

Tài liệu giảng dạy môn Khai thác dữ liệu Web

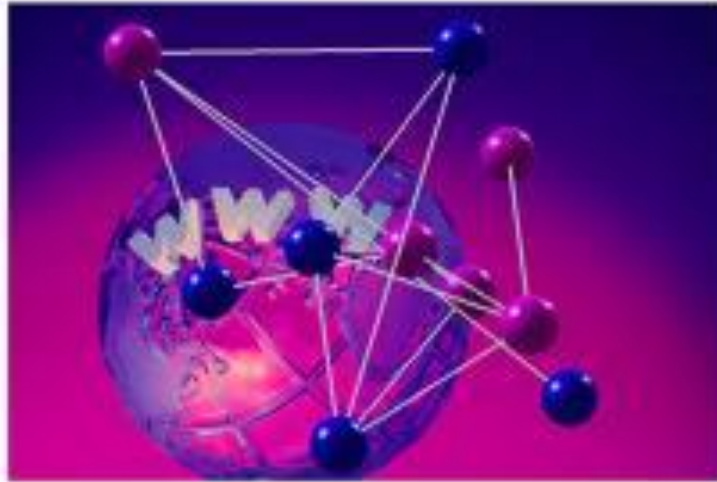
PHÂN TÍCH MẠNG XÃ HỘI

TS. Nguyễn Ngọc Thảo – ThS. Lê Ngọc Thành
Bộ môn Khoa học Máy tính, FIT HCMUS, VNUHCM

Thành phố Hồ Chí Minh, 02/2019

Nội dung bài giảng

- Khai thác cấu trúc Web
 - Mục tiêu và ứng dụng
- Khái niệm đồ thị Web
- Phân tích mạng xã hội
 - Tính trung tâm
 - Tính uy tín
- Phân tích trích dẫn
 - Đồng trích dẫn
 - Liên kết thư mục

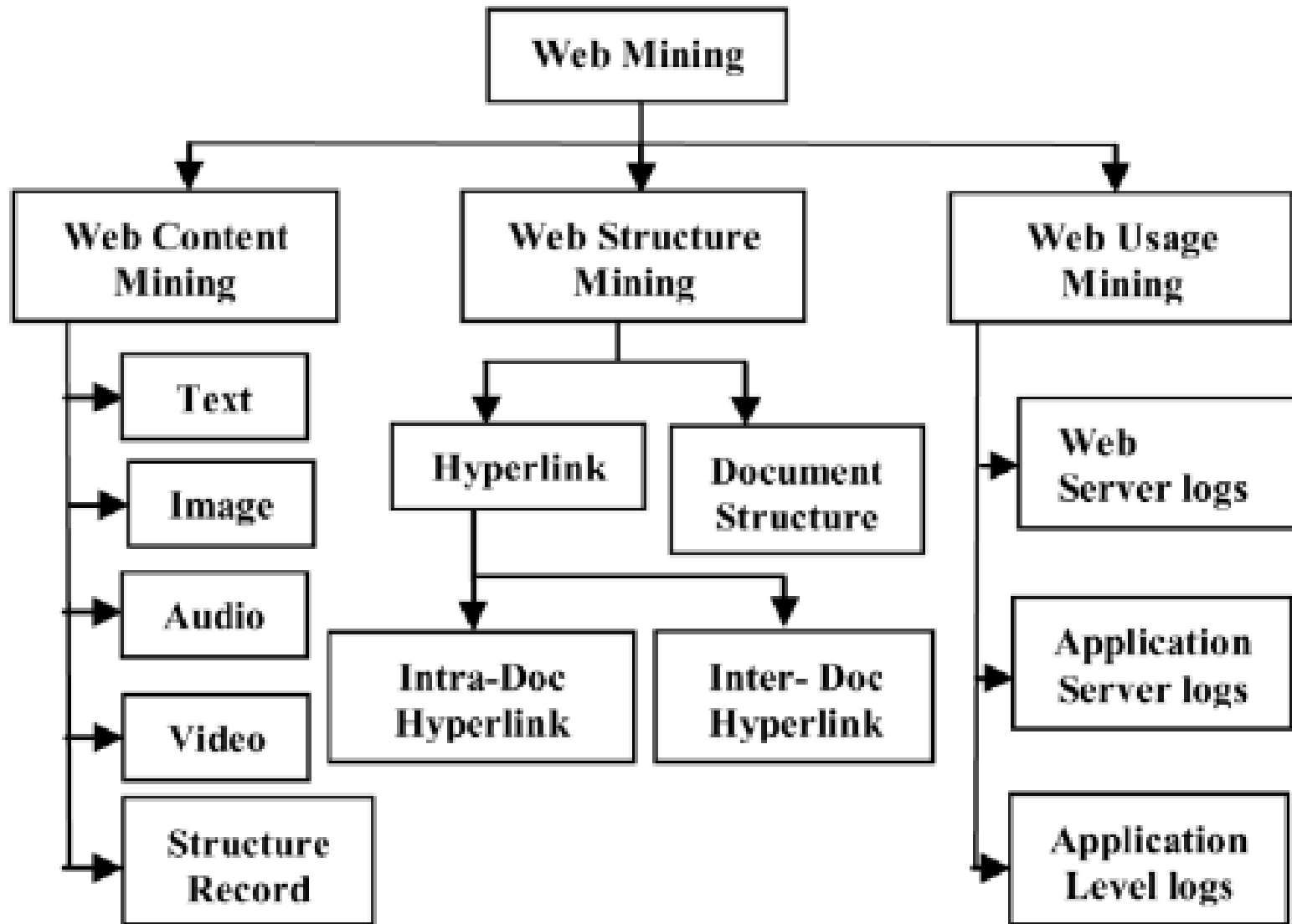


Khai thác cấu trúc Web

Khai thác cấu trúc Web

- Xem mạng lưới các trang Web như đồ thị và ứng dụng lý thuyết đồ thị để phân tích
- Phân loại bài toán dựa theo loại dữ liệu cấu trúc Web
- Rút trích hình mẫu từ siêu liên kết trong Web
 - Siêu liên kết (hyperlink): một dạng thành phần cấu trúc để liên kết từ một trang Web đến một vị trí khác
- Khai thác cấu trúc của tài liệu
 - Phân tích cấu trúc dạng cây của các trang sử dụng HTML hay XML

Hệ thống bài toán Khai thác Web



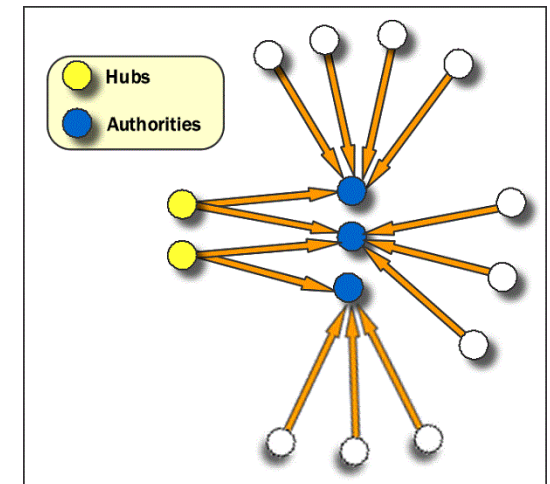
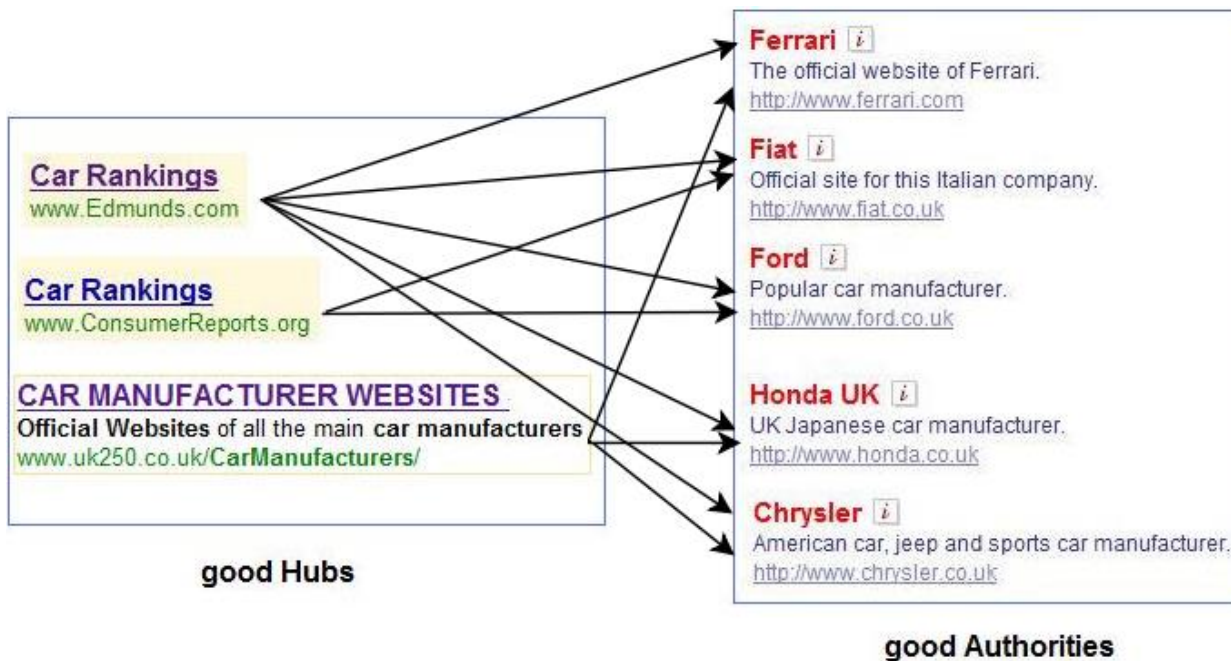
Rút trích hình mẫu từ siêu liên kết

- Phân tích các liên kết hình ảnh theo chủ đề “các loài chim”.



