Social Network Analysis

Centrality I

Robin Burke

DePaul University

Chicago, IL

Project

- Getting started on the project
  - Form groups (ASAP)
  - Proposal (due in 2 weeks)

Project Proposal example

Name(s):	**Robin Burke**
Brief description of network / data set:	Bi-partite affiliation network of professors at DePaul University and the University-level committees on which they have served from 2001-2011.
Data source:	Does not exist, but it would be cool if it did
Network size (if known) nodes / edges:	357 faculty / 87 committees, 8,342 edges
Edge relation(s):	Faculty member -> committee served
Vertex attributes:	Professors: ID (anonymized), age, sex, ethnicity, race, college as of 2011, years at DePaul (integer 0-35), rank, Committees: ID, name, standing, elected, size, date of creation (2001-1-1 to 2011-12-31), date of dissolution (2001-1-1 to 2011-12-31)
Edge attributes:	Date of start of committee service (2001-1-1 to 2011-12-31)
Analysis question(s):	• Who are the influential faculty members with respect to university committees? How are these individuals distributed by college? • Is there a correlation between academic rank and committee service? • How are committee assignments distributed by college and by demographic factors? • Are there clusters of individual who co-serve on the same committees? • What committees are linked by co-membership? Does this association change over time?
Notes:	End of service data was not available is most cases. I am still working to calculate this from other sources.

Ideas

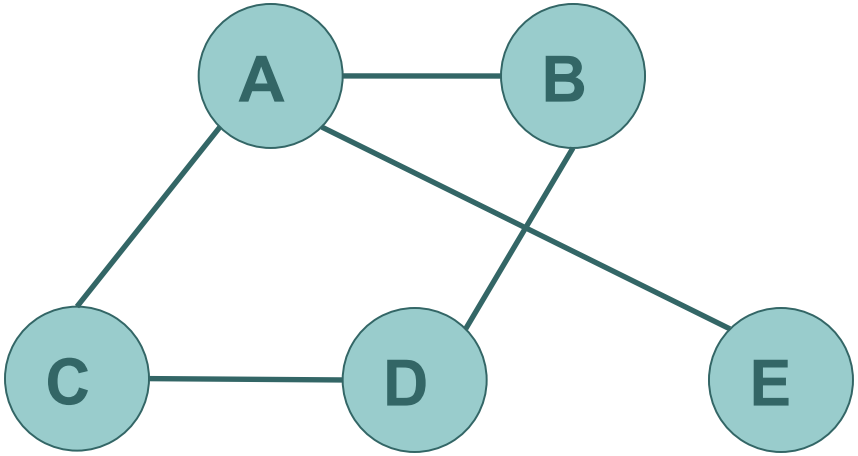
- Links to sources network data sets on D2L
- Examples
  - Included on D2L are two projects from previous quarters
  - To give you some ideas
  - To show what an excellent project looks like

Outline

- Terminology
  - Paths
  - Components
- Linear Algebra review
- Centrality
  - Degree
  - Betweenness
  - Closeness
- Break
- Eigenvector and related measures
  - Eigenvector centrality
  - PageRank
- Example

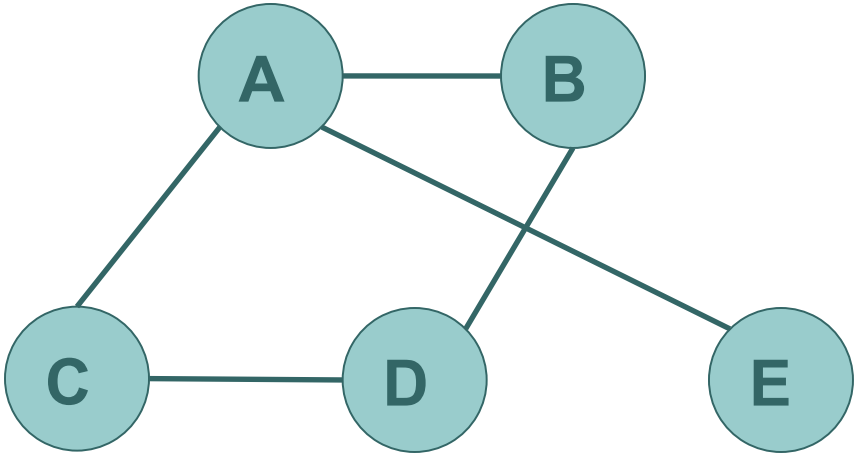
Movement on a graph

- On a graph you can move from node to node
  - via edges
- Might be actual movement
  - messages in a communication network
- Or it might be more implicit
  - social influence



Walk

- Sequence of edges
  - such that the node ending each step is the beginning of the next
- B-A-C-A is a walk...
- Length
  - # of edges

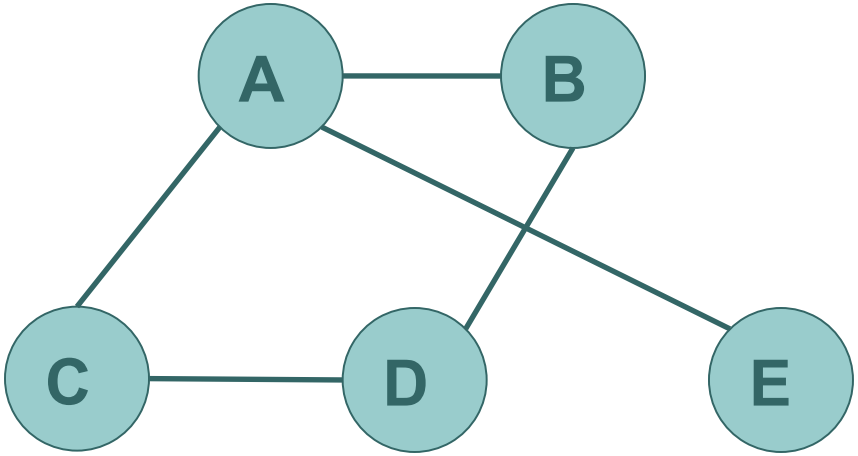


Random walk

- A random walk is a sequence of edges
  - chosen randomly at each vertex
- Turns out to be a very important construct
  - more later tonight

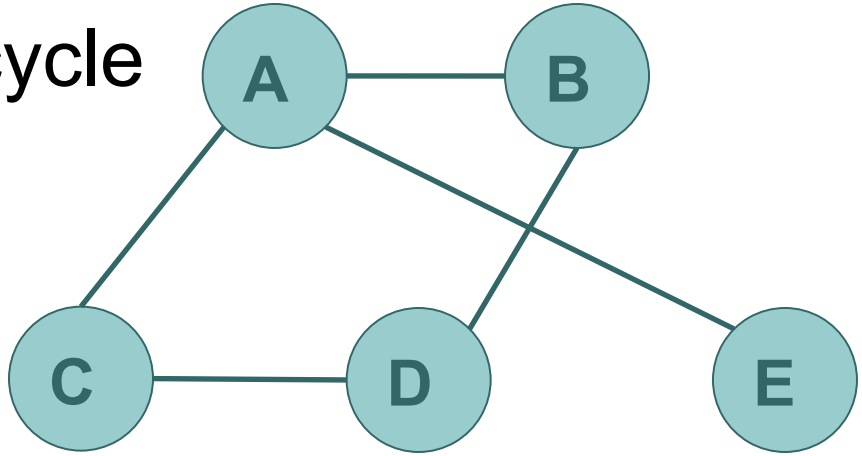
Path

- A walk where no vertex appears more than once
- B-A-C-A is not a path
- B-A-C-D is a path



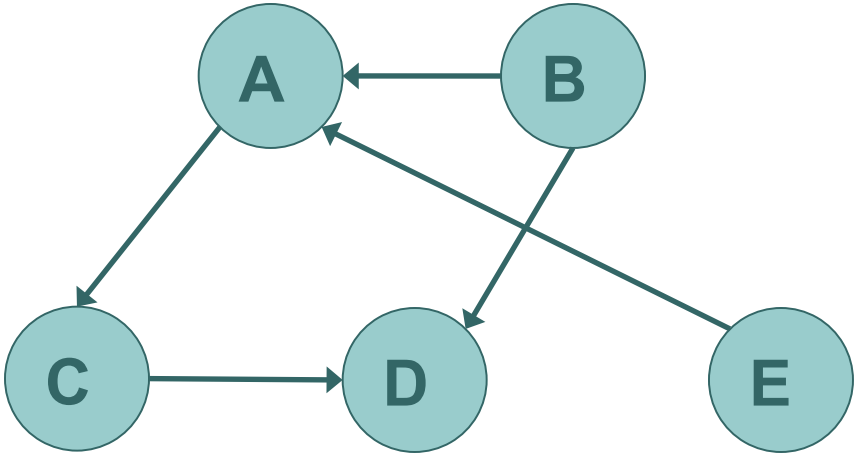
Cycle

- A walk where the beginning and ending nodes are repeated
  - but no others
- B-A-B is a cycle
- B-A-C-A is not a cycle
- B-A-C-D-B is a cycle



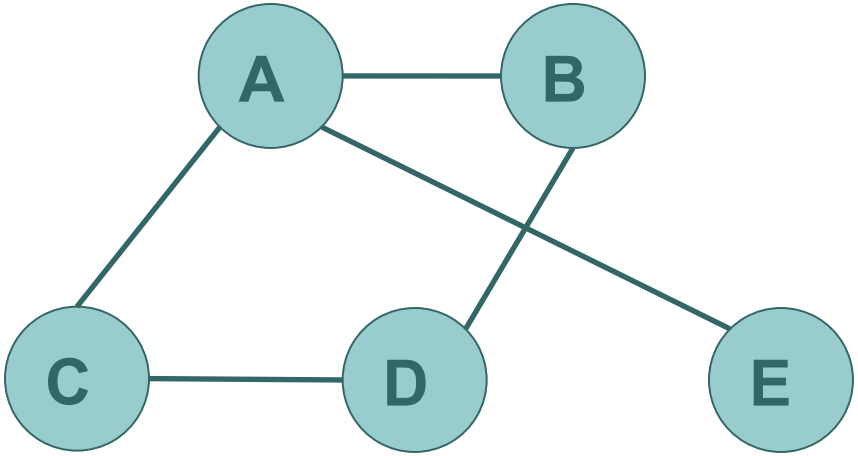
Cyclic / acyclic graphs

- Distinction for directed graphs
- Cyclic graph is one with cycles
- Acyclic has no cycles



