PP291B, Network Science Using R
Professor Zachary Steinert-Threlkeld
Winter 2022
Tuesday and Thursday, 14:00-15:15 in Public Affairs 2214

Per UCLA orders, class is remote through Monday January 17.

Office hours: Monday 13:00-15:00 in Public Affairs 6347
Zoom: https://ucla.zoom.us/j/93008888979

Use this link for remote classes and office hours.

E-mail: zst@luskin.ucla.edu

Course Description

Network analysis offers a framework for understanding relationships between entities such as people, places, or cultural objects. For example, why individuals decide to protest or vote, the amount of education they pursue, or the effect of human interference in an ecosystem can all be considered using network analysis. Network analysis also is the foundation of many content recommendation algorithms. The course is structured around a weekly introduction of concepts from network analysis, followed by working through it using R, a popular programming language. No prior knowledge of R is required.

Textbooks

This course uses a combination of readings from Douglas A. Luke's *A User's Guide to Network Analysis in R*, Eric D. Kolacyzk and Gábor Csárdi's *Statistical Analysis of Network Data with R*, and *Network Science* by Alberto-Lásló Barábasi. The first two are available as ebooks through the library, and the latter is available via its own website.

Other useful books include:

- *Networks: An Introduction* by M.E.J. Newman. Oxford University Press. Another very good textbook; more thorough than Barabasi's but about as technical.
- *Social Network Analysis, Second Edition* by David Knoke and Song Yang. SAGE Publications.
- *Social Network Analysis in Program Evaluation* edited by Maryann M. Durland and Kimberly A. Fredericks.
- *Social Network Analysis: A Handbook* by John Scott. SAGE Publications.
- *Social Network Analysis: Methods and Applications* by Stanley Wasserman and Katherine Faust. Cambridge University Press.

- *The SAGE Handbook of Social Network Analysis* edited by John Scott and Peter J. Carrington. SAGE Publications.

Software

R is an open source, Turing complete language designed for statistical analysis. While public affairs does not traditionally require programming, we increasingly live in a world in which facility with it and its concepts affects your future success. While I do not expect you to have any programming knowledge or even particularly enjoy working with computers, I do expect you to work hard to learn how to conduct network analysis in R.

I have made an introductory guide to R and posted it to the Readings section of the course website. It includes information on how to install R and RStudio, which is an integrated development environment that many people find easier than base R. The guide also provides links to several excellent tutorials.

This course uses RStudio to compile R Markdown documents, which combine R code and prose into HTML or .pdf documents. Markdown is a markup language, like HTML: you write text, add some commands for how the text should be displayed, and then compile the text to get it displayed how you want. (This approach is in contrast to what-you-see-is-what-you-get (WYSIWYG) approaches like Microsoft Word.) R Markdown is the version of Markdown the developers of RStudio created, and it makes it easy to combine prose with R code in one document. See the Assignments section for more detail.

Coding is hard, and you will get frustrated. It is common for all programmers to write buggy code, especially when a novice. It does not help that most of R's error messages are hard to decipher One skill you will therefore develop in this class is learning what to ask Google to get directed to the right StackOverflow questions. In all of my coding, I have errors that slow me down and many that send me to the internet; struggling with programming is programming. At the same time, do not be shy about asking me coding questions, either via the Discussion section of the course website or e-mail. Previous classes have found the forum a useful way to ask questions because then others realize they are not struggling alone. As I tell every class I teach, ***THERE ARE NO STUPID QUESTIONS***.

Class

Broadly speaking, each week is structured so that the Tuesday is a lecture on a network concept and Thursday is an in-class lab where we program what we learned. You will need your computer for the lab classes. In the classes where I introduce network concepts, you can choose whether to take notes on your computer or by hand. I encourage you to take notes by hand, as growing evidence suggests that doing so leaders to better retention of course material.

