USC Annenberg

COMM 599: Special Topics: Data Science for Communication & Social Networks

4 Units

Spring 2020 – Tuesday 12:30-3:20pm
Section: 20882D
Location: ASC 228

Instructor: Emilio Ferrara
Office: ISI 933 (Marina del Rey)
Office Hours: On request
Contact Info: emiliofe@usc.edu

I. Course Description

Learn how to unleash the full power and potential of Social Web data for research and business application purposes!

The Social Web pervades all aspects of our lives: we connect and share with friends, search for jobs and opportunities, rate products and write reviews, establish collaborations and projects, all by using online social platforms like Facebook, LinkedIn, Yelp and GitHub. We express our personality and creativity through social platforms for visual discovery, collection and bookmarking like Tumblr and Pinterest. We keep up-to-date, communicate and discuss news and topics of our interest on Twitter and Reddit.

In this course, we will explore the opportunities provided by the wealth of social data available from these platforms. You will learn how to acquire, process, analyze and visualize data related to social networks and media activity, users and their behaviors, trends and information spreading. This journey will bring through the lands of data mining and machine learning methods: supervised and unsupervised learning will be applied to practical problems like social link analysis, opinion mining, and building smart recommender systems. We will explore open-source tools to understand how to extract meaning from human language, use network analysis to study how human connect and discover affinities among people's interests and tastes by building interest graphs.

II. Student Learning Outcomes

Taking this course, you should expect to learn about:

- Networks.
  - Statistical descriptors of networks: link analysis, centrality, and prestige.
  - Network clustering: modularity and community detection.
  - Dynamics of information and epidemics: threshold and information cascade models.
  - Network biases and network manipulation: paradoxes, bots, disinformation.
  - Network visualization algorithms: spring-like layouts, multidimensional scaling, Gephi.

- Applications of texts and documents analysis.
  - Natural Language Processing and Part-of-speech tagging.
  - Sentiment Analysis.
  - Topic Modeling.

- Supervised learning: Crush course on Data Classification.
  - Eager vs. Lazy learning: Decision Trees
  - Ensemble methods: Random Forest
  - Classification performance evaluation: Precision/Recall/F1, Accuracy and ROC Curves.

- Unsupervised learning: Crush course on Clustering Data.
  - Distance and similarity measures & K-means clustering.
  - Hierarchical Clustering and Dendrograms.
  - Clustering performance evaluation.

All topics will be explored from an applied, practical, computational perspective. This will allow the interested student to deepen the rigorous theoretical implications of the methods in other courses offered by USC (for example, CSCI-567 Machine Learning). Throughout the course, we will deliver several "hands-on" sessions with live coding, data analysis, and problem-solving!

III. Course Notes

Class participation and engagement are essential ingredients for success in your academic career, therefore during class turn off cell phones and ringers (no vibrate mode), laptops and tablets. The only exception to use laptops during class is to take notes and during live coding sessions. In this case, please sit in the front rows of the classroom: no email, social media, games, or other distractions will be accepted. Students will be expected to do all readings and assignments, and to attend all meetings unless excused, in writing, at least 24 hours prior.

The following misconducts will automatically result in a zero weight for that component of the grade: (1) failing to attend class on the day of your presentation; (2) failing to attend meetings of your group's Hackathon and/or final presentation; (3) failing to submit your final paper by the expected date. Extenuating circumstances will normally include only serious emergencies or illnesses documented with a doctor's note.

IV. Description and Assessment of Assignments

Assignments

Reaction debriefs: Each student will prepare a "synthesis and reaction" debrief in response to the weekly readings. This will be a brief note, aimed at summarizing in one paragraph the gist of the paper, and provide comments or inputs for discussions, including questions, critiques, and/or theoretical and methodological concerns or ideas. These will be used to guide the discussion session of each class. (Reaction debriefs are not graded)

Readings & discussion

During each lecture (starting lecture 2), one student will hold a 10m presentation on one of the daily reading of choice and will help moderate a discussion session about it. The list of readings is available at the end of the syllabus.

Mid-Term Hackathon
The mid-term exam is in the form of a collaborative hackathon project. The goal is to develop crucial abilities such as:

- Intellectual development: leveraging expertise and multidisciplinary backgrounds, sharing ideas and knowledge.
- Teamwork skills: effective brainstorming, communication and presentation, and group problem-solving.
- Project management skills: ability to set goals, map progress, prototyping-delivery, and matching deadlines.

