Data Analytics in Business

Social Networks

Sridhar Narasimhan
Professor
Scheller College of Business

Social media

GTx

Lessons

A. Social Media
B. What is a Social Network?
C. Walks, Trails, Paths
D. Directed Graphs (Digraphs)
E. Centrality

GTx

Useful Resources for Social Network Analysis

- Gephi.org (download Gephi 0.9.2)
- igraph, tnet, sna packages in R

GTx

Social Media

- What do social media tools, services, and platforms enable?
  - Interactions among individuals
  - Interactions among individuals and businesses
- Which could lead to
  - Creation, sharing, and/or exchange of news, content, and ideas in virtual networks

GTx

Examples of Social Media Services

- Wikis – e.g., Wikipedia
- Blogs – e.g., Wordpress
- Microblogs – e.g., Twitter
- User generated content hosting and sharing sites – e.g., YouTube, flickr, Instagram
- Messenger tools – e.g., Whatsapp, WeChat, etc.
- Social Networks – e.g., Facebook
- And others!

GTx

Some Aspects of Social Media

- Viral dissemination of information
- Used by individuals, businesses, organizations, government, etc.
- Social influence
  - How can we measure it?
- User generated content impacting consumer behavior

GTx

How Do Companies Leverage Social Media for Business?

- They can use it within their websites, especially when they want to promote online interactions
- Example: eCommerce sites that offers online review tools, such as ratings, uploading of images/videos, reviews, comments on reviews, etc.

★★★★★ **Product works as advertised.**
by user123 on April 11, 2018

1000 of 1800 people found this review helpful
10 comments. Was this review helpful to you? Yes No

GTx

How Do Companies Leverage Social Media for Business?

- They can use or monitor the established social media platforms – e.g., Facebook pages for businesses.

GTx

How to Analyze Social Media Data

- Social network analysis
  - Understand how different players using the social media are connected
  - Determine the business value and implications of these connections
- Text Analytics
  - Understand the content of the textual data encountered in social media
  - Determine what businesses can learn from textual data in social media

GTx

Data Analytics in Business

Social Networks

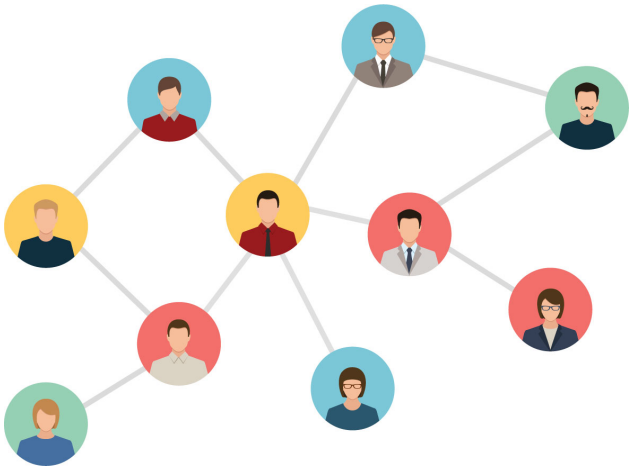
Sridhar Narasimhan
Professor
Scheller College of Business

What is a Social Network?

GTx

What is a Social Network?

- Nodes = social entities or "actors"
  - Individuals
  - Organizations
  - Groups (e.g., departments)
  - Other social units
- Edges = social ties



GTx

How are Social Networks Modeled?

- Graphs
- Directed graphs
- Matrices (we don't study these)

GTx

Undirected Graphs

- **Undirected** links that capture dichotomous relationships between **nodes**
- A link between nodes *i* and *j* represents a relationship
- A relationship could represent formal or informal relationships such as friendship, kinship (e.g., marriage), membership in an organization, proximity, etc.
- Nodes are also referred to as **vertices** or **points**
- Links are also referred to as **lines or edges**
- The maximum # of possible edges in a graph with N nodes = N(N-1)/2
- A graph with N(N-1)/2 edges is called a **complete** graph

GTx

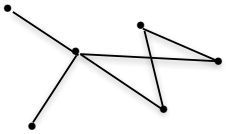
Graphs

- Each edge in a graph is an unordered pair of nodes.
- So each link, or edge, l_k can be written as (n_i, n_j) and this is identical to (n_j, n_i). These two nodes n_i and n_j are said to be **adjacent**. Node n_i (and node n_j) are said to be **incident** with edge l_k.
- We do not permit more than one link between two nodes.
- We typically do not include loops – i.e., we don't have a line (n_i, n_i)
- Graphs with these properties are called **simple** graphs
- A graph with only one node is **trivial**; otherwise it is **non-trivial**.
- A graph with no links is **empty**.

