Analytics and Visualization of Big Data

Fadel M. Megahed

Lecture 21: Mining Social Network Graphs

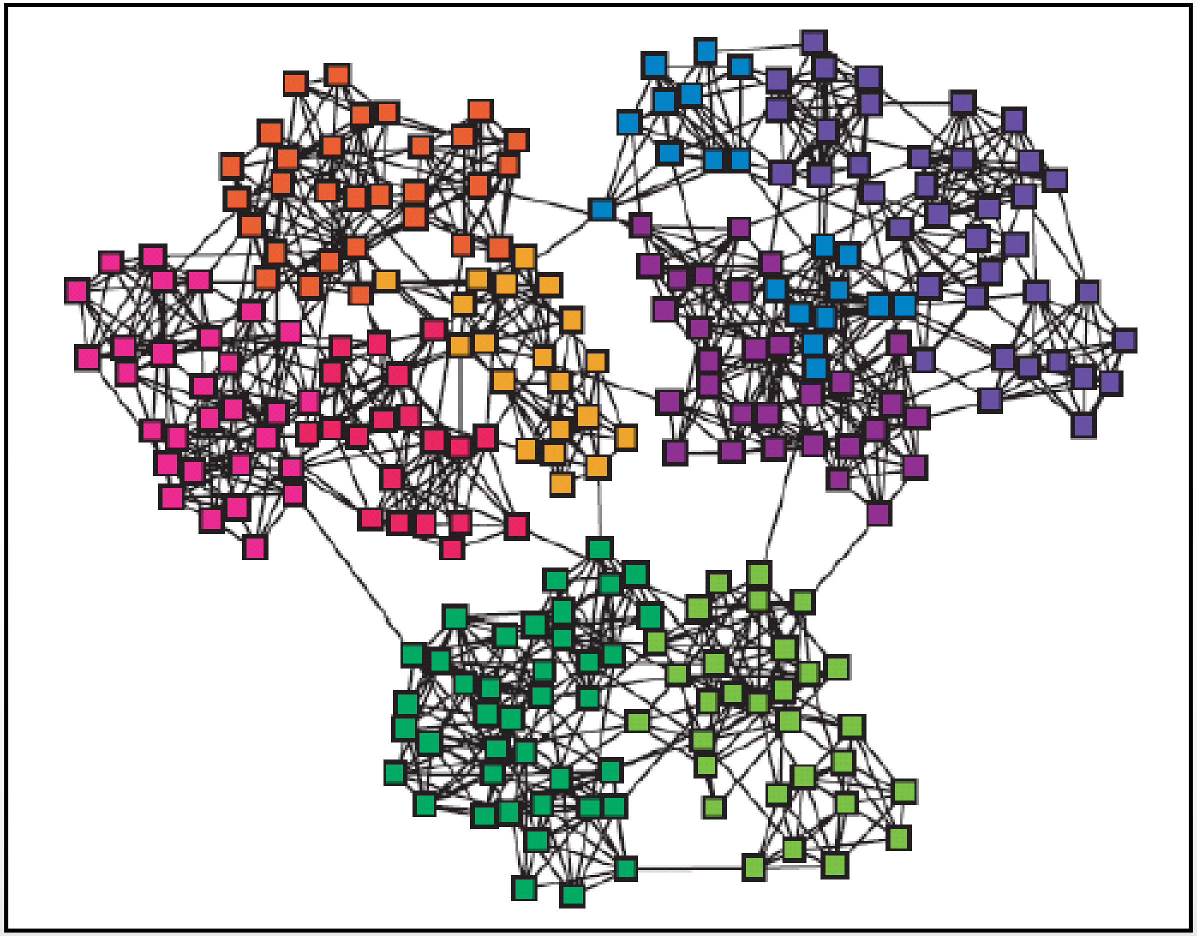
AUBURN UNIVERSITY
SAMUEL GINN
COLLEGE OF ENGINEERING

