such as flying or using

social media. Major attempts to de-bias and illicit scientifically useful information from this data sources have led to new and improved statistical and machine learning methods.

Text analysis and Natural Language Processing

Internet search terms have been analyzed to measure racism and mortality (Chae et al., 2015), estimate and predict disease trends like flu and syphilis (Clemente, Lu, & Santillana, 2019; Ginsberg et al., 2009; Young et al., 2018), and understand suicide trends (Jimenez, Santed-Germán, & Ramos, 2020).

Promising studies published based on textual analysis of social media data using NLP methods to trace health-related behaviors. The outcome of interests are quite varied, ranging from food consumption (Shah, Srivastava, Savage, & Mago, 2020), noise pollution (Peplow, Thomas, & AlShehhi, 2021), vaccine-hesitancy (Wilson & Wiysonge, 2020), mental health (Valdez, ten Thij, Bathina, Rutter, & Bollen, 2020; Zaydman, 2017). There is a large body of work on using text as data in the social sciences broadly (Grimmer & Stewart, 2013), and the potential uses in demography and health are just coming to fruition as the availability and usability of these methods mature.

Database linkages

Data linked between multiple sources (e.g., US Census data with Internal Revenue Service migration data) is a powerful resource for modern demographic and health research. Data is linked by many different methods, from hard linkages (e.g., Ferrie (1996)’s work connecting 1850 to 1860 US Census data) to probabilistic linkage methods (e.g., Steorts, Hall, and Fienberg (2016) has developed Bayesian methods for entity resolution, i.e., record linkages). This is an area that is changing how we think about data in the social sciences as we are able to link demographic data to major health databases (e.g., Kaiser Permanente health data records (Paxton, Inacio, Slipchenko, & Fithian, 2008)) or other administrative data such as the IRS migration data Hauer and Byars (2019).

Public health and demography

Public health policy researchers have studied governmental regulations and campaigns (including tweets from experts) to derive data-driven policy recommendations for public health experts. For example, Ojo, Guntuku, Zheng, Beidas, and Ranney (2021) compared healthcare-led Twitter hashtags with non-healthcare-led hashtags during Gun Violence and COVID-19 epidemics and showed that the former expressed more positive and action-oriented language than the latter. Lohmann et al. (2018) analyzed the tweets by HIV experts and their engagement and found that fear-related language, longer messages, and including images (e.g., photos, GIFs, or videos) resulted in more shares. And, Sumner, Burke, and Kooti (2020) analyzed suicide reporting news on Facebook across countries and found that closer adherence to safe-reporting practices is associated with more engagements (e.g., likes and shares). Gozzi, Perrotta, Paolotti, and Perra (2020) conducted online surveys via Influweb, a web platform for participatory surveillance in Italy, and utilized machine learning algorithms to examine the factors driving an individual’s behavior.

Social Networks and Diffusion of Information and Infection

A major growth area in computational demography is the application of empirical and simulated networks for obtaining population-level estimates of diffusion processes for public health contexts. For example, Almquist (2020) demonstrated that it is possible to simulate homeless-to-homeless networks for understanding the diffusion process of information or infection for people experiencing homelessness in major metropolitan areas using very limited information such as the Point-in-Time survey conducted yearly on people experiencing homelessness (see Almquist, Helwig, & You, 2020, for more details). Recent work by Thomas et al. (2020) has extended this work and that of others to pull apart the network mechanisms that underlie COVID-19 transmission dynamics. Overall, there is extensive literature on network diffusion methods (see the review by Zhang & Centola, 2019).

Synthetic world migration data

Historically, there has been minimal data on worldwide international migration. In the last decade, there have been two major developments in creating global migration data. Abel and Sander (2014) generated 5-year migration flow estimates from sequential stock tables published by the United Nations (UN) for 1990, 2000, and 2010. More recently, Azose and Raftery (2019) estimated 5-year bilateral migration data using return and transit components with a Bayesian statistical model. Specifically, Azose and Raftery (2019) finds the “total number of individuals migrating internationally has oscillated between 1.13 and 1.29% of the global population per 5-year period since 1990.”

Fertility and mortality forecasting

Major updates to population forecasting have been developed in the last 30 years. Raftery, Li, Ševčíková, Gerland, and Heilig (2012) has generated probabilistic forecasts for all UN countries. This has been further improved to consider uncertainty in migration counts (Azose, Ševčíková, & Raftery, 2016). This is an important growing field of research within the computational demography community.

Applications to demography and health processes

Work by Kim et al. (2020) and others have demonstrated that social media populations and census demographics do not differ wildly. Providing evidence that data collected from social media and other online sources can provide useful evidence in a demographic and health context. Noteworthy health applications include using Twitter to monitor syphilis (Young et al., 2018), examining the effect of lock-down orders during an Ebola outbreak on mobility (Peak et al., 2018), and using Facebook data with MRP to improve behavioral health response for spatial units like the county (Loux et al., 2019). This is, obviously, a non-exhaustive list, but it provides a basic sense of this area’s extensive coverage and future importance to demographic and public health research.

A Note about COVID-19

Major innovations to computational demography and health have arisen in the last three years due to COVID-19. This includes major surveillance activities such as the partnership with Apple and Google for contact tracing (Michael & Abbas, 2020) and the Facebook COVID-19 Symptoms Survey (Reinhart et al., 2021). Other significant works include models and predictions of the epidemic spread (for a review see Jones, Hazel, & Almquist, 2020) and estimates of COVID-19's impact on racial inequities (Thomas et al., 2022; Wrigley-Field, 2020).

Conclusion

Computational demography is an emerging field in demography and public health with the potential to change how we measure, model, and theorize about population and health processes. This area includes major developments in data, such as physical tracking, mobile phone data, geolocated information, specialized app development, historical text data, and classic administrative data. Statistical and computational methods are improving our analysis, prediction, and ability to harmonize this data (e.g., probabilistic record linkages). Significant work is being done to de-bias or minimally bias statistical estimates, such as the research in de-biasing and extrapolating online trace data, such as the work done in re-weighting activity-based data (Almquist et al., 2019), Multilevel-Regression with Poststratification for public health data (Loux et al., 2019) or demographic re-weighting for COVID-19 symptoms survey (Reinhart et al., 2021). These methods are not without controversy and detractors; see Bradley et al. (2021) work comparing a traditional address-based sample with a large-scale Facebook sample and found that the Facebook sample with basic demographic adjustments was consistently biased for the general population. Other promising research areas are real-time response to things like natural disasters (Maas et al., 2020) and health monitoring of Google Flu, Twitter trends, and Facebook surveys.

Further readings

Readers interested in a broader review of ethical considerations in computational social science can look at the Ethics chapter in Salganik (2019). A deeper discussion on data privacy and best practices can be found in Kearns and Roth (2019). Further readings on using big data for health research: on privacy issues related to social media data (Ford, Shepherd, Jones, & Hassan, 2021) and concerns about marginalized communities (Wesson, Hswen, Valdes, Stojanovski, & Handley, 2022).

Other data sources are used in demographic and health research: LinkedIn for economic migration (Perrotta et al., 2022) or gender gaps (Kashyap & Verkroost, 2021), bibliometric data for high-skilled return migration (Zhao, Aref, Zagheni, & Stecklov, 2022), and Google Search Trends for contraceptive behavior (Berger, Ferrari, Leturcq, Panico, & Solaz, 2021) and intimate partner violence (Köksal, Pesando, Rotondi, & Şanlıtürk, 2022). A review on machine learning methods for social sciences can be found in Molina and Garip (2019).

Further readings on online surveys (social media, web-based, panel): an empirical assessment of different types (Lehdonvirta, Oksanen, Räsänen, & Blank, 2021) and meta-analysis comparing response rates (Daikeler, Bošnjak, & Lozar Manfreda,

2020). Recent publications assessed the representativeness of social media samples, such as Facebook Grow et al. (2022) and Twitter Morstatter, Pfeffer, Liu, and Carley (2021). The reviews on the advances on network sampling methods: respondent-driven sampling (Goel & Salganik, 2010) and ego-centric sampling (Krivitsky, Morris, & Bojanowski, 2022).

For a literature review on computational social science (Edelmann, Wolff, Montagne, & Bail, 2020) and demography (McCormick, Lee, Cesare, Shojaie, & Spiro, 2017). Readers interested in epistemological discussions on existing problems and broad recommendations (D. M. J. Lazer et al., 2020), deriving scientific measurements from trace data (D. Lazer et al., 2021), methodological integration across social sciences and computational science (Hofman et al., 2021), and mixed-methods framework for text analysis for demographic research (Chakrabarti & Frye, 2017).

For a review of recent advancements in computational approaches in migration research, see Drouhot, Deutschmann, Zuccotti, and Zagheni (2022). Crowd-sourced online genealogies present ample opportunities for demographers, for example, mortality and kinship (Alburez-Gutierrez, 2022). However, one must consider the embedded bias in family lineages (Stelter & Alburez-Gutierrez, 2022). Another rising area of research is Genome-Wide Association Studies (GWAS), thanks to large-scale open-access genomic data (Mills & Tropf, 2020; Uffelmann et al., 2021). Recent studies published on emerging health fields, a survey of personalized medicine literature (Bershteyn, Kim, & Braithwaite, 2022), and disparate effects of digitalized health (Timmermans & Kaufman, 2020).

Acknowledgments

Partial support for this research came from a Eunice Kennedy Shriver National Institute of Child Health and Human Development research infrastructure grant, P2C HD042828, and Shanahan Endowment Fellowship, and a Eunice Kennedy Shriver National Institute of Child Health and Human Development training grant, T32 HD101442, to the Center for Studies in Demography & Ecology at the University of Washington. Partial support for this research also came from an NSF CAREER Award #2142964 and an ARO Award #W911NF-19-1-0407. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, NSF, or ARO.

References

- Abel, G. J., & Sander, N. (2014). Quantifying global international migration flows. *Science (New York, N.Y.)*, *343*(6178), 1520–1522.
- Ajana, B. (2017). Digital health and the biopolitics of the Quantified Self. *Digital Health*, *3*, 2055207616689509.
- Alburez-Gutierrez, D. (2022). The Demographic Drivers of Grief and Memory After Genocide in Guatemala. *Demography*, *59*(3), 1173–1194.
- Alexander, M., Polimis, K., & Zagheni, E. (2019). The Impact of Hurricane Maria on Out-migration from Puerto Rico: Evidence from Facebook Data. *Population and Development Review*, *45*(3), 617–630.
- Alexander, M., Polimis, K., & Zagheni, E. (2020). Combining Social Media and Survey Data to Nowcast Migrant Stocks in the United States. *Population Research and Policy Review*.

- Almquist, Z. W. (2020). Large-scale spatial network models: An application to modeling information diffusion through the homeless population of San Francisco. *Environment and Planning B: Urban Analytics and City Science*, 47(3), 523–540.
- Almquist, Z. W., Arya, S., Zeng, L., & Spiro, E. (2019). Unbiased sampling of users from (online) activity data. *Field Methods*, 31(1), 23–38.
- Almquist, Z. W., Helwig, N. E., & You, Y. (2020). Connecting Continuum of Care point-in-time homeless counts to United States Census areal units. *Mathematical Population Studies*, 27(1), 46–58.
- Andersen, M. (2020). Early evidence on social distancing in response to COVID-19 in the United States. *Available at SSRN 3569368*.
- Aramburu, M. J., Berlanga, R., & Lanza, I. (2020). Social Media Multidimensional Analysis for Intelligent Health Surveillance. *International Journal of Environmental Research and Public Health*, 17(7), 2289.
- Astley, C. M., Tuli, G., Mc Cord, K. A., Cohn, E. L., Rader, B., Varrelman, T. J., ... others (2021). Global monitoring of the impact of the COVID-19 pandemic through online surveys sampled from the Facebook user base. *Proceedings of the National Academy of Sciences*, 118(51).
- Azose, J. J., & Raftery, A. E. (2019). Estimation of emigration, return migration, and transit migration between all pairs of countries. *Proceedings of the National Academy of Sciences*, 116(1), 116–122.
- Azose, J. J., Ševčíková, H., & Raftery, A. E. (2016). Probabilistic population projections with migration uncertainty. *Proceedings of the National Academy of Sciences*, 113(23), 6460–6465.
- Bailey, M., Cao, R., Kuchler, T., Stroebel, J., & Wong, A. (2018). Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives*, 32(3), 259–80.
- Basellini, U., Alburez-Gutierrez, D., Del Fava, E., Perrotta, D., Bonetti, M., Camarda, C. G., & Zagheni, E. (2021). Linking excess mortality to mobility data during the first wave of COVID-19 in England and Wales. *SSM - Population Health*, 14, 100799.
- Batbaatar, E., & Ryu, K. H. (2019). Ontology-Based Healthcare Named Entity Recognition from Twitter Messages Using a Recurrent Neural Network Approach. *International Journal of Environmental Research and Public Health*, 16(19), 3628.
- Berger, L. M., Ferrari, G., Leturcq, M., Panico, L., & Solaz, A. (2021). COVID-19 lockdowns and demographically-relevant Google Trends: A cross-national analysis. *PLOS ONE*, 16(3), e0248072.
- Bershteyn, A., Kim, H.-Y., & Braithwaite, R. S. (2022). Real-Time Infectious Disease Modeling to Inform Emergency Public Health Decision Making. *Annual Review of Public Health*, 43(1), 397–418.
- Blondel, V. D., Decuyper, A., & Krings, G. (2015). A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(1), 10.
- Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264), 1073–1076.
- Blumenstock, J. E. (2012). Inferring patterns of internal migration from mobile phone call records: Evidence from Rwanda. *Information Technology for Development*, 18(2), 107–125.
- Bohon, S. A. (2018). Demography in the Big Data Revolution: Changing the Culture to Forge New Frontiers. *Population Research and Policy Review*, 37(3), 323–341.
- Bouwer, L. M. (2011). Have disaster losses increased due to anthropogenic climate change? *Bulletin of the American Meteorological Society*, 92(1), 39–46.
- Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L., & Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, 600(7890), 695–700.
- Cas, A. G., Frankenberg, E., Suriastini, W., & Thomas, D. (2014). The impact of parental death on child well-being: Evidence from the Indian Ocean tsunami. *Demography*, 51(2),

- 437–457.
- Cesare, N., Lee, H., McCormick, T., Spiro, E., & Zagheni, E. (2018). Promises and Pitfalls of Using Digital Traces for Demographic Research. *Demography*, 55(5), 1979–1999.
- Chae, D. H., Clouston, S., Hatzenbuehler, M. L., Kramer, M. R., Cooper, H. L. F., Wilson, S. M., ... Link, B. G. (2015). Association between an Internet-Based Measure of Area Racism and Black Mortality. *PLOS ONE*, 10(4), e0122963.
- Chakrabarti, P., & Frye, M. (2017). A mixed-methods framework for analyzing text data: Integrating computational techniques with qualitative methods in demography. *Demographic Research*, 37, 1351–1382.
- Cherian, R., Westbrook, M., Ramo, D., & Sarkar, U. (2018). Representations of Codeine Misuse on Instagram: Content Analysis. *JMIR Public Health and Surveillance*, 4(1), e22.
- Chiolerio, A., & Buckeridge, D. (2020). Glossary for public health surveillance in the age of data science. *Journal of Epidemiology and Community Health*, jech-2018-211654.
- Chisholm, E., & O’Sullivan, K. (2017). Using Twitter to Explore (un)Healthy Housing: Learning from the #Characterbuildings Campaign in New Zealand. *International Journal of Environmental Research and Public Health*, 14(11), 1424.
- Clemente, L., Lu, F., & Santillana, M. (2019). Improved Real-Time Influenza Surveillance: Using Internet Search Data in Eight Latin American Countries. *JMIR Public Health and Surveillance*, 5(2), e12214.
- Daikeler, J., Bošnjak, M., & Lozar Manfreda, K. (2020). Web Versus Other Survey Modes: An Updated and Extended Meta-Analysis Comparing Response Rates. *Journal of Survey Statistics and Methodology*, 8(3), 513–539.
- Davies, N. G., Barnard, R. C., Jarvis, C. I., Russell, T. W., Semple, M. G., Jit, M., ... ISARIC4C investigators (2021). Association of tiered restrictions and a second lockdown with COVID-19 deaths and hospital admissions in England: A modelling study. *The Lancet. Infectious diseases*, 21(4), 482–492.
- Delpierre, C., & Kelly-Irving, M. (2018). Big Data and the Study of Social Inequalities in Health: Expectations and Issues. *Frontiers in Public Health*, 6, 312.
- Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., ... Tatem, A. J. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45), 15888–15893.
- DeWaard, J., Johnson, J. E., & Whitaker, S. D. (2020). Out-migration from and return migration to Puerto Rico after Hurricane Maria: Evidence from the consumer credit panel. *Population and Environment*, 42(1), 28–42.
- Drouhot, L. G., Deutschmann, E., Zuccotti, C. V., & Zagheni, E. (2022). Computational approaches to migration and integration research: Promises and challenges. *Journal of Ethnic and Migration Studies*, 1–19.
- Edelmann, A., Wolff, T., Montagne, D., & Bail, C. A. (2020). Computational Social Science and Sociology. *Annual Review of Sociology*, 46(1), 61–81.
- Feehan, D. M., & Cobb, C. (2019). Using an Online Sample to Estimate the Size of an Offline Population. *Demography*, 56(6), 2377–2392.
- Ferrie, J. P. (1996). A new sample of males linked from the public use microdata sample of the 1850 US federal census of population to the 1860 US federal census manuscript schedules. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 29(4), 141–156.
- Ford, E., Shepherd, S., Jones, K., & Hassan, L. (2021). Toward an Ethical Framework for the Text Mining of Social Media for Health Research: A Systematic Review. *Frontiers in Digital Health*, 2, 592237.
- Frankenberg, E., Sikoki, B., Sumantri, C., Suriastini, W., & Thomas, D. (2013). Education, Vulnerability, and Resilience after a Natural Disaster. *Ecology and Society*, 18(2).
- Fussell, E., Curtis, K. J., & DeWaard, J. (2014). Recovery migration to the City of New Orleans after Hurricane Katrina: A migration systems approach. *Population and Environment*, 35(3), 305–322.

- Gething, P. W., & Tatem, A. J. (2011). Can mobile phone data improve emergency response to natural disasters? *PLoS medicine*, 8(8), e1001085.
- Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232), 1012–1014.
- Gittelman, S., Lange, V., Gotway Crawford, C. A., Okoro, C. A., Lieb, E., Dhingra, S. S., & Trimarchi, E. (2015). A New Source of Data for Public Health Surveillance: Facebook Likes. *Journal of Medical Internet Research*, 17(4), e98.
- Goel, S., & Salganik, M. J. (2010). Assessing respondent-driven sampling. *Proceedings of the National Academy of Sciences*, 107(15), 6743–6747.
- Gozzi, N., Perrotta, D., Paolotti, D., & Perra, N. (2020). Towards a data-driven characterization of behavioral changes induced by the seasonal flu. *PLOS Computational Biology*, 16(5), e1007879.
- Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3), 267–297.
- Grow, A., Perrotta, D., Del Fava, E., Cimentada, J., Rampazzo, F., Gil-Clavel, S., & Zagheni, E. (2020). Addressing Public Health Emergencies via Facebook Surveys: Advantages, Challenges, and Practical Considerations. *Journal of Medical Internet Research*, 22(12), e20653.
- Grow, A., Perrotta, D., Del Fava, E., Cimentada, J., Rampazzo, F., Gil-Clavel, S., . . . Weber, I. (2022). Is Facebook’s advertising data accurate enough for use in social science research? Insights from a CROSS-NATIONAL online survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185(S2).
- Harris, J. K., Hawkins, J. B., Nguyen, L., Nsoesie, E. O., Tuli, G., Mansour, R., & Brownstein, J. S. (2017). Using Twitter to Identify and Respond to Food Poisoning: The Food Safety STL Project. *Journal of Public Health Management and Practice*, 23(6), 577–580.
- Hauer, M., & Byars, J. (2019). IRS county-to-county migration data, 1990–2010. *Demographic Research*, 40, 1153–1166.
- Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., . . . Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, 595(7866), 181–188.
- Hunter, R. F., Gough, A., O’Kane, N., McKeown, G., Fitzpatrick, A., Walker, T., . . . Kee, F. (2018). Ethical Issues in Social Media Research for Public Health. *American Journal of Public Health*, 108(3), 343–348.
- Jimenez, A., Santed-Germán, M.-A., & Ramos, V. (2020). Google Searches and Suicide Rates in Spain, 2004-2013: Correlation Study. *JMIR Public Health and Surveillance*, 6(2), e10919.
- Jones, J. H., Hazel, A., & Almquist, Z. (2020). Transmission-dynamics models for the SARS Coronavirus-2. *American Journal of Human Biology*, 32(5).
- Joshi, A., Sparks, R., McHugh, J., Karimi, S., Paris, C., & MacIntyre, C. R. (2020). Harnessing Tweets for Early Detection of an Acute Disease Event. *Epidemiology*, 31(1), 90–97.
- Kashyap, R., & Verkroost, F. C. J. (2021). Analysing global professional gender gaps using LinkedIn advertising data. *EPJ Data Science*, 10(1), 1–32.
- Kearns, M., & Roth, A. (2019). *The Ethical Algorithm : The Science of Socially Aware Algorithm Design*. Oxford, UNITED STATES: Oxford University Press USA - OSO.
- Keralis, J. M., Javanmardi, M., Khanna, S., Dwivedi, P., Huang, D., Tasdizen, T., & Nguyen, Q. C. (2020). Health and the built environment in United States cities: Measuring associations using Google Street View-derived indicators of the built environment. *BMC Public Health*, 20(1), 215.
- Kim, E. S., James, P., Zevon, E. S., Trudel-Fitzgerald, C., Kubzansky, L. D., & Grodstein, F. (2020). Social Media as an Emerging Data Resource for Epidemiologic Research: Characteristics of Regular and Nonregular Social Media Users in Nurses’ Health Study II. *American Journal of Epidemiology*, 189(2), 156–161.
- Köksal, S., Pesando, L. M., Rotondi, V., & Şanlıtürk, E. (2022). Harnessing the Potential of

- Google Searches for Understanding Dynamics of Intimate Partner Violence Before and After the COVID-19 Outbreak. *European Journal of Population*, 38(3), 517–545.
- Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24), 8788–8790.
- Krivitsky, P. N., Morris, M., & Bojanowski, M. (2022). Impact of survey design on estimation of exponential-family random graph models from egocentrically-sampled data. *Social Networks*, 69, 22–34.
- Lazer, D., Hargittai, E., Freelon, D., Gonzalez-Bailon, S., Munger, K., Ognyanova, K., & Radford, J. (2021). Meaningful measures of human society in the twenty-first century. *Nature*, 595(7866), 189–196.
- Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, 343(6176), 1203–1205.
- Lazer, D., & Radford, J. (2017). Data ex Machina: Introduction to Big Data. *Annual Review of Sociology*, 43(1), 19–39.
- Lazer, D. M. J., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., . . . Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science*, 369(6507), 1060–1062.
- Lee, V. R. (2014). What’s happening in the” Quantified Self” movement? *ICLS 2014 Proceedings*, 1032.
- Lehdonvirta, V., Oksanen, A., Räsänen, P., & Blank, G. (2021). Social Media, Web, and Panel Surveys: Using Non-Probability Samples in Social and Policy Research. *Policy & Internet*, 13(1), 134–155.
- Lohmann, S., White, B. X., Zuo, Z., Chan, M.-P. S., Morales, A., Li, B., . . . Albarracín, D. (2018). HIV messaging on Twitter: An analysis of current practice and data-driven recommendations. *AIDS*, 32(18), 2799–2805.
- Loux, T., Nelson, E. J., Arnold, L. D., Shacham, E., & Schootman, M. (2019). Using multilevel regression with poststratification to obtain regional health estimates from a Facebook-recruited sample. *Annals of Epidemiology*, 39, 15-20.e5.
- Lupton, D. (2016). *The quantified self*. John Wiley & Sons.
- Maas, P. (2019). Facebook Disaster Maps: Aggregate Insights for Crisis Response & Recovery. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 3173–3173). Anchorage AK USA: ACM.
- Maas, P., Almquist, Z., Giraudy, E., & Schneider, J. (2020). Using Social Media to measure demographic responses to natural disaster: Insights from a large-scale Facebook survey following the 2019 Australia Bushfires. *arXiv preprint arXiv:2008.03665*.
- Martín, Y., Cutter, S. L., Li, Z., Emrich, C. T., & Mitchell, J. T. (2020). Using geotagged tweets to track population movements to and from Puerto Rico after Hurricane Maria. *Population and Environment*, 42(1), 4–27.
- McCormick, T. H., Lee, H., Cesare, N., Shojaie, A., & Spiro, E. S. (2017). Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing. *Sociological Methods & Research*, 46(3), 390–421.
- Mejova, Y., Weber, I., & Fernandez-Luque, L. (2018). Online Health Monitoring using Facebook Advertisement Audience Estimates in the United States: Evaluation Study. *JMIR Public Health and Surveillance*, 4(1), e30.
- Michael, K., & Abbas, R. (2020). Behind COVID-19 contact trace apps: The Google–Apple partnership. *IEEE Consumer Electronics Magazine*, 9(5), 71–76.
- Mills, M. C., & Tropf, F. C. (2020). Sociology, Genetics, and the Coming of Age of Sociogenomics. *Annual Review of Sociology*, 46(1), 553–581.
- Molina, M., & Garip, F. (2019). Machine Learning for Sociology. *Annual Review of Sociology*, 45(1), 27–45.
- Mooney, S. J., Westreich, D. J., & El-Sayed, A. M. (2015). Commentary: Epidemiology in the Era of Big Data. *Epidemiology*, 26(3), 390–394.
- Morrison, P. A., Edmondson, B., Ferrantella, K., Lockhart, D., Reis, S., & Tapp, A. (2020).

- Estimating Nantucket’s Effective Population. *Population Research and Policy Review*, 39(3), 577–604.
- Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. (2021). Is the Sample Good Enough? Comparing Data from Twitter’s Streaming API with Twitter’s Firehose. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1), 400–408.
- Newing, A., Anderson, B., Bahaj, A., & James, P. (2016). The Role of Digital Trace Data in Supporting the Collection of Population Statistics - the Case for Smart Metered Electricity Consumption Data: Collecting Population Statistics via Smart Metered Electricity Load Data. *Population, Space and Place*, 22(8), 849–863.
- Ojo, A., Guntuku, S. C., Zheng, M., Beidas, R. S., & Ranney, M. L. (2021). How Health Care Workers Wield Influence Through Twitter Hashtags: Retrospective Cross-sectional Study of the Gun Violence and COVID-19 Public Health Crises. *JMIR Public Health and Surveillance*, 7(1), e24562.
- Palmer, J. R. B., Espenshade, T. J., Bartumeus, F., Chung, C. Y., Ozgencil, N. E., & Li, K. (2012). New Approaches to Human Mobility: Using Mobile Phones for Demographic Research. *Demography*, 50(3), 1105–1128.
- Paxton, E. W., Inacio, M., Slipchenko, T., & Fithian, D. C. (2008). The Kaiser Permanente national total joint replacement registry. *The Permanente Journal*, 12(3), 12.
- Peak, C. M., Wesolowski, A., zu Erbach-Schoenberg, E., Tatem, A. J., Wetter, E., Lu, X., . . . Bengtsson, L. (2018). Population mobility reductions associated with travel restrictions during the Ebola epidemic in Sierra Leone: Use of mobile phone data. *International Journal of Epidemiology*, 47(5), 1562–1570.
- Peplow, A., Thomas, J., & AlShehhi, A. (2021). Noise Annoyance in the UAE: A Twitter Case Study via a Data-Mining Approach. *International Journal of Environmental Research and Public Health*, 18(4), 2198.
- Perrotta, D., Johnson, S. C., Theile, T., Grow, A., de Valk, H., & Zagheni, E. (2022). Openness to Migrate Internationally for a Job: Evidence from LinkedIn Data in Europe. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1), 759–769.
- Raftery, A. E., Li, N., Ševčíková, H., Gerland, P., & Heilig, G. K. (2012). Bayesian probabilistic population projections for all countries. *Proceedings of the National Academy of Sciences*, 109(35), 13915–13921.
- Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. *Health Information Science and Systems*, 2(1), 3.
- Reinhart, A., Brooks, L., Jahja, M., Rumack, A., Tang, J., Agrawal, S., . . . others (2021). An open repository of real-time COVID-19 indicators. *Proceedings of the National Academy of Sciences*, 118(51).
- Ribeiro, F. N., Benevenuto, F., & Zagheni, E. (2020). How biased is the population of facebook users? Comparing the demographics of facebook users with census data to generate correction factors. In *12th ACM conference on web science* (pp. 325–334). New York, NY, USA: Association for Computing Machinery.
- Ruggles, S. (2014). Big Microdata for Population Research. *Demography*, 51(1), 287–297.
- Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, 346(6213), 1063–1064.
- Salerno, J., Knoppers, B. M., Lee, L. M., Hlaing, W. M., & Goodman, K. W. (2017). Ethics, big data and computing in epidemiology and public health. *Annals of Epidemiology*, 27(5), 297–301.
- Salganik, M. J. (2019). *Bit by bit: Social research in the digital age*. Princeton University Press.
- Shah, N., Srivastava, G., Savage, D. W., & Mago, V. (2020). Assessing Canadians Health Activity and Nutritional Habits Through Social Media. *Frontiers in Public Health*, 7, 400.
- Stelter, R., & Alburez-Gutierrez, D. (2022). Representativeness is crucial for inferring demographic processes from online genealogies: Evidence from lifespan dynamics. *Proceedings*

- of the *National Academy of Sciences*, 119(10), e2120455119.
- Steorts, R. C., Hall, R., & Fienberg, S. E. (2016). A Bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association*, 111(516), 1660–1672.
- Stevens, R. C., Brawner, B. M., Kranzler, E., Giorgi, S., Lazarus, E., Abera, M., . . . Ungar, L. (2020). Exploring Substance Use Tweets of Youth in the United States: Mixed Methods Study. *JMIR Public Health and Surveillance*, 6(1), e16191.
- Stewart, I., Flores, R. D., Riffe, T., Weber, I., & Zagheni, E. (2019). Rock, rap, or reggaeton?: Assessing mexican immigrants’ cultural assimilation using facebook data. In *The world wide web conference* (pp. 3258–3264).
- Sumner, S. A., Burke, M., & Kooti, F. (2020). Adherence to suicide reporting guidelines by news shared on a social networking platform. *Proceedings of the National Academy of Sciences*, 117(28), 16267–16272.
- Thomas, L. J., Huang, P., Yin, F., Luo, X. I., Almquist, Z. W., Hipp, J. R., & Butts, C. T. (2020). Spatial heterogeneity can lead to substantial local variations in COVID-19 timing and severity. *Proceedings of the National Academy of Sciences*, 117(39), 24180–24187.
- Thomas, L. J., Huang, P., Yin, F., Xu, J., Almquist, Z. W., Hipp, J. R., & Butts, C. T. (2022). Geographical patterns of social cohesion drive disparities in early COVID infection hazard. *Proceedings of the National Academy of Sciences*, 119(12), e2121675119.
- Timmermans, S., & Kaufman, R. (2020). Technologies and Health Inequities. *Annual Review of Sociology*, 46(1), 583–602.
- Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., . . . Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1), 1–21.
- Valdez, D., ten Thij, M., Bathina, K., Rutter, L. A., & Bollen, J. (2020). Social Media Insights Into US Mental Health During the COVID-19 Pandemic: Longitudinal Analysis of Twitter Data. *Journal of Medical Internet Research*, 22(12), e21418.
- Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3), 980–991.
- Wesson, P., Hswen, Y., Valdes, G., Stojanovski, K., & Handley, M. A. (2022). Risks and Opportunities to Ensure Equity in the Application of Big Data Research in Public Health. *Annual Review of Public Health*, 43(1), 59–78.
- Wilson, S. L., & Wiysonge, C. (2020). Social media and vaccine hesitancy. *BMJ Global Health*, 5(10), e004206.
- Wrigley-Field, E. (2020). US racial inequality may be as deadly as COVID-19. *Proceedings of the National Academy of Sciences*, 117(36), 21854–21856.
- Yildiz, D., Munson, J., Vitali, A., Tinati, R., & Holland, J. A. (2017). Using Twitter data for demographic research. *Demographic Research*, 37(46), 1477–1514.
- Young, S. D., Torrone, E. A., Urata, J., & Aral, S. O. (2018). Using Search Engine Data as a Tool to Predict Syphilis. *Epidemiology*, 29(4), 574–578.
- Younis, J., Freitag, H., Ruthberg, J. S., Romanes, J. P., Nielsen, C., & Mehta, N. (2020). Social Media as an Early Proxy for Social Distancing Indicated by the COVID-19 Reproduction Number: Observational Study. *JMIR Public Health and Surveillance*, 6(4), e21340.
- Yum, S. (2020). Mining Twitter Data to Understand the Human Sentiment on Hurricane Florence. *Journal of Disaster and Emergency Research*, 3(2), 74–86.
- Zagheni, E., & Weber, I. (2015). Demographic research with non-representative internet data. *International Journal of Manpower*, 36(1), 13–25.
- Zagheni, E., Weber, I., & Gummadi, K. (2017). Leveraging Facebook’s Advertising Platform to Monitor Stocks of Migrants: Leveraging Facebook’s Advertising Platform. *Population and Development Review*, 43(4), 721–734.
- Zaydman, M. (2017). *Tweeting About Mental Health: Big Data Text Analysis of Twitter for Public Policy*. RAND Corporation.
- Zeng, L., Almquist, Z. W., & Spiro, E. S. (2017). Let’s workout! Exploring social exercise in an online fitness community. *iConference 2017 Proceedings Vol. 2*.

