From Graph to Knowledge Graph: Algorithms and Applications

Module 2: Graph Properties and Applications

Module 2: Graph Properties and Applications

- Graph basics
  - Graph history
  - Basic node centralities
  - Eigenvector, HITS, & PageRank
- Graph applications
  - Node label classification
  - Community detection
  - Link prediction
- What will not be covered
  - Graph Theory

As of Jan. 2018

- World Population
  - **7.593 Billion**
- Internet Users
  - 4.021 Billion, 53% of global penetration
- Active Social Media Users
  - 3.196 Billion, 42% of global penetration
- Unique Mobile Users
  - 5.135 Billion, 68% of global penetration
- Active Mobile Social Users
  - 2.958 Billion, 39% of global penetration

Data source: https://wearesocial.com/blog/2018/01/global-digital-report-2018

One second of Jan. 2017

- 3,271 Skype calls in 1 second
- 7,871 Tweets sent in 1 second
- 75,494 YouTube videos viewed in 1 second
- 2,801,324 Emails sent in 1 second

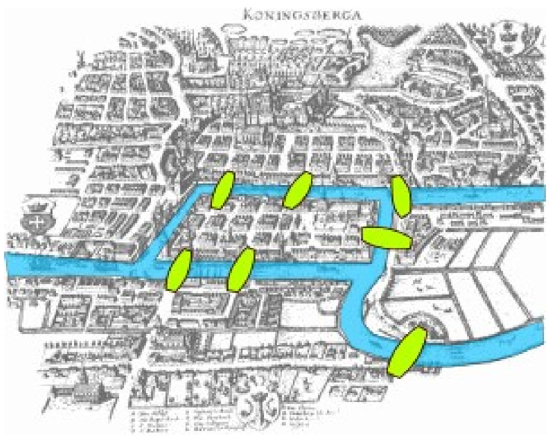
Data source: http://www.internetlivestats.com/one-second/

The era of (digitally) connected world

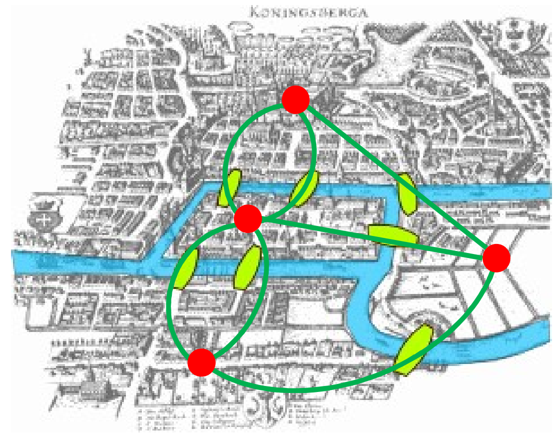
When and how did the mind of graphs start?

Leonhard Euler

(1707--1783)



Seven Bridges of Königsberg (1736)



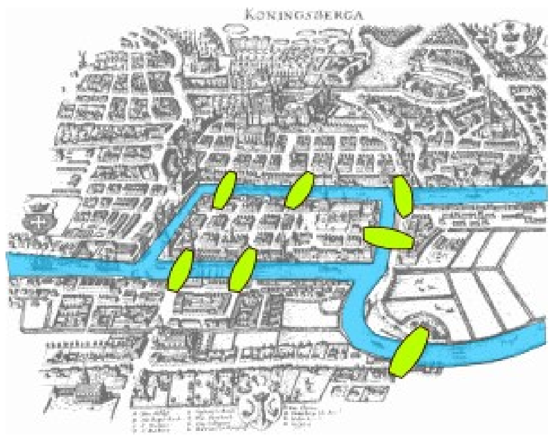
Can we design a routine to walk through each bridge once and only once?

Image Credit: https://en.wikipedia.org/wiki/Leonhard_Euler

When and how did the mind of graphs start?

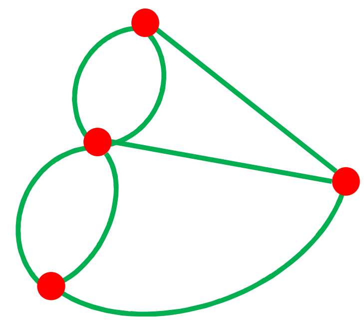
Leonhard Euler

(1707--1783)



Seven Bridges of Königsberg (1736)

Can we design a routine to walk through each bridge once and only once?

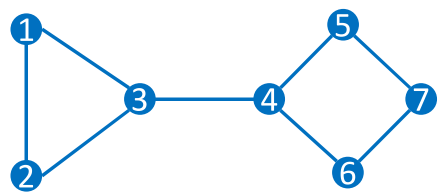


When did the term “graph” start?

James J Sylvester

(1814--1897)

The term “graph” (1878)



This is a graph!

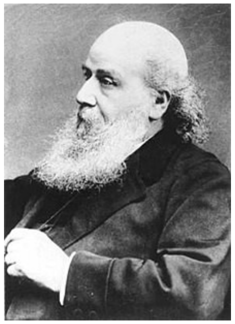
G = (V, E), where *V* is the node set and *E* denotes the edge set.

- V: $v_1, v_2, v_3, v_4, v_5, v_6, v_7$
- E: $e_{12}, e_{13}, e_{23}, e_{34}, e_{45}, e_{46}, e_{57}, e_{67}$
- $E \subseteq V \times V$

- #nodes: n = $|V| = 7$
  - The **order** of the graph *G*
- #edges: m =$|E| = 8$
  - The **size** of the graph *G*

Image Credit: https://en.wikipedia.org/wiki/James_Joseph_Sylvester

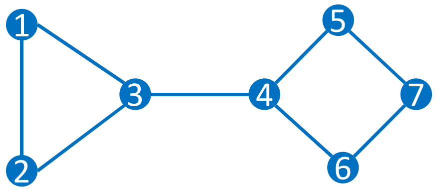
Matrix representations of graphs



James J Sylvester

(1814--1897)

The term "graph" (1878)
The term "matrix" (1850)



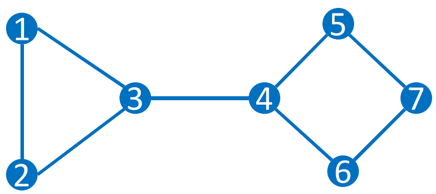
This is a graph!

G = (V, E), where *V* is the node set and *E* denotes the edge set.

- V: $v_1, v_2, v_3, v_4, v_5, v_6, v_7$
- E: $e_{12}, e_{13}, e_{23}, e_{34}, e_{45}, e_{46}, e_{57}, e_{67}$
- $E \subseteq V \times V$

The graph G can be represented as a matrix!

Matrix representations of graphs

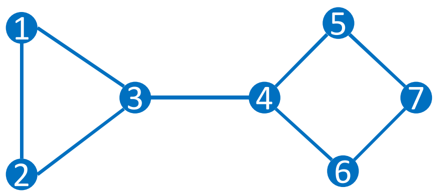


Adjacency matrix $\boldsymbol{A} = \{a_{ij}\}_{n\times n}$

- $a_{ij} = \begin{cases} 1 & if\ e_{ij} \in E \\ 0 & otherwise \end{cases}$

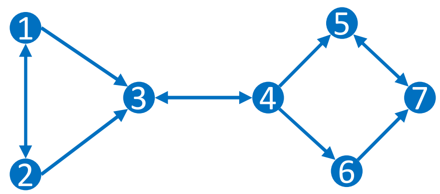
$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

Graph Type



Undirected graph

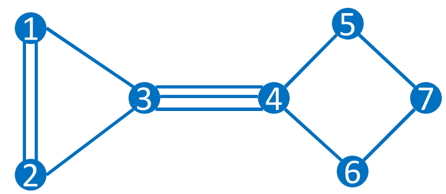
$\forall\ e_{ij} \in E \Rightarrow e_{ji} \in E$



Directed graph

$e_{ij} \in E \nRightarrow e_{ji} \in E$

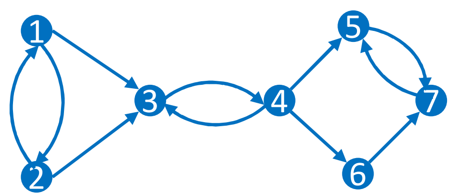
Simple graph
- *No self-loops:* $e_{ii} \notin E, \forall\ i \in V$
- *No multiple edges between any two nodes*



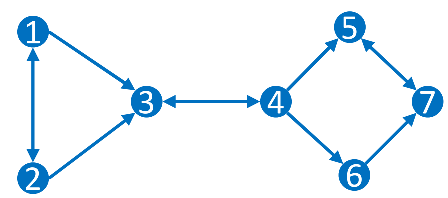
Multigraph
- *No self-loops:* $e_{ii} \notin E, \forall\ \mathrm{i} \in V$

How about this one?

Simple or multigraph?



