IC  📖 **jacobeisenstein** / **gt-css-class**

★ Star  5 | ⑂ Fork  0

⑂ branch: master ▾  **gt-css-class** / **README.md** 📋

👤 **jacobeisenstein** just now changed assignment details

**1 contributor**

📄 file | 195 lines (158 sloc) | 19.701 kb

Edit | Raw | Blame | History | Delete

# Georgia Tech CS 8803-CSS: Computational Social Science
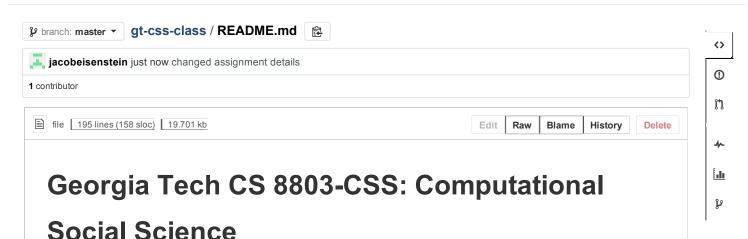
The principle aim for this graduate seminar is to develop a broad understanding of the emerging cross-disciplinary field of Computational Social Science. This includes:

- Methodological foundations in network and content analysis: understanding the mathematical basis for these methods, as well as their practical application to real data.
- Best practices and limitations of observational studies.
- Applications to political science, sociolinguistics, sociology, psychology, economics, and public health.

## Assignments and Grading

This is a seminar-style class, and will emphasize classroom discussion. For this reason, it is essential that students do the readings in advance of the lecture.

- Classroom participation: 10%
- Weekly blogposts about the reading, where you should raise relevant issues for classroom discussion: 30%
- Two research blogposts, in which students apply techniques from the course to shared datasets: 30%
- One indepedendent project, in which students apply techniques from the course to a dataset of their choice: 30%

The weekly blogposts should demonstrate your understanding of the assigned reading, and raise questions for classroom discussion.

The shared-data research blogposts should use the techniques in the class to explore a shared dataset, and attempt to answer one or two arguable questions about the data. (Excellent) examples of the style of work that I'm looking for can be found here, here, and here. The shared-data projects must be performed independently, and will require both original work as well as a small number of compulsory analyses that cover key concepts from the course.

The independent project should be substantive, original work in the area of computational social science. This can include: a new study using techniques described in the course; a refinement of the techniques described in the course; a novel survey paper that provides a unified treatment of an area of computational social science. This should represent roughly the same amount of work as the two shared-data blogposts, and can be done in teams of up to three students.

Students may audit the course, but all students who attend must perform the weekly blogposts about the reading, to facilitate discussion.

# Course schedule

E&K refers to the textbook Networks, Crowds, and Markets by Easley and Kleinberg. Free PDFs of each chapter are available by following the link.

**Week 1: Foundations**

- January 7. Welcome. Course aims, expectations, and ground rules. Reading: Michael Scherer on the Obama campaign
- January 9. Foundations of CSS. Reading: Computational social science by Lazer et al, 2009; Six provocations for Big Data by boyd and Crawford, 2011. Reminder: post your response to the class Tumblr by 1pm on Thursday.

## Part 1: Networks

**Week 2: Graphs**

- January 13. Reading: E&K ch. 1 and 2.
- January 15. Reading: Inferring friendship network structure using mobile phone data by Eagle, Pentland, and Lazer, 2009; Reply by adams; Rejoinder by Eagle et al. Optional but recommended light reading: Using metadata to find Paul Revere

**Week 3: Strong and weak ties**

- January 20. Reading: E&K ch. 3.
- January 22. Reading: The Role of Social Networks in Information Diffusion by Bakshy et al, 2012; Structure and tie strength in mobile communication networks by Onnela et al, 2007.

**Week 4: Networks in their surrounding contexts**

- January 27: Reading: E&K ch 4.
- January 29. Reading: Classifying political leanings by Zhou, Resnick, and Mei, 2011. Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity by Backstrom, Sun, and Marlow, 2010.

**Week 5: Structural balance**

- February 4: Reading: E&K ch 5.
- February 6: Reading: Signed networks in social media and Predicting positive and negative links in online social networks by Leskovec, Huttenlocher, and Kleinberg, 2010. Project 1 out.

