# STA 4513H F

# Statistical Models of Networks, Graphs, and Other Relational Structures

## Instructor

Prof Daniel M. Roy
daniel.roy@utoronto.ca (please include "STA4513" in your email's subject line or body)
Office hours: SS 6026, Tuesdays 9--11am, or by appointment.

## First Class

Because the class was approved rather late, students who miss the first class but still want to take the course are encouraged to come to the second class. Please contact me so that you can receive notes for the first class.

## Drop Date

For short courses, like this one, the drop date is the date of the second class.

## Time and Location

Date/Time: Fall 2014, Wednesdays, 2--5pm from Sept. 10 through Oct. 15
Room: WE 69 (Wetmore Hall, New College, across Huron St from Sidney Smith)

## Overview

This course is a survey of topics on the statistical analysis of network, graph, and relational data, with an emphasis on network modeling and random graphs. In particular, the role of probabilistic symmetries---including exchangeability and stationarity---will be our primary concern, at least in the second half.

Our understanding of graph- and network-valued data has undergone a dramatic shift in the past decade. We now understand there to be fundamentally different regimes that relate to the prevalence of edges. The best understood is the dense regime, where, informally speaking, we expect to see edges among vertices chosen uniformly at random from a large graph. The mathematical foundations of this area can be traced back to work by Aldous and Hoover in the early 1980s, but work in graph theory over the past decade has enriched our understanding considerably. Most existing statistical methods, especially Bayesian ones, work implicitly in the dense regime. Real-world networks, however, are not dense. A growing community is now focused on the structure of large sparse graphs. The sparse regime, however, is not well understood: key mathematical notions continue to be identified. We will work through key papers in probability, statistics, and graph theory in order to gain the broader perspective necessary to identify opportunities to contribute to our understanding of statistical methods on graphs and networks.

## Structure of the course

The course will be structured around weekly readings and student presentations of papers. Each week there will be a lecture on a new subject. Student presentations on each topic will follow the subsequent week.

Each week there will be a *lecture on a topic* and a **student presentation** on the previous week's topic.

- Week 1
  - Student background survey
  - High level overview of course
  - *Latent Variable Models*
  - Initial paper assignments
- Week 2
  - *Exponential Random Graph Models*
  - **Student Presentations on Latent Variable Models**
- Week 3
  - *Exchangeability of Sequences and Graphs*
  - **Student Presentations on Exponential Random Graph Models**
- Week 4
  - *Graphons*
  - **Student Presentations on Exchangeability of Sequences and Graphs**
- Week 5
  - *Sparse Graphs*
  - **Student Presentations on Graphons and Exchangeable Random Graphs**
- Week 6
  - *Unimodular Random Networks*
  - **Student Presentations on Sparse Graphs**

See *Reading* (below) for papers associated with each week's topic.

## Grading

The grade in the course will be based on:

- 25% class presentation
- 25% second class presentation *or* research project;
- 50% class participation, including scribing*.

Each student will be expected to make at least one class presentations. The second presentation can be replaced by a research project. Those auditing will be expected to make at least one class presentation, and potentially scribe if there are too few registered students.

**Scribing* requires that the student take detailed notes during lecture, and produce a LaTeX version of these notes. The notes should be free from typos, and should allow a reader to follow the logic of the presentation. These will be distributed to the students. A LaTeX template for scribing will be made available on the course website and should be used.

### Class participation

Students should remain attentive, ask questions, and answer questions during class. Participation also involves reading over those papers being presented that week, and coming with questions. Scribing is a clear way to participate.

### Presentations

Students are expected to make paper presentations, which can be tackled in groups of up to 2 students collaboratively. (Groups of 3 should receive my prior approval.) A list of scheduled presentations will be available on the course website. Papers other than those that have been marked with yellow astericks (*) require prior approval. Papers should be chosen from the appropriate topic for the week, but exceptions will be considered if proposed more than a week in advance. Students are also encouraged to find papers in application areas of interest to them.