Department of Industrial and Systems Engineering

Spring 13

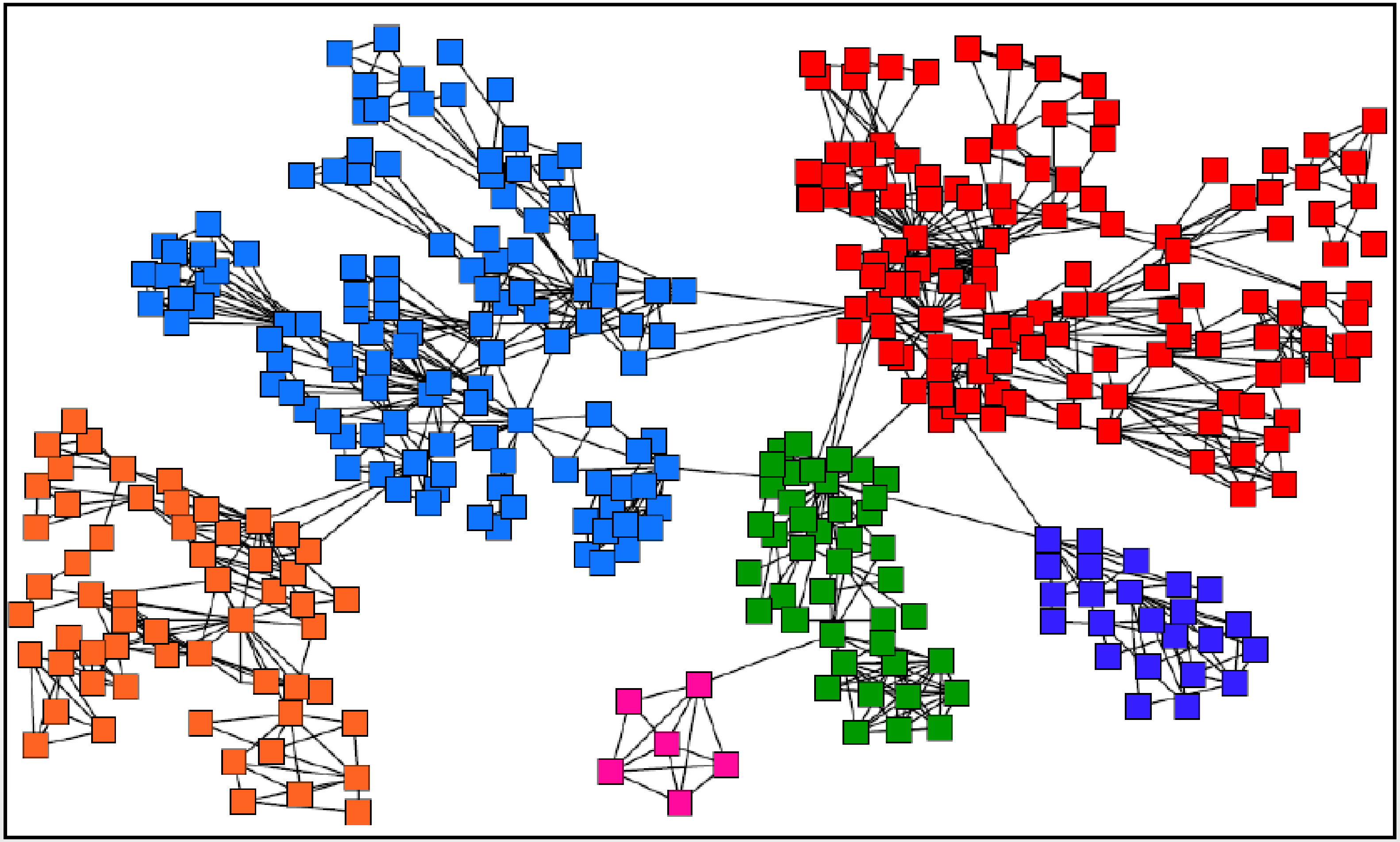
Preface: Network and Communities

We can think of networks as something looking like this:



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

Goal: Finding Densely Linked Clusters



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

Outline for Topics Covered in Chapter 10

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

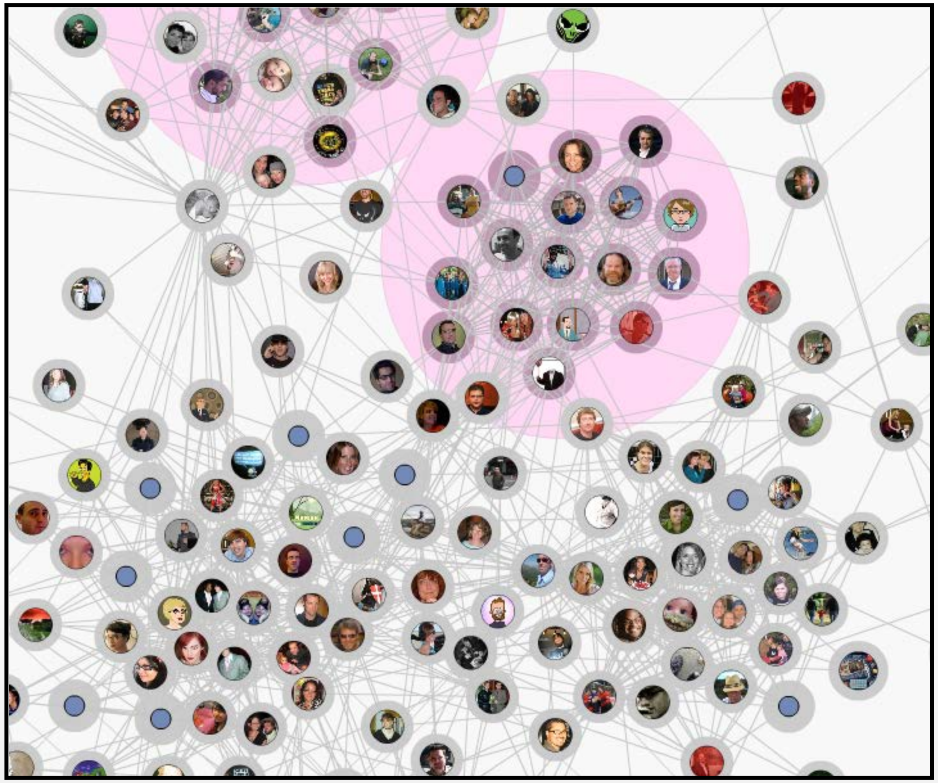
Partitioning of Graphs

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

Social Networks as Graphs

- Essential Characteristics of a social network are:
  - Building Block → Entities (Entities are typically people)
  - At least one relationship between entities of a network
  - Assumption: **<u>nonrandomness</u>**

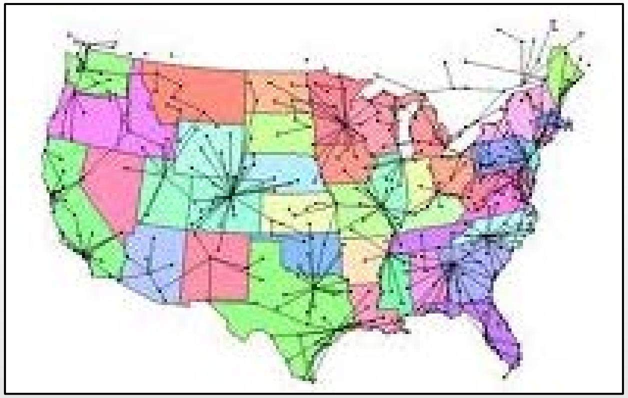
- Naturally modeled as undirected graphs:
  - Entities → Nodes
  - Nodes connected if there is a relationship between entities
  - Degree → labeling the edges



Source: Click here

Are Facebook/Twitter the only Networks that are Social?

- Other than "friends" networks, there are many examples that exhibit locality of relationships:
  - Telephone Networks
  - Email Networks
  - Collaboration Networks
  - Airport Networks

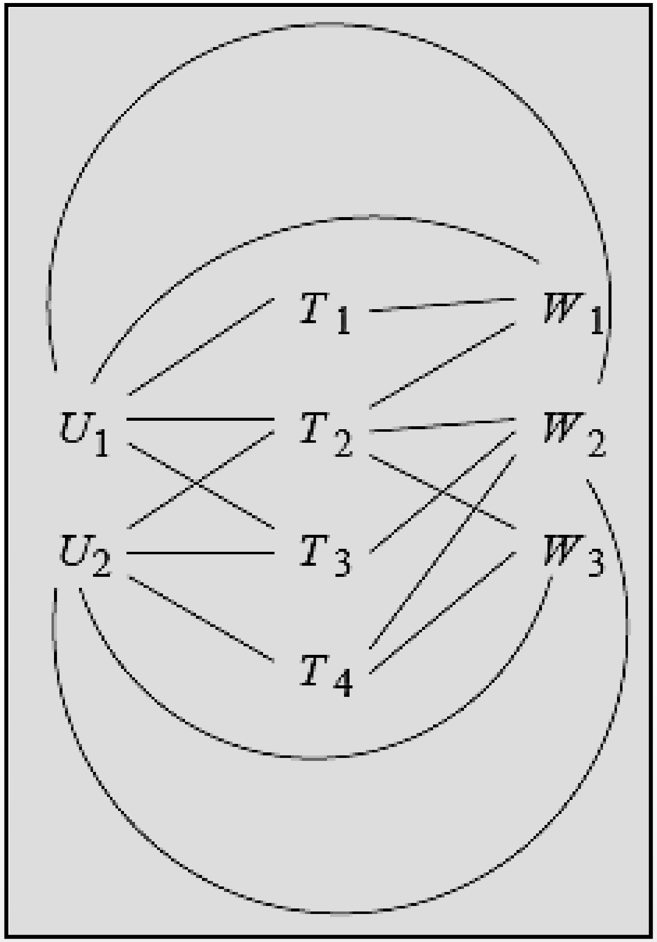


Graphs with Several Node Types

- Entities can be of different types (e.g. Facebook has people, pages, and network)

- A natural way to represent that is through a **k-partite graph**
  - Consists of k sets of nodes (no edges between nodes of same set)

- Figure: Deli.cio.us network



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

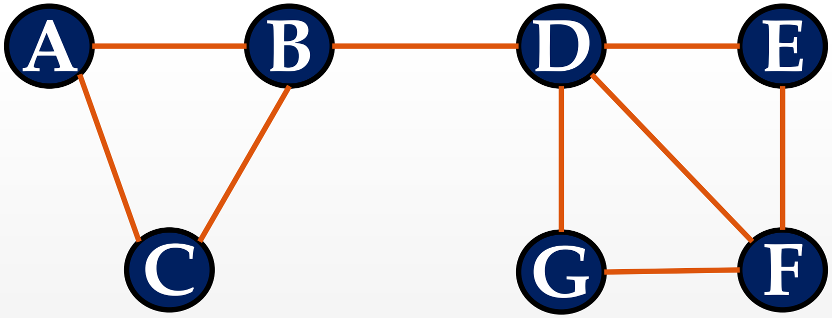
Partitioning of Graphs

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

How do you define a distance measure for a social network?

