

# Syed Arefinul Haque

✉ syedarehaq@gmail.com  
🌐 linkedin.com/in/syedarehaq 📄 Google Scholar

## SUMMARY

Data scientist with a background in network science, machine learning, and natural language processing. Harnesses training in network science to extract, analyze and visualize vital nodes, hidden communities, vulnerabilities, and diffusion dynamics in large scale networks. Scientific curiosity and rigor blended with communication skills and pragmatism borrowed from business background provides an extra edge in collaborative environment. Looking forward to utilizing network science and statistics skills to model and solve real-world problems.

## TECHNICAL STRENGTHS

**Analytical Skills:** Network Analysis, Network Visualization, Regression Analysis, Natural Language Processing, Agent Based Modelling, Epidemiology, Bayesian Statistics, Clustering Techniques, Embedding Methods, Logistic Regression, Markov Chain Monte Carlo (MCMC), Retrospective Analysis

### Computing Skills:

Programming: Python, R, JavaScript, Bash

Database: SQL, BigQuery, MongoDB, neo4j

Visualization: Matplotlib, Seaborn, D3, Cytoscape, Gephi, GnuPlot, Adobe Illustrator

Network Analysis: NetworkX, Graph-tool, iGraph

Other: Unix, Git, SLURM, Google Cloud, L<sup>A</sup>T<sub>E</sub>X

## EDUCATION

### Northeastern University

Ph.D. in Network Science

Boston, MA, USA

Expected April, 2021

### United International University

M.Sc. in Computer Science and Engineering, *summa cum laude*

Dhaka, Bangladesh

2015

Thesis : "Virtual P2P client: accessing P2P applications using virtual terminals"

### Institute of Business Administration, University of Dhaka

B.B.A. in Finance (Minor in Marketing)

Dhaka, Bangladesh

2013

## EXPERIENCE

### Graduate Researcher, Northeastern University

2015-Present

#### Evolution and impact of revisions related to gender bias and sexism in Wikipedia

- Extracted all the revision comments from the revision history of all the available Wikipedia edits. This required efficient extraction of specific fields from the large scale Wikipedia xml dumps of around 50TB.  
jq, xml, BigQuery
- Working on using phrase embedding to identify the gender bias correction related edits in Wikipedia. Using regression discontinuity to detect the the impact of such gender bias related edits on the over all tone of the articles. Word Embedding, Phrase Embedding, Regression Discontinuity Analysis

#### Diversity of COVID-19 experts in news media

- Identified experts mentioned in COVID-19 news collected from Media Cloud API using named entity recognition techniques. Aim of this project is to understand who are getting represented as spokespersons of COVID-19 related research in the news media, and whether its informational value gets diluted with the co-mention of politicians. NLP, NER
- Organized a hackathon where interested volunteers worked on identifying the race, gender and expertise of 5500 people mentioned in COVID-19 related news. SOP Design, Project Management

#### Mapping organizational change for gender equity using NSF grant data

- Visualized co-occurrence of gender equity related keywords found the grant data to show the change of language related to implicit gender bias. NLP, Network Visualization

- Managed large scale bibliographic database using Google Cloud Platform to identify researchers who worked on projects that ameliorate gender bias in academia. Google-BigQuery

### Reconstructing pathways of Zika virus epidemic in Americas

- Collected genomic inferences and surveillance data on the Zika virus and applied statistical techniques to compare them with model generated data to learn how the disease spread throughout Americas. Epidemiology, Simulation Google-BigQuery; Python Cross Correlation, Linear Regression
- Developed a web based interactive visualization which illustrates the simulated imported Zika cases in more than 3000 urban areas throughout the world. D3; MongoDB; ExpressJS

### Research Assistant, United International University

2013-2015

- Developed a virtualized Peer-to-Peer (P2P) client of BitTorrent protocol that was deployed in web server to measure how well P2P clients perform in the edge clouds like AWS and Linode. Amazon AWS; WebRTC; ExpressJS

### Business Development Executive, Mukto Software Limited

2013 - 2015

- Served as a liaison between the corporate customer and the software development team by outlining requirements of enterprise resource planning (ERP) software projects. Project Management; Kanban

## SELF DIRECTED PROJECTS

### Bias in Bangladeshi newspaper portrayal

2018

Crawled and curated newspaper data from six Bangladeshi newspapers and used named entity recognition (NER) tools to find actors in those news articles. Through this analysis we were able to show the bias towards political actors in news reporting. NLP; NER

## SELECTED COURSE PROJECTS

### Bayesian and Network Statistics

2017

Analyzed dataset from Second Life online platform to explore the factors behind formation of trust relationship between users. Social Network Analysis; ERGM

### Dynamical Processes on Complex Networks

2016

Developed a predictive epidemiological model of 2016 US presidential election using social media and poll data. Mechanistic Modeling; SIR Process

### Network Science Data

2016

Crawled GitHub social network and analyzed how the reputation of a user increases based on their collaboration in diverse projects. Network Analysis; Web Crawling

## PUBLICATIONS

Chowdhury, S. S., Saquib, N., Zawad, N., Mandal, M.K. & **Haque, S. A.** (2018), Statement networks: a power structure narrative as depicted by newspapers. *Proceedings of NeurIPS 2018 workshop on Machine Learning for the Developing World*

Hassan, M. K., Islam, L. & **Haque, S. A.** (2017), Degree distribution, rank-size distribution, and leadership persistence in mediation-driven attachment networks. *Physica A: Statistical Mechanics and its Applications*, 469, 23-30

**Haque, S. A.**, Islam, S., Islam, M. J., & Grégoire, J. C. (2016). An architecture for client virtualization: A case study. *Computer Networks*, 100, 75-89.

Mistry, D., Litvinova, M., Pastore y Piontti, A., Mu, K., Xiong, X., **Haque, S. A.**, Quan-Hui, L., Gomes, M. F. C., Fumanelli, L., Longini Jr, I. M., Halloran, M. E., Merler, S., Ajelli, M. & Vespignani, A. Inferring high-resolution disease-specific human mixing patterns. (Under review)

## ADVANCED TRAININGS AND CERTIFICATES

9th Summer Institute in Statistics and Modeling in Infectious Diseases (SISMID), University of Washington

2017

Complex System Summer School (CSSS 2016), Santa Fe Institute, New Mexico

2016