#### *Presentation structure and notes*

Presentations can either be slide presentations or chalk talks. They should be planned to take 45 minutes, with 15 minutes for questions. (Rehearse your presentations to check for timing. A rough guide is no more than one slide per minute.) In the case of a slide presentation, PDF slides will be placed on the website after the class. In the case of a chalk-talk, the presenter is responsible for producing notes for the talk. These can be the hand-written notes if they are clearly written and suitable for publication on the website, otherwise they should be LaTeX'd. Notes should be submitted by Friday of the same week of the presentation.

### Research Projects

Research projects will be simulations and/or theory on a model that I will choose. Unless I give approval otherwise, research projects should be completed individually. Students should decide whether they will pursue a research project before the third class.

### Policy on Late

Research projects are due by email on the last day of class (11:59pm Toronto time, Oct 15). Every day of delay thereafter results in a 10% deduction. Presentations cancelled later than Monday noon will be counted as missed. Missed presentations can be made up if there are slots available for 75%. Extenuating circumstances will be handled on a case by case basis.

## Pre-requisites

Mathematical maturity and some background in linear algebra, analysis, measure theory, and probability theory recommended.

## Course Webpage

Blackboard will be used to manage the course list and grades, but the course information and links will be available at http://danroy.org/teaching/2014/STA4513/

## Accessibility

Students with diverse learning styles and needs are welcome in this course. Please feel free to approach me or Accessibility Services so we can assist you in achieving academic success in this course. If you have not registered with the Accessibility Services and have a disability, please visit the Accessibility Services website at http://www.accessibility.utoronto.ca for information on how to register.

## Reading

There is no required course text. Students should read those papers being presented in each week. Papers marked with * are approved for presenting. Students may suggest other articles (listed here, or otherwise), but my prior approval is required.

### Surveys/books on Statistical Network Analysis (no presentations)

- *Statistical Models for Social Networks*
  Tom A.B. Snijders
  http://www.stats.ox.ac.uk/~snijders/StatModelsSocNetworks.pdf
- *Statistical Analysis of Network Data* **[book]**
  Eric Kolaczyk
  http://www.springer.com/computer/communication+networks/book/978-0-387-88145-4

### Latent Variable Models

- *Learning systems of concepts with an infinite relational model*
  Charles Kemp, Josh Tenenbaum, Tom Griffiths, Takeshi Yamada, Naonori Ueda
  http://www.psy.cmu.edu/~ckemp/papers/KempTGYU06.pdf
- * Modeling homophily and stochastic equivalence in symmetric relational data
  Peter Hoff
  http://papers.nips.cc/paper/3294-modeling-homophily-and-stochastic-equivalence-in-symmetric-relational-data.pdf
- * *Learning Annotated Hierarchies from Relational Data*
  Daniel M. Roy, Charles Kemp, Vikash Mansinghka, Joshua B. Tenenbaum
  http://danroy.org/papers/RoyKemManTen-NIPS-2007.pdf
- *The Mondrian Process*
  Daniel Roy, Yee Whye Teh

http://danroy.org/papers/RoyTeh-NIPS-2009.pdf
(Could be presented alongside paper on Annotated Hierarchies.)

- * *Mixed Membership Stochastic Blockmodels*
  Edoardo Airoldi, David Blei, Stephen Fienberg, Eric Xing
  http://jmlr.org/papers/volume9/airoldi08a/airoldi08a.pdf
- * *Inference and Phase Transitions in the Detection of Modules in Sparse Networks*
  Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborova
  http://journals.aps.org/prl/abstract/10.1103/PhysRevLett.107.065701
- * *Matrix estimation by Universal Singular Value Thresholding*
  Sourav Chatterjee
  http://arxiv.org/abs/1212.1247

### Exponential Random Graph Models

- *Statistical Models for Social Networks*
  Tom A.B. Snijders
  http://www.stats.ox.ac.uk/~snijders/StatModelsSocNetworks.pdf
  (Contains a number of references to other articles on exponential random graphs, including pathologies of working with this family.)
- * *New specifications for exponential random graph models*
  Tom A. B. Snijders, Philippa E. Pattison, Garry L. Robins, Mark S. Handcock
  http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9531.2006.00176.x/abstract
- * *Estimating and understanding exponential random graph models*
  Sourav Chatterjee, Persi Diaconis
  http://arxiv.org/abs/1102.2650