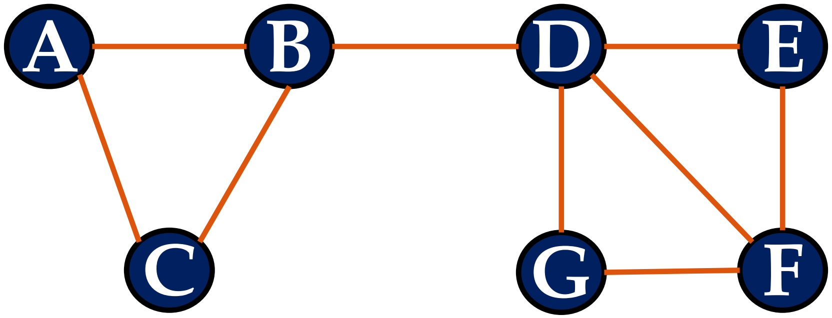
- If we are to apply standard clustering techniques, our **first step** would be to define a **distance measure**.

- **Question: What would be a suitable distance measure for the graph below?**
  - **Hint: The closer the nodes, the better ☺**



Problems in Hierarchical Clustering

- Consider the graph in the Figure below.



Questions:

1. **Based on the geometry of the graph, identify some of the large clusters that we can obtain.**

2. **By using hierarchical clustering (bottom-up approach), what is the probability that we cluster B and D together first?**

Problems with Point Assignment (*k*-Means Clustering)

- If we start by picking two points at random, they might be in the same cluster.

- If we start by picking one point at random, and select a point furthest away from it, we can still mess it up – **Example??**

- Even if we get two reasonable starting points, e.g. B and F, how will we assign D?

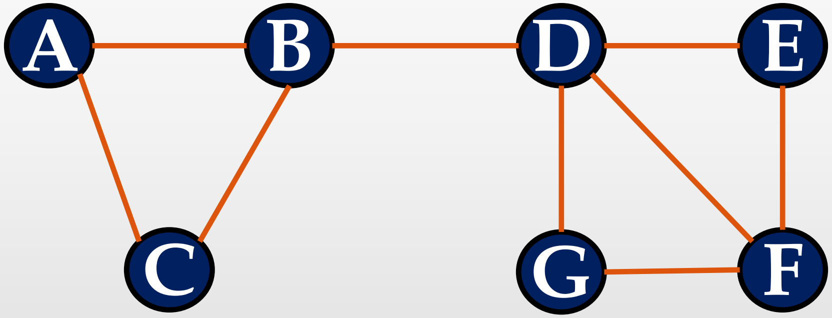
With a larger # nodes, the problem gets more complicated as you can imagine ☺

Betweenness

- **Definition:** Betweenness of edge (a,b) is defined as the number of pairs of nodes (x, y) such that edge (a,b) lies on the shortest path between them.

- **Properties:**
  - High Betweenness is bad!!
    - **Interpretation: High scores suggest that (a,b) runs between two different communities**

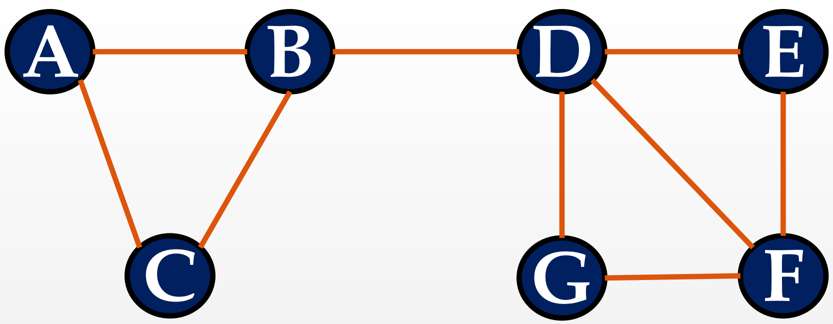
- Example/Exercise: Calculate the betweenness



Betweenness – An Observation

The betweenness score for edges of a graph behave something like (i.e. not exactly) a distance measure on the nodes of a graph. Therefore, we can cluster by taking out the edges in an increasing order of betweenness!!

Example:



What is the interpretation of the first and last sets of clusters?