Ứng dụng của Khai thác cấu trúc Web


- Khái niệm **hub** và **authority** thiết lập nền tảng cho nhiều giải thuật xếp hạng kết quả tìm kiếm.
 - Ví dụ, Page Rank, HITS, v.v.




Query: **Top automobile makers**

Ứng dụng của Khai thác cấu trúc Web


- Các trang Web cùng chủ đề thường có cấu trúc tương tự.



Jiawei Han
Professor, Department of
Univ. of Illinois at Urbana-
Rm 2132, Siebel Center f
201 N. Goodwin Avenue
Urbana, IL 61801, USA
E-mail: hanj[at]cs.uiuc.edu
Ph.D. (1985), Computer



Gerald DeJong (a.k.a. Mr. EBL)
Professor of Computer Science
Affiliate of the Electrical and Computer Engineering Department



PROFESSOR MICHAEL T. HEATH
Professor Michael T. Heath is Fulton Watson Copp Chair in the Department of
Computer Science at the University of Illinois at Urbana-Champaign, where he is
also
Cent
anal
comp
Acad

Current Research (Selected Publications)

- Information Network Analysis and Discovery
- Sequential and Structured Pattern Discovery
- Discovery of the Dynamics of Data Stream
- Ranking and Multidimensional Analysis in
- Analysis of Spatiotemporal Trajectories, a
- Knowledge Discovery in Cyberphysical Sys
- Accured Information Sharing Lifecycle (AI
- Software Bug Detection in Sensor Networ
- CS-BibCube: OLAPing and Analysis of Co

Teaching

- UIUC CS512: Data Mining: Principles and
- UIUC CS412: An Introduction to Data Wa
- UIUC CS591 Han Advanced Topics in Data
- UIUC CS591 Yahoo! DAIS (Data and Info

I received my Ph.D. i
Shin. I was an Asista
Who is Mr.
I then joined the Univ
temure. My interests i
1995-present Pr
University of Ill
1985 1995 Ass
Engineering, Uni
1981-1985 Assi
1980 Instructor,
1979 Ph.D., Co
1974 B.S., Phys

ENGINEERING AT ILLINOIS

PROFESSOR MICHAEL T. HEATH

Basic Website Layout

Home Page
(Index page)

Main Sections
(site index)

Subsections
(content)

8

Ứng dụng của Khai thác cấu trúc Web

- Tìm ảnh trên Flickr có nội dung liên quan đến ảnh truy vấn

Which pictures are most similar to this one?



flickr

Evaluate the similarity between images according to their linked tags

Meta-Path: *Image-Tag-Image*



(a) top-1 (b) top-2 (c) top-3



(d) top-4 (e) top-5 (f) top-6

Evaluate the similarity between images according to tags and groups

Meta-Path: *Image-Tag-Image-Group-Image-Tag-Image*



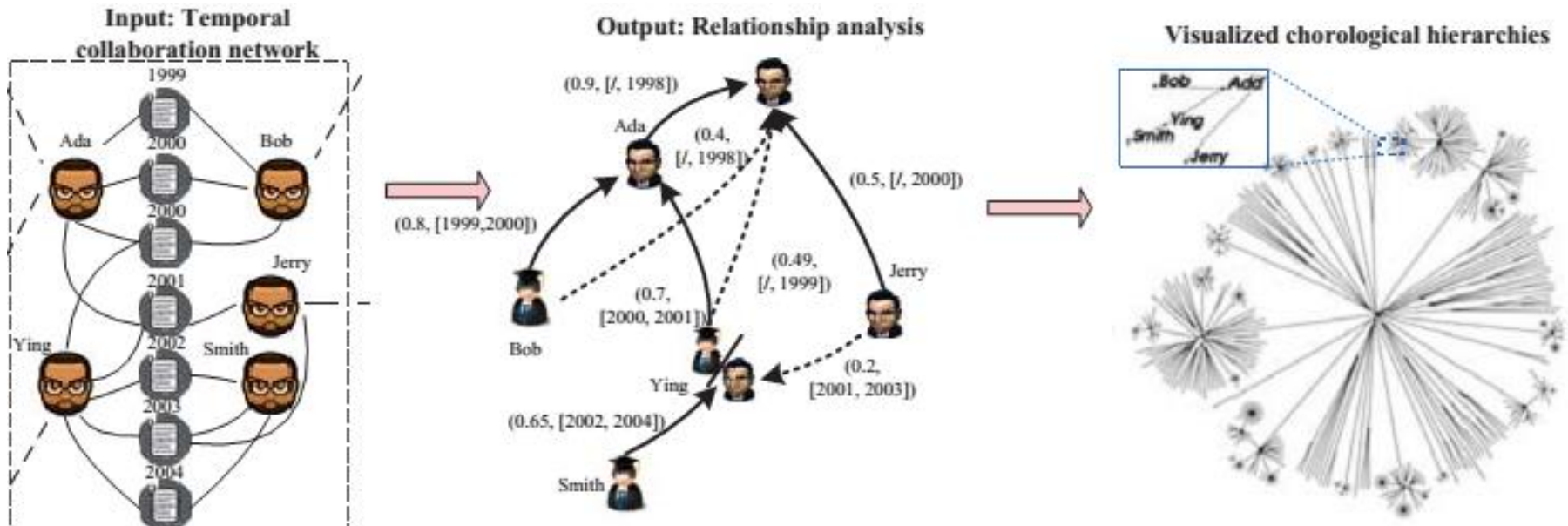
(a) top-1 (b) top-2 (c) top-3



(d) top-4 (e) top-5 (f) top-6

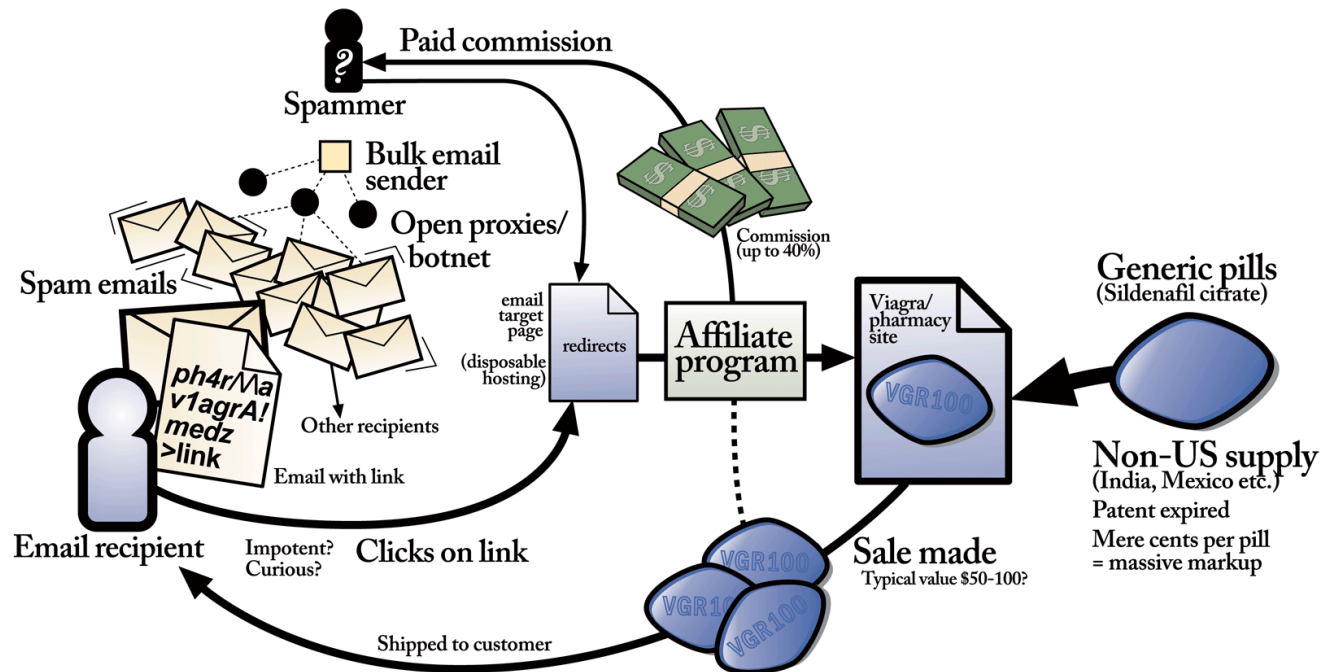
Ứng dụng của Khai thác cấu trúc Web

- C. Wang, J. Han, et al., “Mining Advisor-Advisee Relationships from Research Publication Networks”, SIGKDD 2010
- Đầu vào: Mạng công bố nghiên cứu khoa học DBLP
- Đầu ra: các mối quan hệ người hướng dẫn – học viên tiềm tàng và thứ hạng của chúng ($r, [st, ed]$)