### Exchangeability of Sequences and Graphs

- * *Bayesian Models of Graphs, Arrays, and other Exchangeable Random Structures*
  Peter Orbanz, Daniel M. Roy
  http://danroy.org/papers/OR-exchangeable.pdf

### Graphons

- *Graph limits and exchangeable random graphs*
  Persi Diaconis and Svante Janson
  http://arxiv.org/abs/0712.2749
- *Large Networks and Graph Limits* **[book]**
  Laszlo Lovasz
  http://www.ams.org/bookstore-getitem/item=COLL-60
  (With approval, students could present several sections or a chapter.)
- * Multivariate sampling and the estimation problem for exchangeable arrays
  Olav Kallenberg
  http://link.springer.com/article/10.1023/A:1021692202530

- * *Nonparametric graphon estimation*
  Patrick J. Wolfe, Sofia C. Olhede
  http://arxiv.org/abs/1309.5936
- * Co-clustering separately exchangeable network data
  David Choi and Patrick J. Wolfe
  http://arxiv.org/abs/1212.4093
- * A Consistent Histogram Estimator for Exchangeable Graph Models
  Stanley H. Chan, Edoardo M. Airoldi
  http://arxiv.org/abs/1402.1888

### Sparse Graphs

- * *Bayesian nonparametric models of sparse and exchangeable random graphs*
  Francois Caron, Emily Fox
  http://arxiv.org/abs/1401.1137
- * *Pseudo-likelihood methods for community detection in large sparse networks*
  Arash A. Amini, Aiyou Chen, Peter J. Bickel, Elizaveta Levina
  http://projecteuclid.org/euclid.aos/1382547514
- * *Asymptotic behavior and distributional limits of preferential attachment graphs*
  Noam Berger, Christian Borgs, Jennifer T. Chayes, Amin Saberi
  http://arxiv.org/pdf/1401.2792.pdf
- * *An Lp theory of sparse graph convergence I: limits, sparse random graph models, and power law distributions*
  Christian Borgs, Jennifer T. Chayes, Henry Cohn, Yufei Zhao
  http://arxiv.org/abs/1401.2906
  (See also part II, http://arxiv.org/abs/1408.0744)

### Unimodular Random Networks

- *Processes on Unimodular Random Networks*
  David Aldous, Russell Lyons
  http://arxiv.org/pdf/math/0603062.pdf

## Additional reading

### Other surveys on Statistical Network Analysis

- *Statistical Analysis of Network Data* **[book]**
  Eric Kolaczyk
  http://www.springer.com/computer/communication+networks/book/978-0-387-88145-4
- *A survey of statistical network models*
  Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, Edoardo M Airoldi
  http://arxiv.org/abs/0912.5410/
- *Statistical Network Analysis: Models, Issues, and New Directions*
  Edited volume: Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, Anna Goldenberg, Eric P. Xing, Alice X. Zheng

http://www.stat.cmu.edu/~fienberg/Stat36-835/LNCS4503-NetworkVolume-3-07(penultimateversion).pdf

- *Statistical Modeling of Social Networks*
  Mark S. Handcock
  http://www.samsi.info/sites/default/files/handcock_august2012.pdf

## Books and reviews on random graphs, networks, etc.

- *Complex Graphs and Networks*
  Fan Chung, Linyuan Lu
  http://www.ams.org/bookstore-getitem/?item=CBMS-107
  (See http://www.math.ucsd.edu/~fan/complex/ for the first 3 chapters.)
- *Random Graph Dynamics*
  Rick Durrett
  http://dx.doi.org/10.1017/CBO9780511546594
- *The structure and function of complex networks*
  Mark Newman
  http://arxiv.org/pdf/cond-mat/0303516.pdf

## Exchangeability

- *Probabilistic Symmetries and Invariance Principles* **[book]**
  Olav Kallenberg
  http://www.springer.com/statistics/statistical+theory+and+methods/book/978-0-387-25115-8
- *Exchangeability and Related Topics*
  David Aldous
  http://www.stat.berkeley.edu/~aldous/Papers/me22.pdf
  http://link.springer.com/chapter/10.1007%2FBFb0099421

