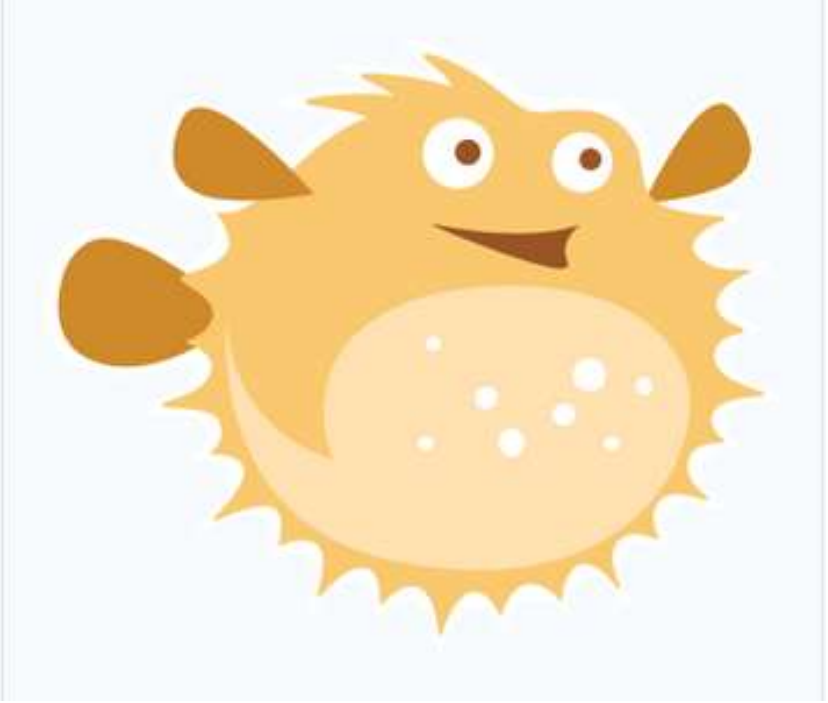
Twitter Network Analysis with NetworkX

Celia La, Sarah Guido
PyCon 2015

@celiala, @sarah_guido

About us: Sarah Guido

- Data scientist at Bitly
- NYC Python and PyGotham organizer
- O'Reilly Media author
- @sarah_guido



About us: Celia La

- Software engineer at Knewton
- PyGotham and Write/Speak/Code
- @celiala

About this talk

- Installation
- Intro to network theory/NetworkX
- Intro to the Twitter API
- Lesson!

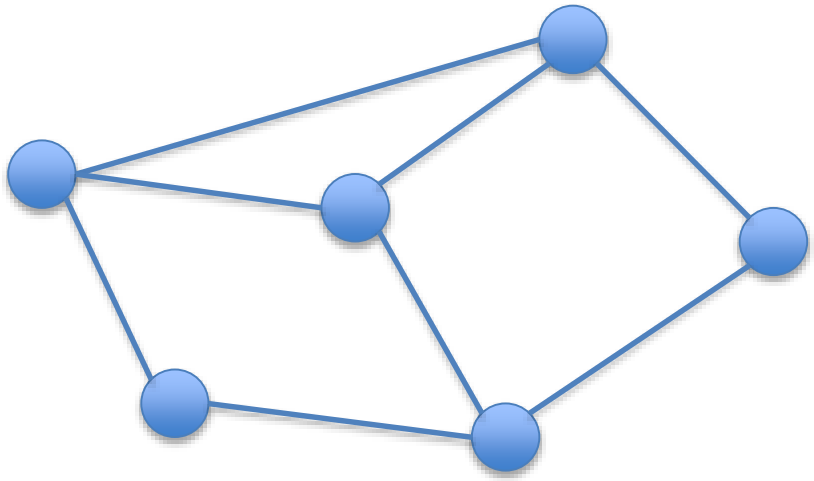
Installation

- Github repo!
- Let’s try opening IPython notebook

The basics of network theory

What is a network?