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

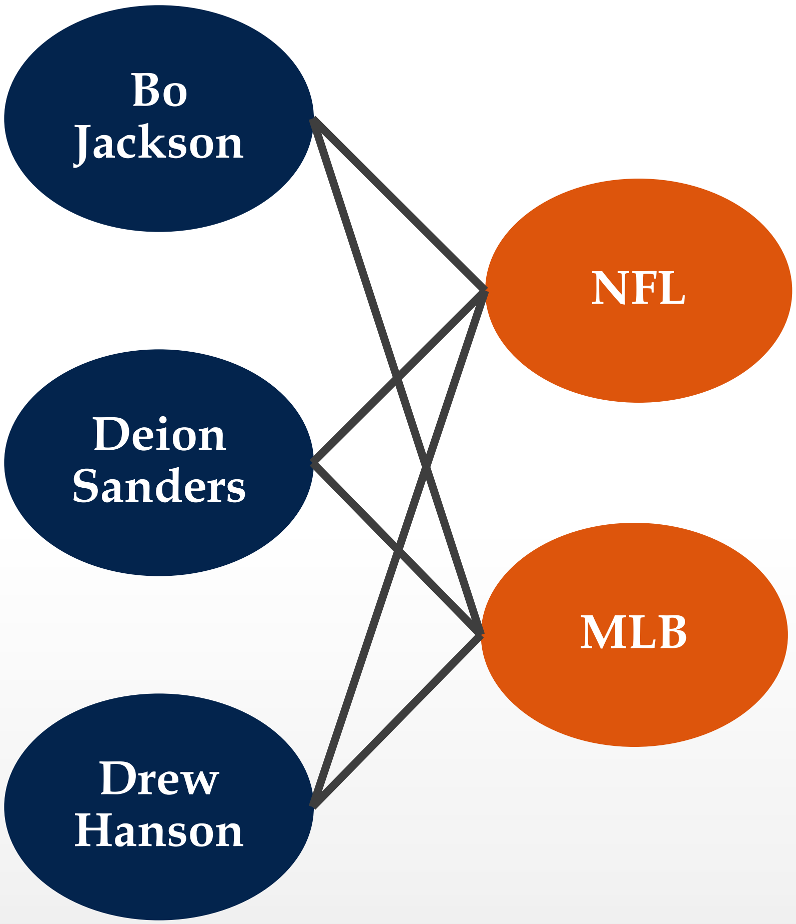
Partitioning of Graphs

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

Side Note: Complete Bipartite Graphs

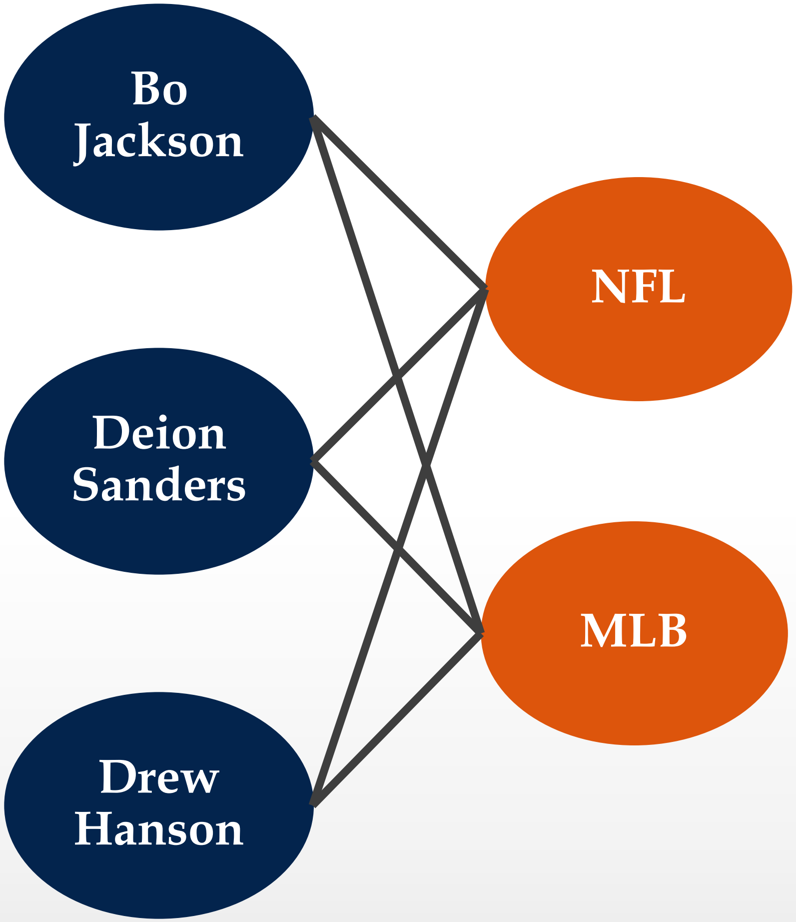
- **Definition:** It consists of ***s*** nodes on one side and ***t*** nodes on the other, with all possible edges between the nodes of one side and the other present.

