CMSC5741 Big Data Tech. & Apps.

Lecture 10: Mining Social-Graphs

Prof. Michael R. Lyu

Computer Science & Engineering Dept.

The Chinese University of Hong Kong

Prediction of 2016 US Election

POLITICS

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

Social Network Prediction of 2016 US Election

Outline

- Introduction
- Social Networks as Graph
  - Basics of Grap
  - Important pro   as graphs
  - Analysis of a real-world social
- Community detection algorit
  - Girvan-Newman algorithm
  - Spectral clustering
- Conclusions

Outline

- Introduction
- Social Networks as Graph

  - Basics of Grap
  - Important pro

as graphs
  - Analysis of a real-world social

- Community detection algorit
  - Girvan-Newman algorithm
  - Spectral clustering
- Conclusions

Networks: Communication

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

Graph of the Internet (Autonomous Systems)
Power-law degrees [Faloutsos-Faloutsos-Faloutsos, 1999]
Robustness [Doyle-Willinger, 2005]

Networks: Media

Connections between political blogs
Polarization of the network [Adamic-Glance, 2005]

Network: Science

Citation networks and Maps of science
[Börner et al., 2012]

Networks: Biology

Protein-Protein Interaction Networks:
Nodes: Proteins
Edges: ‘physical’ interactions

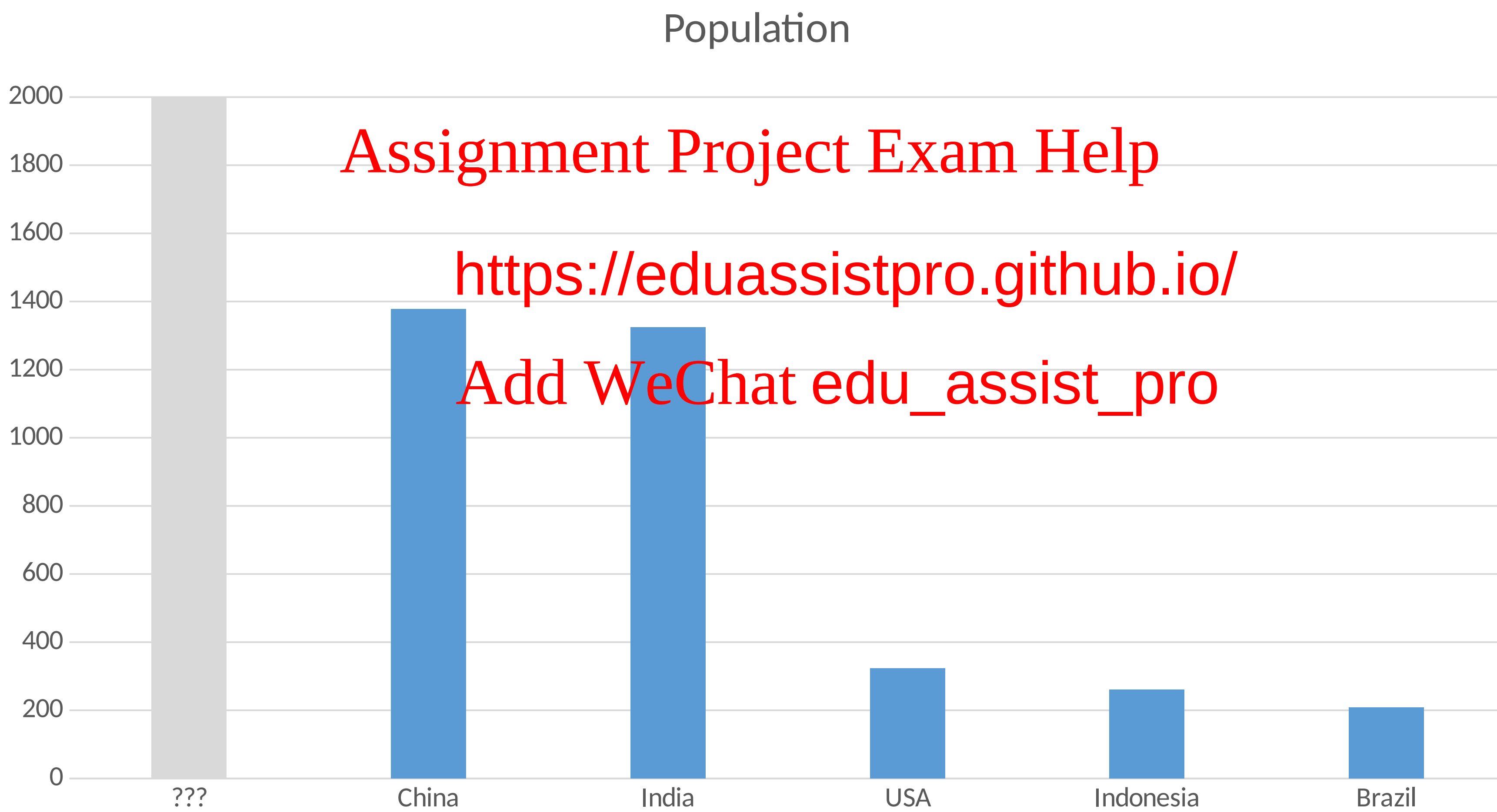
Metabolic networks:
Nodes: Metabolites and enzymes Edges: Chemical reactions

Networks: Brain

Human brain has between 10-100 billion neurons [Sporns, 2011]

Networks: Social

Country with Largest World Population?



Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

If Facebook were a Country..

4

Acquired by Facebook

Networks!

- Behind each such system there is an intricate wiring diagram, **a network**, that defines the **interactions** between the components

- **We will never understand th s unless we understand the networks be !**

Social Web - The Lab of Knowledge and Humanity

The web is " for knowledge and unde ng humanity

Networks: Impact

Cisco Market cap: $211b
billion (2y ago it was 150b)

oogle Market cap: $740b
lion (2y ago it was 536b)

ebook Market cap: $400b
on (2y ago it was 336b)

Networks: Problems

- **Social networks:**
  - Link prediction, friend recommendation

  - Social circle detection, community detection
  - Social recomm

  - Identifying influential nodes, v g

- **Communication networks:**
  - Intrusion detection, fraud
  - Churn prediction
- **Information networks:**
  - Navigational aids

Outline

- Introduction
- Social Networks as Graph
  - Basics of Grap
  - Important pro
  - Analysis of a real-world social
- Community detection algorit
  - Girvan-Newman algorithm
  - Spectral clustering
- Conclusions

Social Network as Graph

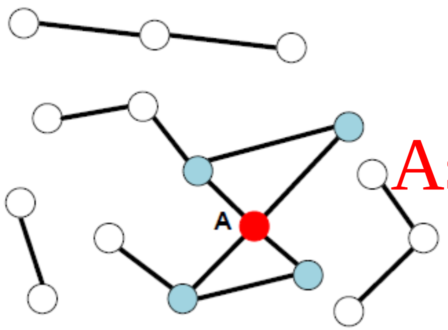
- **Undirected graphs**
  - **Links:** undirected (symmetrical, reciprocal relations)
- **Directed graphs**
  - **Links:** directed (asymmetrical relations)

- Undirected links:
  - Collaborations
  - Friendship on Facebook
- Directed links:
  - Phone calls
  - Following on Twitter

Adjacency Matrix

Node Degrees

Undirected



- Node degree : the number of edges adjacent to node

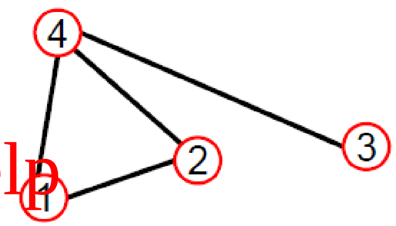
- In directed networks we define an **in-degree** and **out-degree**.
  - The (total) degree of a node is the sum of in- and out-degrees.

$$k_C^{in} = 2 \qquad k_C^{out} = 1 \qquad k_C = 3$$

$$\bar{k} = \frac{E}{N} \qquad \overline{k^{in}} = \overline{k^{out}}$$

Degree Matrix

- This graph has entries only on the diagonal.
- The entry for r column *i* is the the *i*th node



$$D = \begin{matrix} & 0 & 0 & 0 \\ & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{matrix}\Bigg)$$

Laplacian Matrix

- Suppose a graph as adjacency matrix and degree matrix .

- Laplacian matri
- the Laplacian m

  - Has the same entries as on th

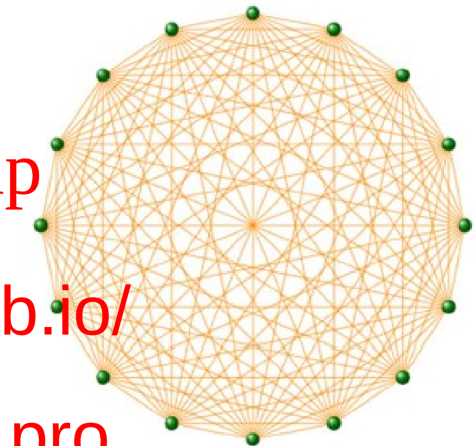
  - Off the diagonal, at row and column ,
    - has −1 if there is an edge between nodes and
    - 0 if not.

