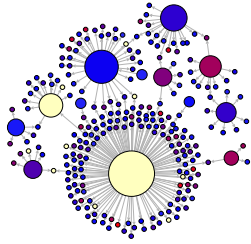


COURSE SYLLABUS

Department of Computer Science and Engineering, College of Engineering

CRN 95954 / 97121, Sec. 001, 3 Credits



Network of Retweets
Source: Shao et al., "The spread of low-credibility content by social bots"

Instructor	Prof. Giovanni Luca Ciampaglia	Term & Year	Fall 2020
Phone	(813) 396-9349	Class Days	Mon, Wed 2-3:15 pm
E-mail	glc3@mail.usf.edu	Class Location	CHE 100 / Blackboard
Website	glciampaglia.com		
Office Hours	T/Th 11-12:30 pm on Blackboard Collaborate Ultra		
TA	Anis Elebiary; 2-4:30 pm on BlackBoard. Email: anis@usf.edu		
Modality	Hybrid remote / F2F. Course can be completed remotely; F2F not required. After Thanksgiving we go fully online.		
First Day Attendance	Attend class on the 1 st day & complete the "Syllabus Quiz" by Friday, Aug. 28 at 10 am. NOTE: If you get less than perfect score in the quiz you will be dropped from the class. You have 3 attempts.		

I Welcome!

How does Facebook know who are the people you went to high school with? Who suggests what we should watch on Netflix tonight? Can we predict if a new meme will go "viral" on Twitter? Learn and understand how the digital trails left behind by millions of users online can be turned into actionable knowledge for a variety of applications.

II University Course Description

This course introduces useful techniques to model, analyze, and understand large-scale social media, with focus on social network analysis, user modeling, bot detection, and dynamical processes over social and information networks.

III Prerequisites (UG section only)

- COP 4530 Data Structures (C- minimum) or equivalent,
- CDA 3201 Computer Logic Design (C- minimum) or equivalent.

IV Course Objectives & Learnings Outcomes

The objective of this course is to provide an introduction to quantitative methods for the analysis of data from social media apps and platforms. The course will cover principles of data acquisition, processing, and mining, including analysis and visualization of networks, text, and maps. An additional objective is to provide an introduction to best practices for development of machine learning and data mining projects.

Students will demonstrate the ability to acquire, process, analyze and visualize social media data; to organize code and workflows with version control systems; and to employ opensource toolkits for data mining and machine learning such as *scikit-learn* or *Gephi*.

V How to Succeed in this Class

You should be familiar with basic descriptive statistics (mean, variance, etc.), probability theory (mostly of the discrete kind, e.g. chain rule, Bayes rule, etc.), and linear algebra, so please refresh them during the first weeks so that you can follow the lectures. Having a data mining or machine learning background is a plus, but not necessary. All programming assignments will be in Python. I have listed some resources to get up to speed with it below. Additional materials are on Canvas.

There will be heavy demand on your time beyond the classroom. The nature of the topic will require you to read papers/books, write reports, understand and implement algorithms, and present to the class. You will need to be programming almost every week for this course.

Last but not least: although social media has implications in many disciplines (including philosophy and literature!), in this course we are going to view it primarily from the lens of STEM. A lot of people think that STEM topics are not for them. The opposite is true: *anyone can learn this stuff*. I recommend a nice essay by Susan J. Fowler (bit.ly/2ZnmzRS) about this small, but important, truth.

VI Required Texts and/or Readings and Course Materials

- Reza Zefarani, Mohammad Ali Abbasi, Huan Liu, "Social Media Mining: An Introduction." Cambridge University Press, 2014. ISBN: 978-1107018853.
- The PDF of the book is available through the USF Library (<https://doi-org.ezproxy.lib.usf.edu/10.1017/CB09781139088510> note: requires login with your USF username and password). You can buy a hardcopy if you want, but it is not required.
- There will be additional *required* readings from the literature; see Section VIII for more information.

VII Course topics

- **Graph Theory:** graphs, walks, paths, graph traversal;
- **Centrality Measures:** degree, eigenvectors, betweenness, etc.;
- **Statistical Properties of Networks:** density, clustering, structural balance, degree distribution;
- **Models of Networks:** Random Graphs (E-R model, Small World), Preferential Attachment (B-A model);
- **Data Mining for Networks:** Classification, TF/IDF, Decision Trees, k -NN, Label Propagation;
- **Information diffusion:** Cascades, Linear Threshold Model, *CAP 6317 only:* Epidemic Spreading (SI, SIR, SIS).