- **It is possible that a bipartite graph with many edges has a large complete bipartite subgraph ☺**



Searching for Small Communities

- We want to enumerate complete bipartite subgraphs ($k_{s,t}$)
  - $k_{s,t}$: ***s*** nodes on the left and ***t*** nodes on the right
  - **Note that the book defines $k_{s,t}$ differently**
- In example, s=3 and t=2;
  - Note it is more efficient to have s<=t (i.e. rotate graph if we were to make any real calculations)



A 3 Step Plan

Two points:

(1) Dense bipartite graph: the signature of a community

(2) Complete bipartite subgraph $K_{s,t}$

- $K_{s,t}$ = graph on s nodes, each links to the same t other nodes

Plan:

How do we solve (2) in a giant graph?

Similar problems were solved on big non-graph data?

Details Regarding Frequent Itemset Enumeration

Setting:

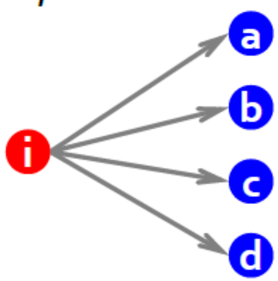
- **Market:** Universe U of n items
- **Baskets:** m subsets of U: $S_1, S_2, \ldots, S_m \subseteq U$ (S_i is a set of items one person bought)
- **Support:** Frequency threshold f

Goal:

- **Find all items in T that were bought together at least f times (T s.t. $T \subseteq Si$ of $\geq$ f sets Si)**
- **What's the connection between the itemsets and complete bipartite graphs?**

From Itemsets to Bipartite $k_{s,t}$

View each node i as a set S_i of nodes i points to



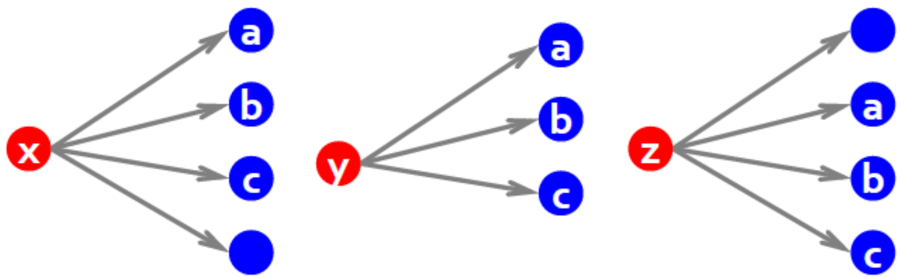
$S_i = \{a,b,c,d\}$

Find frequent itemsets:

- s … minimum support
- t … itemset size

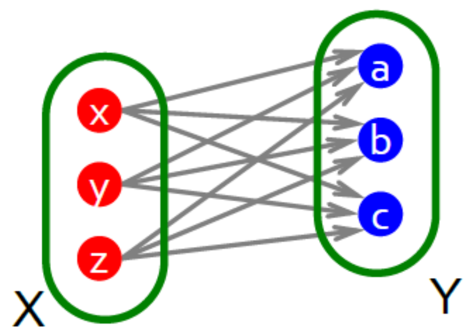
Say we find a **frequent itemset** $Y=\{a,b,c\}$ of supp. s

So, there are s nodes that link to all of {a,b,c}:



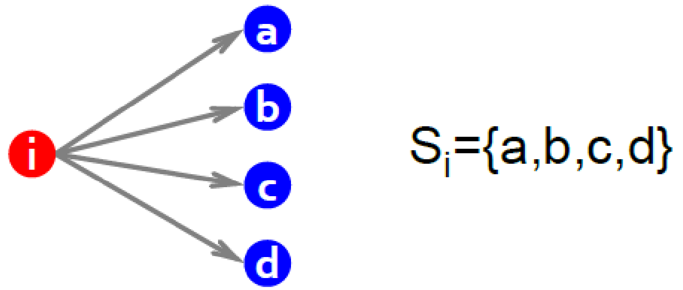
We found $K_{s,t}$!