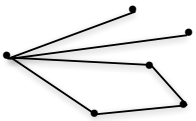
GTx

Visualizing Graphs

- A graph can be represented by a diagram in which the nodes are shown as points and the edges by lines.
- A line is drawn between two points if there is a link between them.
- The location of the points on a diagram is *arbitrary*.
- Graphs A and B below are **isomorphic** and are identical on all graph theoretic properties. They have the same # of nodes and links, network, diameter, etc.



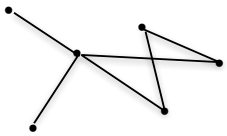
Graph A



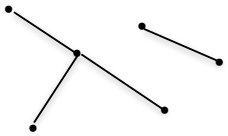
Graph B

GTx

Connected vs. Disconnected Graphs



Connected Graph



Disconnected Graph
with Two Components

- A graph is called **connected** if there is a path (a set of links) between every pair of nodes.
- The example to the right has two **components**. So nodes in one component can't communicate with nodes in the other component.
- The nodes *within* each component of a disconnected graph are connected.

GTx

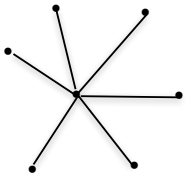
Dyad vs. Triad

- A **dyad** refers to a pair of nodes and the possible link between them in a graph. The two nodes in a dyad are either adjacent or not adjacent.
- A **triad** refers to three nodes in a graph. There are four possible states as show below:

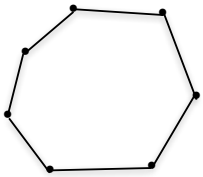
0 edges 1 edge 2 edges 3 edges

GTx

Star, Circle, and Line Graphs



Star Graph



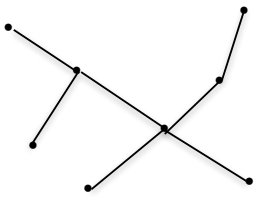
Circle Graph

Line Graph

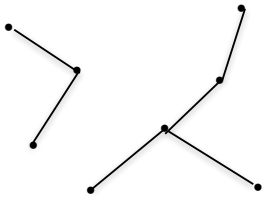
- One node in the **star** graph is more central than the others.
- All nodes in the **circle** graph are identical on centrality.
- In a **line** graph, the nodes in the middle are more central than peripheral nodes.

GTx

Trees vs. Forest



Tree



Forest

- A connected graph with no cycles is called a **tree**.
- The number of lines (edges) in a tree is equal to the # of nodes – 1
- There is only **one** path between any two nodes in a tree.
- A disconnected graph where each component is a tree is called a **forest**.

GTx

Degree

- The **degree**, *d(i),* of a node *i* is equal to the # of edges that are incident with it.
- *d(i)* can range from 0 to (N - 1), where N is the # of nodes in the graph
- **Average nodal degree**, $\bar{d}$, is equal to the average degree of the nodes in a graph

  $\bar{d} = \frac{\sum_{i=1}^{N} d(i)}{N} = \frac{2L}{N}$, where L = # of edges in the graph

- **Variance** of the node degrees = $\frac{\sum_{i=1}^{N}(d_i - \bar{d})^2}{N}$
- If all nodes have the same degree, then variance = 0.

GTx

Density

- The **density** of a graph is the ratio of the actual # of edges (L) to the maximum # of possible edges (N(N-1)/2)

  Density = $\frac{L}{(\frac{N(N-1)}{2})} = \frac{2L}{N(N-1)}$

- The density of a graph equals 0 if there are no links present and has a maximum value of 1 if the graph is complete.
- We know that average degree $\bar{d} = \frac{2L}{N}$, therefore, density $= \frac{\bar{d}}{(N-1)}$

GTx

Quiz (True/False)

- One node in the star graph is more central than the others.
  Answer: **TRUE**

- **Links** are also referred to as **vertices** or **points**.
  Answer: **FALSE**. Nodes are also referred to as **vertices** or **points** and links are also referred to as **lines** *or* **edges**.

