Computational Social Science
Research Design and Data Analytics [17-756]

Patrick Park

Spring, 2023

E-mail: patpark@cmu.edu
Office Hours: W 3:30-4:30pm
Office: TCS 324

Web: **https://patpark.org**
Class Hours: T/Th 5-6:20pm
Class Room: WEH 3203

Course Description

This course surveys how digital trace data of human activity online, combined with careful research design and data analytics, have led to surprising discoveries and development of novel theoretical explanations in the field of computational social science (CSS). The course has three aims. (a) It is intended to stimulate the capacity to formulate research questions on pressing/emerging societal phenomena of interest (e.g., political polarization, gender representation in open-source software development, decentralized governance based on blockchain technology), to construct explanations through theorizing, and to explore strategies to test the explanations with the new possibilities afforded by a wide array of digital trace data (e.g., social media, open-source software collaboration, cryptocurrency transactions, satellite imagery). This course will also (b) cover a range of methods employed in CSS that have led to novel empirical discoveries and/or theoretical breakthroughs. Specifically, we will evaluate the adequacy of analytic approaches (statistical modeling, network analysis, online experiments, simulations) for a given research question, and useful tools for big data processing and analytics (distributed data processing frameworks, visualization software). Throughout the course, we will (c) discuss practical considerations that commonly arise in a CSS research project cycle. These issues may include, but are not limited to, researcher biases in social science research, data collection (sampling methods and bias), repurposing of existing datasets that were collected in a different research context, constructing metrics based on theoretical insights while considering construct validity, ethical considerations, and data sharing. In the end, participants will develop "good taste" for theoretically insightful, methodologically creative, and carefully designed research and apply it towards initiating or developing their own research projects.

Course Learning Goals

1. Understand the central objective of explanation in CSS research.
2. Understand the basic logic and application of methods commonly employed in CSS research.
3. Demonstrate operating knowledge of commonly used tools for data collection, data processing, network analytics, and simulation.
4. Learn how to design a CSS research project proposal that develops a highly refined question and an explanation that can be evaluated with computational methods.
5. Learn to identify and address the commonly encountered issues in CSS research.

Reading Materials

Most reading materials are directly downloadable from online sources. Excerpts from books or other readings that are not readily available online will be posted in Canvas.

Prerequisites

There is minimal requirement for technical preparedness for the average graduate student. That being said, this course assumes working knowledge of statistical hypothesis testing (experiments), basic linear algebra (social network analysis), and coding experience. Furthermore, participants are strongly expected to be self-motivated grasping how the mathematic and quantitative operationalizations discussed throughout the course capture the social scientific insights behind them. As apparent from the readings in the course schedule, we will discuss the intersection of computational methods and social science research. As such, some of the foundational technical readings for a given methodology (e.g., Word2Vec) that may need to be read up on will be the responsibility of the participant.

Course Structure

Biweekly Seminars

For each biweekly seminar session, participants are expected to thoroughly read the assigned readings and submit a one-page reaction memo before the beginning of the seminar (Details on how to write the reaction memo can be found in Canvas). The seminar will consist of the instructor's overview and orientation for the selected readings and the participants' leading the discussion for each reading. Participants will sign up at this spreadsheet in advance for taking charge summarizing one of the assigned papers and leading the discussion (10 15 minutes each).

Final Project

Participants will use the readings and seminars towards writing their research papers. Towards this final deliverable, participants will draft a project proposal by the end of Spring Break (Week

8) and present their proposals at the beginning of Week 9. The final projects will be presented in Week 15. **Deadline for the final paper is May 8**.

Grading Policy

- **10%** assignments
- **15%** class participation (attendance, leading/participating in discussions)
- **25%** project proposal
- **50%** final project paper

Lab

This course is designed to focus on how computational methods can serve to address questions about human behavior within the context of socio-technical systems, by considering the entire research process in computational social science. Therefore, lab sessions that are designed to solve practical technical problems are not the main focus and will be limited to one week towards the end of the semester.

Course Schedule

The schedule is tentative and subject to change. The first half of the schedule begins with an overview of the field of computational social science and subsequently covers three widely used methodological approaches, social network analysis, online experiments, and agent-based models (Text analytic approaches, which also constitute a growing share of representation in CSS will be covered towards the end of the semester). In the second half of the course, attention will be given to practical issues that arise in research, from research ethics, computational tools, metrics constructions, ways to creatively appropriate and join data, to the limitations of big data. Week 14 will be left undecided for covering topics of interest to the seminar participants as we proceed.