If possible, we suggest that participants form groups of 2 members with the goal of solving a single problem. Students are encouraged to form groups with members from different academic background when possible. Each group will propose or receive a different problem.

We will propose several problems of interest for the course, as well as receive your explicit solicitations, that should be agreed upon with the Instructor during the first 4 weeks, in the form of a short one-page proposal clearly stating:

- What is the problem?
- Why it is deemed relevant.
- How the group plans to solve the problem.
- Bibliographic references to at least one relevant related paper.

All project proposals will be subject to our approval. Groups will be assigned an approved project, either selected among those proposed by the Instructor, or by the group itself.
Each group will receive a 30m slot for the presentation of their results, in which each member of the group is expected to discuss at least one critical task of the project. The grading of the projects will be in part based on crowd-sourced ratings attributed by other fellow students and submitted in anonymous form at the end of each presentation day.

Final Paper
A serious final paper will be expected. The manuscript will be at least 3,000 words (excluding references) and no more than 4,000 (excluding references) and will include appropriate figures and tables, and unlimited number of references. The work should cover the following points:

- Statement of the problem & Why the problem is important.
- How the problem was faced —including a description of methodology and dataset(s).
- Discussion of results, findings, and limitations of the study.
- Related literature & Final remarks/conclusions.

The final paper should be ideally based on the student's mid-term hackathon project. Text with other group members cannot be shared, figures/tables can be shared when appropriate with proper credit attribution. Grading will be based on soundness (both quality and quantity of original work). Groups of 2 students will be allowed to turn in a single joint-authored manuscript, in the format of a submission for an appropriate peer-reviewed journal or

conference. Each author must contribute sufficient material to justify his/her "equal contribution" in the work. Both authors will receive the same grade for such manuscript.

V. Grading
a. Breakdown of Grade

Assignment	Points	% of Grade
Participation	15	15
Midterm exam	35	35
Final exam	50	50
TOTAL	**100**	**100%**

b. Grading Scale

95% to 100%: A	80% to 83%: B-	67% to 69%: D+
90% to 94%: A-	77% to 79%: C+	64% to 66%: D
87% to 89%: B+	74% to 76%: C	60% to 63%: D-
84% to 86%: B	70% to 73%: C-	0% to 59%: F

VII. Assignment Submission Policy
A. All assignments are due on the dates specified. Lacking prior discussion and agreement with the instructor, late assignments will automatically be given a grade of F.
B. Assignments must be submitted via Blackboard.

VIII. Required Readings and Supplementary Materials
Recommended textbooks (total Amazon price [new/used]: $100/$60)

1. Web Data Mining (2nd Ed.) —by Bing Liu (Amazon price [new/used]: $48/$35)
2. Mining the Social Web (2nd Ed.) —by Matthew A. Russell (Amazon price [new/used]: $27/$15)
3. Programming Collective Intelligence —by Toby Segaran (Amazon price [new/used]: $25/$10)
4. Network Science Book —by Laźzló Barabási (FREE: http://barabasilab.neu.edu/networksciencebook/)
5. Dive into Python —by (FREE: http://www.diveintopython.net/)

Some details: Book 1 will provide insights on methods and approaches studied throughout the course from a machine learning perspective; Books 2 and 3 will serve as recipe books to effectively design and make those methods work with Social Web data; Books 4 and 5 are free resources we will exploit to gather additional material on networks and Python programming. Technical, recommended (non-required) Python "cookbooks":

- Python Data Visualization Cookbook —by Igor Milovanović (ebook: $14)
- Learning IPython for Interactive Computing and Data Visualization —by Cyrille Rossant (ebook: $10)
- Learning scikit-learn: Machine Learning in Python —by Raúl Garreta and Guillermo Moncecchi (ebook: $10)

IX. Laptop Policy

All undergraduate and graduate Annenberg majors and minors are required to have a PC or Apple laptop that can be used in Annenberg classes. Please refer to the **Annenberg Digital Lounge** for more information. To connect to USC's Secure Wireless network, please visit USC's **Information Technology Services** website.