## Part 2: Content

**Week 6: Text classification and regression**

- February 11. Reading: Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts by Grimmer and Stewart, 2013.
- February 13. Reading: A few useful things to know about machine learning by Domingos, 2012. A computational approach to politeness by Cristian Danescu-Niculescu-Mizil et al, 2013.
- February 14. Deadline to register for PoliInformatics Unshared Task and get free data, with great final project potential.

**Weeks 7 and 8: Topic models**

- February 18. Reading: Probabilistic topic models by David Blei, 2012. Optional helpful tutorial on topic models from a Digital Humanities perspective.
- February 20. Reading: The Issue Adjusted Ideal Point Model by Gerrish and Blei, 2012. You can also watch the video.
- February 21. Project 1 due at 5pm.
- February 25. No class. Watch Probabilistic topic models by Blei and The impression of influence by Grimmer. Reading: You Are What You Tweet: Analyzing Twitter for Public Healthby Paul and Dredze, 2011. You can optionally watch a whole course on Bayesian Generative Models (which generalize topic models) here.

- February 27. Topics and networks. Reading: Hierarchical relational models for document networks by Chang and Blei, 2010. (Warning: mathy! It's okay if you can't follow all of section 3, but do make sure you read and understand the rest of the paper.)
- February 28. **Drop deadline**. Project 2 out.

**Week 9: Memes, text reuse, and censorship**

- March 4. Reading: Meme-tracking and the dynamics of the news cycle by Leskovec, Backstrom, and Kleinberg, 2009.
- March 6. Reading: Censorship and deletion practices in Chinese social media by Bamman, O'Connor, and Smith, 2012. Tracing the flow of policy ideas in legislatures by Wilkerson, Smith, and Stramp, 2013.

## Part 3: Methods

**Week 10: Time series and elections**

- March 11. Reading: Predicting the present by Choi and Varian (video). From tweets to polls by O'Connor et al., 2010.
- March 13. Reading: More tweets, more votes by Joseph DiGrazia et al, 2013. How (not) to predict elections by Metaxas et al, 2011.
- March 14. Project 2 due at 5pm.

**Weeks 11 and 12: Observational and experimental studies**

- March 25. Reading: Linguistic Diversity and Traffic Accidents. Roberts and Winters, 2013. Blog post
- March 27. Reading: Homophily and Contagion Are Generically Confounded in Observational Social Network Studies by Shalizi and Thomas, 2011; Distinguishing contagion and homophily by Aral and Muchnik, 2012.
- April 1. Final project proposal presentations and discussion.
- April 3. Reading: E&K ch 20-20.6 (20.7 is optional); A 61-million-person experiment by Marlowe et al, 2013.
- April 4. Final project proposals due at 5pm.

## Part 4: Applications

- April 8. **Economics**. Reading: Twitter mood predicts the stock market by Bollen, Mao, and Zeng, 2011... or does it?. The blogpost mentions the Bonferroni correction for multiple comparison, which is best motivated by this XKCD cartoon.
- April 10. **Psychology**. Reading: Predicting Postpartum Changes in Emotion and Behavior via Social Media by De Choudhury, Counts, and Horvitz (2013). The psychological meaning of words: LIWC and computerized text analysis methods by Tausczik and Pennebaker (2010).
- April 15. **Interpersonal dynamics**. Reading: Extracting Social Meaning: Identifying Interactional Style in Spoken Conversation by Jurafsky et al, 2009. Shows use of prosodic features (tone of voice). Utterance-Level Multimodal Sentiment Analysis by Perez-Rosas, Mihalcea, and Morency, 2013.
- April 17. **Social media analysis**. Reading: Who says what to whom on Twitter by Wu et al, 2011. What to do about bad language on the internet by Eisenstein, 2013.
- April 22. **Predicting things about authors**. Reading: Discriminating gender on Twitter by Burger et al, 2011. Gender identity and lexical variation in social media by Bamman, Eisenstein, and Schnoebelen, 2014 (in review).
- April 24. Final project presentations.

## Additional readings

There are many, many more interesting papers than what we can cover in this class. Here are just a few.

**Overviews**

- Computational Text Analysis for Social Science: Model Assumptions and Complexity by O'Connor, Bamman, and Smith, 2012. Nice overview and bibliography.

**Discourse, dialogue, and pragmatics**

- Echoes of power: Language effects and power differences in social interaction by Danescu-Niculescu-Mizil, Lee, Pang,

Kleinberg, 2012.

