Analytics and Visualization of Big Data

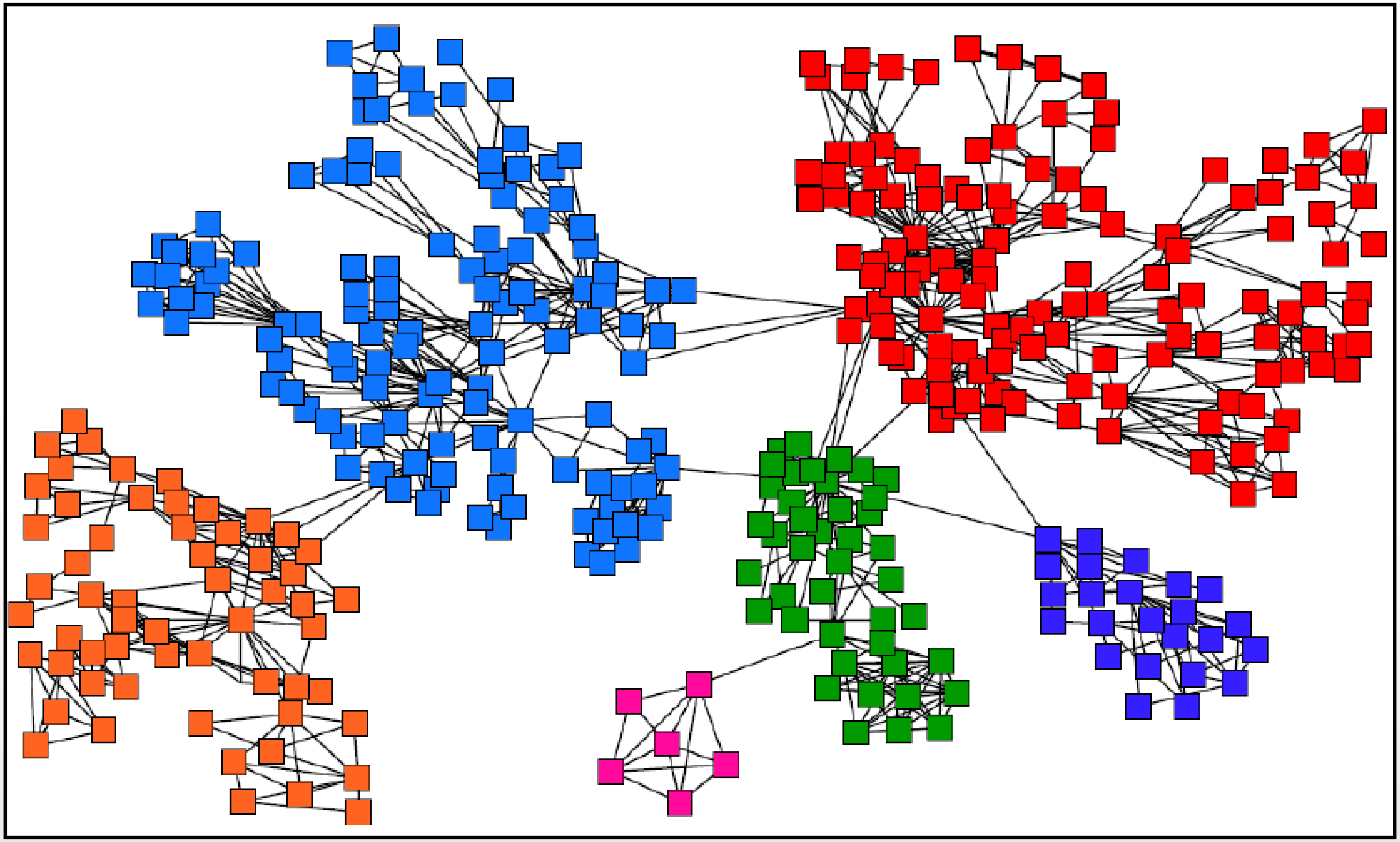
Fadel M. Megahed

*Lecture 22: Mining Social Network Graphs (Cont.)**

*Slides from Jeremy Ezell , PhD Candidate (College of Business) will be made available after seeking his permission

AUBURN UNIVERSITY
SAMUEL GINN
COLLEGE OF ENGINEERING

Goal: Finding Densely Linked Clusters



Source: Slide Adapted Jure Leskovic, Stanford CS246, Lecture Notes, see http://cs246.stanford.edu