In-class Practice 1

- Go to practice

Complete Graph

- The **maximum number of edges** in an undirected graph on ***N*** nodes is



- A graph with the number of edges is a **complete graph**
  - and its average degree is
  - ***N*-1**

Networks are Sparse Graphs

- **Most real‑world networks are sparse**
  - **$E << E_{max}$ (or $k << N-1$)**

- **Consequence: Adjacency matrix is filled with zeros!**
  - **Density of the matrix (E/N^2):**

  WWW=$1.51*10^{-5}$, MSN IM Instant Messenger = $2.27*10^{-8}$

More Types of Graphs

Unweighted

Weighted

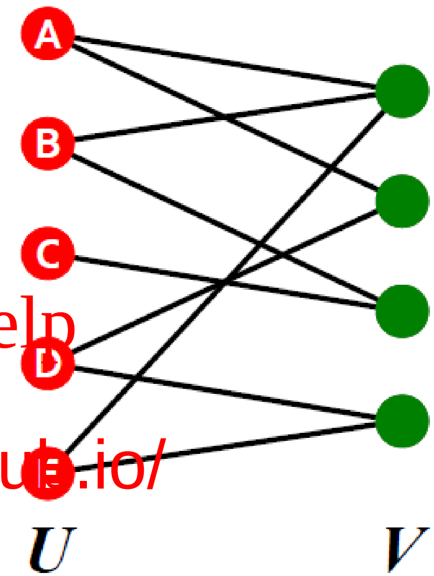
More Types of Graphs

Self-edges (Self-loops) **Multigraph**

Network Representations

Bipartite Graph

- **Bipartite graph** is a graph whose nodes can be divided into two disjoint sets ***U*** and ***V*** such that every link connects a node in ***U*** to one in that is, ***U*** and ***V*** are i



- **Examples:**
  - Authors-to-papers (they authored)
  - Actors-to-Movies (they appeared in)
  - Users-to-Movies (they rated)
- **"Folded" networks:**
  - Author collaboration networks
  - Movie co-rating networks

Outline

- Introduction
- Social Networks as Graph

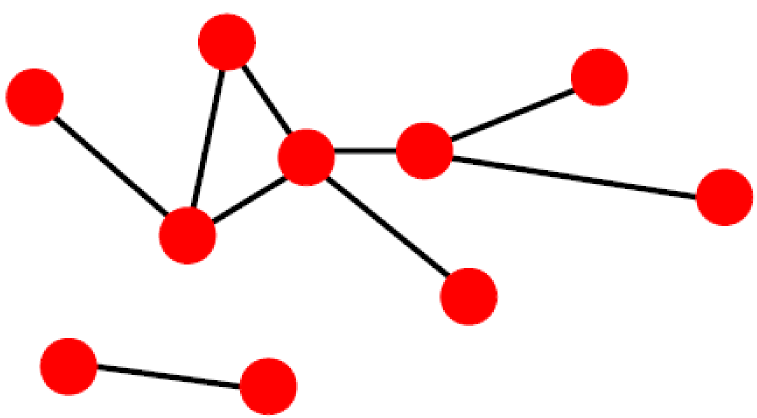
  - Basics of Grap
  - Important pro

as graphs
  - Analysis of a real-world social

- Community detection algorit
  - Girvan-Newman algorithm
  - Spectral clustering
- Conclusions

Degree Distribution

- Degree distribution : Probability that a randomly chosen node has degree ***k***
  - N_k = # nodes with degree ***k***
- Normalized:
  - $P(k) = N_k / N$ ➔ **plot**

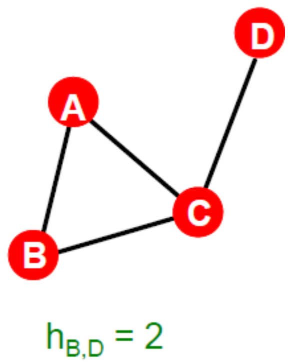


Paths in a Graph

- A ***path*** is a sequence of nodes in which each node is linked to the next one

- Path can intersect itself and pass through the same edge multiple times
  - E.g.: ACBDCDEG
  - In a directed graph a path can only follow the direction of the "arrow"

Distance in a Graph



- **Distance (shortest path, geodesic)** between a pair of nodes is defined as the number of edges along the shortest nodes
  - *If the tw disconnected, the distance i ned as infinite
- In **directed graphs** paths need to follow the direction of the arrows
  - Consequence: Distance is **not symmetric**: $h_{A,C} \neq h_{C,A}$

Network Diameter

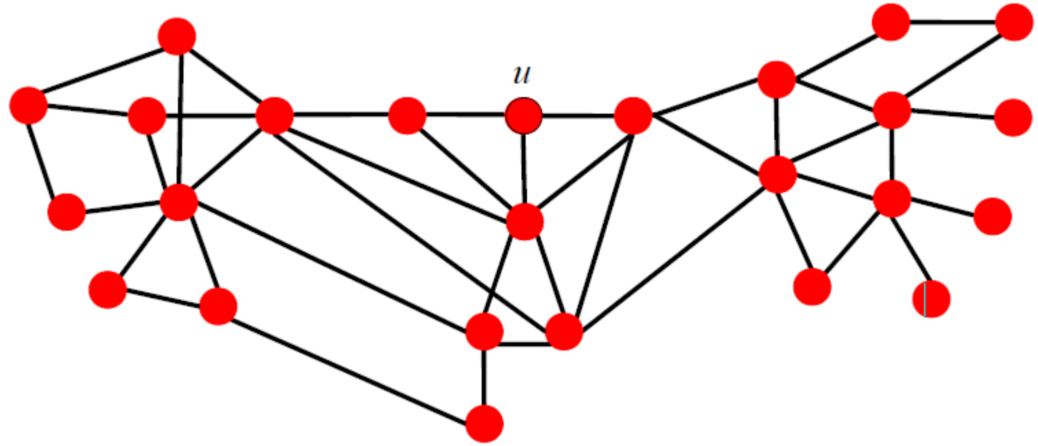
- **Diameter**: the maximum (shortest path) distance between any pair of nodes in a graph