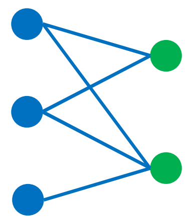
≡



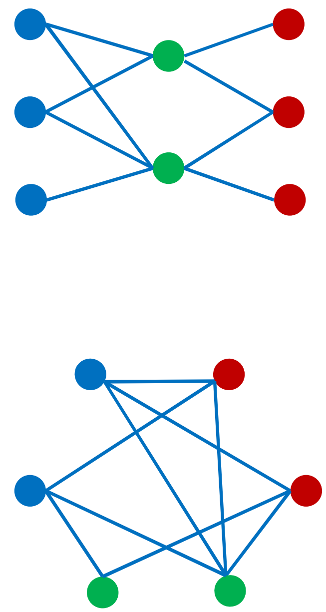
≡

Simple directed graph

K-partite graph

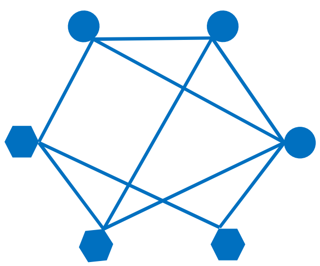


bipartite

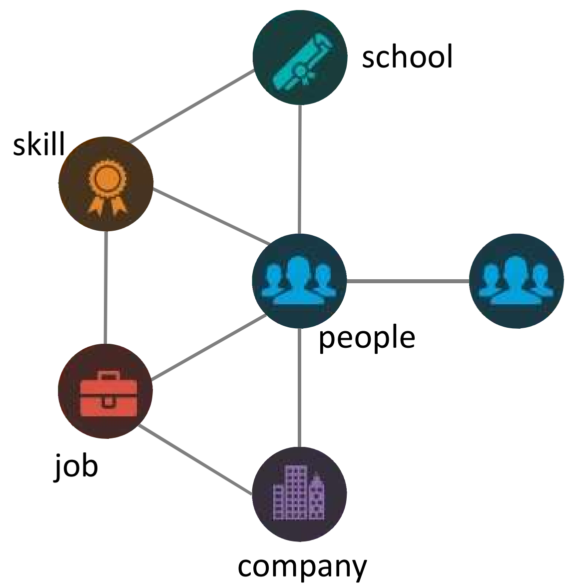


tripartite

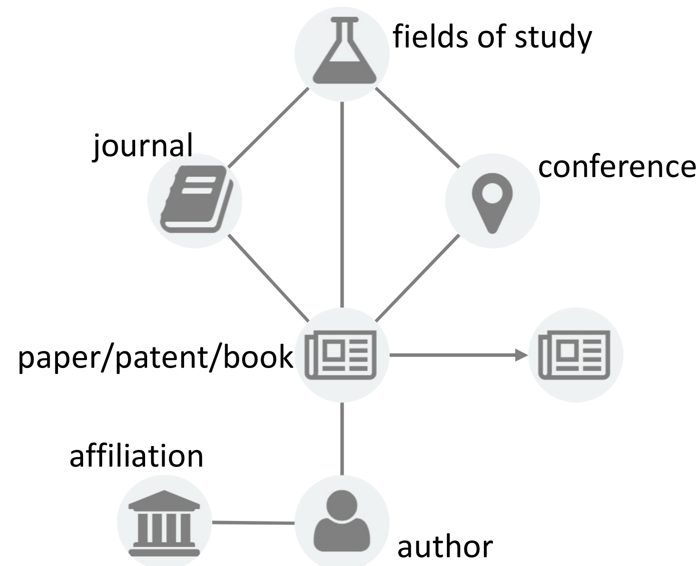
Heterogeneous graphs



Many real-world networks
are heterogeneous



LinkedIn Economic Graph



Microsoft Academic Graph

Microsoft Academic Graph

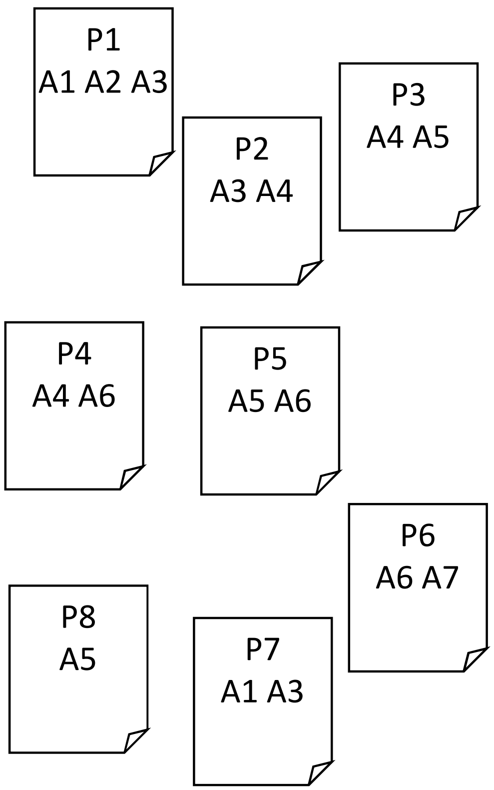
Publications	Authors
172,037,947	209,404,413

Fields of Study	Conferences	Journals	Affiliations
228,563	4,027	47,927	18,720

https://academic.microsoft.com/

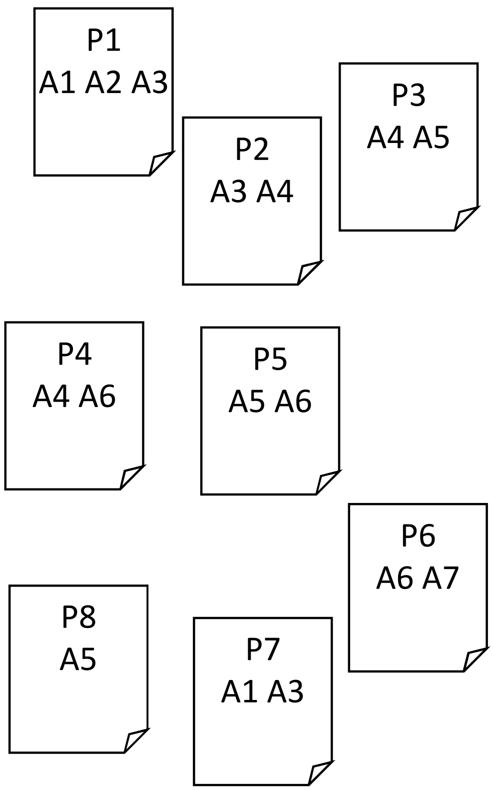
As of March 06, 2018

Build an author-collaboration graph

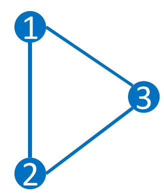


P1 A1
P1 A2
P1 A3
P2 A3
P2 A4
P3 A4
P3 A5
P4 A4
P4 A6
P5 A5
P5 A7
P6 A6
P6 A7
P7 A1
P7 A3
P8 A5

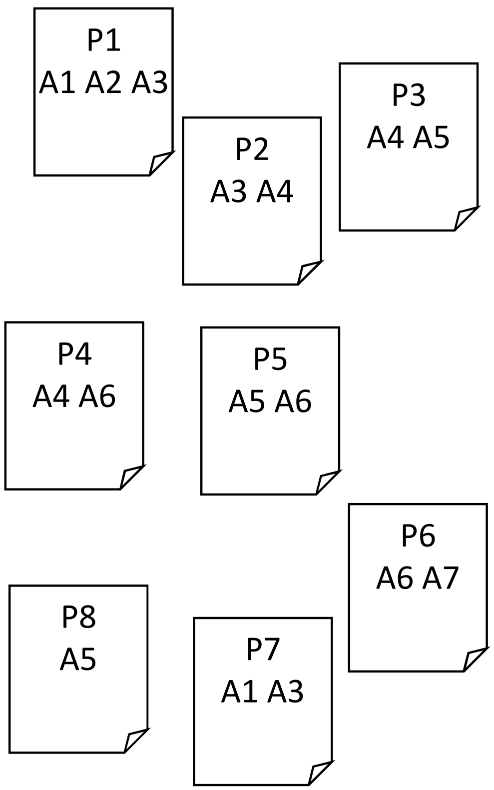
Build an author-collaboration graph



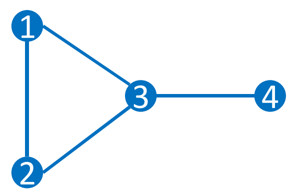
P1	A1
P1	A2
P1	A3
P2	A3
P2	A4
P3	A4
P3	A5
P4	A4
P4	A6
P5	A5
P5	A7
P6	A6
P6	A7
P7	A1
P7	A3
P8	A5



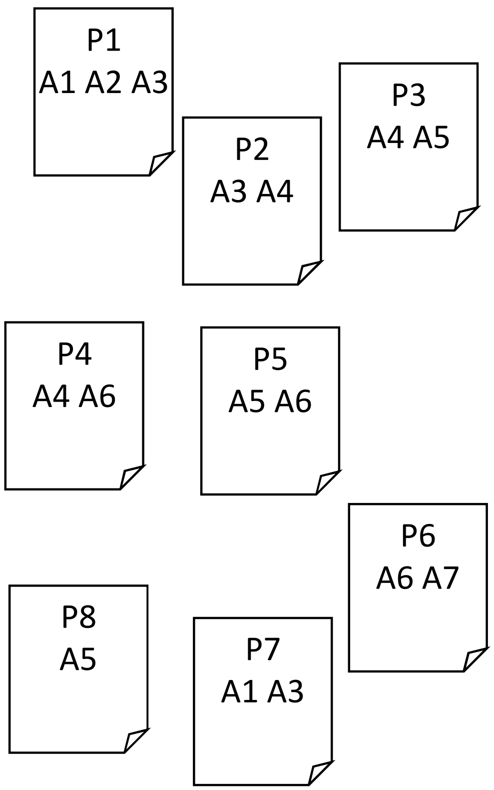
Build an author-collaboration graph



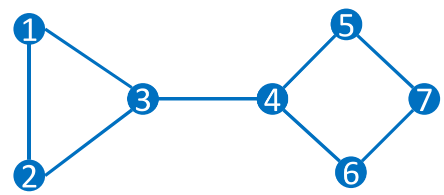
P1	A1
P1	A2
P1	A3
P2	A3
P2	A4
P3	A4
P3	A5
P4	A4
P4	A6
P5	A5
P5	A7
P6	A6
P6	A7
P7	A1
P7	A3
P8	A5



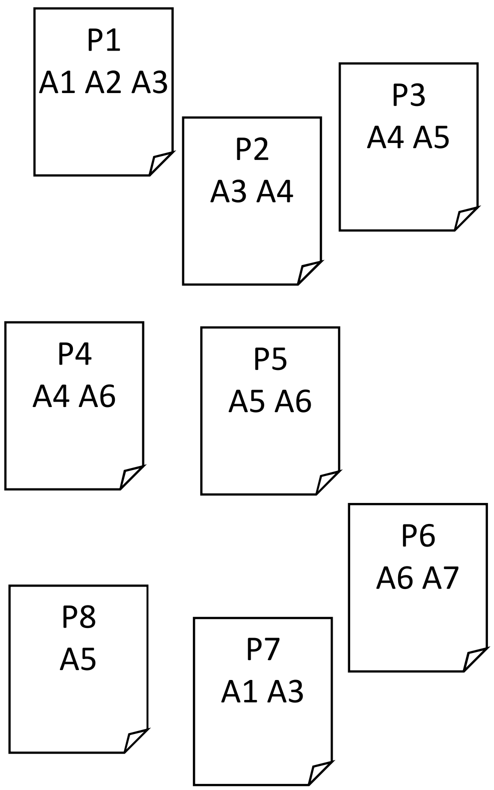
Build an author-collaboration graph



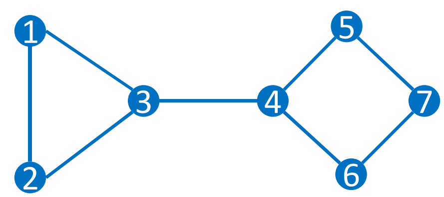
P1 A1
P1 A2
P1 A3
P2 A3
P2 A4
P3 A4
P3 A5
P4 A4
P4 A6
P5 A5
P5 A7
P6 A6
P6 A7
P7 A1
P7 A3
P8 A5



Build an author-collaboration graph



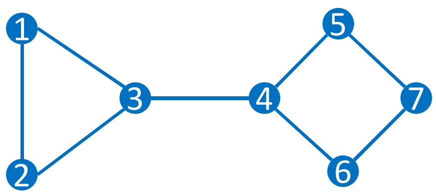
P1 A1
P1 A2
P1 A3
P2 A3
P2 A4
P3 A4
P3 A5
P4 A4
P4 A6
P5 A5
P5 A7
P6 A6
P6 A7
P7 A1
P7 A3
P8 A5



$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

Degree

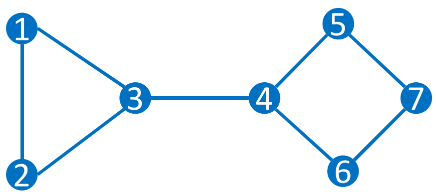
How many neighbors does each node have?



v_1	$: v_2, v_3$	$d(v_1) = 2$
v_2	$: v_1, v_3$	$d(v_2) = 2$
v_3	$: v_1, v_2, v_4$	$d(v_3) = 3$
v_4	$: v_3, v_5, v_6$	$d(v_4) = 3$
v_5	$: v_4, v_7$	$d(v_5) = 2$
v_6	$: v_4, v_7$	$d(v_6) = 2$
v_7	$: v_5, v_6$	$d(v_7) = 2$

Degree

How many neighbors does each node have?



