Department of Computer and Information Science
University of Pennsylvania
+1 720-717-3398
upasanad@seas.upenn.edu
upasanadutta98.github.io
UpasanaDutta98

Upasana Dutta

Education

2022-present **PhD in Computer and Information Science, University of Pennsylvania** GPA: 3.88/4.0
Advisors: Duncan Watts, Aaron Clauset

2019-2022 **Master of Science in Computer Science, University of Colorado Boulder** GPA: 4.0/4.0

2015-2019 **B.Tech in Computer Science and Engg., Heritage Institute of Technology, India** GPA: 9.36/10

Publications and Research

2023-present **Active Learning with Pretrained Embeddings: Measuring Partisanship in TV News**
Upasana Dutta, Homa Hosseinmardi, Amir Ghasemian, Aaron Clauset, Duncan Watts
Leveraged LLMs and Active Learning for training a neural network using transformer-based embeddings with partisanship labels to predict media bias in 330,000+ TV news episodes across 2013-2022.

2024-present **Production and Consumption patterns in the News Ecosystem on YouTube**
Amir Ghasemian, Homa Hosseinmardi, Upasana Dutta, Duncan Watts
Applied unsupervised clustering to group YouTube viewers based on consumption patterns, analyzing engagement with both traditional media organizations and content creators across predefined topics.

2024-present **Classifying online news using Weakly Supervised Text Classification**
Atieh Armin, Homa Hosseinmardi, Amir Ghasemian, Upasana Dutta, Duncan Watts
Using weakly supervised text classification to classify news articles from 10 major online publishers (New York Times, Wall Street Journal, etc) for examining thematic distinctions across publishers.

2024-present **Epistemic Audit of Researcher Bias in Meta-Analytic Studies**
Upasana Dutta Hanzhao Kuang, Barbara Ann Mellers, Cory Clark, Philip Tetlock, Eric Luis Uhlmann
Analyzing meta-analytic databases with LLMs to explore the influence of researchers' political inclinations (liberalism-conservatism) on study outcomes and methodological choices.

2023 **Sampling random graphs with specified degree sequences**
Upasana Dutta, Bailey K. Fosdick, Aaron Clauset [Preprint] [Code] [Python Package]
Forthcoming, **Journal of Computational and Graphical Statistics 2024**
Developed a method for sampling networks from the degree-preserving configuration model by detecting convergence in a double-edge swap Markov chain sampler.

2021 **Analyzing Twitter Users' Behaviour Before and After Contact by Russia's Internet Research Agency**
Upasana Dutta, Rhett Hanscom, Jason Zhang, Richard Han, Tamara Lehman, Qin Lv, Shivakant Mishra [Publication] [Code] [CU Boulder Today]
Published in the **Proceedings of the ACM on Human-Computer Interaction, CSCW 2021**
Employed statistical methods to analyze changes in user behavior on Twitter after they engaged with Russian bots backed by the Internet Research Agency (IRA) during the US 2016 presidential election.

Work Experience

May-August 2022 **Research Assistant, Clauset Lab, University of Colorado Boulder**
Analyzed the scaling trends of network statistics (average degree, clustering, mean geodesic distances) in a large corpus of real-world social, biological, technological, and informational networks, and assessed the extent to which null models (for eg. configuration model) explain the scaling behavior.

Technical Projects

Spring 2024 **Effects of Exposure to Ideologically-biased TV news** [Slides]
Developed causal model to detect if partisan TV exposure affects future Fox News/MSNBC viewing.

Fall 2023 **Bullets and Bytes: Evaluating the Performance of LLMs on the Gun Violence Database** [Report] [Presentation]
Evaluated LLMs (zero/few-shot T5, BERT, GPT-3, GPT-4) for gun violence reports data extraction.

Fall 2022 **Restuarant Recommender System** [Report]
Analyzed ML models (KNN, matrix factorization, neural networks) for restaurant recommendations.

Spring 2021 **Underrepresentation of rural undergraduate students in CU Boulder** [Slides] [Report] [Video]
Leveraged Data Analytics to evaluate underrepresentation of small/rural communities in CU Boulder.

Fall 2019 **Study of user activity on Question-Answering Platform : Stack Exchange** [Slides] [Report]
Analyzed user activity on Q/A platform for latent community structure with Stochastic Block Modeling.

Talks

Aug' 2024 **ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2024)** [Slides]

Aug' 2024 **Bernoulli-IMS 11th World Congress in Probability and Statistics 2024** [Slides]

July 2024 **International Conference on Computational Social Science (IC2S2 2024)** [Slides]

Nov' 2021 **The Mitchell Centre for Social Network Analysis, University of Manchester** [Slides] [Talk]

July 2021 **A Joint Sunbelt and NetSci Conference (Networks 2021)** [Abstract] [Slides] [Talk]

May 2021 **International Conference on Complex Networks (CompleNet 2021)** [Abstract] [Slides] [Talk]

Professional Activities

Fall 2024 Teaching Assistant for CIS 5300 Natural Language Processing at UPenn

July 2024 Gave a Networks Tutorial at the Summer Institute of Computational Social Science 2024 [Code]

Spring 2022 Google CS Research Mentorship Program 2022

Jan' 2021 Complex Networks Winter Workshop 2021, University of Vermont [Slides]

Awards

Apr' 2022 **Abel Lukens Stout Fellowship**, Department of Computer and Information Science, UPenn

Apr' 2022 **Bell Foundation Outstanding Research Award**, Department of Computer Science, CU Boulder

Mar' 2022 **CS Annual Research Expo Award**, Department of Computer Science, CU Boulder [Poster]

Dec' 2021 **NCWIT Collegiate Award 2022 Finalist** [Video]

Mar' 2021 **CS Annual Research Expo Award**, Department of Computer Science, CU Boulder [Poster]

Feb' 2021 **CS Publication Recognition Award**, Department of Computer Science, CU Boulder [Publication]

Skills

Computer Languages: Python, C, SQL, R

Operating Systems: MacOS, Linux, Windows environments

Toolkits: Machine Learning Frameworks (PyTorch/Tensorflow, Huggingface, etc), Statistical Tools (pandas, scikit-learn, etc), Visualization Tools (matplotlib, seaborn)

Data Science: Data Preprocessing, Statistical Inference, Anomaly Detection, Time-Series Analysis
Machine Learning: Regression Analysis, Supervised/Unsupervised Learning, Ensemble Methods, Deep Learning, Model Training and Fine-Tuning, Model Evaluation
Natural Language Processing: Text Preprocessing, LLMs, Topic Modeling, Prompt Engineering
Network Analysis: Centrality Measures, Null models, Community detection, Cascade models
Causal Inference: A/B testing, RCTs, Difference-in-Difference, Instrumental Variables