Week 04

Node-level metrics

Tuesday, September 14

INFO 5613: Network Science

Brian Keegan
brian.keegan@colorado.edu

University of Colorado Boulder

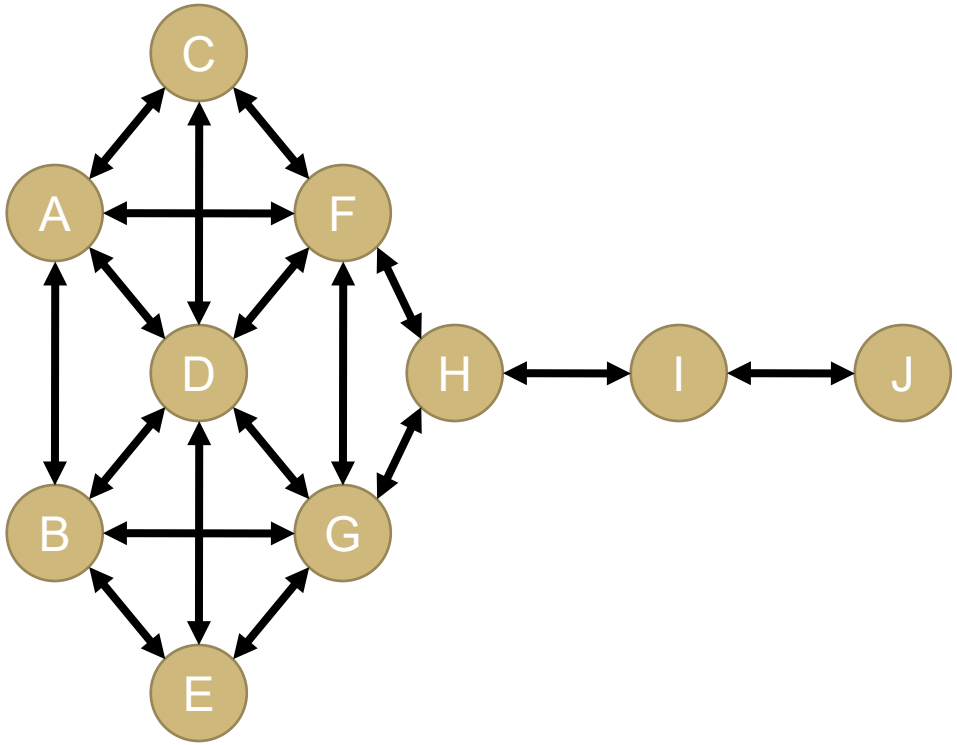
Agenda

- Lecture: Centrality metrics
- Activity: Measuring centrality with networkx and visualizing Boulder’s local politician network

Lecture

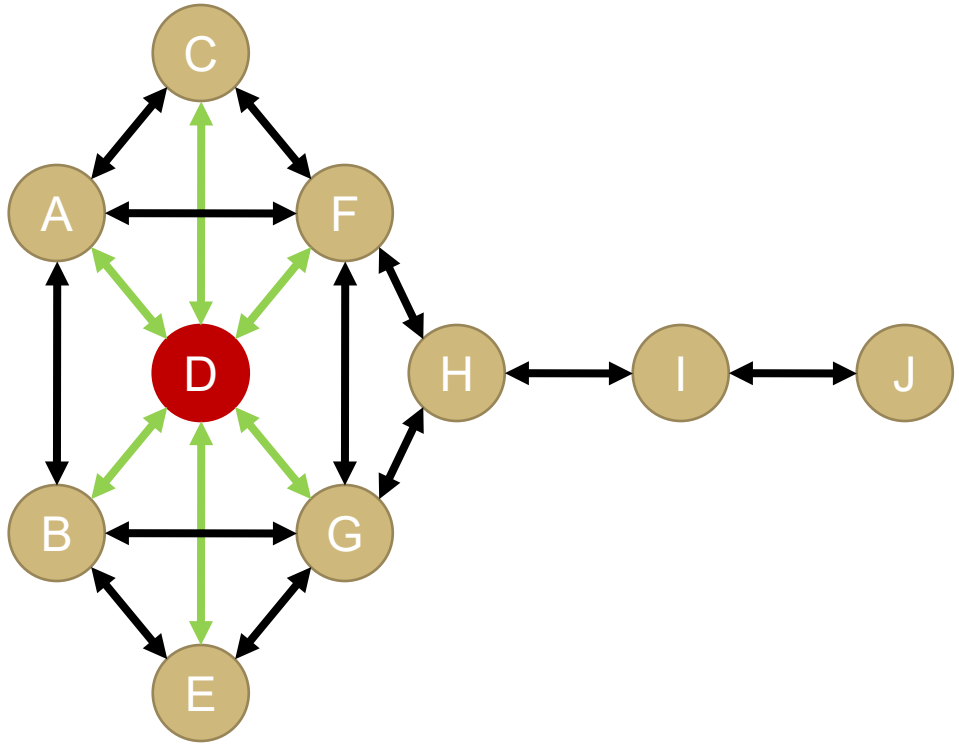
Centrality

- Which actor is the ***most important*** in this network?
- This is the “Krackhardt Kite”