$v_1 \quad : v_2, v_3 \qquad d(v_1) = 2$

$v_2 \quad : v_1, v_3 \qquad d(v_2) = 2$

$v_3 \quad : v_1, v_2, v_4 \qquad d(v_3) = 3$

$v_4 \quad : v_3, v_5, v_6 \qquad d(v_4) = 3$

$v_5 \quad : v_4, v_7 \qquad d(v_5) = 2$

$v_6 \quad : v_4, v_7 \qquad d(v_6) = 2$

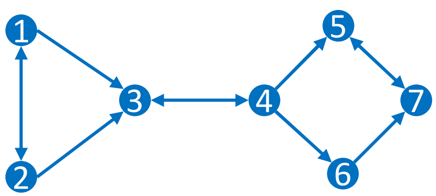
$v_7 \quad : v_5, v_6 \qquad d(v_7) = 2$

$$A = \{a_{ij}\}_{7\times 7}$$

$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

$$d(v_i) = \sum_{j=1}^{j\leq 7} a_{ij}$$

Degree



Out-degree

v_1	$: v_2, v_3$	$d^{out}(v_1) = 2$
v_2	$: v_1, v_3$	$d^{out}(v_2) = 2$
v_3	$: v_4$	$d^{out}(v_3) = 1$
v_4	$: v_3, v_5, v_6$	$d^{out}(v_4) = 3$
v_5	$: v_7$	$d^{out}(v_5) = 1$
v_6	$: v_7$	$d^{out}(v_6) = 1$
v_7	$: v_5$	$d^{out}(v_7) = 1$

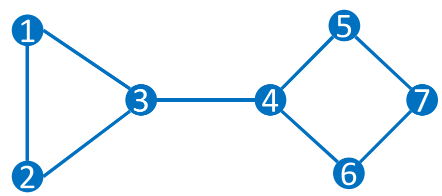
in-degree

v_1	$: v_2$	$d^{in}(v_1) = 1$
v_2	$: v_1$	$d^{in}(v_2) = 1$
v_3	$: v_1, v_2, v_4$	$d^{in}(v_3) = 3$
v_4	$: v_3$	$d^{in}(v_4) = 1$
v_5	$: v_4, v_7$	$d^{in}(v_5) = 2$
v_6	$: v_4$	$d^{in}(v_6) = 1$
v_7	$: v_5, v_6$	$d^{in}(v_7) = 2$

Degree

The handshaking Lemma

$$\sum_{i=1}^{n} d(v_i) = 2m$$



v_1	$: v_2, v_3$	$d(v_1) = 2$
v_2	$: v_1, v_3$	$d(v_2) = 2$
v_3	$: v_1, v_2, v_4$	$d(v_3) = 3$
v_4	$: v_3, v_5, v_6$	$d(v_4) = 3$
v_5	$: v_4, v_7$	$d(v_5) = 2$
v_6	$: v_4, v_7$	$d(v_6) = 2$
v_7	$: v_5, v_6$	$d(v_7) = 2$

$\sum_{i=1}^{n} d(v_i) = 2 + 2 + 3 + 3 + 2 + 2 + 2 = 16$

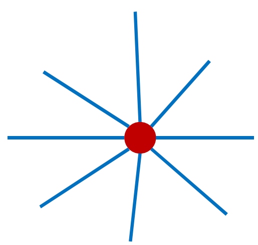
$2m = 2 \times 8 = 16$

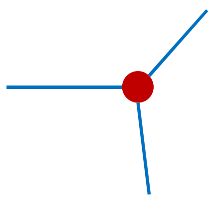
Degree in Microsoft Academic Graph (MAG)

- Author collaboration graph (undirected graph)
  - Degree represents the number of collaborators each author has
- Paper citation graph (directed graph)
  - In-degree represents the number of citations each paper collects
  - Out-degree represents the number of references each paper covers
- Institution collaboration graph
- Institution citation graph
- Field-of-Study citation graph
- Venue citation graph
- … …

Degree centrality

How to measure the importance of nodes in a graph?



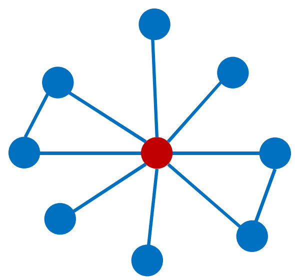


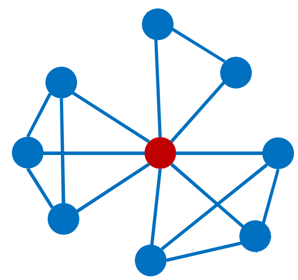
One measurement: More neighbors (connections), more important!

- *We can rank all nodes based on their degree*

Clustering coefficient centrality

How to measure the importance of nodes in a graph?

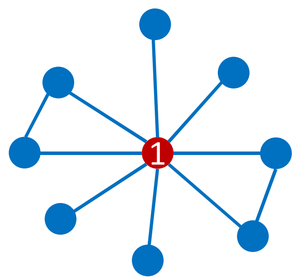


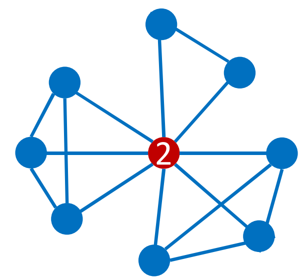


Both red nodes have the same number of neighbors, which one is more important?

Clustering coefficient centrality

$$cc(v_i) = \frac{|\{e_{ik}, e_{ij}, e_{kj} \in E\}|}{d(v_i) \times (d(v_i) - 1)/2} = \frac{\#triangles\ formed\ by\ v_i\ \&\ its\ neighbors}{\#possible\ triangles\ by\ v_i\ \&\ its\ neighbors}$$





$$cc(v_1) = \frac{2 \times 2}{8 \times (8-1)} \approx 0.0714$$

$$cc(v_2) = \frac{2 \times 7}{8 \times (8-1)} \approx 0.25$$

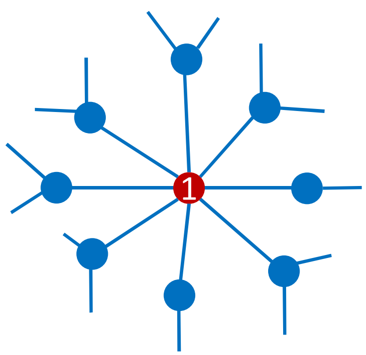
Neighborhood connectivity

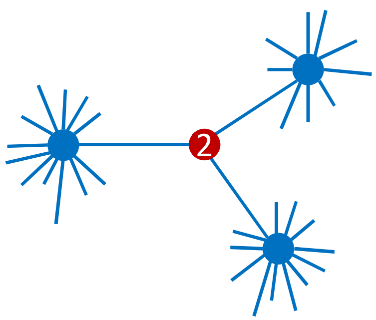
How to measure the importance of nodes in a graph?

What if your neighbors are very important?

Neighborhood connectivity

$$nc(v_i) = \frac{\sum_{e_{ij} \in E} d(v_j)}{d(v_i)} = the\ average\ degree\ of\ v_i's\ neighbors$$





$$nc(v_1) = \frac{3+3+2+3+2+3+3+3}{8} = 2.75$$

$$nc(v_2) = \frac{13+13+10}{3} = 12$$

Neighborhood connectivity

$$nc(v_i) = \frac{\sum_{e_{ij} \in E} d(v_j)}{d(v_i)} = the\ average\ degree\ of\ v_i's\ neighbors$$

Your neighbors' importance is determined by their degree centralities?

Eigenvector centrality

- Let x_i denote the importance of node v_i and let us iteratively achieve x_i

  $$x_i^{t+1} = \frac{1}{\lambda} \sum_{j=1}^{|V|} a_{ij} x_j^t$$

  $\Rightarrow \boldsymbol{Ax} = \lambda \boldsymbol{x}$, where $\boldsymbol{x}$ is one of $\boldsymbol{A}$'s eigenvector and λ is its associated eigenvalue

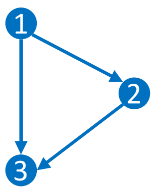
- By definition, we need the importance of each node to be non-negative

  $\Rightarrow \boldsymbol{Ax} = \lambda_1 \boldsymbol{x}$, where λ_1 is the largest eigenvalue of $\boldsymbol{A}$

- x_i represents the eigenvector-centrality-based importance of node v_i in a graph G

Eigenvector centrality

- How about directed networks?



$\Rightarrow v_1$ has no in-links $\Rightarrow x_1 = 0$

$$x_i = \frac{1}{\lambda_1} \sum_{j=1}^{|V|} a_{ij} x_j$$

$\Rightarrow x_2 = 0$

$$x_i = \frac{1}{\lambda_1} \sum_{j=1}^{|V|} a_{ij} x_j$$

$\Rightarrow x_3 = 0$

In directed networks, $a_{ij} = 1$ denotes that there exists a link from node v_j to v_i.

- Eigenvector centrality does not work well for Directed Acyclic Graph.

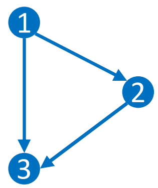
PageRank centrality

$$x_i^{t+1} = \sum_{j=1}^{|V|} \frac{a_{ij}}{d_j^{out}} x_j^t$$

$$\Rightarrow \boldsymbol{X} = \boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{X}$$

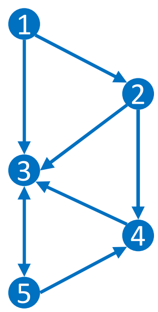
$$\boldsymbol{D} = diag(\{d_i^{out}\}_{j=1}^{|V|})$$

In directed networks, $a_{ij} = 1$ denotes that there exists a link from node v_j to v_i.



v_3 has no out-links, there are two issues:

- $\frac{a_{ij}}{d_j^{out}}$ with d_j^{out}=0, one way to solve it is to set $\frac{a_{ij}}{\max(d_j^{out},1)}$
- **Node importance is leaked out as v_1 and v_2 “give” importance to v_3 but v_3 never distributes importance out to others**

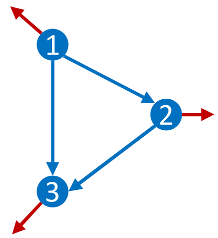


v_3, v_4, v_5 have no out-links as a whole (group):

- **Node importance is absorbed by the sub-group v_3, v_4, v_5**

Page, Brin, Motwani, Winograd. The PageRank citation ranking: bring order to the Web. 1999.

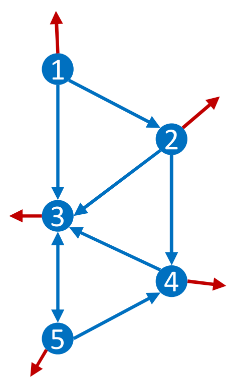
PageRank centrality



$$x_i^{t+1} = \sum_{j=1}^{|V|} \frac{a_{ij}}{d_j^{out}} x_j^t$$

$$\Rightarrow \boldsymbol{X} = \boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{X}$$

$$\boldsymbol{D} = diag(\{d_i^{out}\}_{j=1}^{|V|})$$



In directed networks, $a_{ij} = 1$ denotes that there exists a link from node v_j to v_i.

Solution: Every node is "virtually" connected with all the nodes in the graph (red links) with an equal chance $\frac{1}{|V|}$ to each of them

$$x_i^{t+1} = \beta \sum_{j=1}^{|V|} \frac{a_{ij}}{d_j^{out}} x_j^t + (1-\beta)\frac{1}{|V|}$$

$$\Rightarrow \boldsymbol{X} = \beta\boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{X} + (1-\beta)\frac{1}{|V|}$$

Interpretation: With probability β, random walk over the original graph structure; with probability $(1-\beta)$, random jump to any node

Hyperlink-Induced Topic Search (HITS)

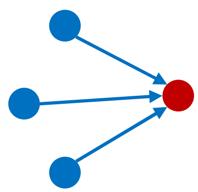
- Some nodes, i.e., web pages, serves as web directories (hubs) that link to pages with authoritative information.
- The importance of each page/node is determined by two scores, one is its Hub centrality and the other one is its Authority centrality.

Search experiences of 1996

- Arts
- Business
- Computers
- Economy
- Education
- Entertainment
- Government
- Health
- International
- Internet
- News & Media
- Reference
- Regional
- Science
- Sports
- Society

Kleinberg. Authoritative sources in a hyperlinked environment. 1997.

