Syllabus for CSCI 7000-001/4830-001
Inference, Models and Simulation for Complex Systems
Fall 2011

Lectures: Tuesdays and Thursdays from 11:00am – 12:15pm in ECCR 131

Lecturer: Aaron Clauset

Office:	ECOT 743
Email:	aaron.clauset@colorado.edu
Web Page:	http://tuvalu.santafe.edu/~aaronc/courses/7000/
Office Hours:	Tuesday 1:30–3:00 or by appointment

Description: This graduate-level topics course will cover a selection of recent developments in computational approaches to doing science with complex systems. It is not a scientific computing course. Topics will include complex networks, statistical inference, probabilistic models, random walks, simulation techniques, computational social science and computational biology. The focus will be on using computational tools (algorithms) to do science (work with data; test hypotheses; build understanding; make predictions).

Prerequisites: CSCI 3104 (undergraduate algorithms), two semesters of calculus (integration and differentiation), and basic statistics or probability theory.

Recommended preparation: CSCI 5454 (graduate algorithms) and APPM 3050 (Scientific Computing in Matlab) and either APPM 3570 (Applied Probability) or APPM 4570 (Statistical Methods).

No biological or social science background is necessary. A quantitative background is mandatory. The concepts and techniques covered in this course depend heavily on algorithms, basic statistics, numerical and analytical programming (`Matlab` and `Mathematica`) and calculus. Students without adequate preparation will struggle to keep up with the lectures and assignments, each of which has a mathematical and a programming component. Students concerned about their preparation are encouraged to instead audit the course.

Required Text: (1) *Networks: An Introduction* by M.E.J. Newman and
(2) *The Sciences of the Artificial* (3rd ed.) by H.A. Simon

Coursework and grading: Attendance to the lectures is required.

The first half of the course will be lecture and problem-set driven. The second half of the class will revolve around crowd-source (student) lectures on advanced topics and a small independent research project. The lecture and project are the major deliverables for the class, and students are expected to spend considerable time outside of class preparing them. There are no written examinations in this class; the lecture and project should be viewed as an oral and a final examination. Many lectures will have an accompanying reading assignment from the required texts or a reading from the primary literature. I'll post the latter readings on the class website.

Problem sets (PS): The four problem sets will each have some mathematical and some programming problems. Programming problems can be done in any reasonable imperative language, but I highly recommend using `Matlab`, or something similar, which has data analysis and visualization capabilities built-in. Familiarity with `Mathematica` will be useful for many of the mathematical problems.

Problem sets will be due roughly two weeks after they are assigned (schedule given below). Solutions must be in PDF format (e.g., typeset using LaTeX), should include all necessary details for me to follow the logic, and must be submitted via email by 11:59pm the day they are due. No late assignments will be accepted. Collaboration is allowed on the problem sets, but you may not copy (in any way) from your collaborators and you must respect University academic policies at all times. To be clear: you may discuss the problems verbally, but you must write up your solutions separately. If you do discuss the problems with someone (and you are encouraged to!), you must then list and describe the extent of your collaboration in your solutions (a footnote is fine). Copying from any source, in any way, including the Web but especially from another student (past or present), is strictly forbidden. If you are unsure about whether something is permitted under these rules, ask me well before the deadline.

Crowd-source lecture (CSL): Students will self-organize into teams of two. Each team will then give a full-length (75 minute) technical lecture that

1. frames a scientific question and motivates an algorithm or model to address it,
2. describes the algorithm or model in appropriate technical detail (at least at the pseudocode level)
3. describes its performance or results, and

4. describes interesting extensions, open question, problems, etc.

Students should notify me via email of their team composition by September 13. Each team should then notify me via email by September 15 of their preferences on lecture topics in the form of a complete ranking of the topics listed below. I will match teams to topics using a matching algorithm. (The number of topics and the precise schedule may change slightly during the semester depending on the number of students in the class, etc.)

Independent project (IP): The assignment here is to develop and complete a small research-style project based around the tools and concepts developed in the course. This independent project should focus on either inference, modeling or simulation. Students will work independently, but are strongly encouraged to get feedback from each other on their ideas. The deliverables here are a short (20 minute) in-class presentation of their results and a 10-page writeup due via email to me by 11:00am on December 15.

A short project proposal is due via email on October 13. The proposal should be two paragraphs: one explaining any background material, including necessary references to the scientific literature, and a second describing what you're going to do. When choosing a topic, students should be cognizant of the need to make reasonable progress in the 9 weeks remaining in the semester. To help facilitate this, students are strongly encouraged to meet with me outside of class prior to October 13 to discuss potential project topics.

Grading: Grades will be assigned based on a 20% attendance, 30% problem sets, 20% lecture and 30% project division.