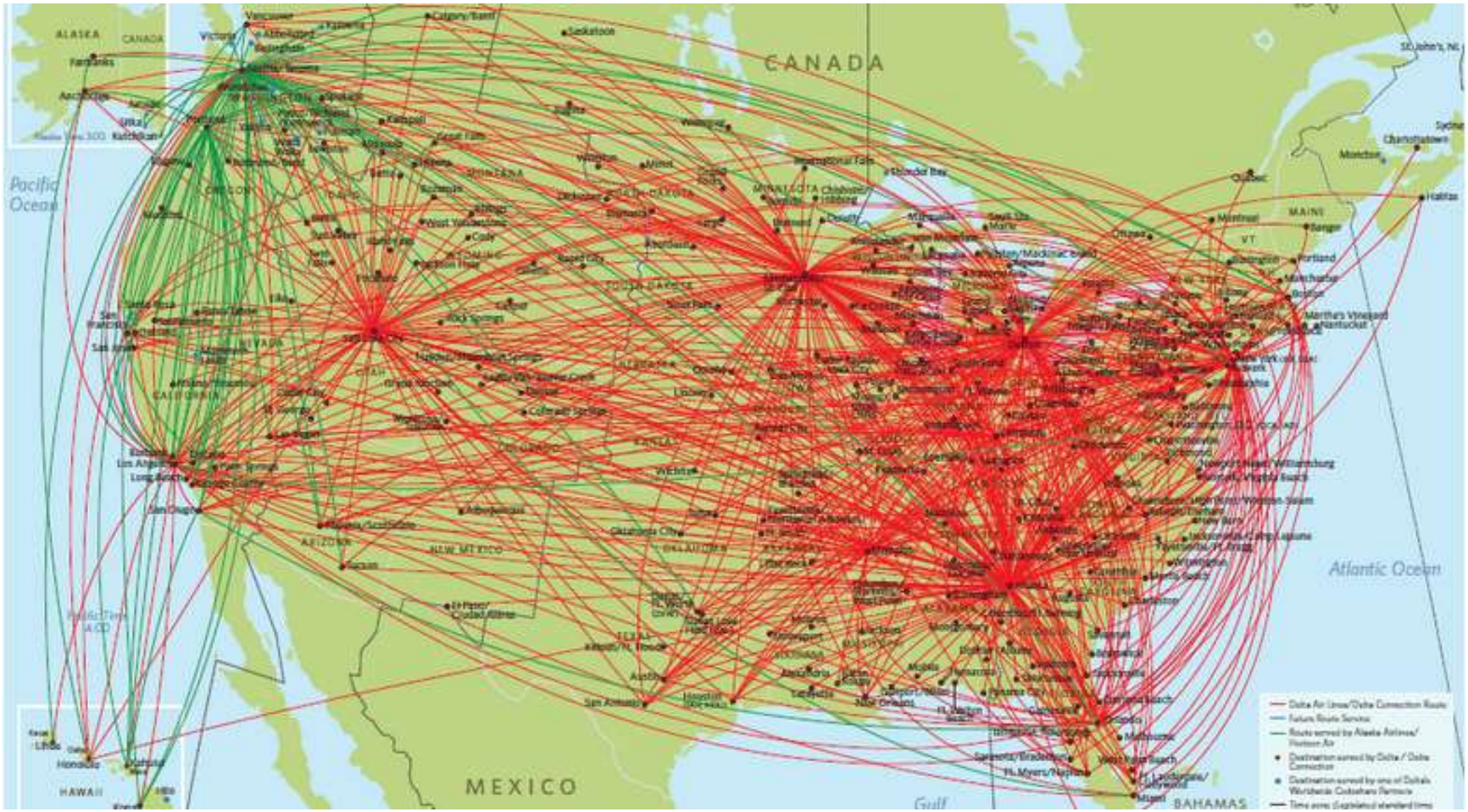
- Collection of points joined by lines
- Mathematically: graph
- Representation of relationships between discrete objects



What is a network?

