WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

UW Reference # MSN251318

American Family Mutual Insurance

PI: Miaoyan Wang

The Good, the Bad, the Pragmatic: Tensor Methods for Social Network Learning

This proposal has been administratively approved on behalf of the Board of Regents of the University of Wisconsin System and is submitted for your consideration. Please keep our office advised as developments occur with regard to this application.

The appropriate programmatic and administrative personnel of each institution involved in this grant application are aware of the sponsor's grant policy and are prepared to establish the necessary inter-institutional agreement(s) consistent with that policy.

All costs cited conform to established institutional policies and procedures. Our DHHS Negotiated Rate Agreement can be found at http://www.rsp.wisc.edu/rates/rates.pdf. Website: http://www.rsp.wisc.edu/

A final agreement is contingent upon the successful negotiation of terms and conditions acceptable to the University of Wisconsin-Madison.

We ask that you use the University's above-referenced proposal number in any future correspondence.

Questions regarding administrative matters should be directed to:

PreAward Services by email: preaward@rsp.wisc.edu or by phone: (608) 262-3822.

Questions regarding the technical nature of this application should be directed to:

The Principal Investigator.

Brenda Egan

Managing Officer 3/16/2021

Office of Research and Sponsored Programs
University of Wisconsin-Madison 21 N. Park St, STE 6401 Madison, Wisconsin 53715
608/262-3822 Fax: 608/262-5111 www.rsp.wisc.edu

Title: The Good, the Bad, the Pragmatic: Tensor Methods for Network Learning

Author: Dr. Miaoyan Wang, Assistant Professor in Department of Statistics, Affiliated with Institute for the Foundations of Data Science (IFDS)

WISPER Record Number: MSN251318

Abstract: The prevailing theme in the proposal is to develop powerful tensor methods for high-dimensional multi-layer network analysis. Rapid developments in modern technologies have made large-scale network datasets readily available. Modern networks are not only large in size, but they also have intricate structure. It is therefore of great importance to find a low-dimensional representation to better understand the key structure buried in noisy observations.

Higher-order tensors provide effective representation of multi-layer networks using multi-way structure. The PI will develop a framework — of tensor models, efficient algorithms, and softwares — to analyze multi-layer networks. Previous literature has advocated unfolding the tensor into a matrix and applying classical methods developed for matrices. Despite the popularity of such techniques, tensor method provides more powerful tools to capture complex structures in data that lower-order methods fail to exploit. The research goal goes beyond the traditional multivariate analysis; we aim to characterize probabilistic distributions over multi-layer edge connections, while taking into accounting the higher-order structures such as transitivity, balance, and community. This will allow researchers to examine complex interactions among entities in a context-specific manner, thereby providing solutions to questions that were previously impossible. The software packages resulting from this proposal, will be released freely, as well as related visualization tools for network analyses.

The Good, the Bad, the Pragmatic: Tensor Methods for Network Learning

My research project is broadly driven by questions related to understanding hidden salient structures of complex network datasets. The questions and techniques that we develop span across the fields of information theory, machine learning, and quantitative social sciences.

Motivations. A central theme in modern data analysis is to find a low-dimensional representation to better understand, compress, and convey the key phenomena buried in noisy observations. Modern datasets are not only large in size, but they also have intricate structure. A typical example is in the form of **network** [2, 10, 19], which is quantitative representation of interactions between entities in complex systems. As real-world networks are huge in nature, dimension deduction is crucial for pattern detection and subsequent specialized tasks.

In network studies, researcher are interested in interpretable low-dimensional structure within the high-dimensional relational data. The question goes beyond the traditional multivariate analysis; we aim to characterize probabilistic distributions over pairwise edge connections, while taking into accounting the higher-order structures such as transitivity ("a friend of a friend is a friend"), balance ("an enemy of a friend and an enemy"), and community ("cohesive subgroups of nodes"). **My project targets at developing higher-order tensor methods for analyzing high-dimensional network data.** The resulting prototypes will facilitate automatic detection of hidden salient structure in network data, thereby providing solutions to questions that cannot be addressed by existing methods. I will elaborate two specific directions for social network learning using tensor methods.