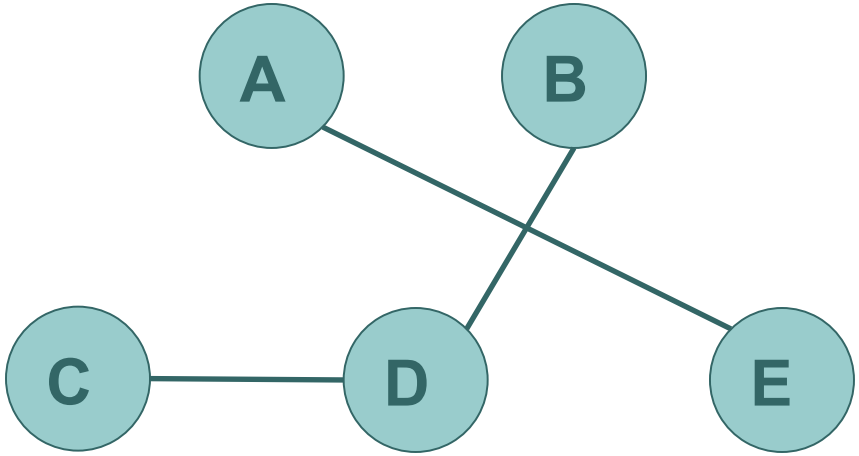
Shortest path / geodesic

- There may be multiple paths from one node to another
- B-A
  - B-A (length 1)
  - B-D-C-A (length 3)
  - B-D-C-D-C-A
    - not a path
- Shortest path aka
  - geodesic



No path

- It is possible that there is no path between two nodes
  - D-E no path
    - length = ∞

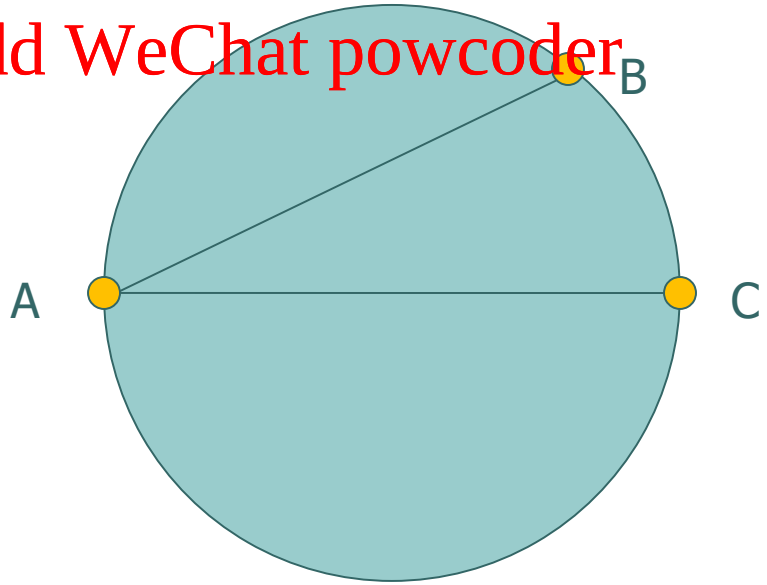


Diameter

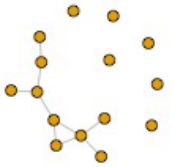
- length of the longest geodesic
- maximum distance inside the network
- Disconnected graph?
  - technically should be ∞ (or undefined)
  - sometimes it is the maximum diameter of any component

Intuition

- Shortest path from A to other nodes
- A-C is the diameter
  - longest of these shortest paths



What is the diameter of this network?



4

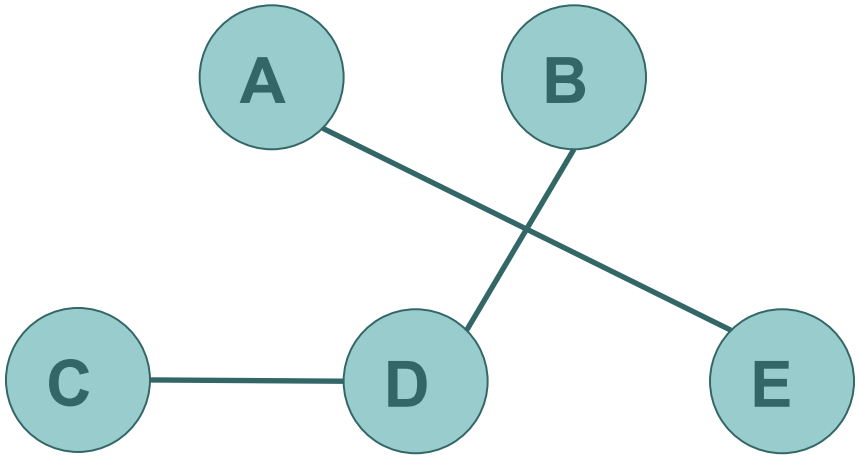
5

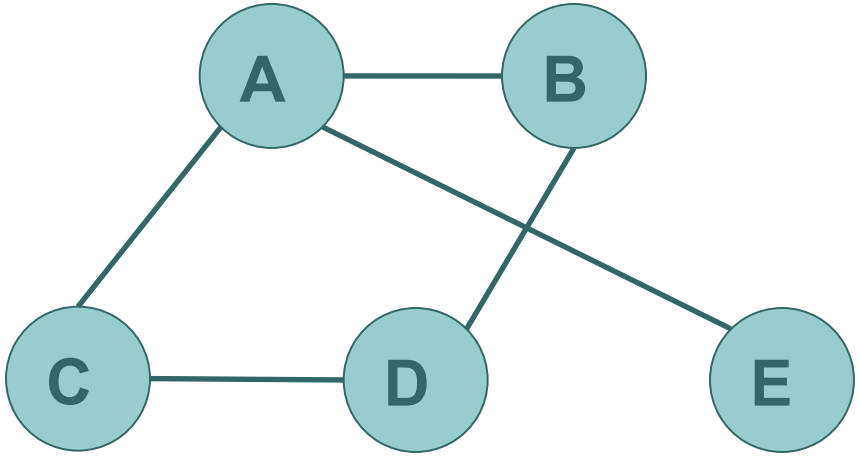
6

No diameter can computed because it contains disconnected nodes

Connected

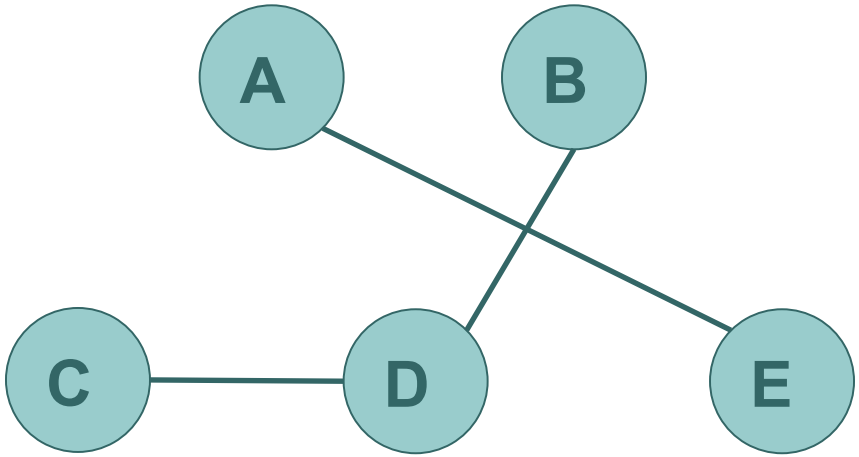
- A network is connected
  - if there is a path between all the nodes in a network

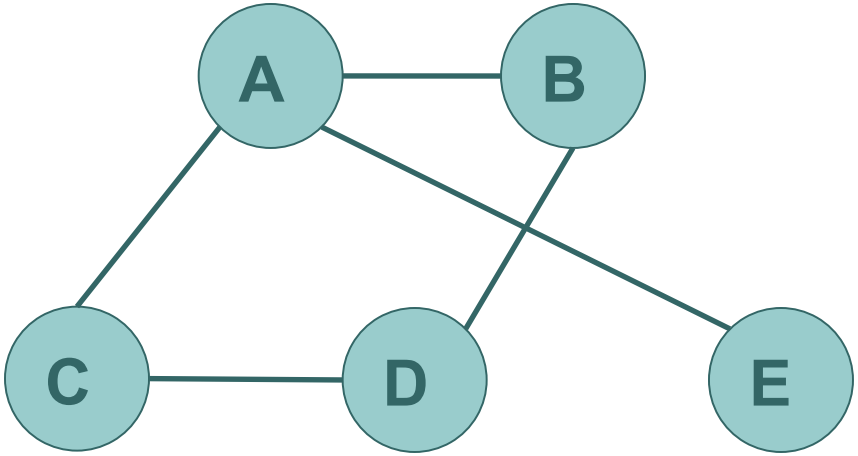




Component

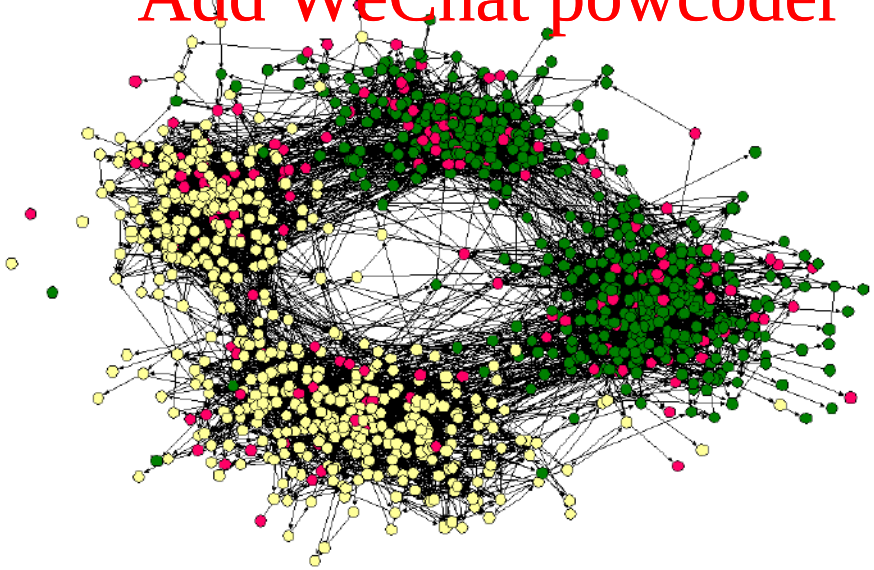
- A component is a connected group of maximal size
- If the network is connected
  - there is one component
- Otherwise there are multiple components