Social Networks as Graphs

- What is a Social Network?
- Social Network as Graphs
- Varieties of Social Networks
- Graphs with Several Node Types

Clustering of Social Network Graphs

- Distance Measures
- Applying Standard Clustering Techniques
- Betweenness
- Using Betweenness to Find Communities

Direct Discovery of Communities

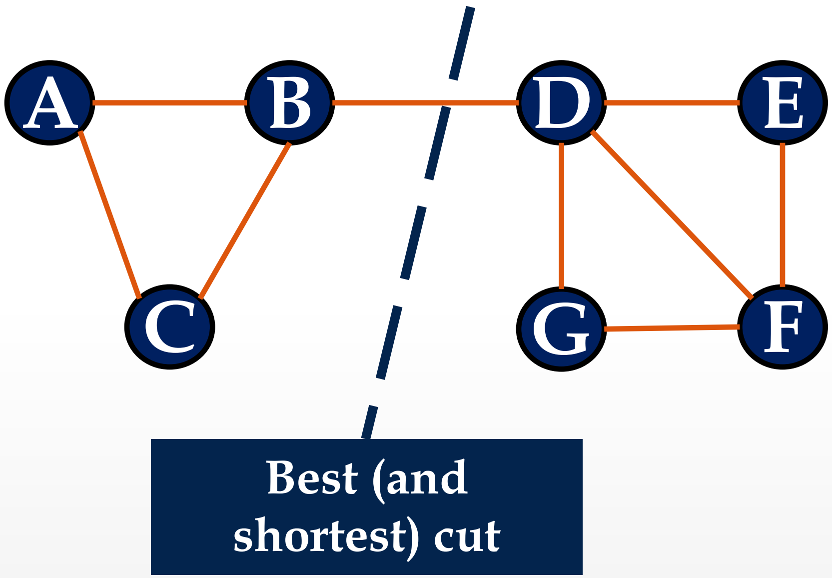
- Finding Cliques
- Complete Bipartite Graphs
- Finding Complete Bipartite Subgraphs

Partitioning of Graphs

- What makes a good partition?
- Normalized Cuts
- Some Matrices that Describe Graphs
- Eigenvalues of the Laplacian Matrix

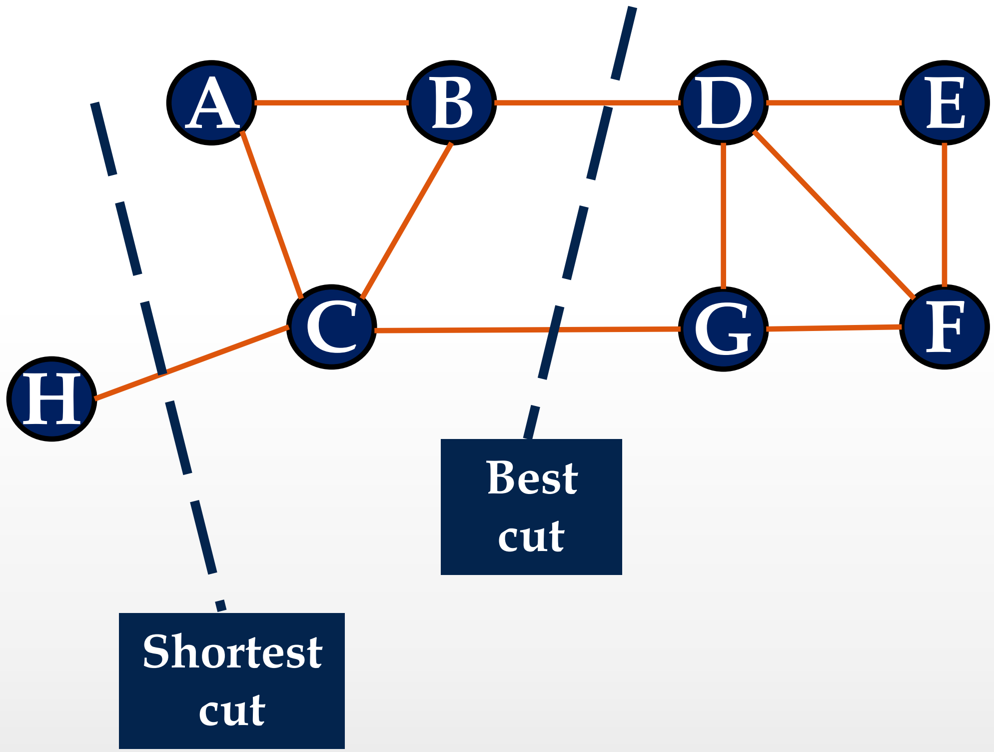
What makes a good partition?

