REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
18-12-2021	Final Report	31-Aug-2015 - 30-Mar-2021

4. TITLE AND SUBTITLE
Final Report: A Spectral Framework for Graph Sampling

5a. CONTRACT NUMBER
W911NF-15-1-0423

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 Wisconsin - Madison
Suite 6401
21 N Park Street
Madison, WI 53715 -1218

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)
65673-MA.16

12. DISTRIBUTION AVAILIBILITY 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 contrued 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	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Karl Rohe
UU	UU	UU	UU		19b. TELEPHONE NUMBER: 608-263-8531

RPPR Final Report
as of 18-Dec-2021

Agency Code: 21XD

Proposal Number: 65673MA **Agreement Number: W911NF-15-1-0423**
INVESTIGATOR(S):

Name: Karl Rohe
Email: karlrohe@stat.wisc.edu
Phone Number: 6082638531
Principal: Y

Organization: **University of Wisconsin - Madison**
Address: Suite 6401, Madison, WI 537151218
Country: USA
DUNS Number: 161202122 EIN: 396006492
Report Date: 30-Jun-2021 Date Received: 18-Dec-2021
Final Report for Period Beginning 31-Aug-2015 and Ending 30-Mar-2021
Title: A Spectral Framework for Graph Sampling
Begin Performance Period: 31-Aug-2015 **End Performance Period:** 30-Mar-2021
Report Term: 0-Other
Submitted By: Karl Rohe Email: karlrohe@stat.wisc.edu
Phone: (608) 263-8531
Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 30 **STEM Participants:** 35

Major Goals: Aim 1: Build a statistical framework for globally-adaptive Markov chain sampling

Sub-aim 1.1:
Using techniques from Markov chains, characterize the “effective sample size” of a Markov chain sample and relate this to the spectral characteristics of the graph. Use these results to estimate standard errors (for confidence intervals and hypothesis testing). Theoretically validate the consistency of these standard error estimates.

Sub-aim1.2:
Develop globally-adaptive Markov chains, prove that they are globally-adaptive, and extend the results from the previous sub-aim to these novel chains.

Sub-aim1.3:
Extend the results from the previous two sub-aims to more general tasks of statistical inference on node/edge contextualizing measures (e.g. two sample tests, linear regression, princi- pal components analysis) and their relationships to network topology (e.g. degree and clustering coefficient).

Aim 2: Characterize the spectral sensitivity to sampling-induced edge dependence for two broad classes of sampling mechanisms.

Sub-aim 2.1:
Characterize the types of link-tracing sampling that produce adjacency matrices that converge in spectral norm. Then, characterize (i) how the limiting object is useful for para- metric tasks and (ii) how sampling-induced dependence manifests in downstream analyses. For example, suppose that the population graph comes from a Stochastic Blockmodel. We will study the performance of the standard standard spectral clustering estimator. More broadly, the techni- cal tools developed in this section will be used to study estimators for networks that result from a dynamic sampling process.

Sub-aim 2.2:
Characterize the types of motifs in motif sampling that create spectral sensitivities. Study the performance of spectral clustering under the motif Stochastic Blockmodel to understand how these spectral sensitivities manifests in down stream analysis.

Aim 3: Develop a class of locally-adaptive sampling mechanisms to test for local structure in massive networks.

Sub-aim 3.1: Develop and study locally-adaptive chains.
Sub-aim 3.2: Develop test statistics and rejection regions.
Sub-aim 3.3: Study the previous steps under a “degree corrected” null model.
Sub-aim 3.4: Detect dense subgraphs with an enriched proportion of high risk nodes/edges.

ADD ON

Add-on Aim 1: Build a data science tool that identifies trustworthy social media accounts and disseminate this tool to information analysts, journalists, and media watchers.

Add-on Aim 2: Study the section of inclusion terms as a problem of statistical estimation under Latent Dirichlet Allocation. Develop fast estimators based upon the word co-occurance graph. Provide R software.