Community detection in multi-layer networks. A multi-layer network consists of multiple undirected graphs (or adjacency matrices), where each graph represents the connection among the same set of vertices (Fig 1a-b). The dataset is naturally organized as an order-3 tensor with the first two modes being vertices and the third mode being the contexts under which the graph is observed. Multilayer networks arise commonly in longitudinal study and multi-relational analysis. While the community structure in each single-layer network has been widely analyzed in the literature, little work has studied the heterogeneous pattern across multiple layers.

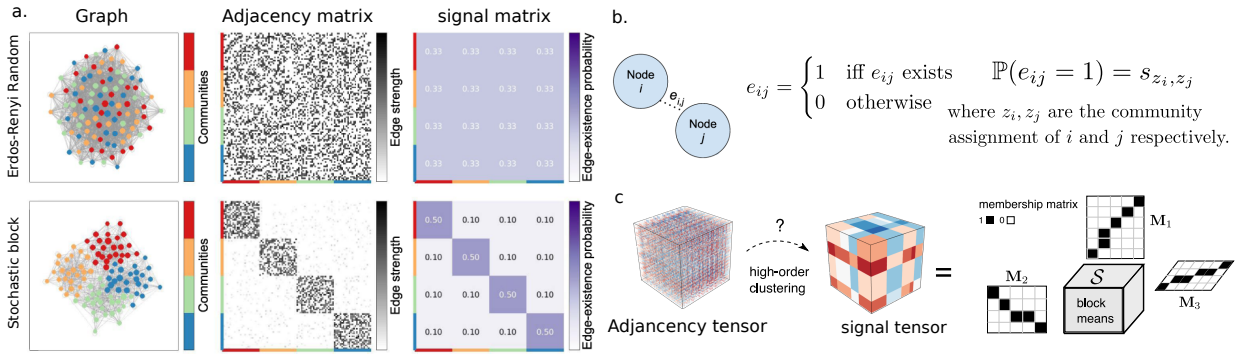

Figure 1: (a) Representation of network data using adjacency matrices [4]. (b) Stochastic block model for single-layer network (i.e., matrix). (c) We propose tensor extension of block model for high-order clustering in the context of multi-layer networks [5, 6].

Methods built on tensors provide generalized tools to capture complex data structure that the off-the-shelf methods may fail to exploit. We develop a tensor stochastic block model [6, 17] for simultaneous clustering of entities along each mode. Specifically, let $\mathcal{Z} = [\![z_{i_1,\ldots,i_d}]\!] \in \{0,1\}^{p\times\cdots\times p}$ denote the order-d adjacency data tensor, where the entries $z_{i_1,\ldots,i_d}$ represent the presence or absence of edge (i_1, i_2) at context $(i_3, \ldots, i_d)$. We model the signal tensor $\mathbb{E}(\mathcal{Z})$ using block structure

$$\mathbb{E}(\mathcal{Z}) = \mathcal{S} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_d \boldsymbol{M}_d,$$

for some low-dimensional core tensor $\mathcal{S}$ and community membership matrices $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_d$ (see Figure 1c). The learning goal is to estimate d-way connection strength tensor $\mathcal{S}$ and community assignment $(\boldsymbol{M}_1, \ldots, \boldsymbol{M}_d)$ from a noisy observation $\mathcal{Z}$. We propose a higher-order Lloyd algorithm, which uses alternating optimization for parameter estimations. Our preliminary analysis shows that the tensor algorithm achieves *exact recovery of communities* under less stringent assumptions than existing algorithms. Surprisingly, we find that the learning performance is fully characterized by signal-to-noise ration (SNR) (see Figure 2). In the strong SNR region A, we prove that the our algorithm achieves exact clustering *in polynomial time*. We also show that the estimation error bound of the target tensor is *free of tensor dimension*. This feature is especially appealing in modern large-scale network analysis. In the weak SNR region B, we provide evidence to the

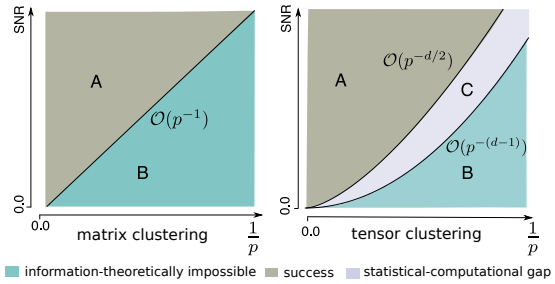

Figure 2: Comparison between matrix and tensor methods.

conjecture of information-theoretical impossibility of consistent clustering. In the modest SNR region C, the problem exhibits a gap between computational and statistical limits, a phenomenon that has drawn much interests in modern learning problems [1,3,18]. Our result provides the first characterization of trade-off for higher-order clustering, and the established results serve the benchmark for algorithm development.