GTx

Data Analytics in Business

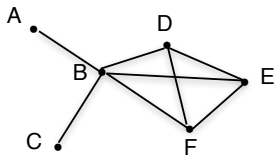
Social Networks

Sridhar Narasimhan
Professor
Scheller College of Business

Walks, Trails, Paths

GTx

Walks, Trails, and Paths in a Graph

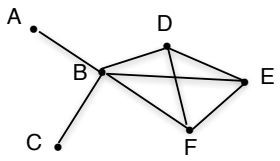


Graph Z

- A **walk** starts at a node, consists of edges (lines) and nodes, and ends at a node. Some nodes and lines may be included more than once.
- The length of a walk is the number of edges. For example, B-D-E-B-F-D-B-A is a walk with a length of 7.
- A **closed walk** begins and ends at the same node, e.g., B-D-E-F-D-B.
- A **cycle** is a walk with at least 3 nodes with distinct lines, e.g., B-D-E-B.
- A **tour** is a walk that includes all the edges at least once.

GTx

Walks, Trails, and Paths in a Graph

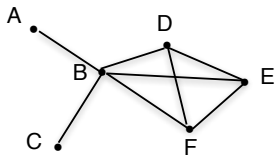


Graph Z

- A **trail** is a special kind of walk with distinct edges (lines). Some nodes may be included more than once.
- The length of a trail is the number of edges. For example, B-D-E-B-A is a trail with a length of 4.
- Note that the walk mentioned on the previous slide – B-D-E-B-F-D-B-A – is not a trail because edge DB appears twice.

GTx

Walks, Trails, and Paths in a Graph

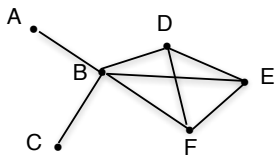


Graph Z

- A **path** is a special kind of walk with distinct edges (lines) and nodes. The length of a path is the number of edges. For example, B-D-E-F is a path with a length of 3.
- Note that the trail mentioned in the previous slide – B-D-E-B-A – is not a path because node B appears twice.

GTx

Geodesic or Shortest Path in a Graph

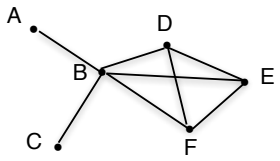


Graph Z

- B-D-E-F is a path between B and F with a length or distance of 3.
- There may be more than one path between a pair of nodes. For example, between B and F, we have B-D-E-F, B-D-F, or B-F, with distances of 3, 2, and 1, respectively.
- A **geodesic** is the shortest path, i.e., the path that has the minimum # of edges between a pair of nodes. The geodesic distance between B and F is d(B,F) = 1.
- Shortest Paths (Geodesics) are used in some centrality metrics

GTx

Geodesic or Shortest Path in a Graph

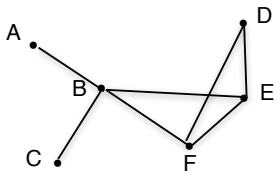


Graph Z

- Note that it is possible that there is more than one shortest path between a pair of nodes.
- If a graph is not connected, then the geodesic between at least one pair of nodes is infinite.

GTx

Diameter of a Graph



Graph Y

- The **diameter** of a graph (that is connected) is the largest geodesic between any pair of nodes

Geodesic Distances

d(A,B) = 1	d(C,D) = 3
d(A,C) = 2	d(C,E) = 2
d(A,D) = 3	d(C,F) = 2
d(A,E) = 2	d(D,E) = 1
d(A,F) = 2	d(D,F) = 1
d(B,C) = 1	d(E,F) = 1
d(B,D) = 2	
d(B,E) = 1	
d(B,F) = 1	

Because the maximum geodesic = 3, the **diameter** of Graph Y = 3

GTx

Quiz (True/False)

- Geodesic is the shortest path between a pair of nodes.
  Answer: **TRUE**.

- A path starts at a node, consists of edges (lines) and nodes, and ends at a node. Some nodes and lines may be included more than once.
  Answer: **FALSE**. A path does not have any repeating nodes and lines.

GTx

Data Analytics in Business

Social Networks

Sridhar Narasimhan
Professor
Scheller College of Business

Directed Graphs (Digraphs)

GTx

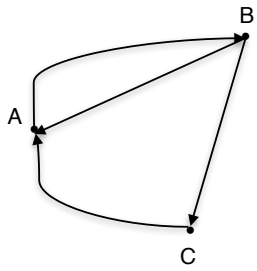
Directed Graphs

- When relations between actors (nodes) are directional, we need to use directed graphs or digraphs.
- These graphs have arcs or directed links between nodes.
- Examples:
  - Imports vs. exports between countries
  - Buyers vs. sellers
  - Influencers vs. followers
- A link (n_i, n_j) has node n_i as the source node and node n_j as the destination node of the link.
- Link (n_i, n_j) does not necessarily imply the existence of link (n_j, n_i)

GTx

Digraph Example

- A likes B, so there is directed link from A to B.
- B likes A, so there is directed link from B to A.
- B likes C, so there is directed link from B to C.
- C likes A, so there is directed link from C to A.
- Note, for example, that there is no link from C to B nor is there a link from A to C.