Undergraduates taking CSCI 4830: Instead of doing both the lecture and the project, undergraduate students may choose to do one or the other. Grades will be assigned based on a 30% attendance, 30% problem sets, and 40% lecture/project division.

Crowd-sourced lecture topics:

1. *The Bootstrap*: Invented by Bradley Efron in the 1970s, the bootstrap is a strikingly simple and yet powerful statistical and computational technique that squeezes more information out of scarce empirical data. It can be used to clarify parameter estimates, create confidence bounds, and better capture the underlying variance in messy empirical data. How does it work? How do we know that it works?

2. *Duplication-mutation network growth models*: The most popular model of gene network evolution is to allow vertices to duplicate themselves and their connections, and to then

rewire some of those connections. What evidence supports this conclusion? What role is left for node deletion, and what impact does it have on network structure or function?

3. *Privacy and algorithms for de-anonymizing human data*: Privacy in the Internet age is complicated. It turns out that it's possible to probabilistically guess a person's Social Security Numbers using only a few pieces of publicly available information, guess an anonymous rater's true identity using Netflix and IMBD data, guess the hidden sexual orientation of a person on Facebook based on the attributes of their friends, etc. What else are we revealing about ourselves simply by leaving a digital trail? What does this say about the future of privacy?

4. *Random walks, Levy flights and models of human or animal mobility*: Some data collected on the movements of animals (deer, albatross, marine predators and some others) and humans suggests that the distribution of step sizes is very heavy-tailed. The most popular model of such behavior is a random walk called a Lévy flight, which assumes power laws in the step-size distribution. Does the data support these conclusions? What are the implications of these for mixing processes, e.g., the spread of epidemics, finding food in a complex landscape, etc.?

5. *Regression*: The most common statistical model in science is the general linear model, better known as "regression." Regression is an extremely flexible way to detect significant correlations between multiple "independent" variables and for extracting a simplified description of covariation. Regression is also something of a black art. When does it work well? What kind of errors does it make, and when does it lead us astray? What problems does non-parametric regression circumvent?

6. *Models of metabolic evolution*: Metabolic networks are the collection of all metabolism-related enzyme-facilitated reactions within a cell. These networks allow a cell to transform inputs into energy and waste products, and research suggests they have a highly conserved core surrounded by flexible periphery composed of somewhat disposable pathways. A recently proposed model argues that prokaryotic networks are like an adaptive "toolbox," in which cells repurpose existing tools and borrow other tools from other species when confronted with a novel environment. Is this a reasonable model? Does the data support it? What does it imply for macro-evolutionary patterns?

7. *Gaussian mixture models and choosing how many clusters*: Spatial data is ubiquitous in science, as is the interest in identifying clusters within such data. Gaussian mixture models are a powerful technique for probabilistically identifying and modeling such inhomogeneous distributions, but to be useful, we must also choose only as many clusters k as the data warrants. How can we do this? When do these techniques fail?

8. *Pedestrian flows and flocking*: People are not particles. But sometimes humans and animals exhibit collective movements strikingly similar to what we'd expect if so. From pedestrians walking through urban environments to birds or fish flocking in groups, models from physics are surprisingly good at describing these collective behaviors. How accurate are such "physics" models of behavior? What does this say about free will?

Tentative schedule:

Week 1	Overview, probability distributions, maximum likelihood, Poisson processes
Week 2	Power-law distributions
Week 3	Hypothesis tests, model plausibility, model comparison
Week 4	Time series analysis, random walks
Week 5	Random walk models of macroevolution
Week 6	Probabilistic models of terrorism and wars
Week 7	Introduction to complex networks
Week 8	Random graph models, small worlds, citation networks
Week 9	Modular and hierarchical structure, missing information, link prediction
Week 10	CSL: (1) The bootstrap and (2) duplication-mutation networks
Week 11	CSL: (3) Privacy and de-anonymizing human data and (4) Lévy flights
Week 12	CSL: (5) Regression and (6) metabolic networks
Week 13	CSL: (7) Gaussian mixture models and (8) people are not? particles
Week 14	Fall break
Week 15	Project presentations
Week 16	Project presentations

Assignments:

	assigned	due
PS 1 (the longest PS)	August 23	September 13
CSL, form team of 2		September 13
CSL, topic preferences		September 15
PS 2	September 14	September 27
PS 3	September 28	October 11
IP, topic proposal		October 13
PS 4	October 12	October 25
CSL, presentation		to be determined
IP, short presentation		to be determined
IP, final write up		December 15

Advice for writing up your solutions:
Your solutions for the problem sets should have the following properties. I will be looking for these when I grade them:

1. **Clarity**: Your solutions should be both clear and concise. The longer it takes me to figure out what you're trying to say, the less likely you are to receive full credit. There is no grader for this class but me and my time is short; thus, the more clear you make your thought process, the more likely you are to get full credit.