- **Average path le** raph (component) or a strongly co f a) directed graph:

  - Many times we compute the average only over the connected pairs of nodes (we ignore "infinite" length paths)

Finding the Shortest Path

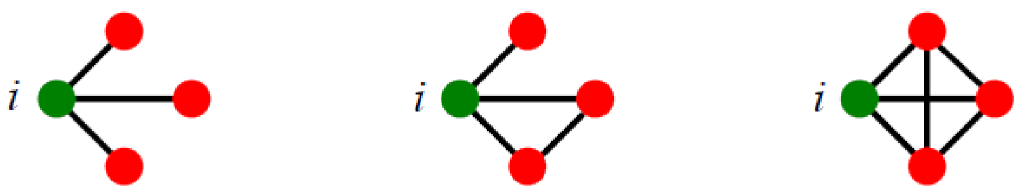
- **Breath‐First Search:**
  - Start with node *u*, mark it to be at distance $h_u(u)=0$, add *u* to the queue
  - While the que
    - Take node *v* off the queue, put it eighbors *w* into the queue and mark $h_u(w)=h_u(v)+1$



Clustering Coefficient

- **Clustering coefficient:**
  - What portion of ***i***'s neighbors are connected?
  - Node ***i*** with de

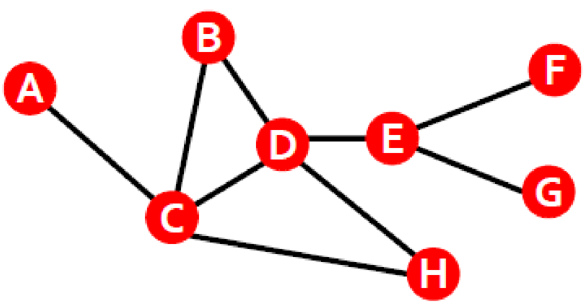
where e_i is the n
between the nei



Clustering Coefficient

- **Clustering coefficient:**
  - What portion of ***i***'s neighbors are connected?
  - Node ***i*** with de

where e_i is the n
between the nei



$k_B=2,\ \ e_B=1,$

$k_D=4,\ \ e_D=2,$

Key Network Properties

- Degree distribution:
- Path length:
- Clustering coe

Outline

- Introduction
- Social Networks as Graph
  - Basics of Grap
  - Important pro as graphs
  - Analysis of a real-world social
- Community detection algorit
  - Girvan-Newman algorithm
  - Spectral clustering
- Conclusions

The MSN Messenger

- **MSN Messenger activity in June 2006:**
  - 150Gb/day (compressed)
  - 4.5Tb / month
  - 245 million us
  - 180 million users engaged in c s
  - More than 30 billion conversa
  - More than 255 billion exchanged messages

Communication: Geography

Communication Network

Messaging as a Network

MSN Network: Connectivity

MSN: Degree Distribution

MSN: Log-Log Degree Distribution

We plot the same data as on the previous slide, just the axes are now logarithmic.

MSN: Clustering

MSN: Diameter

Number of links between pairs of nodes

nodes as we do BFS out of a random node

Steps	#Nodes
0	1
1	10
2	78
3	3,96
4	8,648
5	3,299,252
6	28,395,849
7	79,059,497
8	52,995,778
9	10,321,008
10	1,955,007
11	518,410
12	149,945
13	44,616
14	13,740
15	4,476
16	1,542
17	536
18	167
19	71
20	29
21	16
22	10
23	3
24	2
25	3