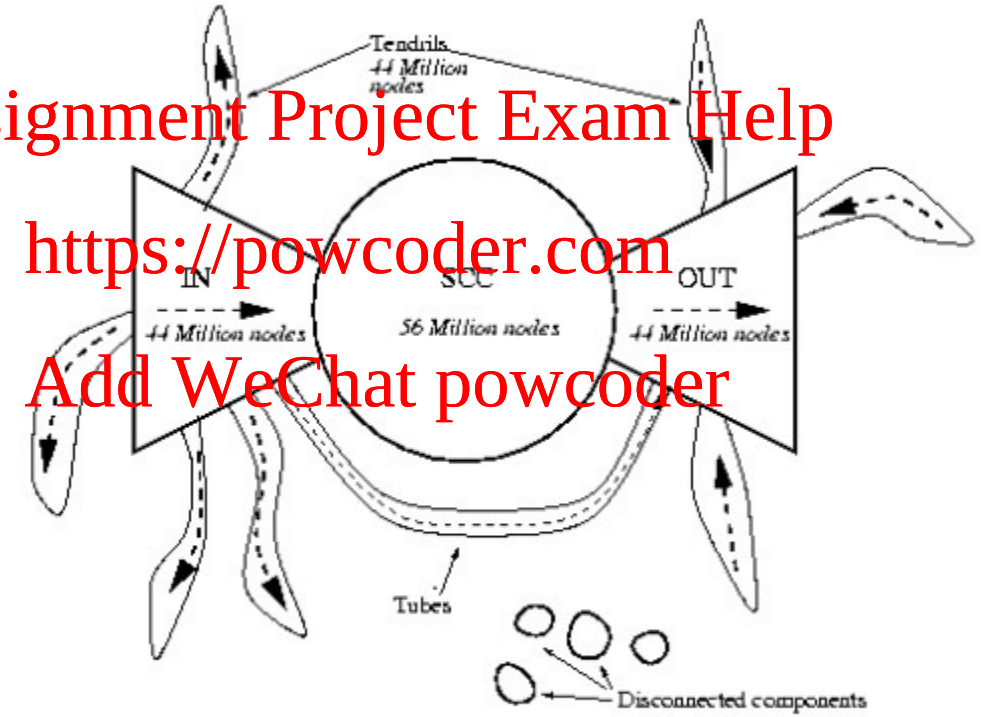


Giant component

- Real-world networks often have a "giant component"
  - which is 80-90% of the network all in one component



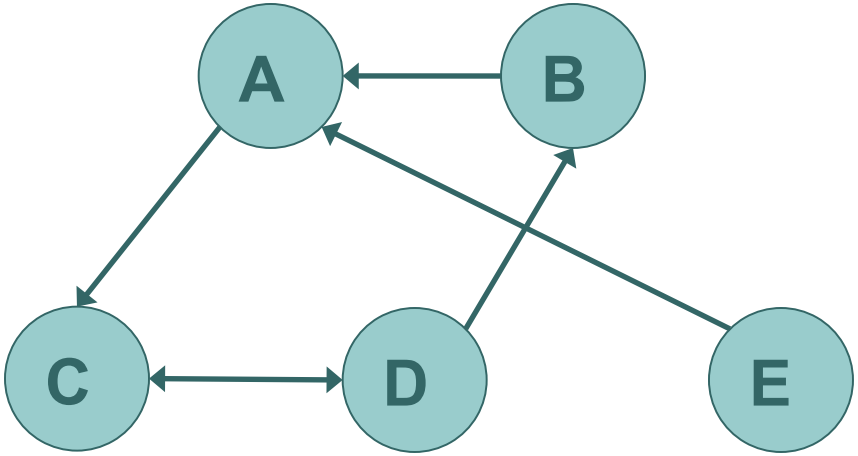
Bow-tie structure of the web



The bow-tie graph structure of the Web (Broder et al 2000)

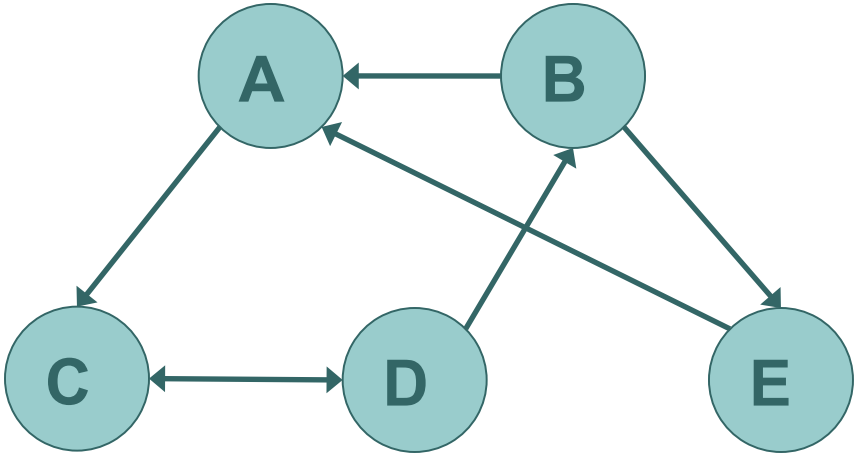
Directed network

- Here there is no path from A to E
  - but there is one from E to A
- We have a new term
  - "strongly-connected"



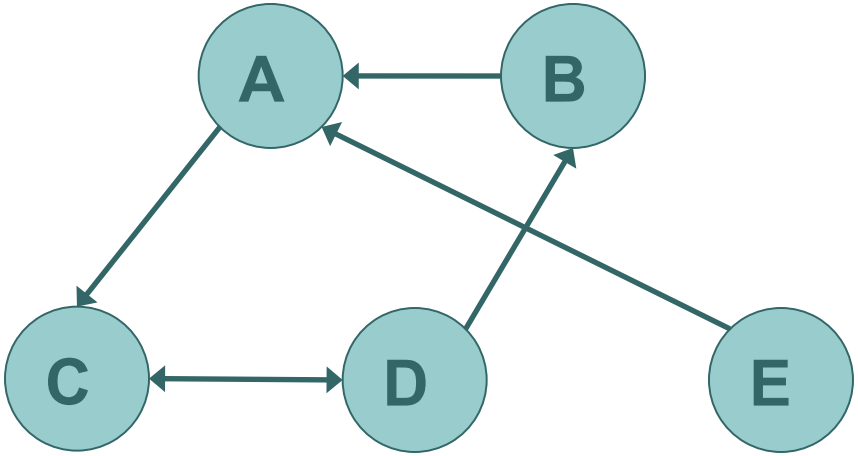
Strongly-connected

- A directed network is strongly-connected
  - if there is a path from every node to every other
- Note that this requires
  - cyclic graph



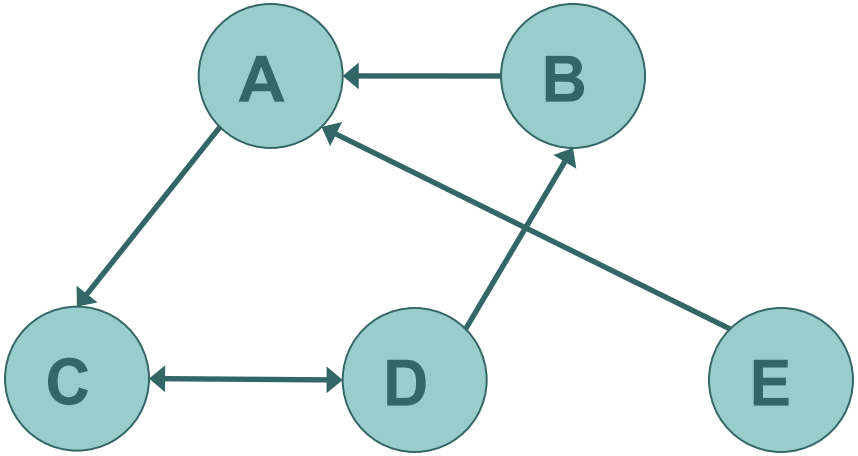
Weakly-connected

- A directed network is weakly-connected
  - if it becomes connected when all edges are made undirected