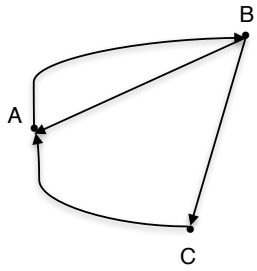


Digraph X

GTx

Node Indegree and Outdegree

- The **indegree** of a node is the number of arcs that terminate at that node:
  - The indegree of A is 2.
  - The indegree of node B is 1.
  - The indegree of node C is 1.
- The **outdegree** of a node is the number of arcs originating at that node:
  - The outdegree of A is 1.
  - The outdegree of node B is 2.
  - The outdegree of node C is 1.



Digraph X

GTx

Density of a Directed Graph

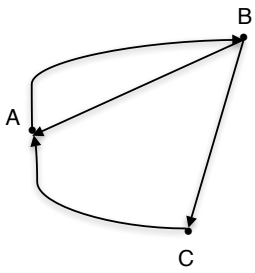
- The maximum # of possible arcs in a directed graph = N(N-1)
- If there are L directed arcs in a directed graph, then

  Density = $\frac{L}{N(N-1)}$
- Density ranges from 0 to 1

GTx

Directed Walks, Directed Paths, Directed Trails in a Directed Graph

- A **directed walk** is a sequence of nodes connected by arcs. Its length is the number of arcs it contains.
  - BABCAB is directed walk of length 5.
- A **directed trail** is a directed walk with no arc used more than once.
  - BABCA is a directed trail with a length of 4.
- A **directed path** is a directed walk with no repeating node or arc.
  - BCA is a directed path with length of 2.



Digraph X

GTx

Reachability and Geodesics

- If there is a directed path from node A to node B then node B is said to be **reachable** from node A.
- In a directed graph, the paths from node A to node B may be different from the paths from node B to node A (due to the direction of the arcs).
- A geodesic is the length of the shortest path between node A and node B (only if B is reachable from A).
- If there is a path between every pair of nodes in a directed graph, then the diameter is the length of the longest geodesic.
- It may be the case that some directed graphs have some non-reachable node pairs, in which case the diameter is undefined.

GTx

Quiz (True/False)

- In a directed graph, link (n_i, n_j) necessarily implies the existence of link (n_j, n_i).
  Answer: **FALSE.**

- The maximum # of possible arcs in a directed graph = N(N-1).
  Answer: **TRUE.**

GTx

Data Analytics in Business

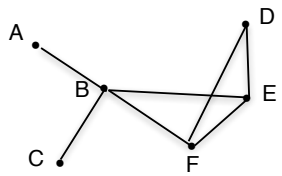
Social Networks

Sridhar Narasimhan
Professor
Scheller College of Business

Centrality

GTx

Social Network Analysis



Graph Y

- Which node (actor in a social network) has more power/influence?
- In Graph Y, nodes B and F are more central, while nodes A and C are peripheral.
- Social networks are complex, so we need to have metrics to better understand them.

GTx

Social Network – Metrics of Influence

- **Centrality** is one of the most common measures of a node (actor) influence in a network.
- It captures reach and social capital.
- We will study various centrality measures.

GTx

Social Capital

- **Social capital** is located in a social relationship.
- One can invest in or earn social capital and also maintain it over time.
- Sometimes one can convert social capital into other forms of capital (e.g., money, goods, ideas).
- Social capital can be a complement to or substitute for other resources.
- Social capital can also be a collective good of the network structure.

GTx

Why Study Centrality?

- We may need to identify the “important” actors or influencers in a social network.
- For example, the folks who may write blogs or tweets about a product or service.
- We want to develop measures to categorize social actors as important or not that important.
- These typically involve knowing the location of a node (actor) in a network.
- These measures are sometimes averaged to get group values.