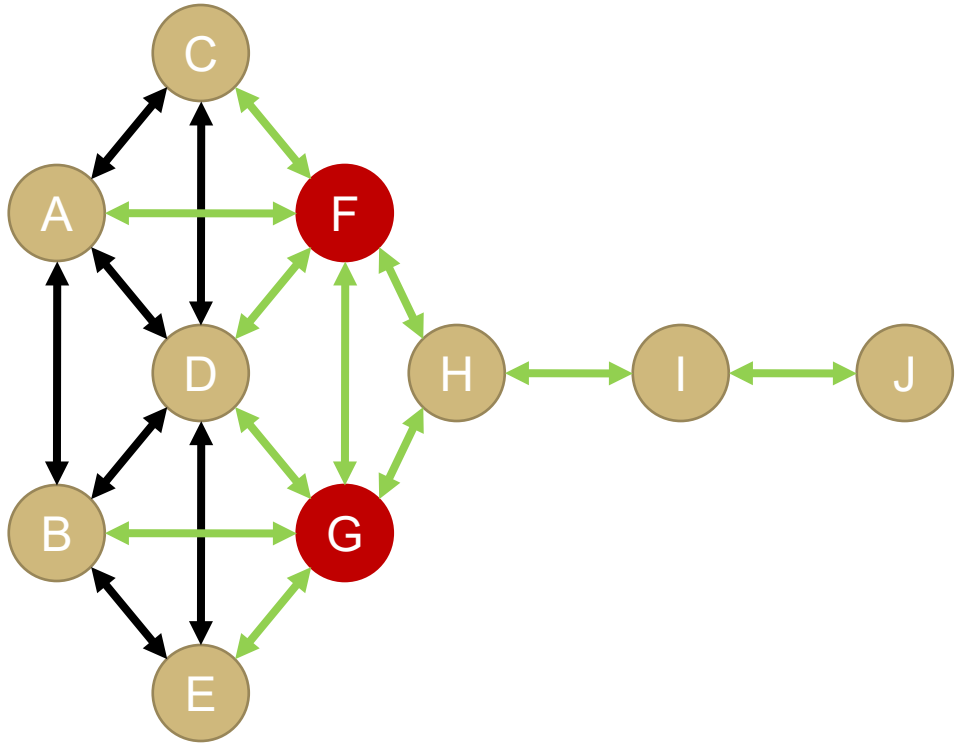
Centrality – Degree

- Degree: total number of links with other actors
  - **In-degree**: Directional links to actor from other actors
  - **Out-degree**: Directional links from actor to other actors
- “Popularity”
- Very fast to compute



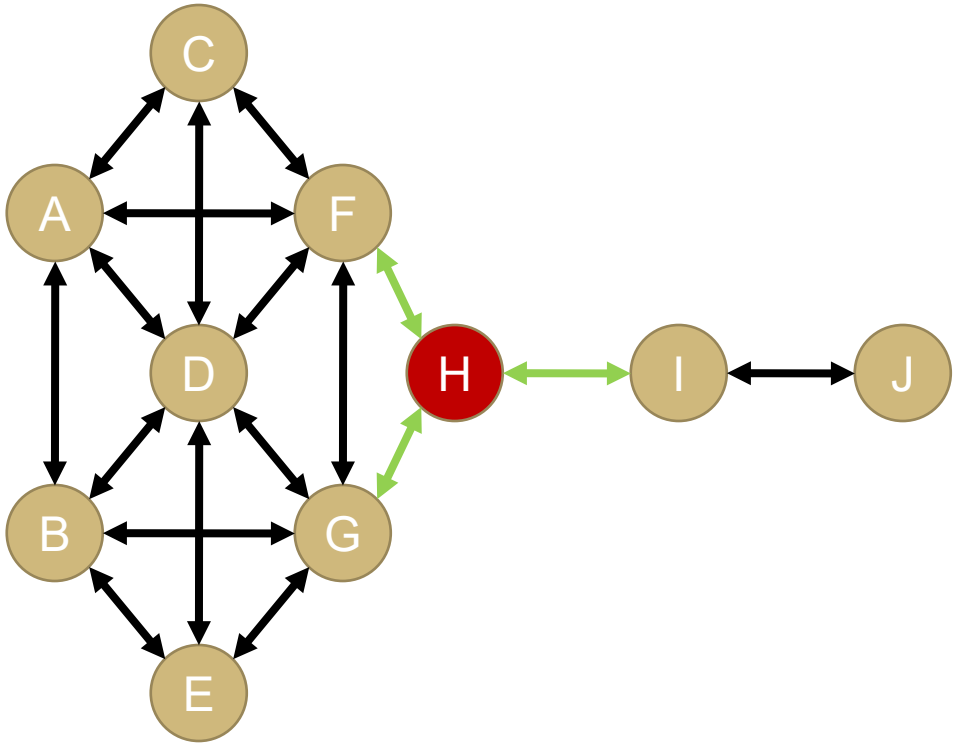
Centrality – Closeness

- How easily one actor can reach rest of network
- Actor with shortest average path length
  - Non-shortest path: F → D → G → H → I
  - Shortest path: F → H → I
- “Pulse-taker”
- Slow to compute and may be undefined
  - Need to enumerate all the shortest paths
  - Some kinds of networks may have undefined paths, meaning closeness can’t be computed for many nodes