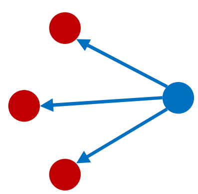
Hyperlink-Induced Topic Search (HITS)



- Many blue nodes consider the red one as an authority

Authority x_i of a node v_i

$$x_i = \alpha \sum_{j=1}^{|V|} a_{ij} y_j$$

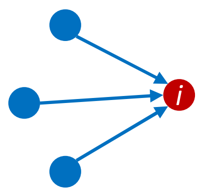


- The blue node knows where to find many authorities

Hubness y_i of a node v_i

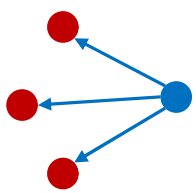
$$y_i = \beta \sum_{i=1}^{|V|} a_{ji} x_j$$

Hyperlink-Induced Topic Search (HITS)



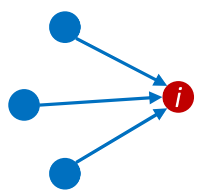
$$x_i = \alpha \sum_{j=1}^{|V|} a_{ij} y_j \qquad \boldsymbol{X} = \alpha \boldsymbol{A}\boldsymbol{Y} \qquad \boldsymbol{X} = \alpha\beta \boldsymbol{A}\boldsymbol{A}^T \boldsymbol{X} \qquad \boldsymbol{A}\boldsymbol{A}^T \boldsymbol{X} = \lambda \boldsymbol{X}$$

$$\lambda = \frac{1}{\alpha\beta}$$



$$y_i = \beta \sum_{i=1}^{|V|} a_{ji} x_j \qquad \boldsymbol{Y} = \beta \boldsymbol{A}^T \boldsymbol{X} \qquad \boldsymbol{Y} = \alpha\beta \boldsymbol{A}^T \boldsymbol{A}\boldsymbol{Y} \qquad \boldsymbol{A}^T \boldsymbol{A}\boldsymbol{Y} = \lambda \boldsymbol{Y}$$

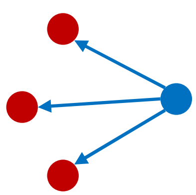
Hyperlink-Induced Topic Search (HITS)



$$\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{X} = \lambda\boldsymbol{X}$$

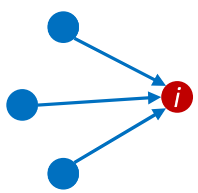
$\boldsymbol{A}\boldsymbol{A}^T$ and $\boldsymbol{A}^T\boldsymbol{A}$ have the same eigenvalues
by looking at their Characteristic polynomial

$$\lambda = \frac{1}{\alpha\beta}$$



$$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{Y} = \lambda\boldsymbol{Y}$$

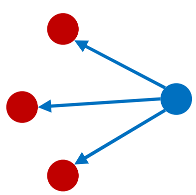
Hyperlink-Induced Topic Search (HITS)



$$\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{X} = \lambda\boldsymbol{X}$$

$\boldsymbol{A}\boldsymbol{A}^T$ and $\boldsymbol{A}^T\boldsymbol{A}$ have the same eigenvalues
by looking at their Characteristic polynomial

$$\lambda = \frac{1}{\alpha\beta}$$



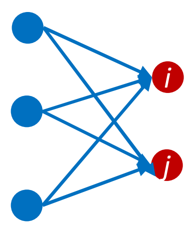
$$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{Y} = \lambda\boldsymbol{Y}$$

Recall the definition the eigenvector centrality

- **Authority** centrality: the eigenvector of $\boldsymbol{A}\boldsymbol{A}^T$ associated with its largest eigenvalue.
- **Hub** centrality: the eigenvector of $\boldsymbol{A}^T\boldsymbol{A}$ associated with its largest eigenvalue.

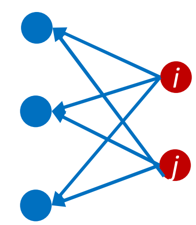
HITS' AA^T and A^TA in MAG

Cocitation



- If v_k points to both v_i & v_j, then v_i & v_j have one cocitation from v_k, that is
$$a_{ik}a_{jk} = 1$$
- How many cocitations do v_i & v_j have?
$$c_{ij} = \sum_{v_k \in V} a_{ik}a_{jk}$$
- Let $C = \{c_{ij}\}$ be the cocitation matrix between any pair of nodes v_i & v_j.
$$C = \boldsymbol{A}\boldsymbol{A}^T$$
- Two papers have many cocitations, meaning the others consider them similar

Coupling



- If both v_i & v_j point to v_k, then v_i & v_j cocite v_k, that is
$$a_{ki}a_{kj} = 1$$
- How many papers do v_i & v_j cocite?
$$c_{ij} = \sum_{v_k \in V} a_{ki}a_{kj}$$
- Let $C = \{c_{ij}\}$ be the coupling matrix between any pair of nodes v_i & v_j.
$$C = \boldsymbol{A}^T\boldsymbol{A}$$
- Two papers have a high coupling score, meaning the author themselves (implicitly) consider these two papers similar

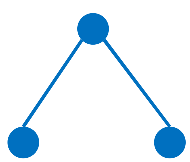
Graph theory

- Graphicality
- Graph isomorphism
- Graph coloring
- Set cover and vertex cover
- Network Flow
- Fractals and scaling
- … …

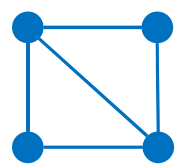
Graphicality

- Degree sequence: given a graph G with n nodes, the list of its nodes' degrees $\{d_1, d_2, d_3, \cdots, d_n\}$ forms its degree sequence.
- Graphicality: A finite sequence of non-negative integers $\boldsymbol{d}$ is graphical if a <u>simple graph</u> G can be constructed with its degree sequence as $\boldsymbol{d}$**;** The graph G is one of $\boldsymbol{d}$'s graphical realization.

$\boldsymbol{d} = \{2,1,1\}$

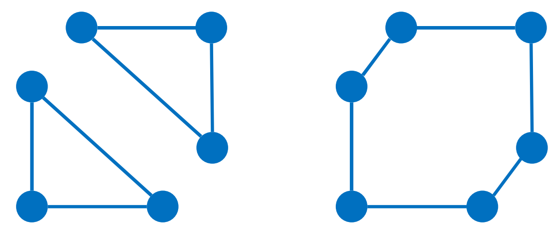


$\boldsymbol{d} = \{3,3,2,2\}$



$\boldsymbol{d} = \{3,2,1\}$

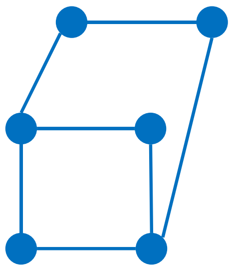
$\boldsymbol{d} = \{2,2,2,2,2,2\}$

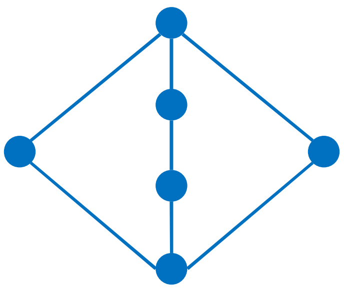


- How can we tell whether a sequence is graphical?
- If one sequence is graphical, how can we have its graphical realization?

Graph isomorphism

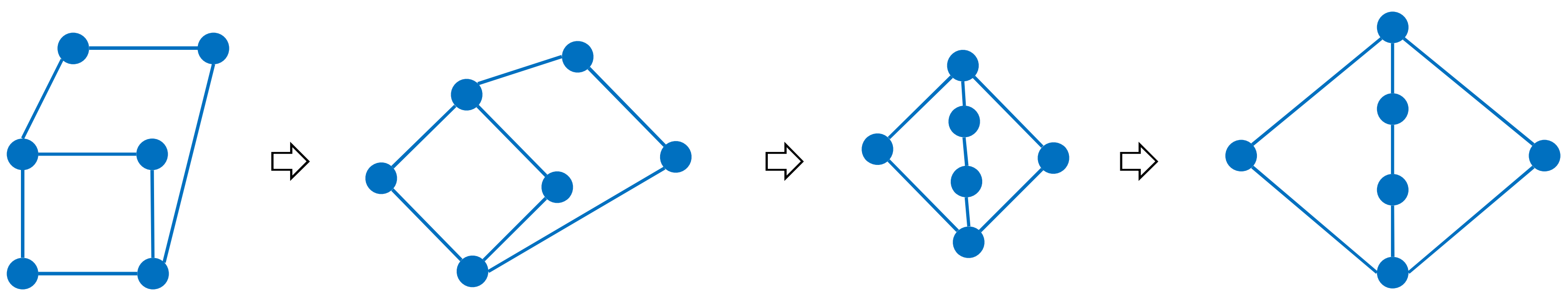
- Two graphs $G = (V, E)$ and $G' = (V', E')$ are isomorphic ($G \simeq G'$), if there exists a projection $\varphi: V \to V'$ such that for all $v_i, v_j \in V$, $\left(v_i, v_j\right) \in E$ if and only if $\left(\varphi(v_i), \varphi(v_j)\right) \in E'$.





Graph isomorphism

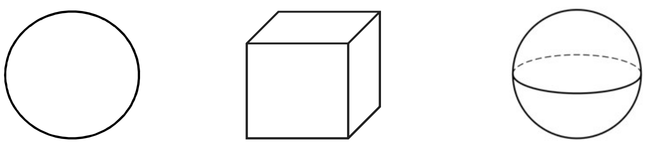
- Two graphs $G = (V, E)$ and $G' = (V', E')$ are isomorphic ($G \simeq G'$), if there exists a projection $\varphi: V \rightarrow V'$ such that for all $v_i, v_j \in V$, $\left(v_i, v_j\right) \in E$ if and only if $\left(\varphi(v_i), \varphi(v_j)\right) \in E'$.



Fractals and scaling

- What are the surface-to-volume ratios for sphere, cube, and circle?
- How about the ratios for objects in nature?

- Why can't we keep growing in body size and weight?
- Why can't we live for 1000 years, but ~100 years?
- Why can't we sleep for 1 hour or 23 hours per day?



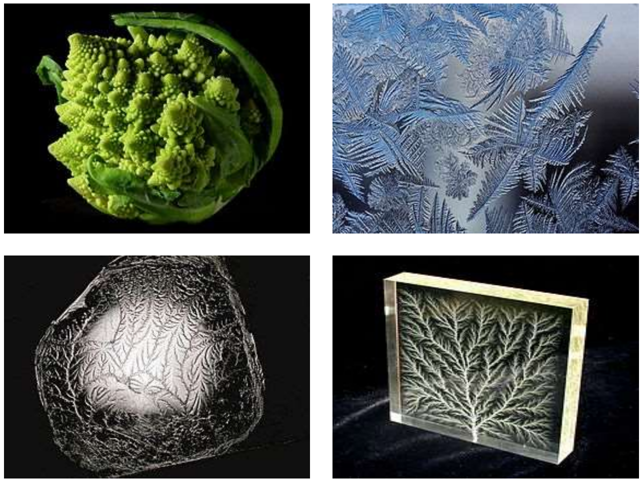
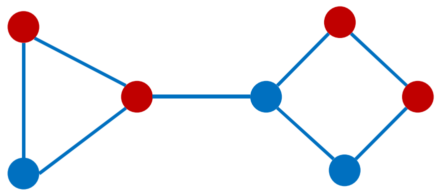


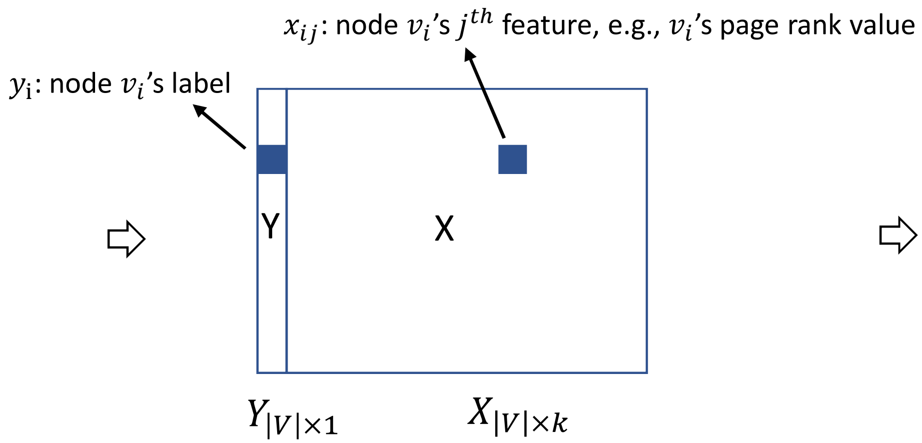
Image Credit: https://en.wikipedia.org/wiki/Fractal