- Zeng, L., Almquist, Z. W., & Spiro, E. S. (2018). Stay connected and keep motivated: Modeling activity level of exercise in an online fitness community. In *International conference on social computing and social media* (pp. 137–147).
- Zeng, L., Almquist, Z. W., & Spiro, E. S. (2019). “Friending” in Online Fitness Communities: Exploring Activity-Based Online Network Structure. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.
- Zhang, J., & Centola, D. (2019). Social networks and health: New developments in diffusion, online and offline. *Annual Review of Sociology*, 45, 91–109.
- Zhao, X., Aref, S., Zagheni, E., & Stecklov, G. (2022). Return migration of German-affiliated researchers: Analyzing departure and return by gender, cohort, and discipline using Scopus bibliometric data 1996–2020. *Scientometrics*, 127(12), 7707–7729.

Small Area Estimation for Demographic Data

Aja M. Sutton, University of Washington, Seattle, USA (amsutton@uw.edu)

Zack W. Almquist, University of Washington, Seattle, USA (zalmquist@uw.edu)

Abstract

In this chapter we will introduce recent developments in the application of small area estimation (SAE) such as combining new statistical methods, e.g. multilevel regression and poststratification (MRP) and neighborhood spatial smoothing models (e.g. a modified Besag-York-Mollié, or BYM2), and consider recent applications for demographic processes such as mortality or migration estimation at local levels. We will discuss recent methodological developments and important applications of SAE such as the United Nations COVID-19 national and subnational mortality estimates. Last, we will provide a clear example of SAE using United States ZIP code-level data for core demographic questions relating to mortality and migration.

Keywords: Small area estimation, forecasting, demographic data, geographic data, areal units

1. Introduction

Demography is a discipline dependent on the accurate enumeration of population-level processes. The foundational unit of measurement for analyzing many demographic processes is the count of individuals; it allows demographers to derive all other statistical descriptions of population dynamics. Therefore, it is crucial to have reliable population counts of the outcome of interest whether they relate to mortality (e.g. number in the population who died), fertility (e.g. number of births in the population), morbidity (e.g. number who contracted a disease), population-level social attitudes (e.g. measurable responses to a sociological survey), migration (e.g. counts of movement between locations) or otherwise. However, unless there is near-perfect enumeration through a census, it is often difficult to establish the count of an outcome in a population. Classic survey methods and modern statistics offer a way to directly estimate these counts. However, when a survey has inadequate sample size, it may not be possible to make reliable direct estimates of the outcome of interest (e.g. counts of people 5-10 years old in Nigeria). Instead, it is often possible to make indirect estimates through methods that are often referred to as small area estimation (SAE). Here, a small area may either refer geographic areas/regions (e.g. counties, states, provinces, census tracts, etc.) or demographic groups (i.e. the various combinations of demographic characteristics present in survey data used to categorize individual participants, including sex, age, race/ethnicity, employment status, etc.) [1]. Small area estimation methods use additional contextual information to produce statistically robust estimates of under-



or unobserved subpopulations or geographic units. Classic spatial models used for public health mapping and small area estimation include the Fay-Herriot model [2], the Besag-York-Mollié model [3], and the modified Besag-York-Mollié model (referred to as the BYM2 model) [4]. While SAE may use such spatial models in producing estimates, recent advancements in computational power mean that in many cases these models are only just beginning to be considered as potentially viable additions to more common SAE methods like multilevel regression and poststratification.

2. Small Area Estimation in Demographic Research

The building blocks of demographic research are census and survey data. As discussed, sometimes these data are unreliable, incomplete or not measured in a timely manner. Increasingly, demographers are using SAE methods to attempt to produce more reliable indirect estimates as part of their analytic processes. This is especially true in high- and middle-income countries where data sparseness may exist in some geographic regions or demographic domains, but data quality and presence are otherwise high. While direct methods of SAE may incorporate estimation needs into survey designs, it is frequently also the case that post-hoc SAE is necessary using data not gathered explicitly for the purpose of generating estimates of the outcome of interest. In this case, indirect estimates may be produced using SAE. Many census authorities and national statistics offices have begun to use SAE for minor applications focused on tracking population trends during inter-census periods among subpopulations that are difficult to enumerate or underrepresented in inter-census data. For example, the United States Census Bureau has several small projects aimed at incorporating both geographic and temporal SAE methods to generate estimates in response to data user demands for increasingly finer-grain geographic areas like census tracts when complete domain data and/or adequate sample size are not available [5]. The American Community Survey (ACS) is perhaps the most well-known United States-based project that uses SAE methods; every month the ACS samples select households in all counties, and produces model-based direct estimates of many outcomes of interest. A significant product of the ACS is indirect estimates of the number of school-aged children living in poverty at the county and school district levels which are published in yearly and five-year aggregates, as are all other ACS estimates. At such a fine-grain geographic level it may be difficult and intrusive to directly enumerate the number of children experiencing poverty, but SAE methods allow the researcher to leverage sub-population level demographic data from sampled households to produce the indirect estimates in a reasonably unobtrusive yet reliable manner. Similar work has been undertaken to produce indirect estimates for public health surveillance, including estimating disease prevalence such as with cancer cases [6], and also health insurance coverage [7].

Just as with census authority work, a survey design or model may be produced to best represent attitudes by demographic characteristics and geographic context of any number of questions from opinion survey responses. In this context, SAE methods may then be used by organizations or groups interested in attitudes or opinions such as in academic contexts (e.g. sociologists, anthropologists, etc.), or business (e.g. market research). While classic use cases regarding public opinion polls are referenced below, some interesting and unique examples of these include estimating forest land ownership for the United States Department of Agriculture National Forest Service [8] and Australian electoral district-level attitudes toward same-sex marriage [9]. However, with these examples there is a presumption of high-quality data and limited “missingness” in terms of even fine-grain geographic



coverage. In countries that may have fewer or limited resources for adequate enumeration efforts through a formal census (often low-income countries), additional information may be generated from observed population data using SAE. For example, the UNICEF Under 5 Year Mortality Estimates are a series of regularly produced estimates of small area mortality rates that use national census and survey data using complex SAE methods for countries of particular concern regarding child survival rates [10]. In this way, SAE offers estimation methods of varying complexity and with varying thresholds of statistical expertise required for implementation but is also a suite of methods that can be applied flexibly to various population enumeration needs given data availability and adequate consideration. However, one SAE method that can be implemented with an accessible threshold level of expertise and is increasingly popular in high- and middle-income contexts is multilevel regression and poststratification.

3. Multilevel Regression and Poststratification: A basic introduction

Multilevel regression and poststratification (MRP) is a model-based statistical method that uses national/high-level administrative survey or census data¹ to adjust for non-representativeness in subnational surveys using national survey or census estimates, and/or to provide small area estimation (SAE) in areas where subnational survey data are sparse or nonexistent. MRP is computationally inexpensive and able to yield quick, consistent, and accurate estimates. Consequently, it is used for resource allocation and policy planning in settings such as public health administration. Prevailing methods of MRP for SAE borrow strength from other administrative areal data through partial pooling. While this approach is powerful, it does not account for any potential existing “neighborhood” spatial structure in the data. Including spatial specifications in MRP can be a valuable way to better handle spatial relationships and generate more accurate area-level estimates.

The traditional MRP method described initially by Gelman and Little (1997), and later expanded upon to consider applications for public opinion polls and surveys (e.g., [11, 12]; see also: [13-15]), is able to estimate the joint posterior distributions of the outcome of interest according to the survey population structure, and using Bayesian methods² to generate marginal estimates for unobserved subpopulations. It then adjusts these distributions using census or large-scale weighted population survey data to reflect the known population structure. If desired, it is then possible to use the mean posterior estimate (i.e., Bayesian methods) for each subpopulation in addition to the known area-level

¹ For the sake of clarity, here “survey data” refers to the unrepresentative data used to build the model, and “census data” are the data representing the known population used to adjust the unrepresentative data and from which population count estimates of the modeled outcome are made. However, it is possible to use large, representative data from the known population such as the American Community Survey 5-year estimates in the United States.

² In this case, we have used Integrated Nested Laplace Approximation (INLA), and so more correctly, our example figures are generated using a method that approximates Bayesian processes.



subpopulation counts from the same census data to predict the number of individuals in each area exhibiting the outcome of interest (the outcome is often presumed to have a binomial distribution, but this is not strictly required). While a relatively straightforward procedure (described more thoroughly below), MRP has strict data requirements. The multilevel regression model built must only use as covariates the variables present in both the survey data used to train the model and also the census data used to poststratify the marginal posterior probabilities. These complementary variables must be organized and coded in the exact same way (e.g., if “age” is used as a covariate, then each dataset must either represent age as a continuous variable, or bin age groups in the same manner). The individual-level survey data should contain at minimum two demographic characteristics per observed participant, in addition to good data availability for the outcome of interest to be modeled. For example, in addition to the respondent’s reported value for the outcome of interest, the sex and age; or sex and race/ethnicity; or sex, age and race/ethnicity must be known for each participant, and these characteristics must also be present in the census data. More concretely, if a model is built with the survey data using sex, age and race/ethnicity as the model covariates on which to later poststratify with the census data, then the census data must contain the known number of individuals belonging to each combination of covariate characteristics even if that number is zero (e.g. the number of people in the state of California who are identified as Black, male and aged 18-24; or the number of people in King County, Washington, USA, who are identified as White, female, and aged 25-34, etc. and every possible combination of these demographic characteristics). To this end, there must be a poststratification table that contains only the variables from the census data used to weight the joint posterior estimates and derive an estimated population count of the outcome. Furthermore, to adjust the estimates using the census data contained in the original poststratification table, it is necessary to calculate the inclusion weights of each poststratification level (e.g., the number of individuals in the poststratification level in ZIP code X divided by the total number of individuals in all poststratification levels in ZIP code X; these levels are clarified below).

For the multilevel model aspect of the MRP method, covariates are typically either modeled as fixed effects or random effects and may include interaction terms between fixed effects. The model is also where the traditional MRP method differs from MRP using spatial specifications. Traditional MRP methods may specify the model as, $\text{logit}^{-1}(y_j) \sim \alpha_j + \sum_{j=1}^J \beta_j + \eta_j$, where y is the outcome of interest, α is the intercept, β represents the covariates (e.g. age, sex, race/ethnicity, etc.) that vary according to poststratification level j , and η is the random error term for each level (specified as independent and identically distributed). In this form, it is possible to specify the geographic area covariate as a random effect for partial local pooling within the context of each of the areas between different categories within a demographic variable (e.g. partial pooling within ZIP code X within different categories of race/ethnicity), but this formulation does not account for spatial structure (i.e. the model does not specify a “neighborhood” spatial structure, and there is no sharing between geographic areas contingent on proximity).

It is possible to use spatial terms in the model if the survey data contain any spatial information also identifiable in the census data (e.g. if the geographic area level of interest is the county level in the context of the United States, these would be United States Census Bureau county FIPS codes which represent each individual county and for which a spatial polygon/shapefile may be easily obtained). While covariates in the model in the traditional MRP method are specified as being independent and identically distributed (IID), the flexible nature of multilevel models can be leveraged such that a spatial random effect may be included in the model (after [16]). A common spatial model used to specify the spatial random



Figure 1: Spatial Adjacency Queen-Style “Neighborhood” Structure

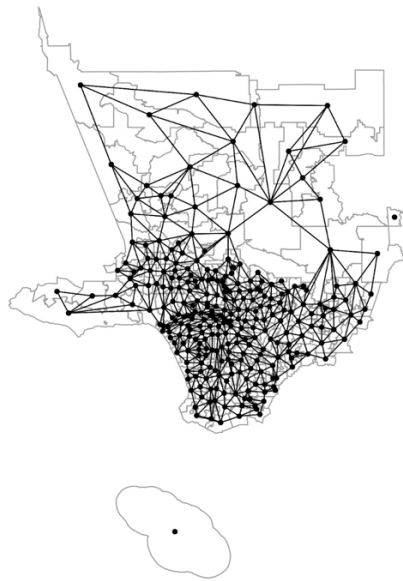


Figure 1 Caption: This simple map represents the United States Census Bureau ZIP Code Tabulation Areas (ZCTAs) roughly approximating Los Angeles County, California, USA. Each ZCTA has a point representing its approximate center, and a queen-style adjacency term has been specified to identify first-order neighbors (represented by the network of lines linking these points, called centroids). Note that the rough shape of Santa Catalina Island at the bottom of the figure has no neighbors identified due to it not sharing a border with any other ZCTA. Image generated in `R` [16] using `tidyverse`-based [17] packages including `tidycensus` [18]; additionally, `R-INLA` [19], and also geographic data sourced from the United States Census Bureau Tiger/Line Shapefiles.

effect is the modified Besag-York-Mollié (BYM2) model [4]. This model uses a spatial grid, or spatial adjacency matrix, generated using the survey data to take advantage of identifiable “neighborhood” structures of the data. Consequently, it also requires all poststratification data to be known for each geographic area of interest (e.g., if the level of geographic interest in the county level in the state of California, then the number of people described by each demographic characteristic combination used in the model in each county in California must be known).³ These are perhaps most simply thought of as poststratification levels (i.e., each possible combination of demographics for each geographic area, such as each county). In this way, while some covariates may be specified as IID, the spatial covariate included in the model is specified as using the modified Besag-York-Mollié (BYM2) model. When this is

³ While any geographic level that is present may be used, such as census tract or state, the present discussion will refer to the county level for simplicity.

Figure 2: Comparison of Number of Vaccinated Individuals: Direct Survey Estimates and Indirect Estimates (Classic MRP and MRP with a Spatial Term) for ZCTAs in Los Angeles County, California, September 2020



Figure 2 Caption: The direct survey estimates (left) represent the most common approach to generating population estimates of an outcome (in this case, the number of individuals vaccinated in Los Angeles County, California in September 2020 according to a large online survey [21]). The direct estimates are slightly higher in certain areas, whereas the indirect estimates appear to be virtually identical in this format. However, when the aggregate estimates are considered (Table 1), it is evident that in this case using MRP with a spatial term subtly improves the estimates. Figure created using `ggplot2` [22] and `viridis` [23] packages in `R`.

done, the data smoothing, and pooling is limited such that only information between immediate or “first-order” neighbors⁴ is shared (Figure 1). This is unlike the traditional MRP method that smooths and pools across the entirety of the data of interest even if the model includes a covariate that might seem spatial but is instead handled as an IID random effect (i.e. though a county is a geographic term, in traditional MRP it would not be handled using a spatial model, but instead specified as IID). In terms of implementation of spatial specifications for MRP, Gao, Kennedy and Simpson [20] proposed a modification to the MRP model, $\text{logit}^{-1}(y_{i,j}) \sim \alpha_{i,j} + \sum_{j=1}^J \beta_{i,j} + \eta_{i,j} + \epsilon_{i,j}$, where i is the area, the intercept is allowed to vary according to poststratification level j and area i , the covariate effects are summed across each area, and in addition to the random effects represented by η , ϵ represents the spatial random effects in area i , level j introduced by the BYM2 model.

When the preferred predictive multilevel model of the outcome of interest has been selected,⁵ the model is fit using Bayesian statistical software to the survey data. The model fit is then used to predict the outcome for each spatial level (e.g. each county) contingent on the population structure represented by the poststratification

⁴ As determined by the first-order neighborhood structure/grid.

⁵ Typical procedures for building a reasonable model of the outcome of interest may be applied here, including visual interaction checks and model selection processes, but these are beyond the scope of the present discussion.

data.⁶ During this process, through the power of the smoothing and pooling specifications used in the model, joint posterior distributions are also predicted for poststratification levels that are not present in the survey data. Then, one thousand draws are made from the resulting prediction object such that a table of the joint posterior distribution for each poststratification level is produced. For each level in each county, each joint posterior estimate is then multiplied by the inclusion weight previously calculated using the poststratification data from the census to create a distribution of weighted posterior estimates. At this point, the mean of each of these distributions is taken to produce the mean weighted probability for each poststratification level in each county. Thus, the first application for MRP is complete: the non-representative survey data have been adjusted using the census data. If small area estimates are desired for each geographic area (e.g., each county), the mean weighted probability for each poststratification level in each county is then multiplied by the number of known individuals in each county's matching poststratification level using the poststratification table. These poststratification level-specific outcome estimates are finally grouped by county and summed to create a population-level estimate of the outcome for each county but may be also informative if subpopulation estimates are desired. It is important to note that MRP with a spatial term is not limited to the county level. Figure 2 compares direct estimates of the number of individuals who were vaccinated in the Los Angeles County, California area in September 2020, at the ZIP code (U.S. Census Bureau ZCTA, or ZCTAs) level using traditional survey methods with indirect estimates generated using classic MRP and MRP with a spatial term. While the indirect estimates appear to be virtually identical in this format, and slightly below the direct estimates, when considering aggregated population results in Table 1 it is evident

Table 1: Comparison of Direct Estimates and MRP Estimates: Aggregated Estimates of the Number of Vaccinated Individuals in ZCTAs approximating Los Angeles County, September 2020

Method	Estimate (95% Confidence Interval)
Direct Estimates: Survey Weights	6807511 (6046260, 7568759)
Indirect Estimates: Classic MRP	6738749 (6119071, 7358432)
Indirect Estimates: MRP with Spatial Term	6740796 (6118052, 7363537)

Table 1 Caption: Direct estimates represent the standard against which to compare the performance of the indirect estimate results. While both indirect estimates from the classic MRP method and the MRP method that uses a spatial term have predicted lower counts of the number of vaccinated individuals in the Los Angeles County, California area in September 2020, it is evident that using MRP with a spatial term has in this instance somewhat improved insofar as the estimate and confidence interval is closer to the direct estimates. While these are small gains, there may be potentially greater gains in analyses where the outcome has even greater spatial dependence.

⁶ The present discussion is informed by building a MRP process with spatial specifications using Integrated Nested Laplace Approximation in R (R-INLA) because of INLA's flexibility in handling spatial models, but similar processes may be undertaken using STAN.



that the spatial term has somewhat improved the estimates, which are closer in total and in variance to the direct estimates.

At this point, it is worth noting that spatial specifications for MRP are only useful if there is spatial structure to the outcome of interest (i.e., the outcome has spatial dependence). Furthermore, it is important to be cautious with the MRP method; whether a spatial term is used, MRP will provide a predicted estimate for every area and every combined demographic category based on the model prediction. It is up to the analyst to determine whether these estimates are reliable or artifacts of predictive modeling using smoothing parameters. Similarly, data availability and richness is important; the spatial model described here requires first-order neighbors, while many missing geographic areas amounting to poor data availability, or else spatial data with little spatial proximity between neighbors (i.e. limited first-order neighbors) will create potentially incorrect estimates. Consequently, the required data quality and richness is why MRP is often only used in high-income country contexts. If these requirements are ignored, it is possible smoothing parameters may potentially “over-smooth” and create estimates for unobserved areas that are unreliable. Perhaps a reasonable threshold for spatial data missingness is roughly 10 percent; for example, if we were conducting an analysis of ZCTAs in the whole of the state of California, approximately only 10 percent should have missing/unobserved values if we are interested in using spatial specifications, lest the model potentially over fit the data and produce spurious results.

4. Multilevel Regression and Poststratification: A basic introduction

Small Area Estimation procedures have been used to estimate migration patterns (e.g., [24]), homeless counts (e.g., [25-26]) and disaster response (e.g., [27]). These applications have a direct impact on policy and general demographic knowledge. More recently, in collaboration with the World Health Organization [28] Msemburi et al. [29] used SAE methods to understand the impact of COVID-19 in hard to measure parts of the world, including low- and middle-income countries, like India, where health and mortality reporting may not be centralized nor records readily available. They found that between 2020 and 2021, there was a global all-cause excess mortality of 14.9 million people due to the COVID-19 pandemic, with 84% of these deaths in Europe, South-East Asia and the Americas.

Conclusion

In this chapter we have reviewed the major importance of small area estimation in demographic and population health fields. Specifically, it is important for demographers and other population scientists to be able to estimate the counts of individuals in arbitrary geographic containers (e.g., neighborhoods or cities). This of particular interest for researchers in the Western world for non-administrative areas and to approximate areas without high quality census in areas of Africa and parts of Southeast Asia such as India. Here, we have reviewed the state-of-the-art methods for SAE such as MRP and other statistical smoothing techniques. We have concluded with applications to migration, homelessness, and COVID-19.

Acknowledgments



Partial support for this research came from a Eunice Kennedy Shriver National Institute of Child Health and Human Development research infrastructure grant, P2C HD042828, and Shanahan Endowment Fellowship, and a Eunice Kennedy Shriver National Institute of Child Health and Human Development training grant, T32 HD101442, to the Center for Studies in Demography & Ecology at the University of Washington. An NSF CAREER Award #2142964 and ARO Award #W911NF-19-1-0407. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, NSF, or ARO.

References

Jiang, J. and Rao, J. S. (2020). Robust Small Area Estimation: An Overview (2020). *Annual Review of Statistics and Its Application*, 7(1): 337-360.
<http://dx.doi.org/10.1146/annurev-statistics-031219-041212>.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to Census data. *Journal of the American Statistical Association*, 74(366): 269-277.

Besag, J., York, J., and Mollié, A. (1991). "Bayesian Image Restoration with Applications in Spatial Statistics (with Discussion)." *Annals of the Institute of Statistical Mathematics* 43: 1-59. <https://doi.org/10.1007/BF00116466>.

Riebler, A., Sørbye, S. H., Simpson, D., & Rue, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, 25(4), 1145-1165. <https://doi.org/10.1177/0962280216660421>

United States Census Bureau. (2022). Small Area Estimation.
<https://www.census.gov/topics/research/stat-research/expertise/small-area-est.html>

National Cancer Institute. (2022). Small Area Estimates for Cancer-Related Measures. <https://sae.cancer.gov/>.

United States Census Bureau. (2022). Health Insurance.
<https://www.census.gov/topics/health/health-insurance.html>

Harris, V., Caputo, J., Finley, A., Butler, B. J., Bowlick, F., & Catanzaro, P. (2021). Small-Area Estimation for the USDA Forest Service, National Woodland Owner Survey: Creating a Fine-Scale Land Cover and Ownership Layer to Support County-Level Population Estimates. *Frontiers in Forests and Global Change*, 4.
<https://doi.org/10.3389/ffgc.2021.745840>.

Jackman, S., Ratcliff, S., Mansillo, L. (2019). Small Area Estimates of Public Opinion: Model-Assisted Post-stratification of Data from Voter Advice Applications [Unpublished Manuscript].
<https://www.cambridge.org/core/membership/services/aop-file-manager/file/5c2f6ebb7cf9ee1118d11c0a>.



UNICEF. (September 6, 2022). Child Mortality. UNICEF DATA.

<https://data.unicef.org/topic/child-survival/under-five-mortality/>

Gelman, A., & Little, T. C.. (1997). Poststratification into Many Categories Using Hierarchical Logistic Regression. *Survey Methodology*, 23(2): 127-35.

Gelman, A., Lax, J., Phillips, J., Gabry, J., & Trangucci, R. (2016). Using multilevel regression and poststratification to estimate dynamic public opinion. *Unpublished manuscript, Columbia University*, 2.

Kastellec, J. P., Lax, J. R., & Phillips, J. H. (2019). Estimating state public opinion with multi-level regression and poststratification using R. *Unpublished manuscript, Princeton University*.
https://jkastellec.scholar.princeton.edu/publications/mrp_primer.

Park, D. K., Gelman, A., & Bafumi, J. (2004). "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls." *Political Analysis*, 12(4): 375-85.

Lax, J. R., & Phillips, J. H. (2009a). "Gay Rights in the States: Public Opinion and Policy Responsiveness." *American Political Science Review*, 103(3): 367-86.

Lax, J. R., & Phillips, J. H. (2009b). "How Should We Estimate Public Opinion in the States?" *American Journal of Political Science*, 53(1): 107-21.

R Core Team. (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Golemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H. (2019). "Welcome to the tidyverse." *Journal of Open Source Software*, 4(43), 1686. doi:10.21105/joss.01686.

Walker, K., Herman, M. (2022). tidycensus: Load US Census Boundary and Attribute Data as 'tidyverse' and 'sf'-Ready Data Frames. R package version 1.3.0.9000, <https://walker-data.com/tidycensus/>.

Rue, H., Martino, S., Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, 71(2): 319-392. www.r-inla.org.

Gao, Y., Kennedy, L., & Simpson, D. (2021). Treatment effect estimation with Multilevel Regression and Poststratification. *arXiv preprint arXiv:2102.10003*.

Salomon, J. A., Reinhart, A., Bilinski, A., Chua, E. J., La Motte-Kerr, W., Rönn, M. M., Reitsma, M., Morris, K. A., LaRocca, S., Farag, T., Kreuter, F., Rosenfeld, R., and



Tibshirani, R. J. (2021). The US COVID-19 Trends and Impact Survey: Continuous real-time measurement of COVID-19 symptoms, risks, protective behaviors, testing, and vaccination. *Proceedings of the National Academy of Sciences*, 118(51): e2111454118. <https://doi.org/10.1073/pnas.2111454118>.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4. <https://ggplot2.tidyverse.org>.

Garnier, S., Ross, N., Rudis, R., Camargo, P. A., Sciaini, M., Scherer, C. (2021). *viridis - Colorblind-Friendly Color Maps for R*. doi:10.5281/zenodo.

Hauer, M., and Byars, J. (2019). IRS county-to-county migration data, 1990–2010. *Demographic Research*, 40: 1153-1166.

Almquist, Z. W., Helwig, N. E., and You, Y. (2020). Connecting Continuum of Care point-in-time homeless counts to United States Census areal units. *Mathematical Population Studies* 27,1: 46-58.

Almquist, Z. W. (2020). Large-scale spatial network models: An application to modeling information diffusion through the homeless population of San Francisco. *Environment and Planning B: Urban Analytics and City Science* 47,3: 523-540.

Maas, P., Almquist, Z., Giraudy, E., & Schneider, J. W. (2020). Using social media to measure demographic responses to natural disaster: insights from a large-scale Facebook survey following the 2019 Australia Bushfires. *arXiv preprint arXiv:2008.03665*.