X. Add/Drop Dates for Session 001 (15 weeks: 1/13/20 – 5/1/20)

Friday, January 31: Last day to register and add classes for Session 001
Friday, January 31: Last day to drop a class without a mark of "W," except for Monday-only classes, and receive a refund for Session 001
Tuesday, February 4: Last day to drop a Monday-only class without a mark of "W" and receive a refund for Session 001
Friday, February 28: Last day to drop a course without a mark of "W" on the transcript for Session 001. [Please drop any course by the end of week three (or the 20 percent mark of the session) to avoid tuition charges.]
Friday, February 28: Last day to change pass/no pass to letter grade for Session 001. [All major and minor courses must be taken for a letter grade.]
Friday, April 3: Last day to drop a class with a mark of "W" for Session 001

XI. Course Schedule

a. A Weekly Breakdown

Important note to students: *Be advised that this syllabus is subject to change - and probably will change - based on the progress of the class, news events, and/or guest speaker availability.*

	Topics/Daily Activities	Readings and Homework	Deliverable/Due Dates
Week 1 **Dates: 1/13-1/17**	Introduction of the course & Planning Part 1— Networks • Crash introduction to Networks— Statistical descriptors of networks		
Week 2 **Dates: 1/20-1/24**	Networks (continued) • Network clustering. Modularity and community detection. Readings: Papers [27], [31], [22] and [4] Recommended Chapters: NSB:1 and NSB:2; NBS:9 and WDM:7.5		

Week 3 **Dates: 1/27-1/31**	Networks (continued): • Dynamics of information and epidemics spreading. Readings: Papers [26], [5], [6], [14] Recommended Chapters: NSB:10.1–10.3[pp.11–29] • Hands-on session: mining Twitter. Readings: Papers [12] Recommended Chapters: MtSW:1[pp.5-26] Documentation: Twitter API (https://dev.twitter.com/)	
Week 4 **Dates: 2/3-2/7**	Networks (continued): • Networks and manipulation: bots, disinformation, emotional contagion Readings: [30], [39] [18], and [20]	
Week 5 **Dates: 2/10-2/14**	Networks (continued): • Network visualization algorithms. Readings: Papers [1] and [28] Recommended Chapters: PCI:12[pp.300–302(MDS)] • Hands-on session: tutorial on Gephi. Readings: Papers [3] Recommended Chapters: NSB:10.4–10.7[pp.30–58] Documentation: Gephi Wiki https://wiki.gephi.org/index.php/Main_Page	
Week 6 **Dates: 2/17-2/21**	Networks (continued): Guest speaker: Prof. Kristina Lerman • Bias in networks: friendship paradoxes & perception bias – network structures bias perception. • Readings: Papers [33], [23], [24] Ask Me Anything (AMA) session with guest	
Week 7 **Dates: 2/24-2/28**	Part 1—Text and Documents • Crash intro to Natural Language Processing: Part-of-Speech Tagging. Readings: Papers [16] Recommended Chapters: WDM:6.5 and MtSW:5.3–5.5[pp.190–222]	

	• Hands-on session: Tutorial on NLP	
Week 8 **Dates: 3/2-3/6**	Text and Documents (continued) • Sentiment Analysis Readings: Papers [15] and [29] Recommended Chapters: MtSW:4[pp.135–180] • Topic Modeling Readings: Papers [2] and [10] Recommended Chapters: WDM:6.7	
Week 9 **Dates: 3/9-3/13**	Mid-term Hackathon week	Mid-term Hackathon presentations
Spring Break **Dates: 3/16-3/20**	No Classes	
Week 10 **Dates: 3/23-3/27**	Part 3—Supervised Learning • Crash intro to Supervised learning. Readings: Papers [17], [19], and [11] Recommended Chapters: WDM:3.1 • Eager vs. Lazy learning—Decision Trees. Readings: Papers [21] Recommended Chapters: WDM:3.2 and WDM:3.9	
Week 11 **Dates: 3/30-4/3**	Supervised Learning (continued) • Ensemble methods & Classification performance evaluation. Readings: Papers [9] and [7] Recommended Chapters: WDM:3.3 and WDM:3.10	
Week 12 **Dates: 4/6-4/10**	Part 4—Unsupervised Learning • Crash introduction to Unsupervised learning—Distance measures & K-means clustering. Readings: Papers [38] and [37] Recommended Chapters: WDM:4.1–4.3[pp.133–147]	
Week 13 **Dates: 4/13-4/17**	Unsupervised Learning (continued) • Hierarchical clustering & Dendrograms. Readings: Papers [25], [32], [34]	

	Recommended Chapters: WDM:4.3–4.5[pp.147–155]	
Week 14 **Dates: 4/20-4/24**	Unsupervised Learning (continued) • Clustering performance evaluation Readings: Papers [35], [36] Recommended Chapters: WDM:4.6–4.10[pp.155–165]	
Week 15 **Dates: 4/27-5/1**		**Project presentations**
FINAL EXAM PERIOD **Dates: 5/6-5/13**		**Final paper submission**

b. Reading list

[1] S. Aral and D. Walker. Identifying influential and susceptible members of social networks. Science, 337(6092):337–341, 2012.

