Network Metrics part 2 & Data Quality

CS 579 Online Social Network Analysis

Dr. Cindy Hood
9/11/25

Homework Assignments

- HW #2 assigned Due by midnight Friday 9/19
- Please contact TAs with questions on hw grading

Exams and Final Project Poster Presentation

- Exam 1 – Oct 9 in class
- Exam 2 – Dec 2 in class

- Final Project Poster Session – Dec 4 in class

- Online students (sections 2 and 3) will have remote options

Teaching Assistants

- Siva Krishna Golla
  - sgolla2@hawk.illinoistech.edu
  - Mondays 2-3pm on zoom
- Khush Dhiren Patel
  - kpatel210@hawk.illinoistech.edu
  - Wednesdays 11-12 online
- Aswith Sama
  - asama@hawk.illinoistech.edu
  - Thursdays 3-4pm on zoom

- **Not yet officially working, waiting for authorization (US govt)**

HW #2 due by midnight 9/19

- In this assignment you will create networks/graph models from 2 different datasets. You may use any tool/platform/language that you like. I have attached a few pages from the Elements of Network Science Book that illustrate basic use of Stata, R and Python. Section 2.3 ElementsofNetworkScience.pdf Download Section 2.3 ElementsofNetworkScience.pdf
- (1) The first dataset is Chicago Community Areas https://en.wikipedia.org/wiki/Community_areas_in_Chicago
- Links to an external site.
- Nodes = Community areas
- Edges = Shared physical boundary (i.e. adjacency) with other community area. Note that you may have to make some assumptions here since you are determining boundaries from the image of the map on the page cited above. State your assumptions.
- You will then create a labelled visualization of this graph and plot the degree distribution of the nodes. You will submit
- (1a) Input file with graph representation,
- (1b) Labelled visualization of network created.
- (1c) Plot of degree distribution.

HW #2 con't

- (2) The second dataset is the CS 579 Class Participant Data Social Network Data collection.xlsx
- Links to an external site.
- Nodes = Class Participants, entities in common
- Edges = Shared entity
- You will create a bipartite graph. Some data cleaning will be necessary. State and justify any assumptions you make during the data cleaning. You will then create a unimodal graph that is a projection of the bipartite graph.
- You will create labelled visualizations of both the bipartite and unimodal graphs and plot the degree distribution of the unimodal graph. You will submit
- (2a) Input file for the bipartite graph.
- (2b) Labelled visualization of bipartite graph.
- (2c) Description of method for projecting bipartite graph to unimodal graph including code.
- (2d) Labelled visualization of unimodal graph.
- (2e) Plot of degree distribution of unimodal graph.

HW #2 – con't

- (3) Compare the degree distributions of the graphs from the two different datasets. What is similar? What is different? Is this what you expected? Why or why not?
- (4) Provide the details of how you did this assignment. What tools did you use to complete the assignment? Why did you choose the tool? Provide citations and links to references and code used. If AI (e.g. ChatGPT, etc.) was used, please include a transcript of the exchange.
- The above can be submitted in a zipped folder that includes
- input files labelled as Input_file1, Input_file2
- pdf report of everything else


Chicago
Central
Far North Side
Far Southeast Side
Far Southwest Side
North Side
Northwest Side
South Side
Southwest Side
West Side
A. Burnside
B. Oakland
C. Montclare
O'Hare
Edison Park
Norwood Park
Jefferson Park
Forest Glen
North Park
Rogers Park
West Ridge
Edge-water
Lincoln Square
Albany Park
Uptown
Portage Park
Irving Park
Dunning
North Center
Lakeview
Avondale
(C)
Belmont Cragin
Hermosa
Logan Square
Lincoln Park
Austin
Humboldt Park
West Town
Near North
West Garf. Park
East Garfield Park
Near West Side
Loop
North Lawndale
Near South
Lower West Side
South Lawndale
Bridge-Port
Armor Square
Douglas
McKinley Park
(B)
Archer Heights
Brigton Park
New City
Fuller Park
Grand Blvd
Ken-Wood
Garfield Ridge
West Elsdon
Gage Park
Wash-ington Park
Hyde Park
Clearing
West Lawn
Chicago Lawn
West Engle-Wood
Engle-Wood
Woodlawn
Greater Grand Crossing
South Shore
Ashburn
Auburn Gresham
Chatham
Avalon Park
South Chicago
(A)
Calumet Heights
Wash-ington Heights
Beverly
Roseland
Pullman
East Side
South Deering
Mount Greenwood
Morgan Park
West Pullman
Riverdale
Hegewisch