World Health Organization (May 5, 2022). 14.9 million excess deaths associated with the COVID-19 pandemic in 2020 and 2021 [News release]. <https://www.who.int/news/item/05-05-2022-14.9-million-excess-deaths-were-associated-with-the-covid-19-pandemic-in-2020-and-2021>

Msemburi, W., Karlinsky, A., Knutson, V., Aleshin-Guendel, S., Chatterji, S., & Wakefield, J. (2022). The WHO estimates of excess mortality associated with the COVID-19 pandemic. *Nature (London)*, 613(7942): 130–137. <https://doi.org/10.1038/s41586-022-05522-2>



[Under peer review Do not Cite or Share]

**Identity and Generosity Norms Among People Experiencing Homelessness in
Nashville, TN: A Dictator Game Experiment**

Authors

Anderson, Mary-Catherine, Stanford University

Perkins, Jessica, Vanderbilt University

Hazel, Ashley, UCSF

Jones, James, Stanford

Almquist, Zack, UW

Acknowledgements

Partial support for this research came from a Eunice Kennedy Shriver National Institute of Child Health and Human Development research infrastructure grant, P2C HD042828, and Shanahan Endowment Fellowship, and a Eunice Kennedy Shriver National Institute of Child Health and Human Development training grant, T32 HD101442, to the Center for Studies in Demography & Ecology at the University of Washington. Partial support for this research also came from an NSF Award #BCS-2028160 and an ARO Award #W911NF-19-1-0407. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, NSF, or ARO.

Abstract

Although society frequently labels people experiencing homelessness (PEH) as a single, marginalized social group—i.e., “the homeless,” it is unknown whether PEH create social identities around their shared characteristic of homelessness.

Using a dictator game, we tested whether PEH display in-group bias, a common feature of social groups, toward other PEH (an in-group) compared with housed people (an out-group). We did not observe in-group bias, suggesting that the shared vulnerability of homelessness may not necessarily foster the formation of a meaningful social group. On average, participants allocated 29% of their game endowment to recipients, suggesting that PEH are not entirely selfish and display behavior consistent with dictator game studies across various cultures. Among men, we found a positive association between game allocations the number of emotional and material supports in one's social network, and the opposite among women. Finally, we found that PEH were more generous in the dictator game when they perceived other PEH would also be more generous in an identical scenario, suggesting PEH act in accordance with perceived norms of fairness.

Over half a million people were experiencing homeless in the United States in 2019 (Department of Housing and Urban Development 2019). Although people experiencing homelessness (PEH) vary widely in their sociodemographic characteristics, behaviors, attitudes, and lived experiences, society tends to label all unhoused people as a single, marginalized social group—"the homeless." However, whether PEH themselves form group, or social, identities around their shared characteristic of homelessness remains unknown.

Past work has found that upon becoming homeless, PEH often replace more traditional social ties with ties to other PEH who provide new sources of emotional, financial, and material support (Anderson et al. 2021; Fitzpatrick, La Gory, and Ritchey

2003; Hertzberg 1992; Irwin et al. 2008; Lee, Tyler, and Wright 2010). While these social network changes may aid in providing resources critical for survival, they may also serve to strengthen a person's identity as "homeless" (Grigsby et al. 1990). Qualitative work on wanderers, and "hobos," or migratory workers, documents people's idealization of homelessness as a lifestyle choice, wherein people choose freedom over the constraints of a traditional, domesticated lifestyle (Anderson 1923). Other PEH may strongly identify as part of a homeless "street family," or a tight-knit social network composed of other PEH.

On the other hand, previously published studies have also found that some PEH are reluctant to self-identify as homeless at all due to stigma associated with being homeless, denial, or determination to immediately exit homelessness (Boydell, Goering, and Morrell-Bellai 2000; Harter et al. 2005; Snow and Anderson 1987). For example, an ethnographic study of street dwellers in Austin, TX, documented a constant struggle among PEH to mentally separate one's personal identity from the stigma that society places on homeless people as a social category (Snow and Anderson 1987:15). While people may acknowledge being unhoused, they may also cognitively distance themselves from the general homeless population. Some PEH may reject stereotypical homeless attributes—such as laziness, greediness, or addiction—when describing their own personal identities (e.g., "I'm not like 'those' homeless people"), while simultaneously exaggerating those same negative stereotypes among the general homeless population (Boydell et al. 2000; Parsell 2011; Snow and Anderson 1987).

Whether or not PEH construct social, or group, identities around their shared characteristic of homelessness has not been quantitatively assessed. Yet,

understanding whether PEH identify as members of a social group can help homeless service providers, advocates, and policymakers better understand the unique social environment of this highly vulnerable population in order to better meet their needs. To explore this research question, we first draw upon Social Identity Theory (SIT) to outline the expected behavioral patterns of social groups. Next, we form a hypothesis about how a homeless social identity could be displayed by PEH as in-group bias, where PEH favor the in-group they belong to (people experiencing homelessness) over an out-group they do not belong to (people not experiencing homelessness). We test for in-group bias using a dictator game experiment among N=173 PEH in Nashville, TN. Finally, as norms of fairness and generosity are largely understudied among PEH, we also use our dictator game experiment as an opportunity to examine these behavioral norms among a homeless population.

SOCIAL IDENTITY THEORY

Social Identity Theory (SIT) posits that human social behavior ranges on a continuum between being influenced by purely interpersonal factors and purely intergroup factors (Tajfel 1974; Turner, Brown, and Tajfel 1979). On the one end, a person's social behavior is entirely determined solely by one's individual characteristics and interpersonal relationships, and not affected by social group membership. On the other end, social behavior is entirely determined by social group membership. SIT defines a social identity as "that part of the individual's self-concept which derives from their knowledge of their membership of a social group (or groups) together with the value and emotional significance attached to that membership (Tajfel 1981)."

According to SIT, people will only identify as members of a social group when doing so serves to improve their self-image (Tajfel 1981). Moreover, people will only identify as members of a group when they identify as more similar to others like them (the in-group) compared with others who are not like them (the out-group) (Stets and Burke 2000; Turner et al. 1994). When a person identifies as a member of a social group, a meaningful distinction emerges between “us,” or the in-group, and “them,” or the out-group. People are motivated to behave according to the perceived norms of the in-group because group membership is tied to self-image. As such, they want to make their group appear superior to other social groups. Thus, in intergroup contexts, people tend to systematically favor their in-group over the out-group—a phenomenon known as in-group bias.

For example, in Tajfel’s classic minimal paradigm experiments, study participants were divided into different groups based on trivial criteria, such as performance rating for completing menial tasks. Subjects were tasked to allocate awards to other anonymous subjects, where the only information known about the other subjects was their group membership. Participants systematically favored the in-group over the out-group, even when group assignment was random. Such experiments found that even the suggestion of an in-group and an out-group is enough to elicit in-group bias even when those groups were artificially constructed (Tajfel 1970). Since the development of SIT, in-group bias has been observed with respect to almost every imaginable trait, ranging from political and religious affiliations (Eriksson 2018; Fowler and Kam 2007; Ruffle and Sosis 2006) to music and food preferences (Ben-Ner et al. 2009).

A Homeless Social Identity and In-Group Bias

Although no correlation between social identity and in-group bias is formally hypothesized in SIT, a critical review of SIT concepts concludes that “if group identification is based on a positive ingroup evaluation, and if people are motivated to maintain a positive social identity, and if ingroups are evaluated primarily in relation to relevant outgroups...one should predict an association between [social] identification and [in-group] bias (Brown 2000:753).” Therefore, we expect that when individuals identify as members of a social group, they will behave in ways that favor their own group over other groups. As such, we assume a correlation exists between social identity and in-group bias. In this study, we hypothesize that if PEH construct social identities around their shared characteristic of homelessness, then we expect PEH to display in-group bias toward other PEH as opposed to an out-group (e.g., people not experiencing homelessness).

THE DICTATOR GAME

The dictator game, a variant of the ultimatum game, has been employed extensively in both field and laboratory settings to assess the extent of prosocial (e.g. stopping to help a stranded motorist on the side of the road), or “other-regarding,” behavior, such as in-group bias and norms of fairness, generosity, and altruism (Benenson, Pascoe, and Radmore 2007; Ben-Ner et al. 2009; Henrich 2009). However, dictator game behavior has yet to be systematically investigated among PEH.

In a standard dictator game, the study participant (the dictator) is given some fixed endowment, most often money or some equivalent. Dictator games are

traditionally played with low monetary stakes, usually between \$5 and \$10 (Fowler and Kam 2007; Piff et al. 2010). The dictator is then tasked with dividing that endowment between themselves and another, anonymous person (the recipient) who initially has no currency. The recipient receives the amount of money the dictator chooses to give them while the dictator keeps the remainder. Typically, the dictator and recipient remain anonymous to each other. Upon completion of the game, the experimenter collects allocations from the dictator and then gifts them to the recipient in a manner that preserves both dictator and recipient anonymity.

Assessing Identity Using the Dictator Game

To test for in-group bias, dictator games can be manipulated wherein the experimenter provides dictators with certain information about the recipient (e.g., demographic or behavioral characteristics about the recipient). A dictator who is more generous toward recipients who share an attribute with the dictator compared with recipients who do not share this attribute (when nothing else is known about the recipient) suggests in-group bias with respect to that attribute. For example, in dictator game experiments, people have been found to display in-group bias with respect to traits such as ethnicity (Friesen et al. 2012; Whitt and Wilson 2007), political affiliation (Rand et al. 2009), and religion (Ben-Ner et al. 2009).

In this study, we examine whether housing status is a relevant dimension to elicit the formation of an in-group (i.e., PEH) in the presence of an out-group (i.e., people not experiencing homelessness). As homeless service providers already use this dimension to distinguish PEH, this dictator game experiment helps us understand whether housing

status is a meaningful dimension that elicits the formation of a social group (Winetrobe et al. 2017).

Assessing Prosocial Norms Using the Dictator Game

Although standard economic theory predicts that people will behave as selfishly as possible to maximize their economic gain (i.e., the “economic man” or “Homo economicus”), dictator game studies demonstrate that this prediction is almost universally violated (Engel 2011). In nearly all cases, people who play the dictator game appear to show regard for perceived social norms, such as fairness and generosity. Although dictator games can be used to assess norms of fairness and generosity, what dictator game behavior is considered “normative” varies substantially depending on the study and cultural context (Henrich et al. 2005). For example, a collection of studies employing the dictator game across populations in North America, South America, Africa, Australia, and Oceania found that dictator game behavior was associated with socio-cultural factors, including the structure of daily social interactions and economic factors, such as market integration (Ensminger and Henrich 2014).

Despite being embedded in Western society, homelessness in the United States represents a unique culture of its own. For example, among PEH, there is a shared understanding that “sleeping rough” refers to sleeping outside and “flying a sign” refers to panhandling (Donley and Wright 2012; Larsen, Poortinga, and Hurdle 2004).

Qualitative research has also found shared attitudes and sentiments surrounding shelter use and resource sharing among homeless populations (Snow and Anderson 1993). However, whether dictator game behavior among PEH will more closely resemble the

selfish “the economic man” or instead reflect the extent to which PEH show a perceived regard for social norms is unknown. To help fill this gap in the literature, our study aims to benchmark normative behaviors of fairness and generosity among PEH using the dictator game. We will examine norms of fairness and generosity among PEH and position our findings within existing field dictator game studies, contributing to a growing body of cross-cultural dictator game literature.

Finally, previous work has found that demographic characteristics, including age (Benenson et al. 2007), gender (Gummerum et al. 2010), ethnicity (Whitt and Wilson 2007), and education (Bekkers 2007) are also associated with how people behave in dictator game experiments. Additionally, people are likely to be more generous in the dictator game when they perceive the recipient as “worthy,” or deserving (Fong 2007). Social support has also been associated with behavior in other kinds of economic games (Fowler and Christakis 2010; O’Malley et al. 2012; Twenge et al. 2007). Dictator game behavior may therefore be influenced by factors other than identifying with a particular in-group. To account for this, we also test for associations between individual PEH characteristics—such as sociodemographic factors, perceived deservedness of the recipient, and social support—and generosity in the dictator game. We describe these measures in more detail in the following section.

METHODS

Study Participants and Recruitment

We recruited 173 participants via convenience sampling at two brick-and-mortar homeless service sites and three street locales in Nashville, TN, from July 2019 to

October 2019 (Anderson et al. 2021). At this time, Nashville-Davidson County had a homeless population of 1,986 (Department of Housing and Urban Development 2019:201) Any person aged 18 or older who self-identified as homeless could participate and was interviewed immediately following recruitment. During the consent process, participants were informed they would be participating in an experiment and would be administered a survey after gameplay. This study was approved by the X University IRB.

Dictator Game Experiment

To test for in-group bias, participants played a single dictator game using 8 single-ride Nashville MTA bus passes as the fixed endowment to keep or gift to an anonymous recipient. The MTA bus system in Nashville is the primary mode of transportation besides walking among PEH. Bus passes, therefore, represent a relevant currency to PEH. At the time of our study, a single-ride bus pass cost \$1.70. Each participant was given 8 bus passes for gameplay because dictator games are traditionally played with low monetary stakes, usually between \$5 and \$10 (Fowler and Kam 2007; Piff et al. 2010). We also gave participants an even number of bus passes because we expected many participants to allocate bus passes in pairs, representing a round-trip journey.

Each participant was randomly assigned to one of three game scenarios using a random number generator without replacement. In all scenarios, participants were told that the recipient in the game was someone in Nashville who relied on the bus for transportation. The participant was then told that the recipient was either (1) housed (2)

unhoused, or (3) no information was provided about the housing status of the recipient, which served as a control. We defined “housed” recipients as people who had a home, and therefore were not currently experiencing homelessness and “unhoused” recipients as people who, like participants, were also currently experiencing homelessness.

Participants were instructed to divide the passes between themselves and the anonymous recipient. Participants could allocate between 0 and 8 bus passes to the recipient and keep the rest following study participation. The participant (dictator) remained anonymous to the recipient and vice versa. To ensure game comprehension, participants were asked to repeat the information they were told about the recipient to the experimenter and asked a series of questions to make sure they understood the game rules.

Participants were then given an envelope containing the 8 bus passes. They were instructed to place however many bus passes, if any, they wanted to give the recipient into a separate envelope and to seal the envelope, even if they chose to allocate nothing. They were told to drop the sealed envelope into a large envelope, which would be distributed to recipients upon study completion. Participants’ behavior in the dictator game is highly sensitive to perceived social cues on how one “should” behave, such as perceived observation or judgement by the experimenter (Haley and Fessler 2005). Therefore, the experimenter either left the room or turned around and stepped aside while the participant divided their bus passes to ensure anonymity during gameplay. The participant was instructed to notify the experimenter after they completed this step.

Each participant's donation envelope was identified by the same randomly generated number that was used to assign them to one of the three treatment groups. By avoiding associating participant names with the donation envelopes, it was clear to participants that the number of bus passes they allocated to the recipient, if any, would remain anonymous to both the experimenter and recipient. This same randomly generated number also appeared on participants' surveys. Upon study completion, participants were given a \$5 gift card of their choosing. To ensure this would not affect the relative stakes of the dictator game, participants were not informed of the gift card amount until after playing the dictator game. Once the game was over, the experimenter administered a survey to participants.

Following study completion, recipients' bus passes were given to homeless service providers in Nashville. These service providers distributed these bus passes to relevant recipients (i.e., housed people, unhoused people, or anyone who takes the bus, regardless of housing status), in the quantities in which they were allocated by participants. Thus, no deception was used in this study.

Survey-Based Data Collection

The experimenter administered a survey to participants after dictator gameplay to collect sociodemographic data including age, gender, ethnicity, education, frequency of bus use, and social support from one's social network. As PEH have been widely documented to rely on members of their social networks for emotional, material, and financial support (Almquist 2020; La Gory, Ritchey, and Fitzpatrick 1991; Lee et al. 2010; Reitzes et al. 2011), we also elicited three indicators of social support within

participants' social networks. We asked participants to think back over the past 30 days and name any individuals who provided them with any kind of (1) financial (2) emotional or (3) material support. We then created variables representing the total number of individuals nominated for each type of support as well as total network size.

Additionally, prior work has linked homelessness duration and propensity to personally reject versus embrace one's identity as "a homeless person" (Grigsby et al. 1990; Snow and Anderson 1987). Therefore, we also had a survey question eliciting participants' best estimates of their lifetime duration. Additionally, when the Department of Housing and Urban Development (HUD) conducts their annual point-in-time (PIT) homeless count, they classify PEH who are residing in shelters as 'sheltered' versus PEH who are not residing in shelters as 'unsheltered' based on where PEH are found sleeping on the night in January when they are counted. To elicit a general indicator of sheltered status, we asked participants to report the number of nights they had spent at a shelter in the past 30 nights. We classified a participant as sheltered if they spent ≥ 14 nights at a shelter, and unsheltered otherwise.

Finally, we wanted to capture participants' perception of fairness and generosity among PEH in Nashville. If a norm of fairness exists within a particular context, people in that context would be expected to act in accordance with that norm. Therefore, we asked participants two questions about norm perception. The first captured how deserving the participant felt the recipient was of free bus passes compared with the participant (Perceived Deservedness of 2nd Player). The second captured how many bus passes, on average, the participant thought other PEH playing the game would allocate in the same game scenario (Perceived PEH Donations).

Statistical Analyses

Testing for a Social Identity: Poisson Regression Models

We began our analysis with a generalized linear model for count data. We initially tested for a treatment group effect on bus pass allocations using a simple Poisson regression model. We then explored additional Poisson regression models to understand the association between participant attributes and participant's perceptions of the recipient and the outcome of the number of bus passes allocated in the game. We built separate models for each treatment group as well as a complete model with the total population under study. We tested four different complete models under key hypotheses (i.e., Sociodemographic Factors, Social Network Composition, and Perceived Deservedness and Generosity of Other PEH). Finally, we used these multivariate Poisson regression models to examine whether (1) PEH displayed in-group bias to other PEH as measured by the dictator game and (2) individual characteristics of PEH and recipients were associated with the quantity of bus pass allocations. All factors that were statistically significant with a p-value <0.05 in at least one of the treatment groups in the univariate models were included in the complete models except for ethnicity and network size. Ethnicity was not significant in the multivariate analyses and its removal resulted in better model fit (model selection performed based on AIC and BIC; (Akaike 1974; Kass and Wasserman 1995)). Network size was highly correlated with both emotional and material support and could not be included in the model (Anderson et al. 2021). Lastly, we excluded participants identifying as non-binary ($n = 3$) in all models due to limited sample size.

We employ an additive model building structure where we use AIC/BIC to adjudicate the best model. In the multivariate models, Model 1 included only sociodemographic characteristics, while in Model 2, we added social network characteristics (i.e., number of emotional and material supports) and variables encapsulating perceptions about PEH. Given the large body of literature citing differences in social network composition between men and women (Benenson 1990; Efferson, Lalive, and Fehr 2008; McLaughlin et al. 2010), we then examined the potential interaction effect of gender and number of emotional supports and material supports on the number of bus passes allocated in the game in Models 3 and 4.

Characteristics Associated with Giving Nothing in Dictator Game

Because we were also interested in using the dictator game to examine norms of fairness and generosity in a homeless population, we constructed an additional binary logistic regression model to test for patterns among participants who gave nothing in the dictator game. We employed this model as a descriptive check for associations between the significant individual-level characteristics in the multivariate models (Gender, Lifetime Homelessness Duration, Rides Bus Every Day, Number of Emotional Supports, and Perceived Deservedness of 2nd Player, and Perceived PEH Donations) and the probability of the outcome of a participant giving zero bus passes (coded as 1) compared with those participants who donated at least one bus pass (coded as 0) in the dictator game. Again, we excluded participants identifying as non-binary ($n = 3$) in this model due to limited sample size.

RESULTS

The distribution of participants across study factors is presented in Table 1.

(Table 1 About Here)

There were 117 (67.6%) men and 53 (30.6%) women comprising our sample population. The mean age of study participants was ~45 years old. Additionally, the mean number of emotional and material support contacts in participants' social networks was 1.3 and 1.8, respectively. The distribution of bus pass allocations in the dictator game stratified by treatment groups is presented in Figure 1.

(Figure 1 About Here).

The mean number of bus passes for participants assigned to the control, unhoused, and housed treatment groups was 2.6, 2.1, and 2.3, respectively. A univariate Poisson regression revealed that neither the unhoused treatment group ($\exp(\beta) = 0.83$, 95% CI (-0.42, 0.10)) nor the housed treatment group ($\exp(\beta) = 0.88$, 95% CI (-0.37, 0.10)) differed significantly from the control group in their bus pass allocations in the dictator game. However, we found that individual characteristics of the dictator were associated with the number of bus passes allocated. Estimates of the univariate association between passes allocated and all potential factors are provided in Supplementary Tables S1 and S2.

Sociodemographic, Bus Use, and Homelessness Duration Correlates of Generosity

Our multivariable Poisson regression models are presented in Tables 2 and 3. Model 1 examined associations between sociodemographic characteristics, bus use frequency, homelessness duration, and the number of bus passes allocated in the dictator game. We found no difference in the amount of bus passes allocated across treatment groups, genders, or educational tertiles. We found that PEH who had experienced 5–10 years of homelessness over their lifetimes allocated 34% fewer bus passes, on average, than PEH who experienced homelessness for 1 year or less ($\exp(\beta) = 0.66$, $p < 0.01$). PEH who experienced homelessness 1—5 years and 10+ years tended to allocate fewer bus passes than PEH who experienced homelessness for 1 year or less, though not significantly so. Finally, PEH who rode the bus more than half the days of a typical month allocated, on average, 36% fewer bus passes in the dictator game than PEH who rode the bus less than half the days of a typical month ($\exp(\beta) = 0.64$, $p < 0.001$).

Social Network Composition and Perception Correlates of Generosity

Model 2 included social network composition and perceived norms measures. There was no association between social network size or the number of material supports and the number of bus passes allocated in the dictator game. However, as the number of participants' emotional supports increased by one, there was an 11% increase in the number of bus passes allocated in the game ($OR = 1.11$, $p < 0.05$). Additionally, participants who felt that the recipient was equally or more deserving of

free bus passes compared with the participant allocated 39% more bus passes on average than PEH who felt the anonymous recipient was less deserving than themselves (OR = 1.39, $p < 0.01$). Finally, there was an association between participants' perceptions of how generous they thought other PEH would be if they were playing the game and the number of bus passes participants allocated. For each additional bus pass participants perceived other PEH would hypothetically give if they were playing the game, PEH gave 9% more bus passes in the game (OR = 1.09, $p < 0.01$).

(Table 2 About Here)

Gender-Emotional and Material Support Interaction Effects

Model 3 included an interaction term between participant gender and emotional support in participants' social networks. Among men, for each additional emotional support, there was a 39% increase in the amount of bus passes allocated compared with women having an additional emotional support (OR = 1.39, $p < 0.001$). Model 4 included an interaction effect between gender and material support in participants' social networks. Similarly, among men, for each additional material support, there was a 28% increase in the amount of bus passes allocated compared with women having an additional material support (OR = 1.28, $p < 0.001$).

(Table 3 About Here).

The relationship between number of emotional supports and number of bus passes allocated in the game stratified by gender is presented in Figure 2.

(Figure 2 About Here)

Logistic Regression on Factors Associated with Giving Nothing in Dictator Game

Our logistic regression model (Model 5) examining associations between participant characteristics and giving nothing in the dictator game is presented in Table 4.

(Table 4 about here)

Among all our sociodemographic predictors included in Models 1–4, only whether the participant rode the bus everyday was significant. We found that people who rode the bus everyday were around 4 times more likely to donate nothing in the dictator game (OR = 1.40, $p < 0.001$). Additionally, we found that for each bus pass participants perceived that other PEH would donate in an identical game scenario, there was a 30% decrease in the odds of donating nothing in the dictator game.

DISCUSSION

Social Identity Among PEH

We did not observe in-group bias in the dictator game among our study population of PEH. However, this finding does not necessarily mean that PEH lack social identities surrounding homelessness. Our experiment was predicated on the

assumption that PEH would relate more to PEH than housed people, and thus would be more likely to form a social identity around their shared homeless experience. However, the lack of in-group bias in our study could be because PEH actually form social identities around categories more salient than homelessness itself.

Our work provides a valuable baseline for future social identity research among homeless populations, such as testing for social identities constructed around homeless subgroups (e.g., gender). Extensive research finds that men and women endure markedly different homeless experiences, with women frequently displaying higher rates of victimization and premature mortality (Auerswald, Lin, and Parriott 2016; Romaszko et al. 2017). It is therefore possible that women have different expectations and norms surrounding their formation of social groups, or communities. A modified version of our experiment where study participants are told that either (1) the recipient is of the same gender or (2) a different gender than the participant could be employed as a simple test of this hypothesis. More detailed metrics of social identity, such as measures of how strongly people identify with certain PEH subgroups, could also be collected to tease apart more nuanced social identities adopted by PEH.

Our work has practical implications for advocacy and service efforts targeted at PEH. First, our finding that PEH do not display increased prosocial behavior toward other PEH compared with housed people suggests that PEH in Nashville may not necessarily consider themselves interchangeable members of a social group with similar values, beliefs, behaviors, and relationships to one another. That is, the shared vulnerability of homelessness does not necessarily foster formation of a meaningful

social group. Service providers, advocates, and allies should therefore take care in the language they use in supporting PEH.

Norms of Fairness and Generosity Among PEH

In our study, PEH allocated an average of 2.35 (29.35%) of their bus passes to recipients in the dictator game. Our results are consistent with past dictator game studies. For example, according to a meta-analysis of 131 scientific papers encompassing 616 dictator game treatments, dictators allocated 28.35% of their game currency to recipients on average (Engel 2011).

To further contextualize our findings, it is useful to compare our results to other field dictator game studies conducted outside laboratory settings. This is because dictator game experiments have mostly been conducted in laboratory settings with university students as participants. On average, participants in these settings tend to allocate ~20% of their endowment to recipients, which has come to be considered a universal norm of dictator game behavior (Levitt and List 2007). However, field dictator games in Africa, South America, Papua New Guinea, Siberia, Oceania, and North America reveal context-specific variation in dictator game behavior, suggesting that different cultural contexts may give rise to different norms surrounding fairness and generosity (Ensminger and Henrich 2014).

For example, the Hadza of Tanzania were found to give a mean 26% of their dictator game endowment to recipients, with a modal offer of 0% (Marlowe 2014). The Sursurunga of Papua New Guinea gave an average of 41% of their dictator game allocations to recipients with modal offers of 50% (Bolyanatz 2014). The Shuar of

Ecuador gave, on average, 35% of their dictator game allocations to recipients, with modal offers at both 20% and 50% (Barrett and Haley 2014). Among a non-student population in rural North America, mean and modal dictator game allocations were 47% and 50%, respectively (Ensminger and Cook 2014). Our finding that PEH, on average, allocated ~29% of their bus passes to recipients in the dictator game contributes to a growing body of literature that finds non-student populations display higher levels of generosity in the dictator game compared with student populations (Ensminger and Cook 2014).

Only 19.9% (n=31) of PEH in our study gave nothing. Our study demonstrates that even among a Western resource-poor group, such as PEH, people do not behave entirely selfishly. Our results also suggest a trend toward fairness among PEH. As seen in Figure 1, there is a trend toward a bimodal distribution in dictator game allocations across all 3 treatment groups, with two apparent modes at 2 and 4 bus passes. This suggests a norm of fair-mindedness among PEH, as an even number of single-ride bus passes allows the recipient to complete at least one round-trip journey.

Our findings can help in combating negative stereotypes against PEH. For example, “the homeless” are frequently labeled as irresponsible, dangers to society, selfish, or greedy (Knecht and Martinez 2009; Snow and Anderson 1987). Future work focusing on the norms of generosity, and fairness among PEH may therefore serve to build empathy among the general population for PEH. Re-framing homelessness as a circumstance rather than a character flaw can help facilitate service and support efforts at local and national levels.

Characteristics Associated with Generosity in the Dictator Game

We found that individual demographic characteristics of participants, including lifetime homelessness duration and monthly bus use, were associated with the number of bus passes participants allocated in the dictator game. While evidence from this study suggests that PEH value their bus passes as an important resource (as those who used the bus more frequently gave fewer bus passes in the dictator game), PEH will still share this resource with recipients who are perceived to also need bus passes.

Additionally, we found that homelessness duration was associated with dictator game allocations. PEH who had experienced between 5–10 years of homelessness over their lifetimes allocated ~31-34% fewer bus passes to recipients compared with PEH who had experienced homelessness less than 5 years and more than 10 years (Models 1-3). People who had experienced homelessness between 1–5 years and 10+ years also allocated less than participants who had only been homeless 1 year or less, though not significantly so. Past work has found that as homelessness duration increases, PEH tend to replace social ties to housed people with ties to unhoused people. In some cases, long-term PEH develop “cliques,” or small, tight-knit social circles with other long-term PEH (Osborne 2002). Therefore, it could be that as homelessness duration increases, people become less likely to share resources with anonymous recipients, as in the dictator game, and more likely to share resources with members of their personal social networks.

We also found that participants were more generous in the dictator game when (1) they perceived the recipient in the dictator game to be equally or more deserving of free bus passes than the participant and (2) they perceived that other PEH in an

identical dictator game scenario would also be more generous. The finding that PEH are more generous toward people they consider deserving is consistent with past studies that find a positive association between the dictator's perception of the recipient's "worthiness" and allocations in the dictator game (Fong and Luttmer 2011). Additionally, our Poisson regression models revealed that PEH allocated between 8-9% more bus passes for each additional bus pass they thought other PEH would hypothetically give in an identical dictator game. Our logistic regression model also revealed that for each bus pass participant thought other PEH would allocate in an identical dictator game scenario, participants were 30% less likely to give nothing in a dictator game (i.e. more likely to donate something). This suggests that PEH aim to act in agreement with a perceived norm of fairness and generosity among other PEH. Our findings build upon decades of work in social psychology showing that people tend to follow perceived norms, further demonstrating that PEH likely act in accordance with the perceived behavior of others (Miller and Prentice 2016; Prentice and Paluck 2020).

