Network Analysis of LinkedIn Connections

Nathan Bick, Angela Threadgill, Pamela Katali

Data

In this project, we performed a network analysis of LinkedIn connections between the group project members. The data obtained contains each of our connections and the company they *currently* work at.

We generated an edge list for a network with edges present between two nodes if they each are currently employed at the same company. Thus, individuals who are not employed at the same company as another individual in the data will not be included in the network.

Network	**Nodes**	**Link**	**Type**
LinkedIn social network	People	Company	Undirected

Challenges

Originally, we were interested in the network where nodes contained edges purely if the node (an individual) was connected to another node (another individual) in the network, regardless of whether they were actually connected by a common element, such as a common employer (that is, simply a Linkedin connection).

This posed a problem for the centrality measures.

- The project member degrees would be meaningfully larger than the degree of the connections
- To combat that, we would have needed to obtain our connections’ connections, but obtaining those now posed additional challenges that were difficult to overcome in the time period allotted for the project

Research questions

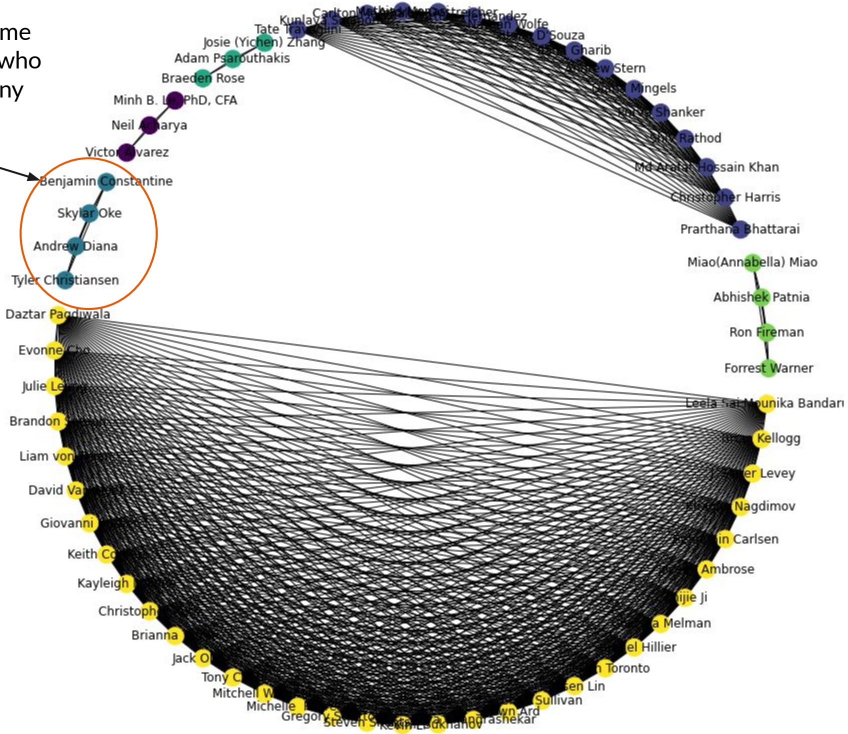
Given our network is now the connection of all individuals within a company (where each company would be its own component), our research questions have been revised.

- Of individuals that we are connected with, which companies are they most likely to *currently* work with?
- What do the different centrality measures tell us about the nodes within each of the components? Do they give us any comparative information across components?
- What is the distance between people and their connections?
- Are there clear communities in the data? If so, to what extent are these communities different?

Network

- **(N, L)** - (324, 5005)
- Average path length (**<l>**) **-** 1 for all components
- **C(k):**
- Average degree - 31
- Number of communities(**C**) - 67
  - Average size of community - 4.8358
- Clustering coefficient - 0.7098

Since these four nodes are the same color, they represent individuals who are employed at the same company



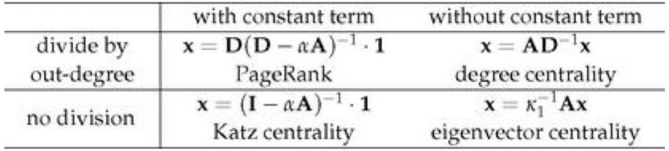
* the visualization above contains a subset of nodes for illustration and readability. We randomly selected components to represent each approximate “order of magnitude” of size, in terms of node count

Comparison to Random Model Network

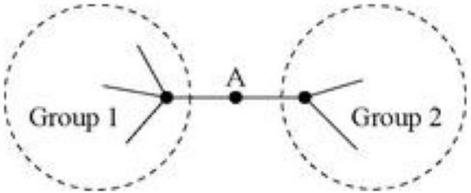
For several of the metrics that summarize the network, these are most useful when used to compare networks. Therefore, we introduced a G(n,m) model random network that resembles our primary network to serve as a point of comparison. This random model has the same number of nodes and edges as our linkedin network and the edges were randomly assigned.