HW #2 Data cleaning

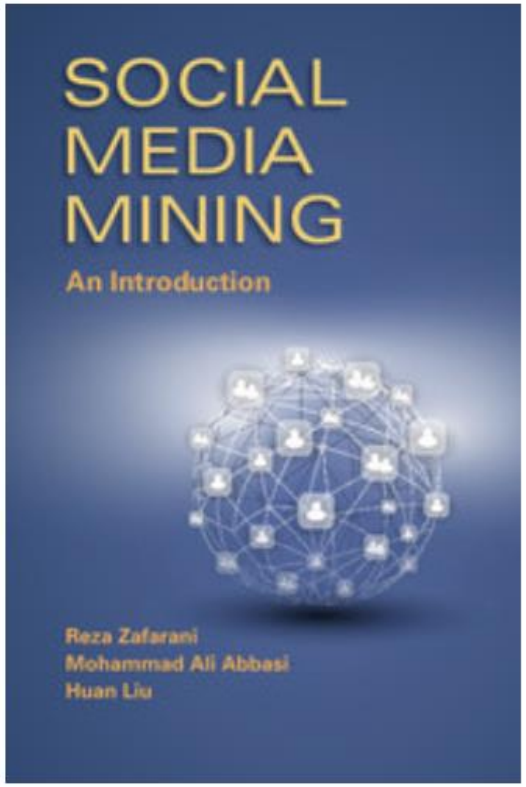
Data cleaning for HW #2

Updated 32 seconds ago by Cindy Hood

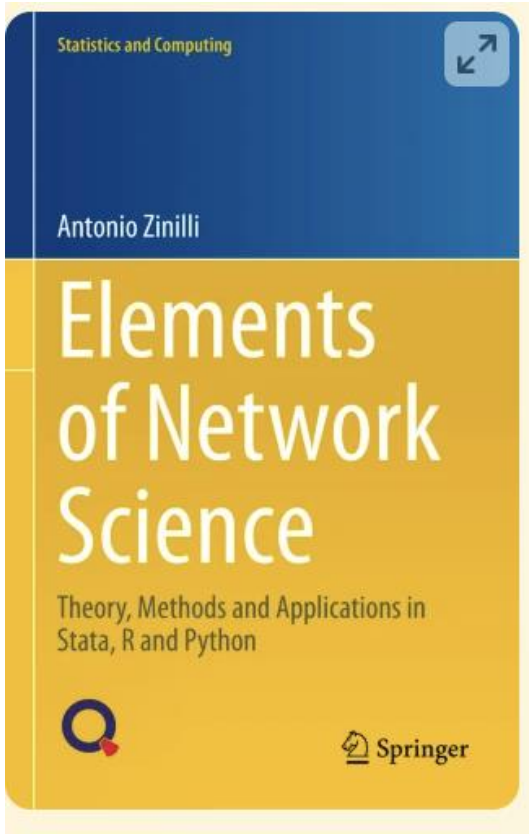
Please feel free to start threads to work on cleaning of student data for hw#2. Substantive contributions will be given extra credit.