Learning latent motifs in network. When analyzing imaging data, one can learn individual patches and pixels (local features) or total variations (global feature) in order to extract useful information from images (Figure 3a). Analogously for networks, one may learn their local structures (e.g., nodes and edges) or global structure (community or core-periphery structures). At the same time, we are facing ever-growing network size due to the explosion of recent technology in measuring, processing, and storing network data. Hence, it would be an important and timely contribution to develop algorithms that compress a large network into a set of multi-scale, interpretable structures.

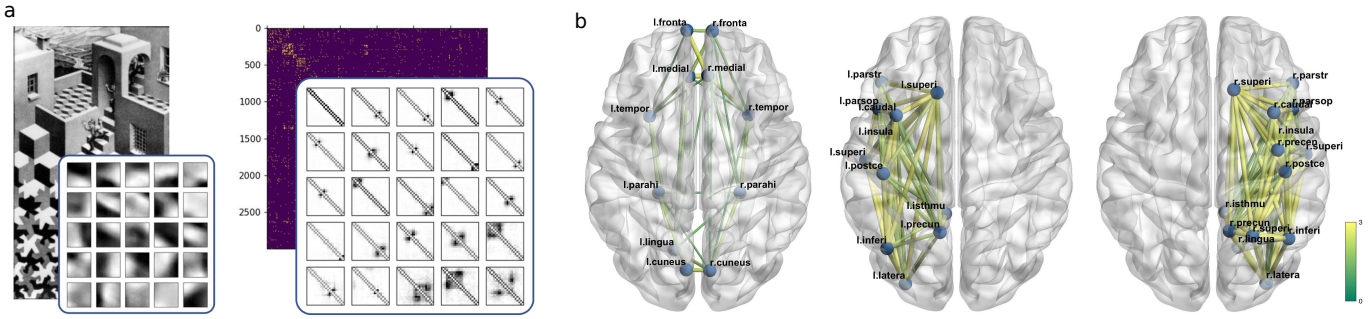

Figure 3: (a) Latent motifs learned from images and social network data [9]. (b) Connectivity motifs learned from human brain networks [7].

We propose to develop a systematic approach to learning latent *network motifs* that form the salient architecture of networks. Our method is based on structured tensor decomposition where each rank-1 tensor is interpreted as a latent motif for multi-way interaction. The ability to encode a network using a set of latent motifs opens up a wide variety of network-analysis tasks, such as network completion, denoising, and link prediction. For example, Figure 3a shows 25 latent motifs learned from images and social network data based on nonnegative matrix decomposition. My group has been generalizing the method to tensors and applying the method to Human Connectome Project (HCP). The result reveals a clear spatial separation among brain node between hemispheres (Figure 3b). The identified latent similarities among nodes *without external labels* highlights the potential power of tensor methods to pattern discovery.

The latent motif learning is challenging because of the extremely high dimensionality in the network data. The PI's previous extensive experience on tensor related work has shown that tensors sought in network applications often possess special structures, such as low-rankness [8,11,13], sparsity [14], non-negativity [12], or orthogonality [16]. We will leverage the formalisms of *intrinsic dimension* to develop efficient statistical methods for analyzing these high-dimensional datasets. We will further develop adaptive, semi-supervised methods that incorporate practical constraints, such as incomplete observation, corrupted distributions, and computation with time and memory constraints.

Deliverables and Milestones. The PI is a young faculty in statistics with affiliation in Institute for the Foundations of Data Science (IFDS). The PI has actively contributed to several multi-institutional consortiums, in collaboration with researchers in both academics and industry. This grant will support the PI's group to further create a diverse working group. Open-source software will be released, as the fruit of the research, that facilitates academia, industry, and society to analyze complicated tensor data. The following table summarizes the planned milestones.