Module 2: Graph Properties and Applications

- Graph basics
  - Graph history
  - Basic node centralities
  - Eigenvector, HITS, & PageRank
- Graph applications
  - Node label classification
  - Community detection
  - Link prediction
- What will not be covered
  - Graph Theory

Node label classification / regression



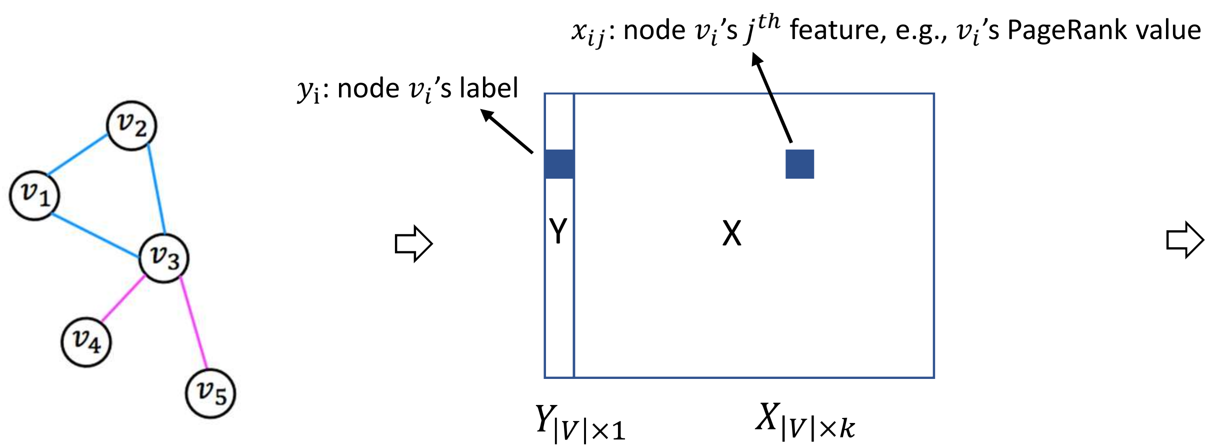


- **Classification algorithms,** such as logistic regression, SVM, and random forest
- **Regression algorithms,** such as linear regression.

Application: Demographic prediction in social networks

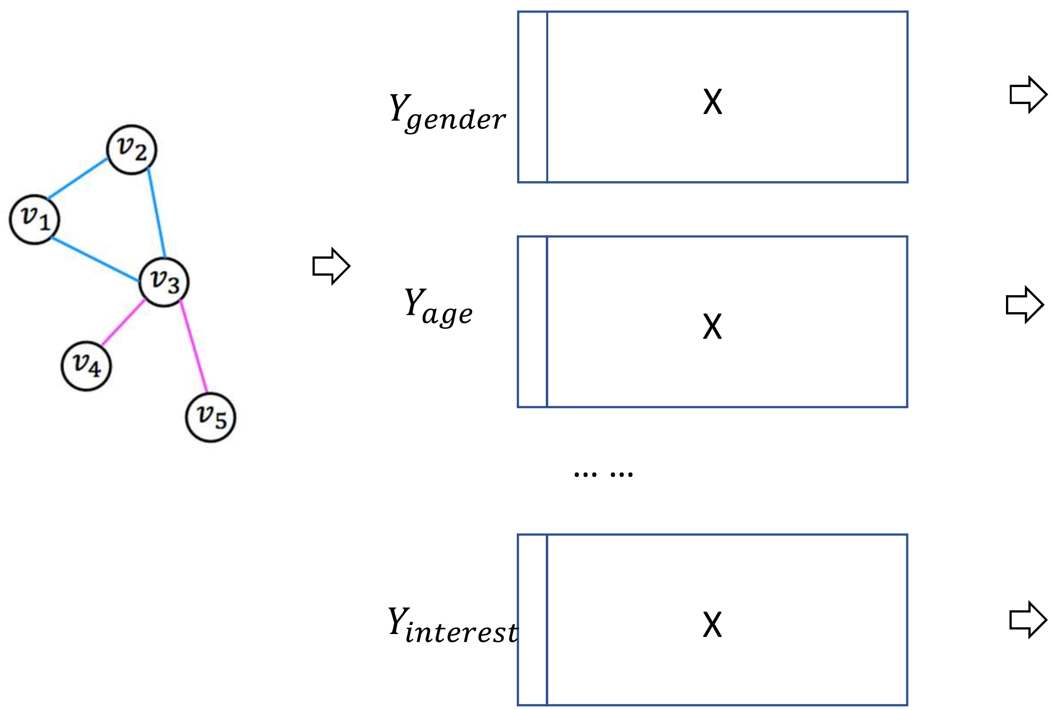
Given a network, we know part of its users' demographic attributes (e.g., gender, age, political party, interest, hometown, etc.), ***can we infer the other users' demographics?***

Application: Demographic prediction in social networks



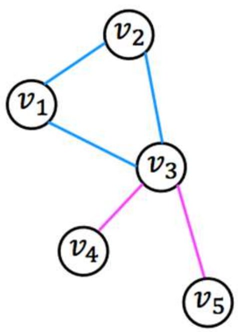
- **Classification algorithms,** such as logistic regression, SVM, and random forest
- **Regression algorithms,** such as linear regression.

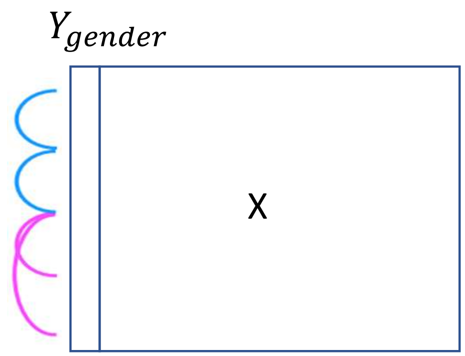
Application: Demographic prediction in social networks



- **Classification**
- **Regression**
- **Classification**

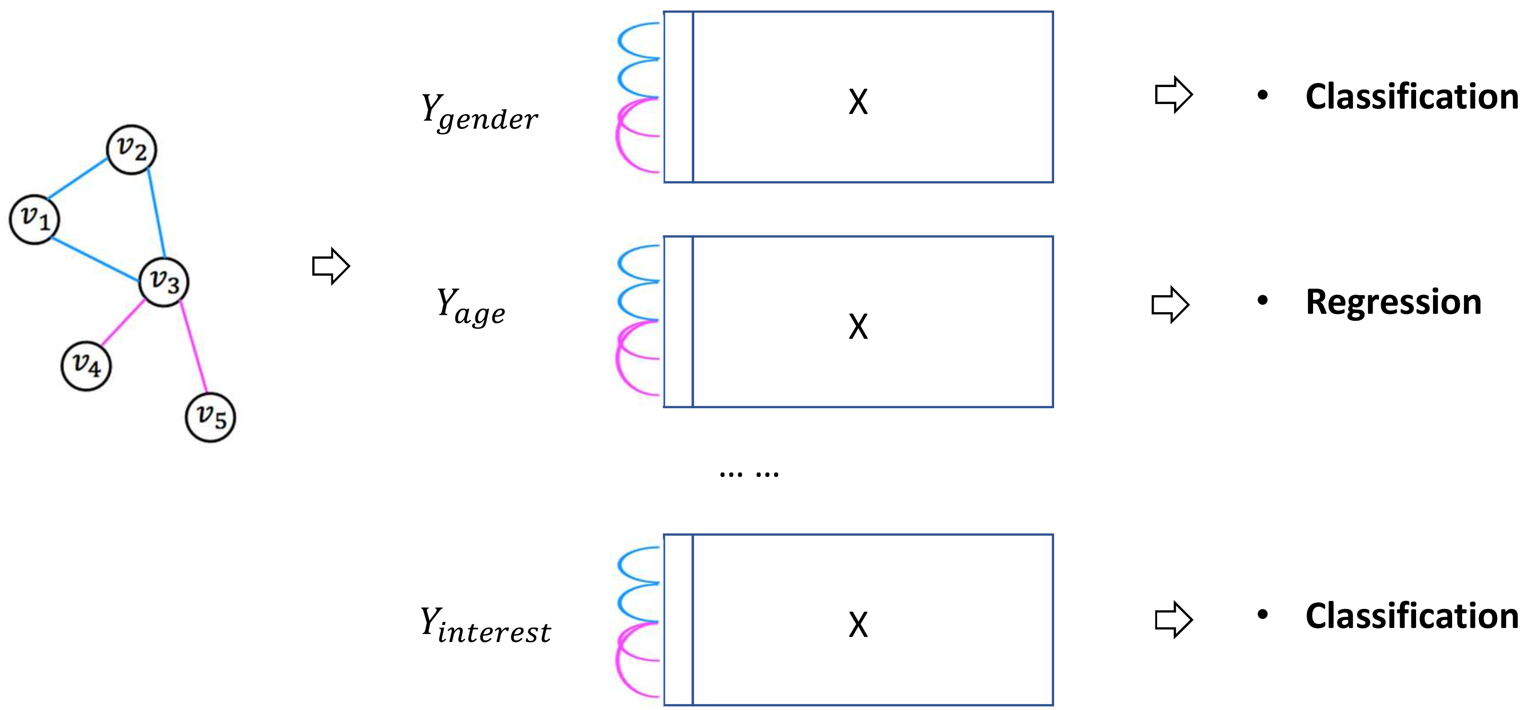
Application: Demographic prediction in social networks



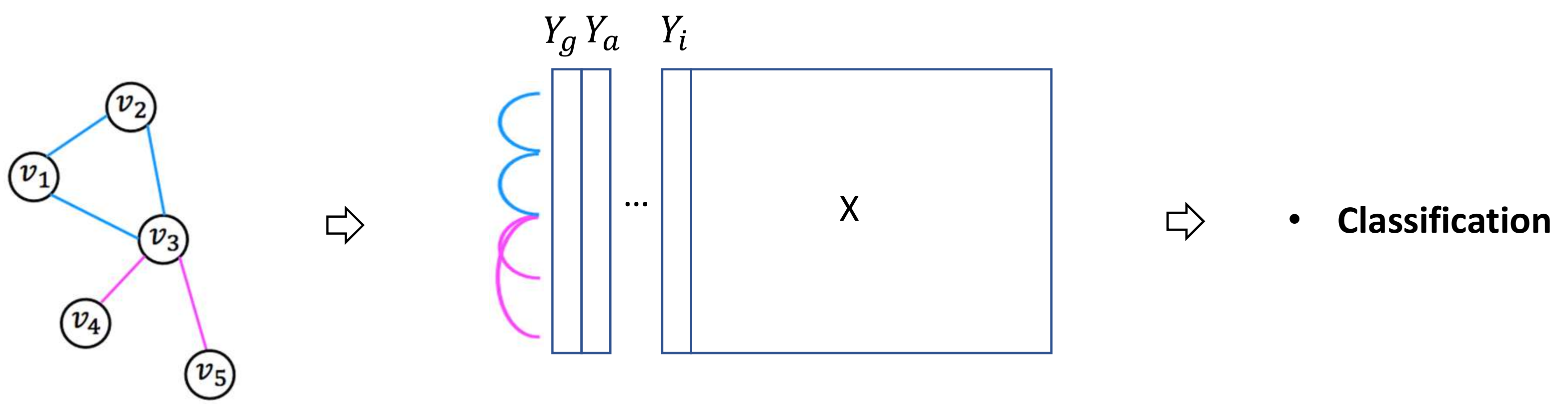


- **Classification**

Application: Demographic prediction in social networks



Application: Demographic prediction in social networks



- In Mobile phone communication networks:
  - Dong et al. Inferring user demographics and social strategies in mobile social networks. In *ACM KDD 2014.*
- In Facebook online social networks:
  - Chakrabarti et al. Joint inference of multiple label types in large networks. In *ICML 2014.*