[2] D. M. Blei. Probabilistic topic models. Communications of the ACM, 55(4):77–84, 2012.

[3] R. M. Bond, C. J. Fariss, J. J. Jones, A. D. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler. A 61-million-person experiment in social influence and political mobilization. Nature, 489(7415):295–298, 2012.

[4] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca. Network analysis in the social sciences. Science, 323(5916):892–895, 2009.

[5] D. Centola. The spread of behavior in an online social network experiment. Science, 329(5996):1194–1197, 2010.

[6] D. Centola. An experimental study of homophily in the adoption of health behavior. Science, 334(6060):1269–1272, 2011.

[7] A. Cho. Ourselves and our interactions: the ultimate physics problem? Science, 325(5939):406, 2009.

[8] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. Proceedings of the National Academy of Sciences, 107(52):22436–22441, 2010.

[9] V. Dhar. Data science and prediction. Communications of the ACM, 56(12):64–73, 2013.

[10] P. S. Dodds, R. Muhamad, and D. J. Watts. An experimental study of search in global social networks. Science, 301(5634):827–829, 2003.

[11] P. Domingos. A few useful things to know about machine learning. Communications of the ACM, 55(10):78–87, 2012.

[12] W. Fan and M. D. Gordon. The power of social media analytics. Communications of the ACM, 57(6):74–81, 2014.

[13] H. Garcia-Molina, G. Koutrika, and A. Parameswaran. Information seeking: convergence of search, recommendations, and advertising. Communications of the ACM, 54(11):121–130, 2011.

[14] Lorenz-Spreen, P., Mønsted, B. M., Hövel, P., & Lehmann, S. (2019). Accelerating dynamics of collective attention. Nature communications, 10(1), 1759.

[15] S. A. Golder and M. W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. Science, 333(6051):1878–1881, 2011.

[16] DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). Big Data & Society, 2(2), 2053951715602908.

[17] N. Jones. Computer science: The learning machines. Nature, 505(7482):146, 2014.

[18] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151.

[19] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 110(15):5802–5805, 2013.

[20] A. D. Kramer, J. E. Guillory, and J. T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences, page 201320040, 2014.

[21] D. Lazer, R. Kennedy, G. King, and A. Vespignani. Big data. the parable of google flu: traps in big data analysis. Science, 343(6176):1203, 2014.

[22] D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al. Life in the network: the coming age of computational social science. Science, 323(5915):721, 2009.

[23] Lee, E., Karimi, F., Wagner, C., Jo, H. H., Strohmaier, M., & Galesic, M. (2019). Homophily and minority-group size explain perception biases in social networks. Nature human behaviour, 3(10), 1078-1087.

[24] Kooti, F., Hodas, N. O., & Lerman, K. (2014, May). Network weirdness: Exploring the origins of network paradoxes. In Eighth International AAAI Conference on Weblogs and Social Media.

[25] D. Liben-Nowell and J. Kleinberg. Tracing information flow on a global scale using internet chain-letter data. Proceedings of the National Academy of Sciences, 105(12):4633–4638, 2008.

[26] P. T. Metaxas and E. Mustafaraj. Social media and the elections. Science, 338(6106):472–473, 2012.

[27] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela. Community structure in time-dependent, multiscale, and multiplex networks. Science, 328(5980):876–878, 2010.

[28] L. Muchnik, S. Aral, and S. J. Taylor. Social influence bias: A randomized experiment. Science, 341(6146):647–651, 2013.

[29] Stella, M., Ferrara, E., & De Domenico, M. (2018). Bots increase exposure to negative and inflammatory content in online social systems. Proceedings of the National Academy of Sciences, 115(49), 12435-12440.

[30] Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. Communications of the ACM, 59(7), 96-104.

[31] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences, 105(4):1118–1123, 2008.

[32] M. J. Salganik, P. S. Dodds, and D. J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. Science, 311(5762):854–856, 2006.

[33] Feld, S. L. (1991). Why your friends have more friends than you do. American Journal of Sociology, 96(6), 1464-1477.

[34] M. Schich, C. Song, Y.-Y. Ahn, A. Mirsky, M. Martino, A.-L. Barabási, and D. Helbing. A network framework of cultural history. Science, 345(6196):558–562, 2014.

[35] C. Staff. Recommendation algorithms, online privacy, and more. Communications of the ACM, 52(5):10–11, 2009.