Finally, our study suggests that social supports function differently for men and women in how they influence generosity. Interestingly, we found that larger material and emotional support networks were associated with greater increases in dictator game generosity among men than among women (Models 3 and 4). A large body of anthropological work finds that men are more likely than women to normatively engage in costly signaling acts that require greater public sacrifice (Bird and Smith 2005; Bird, Smith, and Bird 2001; Sosis 2000). Conversely, women tend to engage in more subtle signaling acts, such as investing in relationships with close friends and acquaintances (Bird, Ready, and Power 2018). In our experiment, it is possible that male PEH with

larger social support networks feel compelled to “pay it forward” to an anonymous third party as a reputation mechanism, signaling their willingness to contribute and cooperate. In contrast, women with larger social support networks may be more likely to share resources with known social network members as opposed to anonymous recipients. Future work into prosocial behavior among homeless populations is needed to improve our understanding of resource allocation across genders among PEH.

Limitations

Our study has several limitations. First, our study relied on the assumption of a correlation between social identity and in-group bias. However, the absence of in-group bias in our experiment does not necessarily indicate the absence of social identities among PEH. While we found no evidence of a social identity constructed around homelessness itself, future work should aim to test for other social identities constructed around certain PEH subgroups, such as gender groups.

Second, dictator games also carry several inherent limitations in that participants may be more generous than they would actually be in real-world scenarios because they feel judged by the experimenter. A perceived lack of anonymity with respect to the experimenter and feelings of being watched are well-documented shortcomings of dictator game methodologies (Haley and Fessler 2005; Lamba and Mace 2010). To limit this bias as much as possible, the experimenter left the room or stepped away from participants while they were allocating their bus passes in the dictator game. We also used randomly generated numbers to identify participants instead of names, further preserving anonymity. Finally, as our study population was small, our findings may not

be generalizable to all homeless populations due to convenience sampling limitations; however our sample as percentage of homeless measured by the HUD Continuum of Care (Almquist, Helwig, and You 2020) represents about one third of the total unsheltered PEH in Nashville, TN (Department of Housing and Urban Development 2020). Furthermore, our sample population was demographically comparable to the population comprising the Nashville-Davidson County PIT count across years 2016-2020 (Anderson et al. 2021).

Summary

Our study produced several key findings. First, we found no evidence of in-group bias in our dictator game experiment. Rather, we found that PEH display behavior consistent with what is typically observed in dictator games among other populations, which is that people are not entirely selfish. On average, participants donated around ~29% of their dictator game allocations to recipients. In addition to testing for in-group bias, we also used our dictator game experiment to investigate norms of fairness and generosity among PEH. We found that certain characteristics of PEH, including gender, the number of emotional supports in one's social network, and the perceived generosity of other PEH in an identical game scenario, were associated with the number of bus passes that participants allocated in the dictator game. Although only 31 out of 173 participants (19.9%) gave nothing in the game, we found that participants who perceived that other PEH would be less generous in an identical dictator game scenario were more likely to give nothing to recipients.

REFERENCES

- Akaike, H. 1974. "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control* 19(6):716–23. doi: 10.1109/TAC.1974.1100705.
- Almquist, Zack W. 2020. "Large-Scale Spatial Network Models: An Application to Modeling Information Diffusion through the Homeless Population of San Francisco." *Environment and Planning B: Urban Analytics and City Science* 47(3):523–40. doi: 10.1177/2399808318785375.
- Almquist, Zack W., Nathaniel E. Helwig, and Yun You. 2020. "Connecting Continuum of Care Point-in-Time Homeless Counts to United States Census Areal Units." *Mathematical Population Studies* 27(1):46–58. doi: 10.1080/08898480.2019.1636574.
- Anderson, Mary-Catherine, Ashley Hazel, Jessica M. Perkins, and Zack W. Almquist. 2021. "The Ecology of Unsheltered Homelessness: Environmental and Social-Network Predictors of Well-Being among an Unsheltered Homeless Population." *International Journal of Environmental Research and Public Health* 18(14):7328. doi: 10.3390/ijerph18147328.
- Anderson, Nels. 1923. *The Hobo: The Sociology of the Homeless Man*. Chicago, Illinois: The University of Chicago Press.
- Auerswald, Colette L., Jessica S. Lin, and Andrea Parriott. 2016. "Six-Year Mortality in a Street-Recruited Cohort of Homeless Youth in San Francisco, California." *PeerJ* 4:e1909. doi: 10.7717/peerj.1909.
- Barrett, H. Clark, and Kevin J. Haley. 2014. "Economic Game Behavior Among the Shuar." Pp. 259–74 in *Experimenting with Social Norms: Fairness and Punishment in Cross-Cultural Perspective*, edited by J. Ensminger and J. Henrich. New York, New York, USA: Russell Sage Foundation.
- Bekkers, Rene. 2007. "Measuring Altruistic Behavior in Surveys: The All-or-Nothing Dictator Game." *Survey Research Methods* 1(3):139–44. doi: 10.18148/srm/2007.v1i3.54.
- Benenson, Joyce F. 1990. "Gender Differences in Social Networks." *The Journal of Early Adolescence* 10(4):472–95. doi: 10.1177/0272431690104004.
- Benenson, Joyce F., Joanna Pascoe, and Nicola Radmore. 2007. "Children's Altruistic Behavior in the Dictator Game." *Evolution and Human Behavior* 28(3):168–75. doi: 10.1016/j.evolhumbehav.2006.10.003.
- Ben-Ner, Avner, Brian P. McCall, Massoud Stephane, and Hua Wang. 2009. "Identity and In-Group/out-Group Differentiation in Work and Giving Behaviors: Experimental Evidence." *Journal of Economic Behavior and Organization* 72(1):153–70. doi: 10.1016/j.jebo.2009.05.007.

- Bird, Rebecca Bliege, Elspeth Ready, and Eleanor A. Power. 2018. "The Social Significance of Subtle Signals." *Nature Human Behaviour* 2(7):452–57. doi: 10.1038/s41562-018-0298-3.
- Bird, Rebecca Bliege, and Eric Alden Smith. 2005. "Signaling Theory, Strategic Interaction, and Symbolic Capital." *Current Anthropology* 46(2):221–48. doi: 10.1086/427115.
- Bird, Rebecca Bliege, Eric Alden Smith, and Douglas W. Bird. 2001. "The Hunting Handicap: Costly Signaling in Human Foraging Strategies." *Behavioral Ecology and Sociobiology* 50(1):9–19. doi: 10.1007/s002650100338.
- Bolyanatz, Alexander H. 2014. "Economic Experimental Game Results from the Sursurunga of New Ireland, Papua New Guinea." Pp. 275–307 in *Experimenting with Social Norms: Fairness and Punishment in Cross-Cultural Perspective*, edited by J. Ensminger and J. Henrich. New York, New York, USA: Russell Sage Foundation.
- Boydell, Katherine M., Paula Goering, and Tammy L. Morrell-Bellai. 2000. "Narratives of Identity: Re-Presentation of Self in People Who Are Homeless." *Qualitative Health Research* 10(1):26–38. doi: 10.1177/104973200129118228.
- Brown, Rupert. 2000. "Social Identity Theory: Past Achievements, Current Problems and Future Challenges." *European Journal of Social Psychology* 30(6):745–78. doi: 10.1002/1099-0992(200011/12)30:6<745::AID-EJSP24>3.0.CO;2-O.
- Department of Housing and Urban Development. 2019. *HUD 2019 Continuum of Care Homeless Assistance Programs Homeless Populations and Subpopulations*.
- Department of Housing and Urban Development. 2020. *HUD 2020 Continuum of Care Homeless Assistance Programs Homeless Populations and Subpopulations*.
- Donley, Amy M., and James D. Wright. 2012. "Safer Outside: A Qualitative Exploration of Homeless People's Resistance to Homeless Shelters." *Journal of Forensic Psychology Practice* 12(4):288–306. doi: 10.1080/15228932.2012.695645.
- Efferson, Charles, Rafael Lalive, and Ernst Fehr. 2008. "The Coevolution of Cultural Groups and Ingroup Favoritism." *Science* 321(5897):1844–49. doi: 10.1126/science.1155805.
- Engel, Christoph. 2011. "Dictator Games: A Meta Study." *Experimental Economics* 14(4):583–610. doi: 10.1007/s10683-011-9283-7.
- Ensminger, Jean, and Kathleen Cook. 2014. "Prosociality in Rural America: Evidence from Dictator, Ultimatum, Public Goods, and Trust Games." Pp. 445–64 in *Experimenting with Social Norms: Fairness and Punishment in Cross-Cultural Perspective*. New York, New York, USA: Russell Sage Foundation.

- Ensminger, Jean, and Joseph Henrich, eds. 2014. *Experimenting with Social Norms: Fairness and Punishment in Cross-Cultural Perspective*. New York, New York, USA: Russell Sage Foundation.
- Eriksson, Kimmo. 2018. "Republicans Value Agency, Democrats Value Communion." *Social Psychology Quarterly* 81(2):173–84. doi: 10.1177/0190272518771060.
- Fitzpatrick, Kevin M., Mark E. La Gory, and Ferris J. Ritchey. 2003. "Factors Associated with Health-Compromising Behavior Among the Homeless." *Journal of Health Care for the Poor and Underserved* 14(1):70–86. doi: 10.1353/hpu.2010.0829.
- Fong, Christina M. 2007. *Evidence from an Experiment on Charity to Welfare Recipients: Reciprocity, Altruism and the Empathic Responsiveness Hypothesis*. Vol. 117. doi: 10.1111/j.1468-0297.2007.02076.x.
- Fong, Christina M., and Erzo F. P. Luttmer. 2011. "Do Fairness and Race Matter in Generosity? Evidence from a Nationally Representative Charity Experiment." *Journal of Public Economics* 95(5–6):372–94. doi: 10.1016/j.jpubeco.2010.07.010.
- Fowler, James H., and Nicholas A. Christakis. 2010. "Cooperative Behavior Cascades in Human Social Networks." *Proceedings of the National Academy of Sciences of the United States of America* 107(12):5334–38. doi: 10.1073/pnas.0913149107.
- Fowler, James H., and Cindy D. Kam. 2007. *Beyond the Self: Social Identity, Altruism, and Political Participation*. Vol. 69. doi: 10.1111/j.1468-2508.2007.00577.x.
- Friesen, Jane, Jasmina Arifovic, Stephen C. Wright, Andreas Ludwig, Lisa Giamo, and Gamze Baray. 2012. "Ethnic Identity and Discrimination among Children." *Journal of Economic Psychology* 33(6):1156–69. doi: 10.1016/j.joep.2012.08.003.
- Grigsby, Charles, Donald Baumann, Steven E. Gregorich, and Cynthia Roberts-Gray. 1990. "Disaffiliation to Entrenchment: A Model for Understanding Homelessness." *Journal of Social Issues* 46(4):141–56. doi: 10.1111/j.1540-4560.1990.tb01803.x.
- Gummerum, Michaela, Yaniv Hanoch, Monika Keller, Katie Parsons, and Alegra Hummel. 2010. "Preschoolers' Allocations in the Dictator Game: The Role of Moral Emotions." *Journal of Economic Psychology* 31(1):25–34. doi: 10.1016/j.joep.2009.09.002.
- Haley, Kevin J., and Daniel M. T. Fessler. 2005. "Nobody's Watching?" *Evolution and Human Behavior* 26(3):245–56. doi: 10.1016/j.evolhumbehav.2005.01.002.
- Harter, Lynn M., Charlene Berquist, B. Scott Titsworth, David Novak, and Tod Brokaw. 2005. "The Structuring of Invisibility among the Hidden Homeless: The Politics of

Space, Stigma, and Identity Construction.” *Journal of Applied Communication Research* 33(4):305–27. doi: 10.1080/00909880500278079.

Henrich, Joseph. 2009. “The Evolution of Costly Displays, Cooperation and Religion: Credibility Enhancing Displays and Their Implications for Cultural Evolution.” *Evolution and Human Behavior* 30(4):244–60. doi: 10.1016/J.EVOLHUMBEHAV.2009.03.005.

Hertzberg, Edwina L. 1992. “The Homeless in the United States: Conditions, Typology and Interventions.” *International Social Work* 35(2):149–61. doi: 10.1177/002087289203500205.

Irwin, Jay, Mark LaGory, Ferris Ritchey, and Kevin Fitzpatrick. 2008. “Social Assets and Mental Distress among the Homeless: Exploring the Roles of Social Support and Other Forms of Social Capital on Depression.” *Social Science and Medicine* 67(12):1935–43. doi: 10.1016/j.socscimed.2008.09.008.

Kass, Robert E., and Larry Wasserman. 1995. “A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion.” *Journal of the American Statistical Association* 90(431):928–34. doi: 10.1080/01621459.1995.10476592.

Knecht, Tom, and Lisa M. Martinez. 2009. “Humanizing the Homeless: Does Contact Erode Stereotypes?” *Social Science Research* 38(3):521–34. doi: 10.1016/j.ssresearch.2009.01.009.

La Gory, Mark, Ferris Ritchey, and Kevin Fitzpatrick. 1991. “Homelessness and Affiliation.” *Sociological Quarterly* 32(2):201–18. doi: 10.1111/j.1533-8525.1991.tb00353.x.

Lamba, Shakti, and Ruth Mace. 2010. “People Recognise When They Are Really Anonymous in an Economic Game.” *Evolution and Human Behavior* 31(4):271–78. doi: 10.1016/j.evolhumbehav.2010.02.002.

Larsen, Larissa, Ernie Poortinga, and Donna E. Hurdle. 2004. “Sleeping Rough: Exploring the Differences between Shelter - Using and Non-Shelter - Using Homeless Individuals.” *Environment and Behavior* 36(4):578–91. doi: 10.1177/0013916503261385.

Lee, Barrett A., Kimberly A. Tyler, and James D. Wright. 2010. “The New Homelessness Revisited.” *Annual Review of Sociology* 36:501–21. doi: 10.1146/annurev-soc-070308-115940.

Levitt, Steven D., and John A. List. 2007. “What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?” *Journal of Economic Perspectives* 21(2):153–74. doi: 10.1257/jep.21.2.153.

- Marlowe, Frank W. 2014. "Better to Receive Than to Give: Hadza Behavior in Three Experimental Economic Games." Pp. 161–76 in *Experimenting with Social Norms: Fairness and Punishment in Cross-Cultural Perspective*, edited by J. Ensminger and J. Henrich. New York, New York, USA: Russell Sage Foundation.
- McLaughlin, Deirdre, Dimitrios Vagenas, Nancy A. Pachana, Nelufa Begum, and Annette Dobson. 2010. "Gender Differences in Social Network Size and Satisfaction in Adults in Their 70s." *Journal of Health Psychology* 15(5):671–79. doi: 10.1177/1359105310368177.
- Miller, Dale T., and Deborah A. Prentice. 2016. "Changing Norms to Change Behavior." *Annual Review of Psychology* 67(1):339–61. doi: 10.1146/annurev-psych-010814-015013.
- O'Malley, A. James, Samuel Arbesman, Darby Miller Steiger, James H. Fowler, and Nicholas A. Christakis. 2012. "Egocentric Social Network Structure, Health, and pro-Social Behaviors in a National Panel Study of Americans." *PloS One* 7(5):e36250. doi: 10.1371/journal.pone.0036250.
- Osborne, Randall E. 2002. "'I May Be Homeless, But I'm Not Helpless': The Costs and Benefits of Identifying with Homelessness." *Self and Identity* 1(1):43–52. doi: 10.1080/152988602317232795.
- Parsell, Cameron. 2011. "Homeless Identities: Enacted and Ascribed." *British Journal of Sociology* 62(3):442–61. doi: 10.1111/j.1468-4446.2011.01373.x.
- Piff, Paul K., Michael W. Kraus, Stéphane Côté, Bonnie Hayden Cheng, and Dacher Keltner. 2010. "Having Less, Giving More: The Influence of Social Class on Prosocial Behavior." *Journal of Personality and Social Psychology* 99(5):771–84. doi: 10.1037/a0020092.
- Prentice, Deborah, and Elizabeth Levy Paluck. 2020. "Engineering Social Change Using Social Norms: Lessons from the Study of Collective Action." *Current Opinion in Psychology* 35:138–42. doi: 10.1016/j.copsyc.2020.06.012.
- Rand, David G., Thomas Pfeiffer, Anna Dreber, Rachel W. Sheketoff, Nils C. Wernerfelt, and Yochai Benkler. 2009. "Dynamic Remodeling of In-Group Bias during the 2008 Residential Election." *Proceedings of the National Academy of Sciences of the United States of America* 106(15):6187–91. doi: 10.1073/pnas.0811552106.
- Reitzes, Donald C., Timothy J. Crimmins, Johanna Yarbrough, and Josie Parker. 2011. "Social Support and Social Network Ties among the Homeless in a Downtown Atlanta Park." *Journal of Community Psychology* 39(3):274–91. doi: 10.1002/jcop.20431.
- Romaszko, Jerzy, Iwona Cymes, Ewa Dragańska, Robert Kuchta, and Katarzyna Glińska-Lewczuk. 2017. "Mortality among the Homeless: Causes and

Meteorological Relationships” edited by J. Shaman. *PLoS ONE* 12(12):e0189938. doi: 10.1371/journal.pone.0189938.

Ruffle, Bradley J., and Richard Sosis. 2006. “Cooperation and the In-Group-out-Group Bias: A Field Test on Israeli Kibbutz Members and City Residents.” *Journal of Economic Behavior and Organization* 60(2):147–63. doi: 10.1016/j.jebo.2004.07.007.

Snow, David A., and Leon Anderson. 1987. “Identity Work Among the Homeless: The Verbal Construction and Avowal of Personal Identities.” *American Journal of Sociology* 92(6):1336–71. doi: 10.1086/228668.

Snow, David A., and Leon Anderson. 1993. *Down on Their Luck: A Study of Homeless Street People*. 1st ed. London, England: University of California Press, Ltd.

Sosis, Richard. 2000. “Costly Signaling and Torch Fishing on Ifaluk Atoll.” *Evolution and Human Behavior* 21(4):223–44. doi: 10.1016/S1090-5138(00)00030-1.

Stets, Jan E., and Peter J. Burke. 2000. “Identity Theory and Social Identity Theory.” *Social Psychology Quarterly* 63(3):224–37. doi: 10.2307/2695870.

Tajfel, H. 1970. “Experiments in Intergroup Discrimination.” *Scientific American* 223(5):96–102. doi: 10.1038/scientificamerican1170-96.

Tajfel, Henri. 1974. “Social Identity and Intergroup Behaviour.” *Social Science Information* 13(2):65–93. doi: 10.1177/053901847401300204.

Tajfel, Henri. 1981. *Human Groups and Social Categories: Studies in Social Psychology*. Cambridge, United Kingdom: Cambridge University Press.

Turner, J. C., R. J. Brown, and H. Tajfel. 1979. “Social Comparison and Group Interest in Ingroup Favouritism.” *European Journal of Social Psychology* 9(2):187–204. doi: 10.1002/ejsp.2420090207.

Turner, John C., Penelope J. Oakes, S. Alexander Haslam, and Craig McGarty. 1994. “Self and Collective: Cognition and Social Context.” *Personality and Social Psychology Bulletin* 20(5):454–63. doi: 10.1177/0146167294205002.

Twenge, Jean M., Natalie J. Ciarocco, Roy F. Baumeister, C. Nathan DeWall, and J. Michael Bartels. 2007. “Social Exclusion Decreases Prosocial Behavior.” *Journal of Personality and Social Psychology* 92(1):56–66. doi: 10.1037/0022-3514.92.1.56.

Whitt, Sam, and Rick K. Wilson. 2007. “The Dictator Game, Fairness and Ethnicity in Postwar Bosnia.” *American Journal of Political Science* 51(3):655–68. doi: 10.1111/j.1540-5907.2007.00273.x.

Winetrobe, Hailey, Harmony Rhoades, Eric Rice, Norweeta Milburn, and Robin Petering. 2017. "'I'm Not Homeless, I'm Houseless': Identifying as Homeless and Associations with Service Utilization among Los Angeles Homeless Young People." *Journal of Social Distress and Homelessness* 26(1):16–24. doi: 10.1080/10530789.2017.1280204.

TABLES

Table 1. Distribution of sociodemographic and social network factors among N=173 people experiencing homelessness (PEH) in Nashville, TN

	Total		Control		Unhoused		Housed	
	N	%	n	%	n	%	n	%
Sociodemographic Factors								
Gender								
Male	117	67.6	38	64.4	40	70.2	39	68.4
Female	53	30.6	20	33.9	16	28.1	17	29.8
Non-Binary	3	1.8	1	1.7	1	1.7	1	1.8
Ethnicity								
White	99	57.2	32	54.2	34	59.6	33	57.8
Non-White	74	42.8	27	45.8	23	40.4	24	42.2
Education								
K-11th Grade	52	30.1	13	22.0	20	35.1	20	35.1
GED or High School	75	43.4	25	42.4	25	43.9	25	43.9
Trade School or Any Higher Ed.	46	26.5	21	35.6	12	21.0	12	21.0
Rides Bus Everyday								
No	79	45.7	30	50.8	27	47.4	22	38.6
Yes	94	54.3	29	49.2	30	52.6	35	61.4
Has Yearlong Bus Pass								
Yes	11	6.4	3	5.1	3	52.6	5	8.7
No	161	93.1	56	94.9	54	47.4	51	89.5
Lifetime Homelessness Duration								
1 year or less	31	17.3	5	8.6	14	24.5	12	21.1
1 year – 5 years	70	40.5	29	49.1	20	35.1	21	36.8
5 years – 10 years	32	18.5	13	22.0	6	10.5	13	22.8
10 years +	40	23.7	12	20.3	17	29.9	11	19.3
Sheltered or Unsheltered								
Sheltered	12	7.5	5	9.3	1	1.8	6	11.8
Unsheltered	161	92.5	54	90.7	56	98.2	51	88.2
Perceived Deservedness of 2 nd Player								
< deserving than participant	52	30.0	15	25.4	23	40.3	14	34.6
≥ deserving than participant	120	69.8	44	74.6	33	59.7	43	65.4
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Age	45.2	11.5	46.3	11.9	45.9	11.2	43.2	11.4
Social Network Factors								
Perceived PEH Allocations	1.8	1.5	2.1	1.5	1.7	1.4	1.6	1.4
Network Size	2.5	1.8	2.5	2.0	2.8	1.8	2.3	1.5
Number of Financial Supports	1.3	1.5	1.4	1.6	1.5	1.5	1.1	1.3
Number of Emotional Supports	1.3	1.4	1.5	1.6	1.4	1.5	1.0	0.9
Number of Material Supports	1.8	1.5	1.8	1.6	1.9	1.5	1.7	1.5

Table 2. Multivariate Poisson regression estimates between number of bus passes allocated in game and sociodemographic factors, social network composition, and community perception among N = 173 PEH in Nashville, TN.

	Model 1		Model 2	
	Exp(β)	95% CI	Exp(β)	95% CI
Treatment Group				
Control (reference)	---	---	---	---
Unhoused	0.81	(0.63, 1.03)	0.93	(0.72, 1.19)
Housed	0.94	(0.74, 1.20)	1.03	(0.80, 1.33)
Sociodemographic Factors				
Gender				
Female (reference)	---	---	---	---
Male	1.04	(0.83, 1.31)	1.01	(0.80, 1.28)
Highest Level of Education				
K-11 th Grade (reference)	---	---	---	---
GED or High School	1.21	(0.94, 1.56)	1.11	(0.86, 1.44)
Trade School or Any Higher Education	1.17	(0.89, 1.56)	1.09	(0.81, 1.45)
Lifetime Homelessness Duration				
1 year or less (reference)	---	---	---	---
1 year – 5 years	0.95	(0.73, 1.25)	1.00	(0.76, 1.31)
5 years – 10 years	0.66*	(0.46, 0.93)	0.69*	(0.48, 0.98)
10+ years	0.87	(0.63, 1.19)	0.86	(0.62, 1.19)
Rides Bus Everyday				
No (reference)	---	---	---	---
Yes	0.64***	(0.53, 0.79)	0.67***	(0.54, 0.82)
Social Network Composition				
# Emotional Supports	---	---	1.11*	(1.02, 1.20)
# Total Material Supports	---	---	0.95	(0.87, 1.02)
Perceived Deservedness & Generosity of Other PEH				
Perceived Deservedness of 2 nd player				
< deserving than participant (reference)	---	---	---	---
≥ deserving than participant	---	---	1.39*	(1.09, 1.79)
Perceived PEH Donations	---	---	1.09**	(1.02, 1.17)
Exp(Intercept)		3.05***		1.35
AIC		656.6		637.6

Significance values: * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$

Table 3. Multivariate Poisson regression estimates between number of bus passes allocated in game and treatment group, sociodemographic factors, social network composition, and community perception among N = 173 PEH in Nashville, TN.

	Model 3		Model 4	
	Exp(β)	95% CI	Exp(β)	95% CI
Treatment Group				
Control (reference)	---	---	---	---
Unhoused	0.88	(0.68, 1.14)	0.84	(0.64, 1.09)
Housed	1.02	(0.79, 1.31)	0.95	(0.74, 1.22)
Sociodemographic Factors				
Gender				
Female (reference)	---	---	---	---
Male	0.66*	(0.48, 0.92)	0.68*	(0.48, 0.95)
Highest Level of Education				
K-11 th Grade (reference)	---	---	---	---
GED or High School	1.13	(0.87, 1.46)	1.11	(0.86, 1.44)
Trade School or Any Higher Education	1.03	(0.77, 1.38)	1.09	(0.82, 1.46)
Lifetime Homelessness Duration				
1 year or less (reference)	---	---	---	---
1 year – 5 years	1.01	(0.77, 1.34)	0.98	(0.75, 1.30)
5 years – 10 years	0.66*	(0.45, 0.93)	0.71	(0.49, 1.01)
10+ years	0.82	(0.59, 1.14)	0.91	(0.66, 1.25)
Rides Bus Everyday				
No (reference)	---	---	---	---
Yes	0.69***	(0.56, 0.84)	0.69***	(0.56, 0.85)
Social Network Composition				
# Emotional Supports	0.82*	(0.68, 0.98)	---	---
# Total Material Supports	---	---	0.82*	(0.70, 0.96)
Perceived Deservedness & Generosity of Other PEH				
Perceived Deservedness of 2 nd player				
< deserving than participant (reference)	---	---	---	---
≥ deserving than participant	1.22	---	1.26	(0.99, 1.63)
Perceived PEH Donations	1.08*	(1.01, 1.16)	1.09*	(1.01, 1.17)
Gender and Emotional Support Interaction				
Female x Total Emotional Supports (reference)	---	---	---	---
Male x Total Emotional Supports	1.39***	(1.14, 1.71)	---	---
Gender and Material Support Interaction				
Female x Total Material Supports (reference)	---	---	---	---
Male x Total Material Supports	---	---	1.28**	(1.08, 1.54)
Exp(Intercept)		2.41*		2.33*
AIC		628.1		635.7

Significance values: *p ≤ 0.05, **p ≤ 0.01, ***p ≤ 0.001

Table 4. Multivariate logistic regression of factors associated with allocating zero bus passes in the dictator game among N = 173 PEH in Nashville, TN.

	Estimate	Model 5	
		OR	95% CI (OR)
Gender			
Female (reference)			
Male	-0.07	0.93	(0.38, 2.38)
Rides Bus Everyday			
No (reference)			
Yes	1.40**	4.07	(1.60, 11.99)
Lifetime Homelessness Duration			
1 year or less (reference)			
1 year – 5 years	0.46	1.58	(0.46, 6.50)
5 years – 10 years	1.10	3.01	(0.80, 13.24)
10+ years	0.44	1.55	(0.39, 6.84)
# Emotional Supports	-0.08	0.92	(0.64, 1.26)
Perceived PEH Donations	-0.32*	0.70	(0.50, 0.97)
Intercept	-2.27**	0.10	(0.02, 0.47)
AIC		157.9	

Significance values: *p ≤ 0.05, **p ≤ 0.01, ***p ≤ 0.001

FIGURES

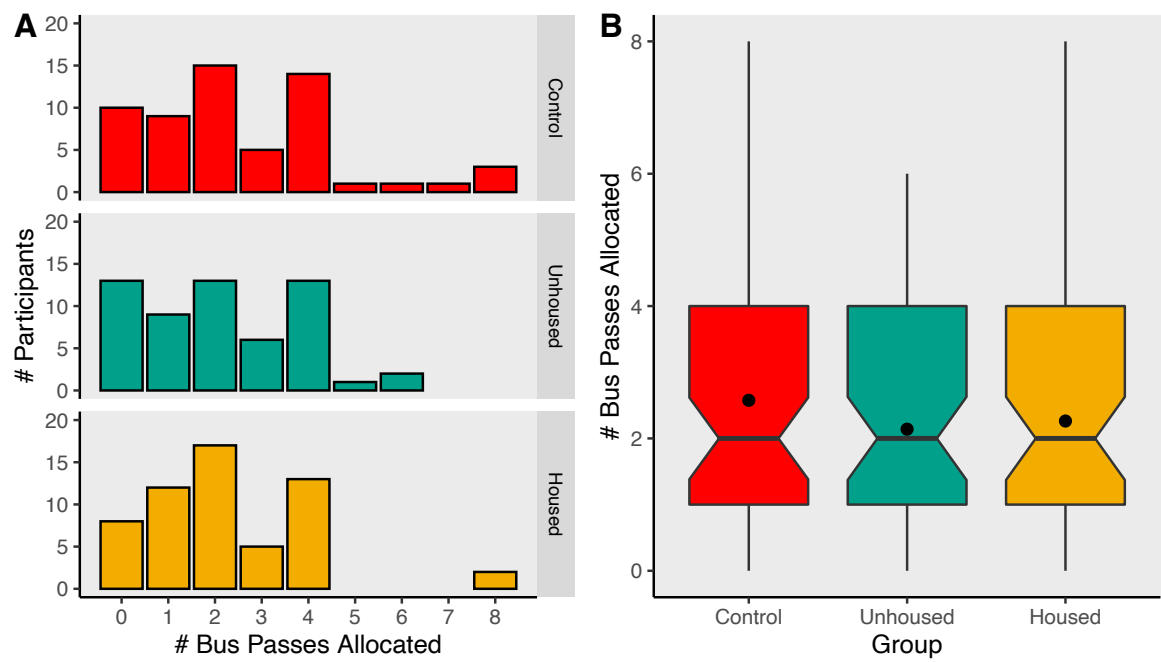


Figure 1.