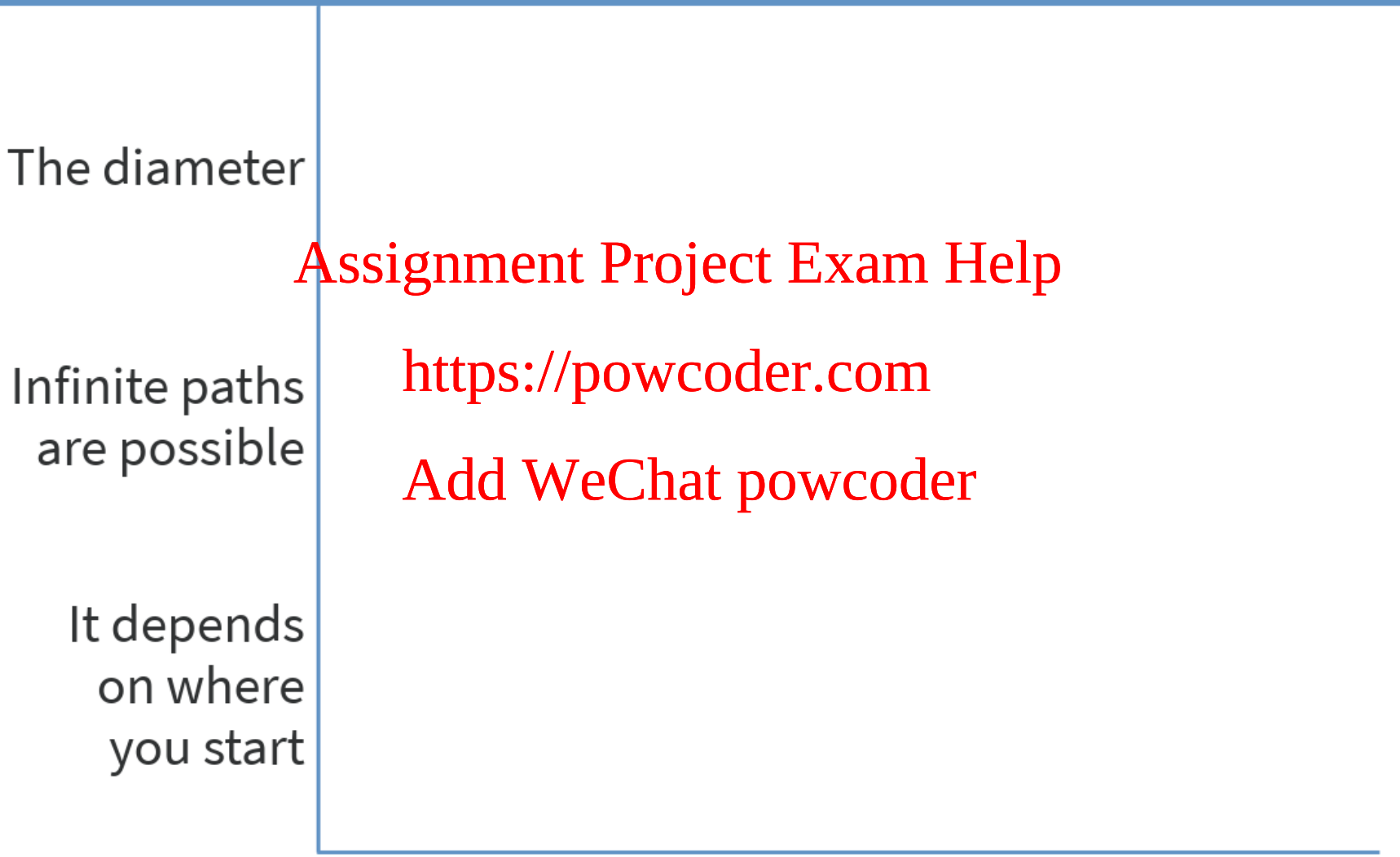


Strongly-connected components

- A weakly-connected network
  - is made up of strongly-connected components
  - chunks of the network that are strongly connected



If a graph is directed and acyclic, what is the longest path it can have?

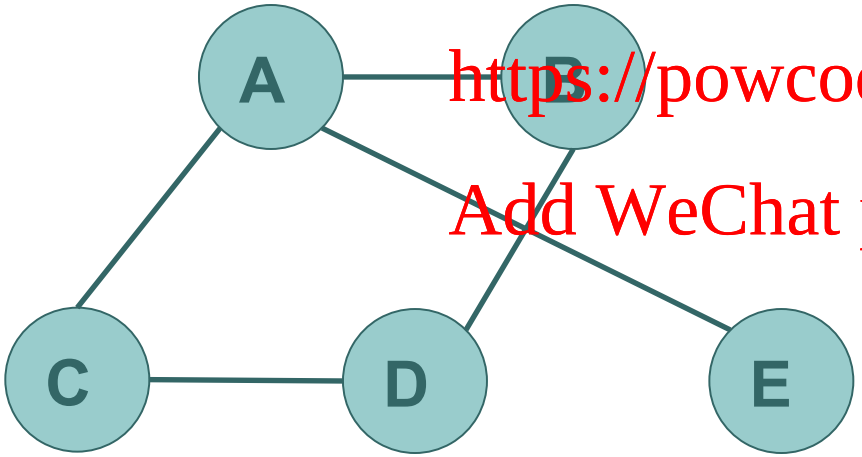


In igraph

- decompose()
  - Returns the components of a graph
    - List format
    - Remember [[1]] type indexing
  - Can specify strong / weak

Linear Algebra Review

Adjacency Matrix



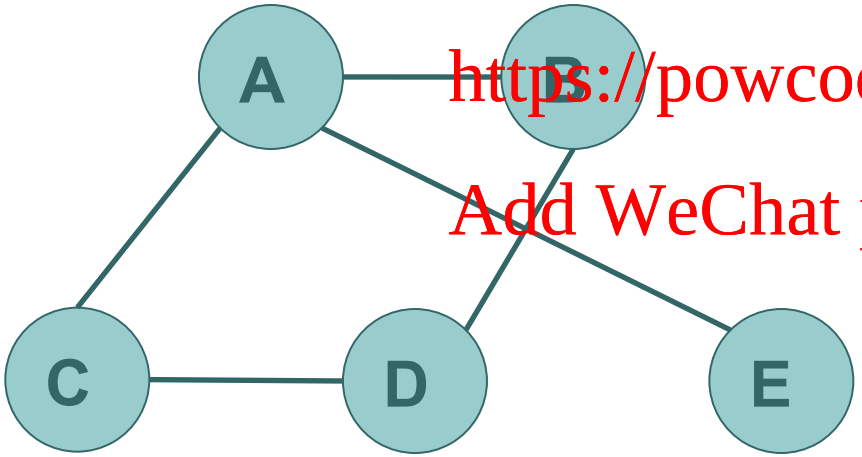
$$\begin{bmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Linear Algebra Review

- An adjacency matrix is a square matrix
  - usually binary valued (0, 1)
- Results from linear algebra very relevant
  - for working with networks

Matrix multiplication

- Equals graph traversal



$$\begin{bmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Start at A

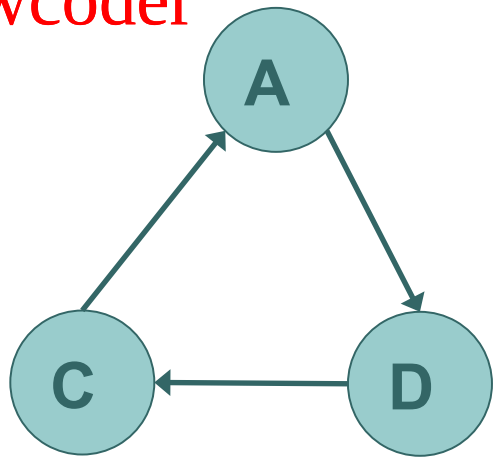
- [1 0 0 0 0] M = [0 1 1 0 1]
  - all 1 step paths from A
- [0 1 1 0 1] M = [3 0 0 2 0]
  - all 2 step paths from A
- [3 0 0 2 0] M = [0 5 5 0 5]
  - all 3 step paths from A
- Note
  - A to A – no three step paths
  - means there are no triangles starting at A

Transformations

- Transformation by the adjacency matrix:
  - represents paths in the network
  - M^k = paths of length k
- What about M^*?
  - arbitrary k

Special case

- Periodicity
  - if the network is periodic
  - no interesting answer to this question



Otherwise

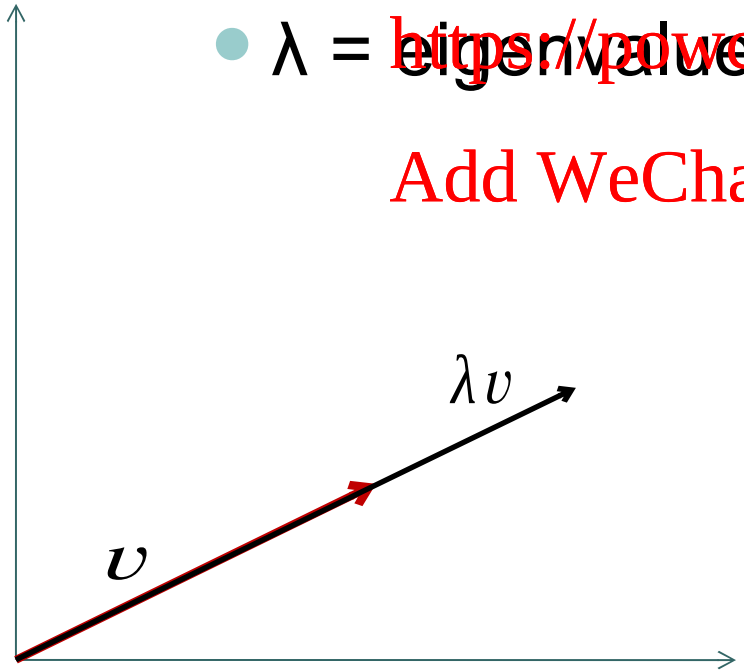
- M* will represent the strongest "tendency" of the matrix
- That might be an interesting thing to know!

Matrices as transformations

- Let M be a nxn square matrix
  - like an adjacency matrix
- Let v be a n dimensional vector
- Mv=w
  - where w is another n dimensional vector
- *(Yes, I switched sides for the multiplication. This is equivalent to using the transpose of the matrix.)*
  - *Or defining "from" as the columns and "to" as the rows*

Eigenvectors

- if $Mv=\lambda v$
  - then v = eigenvector of M
  - λ = eigenvalue of M



Eigenvalues

- An nxn matrix has at most n eigenvalues
- $\det(A - \lambda I) = 0$
  - given an n-degree polynomial that defines the eigenvalues
- Sometimes there are multiple eigenvectors with the same λ
  - "defective" matrices
- Symmetric matrix
  - has real eigenvalues
  - non-symmetric matrices might have complex eigenvalues

Ordering eigenvalues

- We can sort the eigenvectors by eigenvalue
  - talk about the 1st eigenvector of a matrix, etc.

Decomposition

- Assume that M has n distinct eigenvalues
  - true in most practical graph applications
- Let Q be a square matrix of the eigenvectors
- $MQ=Q\Lambda$
  - where Λ is the matrix with the corresponding eigenvalues on the diagonal
- Means
  - $MQQ^{-1}=Q\Lambda Q^{-1}$
  - $M =Q\Lambda Q^{-1}$
- So the matrix M can be expressed in terms of just its eigenvalues and eigenvectors

Powers of M

- $M^k = (Q \Lambda Q^{-1})^k$
- $= Q \Lambda^k Q^{-1}$
- Meaning that the largest eigenvalue will dominate
  - as k -> infinity

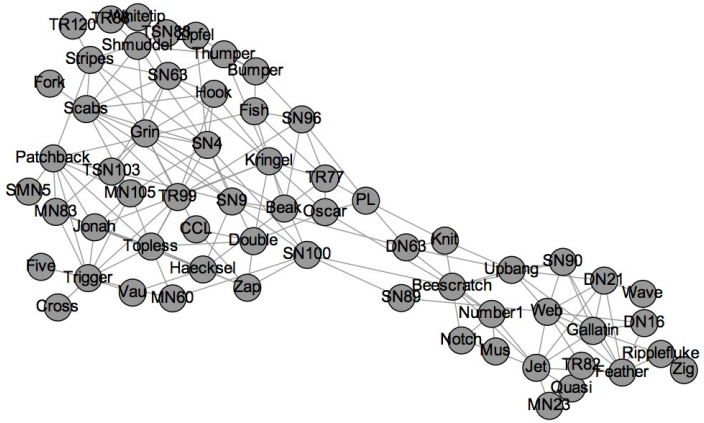
Steady state tendency

- Leading eigenvector
  - corresponding to largest eigenvalue
- Represents
  - where paths through the network
  - will tend to stay as they get longer and longer
- Interesting way of ranking nodes
  - which are more or less probable?
- We will come back to this insight!