Accomplishments: Aim 1.1: This is accomplished in these papers:
- Rohe 2019: "A critical threshold for design effects in network sampling"
Annals of Statistics

- Li and Rohe 2017: "Central limit theorems for network driven sampling"
Electronic Journal of Statistics

Aim 1.2: This is accomplished in this paper:
- Khabbazian et al 2017: "Novel sampling design for respondent-driven sampling"
Electronic Journal of Statistics

Aim 1.3: Our first efforts show that there is much more work to be done than initially proposed. In particular, previous estimators are not \sqrt{n}-consistent. Their standard errors converge much more slowly. We have started to address Aim 1.3 in this paper:
- Roch and Rohe 2018: "Generalized least squares can overcome the critical threshold in respondent-driven sampling"
Proceedings of the National Academy of Sciences

- Zhang, Rohe, Roch 2018: "Reducing Seed Bias in Respondent-Driven Sampling by Estimating Block Transition Probabilities"
arXiv:1812.01188 (in revision)

- Yan, Hanlon, Roch, Rohe 2020: "Asymptotic Seed Bias in Respondent-driven Sampling"
Electron. Journal of Statistics

Aim 2:
This aim led to unexpected results. The first provides a surprisingly simple understanding of regularized spectral clustering and was at NeurIPS. The second provides a simple and fast way to sample from a broad class of random graph models. It was published at JMLR.
- Zhang, Rohe 2018: "Understanding regularized spectral clustering via graph conductance"
Advances in Neural Information Processing Systems (NeurIPS)

-K Rohe, J Tao, X Han, N Binkiewicz 2018: "A note on quickly sampling a sparse matrix with low rank expectation"
The Journal of Machine Learning Research (JMLR)

Aim 3: This aim has been accomplished in two very different ways. The first is theoretical/methodological and is represented in a paper that has been resubmitted with minor revisions to JRSS-B.

- Chen, Zhang, Rohe 2019: "Targeted sampling from massive Blockmodel graphs with personalized PageRank" JRSS-B

To ensure that this aim did not become a simply theoretical exercise, it was paired with empirical research joint with colleagues in the School of Journalism.

- Zhang, Poux-Berthe, Wells, Koc-Michalska, Rohe 2018: "Discovering political topics in Facebook discussion threads with graph contextualization"
Annals of Applied Statistics

This resulted in the initial work on murmuration.wisc.edu, which was included in a recent ARO grant.

Add-on Aim 1: This aim has been accomplished and dissemination is ongoing. Our website https://murmuration.wisc.edu describes daily features of the elite layer of twitter discourse on current events.

The manuscript that describes this website:

“Social Media Public Opinion as Flocks in a Murmuration: Conceptualizing and Measuring Opinion Expression on Social Media”

Is accepted at the highest impact factor journal in Communication, called the Journal of Computer Mediated Communication.

This applied work builds on three pieces of “fundamental science” that were developed with ARO funding. First, the article "Targeted sampling from massive Blockmodel graphs with personalized PageRank" published in the highest impact factor journal in Statistics, JRSS-B. Second, the twittercache software developed in our lab (https://github.com/alexpghayes/twittercache) which enables sampling billions of edges in the Twitter graph. Third, the PI’s work has enabled a theoretical justification for the use of Varimax.

Data is still being collected and analyzed by the PIs colleagues (former students) that are now tenure track faculty in the departments of Journalism at University Buffalo and UT Austin.

Add-on Aim 2:

This aim has been accomplished by PhD student Fan Chen in his dissertation. This work shows how to find additional “inclusion keywords” for targeted document assembling. We study a network-based method to prioritize words. In the network of words, two words are connected if and only if they co- occur in at least one document. On the network of words, we propose to rank words using personalized PageRank (PPR). We prove that WordPPR is accurate and demonstrate it is computationally efficient.

Training Opportunities: A description of Opportunities for training during the reporting period.

This award has provided a bridge from assistant, to associate, to full professor for PI Rohe. He relied upon this award to support various types of research expenses that were fundamental to his success. He remains very thankful for the continued support of the Army Research Office.

Since 2015, the Rohe lab has graduated 9 PhD students, both domestic and international, including three women.

Since 2015, the Rohe lab has included 23 undergraduates in research (ranging from published research, to data collection, to simulations). These students have gone on to Statistics PhD programs at Berkeley, Princeton, U Chicago, NYU, U Michigan, Carnegie Mellon, Yale, Harvard, and Penn State.