Ứng dụng của Khai thác cấu trúc Web

- Phát hiện spam



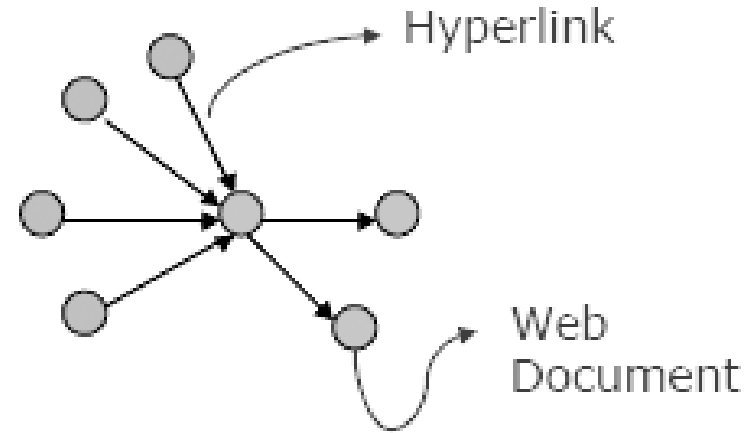
How Viagra spam works

(modern life is rubbish) <http://www.modernlifeisrubbish.co.uk>

Thuật ngữ về cấu trúc Web

- **Đồ thị Web:** Đồ thị có hướng biểu diễn Web

- **Nút:** Mỗi trang Web là một nút của đồ thị Web.
- **Liên kết:** Mỗi siêu liên kết trên Web là một cạnh có hướng của đồ thị Web

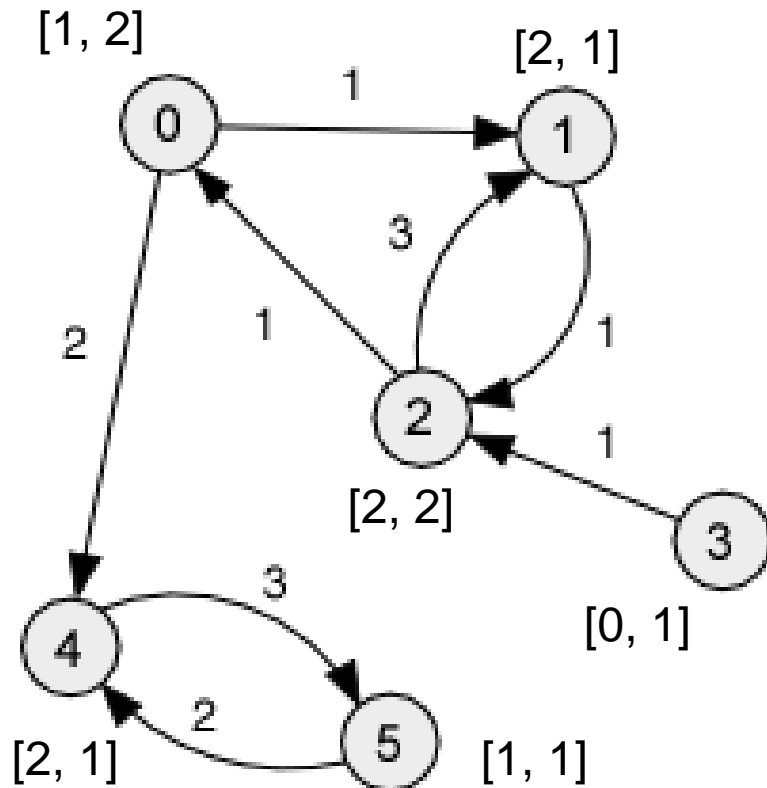


- **Bậc trong** (in-degree) của nút p là số liên kết phân biệt trở đến p .
- **Bậc ngoài** (out-degree) của nút p là số liên kết phân biệt xuất phát từ p trở đến các nút khác.

Thuật ngữ về cấu trúc Web

- **Đường đi có hướng** (Directed path): Một chuỗi các liên kết, bắt đầu từ nút p và lần theo đó có thể đến nút q .
- **Đường đi ngắn nhất** (Shortest path): Đường đi có **độ dài ngắn nhất**, tức là số liên kết trên nó, trong mọi đường đi giữa hai nút p và q .
- **Đường kính** (Diameter): Giá trị **cực đại** trong các đường đi ngắn nhất giữa mọi cặp nút p và q trên đồ thị Web.
- **Khoảng cách kết nối trung bình** (Average connected distance): Giá trị **trung bình** trong các đường đi ngắn nhất giữa mọi cặp nút p và q trên đồ thị Web.

Thuật ngữ về cấu trúc Web: Ví dụ



*[x, y]: [in-degree, out-degree]

| $i \setminus j$ | 0 | 1 | 2 | 3 | 4 | 5 |
|-----------------|---|---|---|----------|---|---|
| 0 | - | 1 | 2 | ∞ | 2 | 5 |
| 1 | 2 | - | 1 | ∞ | 4 | 7 |
| 2 | 1 | 2 | - | ∞ | 3 | 6 |
| 3 | 2 | 3 | 1 | - | 4 | 7 |
| 4 | 0 | 0 | 0 | ∞ | - | 3 |
| 5 | 0 | 0 | 0 | ∞ | 2 | - |