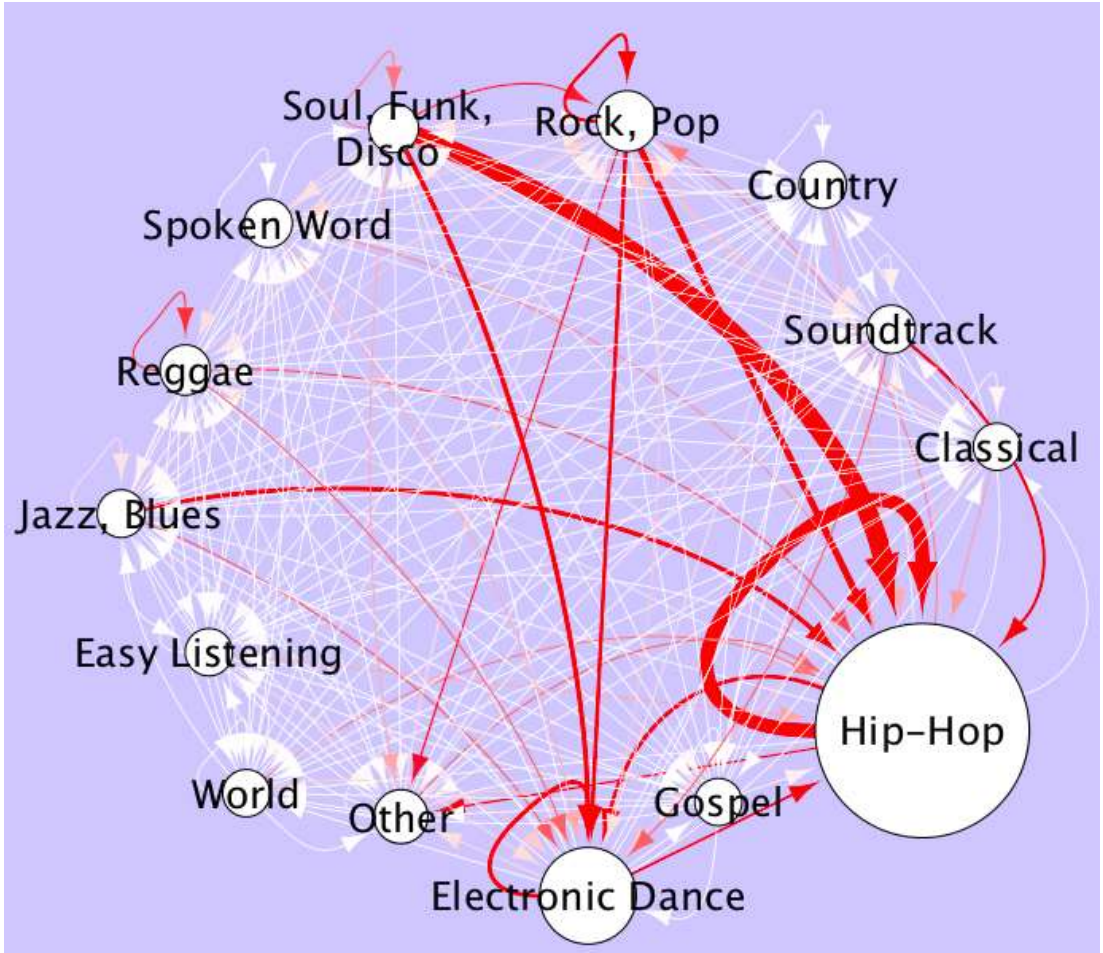
- Can be thought of as
  - a complicated data structure
  - a complex system
  - a way of exploring data

What is a network?