hw2

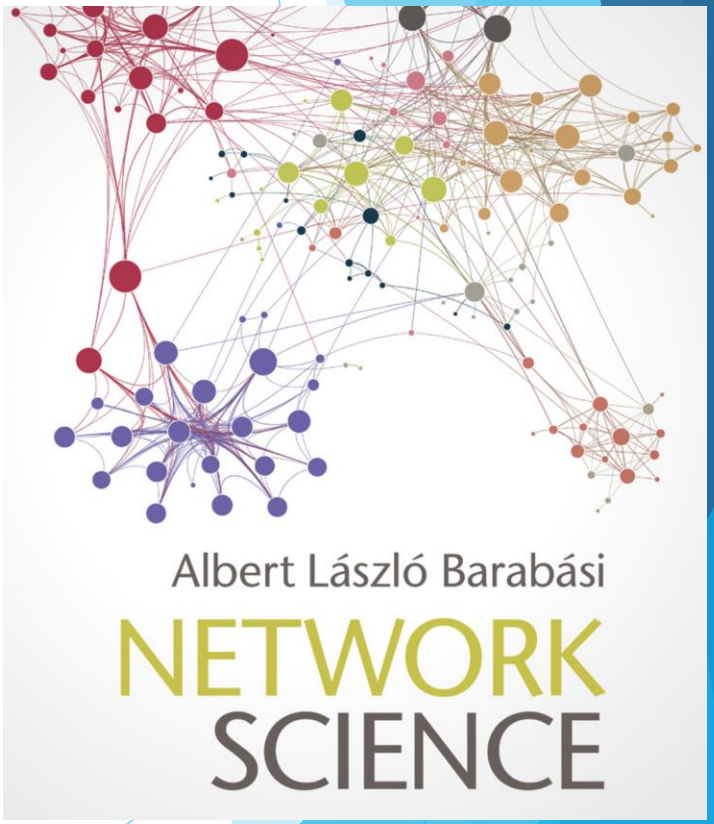
References



http://www.socialmediamining.info



https://link.springer.com/book/10.1007/978-3-031-84712-7



http://networksciencebook.com

Closeness Centrality

- Captures the average distance between a vertex and every other vertex in the network
- Can be considered a measure of how easily information can spread in a network
  - The lower the value, the “nearer” a vertex is to other vertices
    - Can spread information more easily
- Sometimes inverse is used for closeness centrality measure
  - Values between 0 and 1
  - In this case, the greater the value, the “nearer” a vertex is to other vertices

Closeness Centrality

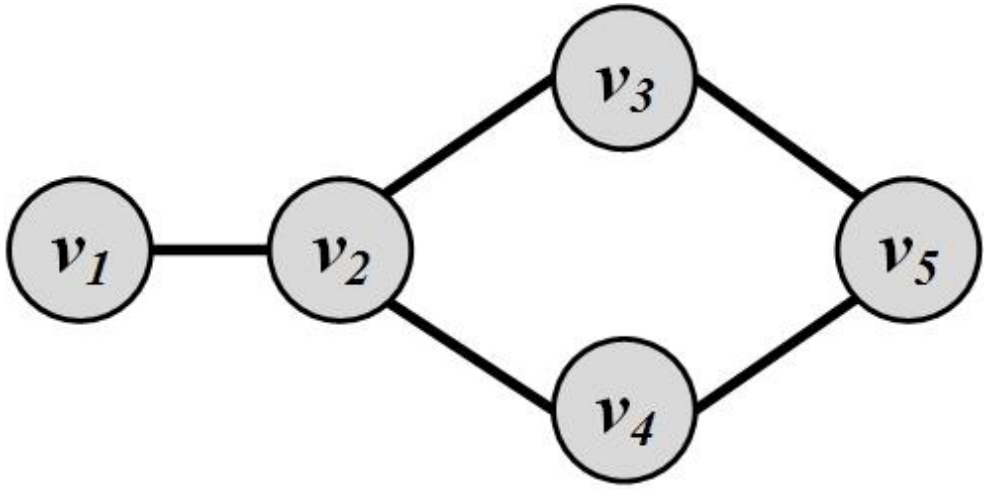
- The intuition is that influential/central nodes can quickly reach other nodes