Diameter = 7

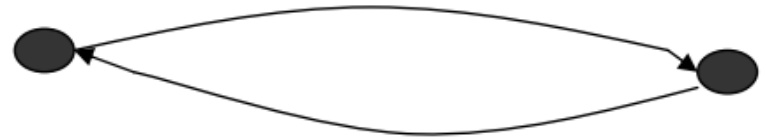
Average connected distance = 58 / 19

Các dạng cấu trúc Web thú vị

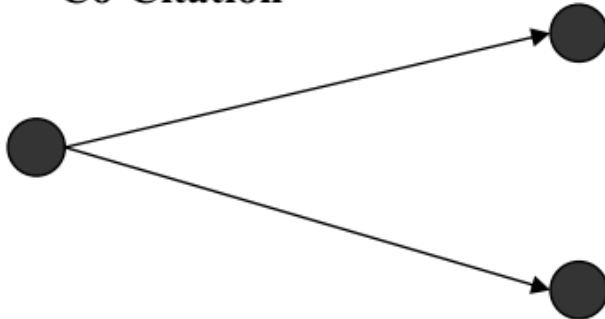
Endorsement



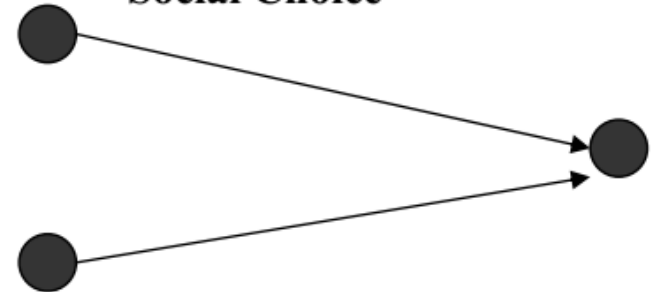
Mutual Reinforcement



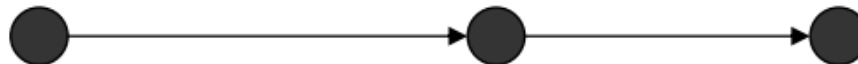
Co-Citation



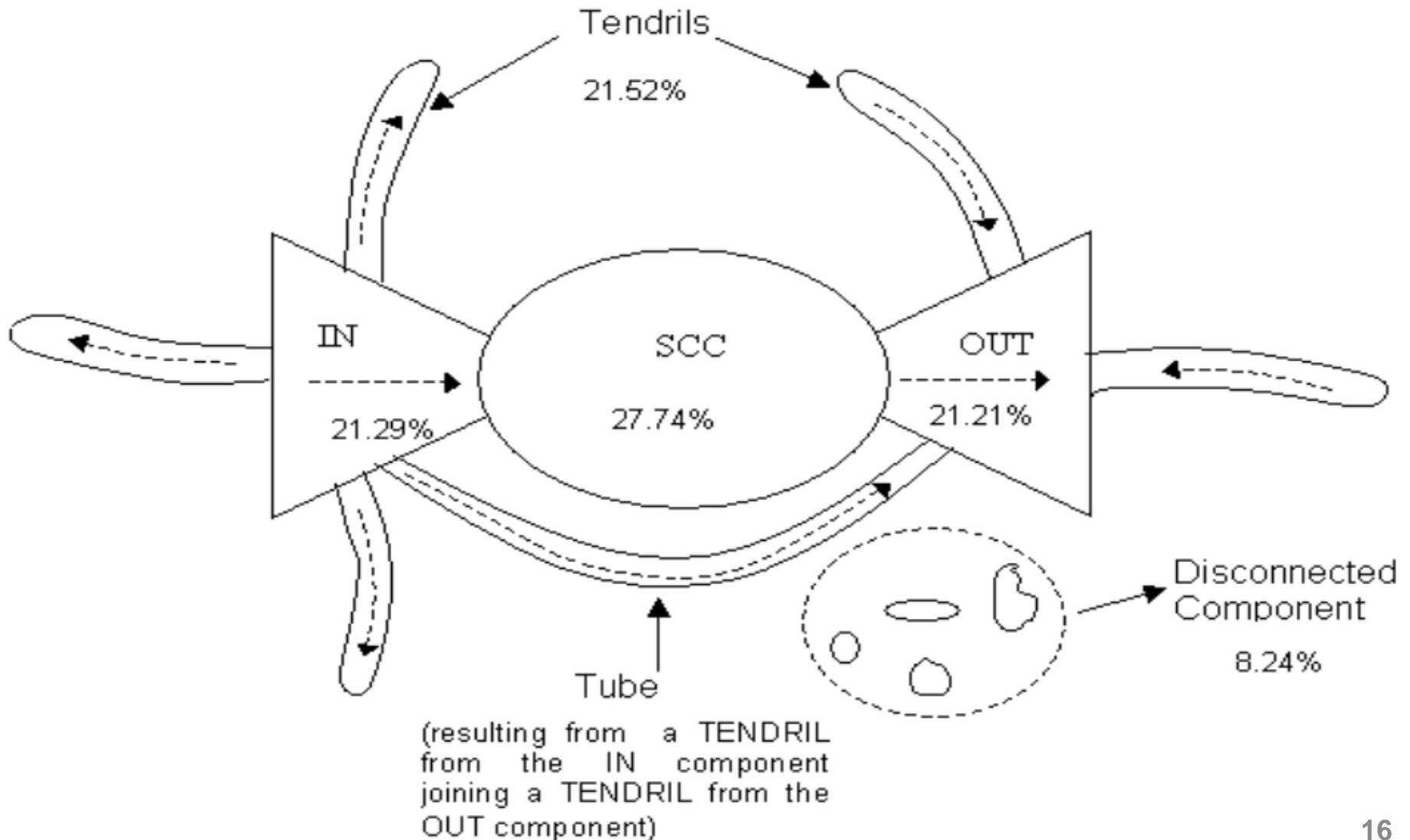
Social Choice



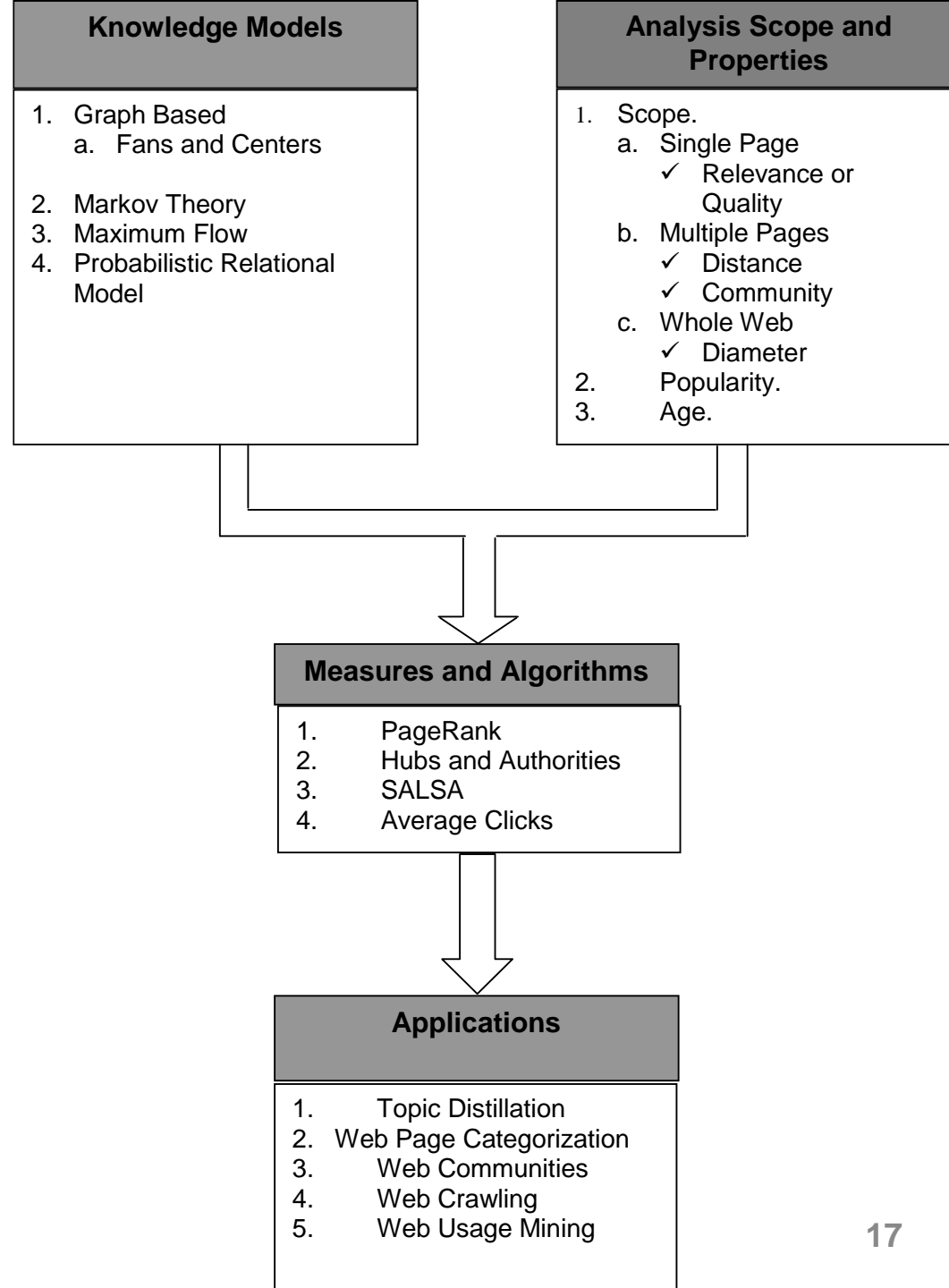
Transitive Endorsement



Mô hình Bow-Tie (2000)



Các kỹ thuật phân tích siêu liên kết



Kỹ thuật phân tích siêu liên kết

- Mô hình tri thức

- Biểu diễn cơ sở giúp hình thành nền tảng để thực hiện những tác vụ đặc thù theo ứng dụng.
- Ví dụ, Graph models, flow models hoặc probabilistic models, v.v.

- Phạm vi phân tích và tính chất

- Đặc tả cho biết tác vụ liên quan đến một trang đơn, một tập hợp các trang hay toàn bộ Web
- Đặc điểm của đối tượng (trang đơn hay nhiều trang) được chọn

- Độ đo và Giải thuật

- Độ đo là các chuẩn về tính chất Web như chất lượng, độ liên quan, hoặc khoảng cách giữa các trang
- Mô hình tri thức + tính chất mong muốn → phá triển độ đo
- Giải thuật được thiết kế để tính toán những độ đo này với hiệu suất tốt.



Phân tích mạng xã hội

Phân tích mạng xã hội

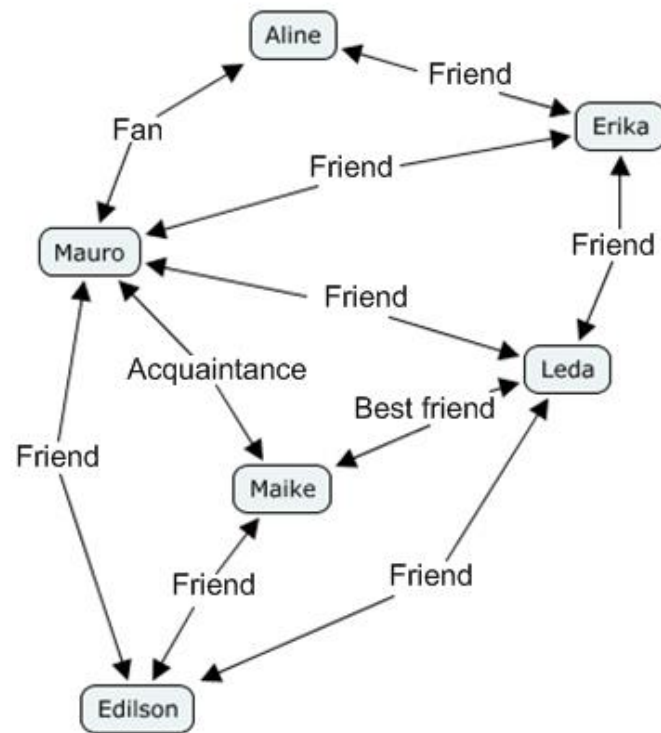
- Nghiên cứu về các đối tượng xã hội (con người trong một tổ chức, gọi là **tác nhân**), và mối quan hệ và tương tác giữa những đối tượng này.

Social
network
analysis



Phân tích mạng xã hội

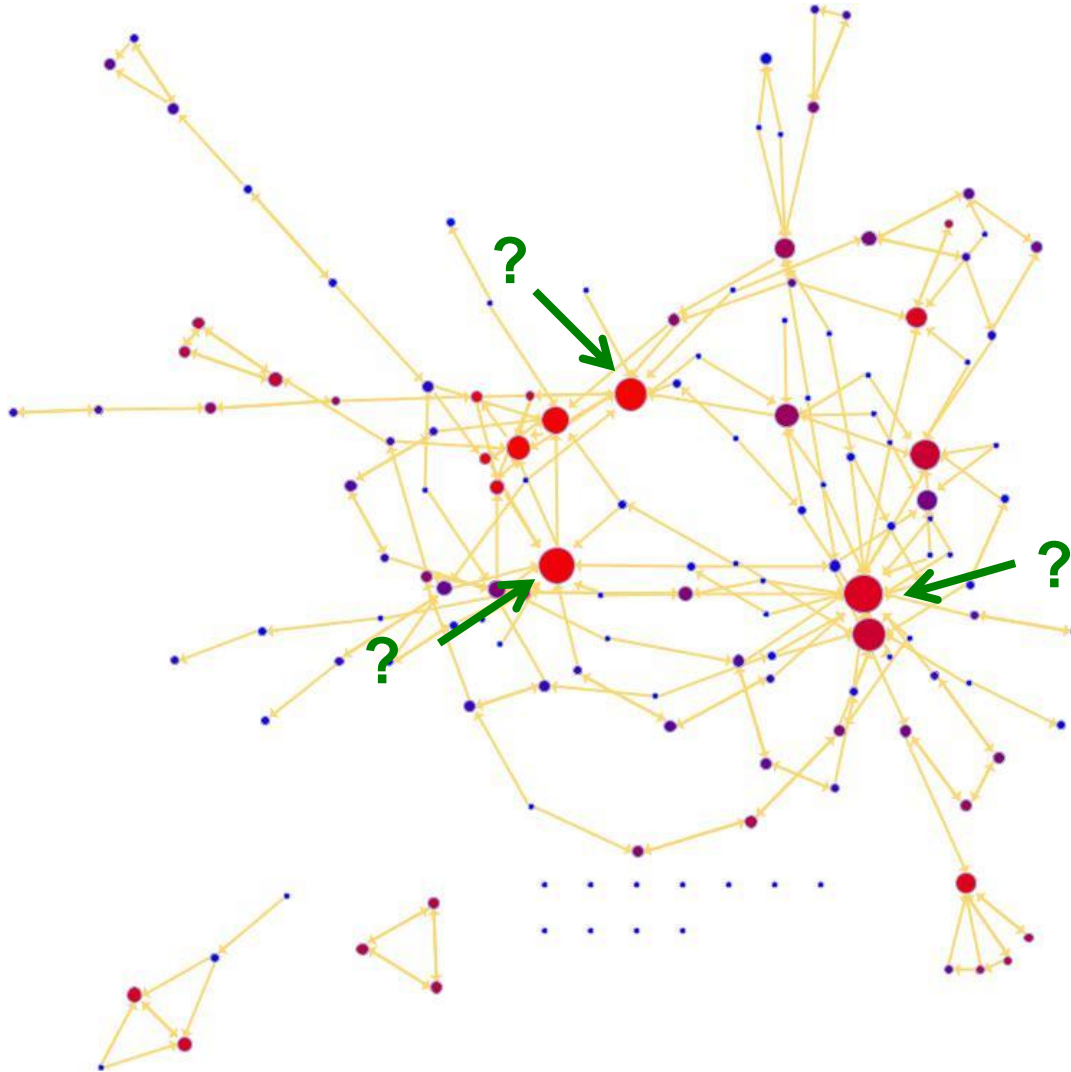
- Các mối quan hệ và tương tác được thể hiện bằng một mạng lưới hay đồ thị trong đó
 - Mỗi **đỉnh** (nút) thể hiện một **tác nhân**, và
 - Mỗi **liên kết** thể hiện một **quan hệ**
- Từ mạng lưới này, ta có thể
 - Nghiên cứu tính chất của cấu trúc mạng, vai trò, vị trí cũng như uy tín của mỗi tác nhân xã hội.
 - Xác định nhiều kiểu đồ thị con khác nhau, ví dụ cộng đồng hình thành bởi một nhóm tác nhân