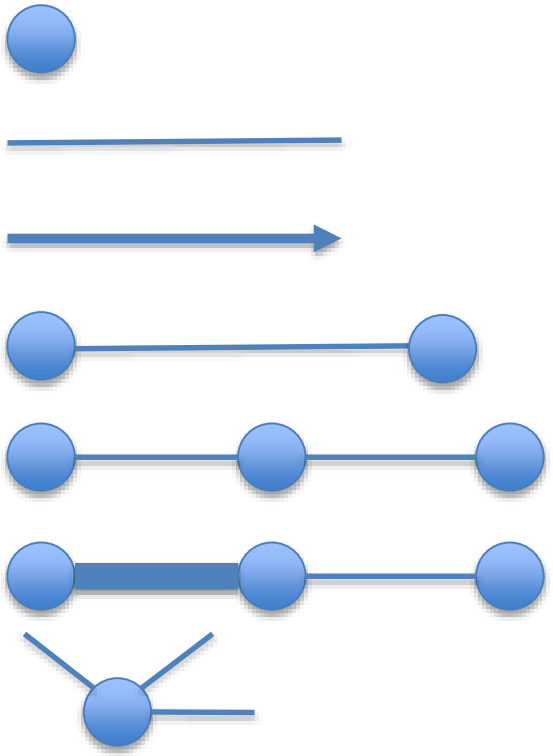
What is a network?

What is a network?