RPPR Final Report
as of 18-Dec-2021

Results Dissemination:
For members of our community, the papers and the technical results have all been posted on arXiv. Any code is published on github. https://github.com/karlrohe

For the broader Statistics audience, we have papers published in Annals of Statistics, Electronic Journal of Statistics, Proceedings of the National Academy of Sciences, Neural Information Processing Systems (top Machine learning conference), Journal of Machine Learning Research, Journal of Computational and Graphical Statistics, Annals of Applied Statistics, Statistica Sinica, and the Journal of the Royal Statistical Society (methodology series)

Honors and Awards: The PI was promoted to Full Professor in Spring 2021

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: PD/PI
Participant: Karl Rohe
Person Months Worked: 15.00 **Funding Support:**
Project Contribution:
National Academy Member: N

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

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

Participant Type: Graduate Student (research assistant)
Participant: Alex Hayes
Person Months Worked: 9.00 **Funding Support:**
Project Contribution:
National Academy Member: N

ARTICLES:

RPPR Final Report
as of 18-Dec-2021

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: arxiv (submitted and under review)
Publication Identifier Type: Other Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: 8/30/18 12:00AM Date Published: 8/11/16 9:22PM
Publication Location:
Article Title: Novel Sampling Design for Respondent-driven Sampling
Authors: Mohammad Khabbazian, Bret Hanlon, Zoe Russek, Karl Rohe
Keywords: Hard-to-reach population; Social network; Trees; Markov chains; Spectral representation; Anti-cluster RDS
Abstract: Respondent-driven sampling (RDS) is a type of chain referral sampling popular for sampling hidden and/or marginalized populations. As such, even under the ideal sampling assumptions, the performance of RDS is restricted by the underlying social network: if the network is divided into weakly connected communities, then RDS is likely to oversample one of these communities. In order to diminish the “referral bottle- necks” between communities, we propose anti-cluster RDS (AC-RDS), an adjustment to the standard RDS implementation. Using a standard model in the RDS literature, namely, a Markov process on the social network that is indexed by a tree, we construct and study the Markov transition matrix for AC-RDS. We show that if the underlying network is generated from the Stochastic Blockmodel with equal block size, then the transition matrix for AC-RDS has a smaller spectral gap and consequently faster mixing properties than the standard random walk model for RDS. In addition, we show ...
Distribution Statement: 3-Distribution authorized to U.S. Government Agencies and their contractors
Acknowledged Federal Support: **Y**

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: PNAS
Publication Identifier Type: Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: 8/27/17 12:00AM Date Published: 8/12/16 12:05AM
Publication Location:
Article Title: Co-clustering directed graphs to discover asymmetries and directional communities
Authors: Karl Rohe, Tai Qin, Bin Yu
Keywords: Spectral Clustering | SVD | Stochastic Blockmodel
Abstract: In directed graphs, relationships are asymmetric and these asymmetries contain essential structural information about the graph. Directed relationships lead to a new type of clustering that is not feasible in undirected graphs. We propose a spectral co-clustering algorithm called DI-SIM for asymmetry discovery and directional clustering. A new Stochastic co-Blockmodel is introduced to show favorable properties of DI-SIM. To account for the sparse and highly heterogeneous nature of directed networks, DI-SIM uses the regularized graph Laplacian and projects the rows of the eigenvector matrix onto the sphere. A node-wise ASYMMETRY SCORE and DI-SIM are employed to analyze the clustering asymmetries in the networks of Enron emails, political blogs, and the C. elegans chemical connectome. In each example, a subset of nodes have clustering asymmetries; these nodes send edges to one cluster, but receive edges from another cluster. Such nodes yield insightful information ...
Distribution Statement: 3-Distribution authorized to U.S. Government Agencies and their contractors
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 18-Dec-2021

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Journal of Computational and Graphical Statistics
Publication Identifier Type: Other Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: 8/30/18 12:00AM Date Published: 8/12/16 12:12AM
Publication Location:
Article Title: Intelligent Initialization and Adaptive Thresholding for Iterative Matrix Completion; Some Statistical and Algorithmic Theory for Adaptive-Impute
Authors: Juhee Cho, Donggyu Kim, Karl Rohe
Keywords: softImpute, generalized-softImpute, non-convex optimization, thresholded sin- gular value decomposition
Abstract: Over the past decade, various matrix completion algorithms have been developed. Thresholded singular value decomposition (SVD) is a popular technique in implementing many of them. A sizable number of studies have shown its theoretical and empirical excellence, but choosing the right threshold level still remains as a key empirical difficulty. This paper proposes a novel matrix completion algorithm which iterates thresholded SVD with theoretically-justified and data-dependent values of thresholding parameters. The estimate of the proposed algorithm enjoys the minimax error rate and shows outstanding empirical performances. The thresholding scheme that we use can be viewed as a solution to a non-convex optimization problem, understanding of whose theoretical convergence guarantee is known to be limited. We investigate this problem by introducing a simpler algorithm, generalized-softImpute, analyzing its convergence behavior, and connecting it to the proposed algorithm.
Distribution Statement: 3-Distribution authorized to U.S. Government Agencies and their contractors
Acknowledged Federal Support: **Y**