For example, we can clearly contrast the highly assorted components of our network to the random network.

Centrality



	with constant term	without constant term
divide by out-degree	$\mathbf{x} = \mathbf{D}(\mathbf{D} - \alpha\mathbf{A})^{-1} \cdot \mathbf{1}$ PageRank	$\mathbf{x} = \mathbf{A}\mathbf{D}^{-1}\mathbf{x}$ degree centrality
no division	$\mathbf{x} = (\mathbf{I} - \alpha\mathbf{A})^{-1} \cdot \mathbf{1}$ Katz centrality	$\mathbf{x} = \kappa_1^{-1}\mathbf{A}\mathbf{x}$ eigenvector centrality



- Random Graph model
  - Nodes - 324
  - Edges - 5005
  - Edges randomly assigned up to m
- Mean of Degree centrality
  - G - 0.09565034590834386
  - random - 0.0913312693498452
  - Noah Murrell
- Mean of Eigenvector centrality
  - G - 0.027950130367362283
  - random - 0.054551356234774764
  - Noah Murrell
- Katz Centrality
  - Did not converge.

- Closeness centrality
  - G - 0.09565034590834386
  - random - 0.5076178817739052
  - Noah Murrell

  Within the cliques closeness is very high, but due to the separate components, the overall closeness is low.
- Mean Betweenness centrality
  - G - 0.0
  - random - 0.003024279463459254

  Betweenness is 0 because there are disconnected components and all shortest paths are length 1
- Clustering coefficient
  - G - 0.7098765432098766
  - random - 0.09088144735273472
  - Forest Warner

  Clustering is an interesting case for us because each of our companies represents highly clustered cliques and components, shown by high coefficient.

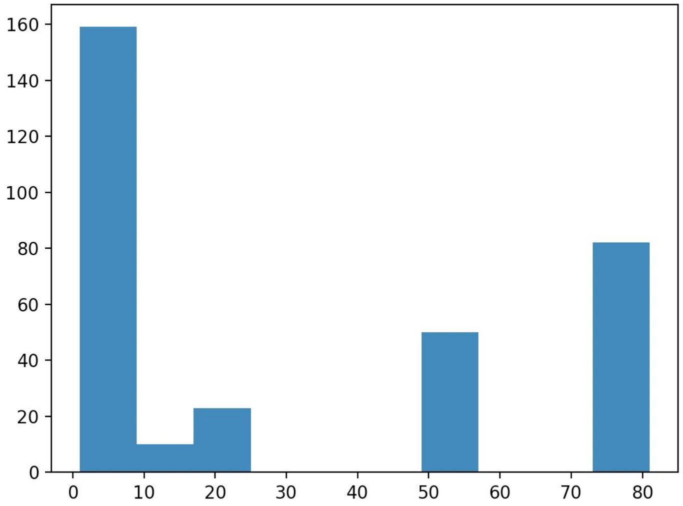
Other Metrics

- Assortativity
  - Degree assortativity = 1
  - Modularity: 0.4971
    - In our network, there are separate components for each company. These components are not connected to each other. This is not the typical case of measuring the modularity (as in the case study of segregation in high schools)

All shortest paths between any two nodes, if they are connected at all, are 1.