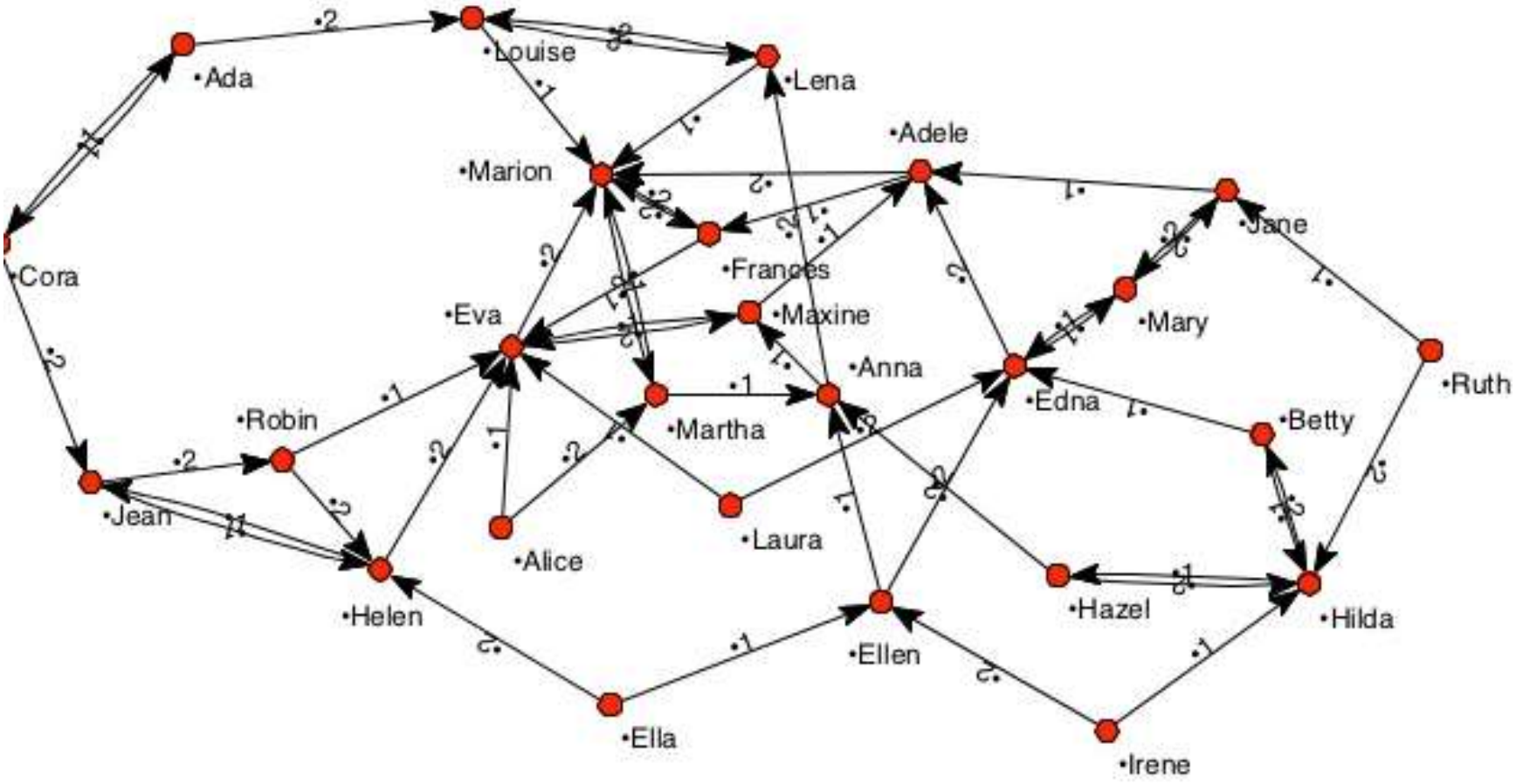


Basics

- Vertex/node
- Edge
- Directed
- Connectivity
- Path
- Weight
- Degree

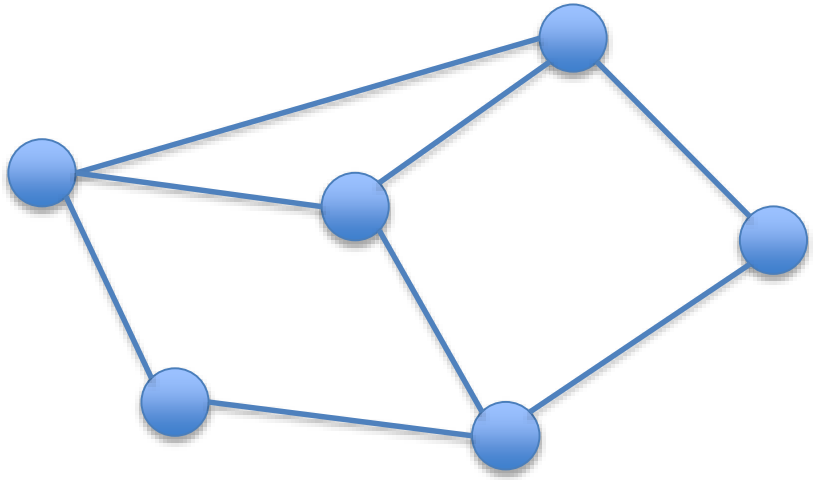


Basics – directed/weighted

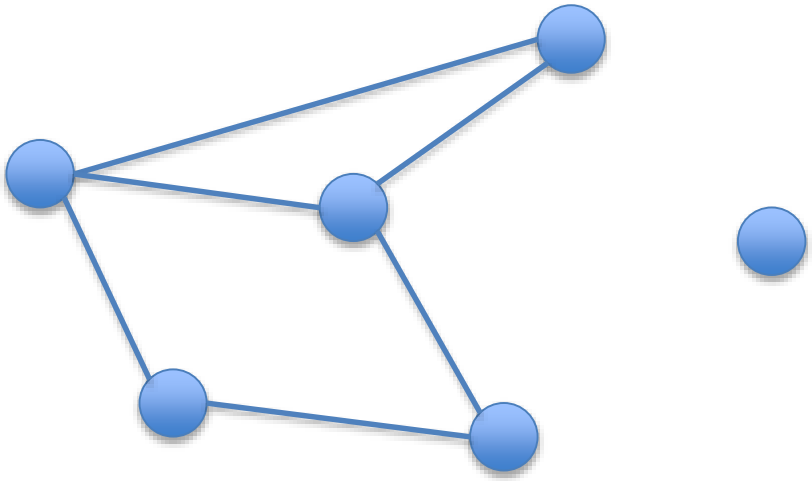


Describing a network

