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, Natural Language Processing, Healthcare NLP, Biomedical Ontology, Regression Analysis, Agent Based Modelling, Epidemiology, Bayesian Statistics, Clustering Techniques, Embedding Methods

Computing Skills:
Programming: Python, R, JavaScript, Bash
Database: SQL, BigQuery, MongoDB, neo4j, ElasticSearch
Visualization: Matplotlib, Seaborn, D3, Cytoscape, Gephi, GnuPlot, Adobe Illustrator
Network Analysis: NetworkX, Graph-tool, iGraph
Other: Unix, Git, HPC SLURM, Google Cloud, Biomedical Onotologies (UMLS, Snomed-CT), LaTeX

Education

Northeastern University — Boston, MA, USA
Ph.D. in Network Science — Expected June, 2022

United International University — Dhaka, Bangladesh
M.Sc. in Computer Science and Engineering, *summa cum laude* — 2015
Thesis : "Virtual P2P client: accessing P2P applications using virtual terminals"

Institute of Business Administration, University of Dhaka — Dhaka, Bangladesh
B.B.A. in Finance (Minor in Marketing) — 2013

Experience

ORISE Fellow of AI & Drug Safety, OTS, CDER, FDA — July 2021- Present

- Created NLP pipeline to extract drug-adverse event information from free text clinical narratives related to mental health disorders, opioid addiction, and opioid overdose and to identify the infectious and non-infectious complications reported in large scale health records such as MIMIC-III.
- Designed novel evaluation technique to compare accuracy of opensource and proprietary Healthcare and Biomedical NER resources such as Spark-NLP, MedCAT, BlueBERT.
- Disambiguated clinical entities using medical ontology such as Unified Medical Language System (UMLS), Medical Dictionary for Regulatory Activities (MedDRA), and visualized the knowledge graph using D3.

LEADER's Co-Op, Merck & Co., Inc. — Jan 2021- May 2021

- Visiting network scientist co-op at Merck through Northeastern's experiential PhD LEADER's program.
- Created a network data structure of the clinical documents and visualized the pathway of information shared between them using neo4j and D3.

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
- Created 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

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

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

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

Cevik, M., **Haque, S. A.**, Manne-Goehler, J., Kuppalli, K., Sax, P.E., Majumder, M.S. and Orkin, C., 2021. Gender disparities in COVID-19 clinical trial leadership. *Clinical Microbiology and Infection*.

Mistry, D., Litvinova, M., y Piontti, A.P., Chinazzi, M., Fumanelli, L., Gomes, M.F., **Haque, S. A.**, Liu, Q.H., Mu, K., Xiong, X. and Halloran, M.E., 2020. Inferring high-resolution human mixing patterns for disease modeling. *Nature Communications*, 12(1), pp.1-12.

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.

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