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Statistica Sinica
Publication Identifier Type: Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: 8/27/17 12:00AM Date Published: 8/12/16 12:16AM
Publication Location:
Article Title: Asymptotic Theory for Estimating the Singular Vectors and Values of a Partially-observed Low Rank Matrix with Noise
Authors: Juhee Cho, Donggyu Kim, Karl Rohe
Keywords: Matrix completion, low rank matrices, singular value de- composition, matrix estimation
Abstract: Matrix completion algorithms recover a low rank matrix from a small fraction of the entries, each entry contaminated with additive errors. In practice, the singular vectors and singular values of the low rank matrix play a pivotal role for statistical analyses and inferences. This paper proposes estimators of these quantities and studies their asymptotic behavior. Under the setting where the di- mensions of the matrix increase to infinity and the probability of observing each entry is identical, Theorem 1 gives the rate of convergence for the estimated singu- lar vectors; Theorem 3 gives a multivariate central limit theorem for the estimated singular values. Even though the estimators use only a partially observed matrix, they achieve the same rates of convergence as the fully observed case. These esti- mators combine to form a consistent estimator of the full low rank matrix that is computed with a non-iterative algorithm. In the cases studied in this paper, this estimator achieves...
Distribution Statement: 3-Distribution authorized to U.S. Government Agencies and their contractors
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 18-Dec-2021

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Electronic Journal of Statistics
Publication Identifier Type: Other Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: 8/30/18 12:00AM Date Published: 8/12/16 12:18AM
Publication Location:
Article Title: Central limit theorems for network driven sampling
Authors: Xiao Li, Karl Rohe
Keywords: Volz-Heckathorn, link-tracing, Central Limit Theorem, Respondent-Driven Sampling
Abstract: Respondent-driven sampling is a popular technique for sampling hidden populations. This paper models respondent-driven sampling as a Markov process indexed by a tree. Theorems 1 and 2 study the asymptotic distribution of an estimator constructed from such a sample. The key technical difficulties stem from (i) the dependence between samples and (ii) the tree structure which characterizes the dependence. The theorems allow the growth rate of the tree to exceed one and suggest that this growth rate should not be too large. To illustrate the usefulness of these results beyond their obvious use, an example shows that in certain cases the sample average is preferable to inverse probability weighting. We provide a test statistic to distinguish between these two cases.
Distribution Statement: 3-Distribution authorized to U.S. Government Agencies and their contractors
Acknowledged Federal Support: **Y**

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Journal of Educational and Behavioral Statistics
Publication Identifier Type: Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: 8/27/17 12:00AM Date Published: 8/12/16 5:26AM
Publication Location:
Article Title: Latent Factors in Student-Teacher Interaction Factor Analysis
Authors: Thu Le, Daniel Bolt, Eric Camburn, Peter Goff, Karl Rohe
Keywords: Achievement, Individual Differences, Teacher Characteristics, Tracking, Factor Analysis
Abstract: Classroom interactions between students and teachers form a two-way or dyadic network. Measurements such as days absent, test scores, student ratings, or student grades can indicate the "quality" of the interaction. Together with the underlying bipartite graph, these values create a valued student-teacher dyadic interaction network. To study the broad structure of these values, we propose using Interaction Factor Analysis (IFA), a recently developed statistical technique that can be used to investigate the hidden factors underlying the quality of student-teacher interactions. Our empirical study indicates there are latent teacher (i.e. teaching style) and student (i.e. preference for teaching style) types that influence the quality of interactions. Students and teachers of the same type tend to have more positive interactions, and those of differing types tend to have more negative interactions. IFA has the advantage of traditional factor analysis in that the types are not presupposed;
Distribution Statement: 3-Distribution authorized to U.S. Government Agencies and their contractors
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 18-Dec-2021