Node label classification in MAG

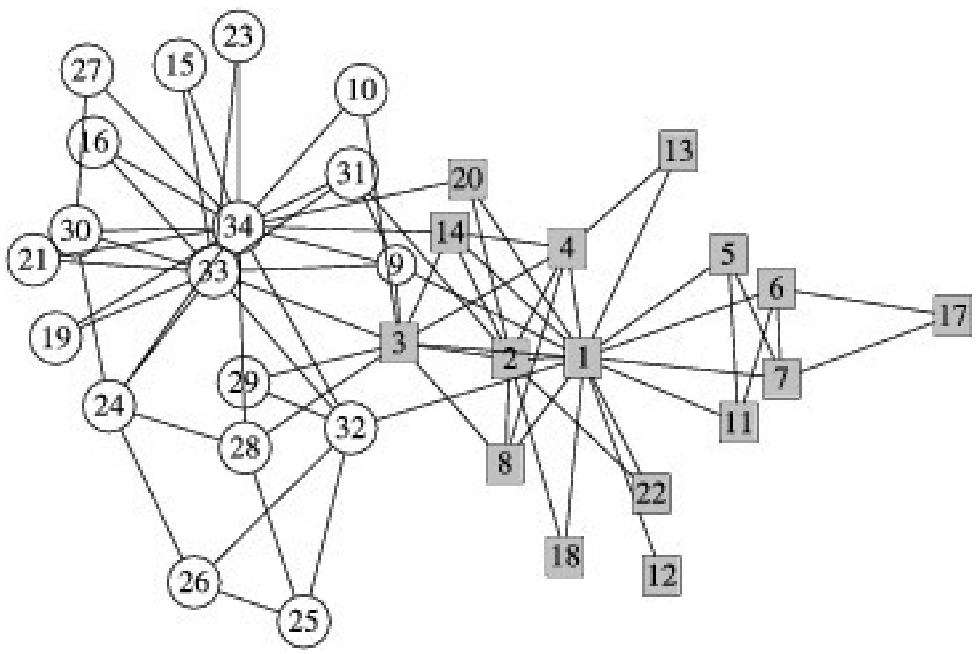
- For each paper in an academic graph, can we infer its fields of study from collaboration/citation network structure?

- For example, the 170 million publications in Microsoft Academic Graph cover 19 high-level fields in *Math, Physics, Computer Science, Chemistry, Biology, Engineering, Arts, Business, Economics, Environmental Science, Geography, Geology, History, Materials Science, Medicine, Philosophy, Political Science, Psychology, and Sociology.*

- The problem is how to decide each paper's fields of study.

- Shen, Ma, Wang. A Web-scale system for scientific knowledge exploration. In *ACL 2018*.

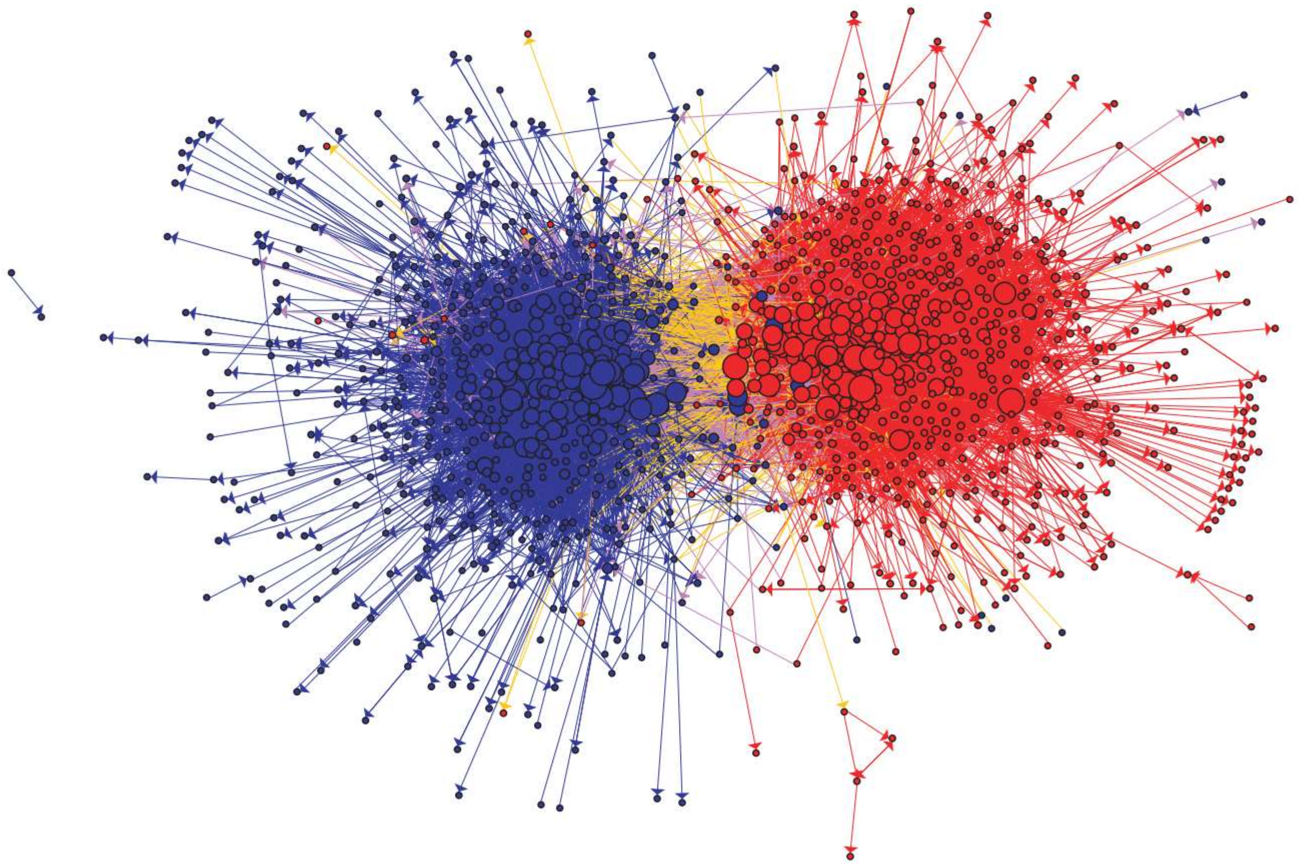
Community detection / node clustering



Zachary's Karate club network

Zachary. An Information Flow Model for Conflict and Fission in Small Groups. In *Journal of Anthropological Research*, 1977.

Community detection / node clustering



Adamic and Glance. The political blogosphere and the 2004 U.S. election: divided they blog. In *LinkKDD* 2005

Community detection / node clustering

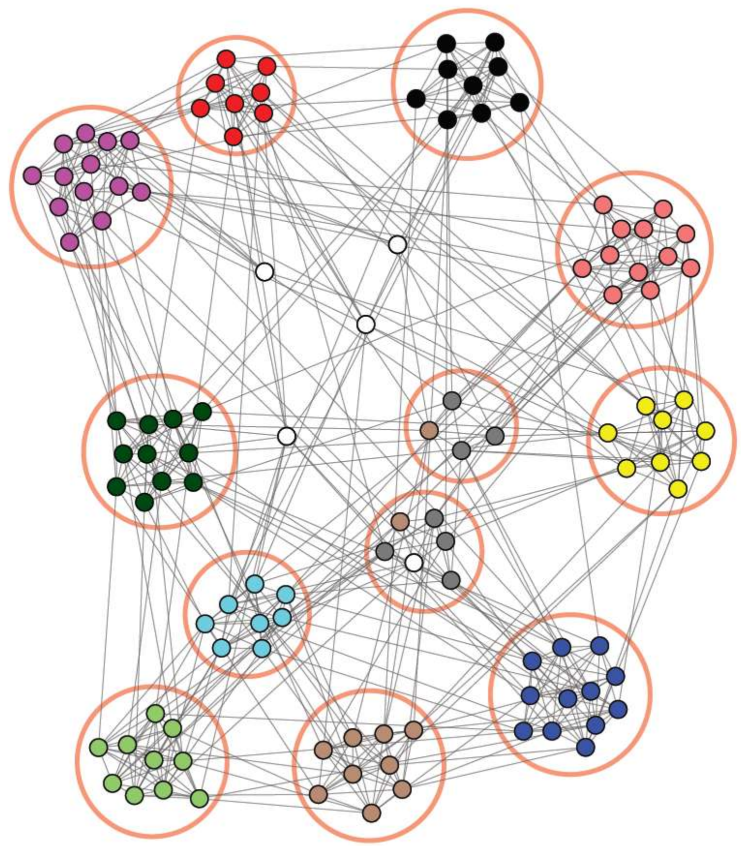
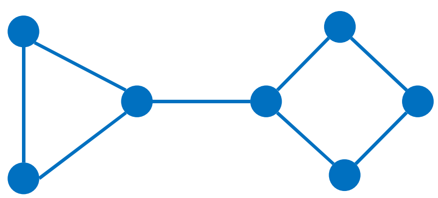
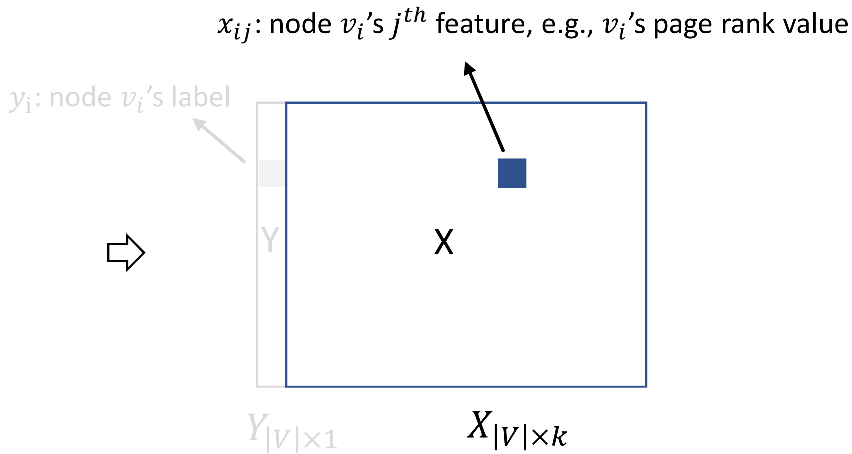