VIII Schedule

This is a tentative schedule of the course, note that it may be subject to change as the term progresses.

* = Required Reading (due 12:00 pm noon before class).

Fall 2020 - CLASS DATES

Mon	Topic	Notes	Wed	Topic	Notes
Aug. 24	Graphs & social media: definitions, typologies	Sec. 2.1	Aug. 26	Graph representation	Sec. 2.2 Syllabus Quiz due (Fri, Aug. 28, 10 am)
Aug. 31	Types of graphs	Sec. 2.3, Parlante, <i>Google's Python Class</i> , *Feld, "Why Your Friends Have More Friends Than You Do"	Sep. 02	Connectivity in graphs	Sec. 2.4, HW0 due (Twitter setup)
Sep. 07	Labor day (no class)	*Quercia, Schifanella, and Aiello, "The Shortest Path to Happiness: Recommending Beautiful, Quiet, and Happy Routes in the City"	Sep. 09	Graph traversal	Sec. 2.6
Sep. 14	Degree & Eigenvector centrality	Sec. 3.1–3.1.2 (incl.), *Page et al., <i>The PageRank Citation Ranking: Bringing Order to the Web.</i> (<i>CAP 4773 only</i>), *Leicht, Holme, and Newman, "Vertex similarity in networks" (<i>CAP 6137 only</i>)	Sep. 16	Katz centrality, PageRank, <i>CAP 6317 only:</i> Similarity	Sec. 3.1.3–3.1.4 (incl.), <i>CAP 6317 only:</i> Sec. 3.4, HW1 due
Sep. 21	Betweenness, closeness centrality	Sec. 3.1.5–3.1.6 (incl.), *Krishnamurthy, Gill, and Arlitt, "A Few Chirps about Twitter" (<i>CAP 4773 only</i>), *Kwak et al., "What is Twitter, a Social Network or a News Media?" (<i>CAP 6317 only</i>)	Sep. 23	LAB 1 Twitter API	Twitter Developer Docs, <i>Introduction to TweetJSON</i> , Tweepy Contributors, <i>Tweepy Documentation</i>



Mon	Topic	Notes	Wed	Topic	Notes
Sep. 28	Transitivity and Reciprocity	Sec. 3.2, *Brzozowski, Hogg, and Szabo, "Friends and Foes: Ideological Social Networking" (<i>CAP 4337 only</i>), *Leskovec, Huttenlocher, and Kleinberg, "Signed Networks in Social Media" (<i>CAP 6317 only</i>)	Sep. 30	Balance and Status	Sec. 3.3, HW2 due
Oct. 05	Degree distributions	Sec. 4.1–4.1.1 (incl.), Adamic, <i>Zipf, Power-laws, and Pareto - a ranking tutorial</i> , *Broder et al., "Graph structure in the Web" Clauset, Shalizi, and Newman, "Power-Law Distributions in Empirical Data" (<i>CAP 6317 only</i>)	Oct. 07	Clustering coefficient, average path length	Sec. 4.1.2–4.1.3 (incl.), HW 3 due
Oct. 12	Random graph model: definition and evolution	Sec. 4.2–4.2.1 (incl.) *Ahn et al., "Analysis of Topological Characteristics of Huge Online Social Networking Services",	Oct. 14	Properties of random graphs and modeling with random graphs	Sec. 4.2.2–4.2.3 (incl.), HW4 due
Oct. 19	Small-world model	Sec. 4.3, *Watts and Strogatz, "Collective dynamics of 'small-world' networks"	Oct. 21	LAB 2 NetworkX	Hagberg, Schult, and Swart, "Exploring Network Structure, Dynamics, and Function using NetworkX", Project Proposal due
Oct. 26	Preferential attachment	Sec. 4.1–4.4.1 (excl.), *Barabási and Albert, "Emergence of Scaling in Random Networks"	Oct. 28	The Barabási-Albert model	Sec. 4.4.1–4.4.2
Nov. 02	Data Mining essentials: TF/IDF, Decision Trees, relational neighbor classifier	Sec 5.1–5.4 (incl.), *Conover et al., "Predicting the Political Alignment of Twitter Users", Hutto and Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text"	Nov. 04	LAB 3 Scikit-Learn	Sec. 5.4.3–5.4.4 (incl.), Buitinck et al., "API design for machine learning software: experiences from the scikit-learn project"
Nov. 09	Information Cascades	Sec. 7.2, *Vosoughi, Roy, and Aral, "The spread of true and false news online", Shao et al., "The spread of low-credibility content by social bots"	Nov. 11	LAB 4 Gephi	Jacomy et al., "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software"
Nov. 16	Epidemic Spreading Models	Sec. 7.3–7.3.2 (incl.) (<i>CAP 4773 only</i>), Sec. 7.3–7.4 (incl.), (<i>CAP 6317 only</i>) *Christakis and Fowler, "Social Network Sensors for Early Detection of Contagious Outbreaks"	Nov. 18	Social Influence and the Linear Threshold Model	Sec. 8.2, Presentation Slides due
Nov. 23	Final Presentations	Groups 1, 2, 3	Nov. 25	Final Presentations	Groups 4, 5, 6
Nov. 30	Final Presentations	Groups 7, 8, 9	Dec. 02	Final Presentations	Groups 10, 11, 12
Dec. 07	Final Report due (9 am on Mon, Dec. 7th, 2020)				