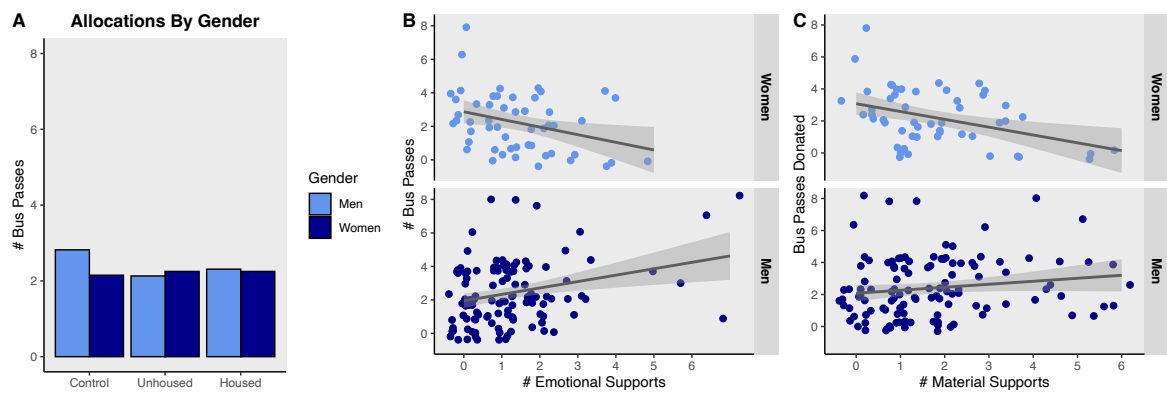


Figure 2.

FIGURE CAPTIONS

Figure 1. Frequency Distributions of Bus Pass Allocations By Treatment Group (A) and Box and Whisker Plot of Bus Pass Allocations By Treatment Group (B).

Figure 2. Frequency Distributions of Bus Pass Allocations By Gender and Treatment Group (A) and Relationship Between Number of Emotional Supports.

[Under Peer Review Do Not Cite or Share]

Barefaced and voting: mask-wearing choice and the Trump vote in the United States during the COVID-19 pandemic

Authors

Aja Sutton (amsutton@uw.edu)

University of Washington

Zack Almquist (zalmquis@uw.edu)

University of Washington

James Holland Jones (jhj1@stanford.edu)

Stanford University

Acknowledgments

Partial support for this research came from a Eunice Kennedy Shriver National Institute of Child Health and Human Development research infrastructure grant, P2C HD042828, and Shanahan Endowment Fellowship, and a Eunice Kennedy Shriver National Institute of Child Health and Human Development training grant, T32 HD101442, to the Center for Studies in Demography & Ecology at the University of Washington. Partial support for this research also came from an NSF Award #BCS-2028160 and an ARO Award #W911NF-19-1-0407. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, NSF, or ARO.

Abstract (300 words max) word count: 300

A poorly understood aspect of the choice to wear a mask during the COVID-19 pandemic in the United States is the role of politicization of mask-wearing along political party lines. We consider the role of voting for Donald Trump in mask-wearing choice during the COVID-19 pandemic in the United States using a logistic choice model of indoor mask-wearing choice and the Trump vote in the 2016 United States presidential election, accounting for sociodemographic and ecological characteristics, trust in government, approximate time and variation in relative space, as well as a Pearson's correlation analysis. Data used are from the American Community Survey 2019 5-year estimates and The Behavioral Change COVID-19 (BCC19) Survey, a nationally representative longitudinal survey of adults (over 18 years) in the United States of opinions and behaviors during the COVID-19 pandemic over three waves from September 2020 to June 2021. We find that voting for Donald Trump in the 2016 federal election had the greatest effect magnitude and negative relationship with indoor public mask-wearing. Trump support also appears to be associated with a lower perception of personal risk from the pandemic and consequently a lower or negative perception of the utility of indoor public mask-wearing. Political identity and partisanship at the population level is an aggregate of how people interact with and are acted on by their environments. The political ecology of health in the United States during the COVID-19 pandemic has shifted as a result of party politics becoming central to individual and community identity, and the polarization of values and choices delineated along these lines. In this case, sociodemographic conditions and group political identity aggregated to create an environment in the United States where voting for Donald Trump was a population-level health risk factor.

Keywords (up to 8): COVID-19, logistic choice model, mask use, political ecology of health,

1. Introduction

Mask-wearing has been a tension point throughout the COVID-19 pandemic. While the federal government of the United States has promoted vaccination and non-pharmaceutical interventions (NPIs) such as mask-wearing as vital components of mitigation, there remains variation in messaging from state and regional governments across geographies and time (White & Hébert-Dufresne, 2020). Individuals and communities decide whether to heed behavior-intervention advice from governments and politicians based on trust (Dai et al., 2020; Siegrist et al., 2021; Ahluwalia et al., 2021). This relationship is complicated: political identity in the United States is increasingly tied not only to preference for a political party (e.g., Republican vs. Democrat) but also for specific presidential candidates (e.g., Donald Trump vs. Hillary Clinton or Joe Biden) (Heltzel & Laurin, 2020). These political identity-centered communities have been shown to be extremely lenient in their evaluation and criticism of claims made by their preferred politicians. Just as political sentiment in the United States has become polarized along party lines and variable geographically (Kaplan et al., 2022), mask-wearing behaviors have been temporospatially variable over the course of the pandemic (Cunningham & Nite, 2021). Differences in mask-wearing behavior are the product of decision-making processes, and these may be influenced by sentiments and preferences formed within the context(s) of the sociopolitical environments in which people live, community political identity, and the trust individuals place in governing bodies to make fair and just decisions for their well-being. A poorly understood aspect of the choice to wear a mask during the COVID-19 pandemic in the United States is the role of the extreme politicization of NPIs along political party lines especially in the lead-up to the 2020 Presidential Election. This study uses a political ecology of health framework to consider how mask-wearing behavior in the United States during the COVID-19 pandemic has been affected by presidential candidate preference, trust in governance, and sociodemographic characteristics.

The places where people live affect their health (Diez Roux & Mair, 2010). The political ecology of health considers health outcomes from a population-level perspective that de-emphasizes individual agency in health. Instead, health outcomes are the result of a complex web of multi-directional factors at

the population and ecological level (Susser & Susser, 1996a; Krieger, 2011), and are constituted of the deep connections between human societies, political decisions and histories with specific temporospatial contexts. An concrete example of this is the building of the Aswan Dam in Egypt in the 1960s (Mayer, 2000); while it improved crop stability, food security, and provided electricity to many Egyptian villages for the first time, there were unforeseen, “unintended” consequences: when the speed of the Nile’s water flow decreased, the population of schistosomiasis-carrying snails living along its banks increased while local populations’ direct contact with the river’s banks persisted, thus increasing exposure to snails and significantly increasing disease incidence. As this example illustrates, population health is an aggregate of how people interact with and are acted on by their physical and socioeconomic environments (Cummins et al., 2007; Krieger, 1994; 2001; 2011). Social and political actions may have unforeseeable or unintended consequences on the well-being and balanced functionality of landscapes, environments and especially human (and non-human animal) health (Turshen, 1977; 1984; Mayer, 1996; 2000; 2010). In this way, the political ecology of health posits that health environments arise when populations contribute to area-level demographic, behavioral and socioeconomic characteristics, and area-level health is at least partially determined by the social, physical and economic characteristics of the place itself (Bambra, 2018).

The political ecology of health framework has been previously applied to COVID-19 (e.g., Tipirneni et al., 2022). Transmission interventions like mask-wearing are upstream to COVID-19 infection; mask-wearing has had a significant impact on reducing the overall transmission of COVID-19 during the pandemic (Howard et al., 2021; Brooks & Butler, 2021; Mello et al., 2022). More abstract political ecology of health mechanisms are affected and driven by the political machinations of policymakers and the politicization of resources, practices and beliefs relating to health outcomes by politicians. These have had unintended consequences on health practices and disease transmission, such as with both the rise of vaccine hesitancy (Bokemper et al., 2021; Jennings et al., 2021) and the increase in anti-masking sentiment (Taylor & Asmundson, 2021). Health risk perceptions and transmission mitigation behaviors are a product of the political ecology of health, varying according to demographic

attributes such as socioeconomic status, race and ethnicity, space and time (Anthonj et al., 2022; Kumar & Encinosa, 2021). Perceived health risk and the uptake of NPIs during the COVID-19 pandemic has varied by age, where older people are more likely to maintain health-protective measures like mask-wearing (Kim & Crimmins, 2020); by race and ethnicity, where people racialized as Black (Kumar & Encinosa, 2021) and Hispanic (Jamieson et al., 2021) are more likely to adhere to protective behavior advice due to a higher perception of risk from COVID-19; and by political affiliation, where left-leaning people and/or Democrats are more likely to wear masks than right-wing people and/or Republicans (Howard et al., 2022). By extension, population-level mask-wearing patterns are likely an aggregate of several political ecological factors which can be studied temporospatially. The current study aims to examine the association between presidential candidate preference and mask-wearing behavior in the United States during the COVID-19 pandemic, while also addressing sociodemographic characteristics, trust in various levels of government (i.e., local, state and federal), and temporospatial variability.

2. Methods

2.1 Data

The Behavioral Change COVID-19 (BCC19) Survey, funded by the National Science Foundation, is a nationally representative longitudinal survey of adults (over 18) that followed the same participants over three waves between September 2020 and June 2021. Each wave of the survey was fielded for two to three weeks through Ipsos KnowledgePanel, an address-based probability sample fielded online through a panel representative of over-18 respondents living in the United States (Ipsos, 2020-2021). Participants responded to a battery of questions focusing on identifying opinions and behaviors within the United States population during the COVID-19 pandemic. This included geographic information (ZIP code, county, state), demographics (e.g., sex/gender, age, race/ethnicity, urbanicity, household and employment characteristics), political identity, trust in government, and social distancing and masking behaviors, among other factors. The *BCC19 survey* data covered a sociopolitically dynamic time in United States history, both with regard to the political events of that period and the concurrent development and

(eventual) availability of multiple effective COVID-19 vaccines. The first wave of data was collected in September 2020, part of the lead-up to the 2020 United States Presidential Election and in a time prior to COVID-19 vaccine availability. The second wave of data was collected in December 2020, shortly after the 2020 Presidential Election between Republican incumbent Donald Trump and Democratic candidate Joe Biden, around the same time the Federal Drug Administration first issued its emergency use authorization for the Pfizer-BioNTech COVID-19 Vaccine. The third wave was collected several months into the Biden administration when COVID-19 vaccines were broadly available to the public.

The first wave of the *BCC19 survey* was benchmarked to be representative of the United States population aged eighteen years or older (reported as weighted survey means); survey respondents are reweighted based on the US Census and American Community Survey by method known as raking (see: e.g., Lumley, 2010). This survey consisted of 1,421 respondents of which there were 586 females (71.9%) and 74.3% identified as White, non-Hispanic (n=1021), 8.8% as Black, non-Hispanic (n=125), 4.36% as Other, non-Hispanic (n=62), 11.3% as Hispanic (n= 160), and 3.7% as two or more races (n= 53). The second survey wave consisted of 1,186 respondents, of which there were 586 females (49.4%), and 72.4% identified as White, non-Hispanic (n=859), 8.3% as Black, non-Hispanic (n=98), 4.5% as Other, non-Hispanic (n=53), 11.2% as Hispanic (n= 133), and 3.6% as two or more races (n= 43). The third survey wave consisted of 1,010 respondents, of which 49.8% were female (n=503), and 74.3% identified as White, non-Hispanic (n=750), 7.4% as Black, non-Hispanic (n=75), 4.4% as Other, non-Hispanic (n=44), 10.2% as Hispanic (n= 103), and 3.8% as two or more races (n= 38). Further descriptive statistics of the survey sample, weighted to reflect the United States population aged 18 years and over, are described in Table 1.

The outcome variable, indoor mask-wearing, was derived from a question in the *BCC19 survey*: “In the last week, how often did you wear a mask indoors in public places?” Responses were originally collected on a five-order ordinal scale (i.e. 1 = always through 5 = never), and were re-coded to a binary response, where “always” was recoded as 1 and all other responses were recoded as 0. Political predictor variables included trust in the federal government, trust in state government, trust in local government,

and if the choice of 2016 United States presidential candidate was Donald Trump (yes: weighted survey mean 0.593 (95% CI: 0.575, 0.61) and no: 0.407 (95% CI: 0.39, 0.425)). Choice of candidate is an aggregate measure of both choice of candidate by respondents who self-reported as having voted in the election, as well as candidate preference of respondents who did not vote. It was not possible to include the candidate choice for the 2020 United States presidential election due to an error in the *BCC19 survey* data product delivered by Ipsos. The sociodemographic predictor variables were age (adults aged eighteen years and over, on a continuous scale); race/ethnicity (collected as White, non-Hispanic; Black, non-Hispanic; Hispanic; Other, non-Hispanic; and Two or More Races, non-Hispanic; but simplified to White, non-Hispanic; Black, non-Hispanic; Hispanic; Other, non-Hispanic due the small number of respondents who identified as belonging to Two or More Races, non-Hispanic); and employment status (working; not working; retired; or disabled). Race/ethnicity includes “Other, non-Hispanic” as an aggregate of respondents who did not self-identify as being either White, non-Hispanic, Black, non-Hispanic, or Hispanic, according to Ipsos’ categorization. In addition, we included ecologic predictor variables: rurality (i.e. the respondent was living in a rural environment; the respondent was not living in a rural environment). Temporal variation was accounted for by including the survey wave as an ordered factor (wave 1, i.e. corresponding to the survey period in September 2020; wave 2, i.e. December 2020; wave 3, i.e. June 2021). Geographic variation was accounted for using the respondents’ ZIP codes provided in the survey data; to correspond with American Community Survey data these were converted to ZIP Code Tabulation Area (ZCTA, a U.S. Census Bureau spatial polygon approximate equivalent to ZIP code) using a crosswalk file from the U.S. Department of Housing and Urban Development (HUD, 2020).

2019 5-year American Community Survey: This study also used supplemental information from the 2019 5-year American Community Survey. We use the 5-year aggregate form of the ACS to obtain ZCTA level estimates with sufficient precision for our research questions. The ACS is a yearly sample of American households aimed at capturing social and behavioral characteristics of the United States population in the years between the Decennial Census; the 5-year version is a survey weighted aggregate

of the previous five years of responses and is considered more representative of trends than the individual one-year versions. Two ecological predictor variables were taken from this survey to improve estimates: proportion of the population racialized as White in the ZCTA, and proportion of the population racialized as Black in the ZCTA. These were included to attempt to account for difficult-to-measure systemic racism and often highly regionalized social inequities and characteristics present in respondents' political ecological contexts for which the survey did not provide more specific data.

2.2 Statistical Analysis

To understand the decision making process of how people select to wear a mask or not we employ a logistic choice model (McFadden, 1986). If we assume that the selection of whether to wear a mask follows a weakly consistent pattern of preferences, in the sense that there exists a *utility function*, u , such that for two alternative states $Y_i = 0$ and $Y_i = 1$, the odds that i will choose $Y_i = 1$ (mask wearing) are strictly increasing in $u_i(Y|Y_i = 1)/u_i(Y|Y_i = 0)$. This general formulation is commonly known as a *stochastic choice* process, and can be viewed as a form of bounded rationality (McFadden, 1986). Although many *stochastic choice* processes exist, we will employ the common *logistic choice* model, which can be written as follows:

$$\Pr(Y_i = 1) = \frac{\text{EXP}[u(Y_i|Y_i=1)]}{\text{EXP}[u(Y_i|Y_i=1)] + \text{EXP}[u(Y_i|Y_i=0)]},$$

or, equivalently, as:

$$\text{logit } \Pr(Y_i = 1) = \ln \frac{\Pr(Y_i=1)}{\Pr(Y_i=0)} = u(Y_i|Y_i = 1) - u(Y_i|Y_i = 0).$$

That is, the log-odds that i will choose to mask is equal to the utility difference associated with wearing (versus not wearing) a mask. Where the utility of one option is substantially greater than the other, then, actor behavior is nearly deterministic: the utility-increasing choice is

selected with very high probability. To use this framework we have to further assume that u can be written as a sum of linearly separable payoff elements, s , such that $u_i(Y) = \theta^T s(Y, i, X)$, so that s represents the factors potentially driving i 's behavior and θ represents the direction and magnitude of the effect these factors have on the propensity to wear a mask or not.

Analysis consisted of a logistic regression model fit using the `lme4` package (Bates et al., 2015) in R statistical language (R Core Team, 2021). The final model was chosen using Bayesian Information Criterion (BIC) model selection criterion (Schwarz, 1978). We interpret any hypothesis that did not make it into the final model test to be a highly unlikely event under the limitations of our data. The outcome “indoor public mask-wearing choice” (hereafter “mask-wearing”) was modeled as a function of survey wave, preferred candidate in the 2016 U.S. presidential election, trust in local government, trust in state government, trust in federal government, proportion of the population in the ZCTA racialized as White, non-Hispanic, proportion of the population in the ZCTA racialized as Black, non-Hispanic, age, rurality (i.e. the inverse of urbanicity; that is, the binary characteristic of being rural or not rural), race and ethnicity, and employment status, with a random effect placed on ZCTA to account for relative geographic variation. The log-odds, 95% confidence intervals, and p-values ($\alpha = 0.05$) are reported below.

A Pearson's correlation coefficient with a bootstrap was calculated to test the association between mask-wearing and preferring Donald Trump in the 2016 presidential election, with results reported below. Three scenarios with different cut-points for binary re-coding of the mask-wearing variable were tested; most strict (1 = “always”; 0 = “most of the time”, “sometimes”, “rarely”, “never”), moderately strict (1 = “always”, “most of the time”, “sometimes”; 0 = “rarely”, “never”), and least strict (1 = “always”, “most of the time”, “sometimes”, “rarely”; 0 = “never”).

3. Results

3.1 Logistic regression results

We highlight results of interest; full results of the logistic regression analysis are available in Table 2 and are illustrated in Figure 1. In terms of sociodemographic and ecological characteristics,

compared to those aged 18-29 years, those aged 60 years and older had 0.976 (95%CI: 0.099, 1.853, $p = 0.029$) the log-odds of choosing to wear a mask; other age categories were not found to be statistically significant. Similarly, race/ethnicity was not found to be significantly predictive of mask-wearing choice: compared to those who were categorized as White/non-Hispanic, respondents who were categorized as Black/non-Hispanic had log-odds of choosing to wear a mask of 0.277 (95%CI: -1.164, 1.717, $p = 0.707$); respondents who were categorized as as Other, non-Hispanic had log-odds of 0.252 (95%CI: -0.77, 1.275, $p = 0.629$); and respondents who were categorized as Hispanic had log-odds of 0.016 (95%CI: -1.001, 1.032, $p = 0.976$). Compared to respondents who identified themselves as currently working, those who were not working had -0.726 (95%CI: -1.408, -0.043, $p = 0.037$) the log-odds of choosing to wear a mask in indoor public spaces, a statistically significant finding, whereas those who were retired had -0.317 (95%CI: -1.076, 0.441, $p = 0.412$) the log-odds of choosing to wear a mask.

Considering results from ecological covariates, compared to those not living in rural areas, those who lived in rural areas had -0.654 (95%CI: -1.207, -0.101, $p = 0.021$) the log-odds of choosing to wear a mask in indoor public spaces. Living rurally was therefore a moderate but statistically significant negative predictor of mask-wearing decision-making. Conversely, proportions of the ZCTA identified as Black, non-Hispanic or White, non-Hispanic were not found to be statistically significant predictors contributing to choosing to wear a mask, though they were both negative trending (log-odds -0.78, 95%CI: -3.2, 1.64, $p = 0.528$, and log-odds -0.878, 95%CI: -2.478, 0.7422 $p = 0.282$, respectively). Similarly, each survey wave was found to be strongly statistically significant in predicting respondent mask-wearing choice (Wave 1/September 2020 log-odds: 4.963, 95%CI: 3.379, 6.547, $p < 0.001$; Wave 2/December 2020 log-odds: 5.541, 95%CI: 3.92, 7.163, $p < 0.001$; Wave 3/June 2021 log-odds: 3.503, 95%CI: 1.942, 5.065, $p < 0.001$).

The effect of respondent trust of government on mask-wearing choice was moderate. Greater trust in federal and state governments was associated with moderate positive trends in the log-odds of choosing to wear a mask (0.486, 95%CI: -0.011, 0.983, $p = 0.055$) and 0.442 (95%CI: -0.145, 1.03, $p = 0.14$), respectively), whereas trust in local government was associated with almost no effect on choice, with

log-odds of -0.096, 95%CI: -0.762, 0.57, $p = 0.778$). The characteristic with the greatest effect magnitude on the decision to wear a mask in indoor public spaces was the choice of 2016 United States presidential candidate. Compared to respondents who did not vote for Donald Trump, those who voted for Donald Trump (“Trump supporters”) had -1.596 (95%CI: 2.189, -1.004, $p < 0.001$) the log-odds of choosing to wear a mask in indoor public spaces.

3.2 Pearson’s correlation analysis

The results of the Pearson’s correlation analysis of the association between mask-wearing and voting for Donald Trump in the 2016 presidential election are available in Table 4. Pearson’s correlation coefficients for all three cut-point scenarios were extremely close to zero, indicating there is not a clear relationship between voting for Donald Trump and mask-wearing choice in any of the scenarios.

3.3 Risk perception among those who preferred Trump for the 2016 presidential election

Over all periods of observation, on average all those who were Trump supporters had known someone who had had a COVID-19 infection, though fewer had known someone who had died from COVID-19 (i.e. survey mean proportions of 0.278; 0.269; 0.338 for each period respectively). On average, in September 2020 those who preferred or voted for Trump perceived the importance of social distancing to mitigate transmission of COVID-19 to be of about the same as what they felt their families’ (3.192, 95%CI: 3.147, 3.238), friends’ (3.107, 95%CI: 3.062, 3.153) and neighbors’ (4.014, 95%CI: 3.941, 4.087) felt about social distancing. At this time, Trump supporters were less likely to want to receive a COVID-19 vaccination if one were to be developed and available (0.474, 95%CI: 0.454, 0.494) in comparison to those who preferred or voted for another candidate (0.694, 95%CI: 0.677, 0.71). Trends were similar in December 2020 (see Table 3 for details and estimates for those who preferred or voted for another candidate).

In June 2021, 160 respondents identified as Trump supporters were undecided or against receiving a COVID-19 vaccination despite vaccines being readily publicly available; we term these respondents either “vaccine hesitant” or “against vaccination”. Among Trump supporters who were vaccine hesitant or against vaccination, three-quarters believed the vaccine was not safe (“Maybe” =

0.742, 95%CI: 0.583, 0.9; “No” = 0.744, 95%CI: 0.658, 0.83), a third-to-half felt the vaccine was unnecessary (“Maybe” = 0.388, 95%CI: 0.214, 0.563; “No” = 0.473, 95%CI: 0.376, 0.57), and a quarter-to-a-third that it would not protect them (“Maybe” = 0.26, 95%CI: 0.11, 0.429; “No” = 0.363, 95%CI: 0.268, 0.458), and a fifth to a quarter believed that it would not protect others (“Maybe” = 0.177, 95%CI: 0.034, 0.32; “No” = 0.268, 95%CI: 0.181, 0.355). Religious beliefs were minimally relevant to an unwillingness to be vaccinated against COVID-19 (“Maybe” = no respondents; “No” = 0.093, 95%CI: 0.041, 0.145). In comparison to Trump supporters, those who preferred or voted for a candidate other than Trump in 2016 and who were also vaccine hesitant or against vaccination had similar responses to Trump supporters with one exception (see Figure 2). Vaccine hesitant respondents who did not prefer or vote for Trump in 2016 had half the proportion of respondents who believed the COVID-19 vaccine unnecessary (0.19, 95%CI: 0.073, 0.308) in comparison to vaccine hesitant Trump supporters (see above).

4. Limitations

As with any analysis, this study has some limitations. We worry that self-reported mask-wearing or vaccine behavior might be exaggerated due to social stigma. Considering the amount of politicization of mask-wearing and politics as identity evident, it is likely that those who responded they “never” wore masks were truthful. Social desirability stigma is likely further reduced because of the anonymity of an online survey.

5. Summary and Discussion

This analysis considered the role of presidential candidate preference with specific attention to the Trump vote in mask-wearing choice during the COVID-19 pandemic in the United States. We modeled the relationship between choosing to wear a mask in indoor public spaces and choosing to vote for Donald Trump (or not) in the 2016 United States presidential election, accounting for sociodemographic and ecological characteristics, trust in different levels of government, approximate time and variation within U.S. Census Bureau ZCTA. Several sociodemographic and ecological characteristics were predictive of mask-wearing choice, including being aged 60 years and over (a protective factor, i.e., greater odds of

mask-wearing), not working, and living rurally (both risk factors, i.e., lesser odds of mask-wearing). Trust in government at the state and federal levels were modestly protective factors in choosing whether to wear a mask in indoor public spaces; trust of governance appears to be part of a broader set of political beliefs and sentiments that may not be summed up to party politics but inform utility in the choice-making process. In terms of perception of risk, we found all respondents to be just as concerned as their social connections about social distancing as a means to prevent COVID-19 transmission in the early part of the pandemic, but less so over time. Respondents who were against vaccination held similarly strong sentiments regarding their reasons for not being vaccinated regardless of their candidate preference. However, while all respondents who were vaccine hesitant in June 2021 indicated concerns around personal safety as being the greatest factor as to why they were undecided about receiving a COVID-19 vaccination, there were proportionately twice as many vaccine-hesitant Trump supporters who believed the vaccine was unnecessary in comparison to vaccine hesitant supporters of other candidates. This speaks to a difference in risk perception and sentiment between Trump supporters and those who support other candidates and bears further analysis. The present study's results suggest those who were vaccine hesitant and did not support Trump were primarily concerned about their personal safety, while attitudes among specifically vaccine-hesitant Trump supporters closely mirrored those held by respondents against vaccination: primarily, that COVID-19 vaccines were unsafe, unnecessary and, to a lesser degree, that they believed would not protect the respondents' health or the health of others. Taken with results relating to comparative social network risk perception (above), the Trump vote appears to be associated with a lower perception of personal risk and consequently a lower or negative perception of the utility of mask-wearing indoor public spaces.

Mask-wearing is a health-protective action found to be a simple and effective way of slowing COVID-19 transmission, though it has come under public scrutiny as a result of "pandemic fatigue" (MacIntyre et al., 2021), and as recent variants have become more easily transmissible with lower viral exposure. The period of the COVID-19 pandemic in the United States has seen extreme politicization and subsequent polarization of many government-sanctioned and often mandated pharmaceutical and

non-pharmaceutical interventions, including but not limited to vaccinations, physical (social) distancing, shelter-at-home orders (and similar), and mask-wearing. As a consequence of the political structure of the United States, politicization of these interventions has occurred along party lines, and the polarization of these political beliefs has led to increasingly strongly delineated political communities where beliefs surrounding COVID-19 and interventions like mask-wearing are given meaning, reinterpreted collectively and mediated or reevaluated through individuals' interactions with these beliefs (see: Blumer, 1969). In this way, Democrats and left-leaning people have emerged as more likely to adhere to mask-wearing advice, whereas Republicans and right-leaning people have not (Howard, 2022). In this study, the characteristic that had the greatest effect magnitude on mask-wearing decision making also had the greatest overall negative association with the choice to wear a mask: of all covariates considered in this analysis, a vote or vote preference for Donald Trump in the 2016 United States presidential election was the single greatest risk factor for choosing to not wear a mask in indoor public spaces. This appears to be due to a low perception of both risk from COVID-19 and the utility of mask-wearing among Trump supporters. Admittedly, it was also the variable for which there was the most statistical certainty (it is one of three variables whose narrow confidence intervals do not contain the possibility of true odds of zero/log-odds of one). Conversely, the results from the Pearson's correlation analysis indicate no clear relationship between mask-wearing and voting for or preferring Trump in the 2016 federal election. Taken together with the results from the logistic regression model, this indicates the utility relationship between Trump support and choosing to wear a mask (or not) is complex, non-linear, and resides in specifics of social context and identity which may be interpreted with a political ecology lens.

Political identity and partisanship at the population level is an aggregate of how people interact with and are acted on by their environments, and these have ultimately affected the political ecology of health in the United States during the COVID-19 pandemic. The collective political identity centered in Trump support has influenced decision making related to mask-wearing by dismissing risk from COVID-19 and reducing the perceived utility of masking in indoor public spaces among its collective, and thus consequently affected the adoption of mask-wearing practices. These results underline the

importance of the conditional relationship between mask-wearing and the Trump vote: mask-wearing had an extremely negative relationship with the 2016 Trump votership conditional on the sociodemographic and ecologic context. Though many Trump voters did receive a COVID-19 vaccination, those who were vaccine hesitant or against vaccination were dismissive of the personal risk posed to them by a COVID-19 infection regardless of whether they had known someone who had been previously infected; Trump supporters' perception of risk during the study period was lower than non-Trump supporters. While the combination of specific social, demographic, geographic and political characteristics in the population of the United States ultimately provide conditions for the Trump vote to have a significantly negative relationship with indoor public mask-wearing, it is the specific combination of these population characteristics and Trump support (or lack thereof) that unite the disparate factors at play in choosing whether to mask up in indoor public spaces (or not). In this way, sociodemographic conditions and group political identity have aggregated to create an environment in the United States where relative to others, voting for Donald Trump was associated with choosing to not wear a mask in indoor public spaces, and to some degree with diminished personal perception of risk from COVID-19. While it is beyond the scope of the present discussion to speculate the degree to which COVID-19 transmission might have been affected or even mitigated if more Trump supporters had chosen to wear masks indoors during the period of this study, it has previously been established that choosing to wear a mask in indoor public spaces has a large effect on mitigating COVID-19 transmission and symptomatic infection rate if mask-wearing compliance is high (Howard et al., 2021). By extension, it is reasonable to surmise that there is a non-zero chance COVID-19 transmission and symptomatic illness did occur because of the increased choice among Trump supporters to not wear a mask. More concretely, we have evidenced Trump support was a population-level health risk factor during the period of observation, and contributed to the political ecology of the COVID-19 pandemic in the United States.