Image Credit: http://snap.stanford.edu/agm/

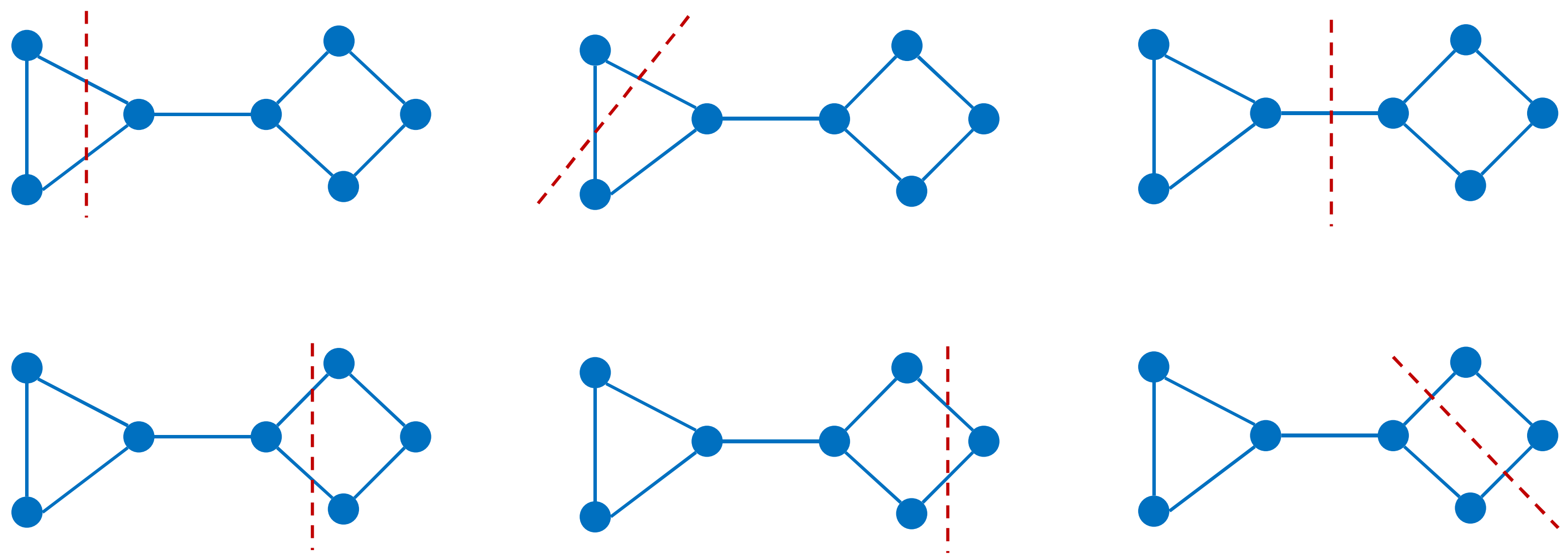
Community detection / node clustering



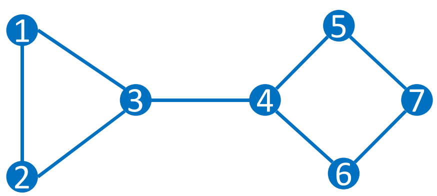


- **Classification algorithms,** such as logistic regression, SVM, and random forest
- **Regression algorithms,** such as linear regression.
- **Clustering algorithms**, such as k-means

Community detection / node clustering



Spectral clustering

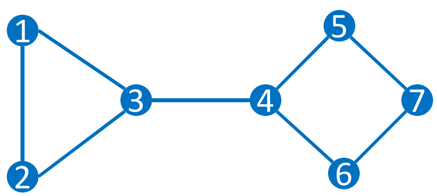


$$\boldsymbol{D} = diag\{d_i\}_{7\times 7} \qquad \boldsymbol{A} = \{a_{ij}\}_{7\times 7}$$

$$L = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix} - \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

Graph Laplacian

Graph Laplacian Matrix



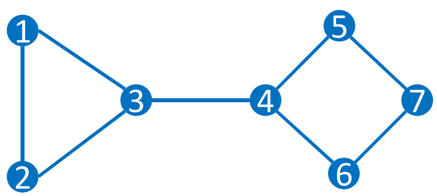
$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$$

$$\begin{bmatrix} 2 & -1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 & 0 \\ 0 & 0 & 0 & -1 & 2 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 2 & -1 \\ 0 & 0 & 0 & 0 & -1 & -1 & 2 \end{bmatrix}$$

Graph Laplacian

- $\boldsymbol{L}$ is Positive semidefinite
- $\boldsymbol{L}$'s Eigenvalues are non-negative
- $\boldsymbol{L}$'s Eigenvectors are real & orthogonal

Graph Laplacian Matrix



$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$$

$$\begin{bmatrix} 2 & -1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 & 0 \\ 0 & 0 & 0 & -1 & 2 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 2 & -1 \\ 0 & 0 & 0 & 0 & -1 & -1 & 2 \end{bmatrix}$$

Graph Laplacian

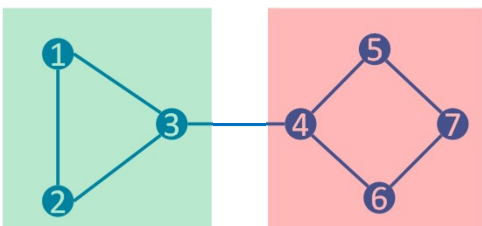
λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
0	0.359	2	2.28	3	3.59	4.48

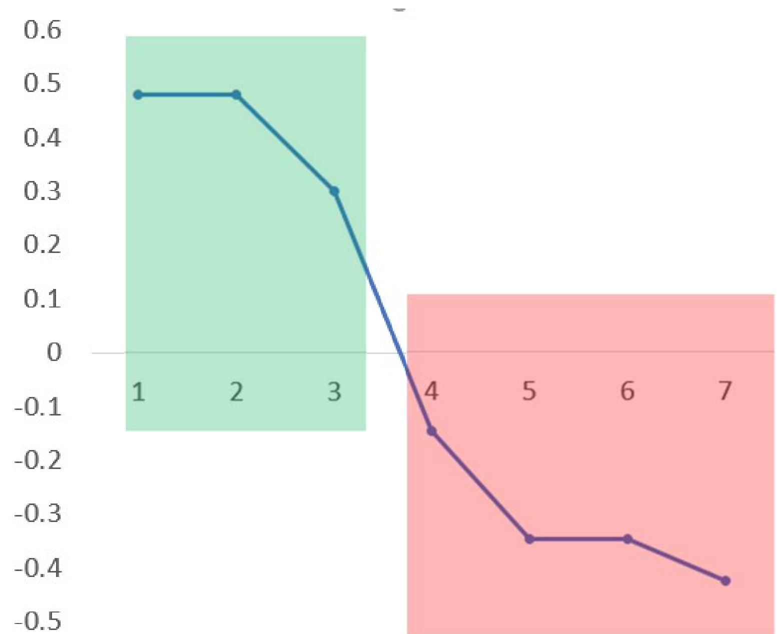
Eigenvalues

x_1	x_2	x_3	x_4	x_5	x_6	x_7
-0.378	0.48	0	0.21	0.71	-0.25	-0.14
-0.378	0.48	0	0.21	-0.71	-0.25	-0.14
-0.378	0.3	0	-0.27	0	0.64	0.53
-0.378	-0.147	0	-0.63	0	0.12	-0.66
-0.378	-0.348	-0.7	-0.09	0	-0.36	0.32
-0.378	-0.348	-0.7	-0.09	0	-0.36	0.32
-0.378	-0.424	0	0.65	0	0.45	-0.23

Eigenvectors

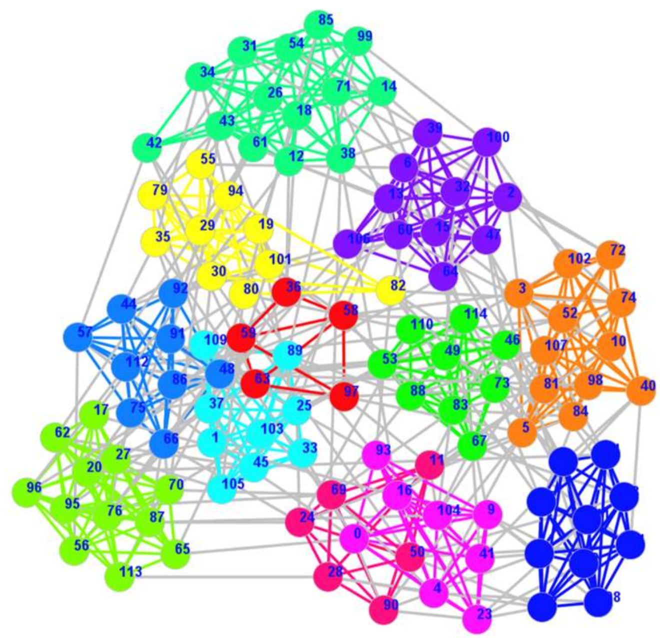
Spectral clustering





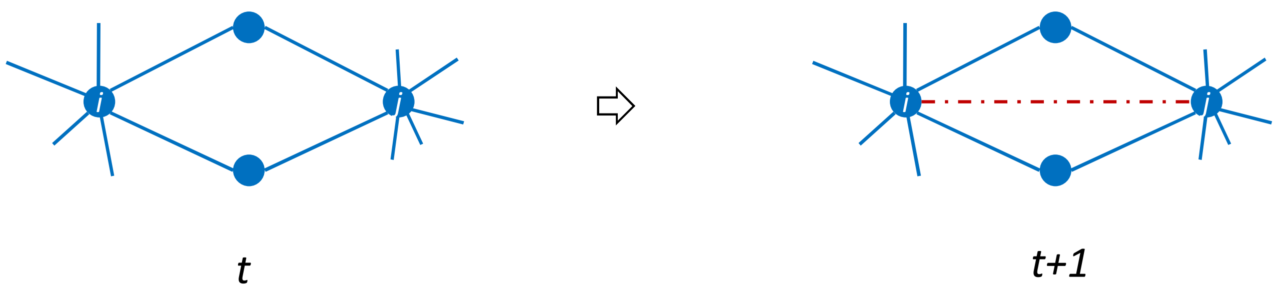
More than two communities?

- How to determine the number of clusters k?
- How to partition a graph into k clusters?



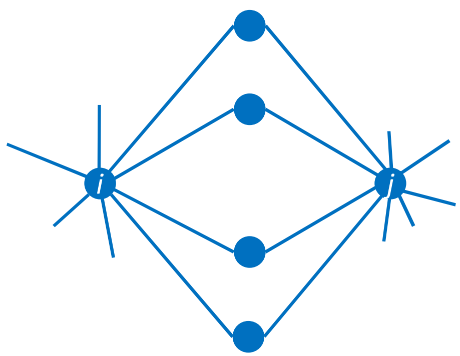
Link prediction

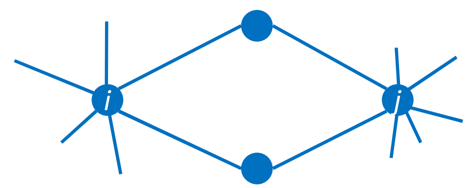
- Given two nodes v_i and v_j that are not connected right now, we aim to infer whether a link will form between them.
  - Friend recommendation, e.g., "People you may know" on LinkedIn or Facebook, "Who to follow" on Twitter
  - Item recommendation, e.g., movies to watch in Netflix, books to buy in Amazon



Similarity in networks

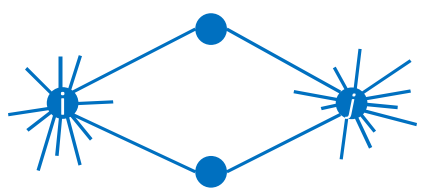
- The number of common neighbors between two nodes
- $S_{ij} = |N(v_i) \cap N(v_j)|$, where $N(v_i)$ represents the neighbors of v_i.

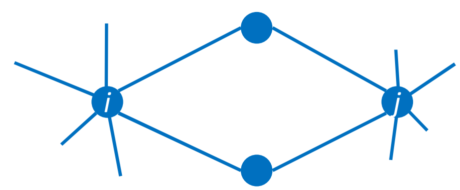




Similarity in networks

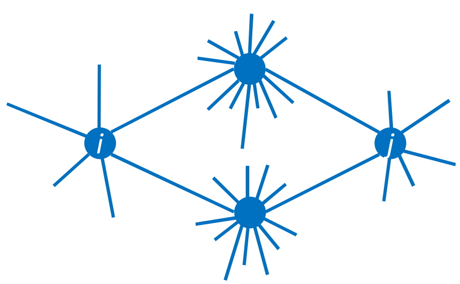
- The intersection of two's neighbors over the union of their neighbors
- $S_{ij} = \frac{|N(v_i) \cap N(v_j)|}{|N(v_i) \cup N(v_j)|}$, where $N(v_i)$ represents the neighbors of v_i.