[36] G. Szabo and B. A. Huberman. Predicting the popularity of online content. Communications of the ACM, 53(8):80–88, 2010.

[37] A. Vespignani. Modelling dynamical processes in complex socio-technical systems. Nature Physics, 8(1):32–39, 2012.

[38] A. Vespignani. Predicting the behavior of techno-social systems. Science, 325(5939):425, 2009.

[39] Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 US Presidential election online discussion. First Monday, 21(11-7).

Statement on Academic Conduct and Support Systems

a. Academic Conduct

Plagiarism

Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in *SCampus* in Part B, Section 11, "Behavior Violating University Standards" policy.usc.edu/scampus-part-b. Other forms of academic dishonesty are equally unacceptable. See additional information in *SCampus* and university policies on scientific misconduct, policy.usc.edu/scientific-misconduct.

b. Support Systems

Counseling and Mental Health - (213) 740-9355 – 24/7 on call
studenthealth.usc.edu/counseling
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.

National Suicide Prevention Lifeline - 1 (800) 273-8255 – 24/7 on call
suicidepreventionlifeline.org
Free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week.

Relationship and Sexual Violence Prevention and Services (RSVP) - (213) 740-9355(WELL), press "0" after hours – 24/7 on call
studenthealth.usc.edu/sexual-assault
Free and confidential therapy services, workshops, and training for situations related to gender-based harm.

Office of Equity and Diversity (OED)- (213) 740-5086 | Title IX – (213) 821-8298
equity.usc.edu, titleix.usc.edu
Information about how to get help or help someone affected by harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants. The university prohibits discrimination or harassment based on the following *protected characteristics*: race, color, national origin, ancestry, religion, sex, gender, gender identity, gender expression, sexual orientation, age, physical disability, medical condition, mental disability, marital status, pregnancy, veteran status, genetic information, and any other characteristic which may be specified in applicable laws and governmental regulations. The university also prohibits sexual assault, non-consensual sexual contact, sexual misconduct, intimate partner violence, stalking, malicious dissuasion, retaliation, and violation of interim measures.

Reporting Incidents of Bias or Harassment - (213) 740-5086 or (213) 821-8298
usc-advocate.symplicity.com/care_report
Avenue to report incidents of bias, hate crimes, and microaggressions to the Office of Equity and Diversity |Title IX for appropriate investigation, supportive measures, and response.

The Office of Disability Services and Programs - (213) 740-0776
dsp.usc.edu
Support and accommodations for students with disabilities. Services include assistance in providing readers/notetakers/interpreters, special accommodations for test taking needs, assistance with architectural barriers, assistive technology, and support for individual needs.

USC Support and Advocacy - (213) 821-4710
uscsa.usc.edu
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

Diversity at USC - (213) 740-2101
diversity.usc.edu
Information on events, programs and training, the Provost's Diversity and Inclusion Council, Diversity Liaisons for each academic school, chronology, participation, and various resources for students.

USC Emergency - UPC: (213) 740-4321, HSC: (323) 442-1000 – 24/7 on call
dps.usc.edu, emergency.usc.edu
Emergency assistance and avenue to report a crime. Latest updates regarding safety, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible.

USC Department of Public Safety - UPC: (213) 740-6000, HSC: (323) 442-120 – 24/7 on call
dps.usc.edu
Non-emergency assistance or information.

Annenberg Student Success Fund
The Annenberg Student Success Fund is a donor-funded financial aid account available to USC Annenberg undergraduate and graduate students for non-tuition expenses related to extra- and co-curricular programs and opportunities.

XIII. About Your Instructor
Dr. Emilio Ferrara is Research Assistant Professor and Associate Director of Informatics & Data Science at the USC Department of Computer Science, Research Team Leader at the USC Information Sciences Institute, and Principal Investigator at the USC/ISI Machine Intelligence and Data Science (MINDS) group.

Ferrara's research interests include using AI for modeling and predicting human behavior in

techno-social systems. Ferrara has published over a hundred articles on social networks, machine learning, and network science, appeared in venues like Proceeding of the National Academy of Sciences, Communications of the ACM, Physical Review Letters, and his research has been featured on all major news outlets.

He was named 2015 IBM Watson Big Data Influencer, he received the 2016 DARPA Young Faculty Award, the 2016 Complex Systems Society Junior Scientific Award, and the 2018 DARPA Director's Fellowship. His research is supported by DARPA, IARPA, the Air Force, and the Office of Naval Research.