	Year 1			Year 2		
	Jan	May	Sept	Jan	May	Sept
Research	build model for supervised learning			build model for unsupervised learning		
		optimization algorithm			non-convex algorithm	
			publication 1			publication 2
Education		new course on data science			create online courses	
	bootcamp meeting			organize workshop		reunion symposia
			student presentation			student presentation
Social Impact		build recommendation systems			build pipelines for multimodal analytics	
			release free software			release free software
		dissemination of research results in academic and industrial conferences				

Budget

Y1:

PI Miaoyan Wang: 1 summer month with 34.6% fringe = $17,894

Graduate Student Research Assistant 1 (TBD): 4.5 academic months with 17.2% fringe = $11,898

Graduate Student Research Assistant 2 (TBD): 2 summer months with 17.2% fringe = $5,288

Graduate Student Research Assistant 3 (TBD): 2 summer months with 17.2% fringe = $5,288

One semester tuition remission for Graduate Student: $6,000

F&A: 55.5% of MTDC = $22,404

Total Y1 costs: $68,772

Y2:

PI Miaoyan Wang: 1 summer month with 34.6% fringe = $18,568

Graduate Student Research Assistant 1 (TBD): 4.5 academic months with 17.2% fringe = $12,360

Graduate Student Research Assistant 2 (TBD): 2 summer months with 17.2% fringe = $5,493

Graduate Student Research Assistant 3 (TBD): 2 summer months with 17.2% fringe = $5,493

Graduate Student Research Assistant 4 (TBD): 2 summer months with 17.2% fringe = $5,493

One semester tuition remission for Graduate Student: $6,000

F&A: 55.5% of MTDC = $26,311

Total Y2 costs: $79,718

Total requested for this project: $148,490

Budget Justification

Senior/Key and Other Personnel:

Miaoyan Wang is an Assistant Professor at UW-Madison and will serve as PI. She will devote 1 summer month to the project each year. The PI will be responsible for designing and developing the computational models used in the project, creating and running the simulation, and analyzing datasets for inference.

One Graduate Student Research Assistant (TBD) will devote 4.5 academic months to the project each year. Three Graduate Student Research Assistants (TBD) will devote 2 summer months to the project in Y1, and Four Graduate Student Research Assistants (TBD) will devote 2 summer months to the project in Y2. The Research Assistants will be responsible for developing software for simulation and data analysis.

Annual inflation is budgeted at 3% for all salaries.

Fringe Benefits:

Fringe benefits for the PI: Y1-34.6%; Y2-35.6%
Fringe benefits for the Research Assistants: Y1-17.2%; Y2-18.2%

Other Direct Costs:

$12,000 is requested for Graduate Student Research Assistant tuition remission ($6,000 in Y1 and $6,000 in Y2).

Indirect Costs:

The F&A rate (MTDC) is 55.5%.

References

[1] Quentin Berthet and Philippe Rigollet, Complexity theoretic lower bounds for sparse principal component detection, *Conference on learning theory (COLT)*, 2013, pp. 1046–1066.

[2] Peter J Bickel and Aiyou Chen, A nonparametric view of network models and Newman–Girvan and other modularities, *Proceedings of the National Academy of Sciences (PNAS)* 106 (2009), no. 50, 21068–21073.

[3] Matthew Brennan, Guy Bresler, and Wasim Huleihel, Reducibility and computational lower bounds for problems with planted sparse structure, *Conference on learning theory (COLT)*, 2018, pp. 48–166.

[4] Joshua Faskowitz, Xiaoran Yan, Xi-Nian Zuo, and Olaf Sporns, Weighted stochastic block models of the human connectome across the life span, *Scientific reports* 8 (2018), no. 1, 1–16.

[5] Jiaxin Hu, Chanwoo Lee, and **Wang, Miaoyan**, Supervised tensor decomposition with interactive side information, Advances in *Neural Information Processing Systems (NeurIPS)* 33 Workshop on Machine Learning and the Physical Sciences, 2020.

This work wins **Best Student Paper Award** from the Statistical Computing and Graphics Section of American Statistical Association (ASA), 2021.

[6] Rungang Han, Yuetian Luo, **Miaoyan Wang**, and Anru R Zhang, Exact clustering in tensor block model: Statistical optimality and computational limit, arXiv preprint arXiv:2012.09996 (2020).

This work wins **Best Student Paper Award** from the Statistical Learning and Data Science Section of the American Statistical Association (ASA), 2021.

[7] Chanwoo Lee and **Miaoyan Wang**, Tensor denoising and completion based on ordinal observations, *International conference on machine learning (ICML)*, 2020, pp. 5778–5788.