- Connected

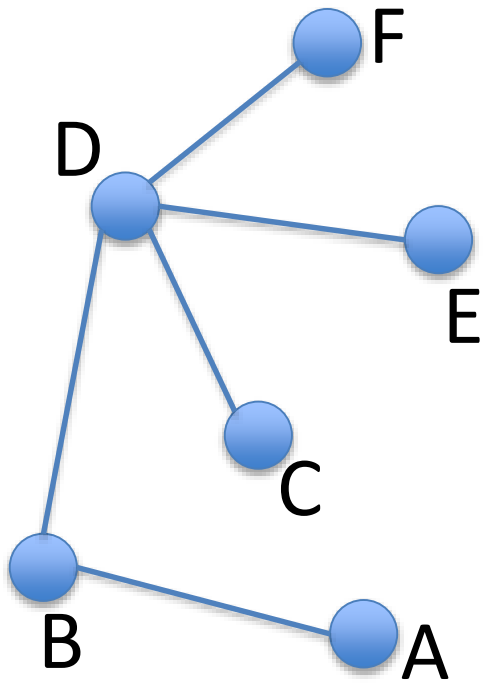


- Unconnected



Describing a network

- Degree distribution



1 node with 4 edges

1 node with 2 edges

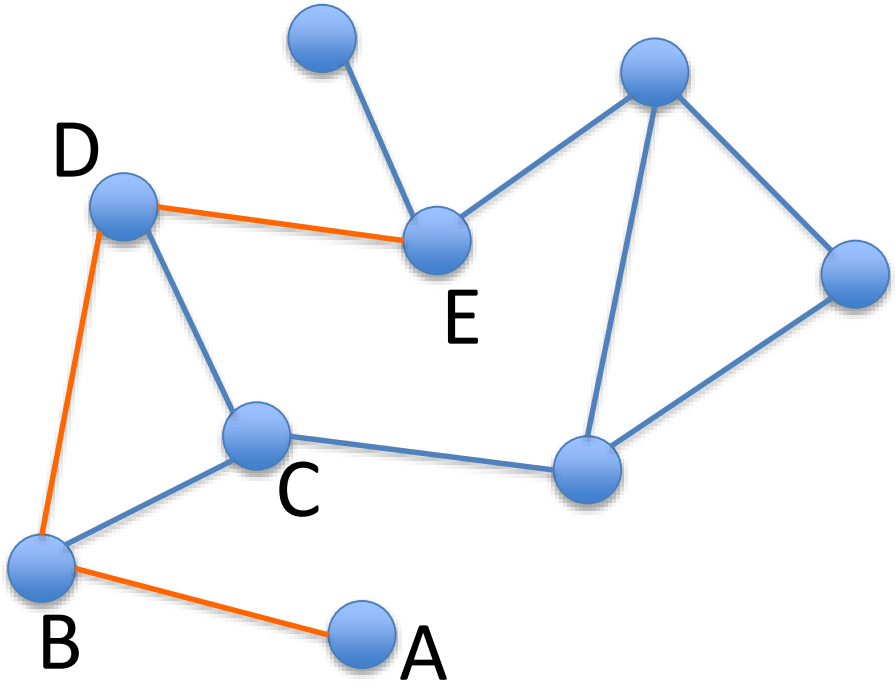
4 nodes with 1 edge

Distribution:

[(1: 4), (1: 2), (4: 1)]

