# Qiaohui (Chelsea) Lin

2317 Speedway, Austin, Texas 78712 https://github.com/qiaohuilin Email: qiaohui.lin@utexas.edu Cell: (919) 433-6310

## SKILLS AND QUALIFICATIONS

- **Statistical Background:** Four years' research experience in Data Sciences and Machine Learning.
- **Statistical Specialty:** Strong quantitative skills in Bayesian predictive modeling, Bayesian hierarchical modeling, cluster analysis, non-parametric methods, and uncertainty quantification, with familiarity in Causal Inference and Time Series.
- **Programming:** Proficient coding in R, Rcpp and Python, with focus on reproducible workflows.
- **Communication:** Strong written and verbal communication skills, including publication in peer-reviewed journals and presentation in international conferences.

## EDUCATION

**University of Texas at Austin** — Austin, TX
*PhD Candidate in Statistics and Data Sciences,* GPA: 4.0/4.0 — Expected May, 2022
Supervisors: Dr. Purnamrita Sarka, Dr. Peter Mueller

**Duke Univerisity** — Durham, NC
*M.A. in Economics*, GPA: 3.80/4.00, MA. Merit Scholar — May, 2018

**Fudan University** — Shanghai, China
*B.A. in Economics*, GPA: 3.81/4.00, Outstanding Graduate of Fudan 2016 — Jun., 2016

## RESEARCH AND PUBLICATION

**Network Statistic Uncertainty Quantification**

- Develop consistent variance estimator for statistics in large scale sparse graph in networks using subsampling methods. -- This approach is more accurate and computationally efficient compared to other algorithms.
- Paper '*On the Theoretical Properties of the Network Jackknife'*, published and presented in *ICML 2020, the 37th international conference on machine learning*.

**Recommender System Under Bayesian Double Feature Allocation Model**

- Develop a Bayesian nonparametric model in a recommendation system to predict a user's rating/ranking to unseen items. -- This approach is bi-clustering users and items and is more explainable than black box prediction.

**Covariate-dependent Distributional Clustering**

- Develop a Bayesian model and successfully identify clusters of proteins with increased abundance with age and condition in neural degeneration disease Ataxia. -- This is a new approach for recognizing clusters that are covariate dependent, with over 95% accuracy in simulation and have possibly many more applications.

## WORK EXPERIENCE

**Amazon** — Seattle, WA
*Applied Scientist Intern*, — Jun. 2020 – Aug. 2020

- Analyze historical customer data using PySpark on Amazon S3 for customer holiday behavior.
- Optimize delivery scheme in holiday seasons for revenue maximization using Gradient Boosted Feature Selection model, a modified and fast gradient boosted tree algorithm.

**Vrbo, Expedia Group** — Austin, TX
*Data Scientist Intern,* — Jun. 2019 – Aug 2019

- Project in Search Engine Optimization: Keyword Clustering and Performance Prediction.
- Cluster Search Keywords based on NLP methods and use a Gaussian Process for time series analysis of conversion rates in each cluster.

**Duke Clinical Research Institute** — Durham, NC
*Research Assistant* — May 2017 – Aug 2017

- Prediction on appointment no-shows using Bayesian hierarchical model (accuracy over 90%). Paper *'Prediction of Appointment No-shows using Electronic Health Records'* published in *Journal of Applied Statistics, 2019.*
- Presented on Women in Machine Learning Conference 2017