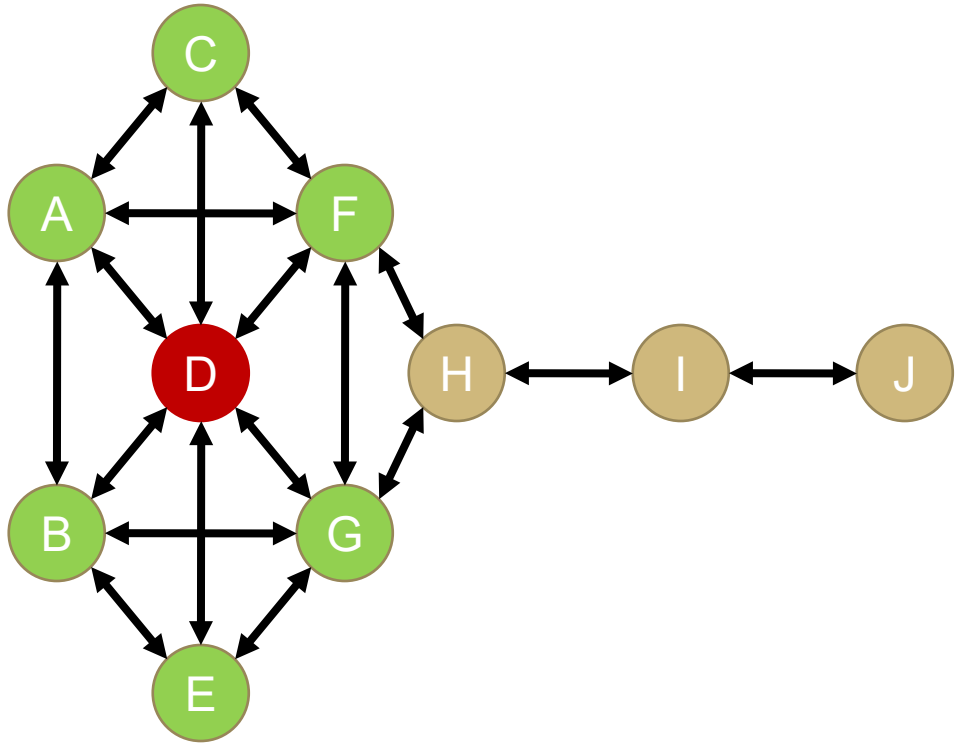
Centrality – Betweenness

- How much an actor lies between distinct groups
- Number of shortest paths passing through actor
  - **A → J**: A → F → H → I → J
  - **E → J**: E → G → H → I → J
  - **A → G**: A → {F,D,G} → G
- "Broker"
- Slow to compute and may be undefined
  - Need to enumerate all the shortest paths
  - Some kinds of networks may have undefined paths, meaning betweenness can't be computed for many nodes



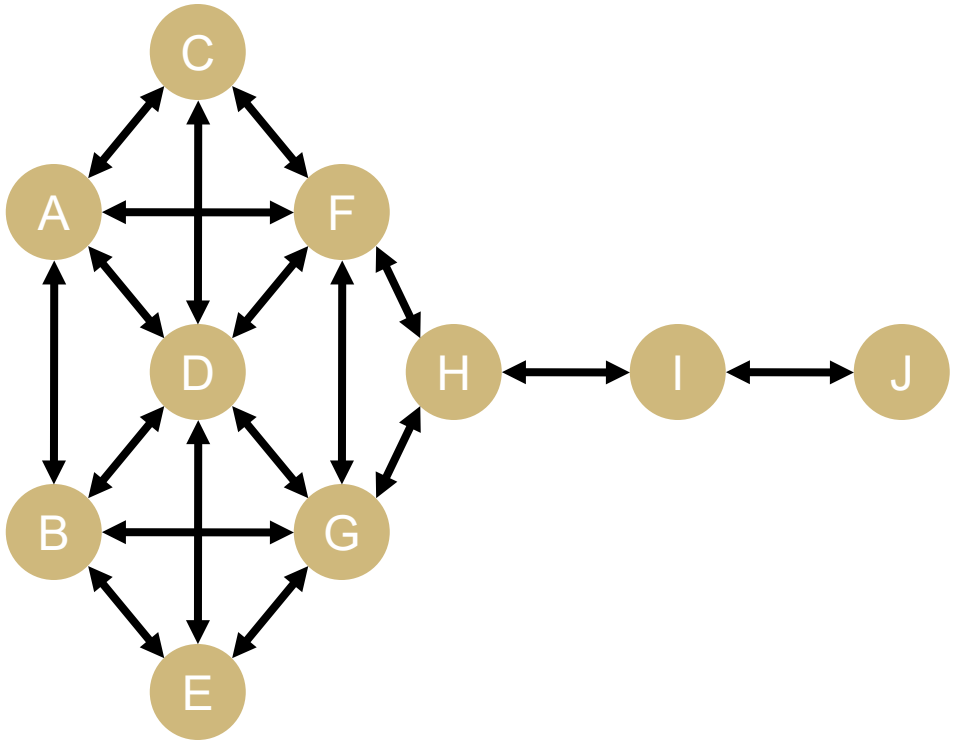
Centrality – Eigenvector / PageRank

- Centrality is a function of neighbor's centrality
- Nodes are central if their neighbors are well-connected
- "Influencer"
- Medium time to compute