Describing a network

- Average shortest path A to E



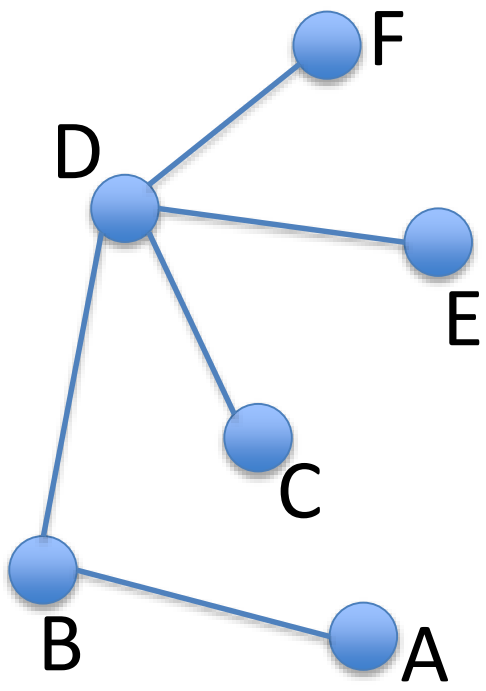
A – B – D – E = 3 hops

A – B – C – D – E = 4 hops

Describing a network

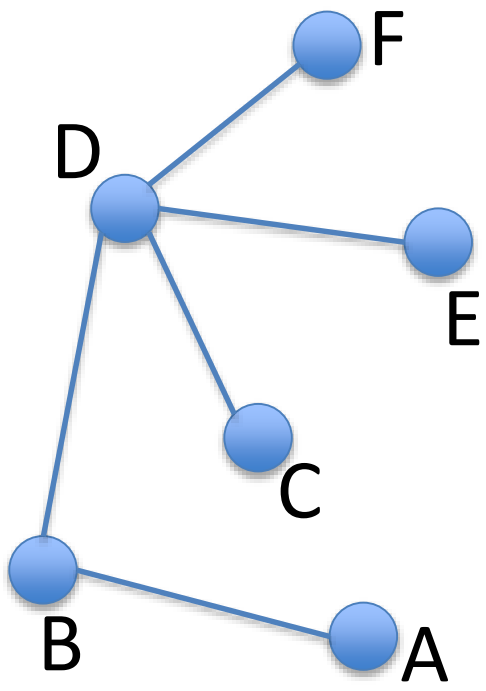
- Centrality
  - Degree: number of connections
  - Betweenness: number of shortest paths from all nodes to all others that pass through a particular node
  - Closeness: average length of the shortest paths between a specific node and all other nodes in the graph

Describing a network



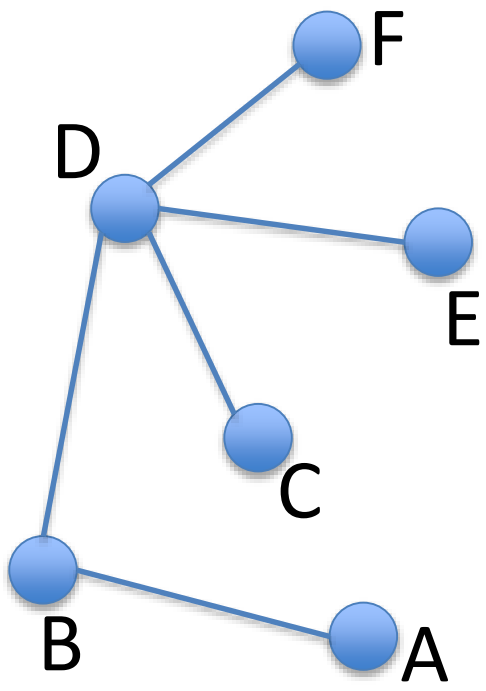
- Degree centrality
  - Most edges == most important
  - D: 4 edges
  - Normalized degree:
    - divide by maximum possible degree (n – 1)
    - 6 nodes means 5 possible connections
    - 4 / 5 = 0.8