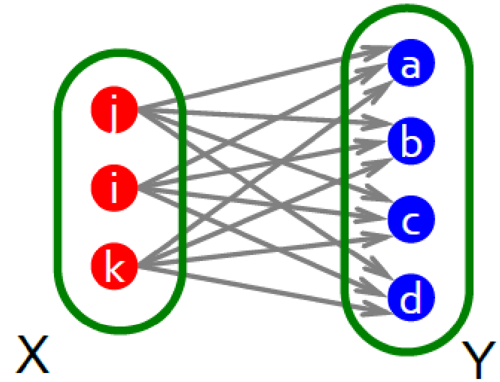
$K_{s,t}$ = a set Y of size t that occurs in s sets S_i



From Itemsets to Bipartite $k_{s,t}$

- **Itemsets finds Complete bipartite graphs!**
- **How?**
  - View each node i as a set S_i of nodes i points to
  - $K_{s,t}$ = a set Y of size t that occurs in s sets S_i
  - Looking for $K_{s,t}$ → set of frequency threshold to s and look at layer t – all frequent sets of size t





s … minimum support (|X|=s)
t … itemset size (|Y|=t)

From Itemsets to Bipartite $k_{s,t}$ – *Summary*

Analytical result:

- Complete bipartite subgraphs $K_{s,t}$ are embedded in larger dense enough graphs (*i.e.*, the communities)
  - Biparite subgraphs act as "signatures" of communities

Algorithmic result:

- Frequent itemset extraction and dynamic programming finds graphs $K_{s,t}$
  - **Method is super scalable**

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

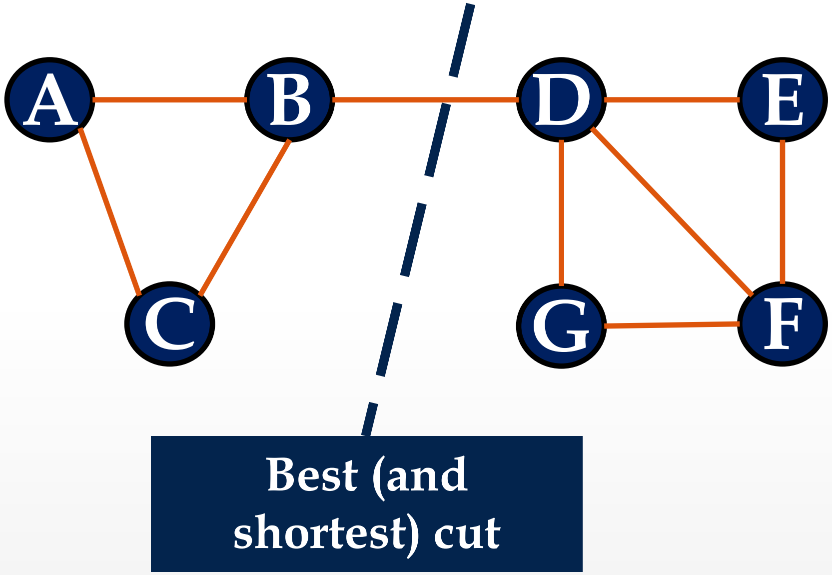
- Finding Cliques
- Complete Bipartite Graphs
- Finding Complete Bipartite Subgraphs

Partitioning of Graphs

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

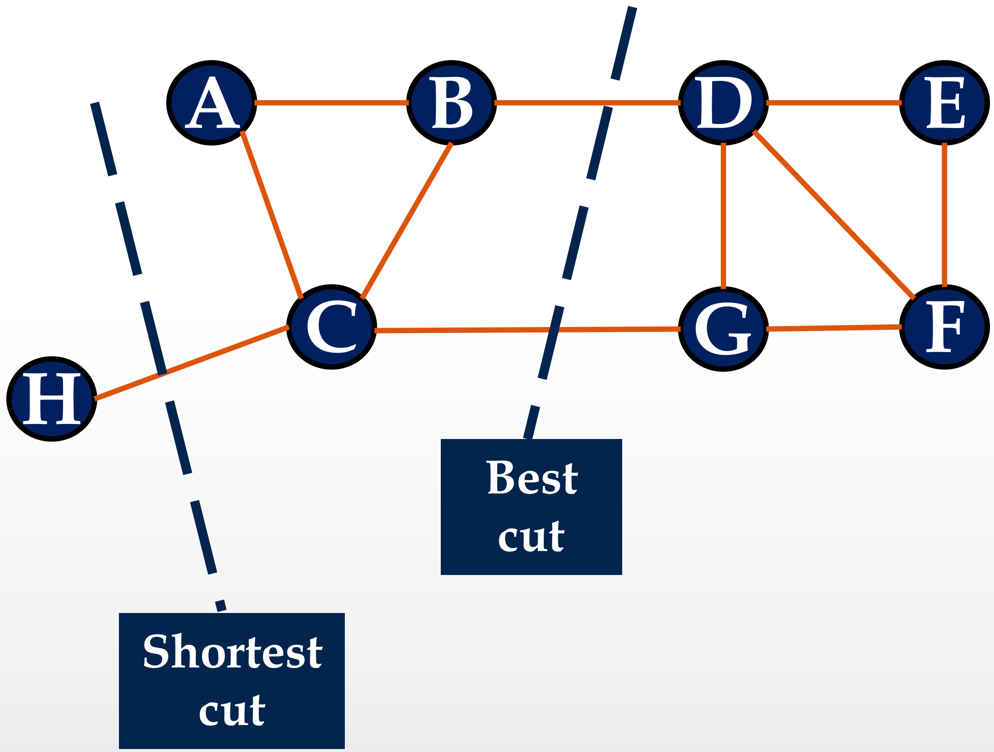
What makes a good partition?