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: PNAS
Publication Identifier Type: Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: 10/3/19 12:00AM Date Published:
Publication Location:
Article Title: Generalized least squares can overcome the critical threshold in respondent-driven sampling
Authors: Sebastien Roch, Karl Rohe
Keywords: Snowball sampling, link-tracing sampling, spectral gap
Abstract: ... Under a Markov model for Respondent-Driven Sampling (RDS), previous research has shown that if the typical participant refers too many contacts, then the variance of common estimators does not decay like $O (n^{-1})$, where n is the sample size. This implies that confidence intervals will be far wider than under a typical sampling design. Here we show that generalized least squares (GLS) can effectively reduce the variance of RDS estimates. In particular, a theoretical analysis indicates that the variance of the GLS estimator is $O(n^{-1})$. We then derive two classes of feasible GLS estimators. The first class is based upon a Degree Corrected Stochastic Blockmodel for the underlying social network. The second class is based upon a rank-two model. It might be of independent interest that in both model classes, the theoretical results show that it is possible to estimate the spectral properties of the population network from the sampled observations. ....
Distribution Statement: 3-Distribution authorized to U.S. Government Agencies and their contractors
Acknowledged Federal Support: **Y**

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Annals of Statistics
Publication Identifier Type: Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: 10/3/19 12:00AM Date Published:
Publication Location:
Article Title: A critical threshold for design effects in network sampling
Authors: Karl Rohe
Keywords: Markov sampling, spectral gap, graph, network
Abstract: Web crawling, snowball sampling, and respondent-driven sampling (RDS) are three types of network sampling techniques used to contact individuals in hard-to-reach populations. This paper studies these procedures as a Markov process on the social network that is indexed by a tree. Each node in this tree corresponds to an observation and each edge in the tree corresponds to a referral. Indexing with a tree (instead of a chain) allows for the sampled units to refer multiple future units into the sample. In survey sampling, the design effect characterizes the additional variance induced by a novel sampling strategy. If the design effect is some value DE, then constructing an estimator from the novel design makes the variance of the estimator DE times greater than it would be under a simple random sample with the same sample size n. Under certain assumptions on the referral tree, the design effect of network sampling has a critical threshold that is a function of the referral rate ...
Distribution Statement: 3-Distribution authorized to U.S. Government Agencies and their contractors
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 18-Dec-2021

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Annals of Applied Statistics
Publication Identifier Type: Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: 8/30/18 12:00AM Date Published:
Publication Location:
Article Title: Discovering Political Topics in Facebook Discussion threads with Spectral Contextualization
Authors: Yilin Zhang, Marie Poux-Berthe, Chris Wells, Karolina Koc-Michalska, Karl Rohe
Keywords: network; Facebook; topic; spectral clustering; node covariate; Stochastic co-Blockmodel
Abstract: We propose a new technique, Spectral Contextualization, to study political engagement on Facebook during the 2012 French presidential election. In particular, we examine the Facebook posts of the eight leading candidates and the comments beneath these posts. We find evidence of both (i) candidate-centered structure, where citizens pri- marily comment on the wall of one candidate and (ii) issue-centered structure (i.e. on political topics), where citizens' attention and ex- pression is primarily directed towards a specific set of issues (e.g. economics, immigration, etc). To discover issue-centered structure, we develop Spectral Contextualization, a novel approach to analyze a network with high-dimensional node covariates. This technique scales to hundreds of thousands of nodes and thousands of covariates. In the Facebook data, spectral clustering without any contextualizing infor- mation finds a mixture of (i) candidate and (ii) issue clusters. The contextualizing text ...
Distribution Statement: 3-Distribution authorized to U.S. Government Agencies and their contractors
Acknowledged Federal Support: **Y**