Describing a network



- Betweenness centrality
  - Between many pairs of nodes
- D: between 9 pairs
  - AC, AE, AF, BC, BE, BF, CE, CF,EF
- Normalized
  - number of shortest paths divided by:
  - $[(n-1)(n-2)/2]$
  - $[(6-1)(6-2)/2] = 10$
  - D: 9/10 = 0.9

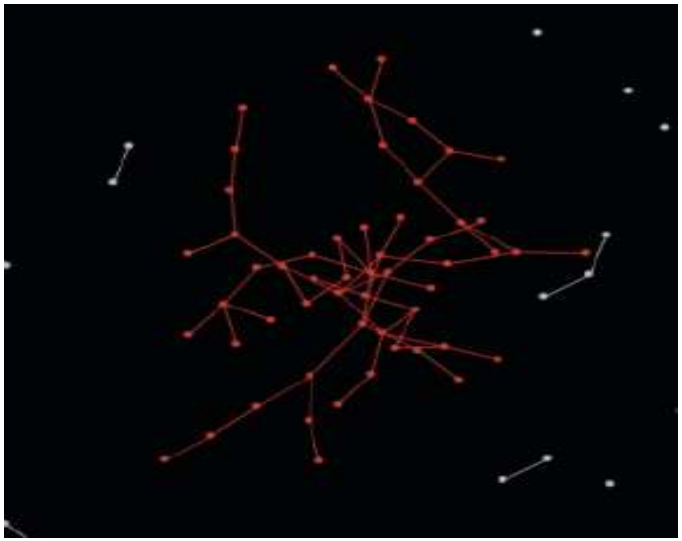
Describing a network

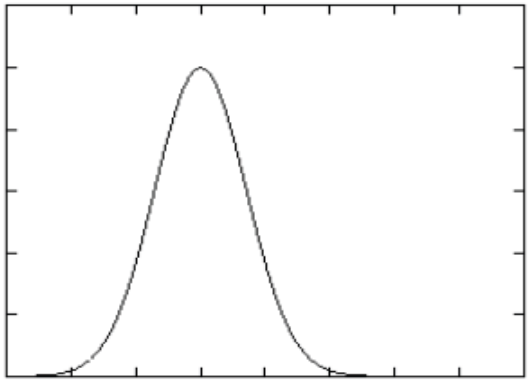


- Closeness centrality
  - Average length of shortest paths
- n – 1 / (sum of all shortest paths)
- D: 6 – 1 / (1 + 1 + 1 + 1 + 2) = 0.83
- A: 6 - 1 / (1 + 3 + 2 + 3 + 3) = 0.43

Modeling networks

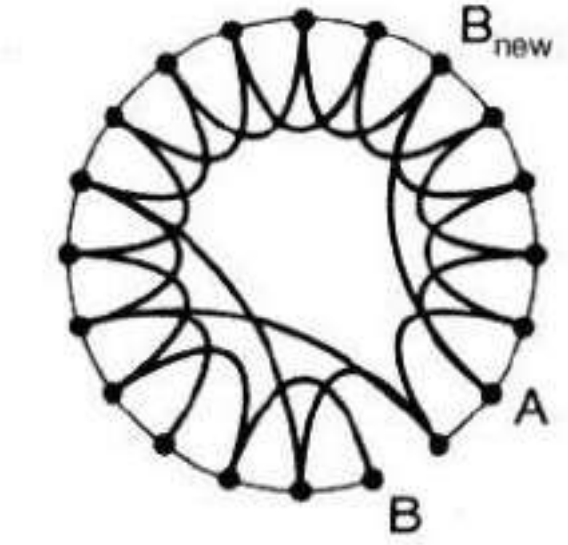
- Random (Erdos-Renyi) network
- Nodes connected at random
- Binomial distribution of edges connected to each node





Modeling networks

- Small world network
- Six degrees of separation
- Dense subgraph



Modeling networks

- Scale-free networks
- Power law distribution when scaled up – looks the same no matter the scale