Clustering

- How many of neighbors are also neighbors?
  - C's neighbors: A, D, F
    - A neighbors with D? Yes.
    - A neighbors with F? Yes.
    - C's clustering = 3 existing ties / 3 possible ties = 1.0
  - I's neighbors: H, J
    - Neighbors? No.
    - I's clustering = 0 existing ties / 1 possible ties = 0.0
- "Embeddedness"
- Slow to compute

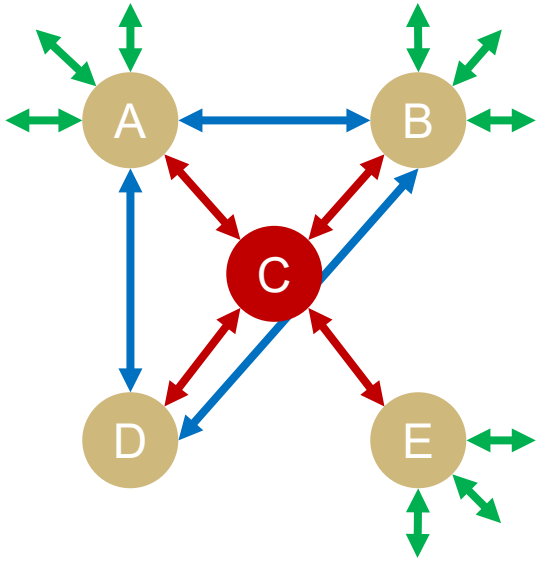


Activity

Online political networks of Boulder

- The Adamic & Glance (2006) political blogs network has a self-evident structure → this is rare!
- Most real-world networks will be “hairballs” with no discernible structure

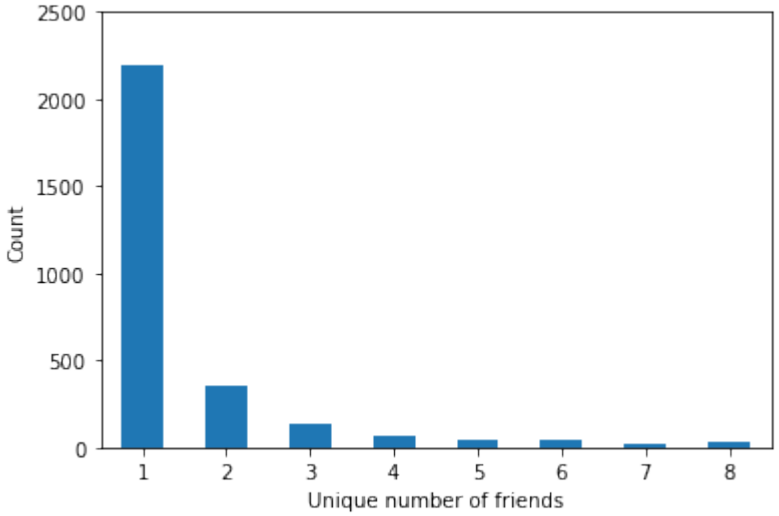
- Who has the most social capital in the online social network of Boulder politics?
  - 6 of 9 current Boulder City Council members has Twitter accounts
  - 5 of 10 candidates for Boulder City Council (don’t forget to vote!) have Twitter accounts
  - **Data considerations are influencing network structure!**



- Build out the 1.5-step ego “friend” network of Boulder’s local politicians
  - People’s choice of who to follow are more agentic than who follows them
  - Twitter is used by politicians for messaging and mobilization
  - Defining network boundaries: candidates/politicians + immediate friends

Data collection: Defining seeds

- Identify the Twitter accounts of six current members and five candidates → “seed accounts”
- Retrieve these 11 accounts’ friends using the Twitter API’s GET friends/ids endpoint
  - Members follow 2,141 unique friends, candidates have 1,158, combined seed accounts have 2,886 unique friends
  - Most of these friends are only followed by 1 seed
- 692 friend accounts are followed by 2 or more seeds
  - **Data considerations are influencing network structure!**
  - 28 accounts are followed by 8 of the seeds
    - This includes other seeds