Assignments

All assignments are released as R Markdown documents. When you complete them, please compile them to HTML or .pdf documents. (Compiling means telling the computer to execute any code and formatting commands in the document.) If you choose to compile to

.pdf and do not have Latex installed on your computer, then make sure to use the `tinytex` package. It is an R package that you will install to so that you can compile R Markdown to .pdf.

Problem Sets. This course features five problem sets to help you learn R by coding network concepts learned up to that point, though focusing on the concepts from the previous week. Since the first time we see code is in Thursday's class, they are released after class and due the following Tuesday by the start of class. This cadence is why my office hours are on Monday.

The problem sets are short, designed to give you practice with R, and should not be a source of grade stress. Each is graded as a ✓-, ✓, or ✓+. ✓- means you did not turn the problem set in; ✓means you turned it in but the code does not work, and ✓+ means everything works. ✓+ sounds intimidating but is not; most problem sets receive that grade. At the end of the quarter, I let ✓+ = 3, ✓= 2, ✓-= 1, add the scores, and calculate the problem set grade. Problem sets turned in late without an excused reason will automatically receive a one point deduction.

Midterm Exam. We will have a midterm exam in Week 6, on Thursday February 10th. The purpose of this exam is test your ability to use R to analyze networks up to that point in the quarter. I will provide you an R Markdown file with questions, your answers will be code, and you will submit a compiled document. The types of questions it will include resemble the following.

- Create a 6x6 matrix. Calculate the degree and clustering coefficient of each node.
- Given the matrix below, tell me how many nodes are in the network. What is the most popular node? How many nodes have fewer than 5 connections? How many have more?
- Create a vector with 10 numbers. Create another vector with 10 numbers. Multiply them. Divide them. Add them together. Subtract the first vector from the second.

Final Exam. The final exam is similar to the midterm but with a network dataset I provide. It occurs from 8:00-11:00 on Wednesday, March 16. Example questions are:

- Plot the degree distribution of this email network.
- After Paul Revere, which four colonists were most influential based on their connections between people?
- What are the three most talked about topics in the tweets of influence operation accounts?

Participation. Since the only way to improve as a programmer is through practice, attendance is expected and will be recorded each class. You are allowed to miss **1** class during the quarter without penalty. Any further unexcused absences will result in 5 percentage point deductions from your participation grade. Absences are excused for medical reasons (with documentation) or family emergencies. Students are responsible for all missed work,

regardless of the reason for absence. It is also the absentee's responsibility to get all missing notes or materials.

Grade Distribution

Problem Sets	40%
Midterm Exam	20%
Final Exam	20%
Attendance	20%

A − is when the grades end in 0-2 and + is 8-9. I am more than happy if everyone earns an A.

Accommodations

If you wish to request an accommodation due to a disability, please contact the Center for Accessible Education as soon as possible. Documentation from them is required for me to provide those accommodations. The CAE is slow, so do not wait until a week or two before an exam to submit your paper work to them. To maintain equity in the classroom, I cannot give students exam accommodation without a letter from the CAE, even if the student is in the process of working with the CAE.

Datasets

The lab sessions and problem sets will make use of the following four datasets.

1. **MovieLense.** The small version of this dataset contains 100,000 ratings of 6,000 movies from 600 users. It is a bipartite network and will be used to learn community detection and edge weighting.
2. **State linked information operations.** Twitter publishes detailed information about accounts it removes from its service that identifies as being engaged in information operation campaigns. We will use tweets from accounts controlled by Iran and Russia to understand centrality, community detection, and visualization.
3. **Offline social networks.** We will use the replication data for "From Chatter to Action: How Social Networks Inform and Motivate in Rural Uganda" (Forthcoming, *British Journal of Political Science*) to understand centrality, connectedness, and contagion.
4. **US airports.** The final dataset details flights between United States of America airports in 2010. We will use it to understand contagion and connectedness.

Below are some archives of network data that are fun to explore.

- Stanford Network Analysis Project. Many large scale datasets primarily for an academic audience.
- Koblenz Network Collection. From here is where I found the airport dataset.