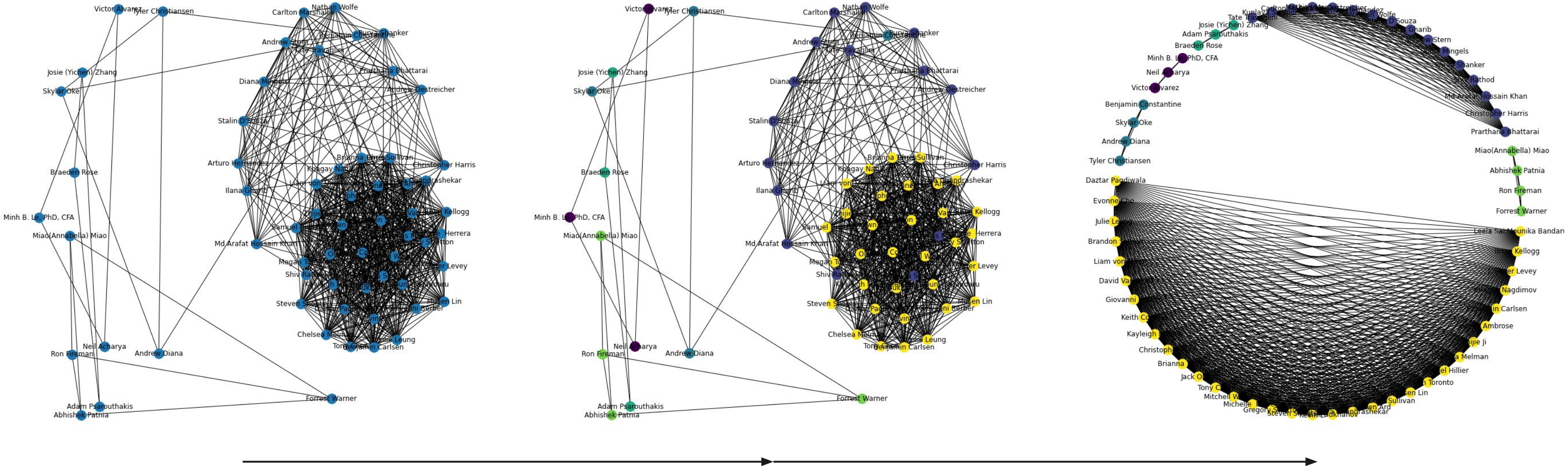
What is diameter? Max shortest path, so 1.

Degree Distribution - P_k



- Largest component has 82 nodes
- Most of the nodes have degree between 1 and 10 or 70 and 80
- There are larger and smaller groups of nodes in the network

Visualization evolution



Color-code nodes by company

Position nodes using the Kamada-Kawai path-length cost-function

Discussion

What was your expectation?

Our expectation was that, given we would most likely be connected to individuals of similar roles and interests, there may be certain companies that our connections would most likely be currently employed at. Additionally, though we would most likely expect the number of nodes (individuals) employed at start-ups to be smaller than that of a fortune 500 company, for example, that expectation may not necessarily be true if one of us works at a startup and is more likely to be connected to individuals currently employed at startups.

How do the results compare to your expectations?

We were able to view our company network from a different perspective since we generally think of our connections by how we knew them, not necessarily by the company that they *currently* work for. We were surprised by the size of networks at certain companies, since it seems as though people we worked with historically are congregating at certain companies (such as Capital One, for example).

Discussion (cont’d)

What did you learn from each quantity you measured?

The average path length wasn’t as useful for this network since every node within a component was connected to every other node. And so this metric will be “1” for every component. However, we were able to determine a couple of different individuals that stood out w.r.t the centrality measures and clustering coefficient.

Importance of different definitions of edge or node

It became clear that the same set of data can be transformed into many alternate definitions of networks, indicating the “art” of network analysis in addition to the “science” of the graph theory etc. As stated the network defined with edges being linkedin connections was not possible for us due to lack of neighbor’s connections, but certainly the results would be quite different and interesting.

A bipartite version of our network would be very useful as well, given more time to expand the analysis.

