Statistics for Social Data

Description

Statistical methods for describing and utilizing nontraditional data modalities arising from social processes. Half of the course will be devoted to network data (communications, recommendations, and transactions) and the other half will be devoted to textual data (words, phrases, and documents). We will survey a broad class of models for dealing with these types of data. Possible topics include the following: word-sense disambiguation; sentiment analysis; Markov and hidden Markov text models; topic discovery; descriptive statistics for networks; community detection; exponential random graph models; latent space models; network sampling; point processes; low-rank matrix approximations; power laws.

Objectives

The objectives of the course are the following:

- to learn about approaches for describing and utilizing network and text data;
- to understand these approaches from within an inferential statistical framework;
- to get hands-on experience in applying these methods, through the homework assignments and the class project.

Prerequisites

The course will assume some prior knowledge of linear algebra, probability, and statistical inference. You should be familiar with eigenvectors, likelihoods, and confidence intervals. Homework assignments will require a small amount of programming in Python and R; you can take the course if you know one of these programming languages and you are willing to learn the other.

Meeting Time and Place

Wednesdays, 2:00-5:00, in the Stat/OM conference room (KMC 8-170).

Course Staff

Instructor:	Patrick Perry
E-mail:	pperry@stern.nyu.edu
Office:	KMC 8-63
Office Hours:	Mondays, 2:00–4:00

Homework Assignments

There will be three homework assignments. Each assignment will involve applied data analysis and theoretical problem solving.

Course Project

For the course project, you will analyze a dataset of your choosing using methods from the class. You will complete the project in three stages: an initial proposal, a midterm progress update, and a final report and presentation.

Readings

TBD. Some mixture of textbook chapters and research articles.

Topics

Topics will be chosen from the following lists.

General

- Power laws
- Low rank matrix approximations

Text

- Word sense disambiguation
- Sentiment analysis
- N-gram models
- Hidden Markov models
- Topic models

Networks

- Descriptive statistics for networks
- Exponential random graph models
- Latent space models
- Community detection
- Agent-based models
- Point process models
- Network sampling

Tentative Schedule

Date	Topic	Assignment Due
9/4	Case Study: *The Federalist*	
9/11	Power Laws	
9/18	Matrix Decompositions	Project Proposal
9/25	N-gram Models	HW1
10/2	Topic Models	
10/9	Sentiment Analysis	
10/16	Exponential Random Graph Models	HW2
10/23	Latent Space Models	
10/30	Community Detection	
11/6	Agent-Based Models	Project Progress Report
11/13	Point Process Models	HW3
11/30	Respondent-Driven Sampling	
11/27	*Thanksgiving Break (No Class)*	
12/4	Projects	Project Presentation