[8] Chanwoo Lee and **Miaoyan Wang**, Beyond the signs: Nonparametric tensor completion via sign series, arXiv preprint arXiv:2102.00384 (2021).

[9] Hanbaek Lyu, Deanna Needell, and Laura Balzano, Online matrix factorization for markovian data and applications to network dictionary learning, *Journal of Machine Learning Research* 21 (2020), no. 251, 1–49.

[10] Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *Annals of Statistics*, Vol 39, 4, 1878–1915, 2011.

[11] **Miaoyan Wang**, Khanh Dao Duc, Jonathan Fischer, and Yun S Song, Operator norm inequalities between tensor unfoldings on the partition lattice, *Linear Algebra and Its Applications* 520 (2017), 44–66.

[12] **Miaoyan Wang**, Jonathan Fischer, and Yun S Song, Three-way clustering of multi-tissue multi-individual gene expression data using semi-nonnegative tensor decomposition, *The Annals of Applied Statistics* 13 (2019), no. 2, 1103–1127.

[13] **Miaoyan Wang** and Lexin Li, Learning from binary multiway data: Probabilistic tensor decomposition and its statistical optimality, *Journal of Machine Learning Research* 21 (2020), no. 154, 1–38.

[14] **Miaoyan Wang**, Fabrice Roux, Claudia Bartoli, Carine Huard-Chauveau, Christopher Meyer, Hana Lee, Dominique Roby, Mary Sara McPeek, and Joy Bergelson, Two-way Mixed-effects methods for joint association analysis using both host and pathogen genomes, *Proceedings of the National Academy of Sciences (PNAS)* 115 (2018), no. 24, E5440–E5449.

[15] **Miaoyan Wang**, Johanna Jakobsdottir, Albert V. Smith, and Mary Sara McPeek. G-STRATEGY: Optimal selection of individuals for sequencing in genetic association studies. *Genetic Epidemiology*, Vol. 40, No. 6, 446–460, 2016.

Highlighted as **Editor's Pick Paper** of this issue. This work wins **ASHG Charles J. Epstein Trainee Award** and **IGES Williams Award**.

[16] **Miaoyan Wang** and Yun Song, Tensor decompositions via two-mode higher-order SVD (HOSVD), *Artificial intelligence and statistics*, 2017, pp. 614–622.

[17] **Miaoyan Wang** and Yuchen Zeng, Multiway clustering via tensor block models, *Advances in Neural Information Processing Systems (NeurIPS)* 32 . (2019).

[18] Yihong Wu and Jiaming Xu, Statistical problems with planted structures: Information-theoretical and computational limits, *Information-Theoretic Methods in Data Science* (2021), 383.

[19] L Zheng and G Raskutti. Testing for high-dimensional network parameters in auto-regressive models. *Electronic Journal of Statistics*, 13(2): 4977-5043 (2019).

Miaoyan Wang
Assistant Professor of Statistics
1300 University Avenue, UW-Madison
miaoyan.wang@wisc.edu

Fudan University	China	Mathematics	B.S. 2010
		Computer Science	2006-2007
University of Chicago	USA	Statistics	PhD, 2015
University of Pennsylvania	USA	Mathematics and Biology	Simons Math+X Postdoc, 2017
UC Berkeley	USA	Computer Science	Postdoc 2018

(a) Professional preparation
(b) Appointment

2018 - Present, Assistant Professor, Department of Statistics, University of Wisconsin–Madison

(c) Top 10 Publications (Students under my supervision are underlined)

J. Hu, C. Lee, and **M. Wang**, Supervised tensor decomposition with interactive side information, Advances in Neural Information Processing Systems (NeurIPS) 33 Workshop on Machine Learning and the Physical Sciences, 2020. **Best Student Paper Award** from the Statistical Computing and Graphics Section of American Statistical Association (ASA), 2021

R. Han, Y. Luo, **M. Wang**, and A. R Zhang, Exact clustering in tensor block model: Statistical optimality and computational limit, arXiv preprint arXiv:2012.09996 (2020). **Best Student Paper Award** from the Statistical Learning and Data Science Section of the American Statistical Association (ASA), 2021.

C. Lee and **M. Wang**. Tensor denoising and completion based on ordinal observations. International Conference on Machine Learning (ICML). 2020.