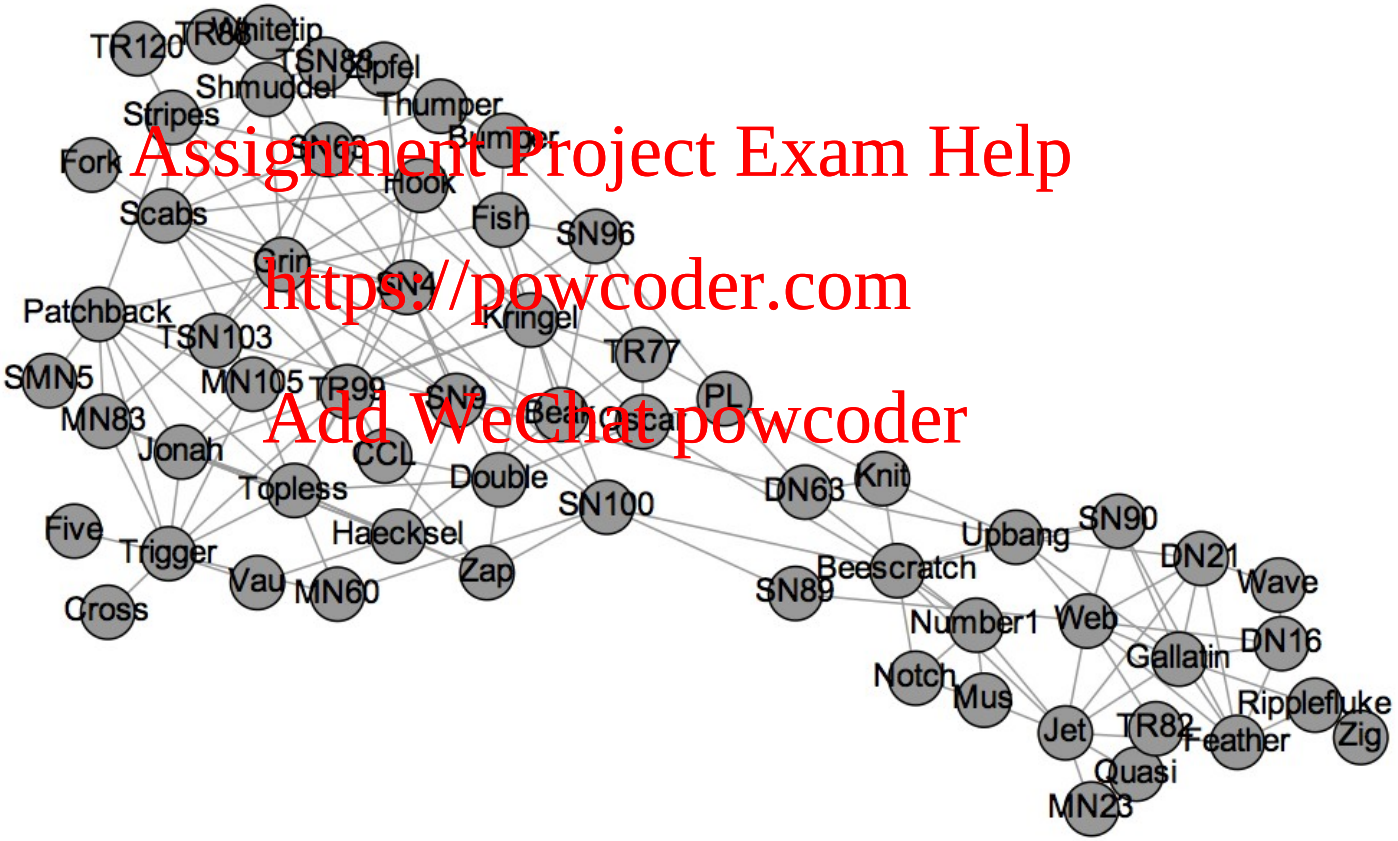
Centrality

The Question

- What are the most important nodes in a network?
- "Important" is vague
  - will vary
  - by network type
  - by analysis task



Dolphin Network

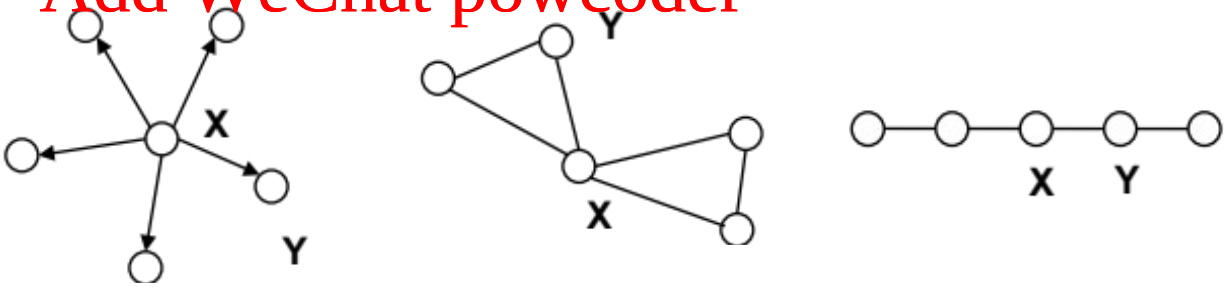


The Answer

- Basic idea
  - apply a node-level metric
  - rank nodes
- What metrics to apply?

Three basic types of centrality

- Different measures capture different ideas of what makes a node important
- Here X > Y by different measures