IX Grading

This is a research-oriented project-based course (total 100 points). Instead of exams, each student will do four individual assignments and a group research project. Each assignment will consist of two exercises: a “theory” problem (based on materials from both the book and the readings) and an auto-graded coding problem. Each student should submit the solutions to both exercises in Canvas (<https://usflearn.instructure.com/courses/1503906>) before 12:00 pm (noon) on Wednesday.

For your assignments you will have to code in Python, due to auto-grading.

Solutions to your **auto-graded coding problems** have to be self-sufficient and not dependent on other data mining / social network analysis packages or code, such as NetworkX or scikit-learn. You may use packages for display graphics or mathematics packages, such as routines for data processing, linear algebra, statistics, and optimization — for example pandas, numpy, or scipy (but not the Compressed Sparse Graph Routines). You may also use Tweepy or other networking library (e.g. Scrapy, requests) to collect data from Twitter.

For your final project you may use other packages or code and a language other than Python.

The class will assign one paper for reading each week. Each student should read the assigned paper and submit a short critique (between 100–200 words) online in Canvas (<https://usflearn.instructure.com/courses/1503906>) before 12:00 pm (noon) on Monday. These reviews should not be simple summaries, but discuss positive aspects of the paper and limitations (see here for some examples: <https://nlpers.blogspot.com/2016/08/some-papers-i-liked-at-acl-2016.html>), or suggestions for how the work could be improved or extended. Submissions that simply state a summary of the paper will be assessed a grade of ‘incomplete’.

- Attendance (individual, 10% of final grade),
- 12× Readings (individual, 10% of final grade),
- 4× Homework assignments (individual, 40% of final grade divided as 10% each),
- Group project (40% of final grade, divided as 30% common + 10% individual; groups of four max):
 - Project Proposal (10%);
 - Presentation (10%);
 - Written Report (20%).

IX (a) Late work policy

There are no make-ups for reading reviews and assignments.

At the beginning of the semester each student has six (6) extension tokens each equivalent to a 24h extension on the due date of a homework assignment or a reading review. To use a token, please contact the TA or the instructor. No reasons need to be provided. Tokens cannot be used on the group project or for attendance.

Please e-mail your reviews or homework to the instructor if there are any technical issues with online submission. Assignments and reviews turned in late will be assessed a 25% penalty of the earned grade each late day.

IX (b) Regrade policy

If you believe an error has been made in the grading of your work, you may resubmit it for a regrade — submit a detailed explanation of which problems you think we marked incorrectly and why. Because we will examine your entire submission in detail, your grade can go up or down as a result of a regrade request.

IX (c) Group Work Policy

Everyone must take part in a group project. All members of a group will receive the same score; that is, the project is assessed and everyone receives this score. However, that number is only 75% of your grade for this project. The final 25% is individual and refers to your teamwork. The instructor will assign a grade that is informed by declaration of work division in the project report and performance during project presentations. Once formed, groups cannot be altered or switched, except for reasons of extended absence due to medical reasons.

Grading scale	
94–100	A
90–93	A-
87–89	B+
84–86	B
80–83	B-
77–79	C+
74–76	C
70–73	C-
67–69	D+
64–66	D
60–63	D-
0–59	F

IX (d) Midterm grade (CAP 4773 only)

There is no midterm exam in this course. However, per USF System Policy 10-504, a midterm grade will be made available to you in OASIS. It will reflect the attendance rate and any graded assignments up to that point.