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Journal of Machine Learning Research (JMLR)
Publication Identifier Type: Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: 10/3/19 12:00AM Date Published:
Publication Location:
Article Title: A note on quickly sampling a sparse matrix with low rank expectation
Authors: Karl Rohe, Jun Tao, Xintian Han, Norbert Binkiewicz
Keywords: simulation; graphs; stochastic blockmodel; random dot product graph
Abstract: Given matrices $X,Y \in \mathds{R}^{n \times K}$ and $S \in \mathds{R}^{K \times K}$ with positive elements, this paper proposes an algorithm \texttt{fastRG} to sample a sparse matrix A with low rank expectation $E(A) = XSY^T$ and independent Poisson elements. This allows for quickly sampling from a broad class of stochastic blockmodel graphs (degree-corrected, mixed membership, overlapping) all of which are specific parameterizations of the generalized random product graph model defined in Section \ref{gRPG}. The basic idea of \texttt{fastRG} is to first sample the number of edges m and then sample each edge. The key insight is that because of the the low rank expectation, it is easy to sample individual edges. The naive ``element-wise" algorithm requires $O(n^2)$ operations to generate the $n\times n$ adjacency matrix A. In sparse graphs, where $m = O(n)$, ignoring log terms, \texttt{fastRG} runs in time $O(n)$. An implementation in \texttt{R} is available on github. ...
Distribution Statement: 3-Distribution authorized to U.S. Government Agencies and their contractors
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 18-Dec-2021

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 0-Other
Journal: Electronic Journal of Statistics
Publication Identifier Type: Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: 8/30/18 12:00AM Date Published:
Publication Location:
Article Title: Novel Sampling Design for Respondent Driven Sampling
Authors: Mohammad Khabbazian, Bret Hanlon, Zoe Russek, and Karl Rohe
Keywords: Hard-to-reach population; Social network; Trees; Markov chains; Spectral rep- resentation; Anti-cluster RDS
Abstract: Respondent-driven sampling (RDS) is a method of chain referral sampling popular for sampling hidden and/or marginalized populations. As such, even under the ideal sampling assumptions, the performance of RDS is restricted by the underlying social network: if the network is divided into communities that are weakly connected to each other, then RDS is likely to oversample one of these communities. In order to diminish the “referral bottlenecks” between communities, we propose anti-cluster RDS (AC- RDS), an adjustment to the standard RDS implementation. Using a standard model in the RDS literature, namely, a Markov process on the social network that is indexed by a tree, we construct and study the Markov transition matrix for AC-RDS. We show that if the underlying network is generated from the Stochastic Blockmodel with equal block sizes, then the transition matrix for AC-RDS has a larger spectral gap and consequently faster mixing properties than the standard random walk model for RDS...
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:** 1-Published
Journal: Neural Information Processing Systems (NeurIPS)
Publication Identifier Type: Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: 10/3/19 12:00AM Date Published:
Publication Location:
Article Title: Understanding Regularized Spectral Clustering via Graph Conductance
Authors: Yilin Zhang, Karl Rohe
Keywords: spectral clustering; regularization; graph cuts; localized eigenvector
Abstract: This paper uses the relationship between graph conductance and spectral clustering to study (i) the failures of spectral clustering and (ii) the benefits of regularization. The explanation is simple. Sparse and stochastic graphs create a lot of small trees that are connected to the core of the graph by only one edge. Graph conductance is sensitive to these noisy “dangling sets”. Spectral clustering inherits this sensitivity. The second part of the paper starts from a previously proposed form of regularized spectral clustering and shows that it is related to the graph conductance on a “regularized graph”. We call the conductance on the regularized graph CoreCut. Based upon previous arguments that relate graph conductance to spectral clustering (e.g. Cheeger inequality), minimizing CoreCut relaxes to regularized spectral clustering. Simple inspection of CoreCut reveals why it is less sensitive to small cuts in the graph. Together, these results show that unbalanced partitions from ...
Distribution Statement: 3-Distribution authorized to U.S. Government Agencies and their contractors
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 18-Dec-2021