M. Wang and Y. Zeng. Multiway clustering via tensor block models. Advances in Neural Information Processing Systems 32 (NeurIPS), 715-725, 2019

M. Wang, J. Fischer, and Y. S. Song. Three-way Clustering of Multi-tissue Gene Expression Data Using Semi-Nonnegative Tensor Decomposition. Annals of Applied Statistics. Vol. 13, No. 2, 1103-1127, (2019).

M. Wang, K. Dao Duc, J. Fischer, and Y.S. Song. Operator Norm Inequalities Between Tensor Unfoldings on the Partition Lattice. Linear Algebra and its Applications, Vol.520, 44-66, (2017).

M. Wang and Y. S. Song. Tensor Decomposition via Two-Mode Higher-Order SVD (HOSVD). Proceeding of Machine Learning Research, Vol 54, 614-622, (2017).

M. Wang and L. Li. Learning from Binary Multiway Data: Probabilistic Tensor Decomposition and Its Statistical Optimality. Journal of Machine Learning Research. 21 (2020), no. 154, 1–38.

M. Wang, F. Roux, C. Bartoli, C. H.-Chauveau, C. Meyer, H. Lee, D. Roby, M. S. McPeek, and J. Bergelson. Two-Way Mixed-Effects Methods for Joint Association Analyses Using Both Host and Pathogen Genomes. Proc. Natl. Acad. Sci. (direct submission), Vol. 115 (24), E5440-E5449, (2018).

M. Wang, J. Jakobsdottir, A. V. Smith, and M. S. McPeek. G-STRATEGY: Optimal Selection of Individuals for Sequencing in Genetic Association Studies. Genetic Epidemiology, Vol. 40, No. 6, (2016) 446-60. Highlighted as **Editor's Pick Paper** of this issue. This work wins **ASHG Charles J. Epstein Trainee** Award and **IGES Williams Award.**

(d) Synergistic Activities

• Member in Women in Probability, Institute of Mathematical Statistics, Society for Industrial and Applied Mathematics, American Society of Human Genetics. 2014 – now.
• Organizer for European Society for Evolution Biology workshop, International Conference on Frontier of Data Science, 2019.
• Reviewer for Journal of the American Statistical Association (JASA), NeurIPS, and Linear Algebra and application, and other applied math/statistics/genetics journals, 2014 – now.
• Statistical Consultant. Provided statistical support for the larger university community at the University of Chicago. 2012-2015.

(e) Current PhD students

Chanwoo Lee (2019 -): BS in Mathematics and Statistics, Seoul National University, 2018.
Jiaxin Hu (2020 -): BS/MS in Statistics, Wuhan University, 2020.
Yuchen Zeng (2019 -): current a PhD student in CS at UW-Madison.
Zhuoyan Xu (2019 -): current a PhD student in Statistics at UW-Madison.

(f) Recent Talks

Department Seminars: Columbia University, Stanford University, UC Berkeley, University of Chicago, CMU, Columbia University, University of Toronto, Fudan University, East China Normal University, Duke University, Johns Hopkins University, Queen's University, University of Massachusetts Amherst, University of Pennsylvania, Boston University.

Conference and Industrial Talks: Eastern North American Region (ENAR), International Conference on Frontiers of Data Science, European Society for Evolutionary Biology, Institute of Mathematical Statistics (IMS), Society for Industrial and Applied Mathematics (SIAM), Joint Statistical Meeting (JSM), American Society of Human Genetics (ASHG), International Genetic Epidemiology Society (IGES).

Industrial Research Lab Talks: Bosch Center for Artificial Intelligence, Takeda Pharmaceutical.

(g) Research Impact and Outreach

• Developed open-source software packages for analyzing tensor datasets in genomics and neuroimaging.
• Faculty feature article ``Women in STEM: 5 Thoughtful Ways to Recruit and Retain Them'' in Course Hero.
• Won **Charles J. Epstein Trainee Award** for Excellence in Human Genetics Research –semifinalist (27 protectoral recipients out of 550 candidates), 2014
• Won **Williams Award** for Best Platform Presentation by Graduate Students – finalist (3 out of 156) in International Genetic Epidemiology Society (IGES), 2013.
• Runner-up for Department of Statistics Consulting Award (ranked as #2 among all PhD students in the departmental vote of 2014). Department of Statistics, The University of Chicago.
• Madison Teaching and Learning Excellence (MTLE) Fellow, 2019 -2020.