Health is political. The United States is in a period of increased political polarization. Mask-wearing is a protective health decision that might have been less contentious had it not been tied to a specific political identity entrenched and embodied in support for Donald Trump. The relationship

between the Trump vote and low odds choosing to wear a mask in indoor public spaces is a clear example of the unintentional but significant ramifications of the politicization of health interventions during the COVID-19 pandemic. The political ecology of health in the United States has shifted as a result of party politics becoming central to individual and community identity, and the polarization of values and choices delineated along these party-identity lines. Population health is an aggregate of individual health, and thus also of individual identity and decision making. As evidenced by the strong risk of not wearing a mask in indoor public spaces associated with the Trump vote, it is critical to consider political sentiment and influence on health behaviors when planning health emergency responses. Acknowledging the ways health in the United States is influenced by its specific political ecologies is crucial to future public health planning efforts. The influence politicians have on individual health-related decision making and subsequently on population health is increasingly apparent.

In the future, public health officials must be mindful of the powerful impact the politicization of public health initiatives can have, and work to ensure that politicians are adequately forewarned of the ramifications of politicizing health in bad faith. Public health must be revered by our political leaders. It is an alarming symptom of the erosion of non-partisanship of the media and other public institutions in the United States that the use of the COVID-19 pandemic transmission mitigation strategies put in place to avoid physical harm to the populous have arisen as political partisan talking points and galvanization tools for deeply polarized political identities and agendas. The current political climate is extremely threatening to the well-being of Americans: given the well-known effectiveness of mask-wearing against COVID-19 infection during the survey period, it is likely that many avoidable infections occurred because of low uptake of mask-wearing behaviors in indoor public spaces, though verification of this is beyond the scope of the present discussion. There is however an intense need for social scientists in public health positions to enable the clear communication of the importance and goals of health initiatives to elected officials of all political affiliations, and especially presidential candidates. Public health officials must attempt to use apolitical language and communication styles, and learn to adapt these to evolving political talking points to minimize politicization of public health goals. If we are to weather the loss of credibility that many

public health institutions have undergone over the course of the pandemic, we must be compassionate, considerate and clear communicators in a political environment hostile toward public health.

References

- Ahluwalia, S., Edelen, M., Qureshi, N., & Etchegaray, J. (2021). Trust in experts, not trust in national leadership, leads to greater uptake of recommended actions during the COVID-19 pandemic. *Risk, Hazards & Crisis in Public Policy*, 12(3), 283-302.
- Anthonj, C., Setty, K. E., Ferrero, G., A. Yaya, A.-M., Mingoti Poague, K. I. H., Marsh, A. J., & Augustijn, E.-W. (2022). Do health risk perceptions motivate water - and health-related behaviour? A systematic literature review. *The Science of the Total Environment*, 819, 1–25. <https://doi.org/10.1016/j.scitotenv.2021.152902>
- Bambra, C. (2018). Placing Health Inequalities: where you live can kill you. In *Routledge Handbook of Health Geography*. Edited by Gavin J. Andrews, Jamie Pearce, & Valorie A. Crooks. Taylor and Francis. <https://doi.org/10.4324/9781315104584>
- Bates D, Mächler M, Bolker B, Walker S (2015). “Fitting Linear Mixed-Effects Models Using lme4.” *Journal of Statistical Software*, 67(1), 1–48. doi: [10.18637/jss.v067.i01](https://doi.org/10.18637/jss.v067.i01).
- Bokemper, S. E., Huber, G. A., Gerber, A. S., James, E. K., & Omer, S. B. (2021). Timing of COVID-19 vaccine approval and endorsement by public figures. *Vaccine*, 39(5), 825–829. <https://doi.org/10.1016/j.vaccine.2020.12.048>
- Bokemper, S. E., Cucciniello, M., Rotesi, T., Pin, P., Malik, A. A., Willebrand, K., Paintsil, E. E., Omer, S. B., Huber, G. A., & Melegaro, A. (2021). Experimental evidence that changing beliefs about mask efficacy and social norms increase mask wearing for COVID-19 risk reduction: Results from the United States and Italy. *PloS One*, 16(10), e0258282–e0258282. <https://doi.org/10.1371/journal.pone.0258282>
- Blumer, H. (1969). Symbolic interactionism; perspective and method. Prentice-Hall.
- Brooks, J. T. & Butler, J. C. (2021). Effectiveness of Mask Wearing to Control Community Spread of SARS-CoV-2. *JAMA : the Journal of the American Medical Association*, 325(10), 998–999. <https://doi.org/10.1001/jama.2021.1505>
- Cummins, S., Curtis, S., Diez-Roux, A. V., & Macintyre, S. (2007). Understanding and representing ‘place’ in health research: A relational approach. *Social Science & Medicine* (1982), 65(9), 1825–1838. <https://doi.org/10.1016/j.socscimed.2007.05.036>
- Cunningham, G. B., and Nite, C. (2021). Demographics, politics, and health factors predict mask wearing during the COVID-19 pandemic: a cross-sectional study. *BMC Public Health*, 21(1), 1–1403. <https://doi.org/10.1186/s12889-021-11424-1>
- Dai, B., Fu, D., Meng, G., Liu, B., Li, Q., & Liu, X. (2020). The Effects of Governmental and Individual Predictors on COVID-19 Protective Behaviors in China: A Path Analysis Model. *Public Administration Review*, 80(5), 797-804.
- Diez Roux, A. V. & Mair, C. (2010). Neighborhoods and health. *Annals of the New York Academy of Sciences*, 1186(1), 125–145. <https://doi.org/10.1111/j.1749-6632.2009.05333.x>
- Glymour, M. M., Avendano, M., Kawachi, I. (2014). Socioeconomic Status and Health. In *Social epidemiology* (Second edition.). Edited by Berkman, Kawachi, Ichirō, & Glymour, M. Maria. Oxford University Press.
- Heltzel, G. and Laurin, K. (2020). Polarization in America: two possible futures. *Current Opinion in Behavioral Sciences*, 34, 179–184. <https://doi.org/10.1016/j.cobeha.2020.03.008>

- Howard, J., Huang, A., Li, Z., Tufekci, Z., Zdimal, V., van der Westhuizen, H.-M., von Delft, A., Price, A., Fridman, L., Tang, L.-H., Tang, V., Watson, G. L., Bax, C. E., Shaikh, R., Questier, F., Hernandez, D., Chu, L. F., Ramirez, C. M., & Rimoin, A. W. (2021). An evidence review of face masks against COVID-19. *Proceedings of the National Academy of Sciences - PNAS*, 118(4), 1–. <https://doi.org/10.1073/pnas.2014564118>
- Howard, M. C. (2022). Are face masks a partisan issue during the COVID-19 pandemic? Differentiating political ideology and political party affiliation. *International Journal of Psychology*, 57(1), 153–160. <https://doi.org/10.1002/ijop.12809>
- Ipsos Public Affairs. (2021). Knowledge Panel. <https://www.ipsos.com/en-us/solutions/public-affairs/knowledgepanel>
- Jamieson, T., Caldwell, D., Gomez-Aguinaga, B., & Doña-Reveco, C. (2021). Race, Ethnicity, Nativity and Perceptions of Health Risk during the COVID-19 Pandemic in the US. *International Journal of Environmental Research and Public Health*, 18(21), 11113. <https://doi.org/10.3390/ijerph18211113>
- Jennings, W., Stoker, G., Bunting, H., Valgarðsson, V. O., Gaskell, J., Devine, D., McKay, L., & Mills, M. C. (2021). Lack of Trust, Conspiracy Beliefs, and Social Media Use Predict COVID-19 Vaccine Hesitancy. *Vaccines (Basel)*, 9(6), 593. <https://doi.org/10.3390/vaccines9060593>
- Kaplan, E., Spenkuch, J. L., & Sullivan, R. (2022). Partisan spatial sorting in the United States: A theoretical and empirical overview. *Journal of Public Economics*, 211. <https://doi.org/10.1016/j.jpubeco.2022.104668>
- Kim, J. K., & Crimmins, E. M. (2020). How does age affect personal and social reactions to COVID-19: Results from the national Understanding America Study. *PloS One*, 15(11), e0241950–e0241950. <https://doi.org/10.1371/journal.pone.0241950>
- Krieger, N. (1994). Epidemiology and the web of causation: Has anyone seen the spider? *Social Science & Medicine*, 39(7), 887–903.
- Krieger, N. (2001). The Ostrich, the Albatross, and Public Health: An Ecosocial Perspective—Or Why an Explicit Focus on Health Consequences of Discrimination and Deprivation is Vital for Good Science and Public Health Practice. *Public Health Reports*, 116(5), 419–423.
- Krieger, N. (2011). *Epidemiology and the people's health: Theory and context*. New York: Oxford University Press.
- Kumar, & Encinosa, W. (2021). Racial Disparities in the Perceived Risk of COVID-19 and in Getting Needed Medical Care. *Journal of Racial and Ethnic Health Disparities*. <https://doi.org/10.1007/s40615-021-01191-5>
- Lumley, T. (2010). *Complex surveys : a guide to analysis using R*. John Wiley. <https://doi.org/10.1002/9780470580066>
- MacIntyre, C. R., Nguyen, P.-Y., Chughtai, A. A., Trent, M., Gerber, B., Steinhofel, K., & Seale, H. (2021). Mask use, risk-mitigation behaviours and pandemic fatigue during the COVID-19 pandemic in five cities in Australia, the UK and USA: A cross-sectional survey. *International Journal of Infectious Diseases*, 106, 199–207. <https://doi.org/10.1016/j.ijid.2021.03.056>
- Mayer, J. D. (1996) The political ecology of disease as a new focus for medical geography. *Progress in Human Geography* 20(4), 441–456.
- Mayer, J. D. (2000). Geography, ecology and emerging infectious diseases. *Social Science & Medicine*, 50(7–8), 937–952. [https://doi.org/10.1016/S0277-9536\(99\)00346-9](https://doi.org/10.1016/S0277-9536(99)00346-9)

- Mayer, J. D. (2010). Medical Geography. In T. Brown, Sara McLafferty & Graham Moon (eds). A companion to health and medical geography. Malden, MA: Wiley-Blackwell.
- Mello, V. M., Eller, C. M., Salvio, A. L., Nascimento, F. F., Figueiredo, C. M., Silva, E. S. R. F., Sousa, P. S. F., Costa, P. F., Paiva, A. A. P., Mares-Guias, M. A. M. M., Lemos, E. R. S., & Horta, M. A. P. (2022). Effectiveness of face masks in blocking the transmission of SARS-CoV-2: A preliminary evaluation of masks used by SARS-CoV-2-infected individuals. *PloS One*, 17(2), e0264389–e0264389. <https://doi.org/10.1371/journal.pone.0264389>
- R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Siegrist, M., Luchsinger, L., & Bearth, A. (2021). The Impact of Trust and Risk Perception on the Acceptance of Measures to Reduce COVID-19 Cases. *Risk Analysis*, 41(5), 787-800.
- Susser, M., & Susser, E. (1996a). Eras and paradigms. *The American Journal of Public Health* 86(5), 674-677.
- Taylor, S., & Asmundson, G. J. G. (2021). Negative attitudes about facemasks during the COVID-19 pandemic: The dual importance of perceived ineffectiveness and psychological reactance. *PloS One*, 16(2), e0246317–e0246317. <https://doi.org/10.1371/journal.pone.0246317>
- Tipirneni, R., Karmakar, M., O'Malley, M., Prescott, H. C., & Chopra, V. (2022). Contribution of Individual- and Neighborhood-Level Social, Demographic, and Health Factors to COVID-19 Hospitalization Outcomes. *Annals of Internal Medicine*. <https://doi.org/10.7326/M21-2615>
- Turshen, M. (1977). The Political Ecology of Disease. *Review of Radical Political Economics*, 9(1), 45-60.
- Turshen, M. (1984). The Political Ecology of Disease in Tanzania. Rutgers University Press.
- White, E. R., & Hébert-Dufresne, L. (2020). State-level variation of initial COVID-19 dynamics in the United States. *PloS One*, 15(10), e0240648–e0240648. <https://doi.org/10.1371/journal.pone.0240648>

Appendix

Fig 1. Odds & Log Odds of Mask-Wearing

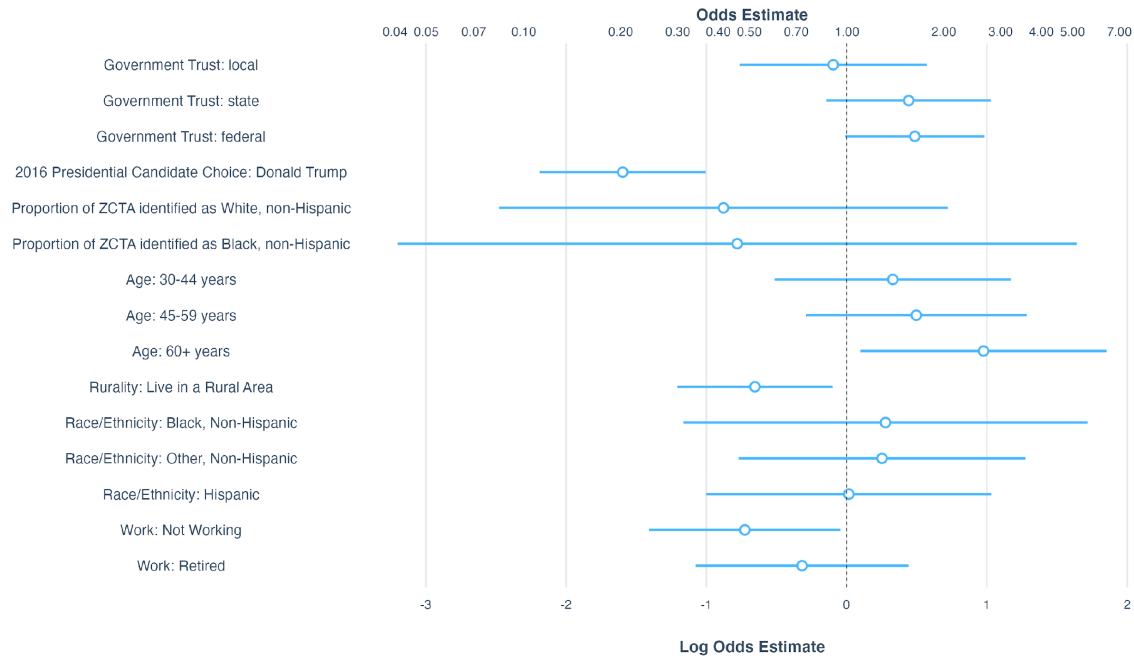


Fig.2: Vaccine Hesitancy Sentiment, June 2021
Trump supporters vs. Other supporters

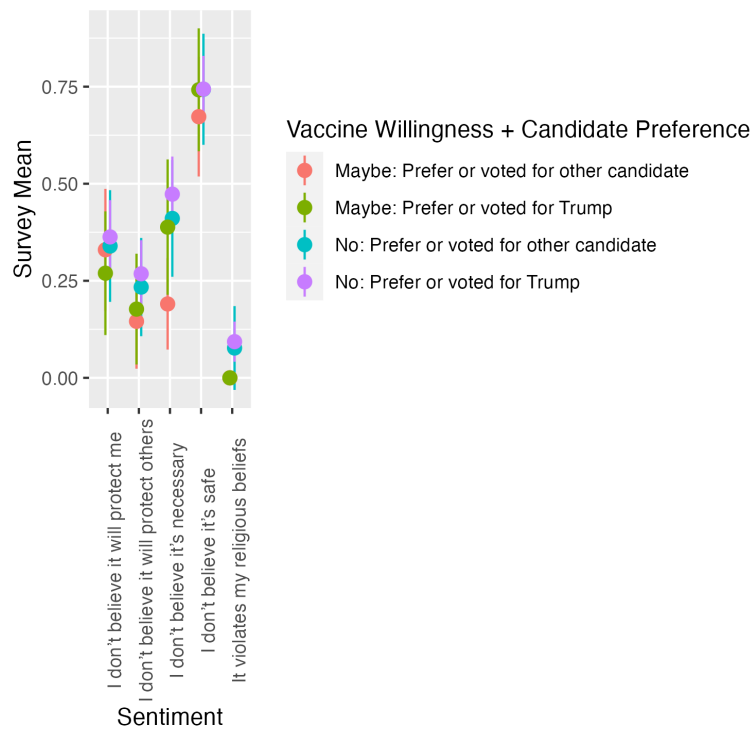


Table 1: Descriptive statistics of the sample population, by survey wave

Survey Wave	Characteristic	Survey Mean*	95% CI: lower	95% CI: upper
Wave 1: September 2020				
	<u>Age</u>			
	18-29 years	0.207	0.180	0.233
	30-44 years	0.254	0.228	0.280
	45-59 years	0.248	0.224	0.273
	60+ years	0.291	0.266	0.316
	<u>Sex</u>			
	Female	0.526	0.497	0.555
	Male	0.474	0.445	0.503
	<u>Race/Ethnicity</u>			
	White, non-Hispanic	0.639	0.610	0.669
	Black, non-Hispanic	0.113	0.093	0.134
	Other, non-Hispanic	0.088	0.071	0.105
	Hispanic	0.160	0.136	0.184
	<u>Work Status</u>			
	Working	0.681	0.654	0.708
	Not Working	0.146	0.124	0.167
	Retired	0.174	0.153	0.194
	<u>Rurality</u>			
	Not Living in a Rural Area	0.825	0.807	0.843

	Living in a Rural Area	0.175	0.157	0.193
	<u>Trust in Government</u>			
	Local Government: Trust	0.166	0.144	0.188
	Local Government: Do not Trust	0.834	0.812	0.856
	State Government: Trust	0.235	0.210	0.259
	State Government: Do not Trust	0.765	0.741	0.790
	Federal Government: Trust	0.419	0.390	0.448
	Federal Government: Do not Trust	0.581	0.552	0.610
Wave 2: December 2020				
	<u>Age</u>			
	18-29 years	0.208	0.177	0.239
	30-44 years	0.251	0.222	0.280
	45-59 years	0.254	0.227	0.281
	60+ years	0.287	0.260	0.314
	<u>Sex</u>			
	Female	0.520	0.487	0.552
	Male	0.480	0.448	0.513
	<u>Race/Ethnicity</u>			
	White, non-Hispanic	0.640	0.607	0.673
	Black, non-Hispanic	0.115	0.091	0.138
	Other, non-Hispanic	0.082	0.063	0.100
	Hispanic	0.164	0.137	0.191
	<u>Work Status</u>			

	Working	0.673	0.643	0.703
	Not Working	0.152	0.127	0.178
	Retired	0.175	0.153	0.197
	<u>Rurality</u>			
	Not Living in a Rural Area	0.827	0.807	0.848
	Living in a Rural Area	0.173	0.152	0.193
	<u>Trust in Government</u>			
	Local Government: Trust	0.149	0.126	0.172
	Local Government: Do not Trust	0.851	0.828	0.874
	State Government: Trust	0.214	0.187	0.240
	State Government: Do not Trust	0.786	0.760	0.813
	Federal Government: Trust	0.353	0.322	0.384
	Federal Government: Do not Trust	0.647	0.616	0.678
Wave 3: June 2021				
	<u>Age</u>			
	18-29 years	0.205	0.171	0.239
	30-44 years	0.248	0.216	0.280
	45-59 years	0.251	0.221	0.280
	60+ years	0.296	0.267	0.326
	<u>Sex</u>			
	Female	0.512	0.476	0.548
	Male	0.488	0.452	0.524
	<u>Race/Ethnicity</u>			

	White, non-Hispanic	0.640	0.603	0.676
	Black, non-Hispanic	0.114	0.087	0.141
	Other, non-Hispanic	0.085	0.064	0.105
	Hispanic	0.162	0.131	0.192
	<u>Work Status</u>			
	Working	0.664	0.630	0.697
	Not Working	0.148	0.120	0.177
	Retired	0.188	0.163	0.213
	<u>Rurality</u>			
	Not Living in a Rural Area	0.824	0.801	0.848
	Living in a Rural Area	0.176	0.152	0.199
	<u>Trust in Government</u>			
	Local Government: Trust	0.156	0.130	0.182
	Local Government: Do not Trust	0.844	0.818	0.870
	State Government: Trust	0.228	0.199	0.257
	State Government: Do not Trust	0.772	0.743	0.801
	Federal Government: Trust	0.308	0.275	0.340
	Federal Government: Do not Trust	0.692	0.660	0.725

*Survey means are weighted to reflect the population of the United States over the age of 18 years

Table 2: Logistic regression results

Variable	Log Odds	95% CI: lower	95% CI: upper	p-value
<u>2016 Presidential Election: Voted for Donald Trump</u>				
No (Voted for/preferred some other candidate)*				
Yes (Voted for/preferred Donald Trump)	-1.596	-2.189	-1.004	< 0.001
<u>Age</u>				
18-29 years*	-	-	-	-
30-44 years	0.329	-0.514	1.172	0.444
45-59 years	0.497	-0.290	1.284	0.216
60 years and over	0.976	0.099	1.853	0.029
<u>Trust in Government</u>				
Federal	0.486	-0.011	0.983	0.055
Local	-0.096	-0.762	0.570	0.778
State	0.442	-0.145	1.030	0.140
<u>Proportion of the ZCTA identified as Black, non-Hispanic</u>	-0.780	-3.200	1.640	0.528
<u>Proportion of ZCTA identified as White, non-Hispanic</u>	-0.878	-2.478	0.722	0.282
<u>Race/Ethnicity</u>				
Black, non-Hispanic	0.277	-1.164	1.717	0.707
Hispanic	0.016	-1.001	1.032	0.976
Other, non-Hispanic	0.252	-0.770	1.275	0.629
White, non-Hispanic*	-	-	-	-
<u>Rurality</u>				
Live in a Rural Area	-0.654	-1.207	-0.101	0.021

Not live in a Rural Area*	-	-	-	-
<u>Survey Wave (Time Period)</u>				
Wave 1: September 2020	4.963	3.379	6.547	< 0.001
Wave 2: December 2020	5.541	3.920	7.163	< 0.001
Wave 3: June 2021	3.503	1.942	5.065	< 0.001
<u>Work Status</u>				
Not working	-0.726	-1.408	-0.043	0.037
Retired	-0.317	-1.076	0.441	0.412
Working*	-	-	-	-

* Variables included in statistical analysis but held as reference.

Table 3: Vaccine hesitancy attitudes and beliefs by Trump support

Voted for Trump in 2016?	Willingness to be vaccinated?	COVID-19 vaccination attitude/sentiment	Survey Mean*	95% CI: lower	95% CI: upper
No	Maybe	I don't it believe it is safe	0.673	0.519	0.827
No	Maybe	I don't it believe it is necessary	0.19	0.073	0.308
No	Maybe	I don't it believe it will protect me	0.33	0.173	0.487
No	Maybe	I don't it believe it will protect others	0.145	0.023	0.267
No	Maybe	It violates my religious beliefs	0	0	0
No	No	I don't it believe it is safe	0.743	0.6	0.886
No	No	I don't it believe it is necessary	0.411	0.26	0.561
No	No	I don't it believe it will	0.34	0.196	0.483

		protect me			
No	No	I don't it believe it will protect others	0.234	0.107	0.36
No	No	It violates my religious beliefs	0.077	-0.031	0.185
Yes	Maybe	I don't it believe it is safe	0.742	0.584	0.9
Yes	Maybe	I don't it believe it is necessary	0.388	0.214	0.563
Yes	Maybe	I don't it believe it will protect me	0.269	0.11	0.429
Yes	Maybe	I don't it believe it will protect others	0.177	0.034	0.32
Yes	Maybe	It violates my religious beliefs	0	0	0
Yes	No	I don't it believe it is safe	0.744	0.658	0.829
Yes	No	I don't it believe it is necessary	0.473	0.376	0.57
Yes	No	I don't it believe it will protect me	0.363	0.268	0.458
Yes	No	I don't it believe it will protect others	0.268	0.181	0.355
Yes	No	It violates my religious beliefs	0.093	0.041	0.145

*Survey means are weighted to reflect the population of the United States over the age of 18 years

[Under Peer Review Do Not Cite or Share]

Political and Educational Dynamics in Religious Group's Mask Noncompliance under COVID-19

Junhe Yang (UW), Zack W. Almquist (UW), James Holland Jones (Stanford)

Acknowledgments

Partial support for this research came from a Eunice Kennedy Shriver National Institute of Child Health and Human Development research infrastructure grant, P2C HD042828, and Shanahan Endowment Fellowship, and a Eunice Kennedy Shriver National Institute of Child Health and Human Development training grant, T32 HD101442, to the Center for Studies in Demography & Ecology at the University of Washington. Partial support for this research also came from an NSF Award #BCS-2028160 and an ARO Award #W911NF-19-1-0407. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, NSF, or ARO.

Abstract

This study investigates the mediation effect of conservative political ideology on the relationship between religiosity and the noncompliance of masks under COVID-19, and the moderated effect of education on the mediation relationship. Using nationally representative data from COVID-19 and Social Distancing Survey, a moderated mediation analysis is applied to examine the pathway from religiosity via political conservativeness to the noncompliance of masks, and the interaction effect between years of education and political ideology. A logistic regression model is used to investigate each path in the mediation analysis. Results suggest that controlling for socio-demographic background, religiosity positively drives mask noncompliance. Further, findings confirm that political orientation is not only an established predictor of the polarized public support of masks as found in existing studies, but is also a key mechanism by which religiosity positively predicts mask resistance. Finally, education is associated with stronger political polarization of public opinions on masks during the pandemic.

Introduction

Since its onset in December 2019, the novel coronavirus has soon become a global pandemic. According to the Center for Disease Control, in the US, COVID-19 has claimed over one million deaths over less than three years (2022). The pandemic has left severe health and financial harm to numerous families in the country, created unemployment, and disproportionately affected people of color and lower-income households (Kochhar and Sechopoulos 2022; Vasquez Reyes 2020). To prevent the spread of the disease, the CDC has recommended public health guidelines such as wearing masks, social distancing, cleaning and disinfecting, and vaccination (CDC 2020). Public opinion on these mandates, however, is divided. Mask-wearing stands at the center of the cultural and ideological debates over the pandemic, becoming a gesture that signals the population's divergent religious and political identities. It is therefore both of theoretical and

practical importance to understand key mechanisms that predict public compliance with masks under COVID-19.

The present study investigates the effect of religiosity on the support for masks under COVID-19, and how conservative political ideology may mediate this effect. Studies mostly find a negative effect of religiosity on mask-wearing (Gonzalez et al. 2021; Hill, Gonzalez, and Burdette 2020; Minton, Wang, and Anthony 2022; Schnabel and Schieman 2022), but fail to empirically identify how religiosity positively predicts mask noncompliance through the mechanism of political orientation, and how this dynamic relationship of religious and political identities on masks differs across people's education and cognitive capabilities. Through the moderated mediation analysis in this study, we quantify answers to the following questions:

- What is the effect of religiosity on mask-wearing noncompliance under COVID-19?
- How does religiosity affect mask-wearing noncompliance through political conservatism?
- How does the indirect effect of religiosity on mask noncompliance via political ideology differ by years of education?

A visual presentation of the modeled relationship of the above questions is presented in section (3) Material and Methods. Through studying these questions, this study provides evidence that political conservativeness as operationalized by the support for Trump channels the pathway from religiosity to an anti-mask stance. It is also the first to test if more education is associated with stronger political polarization on an issue as straightforward as mask-wearing, a relationship that is previously only identified with more complex and contentious science topics such as climate change, and the Big Bang (Drummond and Fischhoff 2017; L. C. Hamilton 2011).

The remainder of the paper is divided into four sections. In section (2) Background and Hypothesis, we summarize the existing studies this research builds upon: religiosity and the resistance of masks; the relationship between religiosity, conservative political ideology, and the polarization of masks under COVID-19; and the multiplicative relationship between political polarization and education on science topics. This section also outlines the key hypotheses generated from the literature. In section (3) Materials and Methods, we introduce the survey dataset used, the model design, the variables of interest, and present descriptive statistics of the variables. We next discuss findings from the models and the moderated mediation analysis in section (4) Findings. In the final section (5), we discuss the implications and limitations of the findings and conclude the study.

Background and Hypotheses

In this section, we lay out the literature and how they are related to our three key hypotheses. These hypotheses, if held, suggest the mechanisms by which we show how religious conservatism leads to politically conservative views, which subsequently increases noncompliance to masks.

Religiosity and mask-wearing under COVID

Studies on the effect of religiosity on mask-wearing under COVID-19 mostly report a negative relationship explainable by multiple reasons. For example, one major factor that explains the conservative protestants' anti-mask attitudes is the group's distrust of science (Alder and Schaeublin 2020; Burge 2020a; 2020b; Evans 2013; Gauchat 2008; Hill, Gonzalez, and Burdette 2020). An empirical study by Evans (2013) finds a declined confidence in scientists' social-moral influence over the society from the conservative protestants that stand with biblical literalism. The conflicts between religion and science evolved beyond the epistemological contentions about the truth of the world, to power struggles between the two as social institutions (Evans and Evans 2008). It is thus believed that with an alternate worldview, the religious group rejects the scientific paradigm. Religious beliefs serve as an ideological foundation against science and public health guidelines under COVID-19 (Gonzalez et al. 2021; Hill, Gonzalez, and Burdette 2020).

Another related explanation for the religious group's reluctance to take precautionary actions against the virus is their perceived lack of control of the matter, and their beliefs in biblical prophecies of the apocalypse (Alder and Schaeublin 2020). Polls show that among religious Americans, two-thirds believe that the virus is "God telling humanity to change the way we are living" (Fowler 2020), and a trust in God will protect them from COVID-19 (Schlor and Fingerhut 2020). Minton et al. (2022) find that religiosity is associated with reduced prosocial behavior because of a perceived lack of personal control over the crisis. Schnabel and Schieman (2022) reach a similar conclusion that the religious group see less of a crisis in COVID-19, suffer from less distress, and are more reluctant to support public health restrictions aiming to contain the spread of the disease.