Publication Type: Journal Article Peer Reviewed: N **Publication Status:** 1-Published
Journal: submitted to Annals of Statistics
Publication Identifier Type: Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: 8/31/18 12:00AM Date Published:
Publication Location:
Article Title: Asymptotic Seed Bias in Respondent-driven Sampling
Authors: Yuling Yan, Bret Hanlon, Sebastien Roch, Karl Rohe
Keywords: Limit distribution, Respondent-driven sampling, Galton-Watson process, Volz- Heckathorn estimator
Abstract: Respondent-driven sampling (RDS) collects a sample of individuals in a networked population by incentivizing the sampled individuals to refer their contacts into the sample. This iterative process is initialized from some seed node(s). Sometimes, this selection creates a large amount of seed bias. Other times, the seed bias is small. This paper gains a deeper understanding of this bias by characterizing its effect on the limiting distribution of various RDS estimators. Using classical tools and results from multi-type branching processes (Kesten and Stigum, 1966), we show that the seed bias is negligible for the Generalized Least Squares (GLS) estimator and non-negligible for both the inverse probability weighted and Volz-Heckathorn (VH) estimators. In particular, we show that (i) above a critical threshold, VH converge to a non-trivial mixture distribution, where the mixture component depends on the seed node, and the mixture distribution is possibly multi-modal. Moreover, ...
Distribution Statement: 3-Distribution authorized to U.S. Government Agencies and their contractors
Acknowledged Federal Support: **Y**

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 4-Under Review
Journal: Journal of the Royal Statistical Society, Series B
Publication Identifier Type: Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: Date Published:
Publication Location:
Article Title: argeted sampling from massive Blockmodel graphs with personalized PageRank
Authors: Fan Chen, Yini Zhang, Karl Rohe
Keywords: Community detection; Degree-Corrected Stochastic Blockmodel; Local clustering; Net- work sampling; Personalized PageRank
Abstract: This paper provides statistical theory and intuition for Personalized PageRank (PPR), a popular technique that samples a small community from a massive network. We study a setting where the entire network is expensive to thoroughly obtain or maintain, but we can start from a seed node of interest and ``crawl'' the network to find other nodes through their connections.By crawling the graph in a designed way, the PPR vector can be approximated without querying the entire massive graph, making it an alternative to snowball sampling. Using the Degree-Corrected Stochastic Blockmodel, we study whether the PPR vector can select nodes that belong to the same block as the seed node. We provide a simple and interpretable form for the PPR vector, highlighting its biases towards high degree nodes outside of the target block. We examine a simple adjustment based on node degrees and establish consistency results for PPR clustering that allows for directed graphs. These results are enabled by...
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 18-Dec-2021

Partners

,

I certify that the information in the report is complete and accurate:
Signature: Karl Rohe
Signature Date: 12/18/21 7:20AM

A suite of spectral techniques and statistical theory for sampling graphs
Karl Rohe
UW Madison
Final Report for W911NF-15-1- 0423.

The first wave of work published in this grant studied the process of obtaining representative samples of "nodes in a graph" by "tree sampling," as illustrated in this diagram:

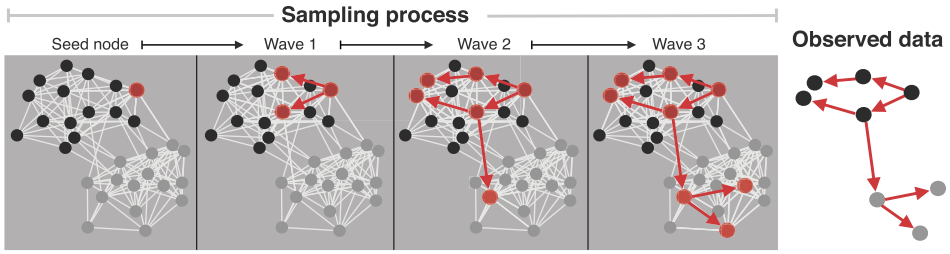

FIG. 1. *Network sampling has two graphs: the underlying social network and the referral tree. Each node in the social network has some feature (e.g., HIV status). In this diagram, the node feature is denoted by color. When we sample a node, we observe* (i) *the node's color and* (ii) *which node referred the node into the sample. In the end, we want to estimate the proportion of nodes that are grey.*

We want to estimate the proportion of grey nodes in the entire network. We sample the network with a tree and we only get to see the tree and the color of the nodes sampled. We show that the variance of standard estimators has two regimes (good and bad). Then, we showed that non-standard estimators, using the same data as previously, could obtain faster rates of convergence. These numerous results were published in Annals of Statistics, PNAS, and the Electronic Journal of Statistics. They are summarized in this table:

Table 1: Summary of properties of IPW and GLS estimators. In the columns, m refers to the number of participants that the typical participant refers into the study and λ_2 is the second eigenvalue of the Markov transition matrix.