- Network Repository. Varied datasets and produces visuals from those data, though the documentation of each dataset often leaves much to be desired.
- UC Irvine Network Data Repository.
- Harvard Dataverse. This website contains replication material, often including data, for over 100,000 academic studies and is not well-indexed by Google. Many of the datasets contain network data; it is how I found the political brokers data, for example.

Course Outline

Week 1 - Introduction to Networks and R

Tuesday, January 4: This class introduces the course, introduces key network terms, and starts working in R.

Problem Set 1 released.

Read: Barábasi Chapter 1; tutorials from the R guide on the course website.

Thursday, January 6: This class explains packages, objects and their types, and calculation. RStudio (software) and RMarkdown (markup language) are introduced.

Read: Tutorials from the R guide on the course website.

Week 2 - R

Tuesday, January 11: Today, we learn how to load data, inspect it, and save data; plot using `ggplot2`; and subset and replace observations.

Read: Tutorials from the R guide on the course website; Luke: Chapter 2; Kolaczyk and Csárdi: Chapter 4.1-4.2, 4.3.2

Thursday, January 13: This class learns advanced concepts such as loops, functions, and merging datasets that we will need later in the quarter.

Read: Tutorials from the R guide on the course website; Luke: Chapter 2; Kolaczyk and Csárdi: Chapter 4.1-4.2, 4.3.2

Week 3: Influence

Tuesday, January 18: This class explains how to measure influence in a network and the different data requirements of these measures.

Problem Set 1 due.

Read: Barábasi Chapter 2.3-2.4, Chapter 7.1-7.5; Luke Chapter 7; Kolaczyk and Csárdi Chapter 4.2.1, 4.2.2; Using Metadata to Find Paul Revere

Thursday, January 20: R.

Problem Set 2 released.

Week 4: Connectedness

Tuesday, January 25: This class introduces the concepts of clustering and density to measure how connected a network is globally and locally.

Problem Set 2 due.

Read: Barábasi Chapter 2.9-2.10

Thursday, January 27: R.

Problem Set 3 released.

Week 5: Visualization

Tuesday, February 1: Principles of Visualization

Problem Set 3 due.

Read: Luke, Chapters 4 and 5; Kolaczyk and Csárdi, Chapter 3.

Thursday, February 3: R

I will release practice midterm problems. They will not be graded.

Week 6: Midterm Exam

Tuesday, February 8: Midterm review, including visualization.

Thursday, February 10: Midterm exam. The midterm resembles problem sets, except it is done in class and therefore timed.

Week 7: Pause

This week is a chance to catch your breath.

Tuesday, February 15: No class.

Thursday, February 17: Gabriel Rossman, a professor of sociology at UCLA, will present simulation and empirical research on cultural contagion.

Week 8: Likeness (Content Recommendation System)

Tuesday, February 22: A primary concern of network analysis is separating like nodes into like groups, such as books that are read together or words that co-occur frequently. This lecture introduces some methods for that task.

Read: Barábasi Chapter 9; Luke, Chapter 8; Kolaczyk and Csárdi, Chapter 4.4.

Thursday, February 24: R

Problem Set 4 released.

Week 9: Contagion

Tuesday, March 1: Networks provide a framework for understanding how outcomes such as behaviors or ideas spread between people or places.

Problem Set 4 due.

Read: Barábasi Chapter 10; Luke Chapter 13.

Thursday, March 3: R

Problem Set 5 released.

Week 10: Analyzing Twitter Data

Tuesday, March 8: Today's lecture explains how researchers use Twitter to study network questions.

Problem Set 5 due.

Read: Newman Chapter 3.

Thursday, March 10: R. Getting data from Twitter and conducting network analysis.

Final exam practice released. It will not be graded; answers will be provided.

Final Exam Week

Wednesday, March 16: The final exam occurs from 08:00-11:00 a.m. in Kaplan A51, southeast of Powell Library.