Phân tích mạng xã hội và Web

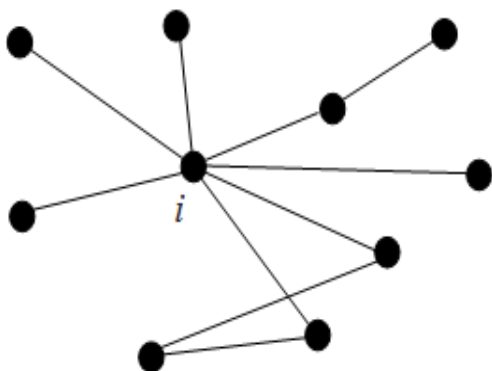
- Web về cơ bản là một mạng xã hội ảo → phân tích mạng xã hội có ích đối với phân tích Web
 - Trang: tác nhân xã hội và siêu liên kết: quan hệ giữa hai tác nhân.
- Các ý tưởng của phân tích mạng xã hội là phương tiện dẫn đến thành công của những cỗ máy tìm kiếm.
 - Ví dụ, khái niệm **centrality** và **prestige** liên hệ mật thiết với bài toán phân tích siêu liên kết và tìm kiếm trên Web

Tính chất của mạng networks: Node nào có tính trung tâm cao nhất?

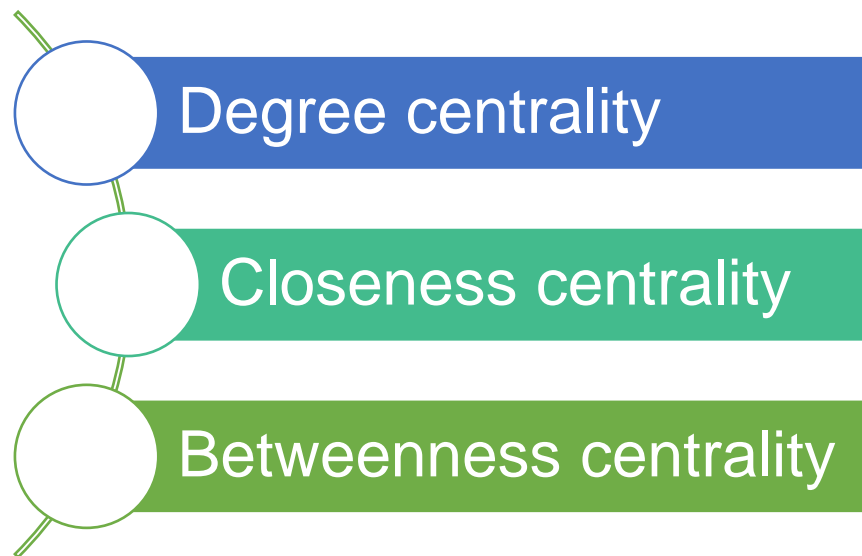


Độ trung tâm (Centrality)

- Tác nhân quan trọng hay nổi bật là đối tượng liên kết hoặc có liên quan rộng rãi với nhiều tác nhân khác.
- Tác nhân trung tâm (central actor) là đối tượng tham gia vào nhiều liên kết (tie).
- Định nghĩa trên cả đồ thị có hướng và vô hướng



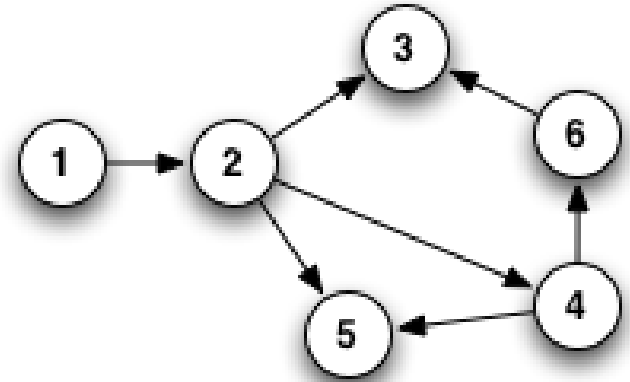
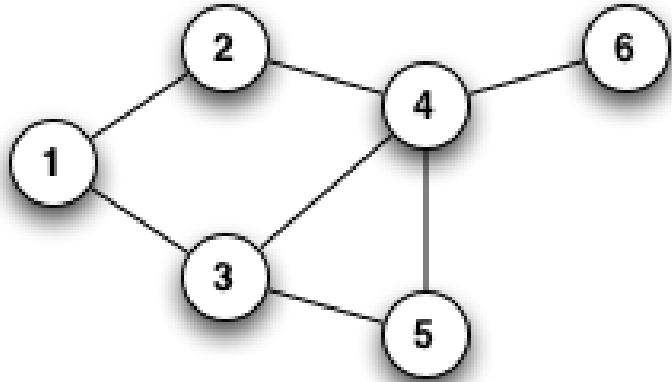
Tác nhân i trung tâm nhất vì có thể giao tiếp với nhiều tác nhân khác nhất.



Degree Centrality $C_D(i)$

- Tác nhân trung tâm là đối tượng năng động nhất, có liên kết đến hầu hết các tác nhân khác.
- Gọi n là số tác nhân trong mạng xã hội đang xét.
- **Đồ thị vô hướng:** $C_D(i) = \frac{d(i)}{n - 1}$
 - Trong đó $d(i)$ là **bậc** của nút i , có giá trị tối đa là $n - 1$.
 - Giá trị của độ đo thay đổi từ 0 đến 1.
- **Đồ thị có hướng:** $C_D(i) = \frac{d_o(i)}{n - 1}$
 - Trong đó $d_o(i)$ là **bậc ngoài** của nút i

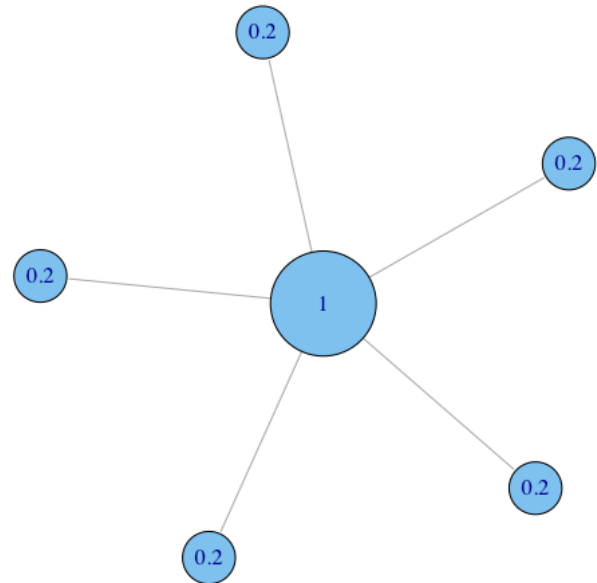
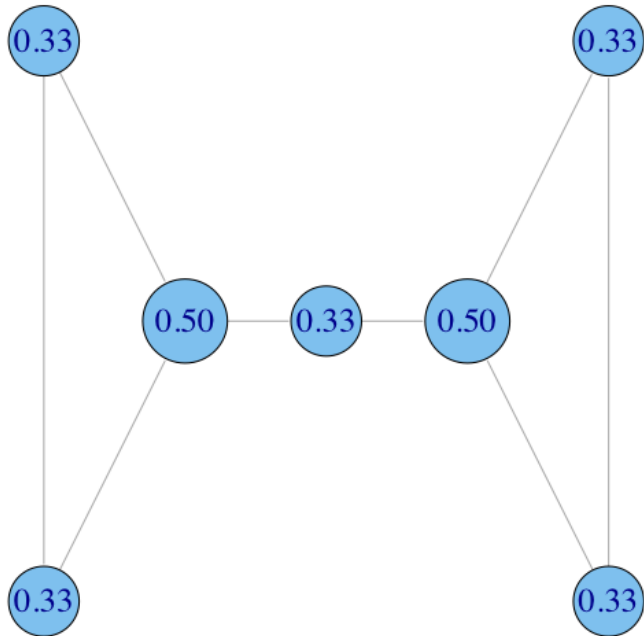
Degree Centrality: Ví dụ



| i | 1 | 2 | 3 | 4 | 5 | 6 |
|----------|-----|-----|-----|-----|-----|-----|
| $C_D(i)$ | 2/5 | 2/5 | 3/5 | 4/5 | 2/5 | 1/5 |