- 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

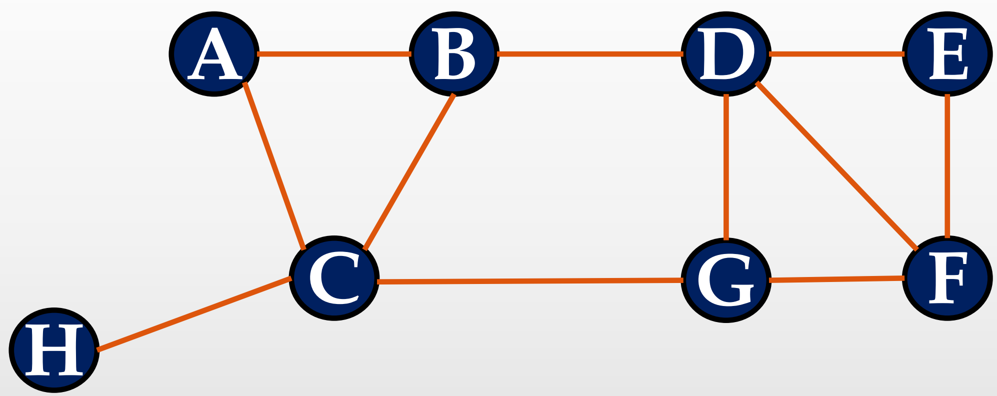
- A proper definition of a "good" cut must produce balanced sets.

- 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*:

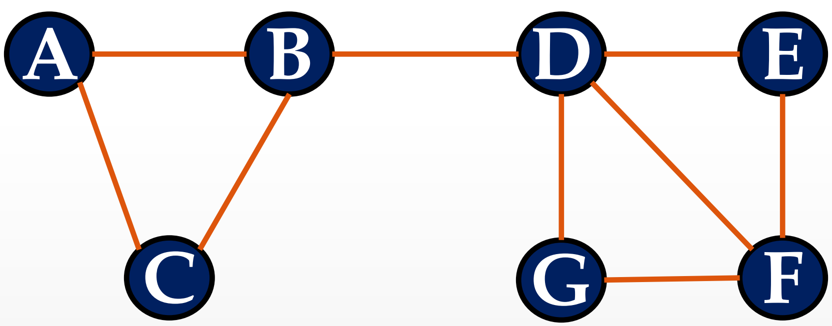
- Smallest cut
- Optimal cut



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

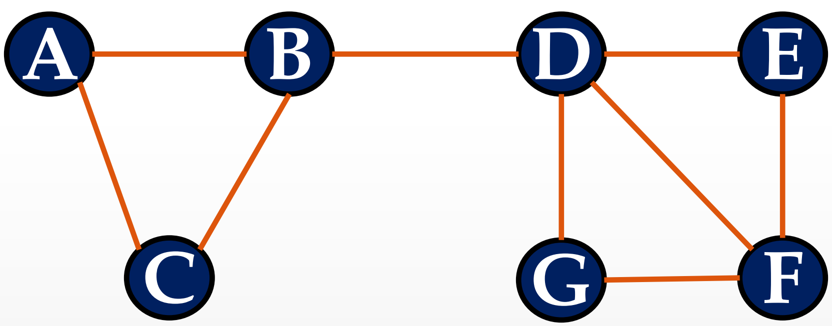
Important Property:

- **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**

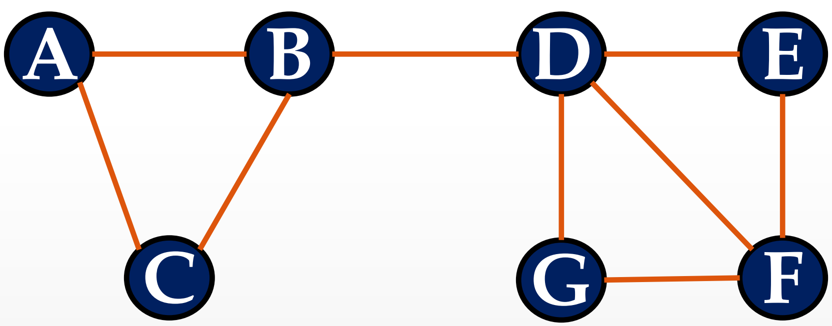
Some Matrices that Describe the Graphs

Laplacian Matrix (L)

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

  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
  - Map vertices to corresponding components of λ_2

$\lambda =$

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

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

How do we now find clusters?

Analytics and Visualization of Big Data

Fadel M. Megahed

Lecture 21: Mining Social Network Graphs

AUBURN UNIVERSITY
SAMUEL GINN
COLLEGE OF ENGINEERING