2. **Completeness**: Full credit is based on (i) sufficient intermediate work and (ii) the final answer. For many problems, there are multiple paths to the correct solution, and I need to understand exactly how you arrived at the solution. A heuristic for deciding how much detail is sufficient: if you were to present your solution to the class and everyone understood the steps and could repeat them themselves, then you can assume it is sufficient.

3. **Succinctness**: Solutions should be long enough to convince me that your answer is correct, but no longer. More than half a page of dense algebra, more than a few figures or more than a page or two per problem is probably not succinct. Clearly indicate your final answer (circle, box, underline, whatever). Rewriting your solutions, with an eye toward succinctness, before submitting will help. Strive for maximum understanding in minimum space.

4. **Numerical experiments**: Many of the programming problems will require you to conduct numerical experiments using stochastic processes. One run is not an experiment. Your goal is to produce beautifully smooth central tendencies and you should average over as many independent trials as is necessary to get it. Further, your results should span several orders of magnitude. I recommend a dozen or so measurement values across the x-axis, distributed logarithmically, e.g., $n = \{2^4, 2^5, 2^6, \ldots\}$.

   No credit will be given if you fail to label your axes and data series, or if you fail to explain your experimental design.

5. **Source code**: Your source code for all programming problems must be included at the *end* of your solutions. Code must include copious comments explaining the sub-algorithms and must be run-able; that is, if I try to compile and run it, it should work as advertised.

Suggestions: Suggestions for improvement are welcome at any time. Any concern about the course should be brought first to my attention. Further recourse is available through the office of the Department Chair or the Graduate Program Advisor, both accessible on the 7th floor of the Engineering Center Office Tower.

Honor Code: As members of the CU academic community, we are all bound by the CU Honor Code. I take the Honor Code very seriously, and I expect that you will, too. Any significant violation will result in a failing grade for the course and will be reported. Here is the University's statement about the matter:

All students of the University of Colorado at Boulder are responsible for knowing and adhering to the academic integrity policy of this institution. Violations of this policy may include: cheating, plagiarism, aid of academic dishonesty, fabrication, lying, bribery, and threatening behavior. All incidents of academic misconduct shall be reported to the Honor Code Council (honor@colorado.edu; 303-735-2273). Students who are found to be in violation of the academic integrity policy will be subject to both academic sanctions from the faculty member and non-academic sanctions (including but not limited to university probation, suspension, or expulsion). Other information on the Honor Code can be found at http://www.colorado.edu/policies/honor.html and at
http://www.colorado.edu/academics/honorcode/

Special Accommodations: If you qualify for accommodations because of a disability, please submit to me a letter from Disability Services in a timely manner (first 3 weeks of class) so that your needs can be addressed. Disability Services determines accommodations based on documented disabilities. Contact: 303-492-8671, Center for Community N200, and http://www.colorado.edu/disabilityservices

If you have a temporary medical condition or injury, see guidelines at
http://www.colorado.edu/disabilityservices/go.cgi?select=temporary.html

Campus policy regarding religious observances requires that faculty make every effort to deal reasonably and fairly with all students who, because of religious obligations, have conflicts with scheduled exams, assignments or required attendance. In this class, I will make reasonable efforts to accommodate such needs if you notify me of their specific nature by the end of the 3rd week of class. See full details at
http://www.colorado.edu/policies/fac_relig.html

Classroom Behavior: Students and faculty each have responsibility for maintaining an appropriate learning environment. Those who fail to adhere to such behavioral standards may be subject to discipline. Professional courtesy and sensitivity are especially important with respect to individuals and topics dealing with differences of race, color, culture, religion, creed, politics, veteran's status, sexual orientation, gender, gender identity and gender expression, age, disability, and nationalities. Class rosters are provided to the instructor with the student's legal name. I will gladly honor your request to address you by an alternate name or gender pronoun. Please advise me of this preference early in the semester (by end of 3rd week) so that I may make appropriate changes to my records. See policies at `http://www.colorado.edu/policies/classbehavior.html` and at `http://www.colorado.edu/studentaffairs/judicialaffairs/code.html#student_code`

Discrimination and Harrassment: The University of Colorado at Boulder Discrimination and Harassment Policy and Procedures, the University of Colorado Sexual Harassment Policy and Procedures, and the University of Colorado Conflict of Interest in Cases of Amorous Relationships policy apply to all students, staff, and faculty. Any student, staff, or faculty member who believes s/he has been the subject of sexual harassment or discrimination or harassment based upon race, color, national origin, sex, age, disability, creed, religion, sexual orientation, or veteran status should contact the Office of Discrimination and Harassment (ODH) at 303-492-2127 or the Office of Student Conduct (OSC) at 303-492-5550. Information about the ODH, the above referenced policies, and the campus resources available to assist individuals regarding discrimination or harassment can be obtained at `http://www.colorado.edu/odh`