## Models of Random Walks

- *Bayesian nonparametric analysis of reversible Markov chains*
  Sergio Bacallado, Stefano Favaro, Lorenzo Trippa
  http://projecteuclid.org/euclid.aos/1369836963

## Asymptotic theory

- *A nonparametric view of network models and Newman-Girvan and other modularities*
  Peter Bickel, Aiyou Chen
  http://www.pnas.org/content/106/50/21068
- *Co-clustering separately exchangeable network data*
  David Choi and Patrick J. Wolfe
  http://arxiv.org/abs/1212.4093

### Multivariate point processes

- *Point process modeling for directed interaction networks*
  Patrick O. Perry, Patrick J. Wolfe
  http://arxiv.org/abs/1011.1703

### Logistic regression

- *Degree-based network models*
  Sofia C. Olhede, Patrick J. Wolfe
  http://arxiv.org/abs/1211.6537
- *Null models for network data*
  Patrick O. Perry, Patrick J. Wolfe
  http://arxiv.org/abs/1201.5871

### Network Sampling

- *Sampling and estimation in large social networks*
  Ove Frank
  http://www.sciencedirect.com.myaccess.library.utoronto.ca/science/article/pii/0378873378900151
- *Statistical properties of sampled networks*
  Sang Hoon Lee, Pan-Jun Kim, Hawoong Jeong
  http://journals.aps.org/pre/abstract/10.1103/PhysRevE.73.016102
- *Estimating the Size of Hidden Populations Using the Generalized Network Scale-Up Estimator*
  Dennis M. Feehan, Matthew J. Salganik
  http://arxiv.org/abs/1404.4009
- *Counting hard-to-count populations: the network scale-up method for public health*
  H Russell Bernard, Tim Hallett, Alexandrina Iovita, Eugene C Johnsen, Rob Lyerla, Christopher McCarty, Mary Mahy, Matthew J Salganik, Tetiana Saliuk, Otilia Scutelniciuc, Gene A Shelley, Petchsri Sirinirund, Sharon Weir, Donna F Stroup
  http://sti.bmj.com/content/86/Suppl_2/ii11.full.pdf
- *What is the real size of a sampled network? The case of the Internet*
  Fabien Viger, Alain Barrat, Luca Dall'Asta, Cun-Hui Zhang, Eric D. Kolaczyk
  http://math.bu.edu/people/kolaczyk/pubs/pre75.pdf

### Statistical Issues

These readings, cited by Snijders, address statistical problems dealing with network data.

- *Formal statistics and informal data analysis, or why laziness should be discouraged*
  I. W. Molenaar
  http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9574.1988.tb01221.x/abstract

- *Alternative Model-Based and Design-Based Frameworks for Inference From Samples to Populations: From Polarization to Integration*
  Sonya K. Sterba
  http://www.vanderbilt.edu/peabody/sterba/pubs/Sterba_2009.pdf
- *On the Validity of Inferences from Non-random Samples*
  T. M. F. Smith
  http://www.jstor.org.myaccess.library.utoronto.ca/stable/2981454

## Other Courses, Tutorials, Workshops

There are several courses available on the statistical *analysis* of networks. They tend not to emphasize the connections with random graphs or exchangeability, hence my emphasis on *models*. Nevertheless, for doing work in this area, it is useful to know the perspective of these courses.

- Peter Hoff's SNA course 567. (Descends from Mark Handcock's course.)
  http://www.stat.washington.edu/hoff/courses/Old567/stat567/
  http://www.stat.washington.edu/~hoff/courses/567/
- Eric Kolaczyk's SNA tutorial at SAMSI
  http://www.samsi.info/sites/default/files/Kolaczyk-CN.pdf
- SAMSI Program on Complex Networks (2010)
  Videos and lectures slides on many interesting topics.
  http://www.samsi.info/workshop/2010-11-program-complex-networks-opening-tutorials-workshop
- David Aldous's course on Random Graphs and Complex Networks
  http://www.stat.berkeley.edu/~aldous/Networks/