- These nodes should have a smaller average shortest path length to others

Closeness centrality: $C_c(v_i) = \frac{1}{\bar{l}_{v_i}}$

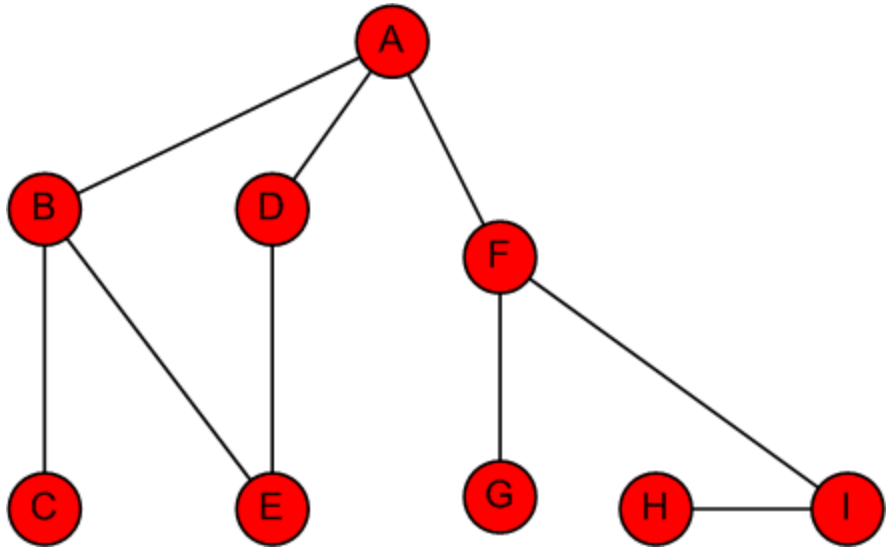
$$\bar{l}_{v_i} = \frac{1}{n-1} \sum_{v_j \neq v_i} l_{i,j}$$

Closeness Centrality: Example 1



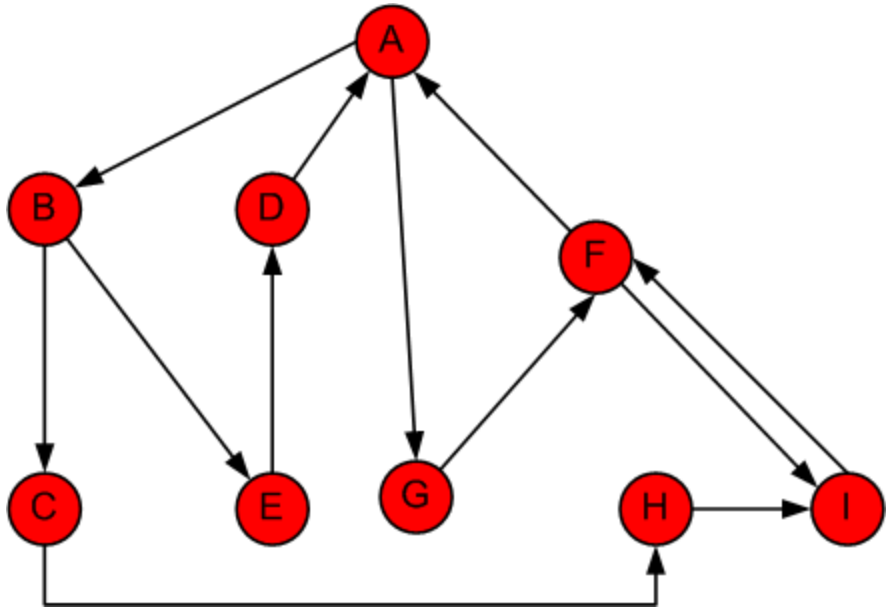
$$
\begin{aligned}
C_c(v_1) &= 1 \,/\, (\,(1+2+2+3)/4\,) = 0.5,\\
C_c(v_2) &= 1 \,/\, (\,(1+1+1+2)/4\,) = 0.8,\\
C_c(v_3) = \quad C_c(v_4) &= 1 \,/\, (\,(1+1+2+2)/4\,) = 0.66,\\
C_c(v_5) &= 1 \,/\, (\,(1+1+2+3)/4\,) = 0.57.
\end{aligned}
$$

Closeness Centrality: Example 2 (Undirected)



Node	A	B	C	D	E	F	G	H	I	D_Avg	Closeness Centrality	Rank
A	0	1	2	1	2	1	2	3	2	1.750	0.571	**1**
B	1	0	1	2	1	2	3	4	3	2.125	0.471	**3**
C	2	1	0	3	2	3	4	5	4	3.000	0.333	**8**
D	1	2	3	0	1	2	3	4	3	2.375	0.421	**4**
E	2	1	2	1	0	3	4	5	4	2.750	0.364	**7**
F	1	2	3	2	3	0	1	2	1	1.875	0.533	**2**
G	2	3	4	3	4	1	0	3	2	2.750	0.364	**7**
H	3	4	5	4	5	2	3	0	1	3.375	0.296	**9**
I	2	3	4	3	4	1	2	1	0	2.500	0.400	**5**