X Course Policies: Technology and Media

Piazza/E-mail We will use Piazza for all announcements and course discussions. You are responsible for checking it regularly to stay abreast of everything that is announced there. To access Piazza you can go on the course page on Canvas, or you can sign up directly at the following link: <https://piazza.com/usf/fall2020/cap4773cis6930>. (Please use your USF e-mail to sign up, not your personal one.)

- Use Piazza for any question about the materials seen in class. Comments and suggestions are also welcome. If you know the answer to something posted by one of your classmates, you are more than welcome to chime in!
- For any other academic or logistical issues feel free to message us either on Canvas, via e-mail, or via a private post on Piazza. We will try to do our best to respond to messages within 24h during regular week days. We cannot promise an answer over the weekend.
- Do NOT share code/solutions on Piazza: this is a breach of academic integrity.

Canvas This course will make use of USF's learning management system (LMS), Canvas. If you need help learning how to perform various tasks related to this course, please view the following videos or consult the Canvas help guides. You may also contact USF's IT department at (813) 974-1222 or help@usf.edu.

Blackboard Collaborate Ultra We may use lecture-capturing. Note that student voices may be heard in the recordings.

Laptop Usage Use of laptop during lecture is allowed only for capturing notes and looking up course/lecture related materials. Use of laptop for other purposes is prohibited.

Classroom Devices/Student Recording Use of tape-recorders or other recording devices is NOT allowed in the class.

You do not have the right to sell notes or tapes of lectures generated from this class.

Phone Usage Use of phone during class is not allowed, including texting or surfing the Internet. Students may not take photos/video/audio recordings of the class. Only pictures of the whiteboard notes are allowed.

XI Course Policies: Student Expectations**XI (a) Standard University Policy**

Policies about disability access, religious observances, academic grievances, academic misconduct, and several other topics are governed by a central set of policies that apply to all classes at USF. These may be accessed at: www.usf.edu/provost/faculty/core-syllabus-policy-statements.aspx.

XI (b) COVID-19 Procedures

All students must comply with university policies and posted signs regarding COVID-19 mitigation measures, including wearing face coverings and maintaining social distancing. Failure to do so may result in dismissal from class, referral to the Student Conduct Office, and possible removal from campus. Additional details are available on the University's Core Syllabus Policy Statements page: www.usf.edu/provost/faculty/core-syllabus-policy-statements.aspx

XI (c) Attendance policy

Students are expected to attend classes. There is no makeup for any in-class assignments or worksheets. See Section IX (a) regarding extension tokens for assignments and readings. Documented excused absences for any group project activity (proposal, presentation, final report) may be allowed by making arrangements ahead of time (when possible) or by providing a reasonable amount of time to make up for the missed work.

XI (d) Turnitin / MOSS

In this course, we will Turnitin (<http://turnitin.com/>), an automated system used by instructors to quickly and easily compare each student's assignment with billions of web sites, as well as an enormous database of student papers that grows with each submission. After the assignment is processed, the TA and I will receive a report showing if and how another author's work was used in the assignment.


We will also use MOSS code checker (<https://theory.stanford.edu/~aiken/moss/>) to check for code plagiarism.

XII Important Dates to Remember (see Section VIII for detailed syllabus)

The dates and assignments in this syllabus are tentative and can be changed at the discretion of the professor.

Event	Date	Time
Syllabus Quiz	Friday, August 28, 2020	10 am
Drop/Add Deadline	Friday, August 28, 2020	
Project Proposals due	Wednesday, October 21, 2020	12 pm (noon)
Presentation Slides due	Wednesday, November 18, 2020	12 pm (noon)
Final Report due	Monday, December 7, 2020	09 am

XIII List of Readings (* = required)

 PDFs of the required readings are available in Canvas.