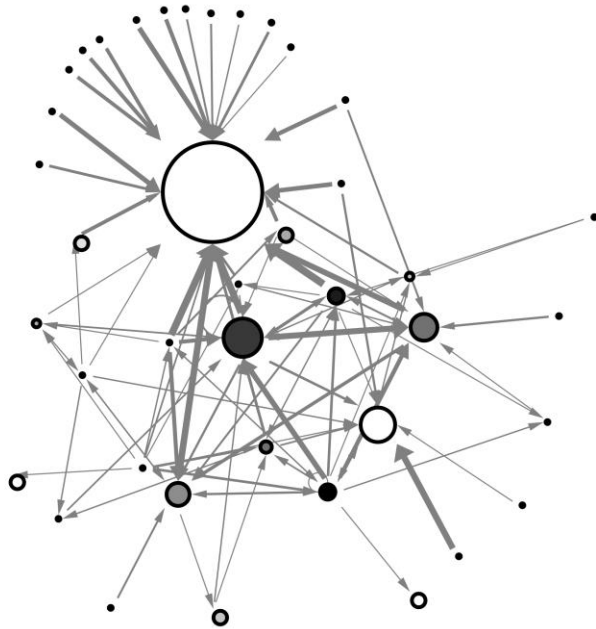
| i | 1 | 2 | 3 | 4 | 5 | 6 |
|----------|-----|-----|-----|-----|-----|-----|
| $C_D(i)$ | 1/5 | 3/5 | 0/5 | 2/5 | 0/5 | 1/5 |

Tính trung tâm bậc (tt.)

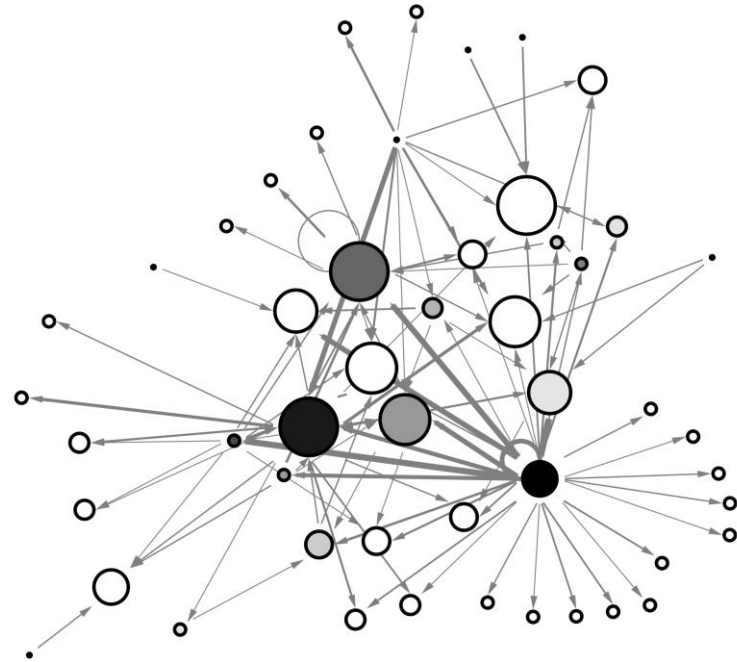


degree centralization examples

example financial trading networks

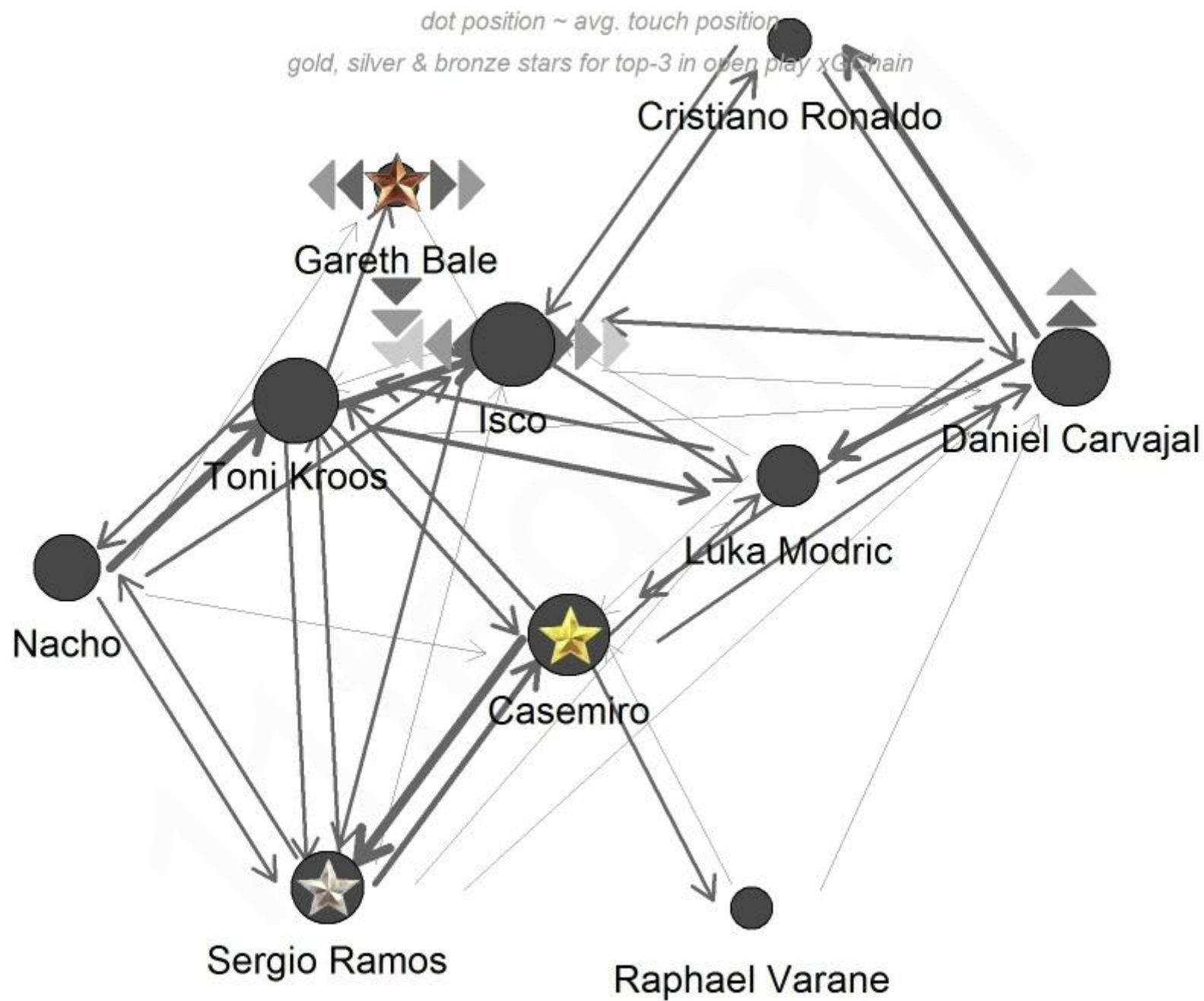


high centralization: one node trading with many others



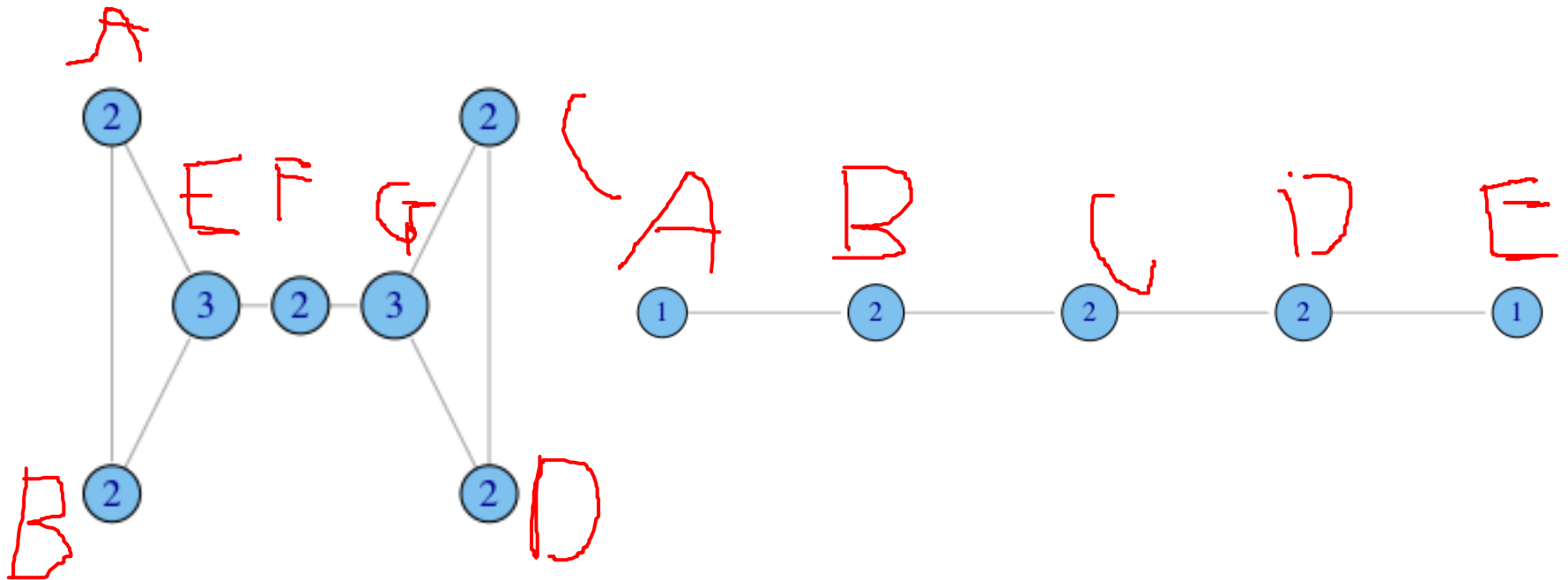
low centralization: trades are more evenly distributed

dot position ~ avg. touch position
gold, silver & bronze stars for top-3 in open play xG chain