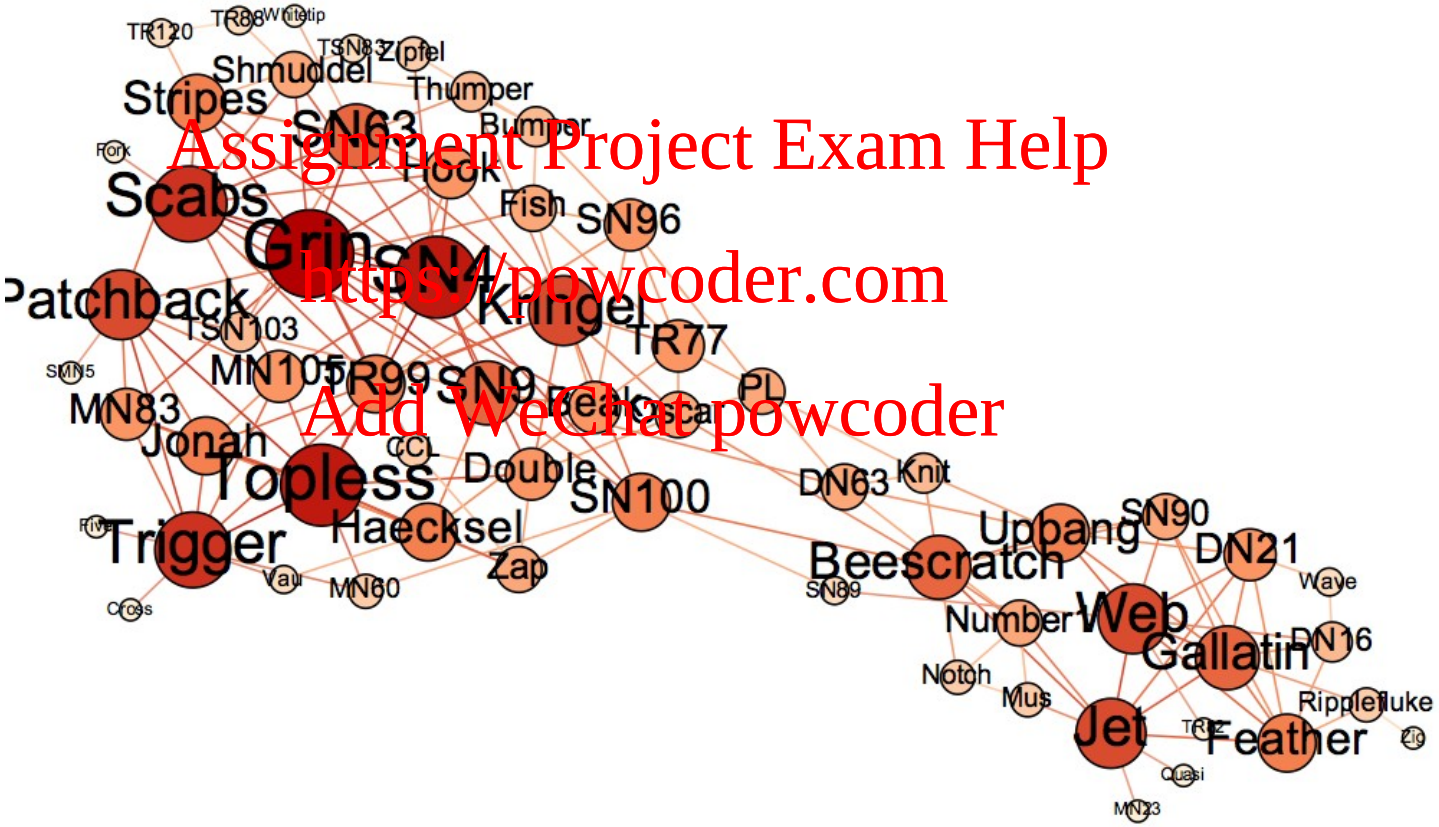


degree betweenness closeness

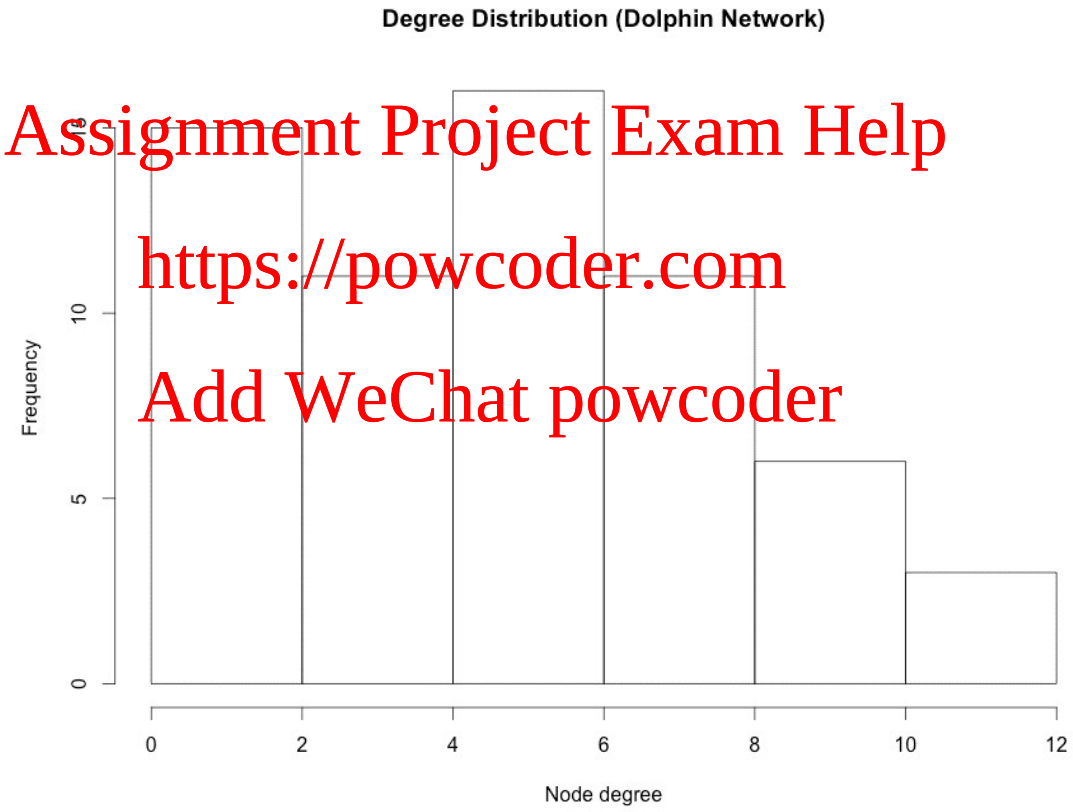
Degree

- We have talked about this before
  - # of incident edges
- High degree makes a node important IF
  - immediate connections are important

Example: Dolphin network



Degree distribution

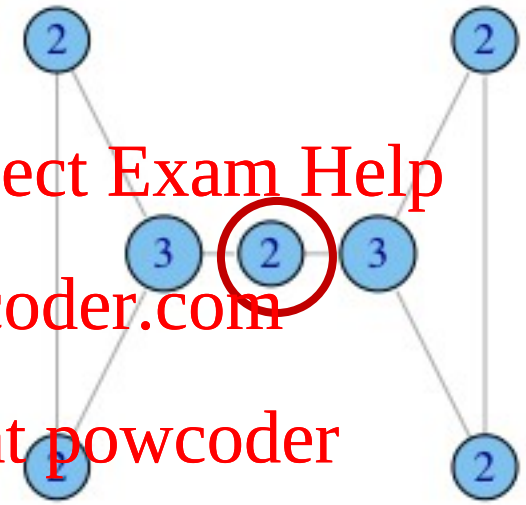


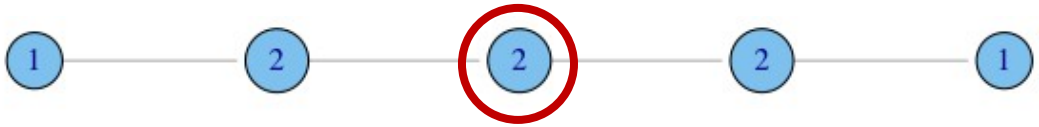
Degree centrality

- Only counts immediate neighbors
- Each neighbor equally valuable
- Many variants
  - weighted
  - in-degree
  - out-degree
  - more exotic extensions

Brokerage

- Not captured by degree

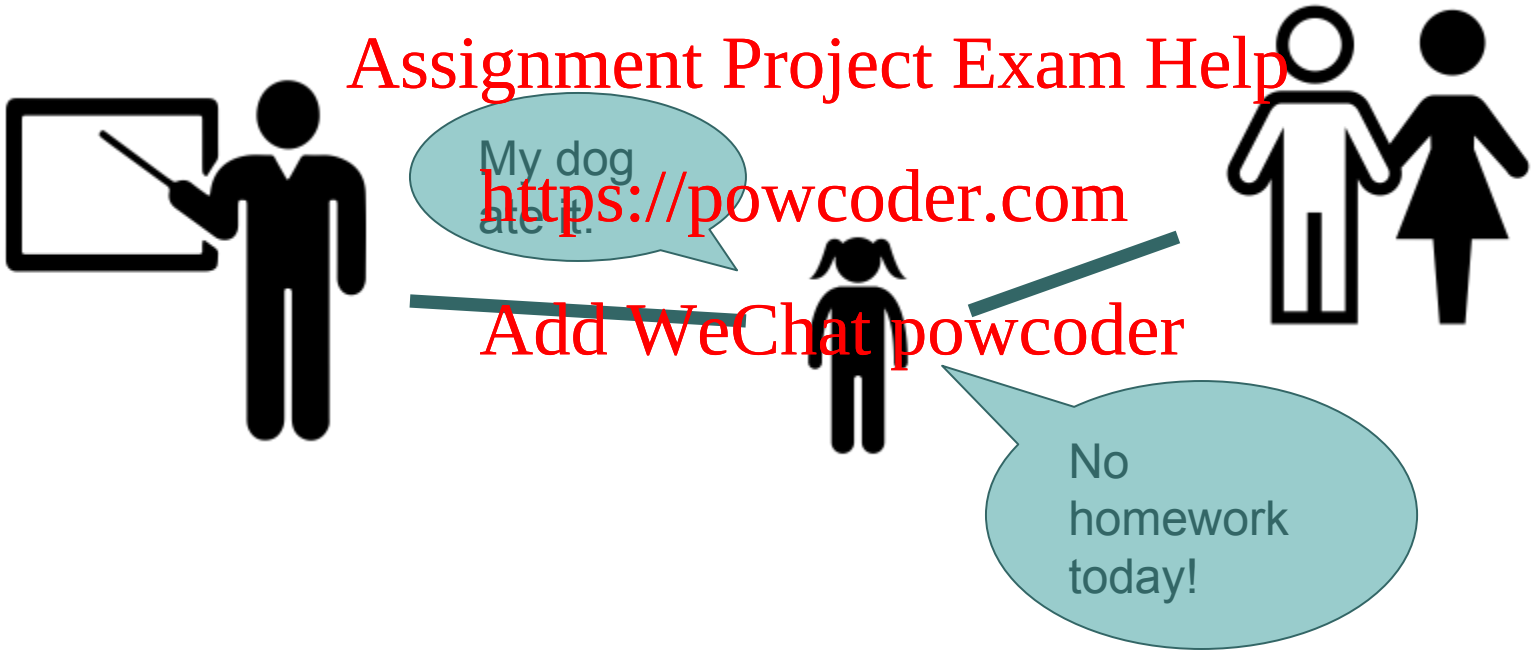




Brokerage

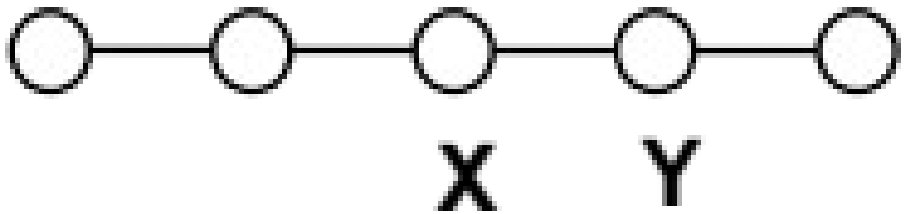
- The extent to which one individual is an intermediary between different groups
- who otherwise lack direct connections

Constraint / Structural holes

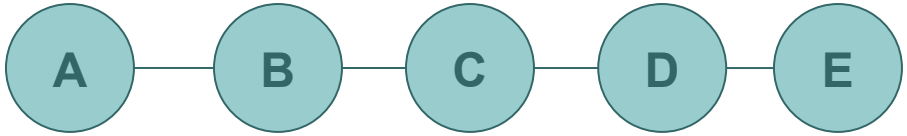


Capturing brokerage

- Intuition: how many pairs of individuals would have to go through you
  - in order to reach one another in the minimum number of hops?



