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 build knowledge graph, and 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, Natural Language Processing, Knowledge Graph, Healthcare NLP, Biomedical Ontology, Network Visualization, 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
Healthcare-NLP: Spark-NLP, MedSpacy, MedCAT, Ontologies (UMLS, Snomed-CT, RxNorm), EHR data (MIMIC-III)
Other: Unix, Git, HPC SLURM, Google Cloud, LaTeX

Education

Northeastern University — Boston, MA, USA
Ph.D. in Network Science — July, 2022
Thesis: “Diversity and gender equity in networks of knowledge production and dissemination”

United International University — Dhaka, Bangladesh
M.Sc., *summa cum laude* in Computer Science and Engineering — 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

Research Scientist, NLP & Advanced Analytics, Independence Blue Cross — September 2022- Present

- Created and visualized knowledge graph from call center transcripts to summarize emerging issues around claims and customer experience.
- Identified important concepts appearing and disappearing in unstructured and noisy text transcripts derived from call center.

ORISE Postdoctoral Fellow of AI & Drug Safety, US FDA — July 2021- September 2022

- During this fellowship at Office of Translational Science, CDER at FDA, I 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 electronic health records such as MIMIC-III using Elasticsearch, Spark-NLP, MedSpacy.
- Designed 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.

Visiting data scientist, 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 2015-2022

Measuring the Scope and Recall of Wikipedia's Coverage of Three Women's Movement Subgroups

- Extracted all the text of the article from the Wikipedia dump. This required efficient extraction of specific fields from the large scale Wikipedia xml dumps of around 50TB. jq, xml, BigQuery
- Created a large-scale fuzzy text search pipeline to calculate the recall of the gender concepts in wikipedia articles. elasticsearch, unix, slurm

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

Working Papers

Haque, S. A., Hasan, M. R., Hasan, M. M., Jermyn, R., Hussein, A., Vega, A., Zembrzuski, K., Ripple, A., Ahadpour, M., Francis, H. & Sorbello, A. Identifying opioid-related adverse events from unstructured text in electronic health records using rule-based algorithms and deep learning methods. (Submitted to *AMIA 2022*)

Publications and conference proceedings

Nelson, L. K., Getman, R. & **Haque, S. A.** (2021). And the Rest is History: Measuring the Scope and Recall of Wikipedia's Coverage of Three Women's Movement Subgroups. *Sociology Methods & Research*, Online First

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

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