Week 01: Introduction

1/17 Course Overview

- Logistics
- Introductions: Getting to know the group

1/19 What is computational social science?

- Lazer et al. 2009. Computational Social Science
- Golder and Macy Digital Footprints
- Evans and Foster Computation and the Sociological Imagination
- Salganik An Introduction to Computational Social Science

Week 02: The Realistic Cycle of Research

1/24 Observe, Describe, and Question

- Watts. *Everything is obvious once you know the answer* Ch. 1 and 2 (Canvas)
- Marwick and boyd. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience
- Golder and Macy. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures

1/26 Theorize, Explain, and Predict

- Swedberg. Before theory comes theorizing or how to make social science more interesting
- Hofman et al. Integrating explanation and prediction in computational social science
- Salganik et al. Measuring the predictability of life outcomes with a scientific mass collaboration

Week 03: Social Networks I

1/31 Graph Theory and the Network Paradigm

- Borgatti. Network Analysis in the Social Sciences
- Hidalgo. Disconnected, fragmented, or united? a trans-disciplinary review of network science
- Barabasi. *Network Science* Ch. 2, 3

2/2 Building Blocks: Social Relationships

- Krackhardt. The Ties That Torture: Simmelian Tie Analysis in Organizations
- Cook et al. The Distribution of Power in Exchange Networks: Theory and Experimental Results
- Backstrom and Kleinberg Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook

Week 04: Social Networks II

2/7 Describing and Explaining Social Structure

- Granovetter. The Strength of Weak Ties
- Park et al. The Strength of Long-Range Ties in Population-Scale Social Networks
- Jahani et al. Long ties, disruptive life events and economic prosperity

2/9 Network Explanations of Individual Outcomes

- Coleman. Social Capital in the Creation of Human Capital
- Burt. Structural Holes and Good Ideas
- Chetty et al. Social capital I: measurement and associations with economic mobility

Week 5: Social Contagion

2/14 Methodological Approaches

- Granovetter. Threshold Models of Collective Behavior
- Centola and Macy. Complex Contagions and the Weakness of Long Ties
- Christakis and Fowler. The Spread of Obesity in a Large Social Network over 32 Years

2/16 Confounding in Observational Data

- Shalizi et al. Homophily and Contagion Are Generically Confounded in Observational Social Network Studies
- Aral and Walker. Tie Strength, Embeddedness, and Social Influence: A Large-Scale Networked Experiment
- Ugander et al. An Experimental Study of Structural Diversity in Social Networks

Week 06: Online Experiments

2/21 Controlled Experiments

- Salganik et al. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market
- Centola et al. Experimental evidence for tipping points in social convention
- Tsvetkova and Macy. The Social Contagion of Antisocial Behavior

2/23 Field and Natural Experiments

- Bond et al. A 61-million-person experiment in social influence and political mobilization
- Kramer et al. Experimental evidence of massive-scale emotional contagion through social networks
- Phan and Airoldi A natural experiment of social network formation and dynamics

Week 07: Agent-Based Models

2/28 Toy Models for Theory Development

- Epstein. *Generative Social Science: Studies in Agent-Based Computational Modeling* Ch. 1 (Canvas)
- Schelling. *Micromotives and Macrobehavior* Ch. 1 (Canvas)
- Axelrod. The Dissemination of Culture: A Model with Local Convergence and Global Polarization
- Baldassarri and Bearman. Dynamics of Political Polarization

3/2 ABM and Counterfactuals

- Centola and Macy. The Emperor's Dilemma: A Computational Model of Self-Enforcing Norms
- Goldberg and Stein. Beyond Social Contagion: Associative Diffusion and the Emergence of Cultural Variation
- DellaPosta and Marjan. The Complexity of Associative Diffusion: Reassessing the Relationship between Network Structure and Cultural Variation

Week 08: Spring Break

3/7 No class

3/9 No class

Week 09: Research Ethics

3/14 Project Proposal Presentation

- Present idea and submit proposal

3/16 Research Ethics

- Salganik. Bit by Bit: Social Research in the Digital Age Ch. 6
- Hunter et al. Ethical Issues in Social Media Research for Public Health
- Krafft et al. Bots as Virtual Confederates: Design and Ethics
- Caliskan et al. Semantics Derived Automatically from Language Corpora Contain Human-like Biases