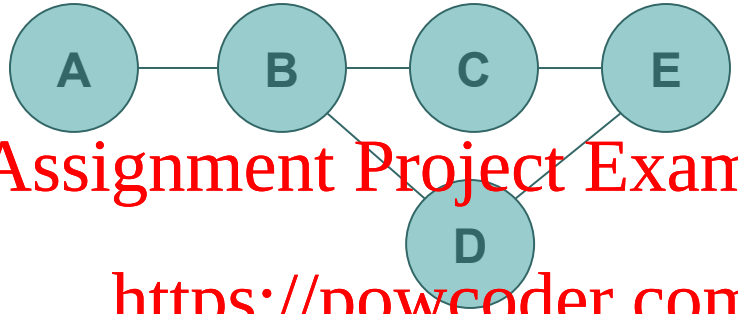
Betweenness



- 10 Pairs
  - A-B, A-C, A-D, A-E, B-C, B-D, B-E, C-D, C-E, D-E

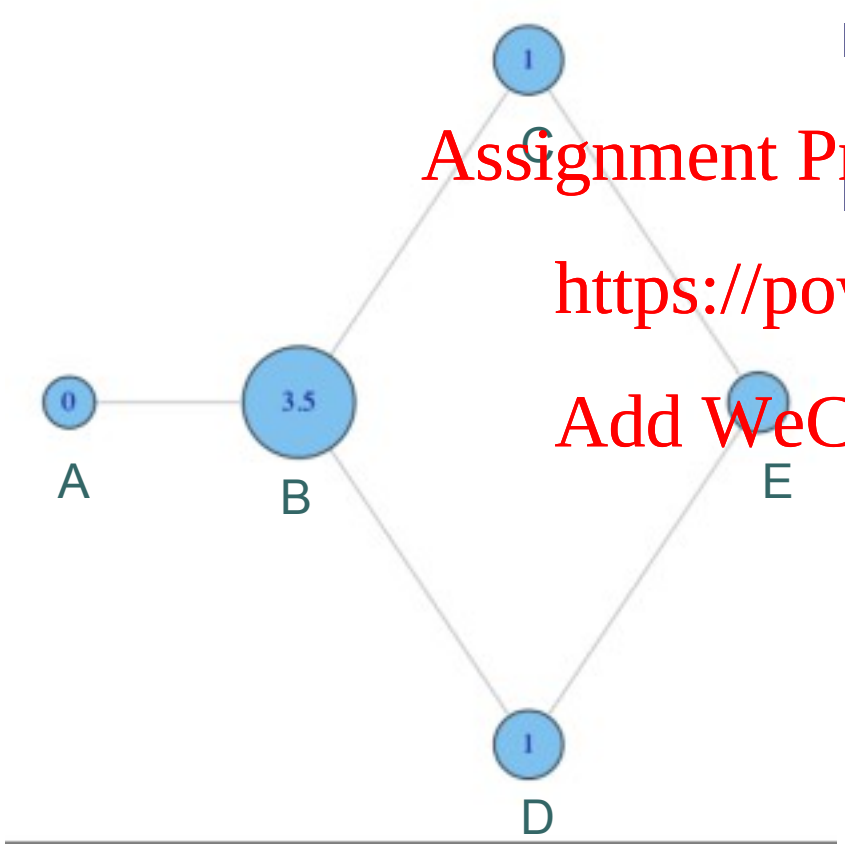
- How many involve C? (without ending at C)
- How many involve D?
- How many involve E?

But



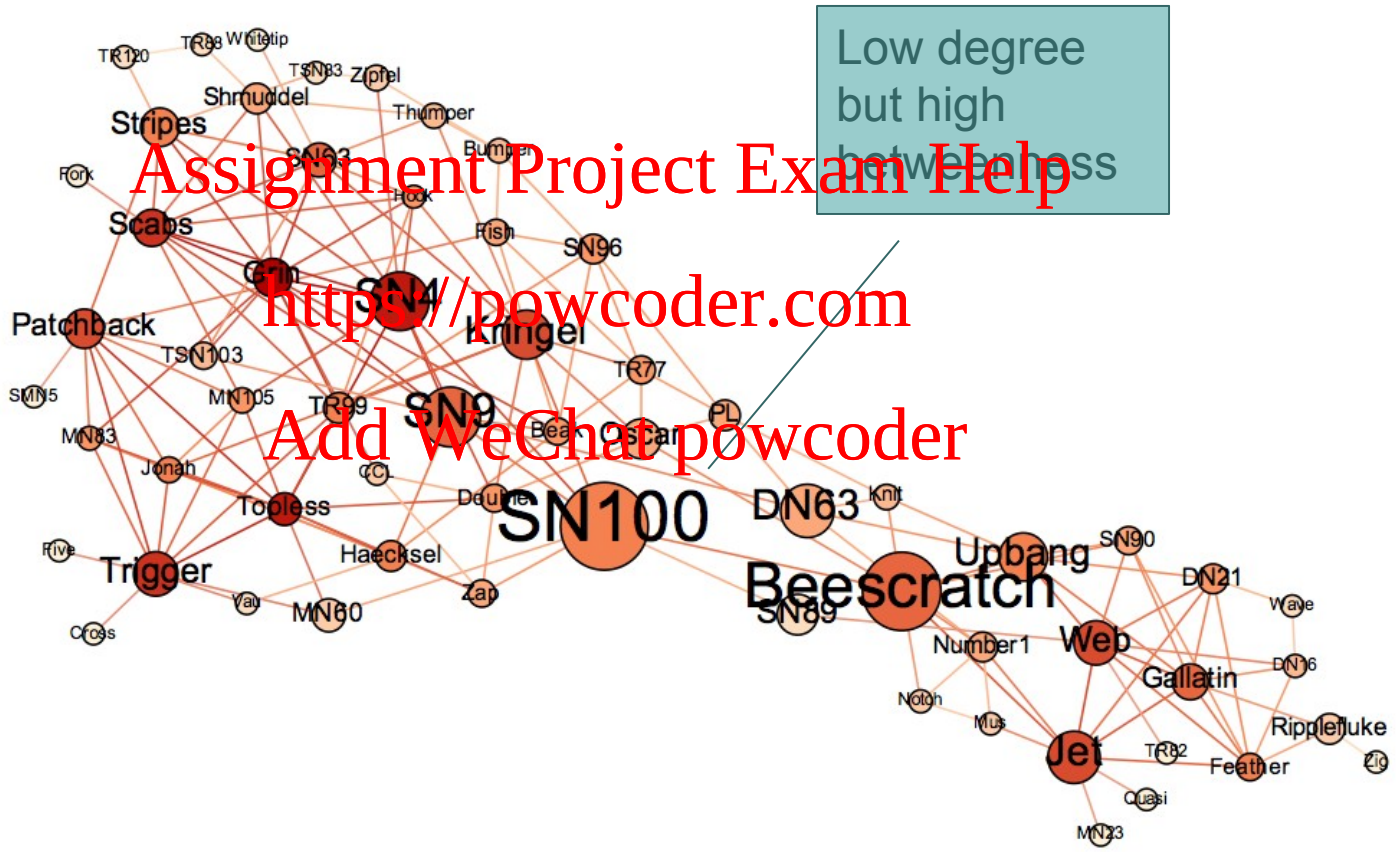
- How to account for multiple shortest paths?
  - C and D “share”
  - divide by the number of shortest paths

Betweenness examples

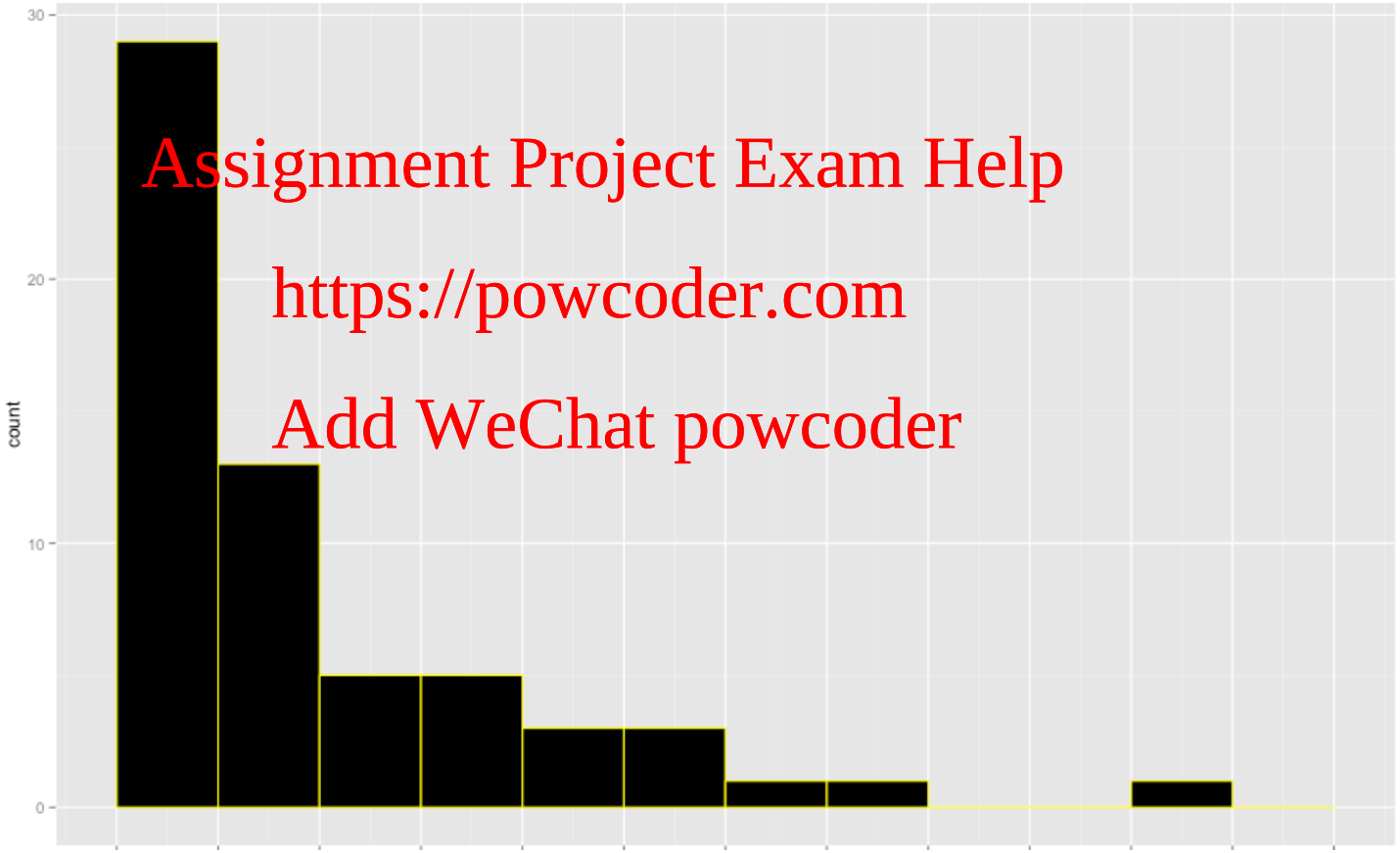


- why do C and D each have betweenness 1?
- They are both on shortest paths for pairs (A,E), and (B,E), and so must share credit:
- ½+½=1

Dolphin Network



Betweenness distribution



Computation

- Betweenness is kind of slow to compute
  - especially if you do it in a naive way
    - $O(n^3)$
- We have to count every path between every pair of nodes
  - best algorithm is $O(n^2)$
  - Compare computing degree = O(n)
- With very large networks
  - might not be able to compute this
  - might prefer measures of lower complexity
- igraph has bounded betweenness
  - only consider paths of a certain length
  - reduces the computational cost greatly