Closing thoughts

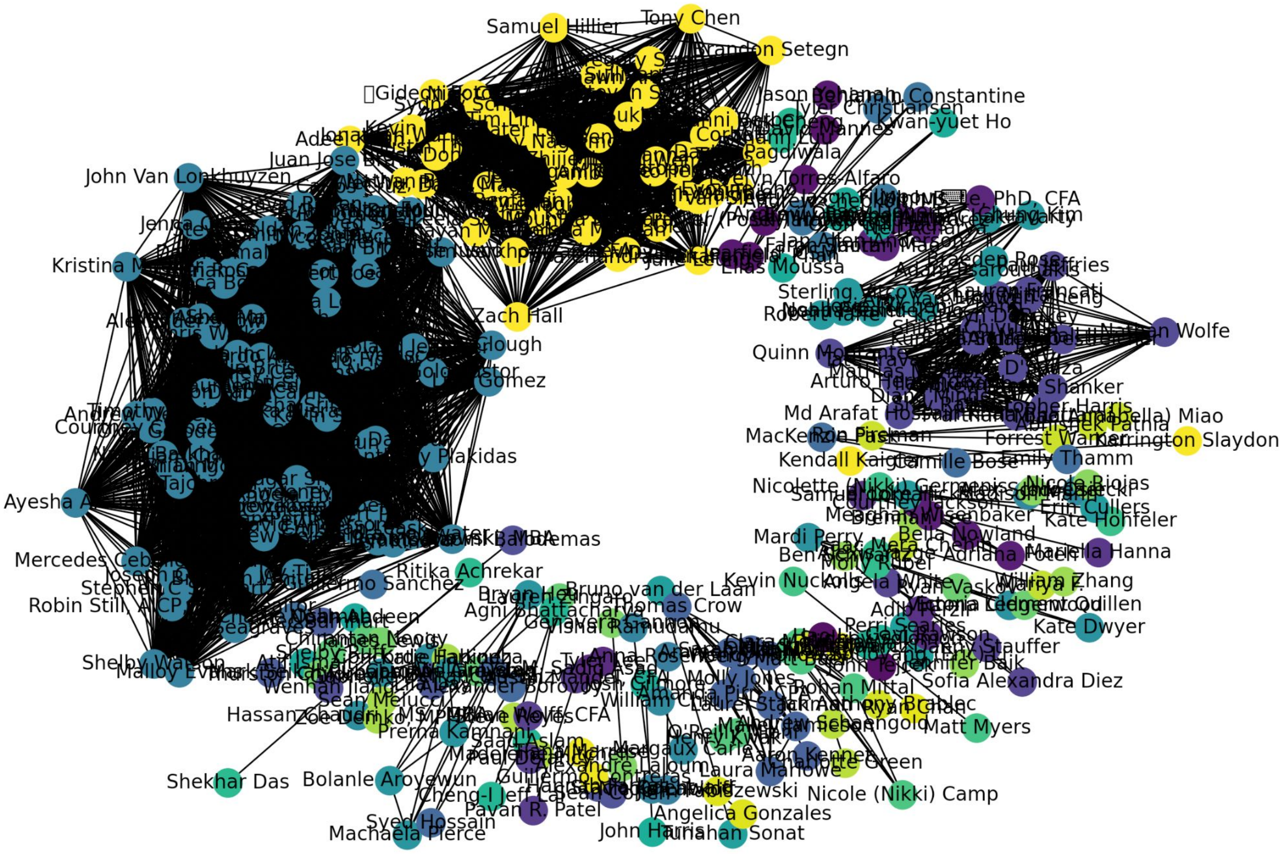
Data can be analyzed in a variety of ways. With additional project and compute time, network analysis of our connections’ connections (and potentially the connections of those connections!) would provide answers to the following questions.

- Which individuals are the most influential across centrality measures?
- How often are individuals connected to roles outside of their direct job family?
- Which individuals are most reachable?
- Which roles are most popular?
- Which roles are associated with the most influential individuals?

Appendix

Full network

To the right, we see a "raw" presentation of the full network. We see there are two large components and several smaller ones. As previously mentioned, we focused on a few components for clarity in the main visualization.


Samuel Hillier
Tony Chen
Brandon Setegn
Hannah Constantine
Tyler Christiansen
Kwan-yuet Ho
Juan Jose
John Van Lonkhuyzen
Kristina
Zach Hall
Pagdiwala
Torres-Alfaro
PhD, CFA
Elias Moussa
Sterling
Wolfe
Quinn
Arturo
Shanker
Md Arafat
Miao
MacKenzie Parkman
Forrest Walker
Kerrington Slaydon
Camille Bose
Emily Thamm
Nicole Riojas
Erin Cullers
Kate Hohfeler
Mardi Perry
Bella Nowland
Mariella Hanna
Molly Ruber
Kevin Nuckolls
William Zhang
Kate Dwyer
Sofia Alexandra Diez
Matt Myers
Nicole (Nikki) Camp
Angelica Gonzales
John Harris
Machaela Pierce
Syed Hossain
Pavan R. Patel
Shekhar Das
Bolanle Aroyewun
Ritika Achrekar
Guillermo Sanchez
Bruno van der Laan
Thomas Crow
Agni Bhattacharya
Prerna Kamnani
Saad Aslam
Laura Marlowe
Charlotte Green
Alexander Borovoy
Sean Melucci
Rohan Mittal
Ayesha A
Robin Still, AICP
Shelby Watson
Courtney
Plakidas
Gomez