- [1] Lada Adamic. *Zipf, Power-laws, and Pareto - a ranking tutorial*. <https://www.labs.hp.com/research/idl/papers/ranking/ranking.html>. 2000.
- [*2] Yong-Yeol Ahn et al. "Analysis of Topological Characteristics of Huge Online Social Networking Services". In: *Proceedings of the 16th International Conference on World Wide Web*. WWW '07. Banff, Alberta, Canada: Association for Computing Machinery, 2007, pp. 835–844. ISBN: 9781595936547.
- [*3] Albert-László Barabási and Réka Albert. "Emergence of Scaling in Random Networks". In: *Science* 286.5439 (1999), pp. 509–512. ISSN: 0036-8075.
- [*4] Andrei Broder et al. "Graph structure in the Web". In: *Computer Networks* 33.1 (2000), pp. 309–320. ISSN: 1389-1286.
- [*5] Michael J. Brzozowski, Tad Hogg, and Gabor Szabo. "Friends and Foes: Ideological Social Networking". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '08. Florence, Italy: Association for Computing Machinery, 2008, pp. 817–820. ISBN: 9781605580111.
- [6] Lars Buitinck et al. "API design for machine learning software: experiences from the scikit-learn project". In: *European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases*. Prague, Czech Republic, Sept. 2013.
- [*7] Nicholas A. Christakis and James H. Fowler. "Social Network Sensors for Early Detection of Contagious Outbreaks". In: *PLoS One* 5.9 (Sept. 2010), pp. 1–8.
- [8] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. "Power-Law Distributions in Empirical Data". In: *SIAM Review* 51.4 (2009), pp. 661–703.
- [*9] M. D. Conover et al. "Predicting the Political Alignment of Twitter Users". In: *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. 2011, pp. 192–199.
- [*10] Scott L. Feld. "Why Your Friends Have More Friends Than You Do". In: *American Journal of Sociology* 96.6 (1991), pp. 1464–1477. ISSN: 00029602, 15375390.
- [11] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. "Exploring Network Structure, Dynamics, and Function using NetworkX". In: *Proceedings of the 7th Python in Science Conference*. Ed. by Gaël Varoquaux, Travis Vaught, and Jarrod Millman. Pasadena, CA USA, 2008, pp. 11–15.
- [12] C. Hutto and Eric Gilbert. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text". In: *Eighth International AAI Conference on Weblogs and Social Media*. Ann Arbor, Michigan, USA: AAAI, June 2014.
- [13] Mathieu Jacomy et al. "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software". In: *PLOS ONE* 9.6 (June 2014), pp. 1–12.



- [*14] Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. "A Few Chirps about Twitter". In: *Proceedings of the First Workshop on Online Social Networks*. WOSN '08. Seattle, WA, USA: Association for Computing Machinery, 2008, pp. 19–24. ISBN: 9781605581828.
- [*15] Haewoon Kwak et al. "What is Twitter, a Social Network or a News Media?" In: *Proceedings of the 19th International Conference on World Wide Web*. WWW '10. Raleigh, North Carolina, USA: Association for Computing Machinery, 2010, pp. 591–600. ISBN: 9781605587998.
- [*16] E. A. Leicht, Petter Holme, and M. E. J. Newman. "Vertex similarity in networks". In: *Phys. Rev. E* 73 (2 Feb. 2006), p. 026120.
- [*17] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. "Signed Networks in Social Media". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '10. Atlanta, Georgia, USA: ACM, 2010, pp. 1361–1370.
- [*18] Lawrence Page et al. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report 1999-66. Previous number = SIDL-WP-1999-0120. Stanford InfoLab, Nov. 1999. URL: <http://ilpubs.stanford.edu:8090/422/>.
- [19] Nick Parlante. *Google's Python Class*. <https://developers.google.com/edu/python/>. July 2015.
- [*20] Daniele Quercia, Rossano Schifanella, and Luca Maria Aiello. "The Shortest Path to Happiness: Recommending Beautiful, Quiet, and Happy Routes in the City". In: *Proceedings of the 25th ACM Conference on Hypertext and Social Media*. HT '14. Santiago, Chile: ACM, 2014, pp. 116–125.
- [21] Chengcheng Shao et al. "The spread of low-credibility content by social bots". In: *Nature Communications* 9.1 (Nov. 2018), p. 4787. ISSN: 2041-1723.
- [22] Tweepy Contributors. *Tweepy Documentation*. <http://docs.tweepy.org/en/latest/>. 2020.
- [23] Twitter Developer Docs. *Introduction to Tweet JSON*. <https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/overview/intro-to-tweet-json>. 2020.
- [*24] Soroush Vosoughi, Deb Roy, and Sinan Aral. "The spread of true and false news online". In: *Science* 359.6380 (2018), pp. 1146–1151.
- [*25] Duncan J. Watts and Steven H. Strogatz. "Collective dynamics of 'small-world' networks". In: *Nature* 393.6684 (June 1998), pp. 440–442. ISSN: 1476-4687.