Result	Estimator	Low variance, i.e. $m < \lambda_2^{-2}$	High variance, i.e. $m > \lambda_2^{-2}$
Variance	IPW	$O(n^{-1})$ (Rohe, 2019)	$O(n^{2\log_m \lambda_2})$ (Rohe, 2019)
	GLS	$O(n^{-1})$ (Roch and Rohe, 2018)	
Distribution	IPW&VH	Asymptotically normal (Li and Rohe, 2017)	Non-trivial mixture [Current paper]
	GLS	Asymptotically normal [Current paper]	

In the above work, the target of the sampling is "a representative sample" of *all nodes.* We studied two other problems of graph sampling; *targeted sampling* and *fast sampling/simulation of graphs.*

For targeted sampling, we studied Personalized Page Rank (PPR), i.e. the Google algorithm. Given a "seed node," we want to find other nodes in the graph that are very similar; this is akin to web search and this algorithm is fundamental to the initial form of Google. We provided the first rigorous statistical results showing how PPR is biased under certain social network models; however, a simple adjustment makes it unbiased; this form of bias is particularly interesting (and not necessarily bad!). We deployed these algorithms to twitter following networks and were hugely motivated by the high quality results that this simple algorithm obtains in sampling massive graphs, even with a very limited API rate.

Table 1. Top 15 handles by PPR clustering†

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

†Column names represent seed nodes, and the sampled nodes are ranked by PPR values, with teleportation constant $\alpha = 0.15$ uniformly. Through the PPR vector, the top 15 handles returned to each of the three seed nodes fit well with the characteristics of the seed nodes. They are popular or high status handles either directly related to the seed nodes or align with their political leanings. This shows the effectiveness of clustering via the PPR vector. It also shows the PPR vector's preference for highly connected nodes.

For sampling large graphs, our code fastRG is on CRAN (the Comprehensive R Archive Network) for simple installation in R. fastRG has three advantages. First, it samples any sparse random matrix whose expectation can be expressed as a low rank product of three matrices ZBY'; this includes Stochastic Blockmodels and all variations, but also a much broader class of models. Second, it is exceedingly fast. Using a laptop and R, this code simulates graphs on a million nodes in under a second. Third, the code is easy to use.

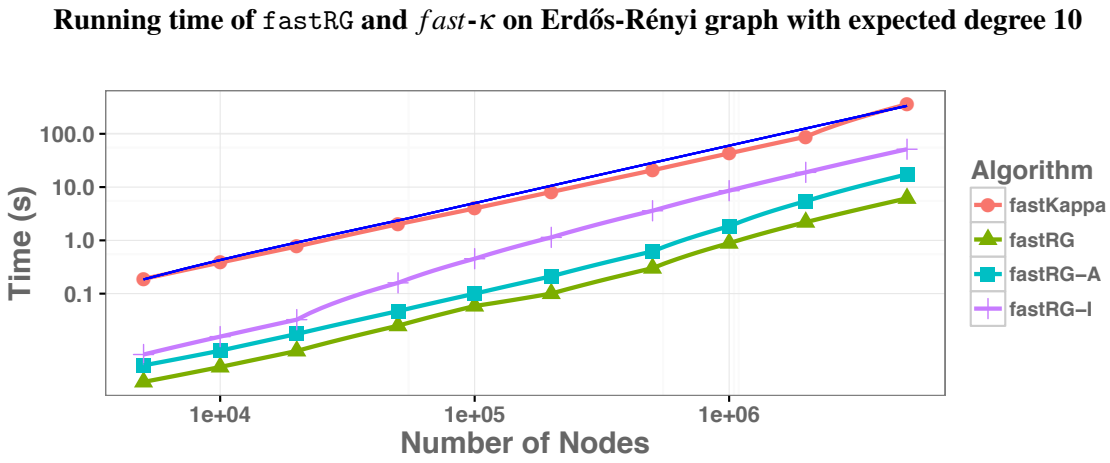

Figure 2: As the number of nodes increases (horizontal axis), all of the running times increase in parallel to the solid blue line which gives the rate $O(n \log n)$. The bottom three lines all correspond to `fastRG`, outputting three different graph types (edge list, sparse adjacency matrix, and igraph). For example, in roughly 8 seconds, *fast*-κ generates a graph with 20k nodes and `fastRG` generates an igraph with 1M nodes. To generate the random edge list on 1M nodes with `fastRG` takes less than 1 second