- A good partition has the following properties:
  - Maximize the number of within-group connections
  - Minimize the number of between-group connections



What makes a good partition?

- A good partition has the following properties:
  - Maximize the number of within-group connections
  - Minimize the number of between-group connections



Typically, we want the clusters to be similar in size

Normalized Cuts

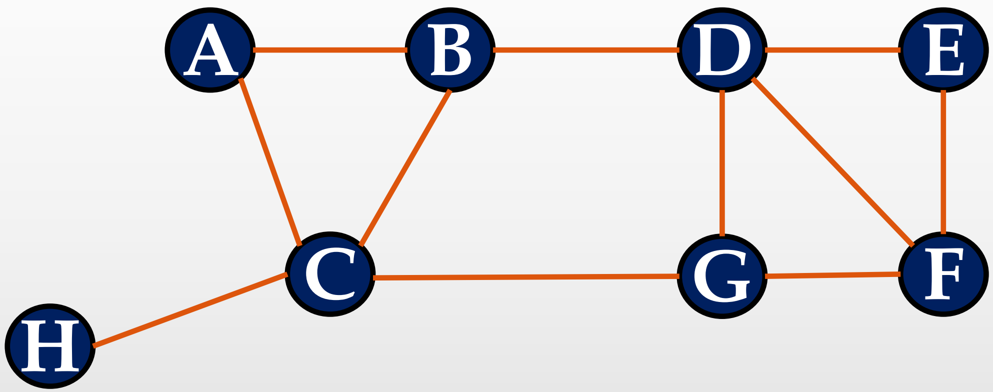
- A proper definition of a "good" cut must produce balanced sets.

- Suppose we want to divide the figure into two distinct sets of nodes: S and T, then the **normalized cut** is:

$$ncut(A,B) = \frac{cut(A,B)}{vol(A)} + \frac{cut(A,B)}{vol(B)}$$

Example: Identify the *ncut*:

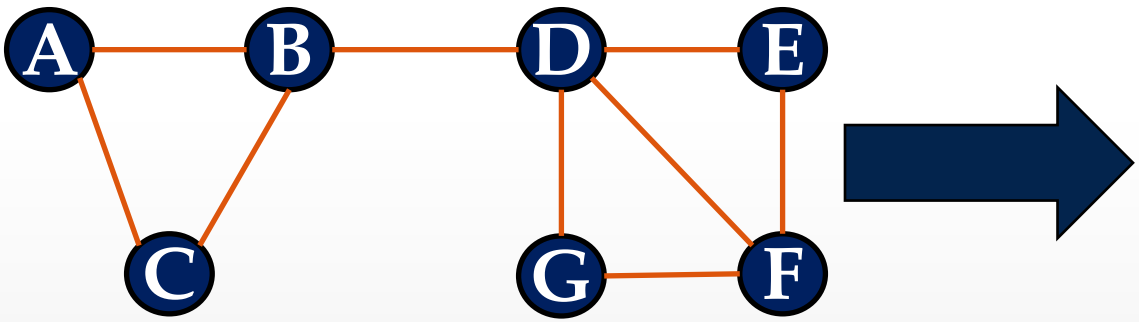
- Smallest cut
- Optimal cut



Some Matrices that Describe the Graphs

Adjacency Matrix (A)

- $n \times n$ matrix
- $A=[a_{ij}]$, a_{ij}= 1 if edge exists between node i and j
  0 otherwise



	A	B	C	D	E	F	G
A	0	1	1	0	0	0	0
B	1	0	1	1	0	0	0
C	1	1	0	0	0	0	0
D	0	1	0	0	1	1	1
E	0	0	0	1	0	1	0
F	0	0	0	1	1	0	1
G	0	0	0	1	0	1	0