Masks are also regarded as an infringement of personal and religious freedom. Kahn (2022) describes mask-wearing resistance as "semiotic expressions of hate". Mask represents a sign of social repression and control in the eyes of the conservatives (Kahn 2022). Individual constitutional rights are viewed as in conflict with indoor mask requirements, the restrictions on indoor assembly and attendance of religious services, and the possible police enforcement of such mandates (Flood et al. 2020; Fowler 2020). Media reports for example, that a father of two teens who refused to wear masks at school claimed "The Bible says we're made in the image of God and Satan tries to cover that up. A mask is a sign of oppression" (Cherry 2021).

Further, religion as an institution and a schema has effectively indoctrinated anti-mask views through religious practices and leadership (Hill, Gonzalez, and Burdette 2020; Schnabel and Schieman 2022). Pastor Locke of the Global Vision Bible Church in Mount Juliet, Tennessee, for example, believed that COVID was a hoax, and threatened to ban mask-wearing parishioners from his service amid the high transmission of the Delta variant in the state (Peiser 2021).

From belief system to practice, religiosity is expected to lead to noncompliance with mask-wearing. Based on the existing academic discussions, we propose the following hypothesis:

Hypothesis 1: Religiosity has direct positive impacts on mask noncompliance under COVID-19.

Religion, Political Conservative Ideology, and Political Polarization of Masks

Political conservatism represents forces that stand against “big government”, the increase of tax, and its possible consequences of reduced personal freedom (Gonzalez et al. 2021). Under Covid-19, the politically conservative republicans hold more negative views against science and public health measures. They are more inclined to believe that efforts to contain the spread of the disease are politically motivated (Hill, Gonzalez, and Burdette 2020). Compared to the liberals, the political conservative group shows stronger resistance to masks. A media report citing an analysis of poll data in mid-2020 shows that 75% of Democrats wore masks in public places while the percentage was greatly lower at 53% for Republicans (Lauter 2020). Studies confirm that in the US, partisanship robustly and powerfully predicts mask use (Albrecht 2022; Boykin et al. 2021; Flaskerud 2020; Gonzalez et al. 2021; Howard 2022; Kahane 2021; Milosh et al. 2021; Perry, Whitehead, and Grubbs 2020). In this study, we operationalize the measurement of political conservatism as the support over the Republican party leader President Donald Trump. The support for Trump consistently and unanimously represents political conservatism that it is treated as a valid indicator of this dimension in the US (Gonzalez et al. 2021).

Christian protestants have strongly aligned with contemporary US political conservatives. This alignment was further strengthened during the Trump era. Trump had targeted outreach to Christian conservatives. In his 2016 campaign, for example, Trump made the promise that “(with) A Trump administration, our Christian heritage will be cherished, protected, defended, like you’ve never seen before” (Politico 2016). His campaign strongly appealed to the white Christian rural population that hoped to gain power against pluralist democracy (Rubin 2022). Scholars have termed the combination of religious and political conservatism as “Christian nationalism” that represents “a hyper-partisan and ultra-conservative ideology that has already been shown to lower Americans’ trust in science and scientific expertise ... promote a view of (conservative Christian) Americans as God’s chosen ... (and) bind them to siding with Trump” (Perry, Whitehead, and Grubbs 2020, 406). Patriot church is an example of Christian nationalism ideology, convening a group that not only believes that America should be a Christian country but also is fervently engaged in seeking political power to rebuild the country to defend its Christian faith (Bailey 2020).

The political conservatives’ anti-big government and anti-science stances align with the religious conservatives’ world views. Trump’s policy aim of pruning the Washington bureaucrats, officials, and lobbyists (“drain the swamp”) speaks to the core conservative principle of constraining the power of government and protecting personal freedom (Gonzalez et al. 2021; Hudson 2018). Anti-science in rhetoric and policy practices, the Trump administration proposed multiple budget cuts to agencies such as National Institutes for Health and the Environmental Protection Agency (Frickel and Rea 2020). In the early stages of COVID-19, the

Trump administration understated the seriousness of the virus, calling it fraudulent, fake, and a hoax fabricated by the “deep state” (Bunch 2020; Gonzalez et al. 2021; Van Bavel 2020). Repeatedly during his presidency and campaign, Trump attacked Dr. Fauci on coronavirus topics, calling him a disaster and an idiot (CNN Newsroom 2020).

Trump’s mixed reactions toward mask use have undoubtedly created misconceptions and encouraged noncompliance with masking (Sami 2020). In an extensive interview with Wall Street Journal, Trump expressed the opinion that people wore masks “not as a preventative measure but as a way to signal disapproval of him” (Bender 2020; Scott 2020). A month later, he wore a mask in public for the first time at Walter Reed, sparking online discussions and positive sentiments toward mask use (Milosh et al. 2021). Trump’s shifted rhetoric and actions on mask use catalyzed masks as a political symbol that further divided public opinions. As a result, masks stand out as the most visible political signal under the background of the bipartisan pandemic.

While scholars often agree that religious and political conservativeness stand as two pillars against mask use under COVID for the reasons mentioned above (Gonzalez et al. 2021), and a group of Christian nationalists share a mingled ideology that is both religiously and politically conservative (Baker, Perry, and Whitehead 2020; Perry, Whitehead, and Grubbs 2020), there have been no studies so far that empirically examined if and how political orientation may have further promoted the anti-science stance of the religious group under COVID-19. In other words, different dimensions of conservatism coexist but how one aligns and further promotes the other against the scientific paradigm remains understudied. Trump’s anti-science stance during COVID-19 has not only appealed to his conservative supporters who distrust scientists but also further aggregated the cultural and partisan polarization among Americans on issues with the pandemic. To white evangelicals, for example, the Trump administration became a more reliable information source than public health officials or the media regarding COVID (Burge 2020a). It is also shown that protestant Christians are less fearful of the pandemic not because of faith, but because of their trust in Trump (Burge 2020b). The study of Gonzalez et al. (2021) using US county-level data finds that the relationship between the percentage of evangelicals in the population and mask resistance is weakened when the Trump effect is adjusted for. Evidence such as these supports the need to empirically examine if political conservative ideology mediates the pathway from religiosity to mask noncompliance, which leads to the following hypothesis:

Hypothesis 2: Adherence to politically conservative ideology mediates the association between religiosity and noncompliance with masks.

Interaction Effect between Political Ideology and Education

Scholars have identified an interaction effect between political ideology and educational achievement over disputed science and environment topics such as climate change, human evolution, and the Big Bang (Drummond and Fischhoff 2017; L. C. Hamilton 2008; L. C. Hamilton, Colocousis, and Duncan 2010; L. C. Hamilton 2011; L. C. Hamilton, Hartter, and

Saito 2015; L. C. Hamilton and Saito 2015; Malka, Krosnick, and Langer 2009; McCright 2011). It is agreed that contentions over these topics are ideologically based and deeply mapped to individual political identities (Drummond and Fischhoff 2017; L. Hamilton and Safford 2020; McCright 2011). For example, studies by Hamilton et al. using different survey data have consistently found the polarization between the Democrats and Republicans on issues such as anthropogenic climate change grows with education (L. C. Hamilton 2008; 2011). Debates on mask-wearing and other public health measurements under COVID-19 are similarly ideologically mapped but few have examined whether an education interaction effect exists here.

Theoretical frameworks in human psychology offer several explanations for the interaction effect between political ideology and education. One category as summarized by Drummond and Fischhoff (2017) involves motivated reasoning that the better educated are more capable of selecting the information that speaks to their prior beliefs, which may lead to *confirmation bias* (Brulle, Carmichael, and Jenkins 2012; L. C. Hamilton, Hartter, and Saito 2015; Kahan et al. 2012; Klayman and Ha 1987; Kunda 1990). The other explanation is the *overconfidence* of the more knowledgeable group (Drummond and Fischhoff 2017; Moore and Healy 2008). A recent study by Light et al. (2022) finds that both on questions with clear scientific consensus and on COVID-19 mitigation measures, the group showing the strongest objection commands the lowest levels of actual knowledge, but the highest level of confidence in their knowledge. On matters as politically polarized as precautionary measures under COVID-19, more education is associated with a higher assessment of subjective knowledge, possibly leading to more extreme views regardless of the actual content of the knowledge (Drummond and Fischhoff 2017; Light et al. 2022; Tversky and Fox 1995).

Based on these existing studies, we propose to test the possible interaction effect of education on the mediated relationship between religiosity and support of masks, coming up with the following hypothesis:

Hypothesis 3: Political conservatives with higher education are more strongly opinionated against compliance to science paradigm, resulting in an educational difference in the indirect effect of religiosity on mask noncompliance through political conservatism.

Materials and Methods

Data and Measurements

We use a nationally representative sample of the over 18 population in the United States from the NSF sponsored Behavioral Change COVID-19 (BCC19) Survey to address the research questions in this study. The survey was fielded by Ipsos Public Affairs (Ipsos) and contains three waves. The main survey period for the first wave is 8/31/2020 to 9/15/2020; for the second wave, 12/11/2020 to 1/11/2021; for the third wave, 5/21/2021 to 6/13/2021. The survey's target population is non-institutionalized adults that reside in the US, with an overrepresentation of rural residents. All household members are enumerated in the survey, but for the purpose of this

study, we keep the population aged 18 and above. Ipsos recruited general population survey respondents using KnowledgePanel, a large online panel platform with probability-based sampling techniques (this is effectively an address based sample of the United States). The sampling frame is developed based on geo-demographic indicators of gender, age, race/ethnicity, education, census region, household income, home ownership status, metropolitan area, and Hispanic origin. Design weights are derived based on sampling selection and nonresponse, and final weights are produced with post-stratification adjustments such as raking, trimming, and scaling. The original survey datasets contain 1421 observations for wave 1, 1186 observations for wave 2, and 1010 observations for wave 3. We choose to include members that appear in all three waves of the survey and are left with a balanced panel of 2,499 observations in total. Table 1 represents an accounting of observations after each step of data processing.

Table 1. Accounting of Number of Records with Steps of Data Preparation

Data preparation steps	Number of total cases	Percentage removed of total	Category of data elimination
Step 0. Raw dataset	3,617	N.A.	N.A.
Step 1. Create a balanced panel	2,934	18.9	Respondents that do not participate in all three waves of the survey
Step 2. Keep only adult cases	2,499	12.0	Household members aged below 18

Mask-wearing noncompliance under COVID-19 is treated as the outcome variable. It is derived from the following survey question: “Given the current state of the world, mask-wearing should be mandatory in my community when people interact within 6ft (e.g. within stores).” Responses were coded so that lower scores indicate “strongly agree” (2) and higher scores indicate “strongly disagree” (6). We recoded the variable by making it binary, where 0 indicates “strongly agree”, and 1 otherwise.

Self-measured religiosity, the focal independent variable, is derived from the question “Do you consider yourself a religious person?” It is binary with 1 representing “yes”, and 0 “no”. Variable devotion to Trump is measured by a question from wave 1 asking which candidate the respondent chose for the 2016 presidential election. It is assumed that the political orientation of respondents stayed consistent across the data collection period of all three waves. The variable assumes 1 if the response is Trump, and 0 otherwise. Years of education is derived from the variable of the highest degree received: “No formal education” corresponds to 0; “5th or 6th grade” corresponds to 5.5; “7th or 8th grade” corresponds to 7.5; “9th grade” corresponds to 9;

“10th grade” corresponds to 10; “11th grade” corresponds to 11; “12th grade No Diploma” corresponds to 11.5; “High school graduate - high school diploma or the equivalent (GED)” corresponds to 12; “Some college, no degree” corresponds to 13; “Associate degree” corresponds to 14; “Bachelor’s degree” corresponds to 16; “Master’s degree” corresponds to 18; “Professional or Doctorate Degree” corresponds to 22. Additional covariates include the respondent’s age, ethnicity, sex, urbanicity, marital status, and income. Variable ethnicity contains four categories: “White, Non-Hispanic”, “Black, Non-Hispanic”, “Hispanic”, and “Other”. Urbanicity assumes a value of 1 when the respondents reside in urban or suburban

Table 2. Descriptive Statistics

	Mean	SE	Min	Max	Count
<i>Outcome variable</i>					
Mask Noncompliance	0.50	0.011	0	1	2,499
<i>Independent variable</i>					
Religiosity	0.58	0.011	0	1	2,499
<i>Mediator</i>					
Trump	0.38	0.010	0	1	2,499
<i>Moderator</i>					
Years of Education	14.05	0.068	0	22	2,499
<i>Controls</i>					
Age	41.25	0.29	18	74	2,499
Married	0.65	0.011	0	1	2,499
Household income ('000)	95.22	1.50	2.5	300	2,499
Female	0.52	0.011	0	1	2,499
Urban	0.82	0.007	0	1	2,499
Ethnicity					

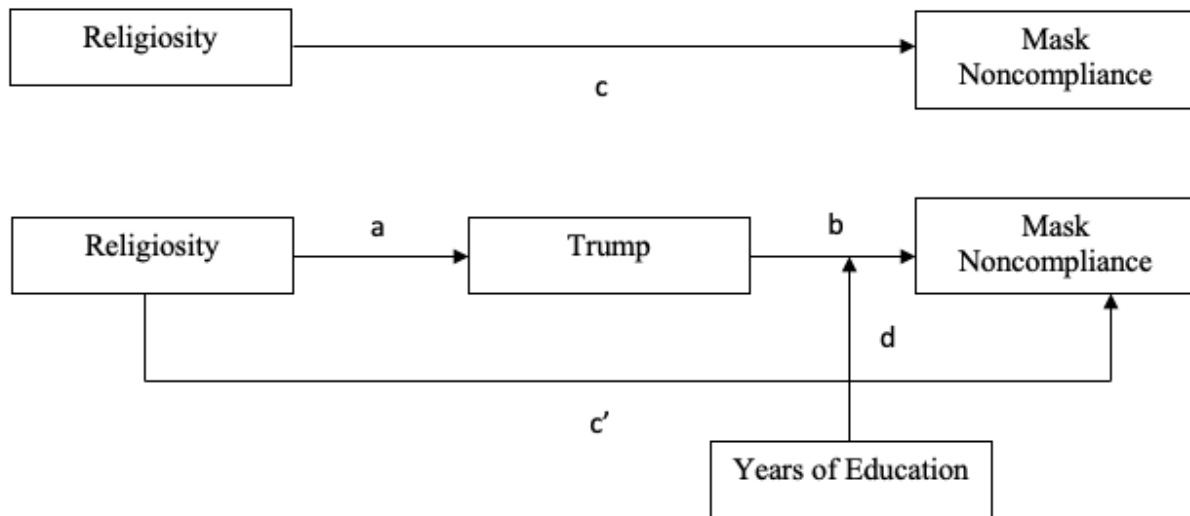
White, Non-Hispanic	0.68	0.011	2,499
Black, Non-Hispanic	0.11	0.008	2,499
Hispanic	0.14	0.009	2,499
Other	0.08	0.006	2,499

areas, and 0 if they are rural. Sex (female) and marital status (married) are binary variables as well. Finally, we encode annual household income ('000) as the mean of the bucket range chosen by survey respondents. Table 2 presents the descriptive statistics of the focal dependent and independent variables, as well as the covariates. Mean and standard errors are weighted estimations derived using the Survey package of the statistical software R.

Methods

Figure 1 below presents the study model as described in this section.

Figure 1. Proposed Moderated Mediation Model



Logistic regression is employed to explore each pathway from religiosity to the support of masks via political conservatism (Trump support). It models the log odds of an event occurring as a linear combination of covariates. Specifically, in our analysis, a four-staged modeling process is employed to test the total, indirect and direct effects of religiosity on mask noncompliance through support over Trump as specified in Figure 1, following an illustration adopted by Yoon, Ro, and Cho (2019).

$$\text{logit}(m_{it}) = \beta_0 + \beta_1 R_{it} + \beta_2 E_{it} + \gamma C_{it} + \delta_2 W2_i + \delta_3 W3_i + \varepsilon_{it} \quad \text{Eq1}$$

Eq1 describes the total effect model (c) where the logit of the probability of mask-wearing noncompliance (m_{it}) is modeled against the focal independent variable religiosity (R_{it}) without adjusting for the Trump effect. Since we are pooling three waves of data together, R_{it} represents the religiosity of respondent i in wave t ; The moderator years of education (E_{it}) is included as one of the controls in this model; C_{it} represents a vector of control variables: age, ethnicity, sex, urbanicity, marital status, as well as income. Two dummy variables of wave 2 and wave 3 are added to control for wave-fixed effects. ε_{it} represents the error term.

$$\text{logit}(t_i) = \beta_0 + \beta_1 R_{it} + \beta_2 E_{it} + \gamma C_{it} + \delta_2 W2_i + \delta_3 W3_i + \varepsilon_{it} \quad \text{Eq2}$$

In the second stage, we model the effect of religiosity on our hypothesized mediator (relationship a) where t_i in this stage is the probability of supporting Trump in the 2016 presidential election for respondent i . Note that the variable is independent of survey wave t . The second model contains the same set of independent variables and control variables on the right-hand side of the equation as Eq1.

$$\text{logit}(m_{it}) = \beta_0 + \beta_1 R_{it} + \beta_2 E_{it} + \beta_3 T_i + \gamma C_{it} + \delta_2 W2_i + \delta_3 W3_i + \varepsilon_{it} \quad \text{Eq3}$$

With the third stage, we model the outcome variable of interest, the probability of mask-wearing (m_{it}) against the same set of predictors, plus the mediator support of Trump (T_i) in addition to the control variables listed in Eq1 and Eq2. In this way, we test the religiosity-adjusted effect of political conservatism on mask-wearing (b) and the political ideology-adjusted effect of religiosity on mask-wearing (c').

A mediation pathway establishes if the total effect (c), the indirect effects (a), and (b) are all found to be significant, and the magnitude of the direct effect (c') is smaller than that of the total effect $|c|$ ($|c'| < |c|$) (Yoon, Ro, and Cho 2019). In this study, we use the Mediation package in the statistical software R to perform the mediation analysis with bootstrapped p value and confidence intervals, relying on the assumption of sequential ignorability (Imai, Keele, and Tingley 2010) that the direct and indirect relationships specified above have no uncontrolled confounders.

$$\text{logit}(m_{it}) = \beta_0 + \beta_1 R_{it} + \beta_2 E_{it} + \beta_3 T_i + \beta_4 T_i E_{it} + \gamma C_{it} + \delta_2 W2_i + \delta_3 W3_i + \varepsilon_{it} \quad \text{Eq4}$$

At the final stage, we model the same set of predictors as Eq3 on the probability of mask-wearing noncompliance, adding the interaction between the mediator Trump (T_i) and

moderator years of education (E_{it}) as one of the controls. β_4 will show how education interacts with the effect of political conservatism (d).

Findings

In this section we report findings from the main models, the mediation analysis, as well as the moderated mediation analysis.

Analysis

Results of logistic regressions of the total effect model, the mediator model, the direct effect model, as well as the moderated mediation model are presented in Table 3. Clustered standard errors are reported to correct for serial correlations. Model 1 presents the results of the total effect model testing if religiosity positively predicts mask noncompliance without adjusting for the Trump effect. Controlling for socio-demographic background, the religious group has a 68.4% higher odds of not complying with masks compared to the non-religious group. Model 2 tests the effect of religiosity on the mediator of support over Trump. It is shown in model 2 that controlling for socio-demographic covariates, compared to the non-religious group, the population self-identified as religious has a 162.6% higher odds to support Trump. As expected, populations with higher educational attainment, being female (compared with male) and urban (compared with rural) are less likely to support Trump. Age is related to slightly stronger support of Trump. Compared to the non-Hispanic Black population, all other three ethnicities show greater alliance with Trump, with the non-Hispanic White population reporting the highest odds ratio of 14.23.

Model 3 reports the results of Eq3 that measures the direct effect of religiosity on mask-wearing noncompliance, and the indirect effect of support over Trump on mask-wearing noncompliance without adjusting for Trump-Education interaction. Controlling for individual socioeconomic backgrounds, results also confirm a positive and significant relationship between religiosity and mask-wearing noncompliance. Specifically, compared to the non-religious group, the population who self-reported to be religious has a 29.7% higher odds of not complying with masks. Note that the effect of religiosity here in the direct effect model is smaller than the effect as found in the total effect model ($|c'| < |c|$), but is still nonnegligible and statistically significant at the level of 0.000. Results of Models 1 to 3 confirm the establishment of the mediation relationship.

Model 4 shows that the odds ratios of the effect of supporting Trump, years of education, and their interaction on mask noncompliance accord with Hypothesis 3 that a deep political polarization exists with public mask-wearing support. Such a gap grows with more education and knowledge. For less of a scientific debate than a political gesture compared to topics of stem cell research, the Big Bang, and human evolution discussed in the previous literature (L. C. Hamilton 2011; Drummond and Fischhoff 2017; Light et al. 2022), public opinion toward mask-wearing is strongly divided along the political line. Specifically, controlling for religiosity and other socio-demographic covariates, for non-Trump supporters, a one-year increase in education

reduces mask-wearing noncompliance by 9.3%. For Trump supporters, however, a one-year increase in education is associated with a 6.74% higher odds of mask noncompliance.

Better-educated Trump supporters are greatly motivated in reasoning against their opposite political stance and are more confident in their beliefs (Drummond and Fischhoff 2017).

Directions of other sociodemographic control variables proved to be as expected. A one-year increase in age is associated with a 2.6% decrease in odds of mask-wearing noncompliance, a reflection of the elderly population's better self-protection given their vulnerability against COVID-19. Compared to males, females have a lower odds of 19.7% of not

Table 3. Modeling Results (Logistic Regression)

	Dependent Variable			
	Mask Noncompliance (1)	Trump (2)	Mask Noncompliance (3)	Mask Noncompliance (4)
Wave 2	0.871 (0.052)	1.017 (0.010)	0.854 (0.056)	0.853 (0.057)
Wave 3	6.026*** (0.590)	1.022 (0.010)	7.063*** (0.773)	7.189*** (0.015)
Religiosity	1.684*** (0.208)	2.626*** (0.397)	1.297*** (0.163)	1.288** (0.162)
Trump			4.201*** (0.590)	0.429 (0.305)
Years of Education	0.934*** (0.023)	0.907*** (0.030)	0.956** (0.023)	0.907*** (0.025)
Age	0.981*** (0.005)	1.019*** (0.006)	0.974*** (0.005)	0.974*** (0.005)
Ethnicity: Hispanic	1.084 (0.357)	5.047*** (3.027)	0.817 (0.277)	0.761 (0.259)
Ethnicity: Non-Hispanic White	2.090*** (0.542)	14.233*** (7.683)	1.211 (0.334)	1.251 (0.346)
Ethnicity: Other	1.266 (0.420)	12.018*** (7.107)	0.742 (0.255)	0.810 (0.283)
Female	0.682*** (0.086)	0.542*** (0.087)	0.797** (0.103)	0.803** (0.104)
Urban	0.679*** (0.096)	0.759*** (0.132)	0.721*** (0.105)	0.711*** (0.104)
Married	1.126 (0.162)	1.111 (0.205)	1.108 (0.168)	1.129 (0.171)
Household Income ('000)	1.002** (0.001)	1.000 (0.001)	1.002** (0.001)	1.002** (0.001)
Trump X Education				1.177*** (0.059)
Constant	0.501** (0.223)	0.092*** (0.065)	1.783* (0.834)	3.592*** (1.782)
Observations	2,499	2,499	2,499	2,499
Log Likelihood	-1,471.561	-1,493.538	-1,369.796	-1,359.583
Akaike Inf. Crit.	2,969.122	3,013.075	2,767.591	2,749.166

Note:

*p<0.1; **p<0.05; ***p<0.01

complying with masking. Compared to rural residents, the population living in urban and suburban areas are with a 28.9% lower odds of rejecting masks. A 1000-dollar higher income is associated with an increased odds of mask noncompliance of 0.2%. No statistically significant differences in mask-wearing noncompliance are observed among different races and ethnicities,

or between married versus the not married.

Mediating Effect of Trump

Results of bootstrapped mediation analysis confirm the role of the support for Trump as a mechanism in explaining the religious group's anti-masking stance, as shown in Table 4. The proportion of the total effect of religiosity on mask-wearing noncompliance mediated by their support for Trump is large at 53%. In other words, by politicizing mask-wearing as a bi-partisan debate, political conservatism plays a salient role in leading the religious group to turn more against mask-wearing. For the religious non Trump supporters, the magnitude of religiosity on their mask resistance is much lower at a log odds of 0.047.

Table 4. Mediation Analysis Results

Total Effects (R->M)		Direct Effect (R->M)		Indirect Effects of R on M				
β	p-value	β	p-value	Hypothesis	β	p-value	95% confidence interval	
							Lower	Upper
0.10	0.000	0.047	0.016	R->T->M	0.053	0.000	0.042	0.07

Note: R: Religiosity; T: Trump; M: mask-wearing

Moderated Mediation Effect of Education

As an additional way to illustrate the moderate effect of years of education on the indirect effect of religiosity on mask-wearing via Trump, we create two explicit levels of years of education variables at one standard deviation below and above the mean ($\mu_e \pm SD_e$). An analysis is then conducted to test if there exist statistically significant differences in the indirect and direct effects from mediation analyses of the models controlling for these two levels of the education variable (Tingley et al. 2014). With this step, we test Hypothesis 3 with a bootstrapped approach using the Mediation package to evaluate the level of indirect effects at two different levels of the moderator (Hayes 2013): the highly educated group with years of education one standard deviation above the mean (17.10), and a lower educated group with years of education one standard deviation below the mean (11.27).

The level of education is found to moderate the indirect effect of religiosity on mask noncompliance through political conservatism. Specifically, for the higher-educated group, the indirect effect is 0.07 (95% CI = 0.05, 0.09); for the lower-educated group, the indirect effect is 0.034 (95% CI = 0.027, 0.05). Moderated mediation analysis further shows that the difference of the indirect effects found between the two groups is significantly different from zero, with the higher educated group seeing a larger effect of religiosity on mask-wearing noncompliance

through Trump by a log odds of 0.035 ($p < 0.05$). In other words, the conditional indirect effect is strongest in the highly educated group and weakest in the lowly educated group.

Table 5. Moderated Mediation Analysis Results

Lower educated group						Higher educated group					
Total effect		Direct Effect		Indirect Effect		Total effect		Direct Effect		Indirect Effect	
β	p-value	β	p-value	β	p-value	β	p-value	β	p-value	β	p-value
0.083	0.000	0.049	0.008	0.034	0.000	0.11	0.000	0.045	0.018	0.07	0.000

Discussion

The world is struggling to recover from the pandemic, but the spread of the virus has not stopped. Even with the majority of the US population having gained COVID-19 antibodies either through vaccination or infection (Crist 2022), masks remain the simplest effective measure against the disease and the harm it brings. In this study, we explore factors that are behind the US population's contention over masks, and more broadly the scientific paradigm it represents. Using individual-level panel data, we first hypothesize that out of an alternative belief system to science about the driver of the pandemic and its possible impacts, the religious group will show resistance to masks compared to the non-religious. Our models in this study consistently support this hypothesis. We then dive further into the political stance of the religious population, and hypothesize that political conservative ideology is a salient mediator in the pathway from religiosity to mask-wearing noncompliance. Results of the mediation analysis support this hypothesis and confirm that support over Trump composes a significant indirect effect that explains the religious group's anti-mask stance, the magnitude of this indirect effect outweighing the direct effect of religiosity itself. Further, with previous studies identify an interaction effect between political polarization and education on contentious science topics (Drummond and Fischhoff 2017; L. C. Hamilton, Colocousis, and Duncan 2010; L. C. Hamilton 2011; L. C. Hamilton, Hartter, and Saito 2015), this study hypothesizes that higher education will enlarge the bipartisan gap on masks, an issue much more straightforward but no less politically contentious than climate change, for example. Our moderated mediation analysis supports this hypothesis by finding that the highly educated group experienced a significantly larger indirect effect of religiosity on the mask-wearing noncompliance through Trump than the lower educated.

Analyses in this study accord with the implications made by Gozalez et al. (2021) that the support for Trump mediates the effect of religious conservativeness against masks. It further extends the findings of existing literature by providing the first empirical evidence of such a mechanism, and by revealing the educational gap in this debate that is greatly tied to the

population's diverged political identities. Theoretically, the conclusions in this study contribute to a deeper understanding of the role of ideologies and human motivations in the face of great hardships in society. Practically, this study points to the perils of over-politicizing the mitigation measure amid a public health crisis, and calls for policies that better speak to the population with less trust in science from the start in the event of future crises.

Limitations of this study point to meaningful next steps of research. In terms of data, a larger sample would enable researchers to better address population heterogeneity, such as a more focused analysis of Christian evangelicals, and identify unique patterns within specific geographic contexts. A larger sample would also bring up the possibility of examining additional mediators, such as the role of news media in forging and strengthening partisan polarization within the country (Gonzalez et al. 2021; Perry, Whitehead, and Grubbs 2020). With a general-purpose survey, our study relies on limited measurements of religious and political ideology, which leaves space for finer-grained research that, for example, could make a distinction between the impacts of political ideology and political party affiliation (Howard 2022), and strong political leaders. More also remains to be identified about the education-political ideology interaction. For example, if the phenomenon is unique within the US, and if the anti-mask stance of the highly educated religious and political conservatives appears to be in congruence with their attitudes on climate change and other contentious science topics. Future research along these directions would not only give us a better understanding of public opinions and reactions to a pandemic, but will also contribute to a more comprehensive academic discussion of the ideological and cultural diversities of the US population.