- Estimating the prevalence of deception in online review communities by Ott, Cardie, and Hancock, 2012.
- Entrainment in spontaneous speech: The case of filled pauses in supreme court hearings by Benus, Levitan, and Hirschberg, 2012.
- Phrases that signal workplace hierarchy by Gilbert, 2012.
- Extracting social power relationships from natural language by Bramsen et al, 2013.
- Public dialogue: Analysis of tolerance in online discussions by Mukherjee et al, 2013.
- The pragmatics of expressive content: evidence from large corpora by Constant et al, 2009.
- Towards a model of formal and informal address in English by Faruqui and Pado, 2012.

**Politics**

- Fighting words by Monroe, Colaresi, and Quinn. Great paper on associating keywords with metadata. For a linguistic / pragmatic take on the same issue, see The pragmatics of expressive content by Constant et al, 2009.
- The political blogosphere and the 2004 US election: divided they blog. Adamic and Glance, 2004. Seminal work on social media analysis, but it's already covered in the Social Computing class.
- A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases by Grimmer, 2009. Interesting application of TMs to politics, but the Von Mises-Fisher stuff is too complicated to discuss in this class.
- Get out the vote: Determining support or opposition from Congressional floor-debate transcripts by Thomas, Pang, and Lee, 2006.
- A Penny for your Tweets: Campaign Contributions and Capitol Hill Microblogs by Yano, Yogatama, and Smith, 2013.
- Lexical and hierarchical topic regression by Nguyen et al, 2013.
- Learning to extract international relations from political context by O'Connor et al, 2013.
- Can machines learn to predict a violent conflict? by Chris Perry, 2013.

**Social media**

- What is Twitter, a social network or a news media? by Kwak et al, 2010. An early look at the structural properties of Twitter.
- Functions of the non-verbal in CMC: Emoticons and illocutionary force by Dresner and Herring, 2010
- Gender and genre variation in weblogs by Paolillo and Herring, 2006.
- Phonological factors in social media writing by Eisenstein, 2013.
- A Latent Variable Model of Geographical Lexical Variation by Eisenstein et al 2010. Good for understanding the data and the problem, but the model in Sparse Additive Generative Models of Text (Eisenstein et al, 2011) is simpler and better.
- Simple supervised document geolocation with geodesic grids by Wing and Baldridge, 2011.

**Networks**

- Chapters 13 and 14 of E&K explain PageRank and hubs and authorities in networks. The focus is on the web, but 14.5 describes very cool applications to legal citation analysis.
- Emergence of scaling in random networks by Barabasi, 1999.
- Inferring social ties from geographic coincidences by Crandall et al, 2010. Social ties between people can be inferred from co-occurrence in time and space.
- Structural diversity in social contagion by Ugander et al, 2012.
- Topic-partitioned multinetwork embeddings by Krafft et al, 2012.
- Topicflow model: Unsupervised learning of topic-specific influences of hyperlinked documents by Nallapati et al, 2011.

**Public health**

- Drug extraction from the web: Summarizing drug experiences with multi-dimensional topic models by Paul and Dredze, 2013. video
- Towards detecting influenza epidemics by analyzing Twitter messages by Culotta, 2010.
- A Generative Joint, Additive, Sequential Model of Topics and Speech Acts in Patient-Doctor Communication by Wallace et al, 2013.

# Related classes

- CSS at Princeton 2012 by Matthew Sagalnick. Nice reading list, especially for applications of large-scale network analysis.
- CSS at Columbia 2013 by Goel, Hofman, and Vassilvitskii. Emphasizes algorithms for big data analysis, including MapReduce and streaming.
- CSS at George Mason 2009 by Geller and Gulden. Emphasizes agent-based simulations.
- NLP and Social Interaction 2013 by Lillian Lee. Great set of NLP papers that touch on social phenomena. Also has links to lots of useful datasets.
- NLP-flavored CSS at Maryland 2013 by Phillip Resnick. Emphasizes links from language to political science and psychology.
- Networks at Michigan 2008 by Lada Adamic.
- Network analysis at Colorado 2013 by Aaron Clauset.
- Digital literacy and cultural studies at CMU 2013 by Bamman and Warren.

# Conferences, symposia, and meetings

- Atlanta Workshop on Computational Social Science
- Chicago Computational Social Science Workshop
- Stanford Computational Social Science Conference