Closeness Centrality: Example 3 (Directed)



Node	A	B	C	D	E	F	G	H	I	D_Avg	Closeness Centrality	Rank
A	0	1	2	3	2	2	1	3	3	2.125	0.471	**1**
B	3	0	1	2	1	4	4	2	3	2.500	0.400	**2**
C	4	5	0	7	6	3	5	1	2	4.125	0.242	**9**
D	1	2	3	0	3	3	2	4	5	2.875	0.348	**3**
E	2	3	4	1	0	4	3	5	5	3.375	0.296	**6**
F	1	2	3	4	3	0	2	4	4	2.875	0.348	**4**
G	2	3	4	5	4	1	0	5	2	3.250	0.308	**5**
H	4	4	5	6	5	2	4	0	1	3.875	0.258	**8**
I	2	3	4	5	4	1	4	5	0	3.500	0.286	**7**

Reference

Harvard Business Review

Analytics And Data Science

Ensure High-Quality Data Powers Your AI

by Thomas C. Redman

August 12, 2024

https://hbr.org/2024/08/ensure-high-quality-data-powers-your-ai

Data Quality is everything

- Good data science + bad data = bad results
- AI models do not need to fail on a global scale to cause enormous damage to individuals, companies, societies
- Models frequently get things wrong, for example
  - Hallucinate
    - An AI hallucination is a response generated by AI which contains false or misleading information presented as fact. https://en.wikipedia.org/wiki/Hallucination_(artificial_intelligence)
  - Bias
    - Systemic errors that create unfair outcomes https://en.wikipedia.org/wiki/Algorithmic_bias
  - Drift
    - Degradation of performance due to changes in data or the relationships between input and output variables https://www.ibm.com/topics/model-drift

Getting data right

- Is it the right data to address the problem?
- Is the data right or correct?

Getting the right data for a problem/project

- Relevance and completeness
  - Data relevant to the problem/task at hand
  - Having as many relevant data attributes as possible
- Comprehensiveness and adequate representation
  - Data adequately covering the population of interest
  - Subpopulations of interest are covered
- Freedom from bias
  - Identifying and reducing historical bias
    - Can be very difficult
      - Many datasets contain historical biases
        - If these biases cannot be removed properly, AI should not be used

Getting the right data for a problem/project

- Timeliness
  - How new must the data be?
- Clear definition
  - Most AI efforts pull data together from different sources
    - Good understanding of sources and data they provide necessary to develop understanding of combined data
      - Clear definitions of sources, data attributes and measurement units
- Appropriate exclusions
  - Certain data should be excluded for legal, regulatory, ethical and intellectual property considerations
    - Ex/ using zip codes can be a proxy for race in loan decisions
    - Must avoid violating laws stipulating how personally identifiabe information (PII) may be used
    - AI models trained on public sources may violate intellectual property rights

Getting the data right

- Accuracy
  - Data values must be correct (i.e. reflect reality)
    - How to assess this?
      - Structured data sets
      - Unstructured data sets (e.g. documents)
- Absence of duplicates
  - Duplicates can skew results
- Consistent identifiers
  - Ex/ John Smith same as J. Smith, J.E. Smith?
- Correct labeling

Limitations of model

- Data quality is not only for training data
- What happens if you feed a well-trained model bad data?
  - Bad/unpredictable results

- Need to understand the limitation of the model
- Need to ensure high-quality data inputs

Extra Credit #3 – You may work in pairs

- Find a recently publicized AI failure (from 2025)
- Write a summary of the failure
- Are the causes of the AI failure mentioned in the article?
- Was quality of the data the cause or one of the causes of the failure?
- If you were hired to do a root cause analysis, what would your approach be?

- Submit written copy to me at the end of class, be sure to put name(s) on it

- Online students submit to canvas