Week 10: Computational Tools

3/21 Data Collection and Processing Lab (tentative)

- API Basics with Twitter
- Distributed Data Processing with Spark

3/23 Data Visualization: A Primer (tentative)

- Visualization Principles
- Visualization Tools

Week 11: Metrics, Validation, and Generalizability

3/28 Measure Construction

- Lazer et al. Meaningful measures of human society in the twenty-first century
- Bonacich Power and Centrality: A Family of Measures
- Bothner et al. A Model of Robust Positions in Social Networks

3/30 Metric Validation and Generalizability

- Wu et al. Large teams develop and small teams disrupt science and technology
- Henrich et al. The Weirdest People in the World?
- Almaatouq et al. Beyond Playing 20 Questions with Nature: Integrative Experiment Design in the Social and Behavioral Sciences

Week 12: Big Data

4/4 Creative Use Cases

- Nagaraj. The Private Impact of Public Data: Landsat Satellite Maps Increased Gold Discoveries and Encouraged Entry

- Park et al. Global labor flow network reveals the hierarchical organization and dynamics of geo-industrial clusters
- Chi et al. Microestimates of wealth for all low- and middle-income countries
- Chen and Rohla. The effect of partisanship and political advertising on close family ties

4/6 Limitations and Constraints

- Lazer et al. The Parable of Google Flu: Traps in Big Data Analysis
- Lewis. Three fallacies of digital footprints
- Bail. Lost in a random forest: Using Big Data to study rare events
- King and Persily A New Model for Industry-Academic Partnerships

Week 13: NLP

4/11 Topic Models and Word Embeddings

- Corritore et al. Duality in Diversity: How Intrapersonal and Interpersonal Cultural Heterogeneity Relate to Firm Performance
- Waller and Anderson. Quantifying social organization and political polarization in online platforms
- Kozlowski et al. The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings
- DiMaggio. Adapting computational text analysis to social science (and vice versa)

4/13 No class (Spring Carnival)

Week 14: Future Topics in CSS

4/18

- TBD

4/20

- TBD

Week 15

4/25 Final Project Presentation

- Up to four projects

4/27 Final Project Presentation

- Up to four projects

Course Policies

Time management

This is a 12-unit course. It is intended that you spend close to 12 hours a week on the course, on average. In general, 3 hours/week will be spent in lecture and 9 hours on reading assignments, homework, and final project preparation.

Writing Skills

For assistance with the written or oral communication assignments in this class, visit the Global Communication Center (GCC). GCC tutors can provide instruction on a range of communication topics and can help you improve your papers, presentations, and job application documents. The GCC is a free service, open to all students, and located in Hunt Library. You can make tutoring appointments directly on the GCC website. You may also browse the GCC website to find out about communication workshops offered throughout the academic year.

Collaboration

Single author projects are allowed, but two-person collaboration is welcome. Collaborative pairs with significantly different (diverse) skillsets and knowledge backgrounds are strongly encouraged. Informal collaboration between collaboration groups is encouraged in terms of conceptual and methodological feedback/brainstorming. However, the final project paper should be strictly the responsibility of the authors.

Academic Integrity and Honesty

The University Policy on Academic Integrity applies. Expectations regarding academic honesty and collaboration for both group work are the same as for individual work, elevated to the level of "group." Within groups, we expect that you are honest about your contribution to the group's work. This implies not taking credit for others' work and not covering for team members that have not contributed to the team.

Accommodations for Disabilities

If you wish to request an accommodation due to a documented disability, please inform the instructor as soon as possible and contact Disability Resources at 412.268.2013 or access@andrew.cmu.edu.

Self Care

Please take care of yourself. Do your best to maintain a healthy lifestyle this semester by eating well, exercising, avoiding drugs and alcohol, getting enough sleep and taking some time to relax. This will help you achieve your goals and cope with stress. All of us benefit from support during times of struggle. You are not alone. There are many helpful resources available on campus and an important part of the college experience is learning how to ask for help. Asking for support sooner rather than later is often helpful. If you or anyone you know experiences any academic stress, difficult life events, or feelings like anxiety or depression, you are strongly encouraged to seek support. Counseling and Psychological Services (CaPS) is here to help: call 412-268-2922 and visit their website at https://www.cmu.edu/counseling. Consider reaching out to a friend, faculty or family member you trust for help getting connected to the support that can help.