Avg. path length **6.6**
90% of the people can be reached in < 8 hops

MSN: Key Network Properties

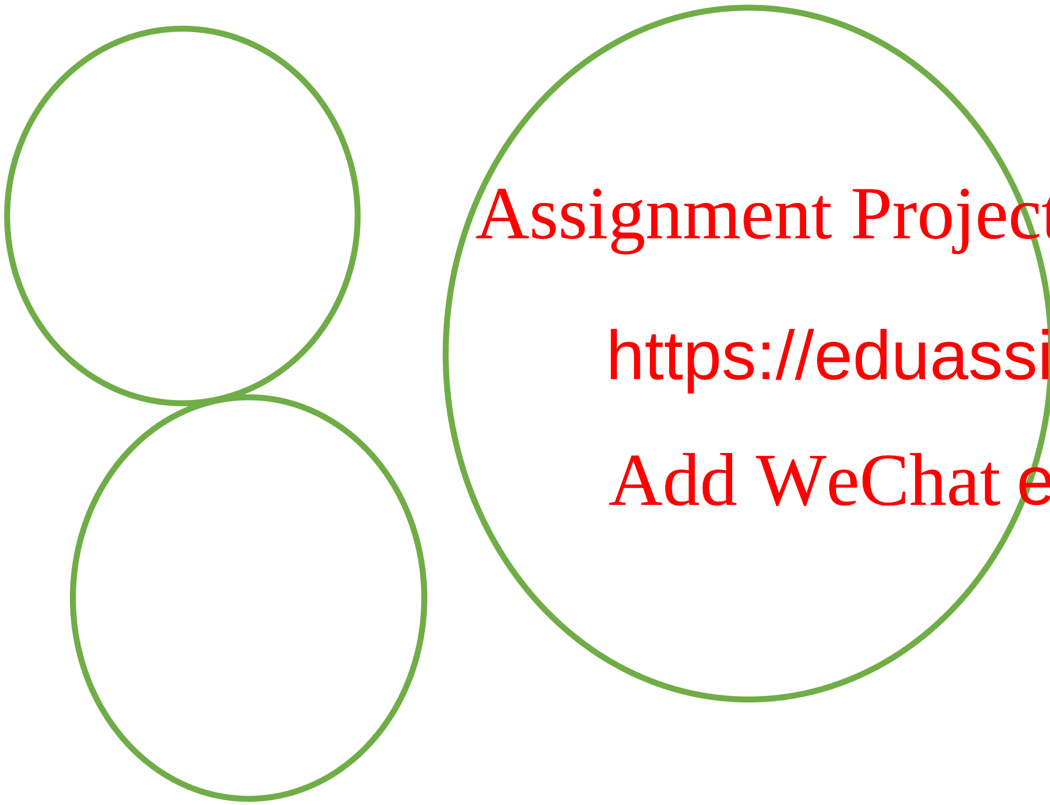
- Degree distribution: heavily skewed. avg degree = 14.4
- Path length: 6.6
- Clustering coe

Outline

- Introduction
- Social Networks as Graph
  - Basics of Grap
  - Important pro as graphs
  - Analysis of a real-world social
- Community detection algorit
  - Girvan-Newman algorithm
  - Spectral clustering
- Conclusions

Community Detection

How to find communities?



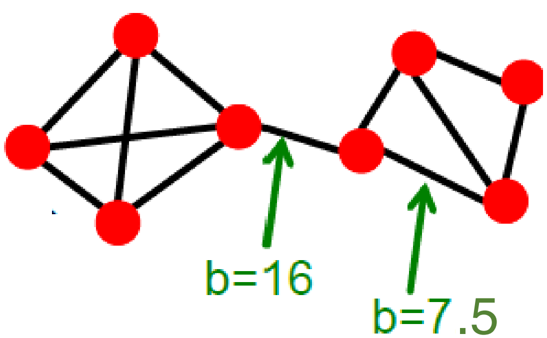
We will work with **undirected** (unweighted) networks

Betweenness: Strength of Ties

- **Edge betweenness**: number of shortest paths (among all pair of vertices) passing over the edge.
- If there is more than one shortest path between a pair of nodes, each path is assigned equal weight such that the total weight of al .

Simple alg culating betweeneness.

1. Repeat for each vertex
   a) compute its shortest path to all other vertices.
   b) for each edge, count the number of the shortest paths passing over the edge
2. Sum up the results from each vertex.



Betweenness: Strength of Ties

Calculation for 1 vertex

Calculation for all vertices

In-class Practice 2

- Go to practice

Method for Community Detection: Girvan-Newman

- Divisive hierarchical clustering based on the notion of edge **betweenness**:

  - **Number of shortest paths passing through the edge**
- **Girvan-Newma**
  - **Undirected unweighted netw**

  - **Repeat until no edges are left**
    - Calculate betweenness of edges
    - Remove edges with highest betweenness
  - Connected components are communities
  - Gives a hierarchical decomposition of the network

Girvan-Newman: Example

Need to re-compute betweenness at every step

Girvan-Newman: Example

Girvan-Newman

Communities in physics collaborations

Problems of Girvan-Newman

- Computing betweenness is slow.
- No theoretical guarantees.
- Not widely appl

Outline

- Introduction
- Social Networks as Graph
  - Basics of Grap
  - Important pro as graphs
  - Analysis of a real-world social
- Community detection algorit
  - *Girvan-Newman* algorithm
  - Spectral clustering
- Conclusions

Graph Partitioning

- **Undirected graph $G\ (V,E)$:**

- **Bi-partitioning task**
  - Divide vertices into two disjoint group

- **Questions:**
  - How can we define a “good” partition of G?
  - How can we efficiently identify such a partition?