GTx

Social Network – Metrics of Influence: Centrality

Centrality measures include:

- **Degree** centrality
- **Closeness** centrality
- **Betweenness** centrality
- Others:
  - Harmonic centrality
  - Eigenvector centrality (Katz and PageRank are variants)
  - Freeman centrality
  - Cross-clique centrality

GTx

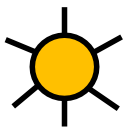
Social Network – Metrics of Influence: Centrality

- We will consider the situation where all edges (links) have weight 1
- This is the situation where all social ties are similar
- One could use a package like Gephi for situations where all edges have similar weight of 1
- Other packages in R allow for the computation of various centrality measures when the edges have different weights, e.g.,
  - tnet package
  - igraph package

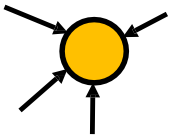
GTx

Node – Degree Centrality

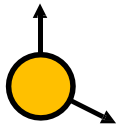
- Degree is the number of edges connecting to a node.
- How is degree centrality related to power?
  - Actors who have more ties (edges) have greater opportunities since they have more choices. They are less dependent on any other specific actor. Hence they are more viewed as more powerful.
  - Actors of higher degree could be approached for their knowledge or for advice, recommendations, money, etc.



Degree = 6



In Degree = 4



Out Degree = 2

GTx

Node – Closeness Centrality

- **Closeness centrality**, CC(*x*), is the inverse of the average geodesic distances (shortest paths) from an actor *x* (node) to all other actors:

$$CC(x) = \frac{N - 1}{\sum_{y \neq x} d(y, x)}$$

where *d(y, x)* is the length of the geodesic (i.e., shortest path) between node *y* and node *x* and *N* is the number of nodes.

- Actors who are able to reach other actors via shorter path lengths are more centrally-located.
- Closeness centrality captures how long it takes an actor to disseminate news over the social network.

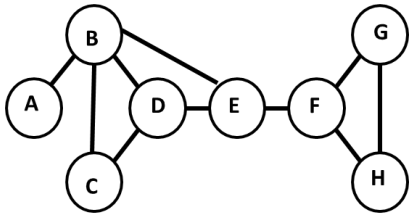
GTx

Node – Closeness Centrality

- Consider this graph.
- What is the closeness centrality of node A?

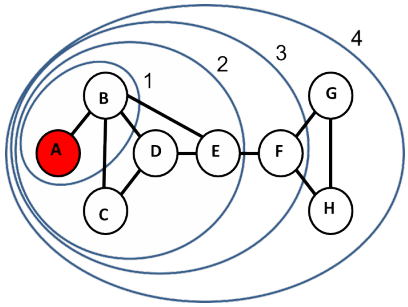
$$CC(A) = \frac{N - 1}{\sum_{y \neq A} d(y, A)}$$

- d(B,A) = 1
- d(C,A) = d(D,A) = d(E,A) = 2
- d(F,A) = 3, d(G,A) = d(H,A) = 4
- Denominator = 1 + 2 + 2 + 2 + 3 + 4 + 4 = 18
- Numerator = N – 1 = 8 − 1 = 7
- Hence CC(A) = 7/18 = 0.39



GTx

Node – Closeness Centrality



- If communication starts at A, it takes 4 stages (time periods) of passing info along before it reaches the entire network.
- Example: Think of stage 1 as an initial tweet to followers, followed by stages 2, 3, 4 as re-tweets.

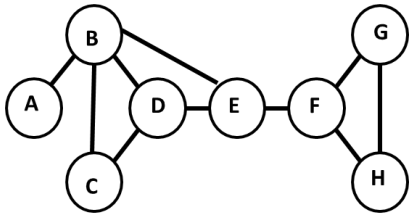
GTx

Node – Closeness Centrality

- Consider this graph
- What is the closeness centrality of node E?

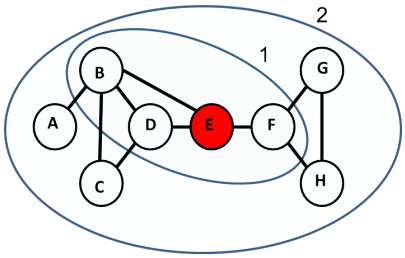
$$CC(E) = \frac{N - 1}{\sum_{y \neq E} d(y, E)}$$

- d(B,E) = d(D,E) = d(F,E) = 1
- d(A,E) = d(C,E) = d(G,E) = d(H,E) = 2
- Denominator = 1 + 1 + 1 + 2 + 2 + 2 + 2 = 11
- Numerator = N – 1 = 8 – 1 = 7
- Hence CC(E) = 7/11 = 0.64



GTx

Node – Closeness Centrality



- If communication starts at E, it takes only 2 stages (time periods) of passing information along before it reaches the entire network.