Bậc không phải tất cả

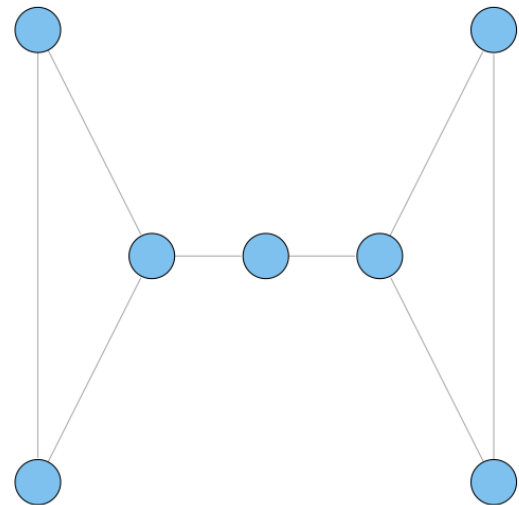
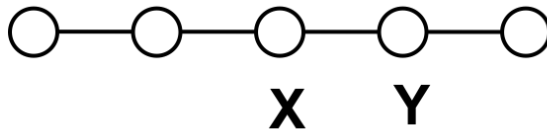
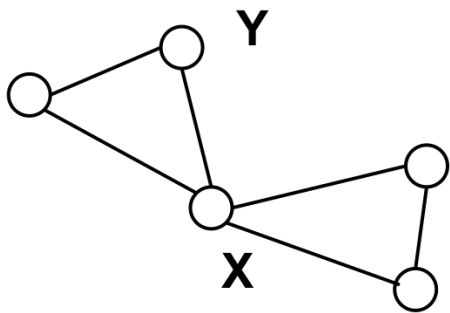
Trong hai đồ thị sau, node nào nên là node trung tâm



- Khả năng trung gian giữa các nhóm
- likelihood thông tin xuất phát từ nơi nào đó trong mạng đến node hiện tại

Tính trung tâm trung gian (Betweenness Centrality)

- **Intuition:** bao nhiêu cặp node mà đường đi ngắn nhất giữa chúng phải đi qua bạn
- Node nào có Betweenness cao hơn, X hay Y



Tính trung tâm trung gian (Betweenness Centrality)

- Nếu hai tác nhân không kề j và k muốn tương tác và tác nhân i nằm trên đường đi giữa j và k , thì i có thể có sự điều khiển trên các tương tác giữa j và k .
- **Độ đo tính trung gian** đo việc điều khiển này của i trên các cặp tác nhân khác.
 - Do đó, nếu i nằm trên các đường đi của nhiều tương tác như vậy, thì i là một tác nhân quan trọng

Betweenness Centrality

- Hai tác nhân không kề nhau, j và k , muốn tương tác với nhau và tác nhân i nằm trên đường đi giữa j và k
- Như vậy, i có thể có quyền điều khiển nào đó lên mỗi tương tác giữa j và k .
- Nếu tác nhân i nằm trên đường đi của nhiều cặp tác nhân thì i là một tác nhân quan trọng.

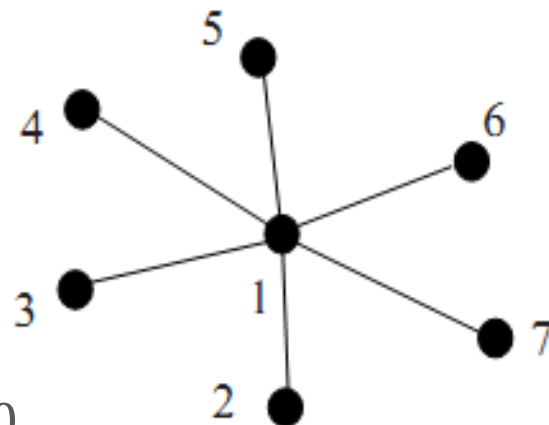
Betweenness Centrality $C_B(i)$

• Đồ thị vô hướng: $C_B(i) = \sum_{j < k} \frac{p_{jk}(i)}{p_{jk}}$

- Trong đó $p_{jk}(i)$ là số đường đi ngắn nhất có qua i ($j \neq i$ and $k \neq i$) và p_{jk} là tổng số đường đi ngắn nhất giữa mọi cặp tác nhân (khác i).
- $C_B(i) = 0$: i không nằm trên đường đi ngắn nhất nào.
- $C_B(i) = (n - 1)(n - 2)/2$: i nằm trên đường đi của mọi cặp tác nhân (không bao gồm i)

$$C_B(i) = 15$$

$$C_B(2) = C_B(3) = C_B(4) = C_B(5) = C_B(6) = C_B(7) = 0$$



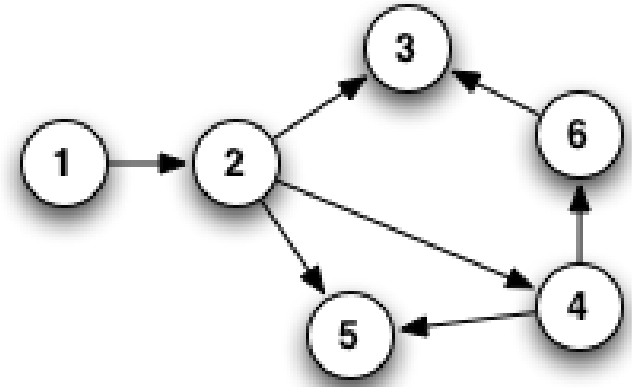
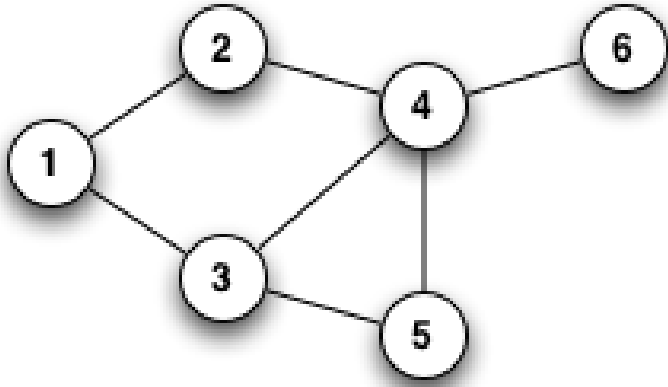
Betweenness Centrality $C_B(i)$

- $C_B(i)$ đôi khi được chuẩn hóa với $(n-1)(n-2)/2$ để đảm bảo miền giá trị thay đổi từ 0 đến 1.

$$C_B(i) = \frac{2 \sum_{j < k} \frac{p_{jk}(i)}{p_{jk}}}{(n-1)(n-2)}$$

- Có thể áp dụng cho đồ thị không liên thông.
- **Đồ thị có hướng:** Sử dụng cùng công thức, nhân cho 2
 - $(n-1)(n-2)$ cặp: đường đi từ j tới $k \neq$ đường đi từ k tới j .
 - p_{jk} phải xét đường đi từ cả hai hướng.

Betweenness Centrality: Ví dụ



| i | 1 | 2 | 3 | 4 | 5 | 6 |
|----------|-----|---|---|------|---|---|
| $C_B(i)$ | 1/2 | 1 | 2 | 11/2 | 0 | 0 |

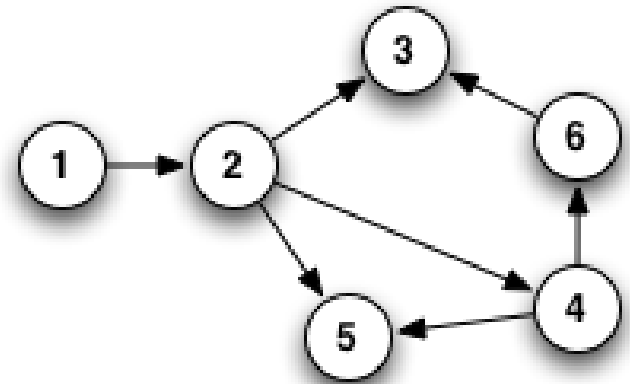
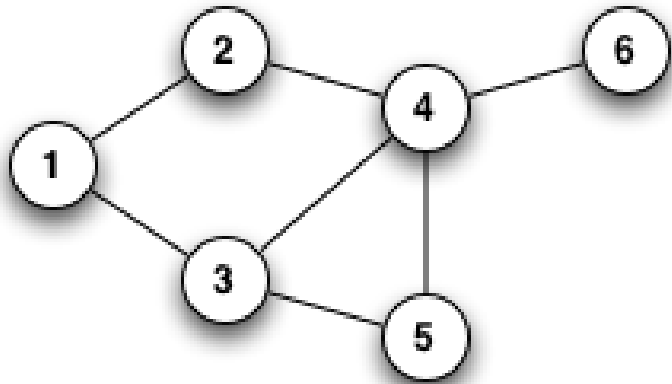
Ví dụ $C_B(1) = 1/2$ $C_B(2) = \frac{1}{2} + \frac{1}{2}$ $C_B(3) = \frac{1}{2} + \frac{1}{1} + \frac{1}{2} = 2$

| i | 1 | 2 | 3 | 4 | 5 | 6 |
|----------|---|---|---|---|---|---|
| $C_B(i)$ | 0 | 4 | 0 | 2 | 0 | 1 |