Graph Partitioning

- **What makes a good partition?**
  - Maximize the number of within-group connections
  - Minimize the number of between-group connections

Graph Cuts

- **Express partitioning objectives as a function of the "edge cut" of the partition**
- **Cut:** Set of edge in a group:

Graph Cut Criterion

- **Criterion: Minimum-cut**
  - Minimize weight of connections between groups

- **Degenerate cas**

- **Problem:**
  - Only considers external cluster connections
  - Does not consider internal cluster connectivity

Graph Cut Criterion

- **Criterion: Normalized-cut** [Shi-Malik, '97]
  - Connectivity between groups relative to the density of each group

  - vol(A): total weight of the ed least one endpoint in A

- **Why use this criterion?**
  - Produces more balanced partitions
- **How do we efficiently find a good partition?**
  - **Problem:** Computing optimal cut is NP-hard

Spectral Graph Partitioning

- ***A***: adjacency matrix of undirected **G**
  - A_{ij} **=1** if (i,j) is an edge, else **0**

Assignment Project Exam Help

- ***x*** is a vector in $\Re^n$ with components (,...,). Think of it as a label/value of eac

https://eduassistpro.github.io/

- **What is the meaning of *Ax*?**

Add WeChat edu_assist_pro

x_j

- **Entry y_i is a sum of labels x_j of neighbors of *i***

What’s the Meaning of Ax?

$$\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

- ***jth* coordinate of *Ax*:**
  - Sum of the ***x***-values of neighbors of ***j***
  - Make this a new value at node ***j***

$$\boldsymbol{A} \cdot \boldsymbol{x} = \boldsymbol{\lambda} \cdot \boldsymbol{x}$$

- **Spectral Graph**

  - Analyze the “spectrum” of ma ting G

  - **Spectrum:** Eigenvectors x_i of a red by the magnitude (strength) of their corresponding eigenvalues λ_i:

$$\Lambda = \{\lambda_1, \lambda_2, \dots, \lambda_n\}$$

$$\lambda_1 \le \lambda_2 \le \dots \le \lambda_n$$

Example: d-Regular Graph

- Suppose all nodes in G have degree d and G is connected

  - **What are some eigenvalues/vectors of G?**
  - $A \cdot x = \lambda \cdot x$ What is

    - **Let's try:** $x = (1,1,\ldots,1)$
    - **Then:** $A \cdot x = d,d,\ldots,d = \lambda \cdot x$. **So:** $\lambda = d$

  - **We found eigenpair of G: $x = (1,1,\ldots,1)$, $=$**

Remember the meaning of $y = A \cdot x$:

$$y_i = \sum_{j=1}^{n} A_{ij} x_j = \sum_{(i,j)\in E} x_j$$

d is the Largest Eigenvalue of A

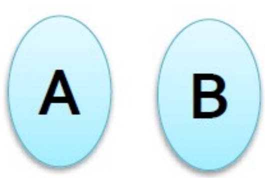
- **G** is **d**-regular connected, **A** is its adjacency matrix
- **Claim:**
  - **d** is largest eigenvalue of A.
  - **d** has multiplicity o associated with eigenvalue **d**)
- **Proof: Why no eigenvalue ’ > ?**
  - To obtain **d** we needed $x_i = x_j$ for every
  - This means $x = c \cdot (1,1,\ldots,1)$ for some const. c
  - **Define:** S = nodes i with maximum possible value of x_i
  - Then consider some vector y which is not a multiple of vector $(\mathbf{1},\ldots,\mathbf{1})$. So not all nodes i (with labels y_i) are in S
  - Consider some node $j \in S$ and a neighbor $i \notin S$ then node j gets a value strictly less than d
  - So y is not eigenvector! And so d is the largest eigenvalue!

Example: Graph on 2 Components

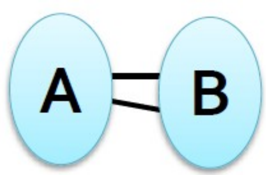
- What if G is not connected?
  - G has 2 components, each d-regular

- **What are some**
  - $\boldsymbol{x}$= Put all $\mathbf{1}$s on $\boldsymbol{A}$ an $\mathbf{0}$ $\boldsymbol{B}$

  - And so in both cases the corresponding
- A bit of intuition



$\lambda_n = \lambda_{n-1}$



$\lambda_n - \lambda_{n-1} \approx \mathbf{0}$

2nd largest eigval. λ_{n-1} now has value very close to λ_n

More Intuition

- More Intuitions

- If the graph is c le) then we already know that $\boldsymbol{x}_n=(\mathbf{1},\ldots,\mathbf{1})$ is an eigenve
  - Since eigenvectors are orthogonal then the components of $\boldsymbol{x}_{n-1}$ sum to **0**.
- So we can look at the eigenvector of the 2nd largest eigenvalue and declare nodes with positive label in **A** and negative label in **B**. **But there is still lots to sort out.**

Matrix Representation

- **Adjacency matrix (A):**
  - $n*n$ matrix
  - $A=[a_{ij}]$, $a_{ij}=1$ if edge between node i and j

- **Important properties:**
  - Symmetric matrix
  - Eigenvectors are real and orthogonal