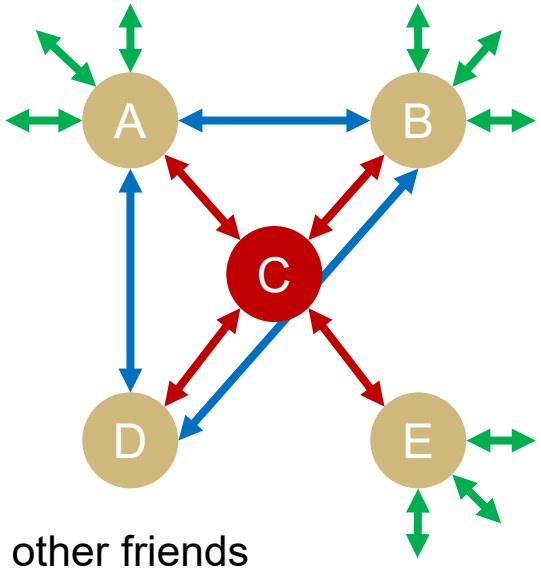


Data collection: Seeds' friends

- Get the information about these 692 "double-nominated friends"
  - Account ID, screen name, location, description, protected, followers, friends, created, verified, statuses
- Getting friends is computationally expensive!
  - Can only retrieve 5,000 friends per request, only 15 requests per 15 minutes
- Some of these friends have extremely high friend counts
  - @BarackObama has 589k friends, @amyklobuchar has 134k friends, @Trish_Zornio has 98k friends
  - We skip their friends → **Data considerations are influencing network structure!**
  - Is this defensible? More for Obama and Klobuchar, less for Zornio
- Some friends still have high friend counts but we will grab their data
  - @MitchellByars 16k, @JaredPolis 6k, @RogerPielkeJr 6k
  - Intermediate API scraping requiring handling pagination/cursors and rate limits

Aside on ego-network terminology

- Ego networks are the networks surrounding a focal node “ego” to “alters”
  - Relationships to/from ego and alters
- Enumerating ego network levels
  - 0-step → Seeds themselves, no relationships
  - 1-step → **Seeds and their friends, relationships from seeds to friends**
  - 2-step → 1-step + friends-of-friends, 1-step + friends to friends-of-friends
- What’s a “1.5-step” ego network?
  - **Some friends-of-friends are seeds and friends**: links from friends back to seeds or to other friends
  - **Many friends-of-friends are neither seeds nor friends**: links from friends to other interests
  - Keep the friends-of-friends relationships involving seeds and friends, throw the rest away
    - **Data considerations are influencing network structure!**



Data collection: Friends-of-friends

- For each of the 692 “double-nominated” friends, get the friends-of-friends
  - Twitter API’s GET friends/id rate limits to ~1 account/min → ~700 accounts is ~700 minutes → **>11 hours**
  - 8 accounts are private → 684 publicly-available accounts remaining
    - **Data considerations are influencing network structure!**
- Friends-of-friends are now ~117k unique accounts
  - Includes seeds, friends, and friends-of-friends and all their relationships with each other
- Retrieve meta-data about all 117k FoF from GET users/lookup API endpoint
  - Account ID, screen name, location, description, protected, followers, friends, created, verified, statuses
  - Can retrieve 100 users/request, 900 requests/15 min = 90k users/15min