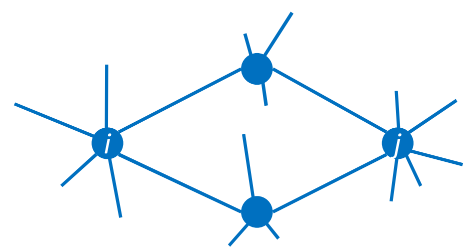




Similarity in networks

- Adamic Adar
- $S_{ij} = \sum_{v_p \in N(v_i) \cap N(v_j)} \frac{1}{\log|N(v_i)|}$

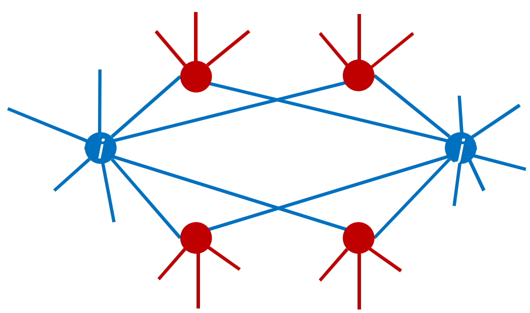


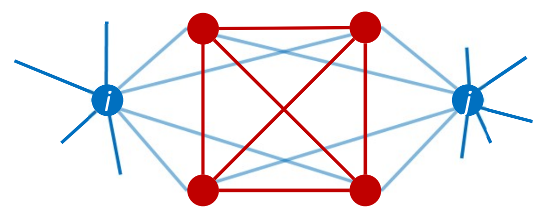


Adamic and Glance. The political blogosphere and the 2004 U.S. election: divided they blog. In *LinkKDD* 2005

Similarity in networks

- Structural diversity of common neighbors

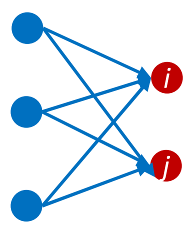




Dong et al. Structural diversity and homophily: A study Across More Than One Hundred Big Networks. In KDD 2017.

Structural similarity in MAG

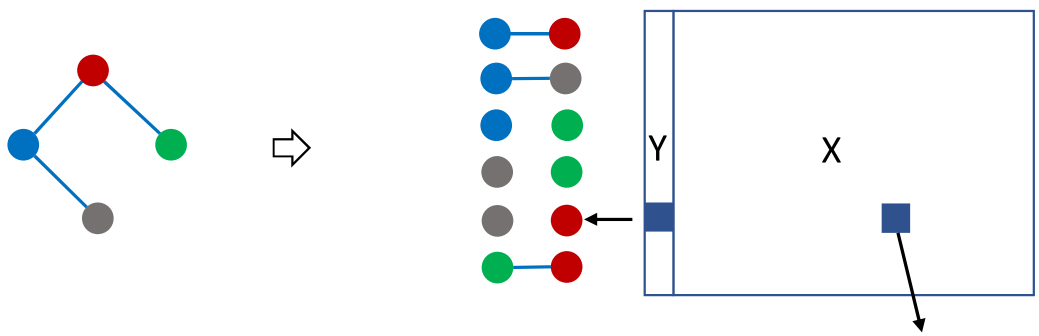
- #common-coauthors of two author collaboration graphs
- #cocitations of two papers in citation graphs



paper citation graph

- If v_k points to both v_i & v_j, then v_i & v_j have one cocitation from v_k, that is
$$a_{ik}a_{jk} = 1$$
- How many cocitations do v_i & v_j have?
$$c_{ij} = \sum_{v_k \in V} a_{ik}a_{jk}$$
- Let $\boldsymbol{C} = \{c_{ij}\}$ be the cocitation matrix between any pair of nodes v_i & v_j.
$$\boldsymbol{C} = \boldsymbol{A}\boldsymbol{A}^T$$
- Two papers have many cocitations, meaning the others consider them similar

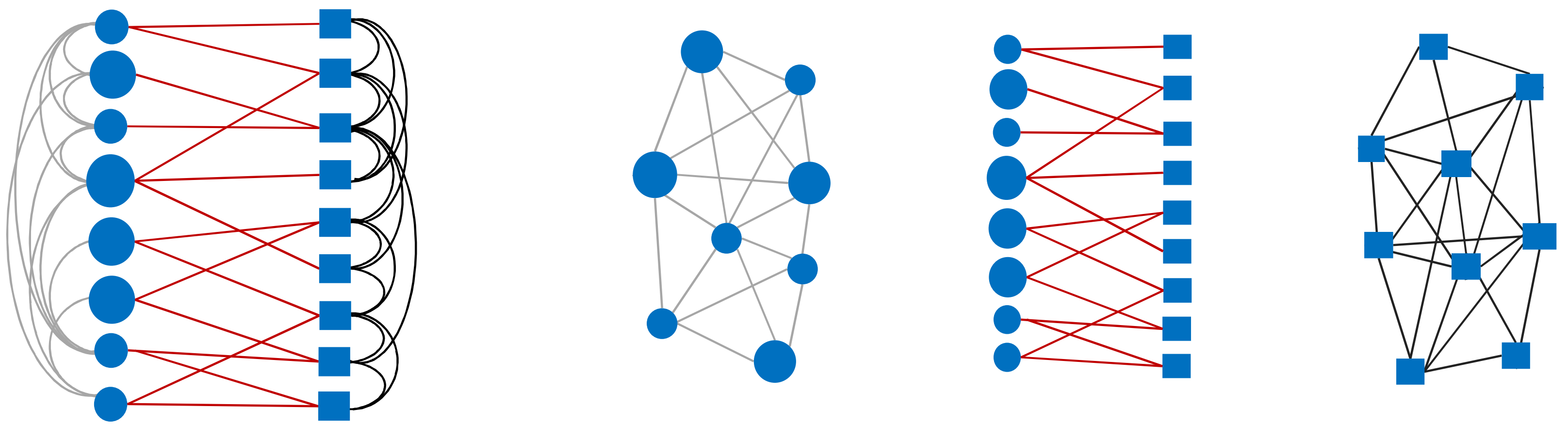
Link prediction



x_{ij}: e_i's j^{th} feature, e.g., e_i's #common-neighbors

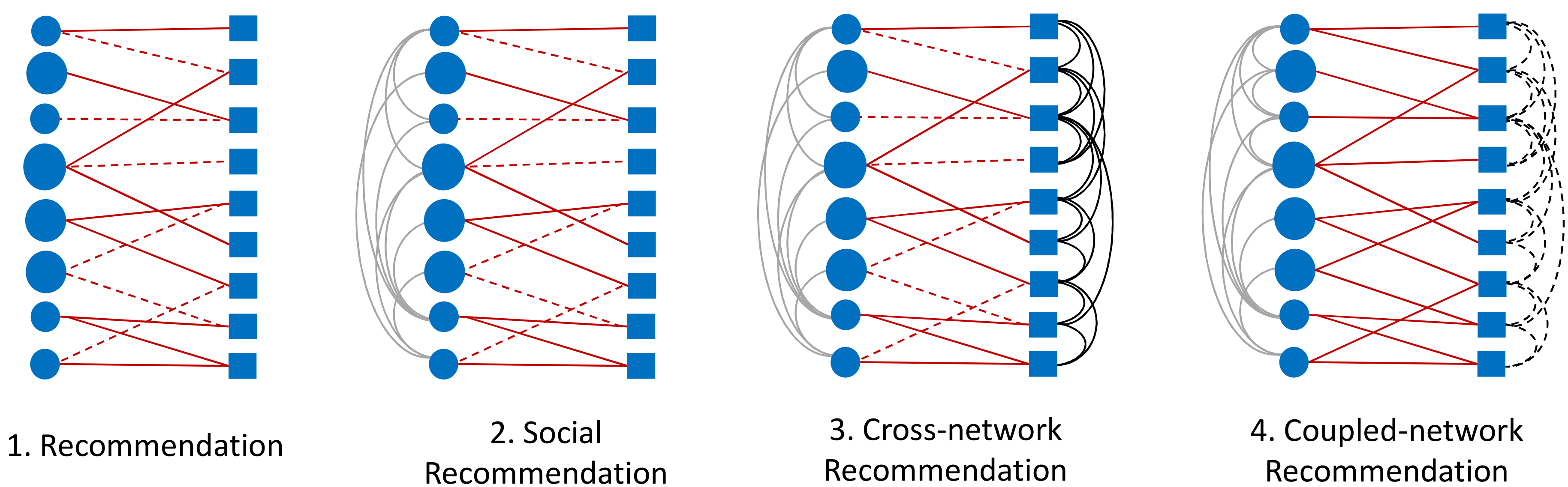
Classification algorithms, such as logistic regression, SVM, and random forest

Link prediction in heterogeneous networks



$G = (V_s, V_t, E_s, E_t, E_{st})$ $G_s = (V_s, E_s)$ $G_{st} = (V_s, V_t, E_{st})$ $G_t = (V_t, E_t)$

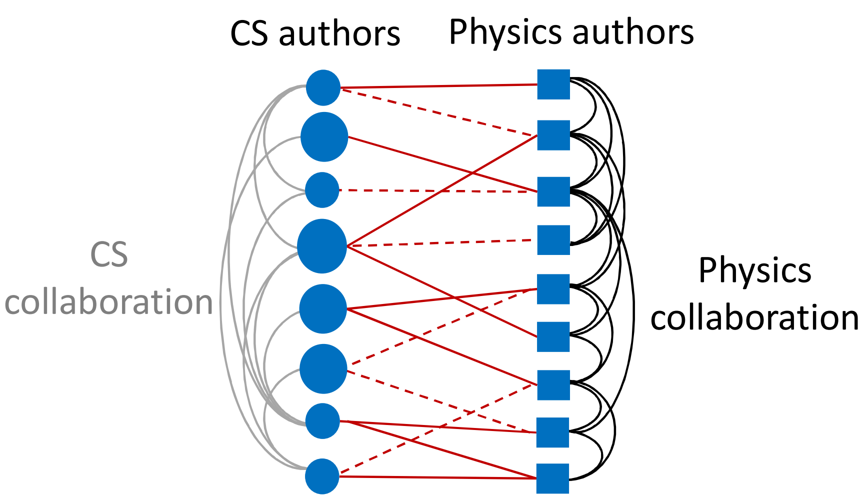
Link prediction in heterogeneous networks



- Koren et al. Matrix factorization techniques for recommendation systems. In *IEEE Computer*, 2009.
- Ma et al. SoRec: social recommendation using probabilistic matrix factorization. In *ACM CIKM* 2008.
- Tang et al. Cross-domain collaboration recommendation. In *ACM KDD* 2012.
- Dong et al. CoupledLP: link prediction in coupled networks. In *ACM KDD* 2015.

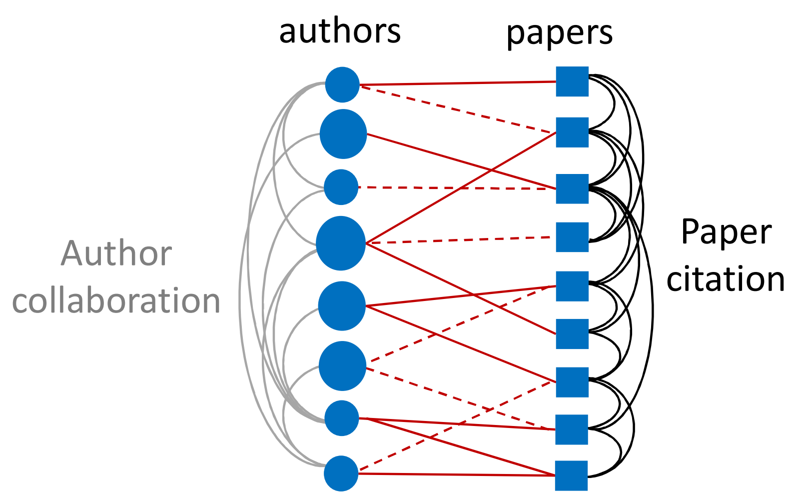
Link prediction in MAG

Which Physicists to collaborate with?



Cross-field Collaborator Recommendation

Which papers to read?



Paper Recommendation

Module 2: Graph Properties and Applications

- Graph basics
  - Graph history
  - Basic node centralities
  - Eigenvector, HITS, & PageRank
- Graph applications
  - Node label classification
  - Community detection
  - Link prediction
- What will not be covered
  - Graph Theory