Matrix Representation

- **Degree matrix (D):**
  - $n*n$ diagonal matrix
  - $D=[d_{ii}]$, d_{ii} = de

Matrix Representation

- **Laplacian matrix (L):**
  - $\boldsymbol{n}^*\boldsymbol{n}$ symmetric matrix

- **What is trivial Eigenpair?**
  - $\boldsymbol{x}=(\mathbf{1},\dots,\mathbf{1})$ then $\boldsymbol{L}\cdot\boldsymbol{x}=\mathbf{0}$ and so $\lambda=\lambda_1=\mathbf{0}$
- **Important properties:**
  - **Eigenvalues** are non-negative real numbers
  - **Eigenvectors** are real and orthogonal

Facts about Laplacian L

- **(a)** All eigenvalues are ≥ 0
- **(b)** for every x
- **(c)**

- That is, L is **positive semi-definite**

as Optimization Problem

- **Fact: For symmetric matrix *L*:**

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

as Optimization Problem

- **What else do we know about *x*?**
  - $\boldsymbol{x}$ is unit vector:
  - $\boldsymbol{x}$ is orthogonal to **1st** eigenvector (,...,) thus:

- Remember

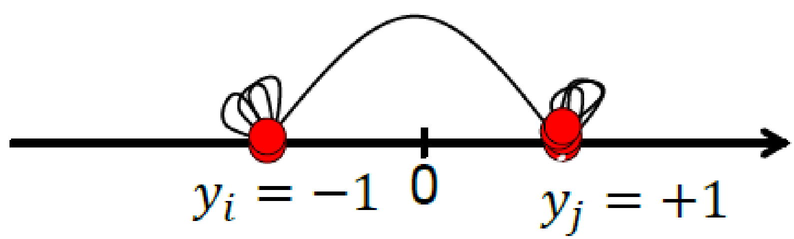
We want to assign values x_i to nodes i such that few edges cross 0. (we want x_i and x_j to subtract each other)

Finding Optimal Cut [Fielder'73]

- Back to finding the optimal cut
- Express partition (A,B) as a vector

- We can minimize the cut of t n by finding a non-trivial vector *x* that **minimizes**

Can't solve exactly. Let's relax *y* and allow it to take any real value.



Rayleigh Theorem

- The minimum value of $\boldsymbol{f}(\boldsymbol{y})$ is give d smallest eigenvalue λ_2 of the Laplacian matrix $\boldsymbol{L}$
- The optimal solution for $\boldsymbol{f}(\boldsymbol{y})$ is given by the corresponding eigenvector $\boldsymbol{x}$, referred as the **Fiedler vector**

Approx. Guarantee of Spectral (details)

- Suppose there is a partition of **G** into **A** and **B** where $A \leq |B|$, s.t. then .
  - This is the approximation guarantee of the spectral clustering.
  - It says the cut from the optimal one of score .

So Far..

- **How to define a "good" partition of a graph?**
  - Minimize a given graph cut criterion

- **How to effi** **such a partition?**
  - Approximate using information provided by the eigenvalues and eigenvectors of a graph
- **Spectral Clustering**

Spectral Clustering Algorithms

- **Three basic stages:**
  1) **Pre-processing**

     - Construct of the graph
  2) **Decompositi**

     - Compute eigenvalues an tors of the matrix

     - Map each point to a low nal representation based on one or more eigenvectors
  3) **Grouping**
     - Assign points to two or more clusters, based on the new representation

Spectral Partition Algorithms

Spectral Partition Algorithms

- **3) Grouping**
  - Sort components of reduced 1-dimensional vector
  - Identify clusters by splitting the sorted vector in two
- **How to choose a splitting point?**
  - Naïve approac
    - Split at **0** or me
  - More expensive approaches:
    - Attempt to minimize normalized ion (sweep over ordering of nodes induced by the eigenvector)

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

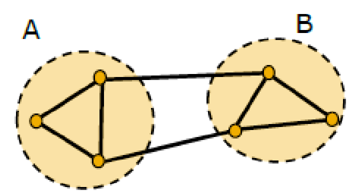
Split at 0:

Cluster A: Positive points

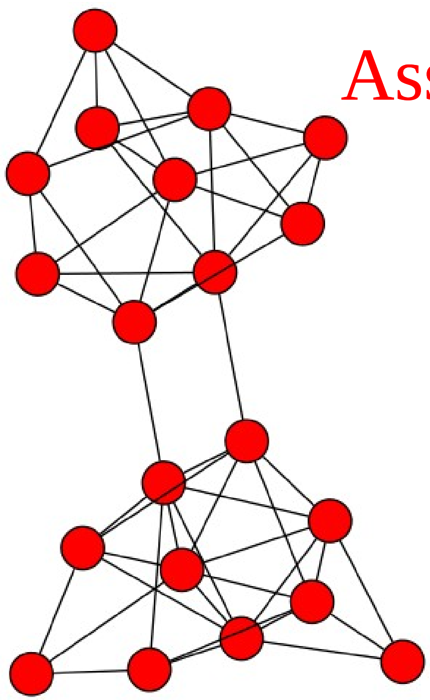
Cluster B: Negative points

1	0.3
2	0.6
3	0.3

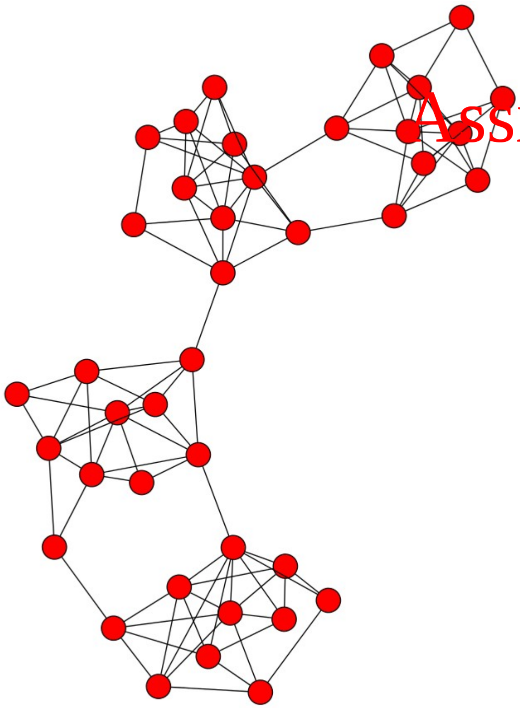
4	-0.3
5	-0.3
6	-0.6



Example: Spectral Partitioning



Example: Spectral Partitioning



Example: Spectral Partitioning

Outline

- Introduction
- Social Networks as Graph
  - Basics of Grap
  - Important pro as graphs
  - Analysis of a real-world social
- Community detection algorit
  - Girvan-Newman algorithm
  - Spectral clustering
- Conclusions

Conclusions

- **How do we reason about networks?**
  - **Empirical:** Study network data to find organizational principles. How do we measure and quantify networks?

- **Mathematic**

  - Graph theory, statistical mode

    - allow us to understand behavior ish surprising from expected phenomena
- **Algorithms**
  - for analyzing graphs
    - Hard computational challenges

Conclusions: This is Just a Beginning

One-Slide Takeaway

- Key concepts of social graphs
  - Undirected/Directed Graph/Adjacency Matrix

  - Shortest path/
  - Degree distrib ing Coefficient
  - Real world graph

    - Degree distribution: high-skewe
    - Most path lengths are small
- Community detection algorithms
  - Purpose of community detection
  - Spectral clustering

Reference

- Adamic, Lada A., and Eytan Adar. "Friends and neighbors on the web." Social networks 25.3 (2003): 211-230.
- Kossinets, Gueorgi, and Duncan J. Watts. "Empirical analysis of an evolving social network." Science 311.5757 (2006): 88-90. APA
- Liben-Nowell, David, et etworks." Proceedings of the National Academy o America 102.33 (2005): 11623-11628.
- Leskovec, Jure, and Eric Horvitz. "Planetary-s large instant-messaging network." Proceedings of the 17th internati on World Wide Web. ACM, 2008.
- J. Shi and J. Malik, “Normalized cuts and image segmentation,” IEEE
- Trans. on Pattern Analysis and Machine Intelligence,” 22:8 (2000), pp. 888–905.
- S. Fortunato, “Community detection in graphs,” Physics Reports 486:3–5(2010), pp. 75–174.
- M. Girvan and M.E.J. Newman, “Community structure in social and bi-ological networks,” Proc. Natl. Acad. Sci. 99 (2002), pp. 7821–7826.

Reference

- U. von Luxburg, “A tutorial on spectral clustering,” Statistics and Computing bf17:4 (2007), 2007, pp. 395–416.
- J. Leskovec, K.J. Lang, A. Dasgupta, and M.W. Mahoney, “Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters,” http://arxiv.org/abs/0810.1355.

- L. Backstrom and J. Leskovec, “Supervised random walks: predicting and recommending links in social networks,” b Search and Data Mining (2011), pp. 635–644.

- Predicting Positive and Negative Links in Online by J. Leskovec, D. Huttenlocher, J. Kleinberg. ACM WWW Interna on World Wide Web (WWW), 2010.

- Learning to Discover Social Circles in Ego Networks by J. McAuley, J. Leskovec. Neural Information Processing Systems (NIPS), 2012.
- Defining and Evaluating Network Communities based on Ground-truth by J. Yang, J. Leskovec. IEEE International Conference On Data Mining (ICDM), 2012.
- The Life and Death of Online Groups: Predicting Group Growth and Longevity by S. Kairam, D. Wang, J. Leskovec. ACM International Conference on Web Search and Data Mining (WSDM), 2012

In-class Practice 1

- For the graph on the right, compute:
  - The adjacency matrix
  - The degree ma
  - The Laplacian

In-class Practice 2

- For the graph on the right, compute:
  - The shortest path between each pairs of n
  - The betweenn edge