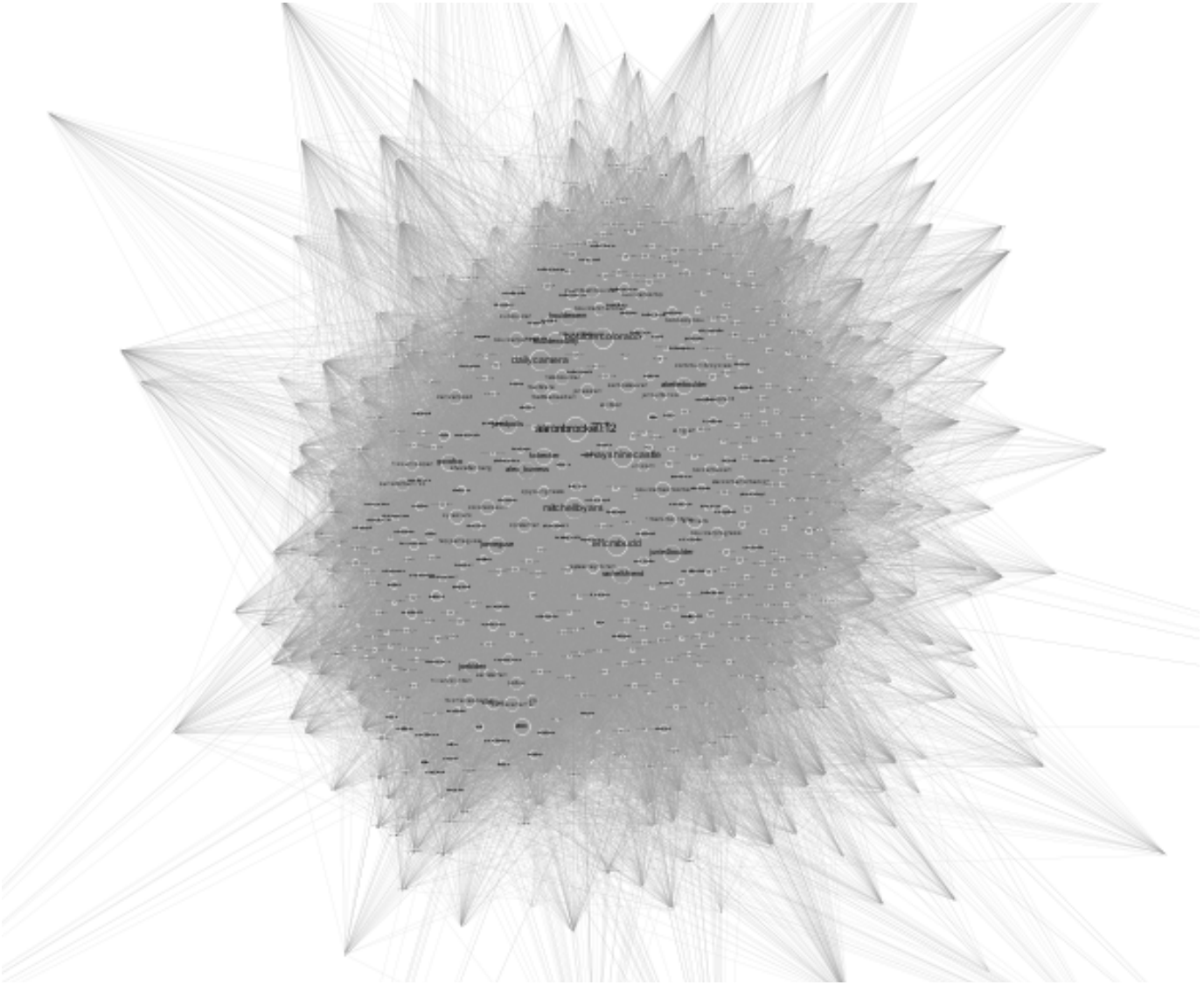
Network construction

- Friend-of-friend edgelist
  - Seeds + friends + friends-of-friends and their friend relationships
  - For each seed, add all their relationships to the network
  - For each friend, add all their relationships to the network IFF it is with a seed or a friend

- This is a directed network: if X follows Y, this doesn't mean that Y follows X

- Network has 683 nodes and 63,697 edges
  - This is an unusually dense social network!
  - 13.6% of possible ties exist → many social networks closer to 1%
  - Potential reasons why?

Network science IRL

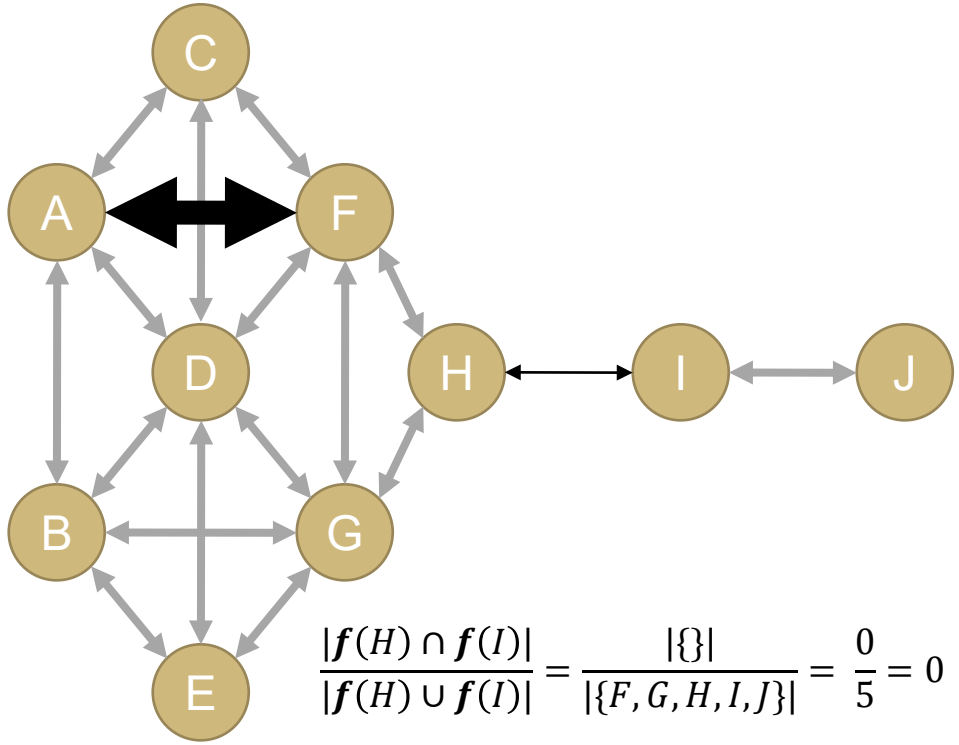


Imputing edge strength

- All the edges are binary → friend or not friend
- But we can intuit that some friend ties are “stronger” because of their “embeddedness”
  - Pressure from neighbors to have a tie
  - Conformity, similarity, exchange, *etc*.
- Embeddedness → Fraction of neighbors also connected
  - The (C,D) relationship is highly embedded
  - The (H,I) relationship is weakly embedded
- Compute edge-wise Jaccard index : size of intersection of neighbors divided by size of union of neighbors
  - 0 (no neighbors in common) → 1 (all neighbors in common)
  - **Data considerations are influencing network structure!**

$$\frac{|\boldsymbol{f}(A) \cap \boldsymbol{f}(F)|}{|\boldsymbol{f}(A) \cup \boldsymbol{f}(F)|} = \frac{|\{B,C,D,F\} \cap \{A,C,D,G,H\}|}{|\{B,C,D,F\} \cup \{A,C,D,G,H\}|} = \frac{|\{C,D\}|}{|\{A,B,C,D,F,G,H\}|}$$