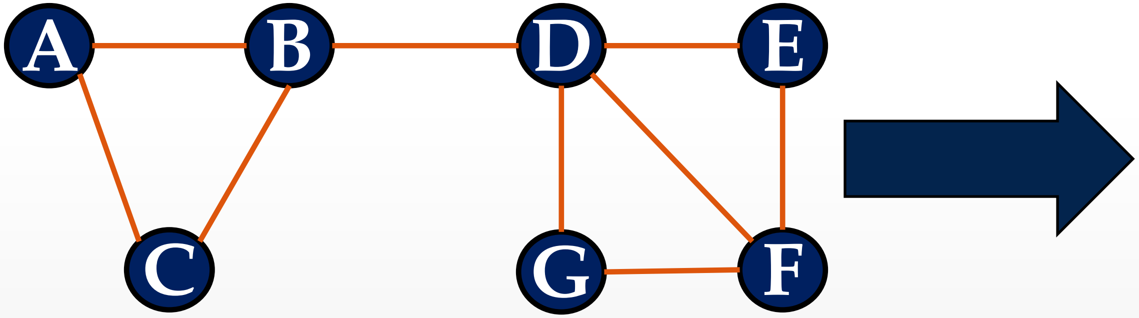
Important Property:

- **Symmetric Matrix**

Some Matrices that Describe the Graphs

Degree Matrix (D)

- $n \times n$ matrix
- $D=[d_{ii}]$, d_{ii}= degree of node



	A	B	C	D	E	F	G
A	2	0	0	0	0	0	0
B	0	3	0	0	0	0	0
C	0	0	2	0	0	0	0
D	0	0	0	4	0	0	0
E	0	0	0	0	2	0	0
F	0	0	0	0	0	3	0
G	0	0	0	0	0	0	2

Important Property:

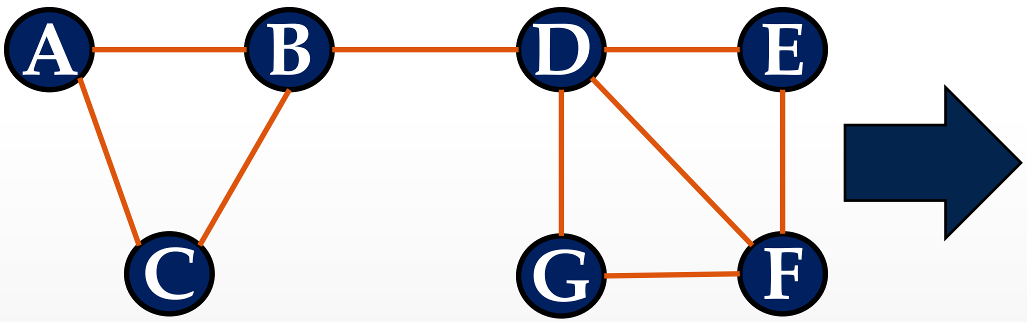
- **Diagonal**

Laplacian Matrix (L)

- $n \times n$ matrix
- $L = D - A = [l_{ij}]$, $l_{ii} = \mathrm{d}_{\mathrm{ii}}$

  l_{ij} = -1 if edge exists

  0 otherwise



	A	B	C	D	E	F	G
A	2	-1	-1	0	0	0	0
B	-1	3	-1	-1	0	0	0
C	-1	-1	2	0	0	0	0
D	0	-1	0	4	-1	-1	-1
E	0	0	0	-1	2	-1	0
F	0	0	0	-1	-1	3	-1
G	0	0	0	-1	0	-1	2

Eigenvalues of the Laplacian Matrix

- Partition based on the smallest eigenvector

Pre-processing:

- Build Laplacian matrix L of the graph

	1	2	3	4	5	6
1	3	-1	-1	0	-1	0
2	-1	2	-1	0	0	0
3	-1	-1	3	-1	0	0
4	0	0	-1	3	-1	-1
5	-1	0	0	-1	3	-1
6	0	0	0	-1	-1	2

Decomposition:

- Find eigenvalues λ and eigenvectors x of the matrix L

λ =

0.0
1.0
3.0
3.0
4.0
5.0

X =

0.4	0.3	-0.5	-0.2	-0.4	-0.5
0.4	0.6	0.4	-0.4	0.4	0.0
0.4	0.3	0.1	0.6	-0.4	0.5
0.4	-0.3	0.1	0.6	0.4	-0.5
0.4	-0.3	-0.5	-0.2	0.4	0.5
0.4	-0.6	0.4	-0.4	-0.4	0.0

- Map vertices to corresponding components of λ_2

1	0.3
2	0.6
3	0.3
4	-0.3
5	-0.3
6	-0.6

How do we now find clusters?

Exam I - Analytics

Deidentified data of the exam, Piazza and the homework are made available at the link below:

https://www.dropbox.com/sh/rnlphui76jqks5z/Ij1FnLgvy8

Feel free to use MATLAB to manipulate and/or generate additional insights ☺

Motivation for Machine Learning