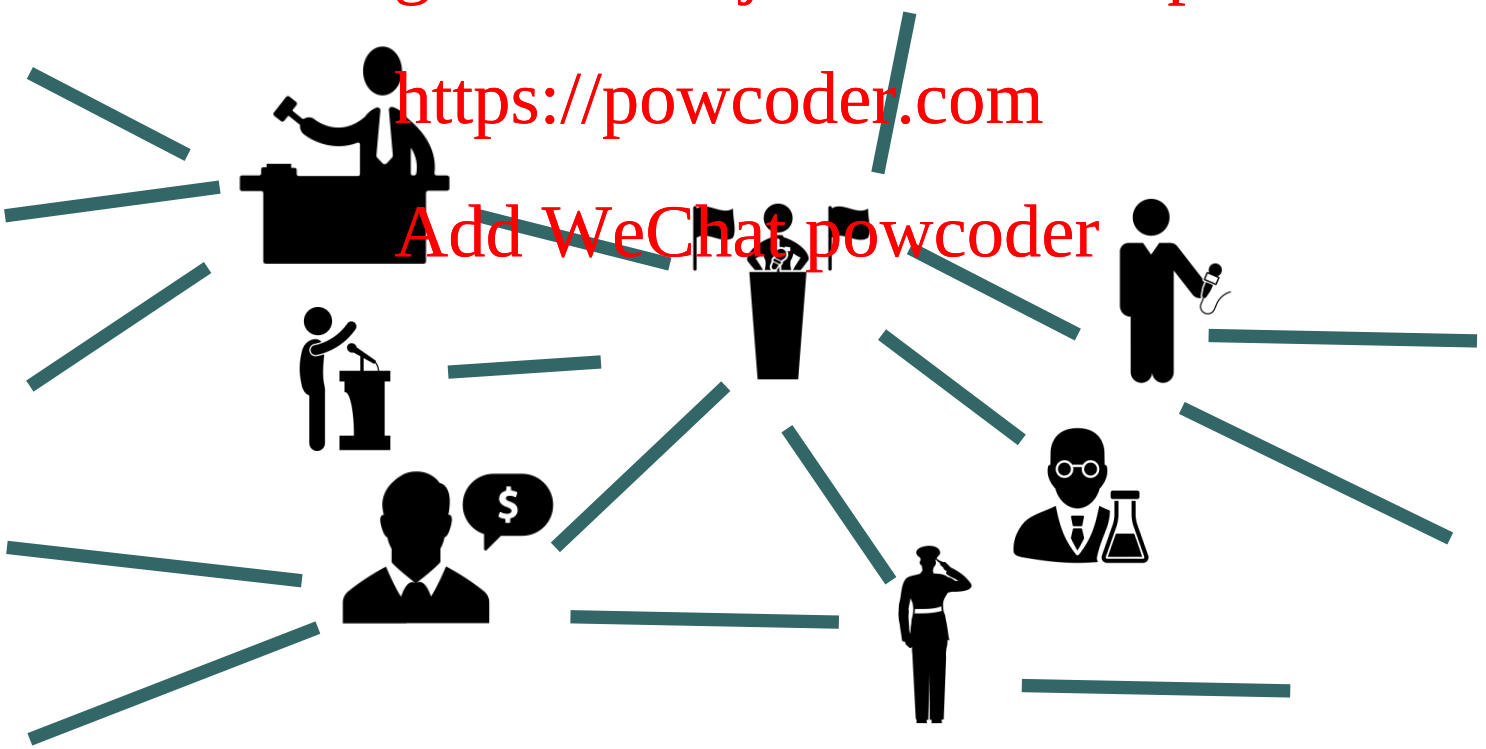
Betweenness

- A node with high betweenness is important IF
  - the purpose of the network is communication or the flow of things
  - and the person benefits from the flow

Closeness

- What if it's not so important to have many direct friends?
- Or be "between" others
- But one still wants to be in the "middle" of things, not too far from the center

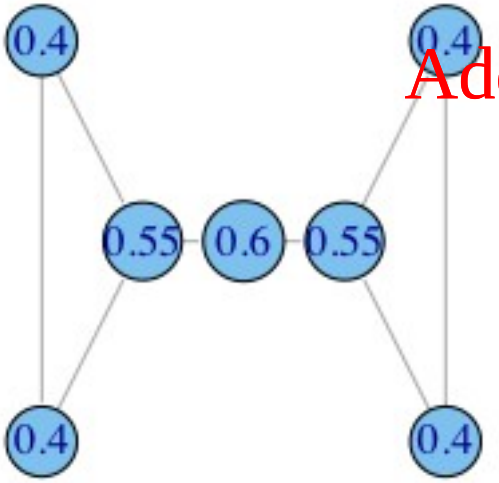
Importance without brokerage

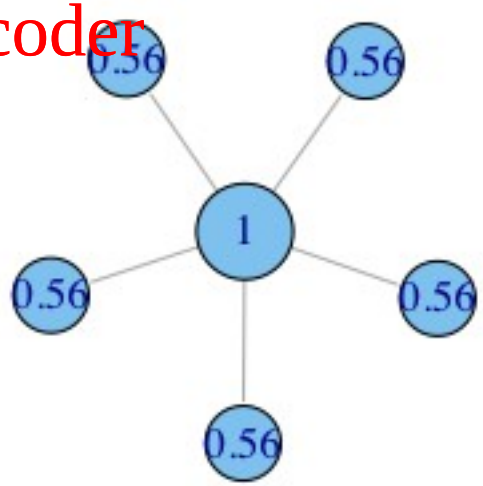


Closeness

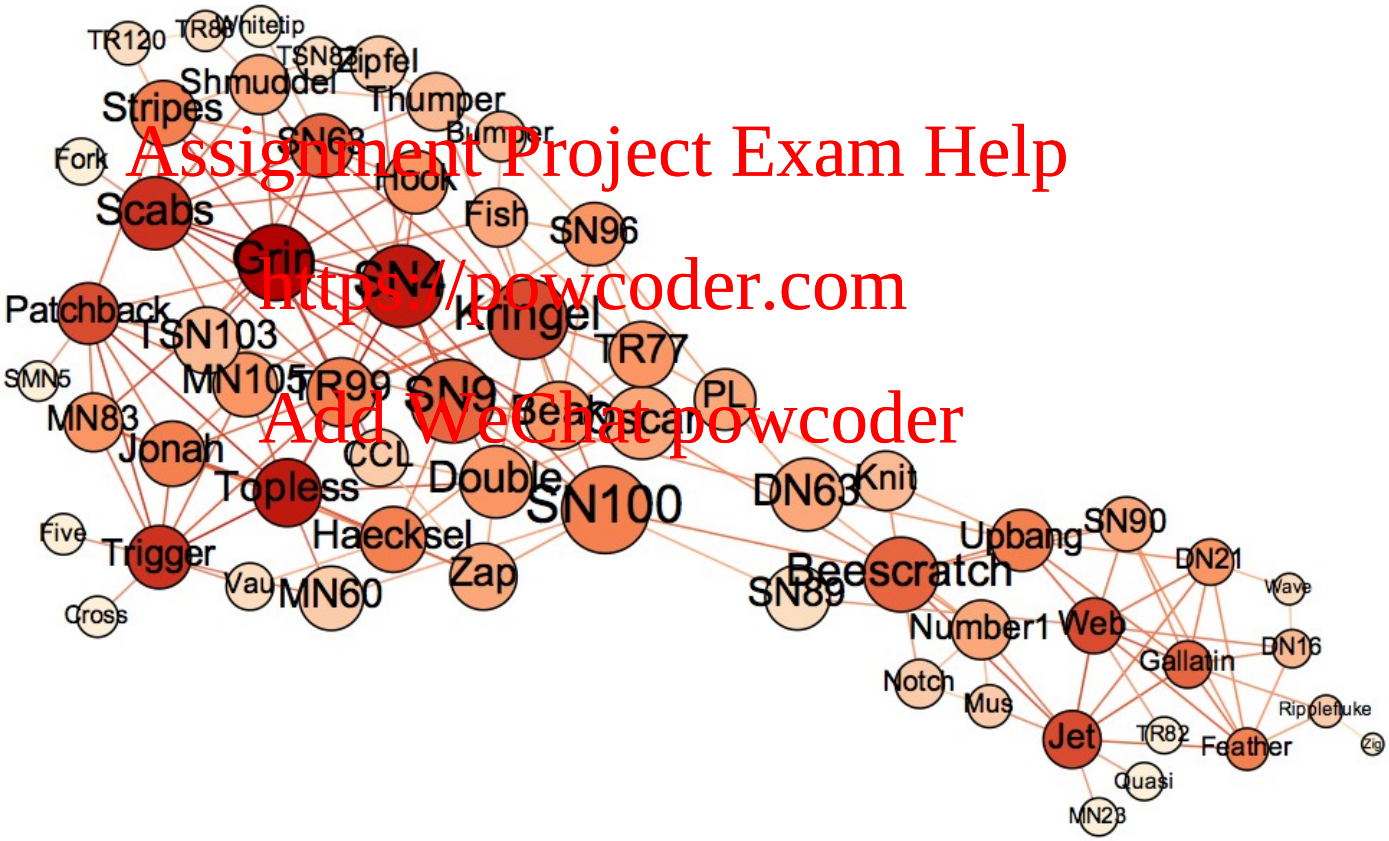
- Average distance to other nodes
- This gives a strange result
  - smaller is more important
- So to match other metrics
  - 1 / (average distance)
- Now the range is 0 to 1
  - close to zero means far from most nodes

Closeness examples





Dolphin Network

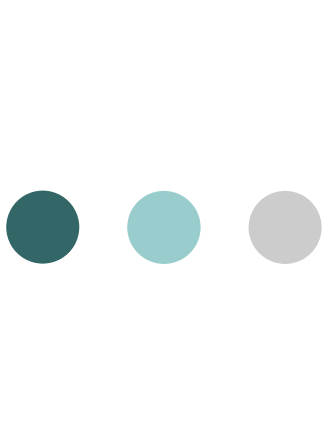


Disconnected graphs

- What to do if there is no path between node A and node B?
- igraph
  - Treats path as if it were longer than the longest possible path
    - |N|
- We need to do this because a singleton node should not have high closeness
  - Should be lowest possible
- igraph will give a warning if the graph is not (strongly) connected

Closeness

- A node with high closeness is important IF
  - reaching others in a small number of steps is important



Break