$$\frac{|\{C,D\}|}{|\{A,B,C,D,F,G,H\}|} = \frac{2}{7} = 0.285$$

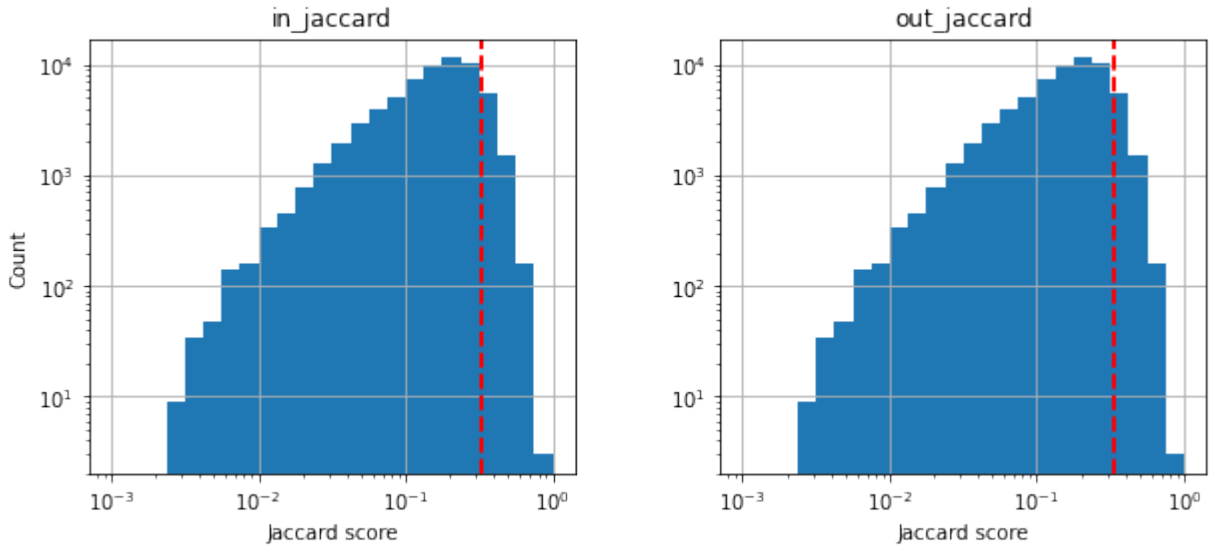


$$\frac{|\boldsymbol{f}(H) \cap \boldsymbol{f}(I)|}{|\boldsymbol{f}(H) \cup \boldsymbol{f}(I)|} = \frac{|\{\}|}{|\{F,G,H,I,J\}|} = \frac{0}{5} = 0$$

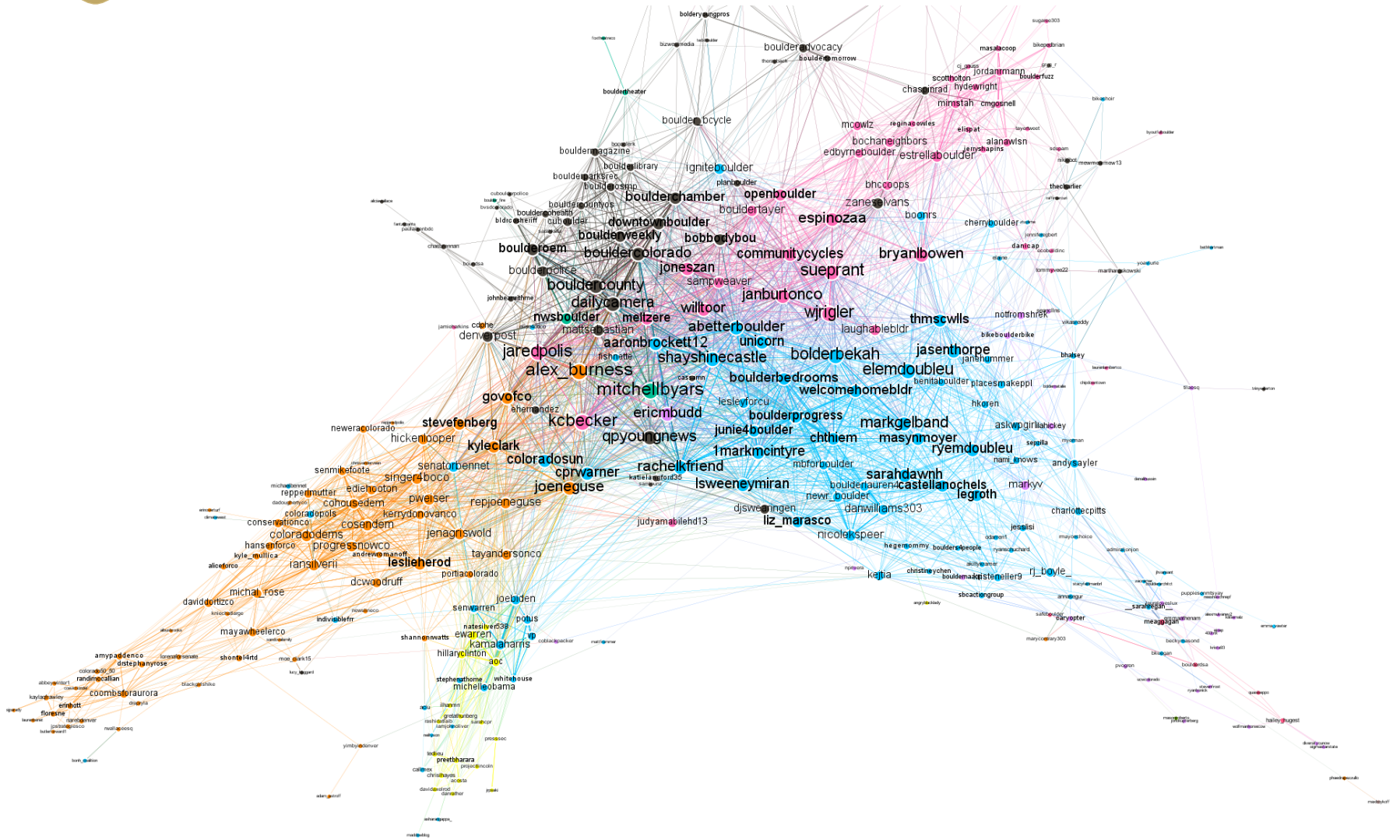
https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.link_prediction.jaccard_coefficient.html