GTx

Node – Betweenness Centrality

- **Betweenness centrality**, *BC*, of a node v is given by the expression

$$BC(v) = \frac{\sigma_{st}(v)}{\sum_{s \neq v \neq t} \sigma_{st}}$$

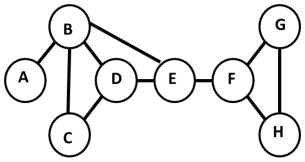
  where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that go through node v
- BC measures the capacity of an actor to "arrange" contacts among other actors, thus this actor has in-between power.

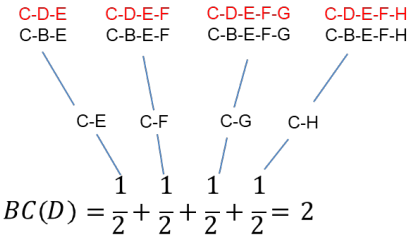
GTx

Node – Betweenness Centrality

Compute the betweenness centrality of node D

- Node D lies on the shortest paths (geodesics) between node C and nodes E,F,G, and H
- There are two shortest paths between C and E. They are CDE and CBE.
- There are two shortest paths between C and F. They are CDEF and CBEF.
- There are two shortest paths between C and G. They are CDEFG and CBEFG.
- There are two shortest paths between C and H. They are CDEFH and CBEFH.
- The shortest paths that go through D are shown in red.



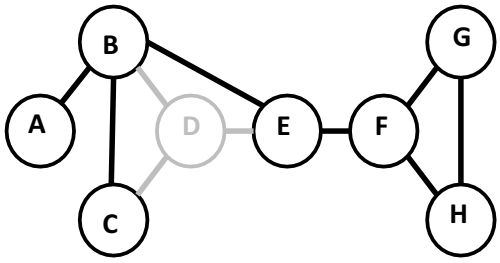


$$BC(D) = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = 2$$

GTx

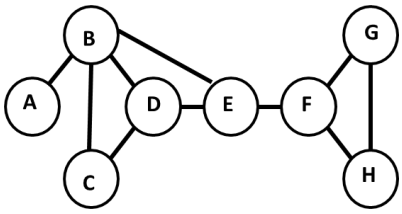
Node – Betweenness Centrality

- If D stops allowing flow (of info, goods, resources) then every other pair of nodes are still connected in some way



GTx

Node – Betweenness Centrality



- Compute the betweenness centrality of node E.

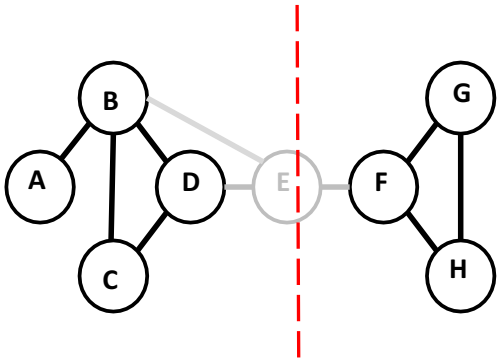
$$BC(E) = 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 = 12$$

A-F A-G A-H B-F B-G B-H C-F C-G C-H D-F D-G D-H

GTx

Node – Betweenness Centrality

- If E stops allowing flow (of info, goods, resources) then communication between any of nodes (A,B,C,D) and any of nodes (F,G,H) will cease.



GTx

Social Network Analysis in R

- Please reference these helpful tools (direct links will be provided on the course website):
  - R package for analysis and visualization of networks – igraph: https://github.com/igraph/rigraph
  - Tutorial: http://kateto.net/network-visualization
  - Book: https://www.amazon.com/Statistical-Analysis-Network-Data-Use/dp/1493909827/

GTx

Quiz (True/False)

- **Centrality** is one of the most common measures of a node's (actor's) influence in a network.

  Answer: **TRUE.**

- The betweenness centrality of a node x is given by $\frac{N-1}{\sum_{y \neq x} d(y,x)}$.

  Answer: **FALSE.** $BC(x) = \frac{\sigma_{st}(x)}{\sum_{s \neq x \neq t} \sigma_{st}}$

GTx

Recap of Lessons

A. Social Media
B. What is a Social Network?
C. Walks, Trails, Paths
D. Directed Graphs (Digraphs)
E. Centrality

GTx

Data Analytics in Business

Social Networks

Sridhar Narasimhan
Professor
Scheller College of Business

End

GTx