References

- Albrecht, Don. 2022. "Politics and the Spread of COVID-19 in the United States." *Medical Research Archives* 10 (2). <https://doi.org/10.18103/mra.v10i2.2681>.
- Alder, Cora, and Emanuel Schaeublin. 2020. "US Evangelicals: From Prophecy to Policy." *Policy Perspectives* 8 (11): 1–4. <https://doi.org/10.3929/ETHZ-B-000438157>.
- Bailey, Sarah. 2020. "Seeking Power in Jesus' Name: Trump Sparks a Rise of Patriot Churches." *Washington Post*, October 26, 2020. <https://www.washingtonpost.com/religion/2020/10/26/trump-christian-nationalism-patriot-church/>.
- Baker, Joseph O., Samuel L. Perry, and Andrew L. Whitehead. 2020. "Crusading for Moral Authority: Christian Nationalism and Opposition to Science." *Sociological Forum* 35 (3): 587–607. <https://doi.org/10.1111/socf.12619>.
- Bender, Michael C. 2020. "WSJ News Exclusive | Trump Talks Juneteenth, John Bolton, Economy in WSJ Interview." *WSJ*. June 19, 2020. <https://www.wsj.com/articles/trump-talks-juneteenth-john-bolton-economy-in-wsj-interview-11592493771>.
- Boykin, Kaitlyn, Mitch Brown, Alicia L. Macchione, Kelsey M. Drea, and Donald F. Sacco. 2021. "Noncompliance with Masking as a Coalitional Signal to US Conservatives in a Pandemic." *Evolutionary Psychological Science* 7 (3): 232–38. <https://doi.org/10.1007/s40806-021-00277-x>.
- Brulle, Robert J., Jason Carmichael, and J. Craig Jenkins. 2012. "Shifting Public Opinion on Climate Change: An Empirical Assessment of Factors Influencing Concern over Climate Change in the U.S., 2002–2010." *Climatic Change* 114 (2): 169–88. <https://doi.org/10.1007/s10584-012-0403-y>.
- Bunch, Will. 2020. "GOP, Fox News Have Waged War on Science. With Coronavirus, Will Their Aging Fans Pay the Price? | Will Bunch." 2020. <https://www.inquirer.com/health/coronavirus/coronavirus-republicans-denial-fox-news-trump-war-on-science-20200315.html>.
- Burge, Ryan P. 2020a. "Where Do White Evangelicals Get Their Coronavirus News? The White House." *News & Reporting*. May 14, 2020. <https://www.christianitytoday.com/news/2020/may/evangelicals-trump-media-public-health-response-coronavirus.html>.
- . 2020b. "White Evangelicals' Coronavirus Concerns Are Fading Faster." *News & Reporting*. June 23, 2020. <https://www.christianitytoday.com/news/2020/june/evangelicals-covid-19-less-worry-social-distancing-masks.html>.
- CDC. 2020. "Coronavirus Disease 2019 (COVID-19)." Centers for Disease Control and Prevention. February 11, 2020. <https://www.cdc.gov/coronavirus/2019-ncov/global-covid-19/global-urban-areas.html>.
- . 2022. "COVID Data Tracker." Centers for Disease Control and Prevention. 2022. <https://covid.cdc.gov/covid-data-tracker>.
- Cherry, Jon. 2021. "Covid Mask and Vaccination Mandates Aren't Christian Persecution." August 31, 2021. <https://www.nbcnews.com/think/opinion/covid-mask-vaccination-mandates-aren-t-christian-persecution-ncna1278067>.

- CNN Newsroom. 2020. "President Trump: People Are Tired Of Hearing Fauci And All These Idiots; President Trump Attacks Dr. Fauci On Call With Campaign Staff; Dr. Anthony Fauci: Certain Parts Of Society 'Disturbingly Anti- Science'; President Trump Tries To Hold Onto Key Rust Belt Swing States; Utah Reports Rise In Test Positivity, Hospitalizations. Aired 12-12:30p ET. - Document - Gale General OneFile." 2020. https://go-gale-com.offcampus.lib.washington.edu/ps/i.do?p=ITOF&u=wash_main&id=GALE%7CA638883027&v=2.1&it=r.
- Crist, Carolyn. 2022. "CDC: Most Americans Have COVID-19 Antibodies." March 30, 2022. <https://www.webmd.com/lung/news/20220330/cdc-most-americans-have-covid-antibodies>.
- Drummond, Caitlin, and Baruch Fischhoff. 2017. "Individuals with Greater Science Literacy and Education Have More Polarized Beliefs on Controversial Science Topics." *Proceedings of the National Academy of Sciences of the United States of America* 114 (36): 9587–92.
- Evans, John H. 2013. "The Growing Social and Moral Conflict Between Conservative Protestantism and Science." *Journal for the Scientific Study of Religion* 52 (2): 368–85. <https://doi.org/10.1111/jssr.12022>.
- Evans, John H., and Michael S. Evans. 2008. "Religion and Science: Beyond the Epistemological Conflict Narrative." *Annual Review of Sociology* 34: 87–105.
- Flaskerud, Jacquelyn H. 2020. "Masks, Politics, Culture and Health." *Issues in Mental Health Nursing* 41 (9): 846–49. <https://doi.org/10.1080/01612840.2020.1779883>.
- Flood, Colleen M., Vanessa MacDonnell, Bryan Thomas, and Kumanan Wilson. 2020. "Reconciling Civil Liberties and Public Health in the Response to COVID-19." *FACETS* 5 (1): 887–98. <https://doi.org/10.1139/facets-2020-0070>.
- Fowler, Hayley. 2020. "Here's How Religious Americans View the Coronavirus Pandemic | Charlotte Observer." May 16, 2020. <https://amp.charlotteobserver.com/news/coronavirus/article242784396.html>.
- Frickel, Scott, and Christopher M. Rea. 2020. "Drought, Hurricane, or Wildfire? Assessing the Trump Administration's Anti-Science Disaster." *Engaging Science, Technology, and Society* 6 (January): 66–75. <https://doi.org/10.17351/ests2020.297>.
- Gauchat, Gordon William. 2008. "A Test of Three Theories of Anti-Science Attitudes." *Sociological Focus* 41 (4): 337–57.
- Gonzalez, Kelsey E., Rina James, Eric T. Bjorklund, and Terrence D. Hill. 2021. "Conservatism and Infrequent Mask Usage: A Study of US Counties during the Novel Coronavirus (COVID-19) Pandemic." *Social Science Quarterly* 102 (5): 2368–82. <https://doi.org/10.1111/ssqu.13025>.
- Hamilton, Lawrence C. 2008. "Who Cares about Polar Regions? Results from a Survey of U.S. Public Opinion." *Arctic, Antarctic, and Alpine Research* 40 (4): 671–78. [https://doi.org/10.1657/1523-0430\(07-105\)\[HAMILTON\]2.0.CO;2](https://doi.org/10.1657/1523-0430(07-105)[HAMILTON]2.0.CO;2).
- . 2011. "Education, Politics and Opinions about Climate Change Evidence for Interaction Effects." *Climatic Change* 104 (2): 231–42. <https://doi.org/10.1007/s10584-010-9957-8>.
- Hamilton, Lawrence C., Chris R. Colocousis, and Cynthia M. Duncan. 2010. "Place Effects on Environmental Views." *Rural Sociology* 75 (2): 326–47. <https://doi.org/10.1111/j.1549-0831.2010.00013.x>.
- Hamilton, Lawrence C., Joel Hartter, and Kei Saito. 2015. "Trust in Scientists on Climate Change and Vaccines." *SAGE Open* 5 (3): 2158244015602752. <https://doi.org/10.1177/2158244015602752>.

- Hamilton, Lawrence C., and Kei Saito. 2015. "A Four-Party View of US Environmental Concern." *Environmental Politics* 24 (2): 212–27. <https://doi.org/10.1080/09644016.2014.976485>.
- Hamilton, Lawrence, and Thomas Safford. 2020. "Ideology Affects Trust in Science Agencies During a Pandemic." The Carsey School of Public Policy at the Scholars' Repository, March. <https://doi.org/10.34051/p/2020.389>https://dx.doi.org/10.34051/p/2020.389.
- Hayes, Andrew F. 2013. *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*. New York: Guilford Publications.
- Hill, Terrence D., Kelsey Gonzalez, and Amy M. Burdette. 2020. "The Blood of Christ Compels Them: State Religiosity and State Population Mobility During the Coronavirus (COVID-19) Pandemic." *Journal of Religion and Health* 59 (5): 2229–42. <https://doi.org/10.1007/s10943-020-01058-9>.
- Howard, Matt C. 2022. "Are Face Masks a Partisan Issue during the COVID-19 Pandemic? Differentiating Political Ideology and Political Party Affiliation." *International Journal of Psychology* 57 (1): 153–60. <https://doi.org/10.1002/ijop.12809>.
- Hudson, Robert B. 2018. "Draining the Swamp While Making America Great Again: Senior Dissonance in the Age of Trump." *Journal of Aging & Social Policy* 30 (3–4): 357–71. <https://doi.org/10.1080/08959420.2018.1462676>.
- Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010. "A General Approach to Causal Mediation Analysis." *Psychological Methods* 15 (4): 309–34. <https://doi.org/10.1037/a0020761>.
- Kahan, Dan M., Ellen Peters, Maggie Wittlin, Paul Slovic, Lisa Larrimore Ouellette, Donald Braman, and Gregory Mandel. 2012. "The Polarizing Impact of Science Literacy and Numeracy on Perceived Climate Change Risks." *Nature Climate Change* 2 (10): 732–35. <https://doi.org/10.1038/nclimate1547>.
- Kahane, Leo H. 2021. "Politicizing the Mask: Political, Economic and Demographic Factors Affecting Mask Wearing Behavior in the USA." *Eastern Economic Journal*, January, 1–21. <https://doi.org/10.1057/s41302-020-00186-0>.
- Kahn, Rob. 2022. "COVID Masks as Semiotic Expressions of Hate." *International Journal for the Semiotics of Law - Revue Internationale de Sémiotique Juridique*, March. <https://doi.org/10.1007/s11196-022-09885-7>.
- Klayman, Joshua, and Young-won Ha. 1987. "Confirmation, Disconfirmation, and Information in Hypothesis Testing." *Psychological Review* 94 (2): 211–28. <https://doi.org/10.1037/0033-295X.94.2.211>.
- Kochhar, Rakesh, and Stella Sechopoulos. 2022. "COVID-19 Pandemic Pinches Finances of America's Lower- and Middle-Income Families." Pew Research Center's Social & Demographic Trends Project (blog). April 20, 2022. <https://www.pewresearch.org/social-trends/2022/04/20/covid-19-pandemic-pinches-finance-of-americas-lower-and-middle-income-families/>.
- Kunda, Ziva. 1990. "The Case for Motivated Reasoning." *Psychological Bulletin* 108 (3): 480–98. <https://doi.org/10.1037/0033-2909.108.3.480>.
- Lauter, David. 2020. "Essential Politics: Can't Mask Partisanship." *Los Angeles Times*. May 15, 2020. <https://www.latimes.com/politics/newsletter/2020-05-15/mask-partisanship-coronavirus-trump-win-essential-politics>.

- Light, Nicholas, Philip M. Fernbach, Nathaniel Rabb, Mugur V. Geana, and Steven A. Sloman. 2022. "Knowledge Overconfidence Is Associated with Anti-Consensus Views on Controversial Scientific Issues." *Science Advances* 8 (29): eabo0038. <https://doi.org/10.1126/sciadv.abo0038>.
- Malka, Ariel, Jon A. Krosnick, and Gary Langer. 2009. "The Association of Knowledge with Concern About Global Warming: Trusted Information Sources Shape Public Thinking." *Risk Analysis* 29 (5): 633–47. <https://doi.org/10.1111/j.1539-6924.2009.01220.x>.
- McCright, Aaron M. 2011. "Political Orientation Moderates Americans' Beliefs and Concern about Climate Change." *Climatic Change* 104 (2): 243–53. <https://doi.org/10.1007/s10584-010-9946-y>.
- Milosh, Maria, Marcus Painter, Konstantin Sonin, David Van Dijke, and Austin L. Wright. 2021. "Unmasking Partisanship: Polarization Undermines Public Response to Collective Risk." *Journal of Public Economics* 204 (December): 104538. <https://doi.org/10.1016/j.jpubeco.2021.104538>.
- Minton, Elizabeth A., Cindy X. Wang, and Carissa M. Anthony. 2022. "Prosocial Responses to Global Crises: Key Influences of Religiosity and Perceived Control." *Journal of Consumer Affairs* 56 (2): 491–511. <https://doi.org/10.1111/joca.12432>.
- Moore, Don A., and Paul J. Healy. 2008. "The Trouble with Overconfidence." *Psychological Review* 115 (2): 502–17. <https://doi.org/10.1037/0033-295X.115.2.502>.
- Peiser, Jaclyn. 2021. "Evangelical Pastor Demands Churchgoers Ditch Their Masks: 'Don't Believe This Delta Variant Nonsense.'" *Washington Post*, July 27, 2021. <https://www.washingtonpost.com/nation/2021/07/27/tennessee-pastor-greg-locke-masks/>.
- Perry, Samuel L., Andrew L. Whitehead, and Joshua B. Grubbs. 2020. "Culture Wars and COVID-19 Conduct: Christian Nationalism, Religiosity, and Americans' Behavior During the Coronavirus Pandemic." *Journal for the Scientific Study of Religion* 59 (3): 405–16. <https://doi.org/10.1111/jssr.12677>.
- Politico. 2016. "Full Text: Trump Values Voter Summit Remarks." *POLITICO*. September 9, 2016. <https://www.politico.com/story/2016/09/full-text-trump-values-voter-summit-remarks-227977>.
- Rubin, Jennifer. 2022. "Opinion | The GOP Is No Longer a Party. It's a Movement to Impose White Christian Nationalism. - The Washington Post." April 27, 2022. <https://www.washingtonpost.com/opinions/2022/04/27/gop-no-longer-a-party-movement-impose-christian-nationalism/>.
- Sami, Shahabir. 2020. "Trump Supporters Are the Majority of Those Continuing to Defy Wearing Masks." *The Science Survey* (blog). November 12, 2020. <https://thesciencesurvey.com/editorial/2020/11/12/trump-supporters-are-the-majority-of-those-continuing-to-defy-wearing-masks/>.
- Schlor, Elana, and Hannah Fingerhut. 2020. "Poll: 63% of Religious Americans Believe COVID-19 Pandemic Is Message from God for Humanity to Change." *Text.Article*. Associated Press. FOX 2 Detroit. May 17, 2020. <https://www.fox2detroit.com/news/poll-63-of-religious-americans-believe-covid-19-pandemic-is-message-from-god-for-humanity-to-change>.
- Schnabel, Landon, and Scott Schieman. 2022. "Religion Protected Mental Health but Constrained Crisis Response During Crucial Early Days of the COVID-19 Pandemic." *Journal for the Scientific Study of Religion* 61 (2): 530–43.

- <https://doi.org/10.1111/jssr.12720>.
- Scott, Dylan. 2020. "Why the US Has Failed in Its Coronavirus Response, in One Trump Interview." Vox. June 18, 2020.
<https://www.vox.com/2020/6/18/21295826/coronavirus-us-update-trump-wsj-interview-masks-tests>.
- Tingley, Dustin, Teppei Yamamoto, Kentaro Hirose, Luke Keele, and Kosuke Imai. 2014. "Mediation : R Package for Causal Mediation Analysis." Journal of Statistical Software 59 (5). <https://doi.org/10.18637/jss.v059.i05>.
- Tversky, Amos, and Craig R. Fox. 1995. "Weighing Risk and Uncertainty." Psychological Review 102 (2): 269–83. <https://doi.org/10.1037/0033-295X.102.2.269>.
- Van Bavel, Jay. 2020. "Democrats and Republicans Are Divided over Coronavirus Risks. That Could Kill People. - The Washington Post." 2020.
<https://www.washingtonpost.com/outlook/2020/03/23/coronavirus-polarization-political-exaggeration/>.
- Vasquez Reyes, Maritza. 2020. "The Disproportional Impact of COVID-19 on African Americans." Health and Human Rights 22 (2): 299–307.
- Yoon, Wonjeong, Young Sun Ro, and Sung-il Cho. 2019. "A Mediation Analysis of the Effect of Practical Training on the Relationship between Demographic Factors, and Bystanders' Self-Efficacy in CPR Performance | PLOS ONE." <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0215432>.

REACH Project: Descriptive Data Summary

The data received from REACH constituted of 913 tent encampment site locations and street outreach locations (i.e., encounters not specifically related to an encampment location) within the Seattle city limits, representing a time period of 02/01/2016 to 09/06/2022. Of these sites, 859 tent encampment locations and 48 street outreach locations were selected as

being suitable for attempted geocoding, and therefore geocoded to obtain specific coordinates using the NAD83(National Spatial Reference System 2011), Washington North (EPSG: 6596) projection specification appropriate for the King County, Washington area. When locations were unable to be identified with geographic specificity, the nearest cross-street (intersection) was identified and geocoded to obtain a relative proximity to the site location. When sites were identified as being on a specific city block, the two intersections which bounded the block were identified and geocoded. To avoid issues with spatial analyses requiring unique geolocations for sites, sites with two geocoded locations were simplified by random selection of one of these geocodes. The number of reliably identified locations for this analysis was thus 742; this accounts for that required two geocodes (i.e. removing duplicates) or were geographically unidentifiable. Data were analyzed globally (i.e., within the bounds of the entire geographic extent of the Seattle city limits) and locally (i.e. within the extent of Seattle city council districts as specified by shapefiles obtained from the City of Seattle GIS portal) (Figure 1). The city of Seattle is divided into 7 distinct city council districts. The boundaries of these may have changed over time, but for simplicity this analysis discusses these districts in terms of the most recent boundaries' shapefiles. The distance to potential service sites were also calculated. Services included churches, medical clinics, DSHS offices, food banks, grocery stores, hospitals, libraries, needle exchange locations, public parks, fast food restaurants with potentially easily accessible hygiene facilities, bus stops, identifiable municipally managed trees (e.g., along pathways or sidewalks). Future analyses consisting of environmental exposure and potential health stressors, and also accounting for the temporal nature of the data, are planned. Some descriptive statistics of distances to services by year and council district are visualized below.



Figure 1: Seattle City Council Districts as of Autumn 2022

Global descriptive statistics

The maximum number individuals observed (coded in the data as “present”) at encampment sites ranged from zero (i.e. sites that were observed but had no individual present at the time of observation) to 225 individuals, with an average of about 17 individuals. The distance from each encampment site to known service locations were calculated, and the minimum, mean, and maximum distances (meters) were, respectively, 0m (most often accounted for by proximity to trees), 1034m, and 11,684m. Further local distances to specific services are described in Table 1, and visualized in Figure 1.

Figure 1: Distance (meters) to specific services by Seattle City Council District

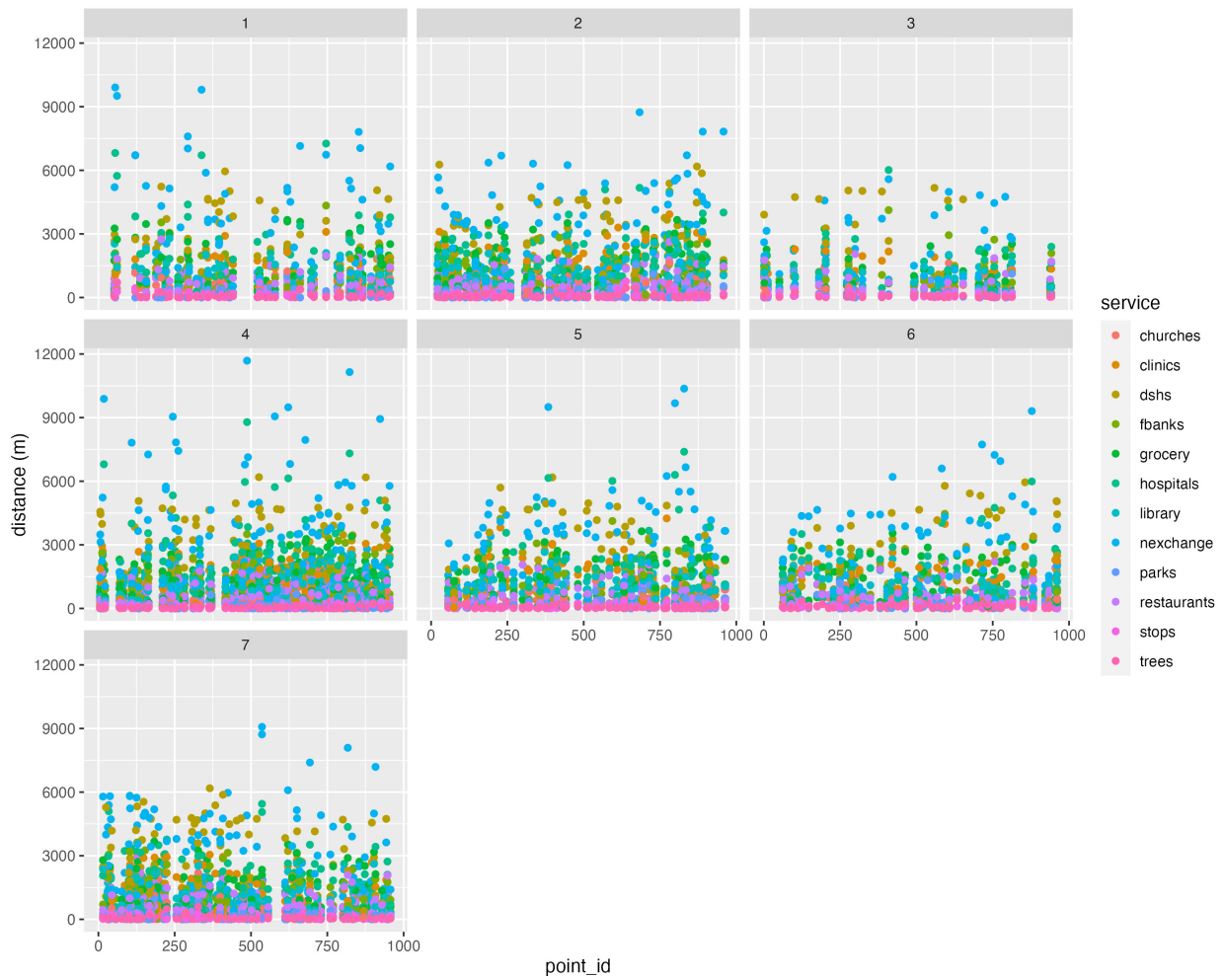


Table 1: Global distances to services

Service	Maximum distance (meters)	Minimum distance (meters)	Mean distance (meters)
churches	2188.92	1.28	501.34
medical clinics	4245.65	3.28	1121.64
DSHS Offices	6268.19	43.30	1985.71
food banks	5007.98	27.38	1005.55
grocery stores	3672.18	60.01	1230.21
hospitals	8788.75	71.28	1664.79
library	2855.40	36.71	1031.44

Service	Maximum distance (meters)	Minimum distance (meters)	Mean distance (meters)
needle exchange locations	11684.27	50.93	2698.37
public parks	1547.35	0.00	253.87
fast food restaurants	2889.71	30.10	728.08
bus stops	764.23	0.88	95.30
municipally identified trees	914.23	1.06	85.84

Descriptive statistics by Seattle city council district

The number of unique encampments in each city council district were: 72 in District 1, 119 in District 2, 46 in District 3, 183 in District 4, 97 in District 5, 76 in District 6 and 149 in District 7. The minimum, mean and maximum distances to all services by Seattle city council district are reported in Table 2. The average distance in meters to all services from encampment sites ranged from 894.41 to 1080.28. Distances to various services by council district and year are visualized in Figure 2.

Table 2: Distances to all services of interest by Seattle city council district

City Council District	Maximum distance (meters)	Minimum distance (meters)	Mean distance (meters)
1	9908.54	0.00	1113.01
2	8737.78	0.00	1084.10
3	6014.49	0.00	894.41
4	11684.27	0.00	1034.99
5	10367.31	0.00	1080.28
6	9308.67	0.00	1046.80
7	9077.03	0.00	958.13

The minimum mean and maximum distances to specific services by Seattle city council district are reported in Table 3.

Figure 2: Distance (meters) from encampment locations to services by city council district, 2016-2022



Table 3: Distances to services of interest by Seattle city council district

City Council District	service	Maximum distance (meters)	Minimum distance (meters)	Mean distance (meters)
1	churches	1238.80	40.50	476.86
1	medical medical medical clinics	3096.30	18.98	1154.51

City Council District	service	Maximum distance (meters)	Minimum distance (meters)	Mean distance (meters)
1	DSHS Offices	5950.36	88.86	2034.40
1	food banks	4336.13	71.26	924.60
1	grocery	3662.90	66.07	1335.32
1	hospitals	7262.14	71.28	1916.06
1	library	2787.23	130.71	1103.77
1	needle exchange locations	9908.54	428.04	3220.94
1	public parks	1010.18	0.00	222.40
1	fast food fast food fast food restaurants	2745.20	30.10	754.01
1	bus stops	444.64	6.95	105.30
1	municipally identified trees	728.86	2.35	96.43
2	churches	1685.96	8.71	549.16
2	medical medical medical clinics	3925.19	40.99	1267.83
2	DSHS Offices	6268.19	95.12	2113.50
2	food banks	5007.98	40.33	988.60
2	grocery	3661.27	100.84	1391.04
2	hospitals	5176.01	247.15	1625.03
2	library	2746.63	74.12	1094.17
2	needle exchange locations	8737.78	159.23	2798.50
2	public parks	1547.35	0.00	249.71
2	fast food fast food fast food restaurants	2624.58	49.83	735.95
2	bus stops	764.23	4.24	105.88
2	municipally identified trees	722.80	2.71	89.81
3	churches	1624.37	15.28	474.72

City Council District	service	Maximum distance (meters)	Minimum distance (meters)	Mean distance (meters)
3	medical medical medical clinics	2272.87	67.53	972.16
3	DSHS Offices	5172.22	115.74	1886.51
3	food banks	4123.70	79.88	978.09
3	grocery	3236.42	127.72	1087.05
3	hospitals	6014.49	185.62	1355.74
3	library	2378.95	58.54	906.30
3	needle exchange locations	5580.40	163.07	2069.35
3	public parks	1201.27	0.00	251.80
3	fast food fast food fast food restaurants	2050.46	58.25	592.93
3	bus stops	361.27	5.59	90.76
3	municipally identified trees	337.47	7.28	67.52
4	churches	2159.12	11.78	506.96
4	medical medical medical clinics	3062.93	63.11	1073.26
4	DSHS Offices	6189.69	78.57	1913.38
4	food banks	3695.50	29.91	1050.97
4	grocery	3402.28	79.25	1160.69
4	hospitals	8788.75	154.60	1853.78
4	library	2573.29	48.16	993.17
4	needle exchange locations	11684.27	225.68	2654.80
4	public parks	1359.14	0.00	258.70
4	fast food fast food fast food restaurants	2311.17	37.05	774.16
4	bus stops	455.20	2.75	85.38
4	municipally identified trees	914.23	2.54	94.69

City Council District	service	Maximum distance (meters)	Minimum distance (meters)	Mean distance (meters)
5	churches	1832.91	23.62	539.90
5	medical medical medical clinics	4245.65	45.84	1127.65
5	DSHS Offices	6182.15	98.67	2099.58
5	food banks	4230.28	36.99	1088.17
5	grocery	3640.20	60.01	1279.51
5	hospitals	7390.61	76.91	1698.91
5	library	2710.02	53.56	1066.99
5	needle exchange locations	10367.31	239.56	2923.68
5	public parks	1219.63	0.00	294.98
5	fast food fast food fast food restaurants	2059.62	34.73	676.51
5	bus stops	405.74	0.88	95.70
5	municipally identified trees	614.81	4.42	71.78
6	churches	2011.07	57.48	550.83
6	medical medical medical clinics	3984.88	51.37	1201.48
6	DSHS Offices	5949.25	43.30	2151.24
6	food banks	4464.35	102.11	919.60
6	grocery	3547.16	84.57	1249.67
6	hospitals	5989.49	165.84	1509.08
6	library	2855.40	36.71	1038.55
6	needle exchange locations	9308.67	50.93	2719.16
6	public parks	1535.89	0.00	292.54
6	fast food fast food fast food restaurants	2209.18	57.13	717.04
6	bus stops	528.09	4.47	108.89

City Council District	service	Maximum distance (meters)	Minimum distance (meters)	Mean distance (meters)
6	municipally identified trees	425.74	5.35	103.52
7	churches	2188.92	1.28	427.12
7	medical medical medical clinics	3080.57	3.28	1049.87
7	DSHS Offices	6180.82	60.07	1821.20
7	food banks	3628.33	27.38	1001.63
7	grocery	3672.18	94.27	1137.37
7	hospitals	5828.61	113.18	1493.90
7	library	2818.54	56.31	1003.70
7	needle exchange locations	9077.03	170.89	2447.55
7	public parks	1290.01	0.00	221.01
7	fast food fast food fast food restaurants	2889.71	92.47	733.26
7	bus stops	447.66	7.00	88.29
7	municipally identified trees	589.08	1.06	72.61

Density of Sites

In terms of global density (Figure 3), the majority of sites are centralized around the downtown Seattle core, especially in places with the greatest population density and/or industrial or semi-industrial areas and parks, including South Lake Union, the International District and SODO, Ballard and the University District. However, when the extent of by-council district density of sites is considered (Figure 4), it is evident that many locations are associated with high car traffic areas (e.g., the I-5 corridor, the SW Spokane Street Bridge to West Seattle, the I-90 interchange, or other major thoroughfares), beaches, parks and trails (e.g. Golden Gardens, The Burke Gilman Trail, Alki Beach, etc.), and other spaces that may have transient foot traffic by the general public, or else places with lower residential density.

Figure 3: density based on city-wide point pattern process analysis

tent encampment density

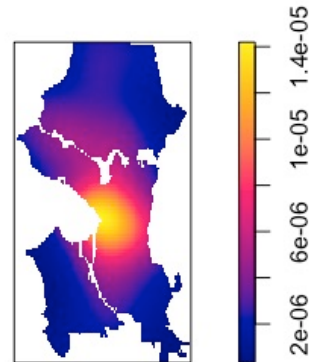


Figure 4: Density plots of encampment locations by City Council District (preliminary findings)