Edge strength

- In a directed network, Jaccard indices might differ based on in versus out ties
  - Out-embeddedness → X following Y is strong if X and Y follow the same people
  - In-embeddedness → X following Y is strong if X and Y are followed by the same people
- Threshold the network at an arbitrary value like 0.33 to keep strongest ties
  - 346 nodes and 6,086 edges → we threw away 50% of nodes and 90% of edges!
  - **Data considerations are influencing network structure!**



But it looks nice!



https://twitter.com/bkeegan/status/1437519577773080577

Is the visualized network valuable? Valid?

- Data considerations structured the network at every step
  - Up to a half of seeds don't have active Twitter accounts
  - Friends needed 2 nominees from seeds
  - Skipping high-friend friends
  - Private accounts excluded
  - Imputing tie strengths using one definition
  - Arbitrary thresholding network to strongest ties
  - Visualization design choices for layout, colors, size, *etc*.

- Download the "boulder_politics.gexf" file from Canvas and open in Gephi

Who has the most social capital?

- Use the different centrality metrics in Gephi to visualize prominent nodes
- In-degree centrality → Friended by many others
- Out-degree centrality → Friends with many others
  - Gephi: Available by default
- Closeness centrality → Close to the rest of the network (trivial in a 1.5-step ego network?)
- Betweenness centrality → Brokering disconnected parts of the network
  - Gephi: Statistics tab > Network Diameter
- PageRank centrality → Connected to influential other nodes
  - Gephi: Statistics tab > PageRank or Statistics > Eigenvector Centrality
- Clustering → Share neighbors in common
  - Gephi: Statistics tab > Avg. Clustering Coefficient

Measure centrality in networkx

- Download “Class 07 – Node-level metrics.ipynb” and launch

- Computing node-level centrality metrics, handling normalized values, inspecting top results
- Correlations across centrality metrics

- Optional exercises applying to boulder_politics data

Module assignment 1

Module Assignment 1

- Due Friday, September 17 by midnight on Canvas

- Find an article around your research interests employing network theories/concepts/methods
- How did you encounter this paper, it or its authors' influence in your research area, *etc*.
- Summarize the key relational constructs used by the paper: what are nodes and links?
- Write a critique using the concepts about data collection, validity, ethics, visual rhetorics
- Outline how the research could've/should've been done if you were driving it

Next class

Next class

- Readings
  - Freeman, L. C. (1978). Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3):215–239
  - Borgatti, S. P., Jones, C., and Everett, M. G. (1998). Network measures of social capital. *Connections*, 21(2):27–36
  - Lin, N. (1999). Building a network theory of social capital. *Connections*, 24(1):28–51
- Don’t let the math be a blocker! Intuitions, abstractions, heuristics more valuable at this stage
- On Thursday, wrap up on notebook/Gephi exercises and discuss readings
- Add terms, phrases, *etc*. to discuss together to the weekly Group Discussion document

Reading responses

- Answer any, all, or none of them if there's something else you'd like to contribute!

- What centrality metrics do you find hard to intuitively understand or apply? What centrality metrics seem more applicable or promising to your research interests?
- What is an example of social capital from your research interests, current events, or history? Is this kind of social capital at the level of individuals or groups, internally-focused or externally? Which of the boxes from Borgatti, *et al*. (1998) contains network measures most relevant to it?
- Social capital as access to resources through social relation is often compared to "classical" forms of financial, human, cultural capital. Information exchange, influence, credentials, reinforcements (pg. 31) differentiate its mechanisms from others. How do you see these mechanisms at work in your research, current events, or historical cases?
- What is it like to read Lin (1999)'s account of "cyber-networks" contributing to a rise of social capital from 2021? What did it get right or wrong and why?