Closeness Centrality $C_c(i)$

- Tác nhân i là trung tâm nếu đối tượng này dễ dàng tương tác với mọi tác nhân khác.
 - Tức là tác nhân i có khoảng cách đến các tác nhân khác ngắn.
- Gọi $d(i, j)$ là khoảng cách ngắn nhất từ i đến j (đo bằng số liên kết có trên đường đi ngắn nhất)
- **Đồ thị vô hướng:**
$$C_c(i) = \frac{n - 1}{\sum_{j=1}^n d(i, j)}$$
 - Giá trị độ đo thay đổi từ 0 đến 1, chỉ áp dụng cho đồ thị liên thông
- **Đồ thị có hướng:** sử dụng cùng công thức
 - Xét đến hướng của liên kết (cạnh) khi tính khoảng cách

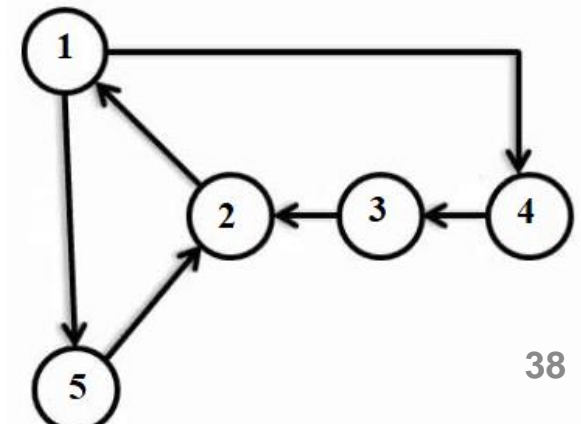
Closeness Centrality: Ví dụ



| i | 1 | 2 | 3 | 4 | 5 | 6 |
|----------|-----|-----|-----|-----|-----|------|
| $C_C(i)$ | 5/9 | 5/8 | 5/7 | 5/6 | 5/8 | 5/10 |

Không tính được vì đồ thị không liên thông mạnh

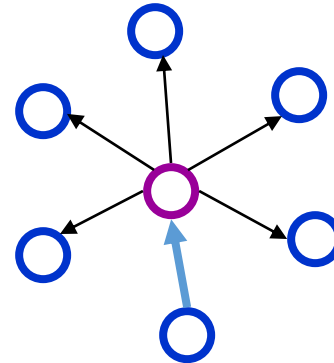
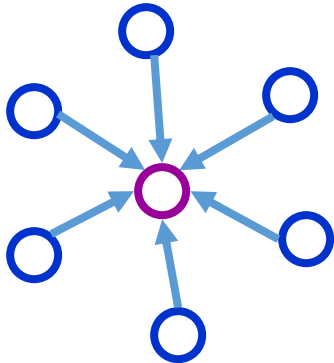
| i | 1 | 2 | 3 | 4 | 5 |
|----------|-----|-----|-----|------|------|
| $C_C(i)$ | 4/6 | 4/8 | 4/9 | 4/10 | 4/10 |



Prestige

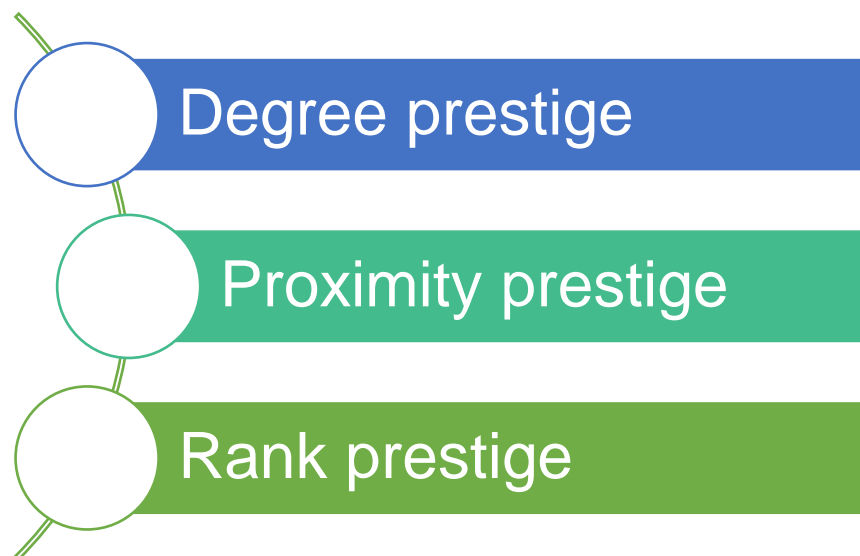
- Uy tín:

- Một bài báo được cited bởi nhiều bài báo khác thì có uy tín cao.
- Một người được đề cử bởi nhiều người khác thì có uy tín cao.



Độ uy tín (Prestige)

- Tác nhân có uy tín được định nghĩa là đối tượng nhận được liên kết rộng rãi từ nhiều tác nhân khác.
- Đo lường sự nổi bật của tác nhân một cách tinh tế hơn độ trung tâm
 - Độ trung tâm tập trung vào **liên kết ngoài** (out-links), trong khi độ uy tín tập trung vào liên kết trong (in-links).
- Chỉ xét đồ thị có hướng



Degree Prestige $P_D(i)$

- Một tác nhân được gọi là có uy tín nếu đối tượng này nhận được nhiều đề cử.
- Gọi n là số tác nhân trong mạng xã hội đang xét.
- Một cách đơn giản nhất, độ uy tín của tác nhân i được định nghĩa là

$$P_D(i) = \frac{d_I(i)}{n - 1}$$

- Trong đó $d_I(i)$ là **bậc trong** của i .
- Giá trị của độ đo thay đổi từ 0 đến 1.

Proximity Prestige

- Xét cả tác nhân liên kết trực tiếp và gián tiếp đến tác nhân i .
 - Proximity = độ gần hay độ xa (khoảng cách) đến tác nhân i .
- Gọi I_i là tập hợp tác nhân có thể đi đến tác nhân i , còn gọi là **miền ảnh hưởng** (influence domain) của i .
- Gọi $d(j, i)$ là khoảng cách ngắn nhất từ tác nhân j đến i .
- Độ đo này sử dụng khoảng cách trung bình

$$P_p(i) = \frac{\sum_{j \in I_i} d(j, i)}{|I_i|}$$

- Trong đó $|I_i|$ là kích thước của tập hợp I_i

Proximity Prestige

- Giá trị độ đo có thể được chuẩn hóa với tỉ lệ số tác nhân có thể đến i .

$$P_p(i) = \frac{|I_i|/(n-1)}{\sum_{j \in I_i} d(j, i) / |I_i|}$$

- $P_p(i) = 1$: mọi tác nhân đều có thể đến tác nhân i ($\frac{|I_i|}{n-1} = 1$) và mọi tác nhân đều liền kề i (mẫu số tiến đến 1).
- $P_p(i) = 0$: không tác nhân nào có thể đến i ($|I_i| = 0$)

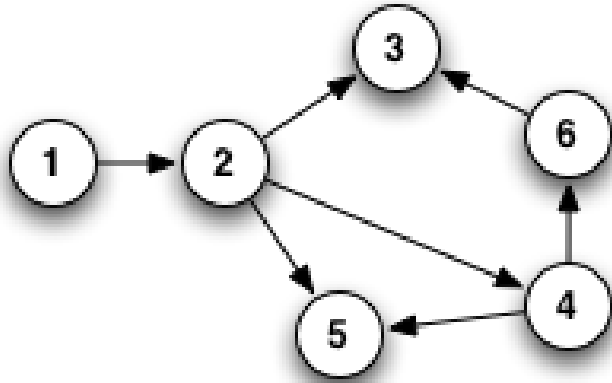
Rank Prestige

- Xét **mức độ nổi bật** của những cá nhân tham gia cuộc “bầu cử” hay “bình chọn”.
- Tác nhân i được chọn bởi một người quan trọng sẽ có nhiều thanh thế hơn là được chọn bởi một người ít quan trọng.
 - Ví dụ, bình chọn của người CEO trong công ty có giá trị hơn bình chọn của một người công nhân bình thường.
- Uy tín của một người chịu ảnh hưởng bởi thứ hạng hay tình trạng của những tác nhân liên quan.
 - Nếu vòng tròn ảnh hưởng của một người toàn những tác nhân uy tín thì tự nhiên người này cũng sẽ trở nên uy tín.

Rank Prestige $P_R(i)$

- Được định nghĩa là tổ hợp tuyến tính các liên kết trở đến i
$$P_R(i) = A_{1i}P_R(1) + A_{2i}P_R(2) + \dots + A_{ni}P_R(n)$$
 - Trong đó $A_{ji} = 1$ nếu j trở đến i , và ngược lại $A_{ji} = 0$.
- $P = (P_R(1), P_R(2), \dots, P_R(n))^T$: vector cột chứa mọi giá trị rank prestige của n
- $P = A^T P \rightarrow P$ là eigenvector của A^T .
- Liên hệ trực tiếp đến các giải thuật nổi tiếng nhất trong Tìm kiếm Web, đó là PageRank và HITS

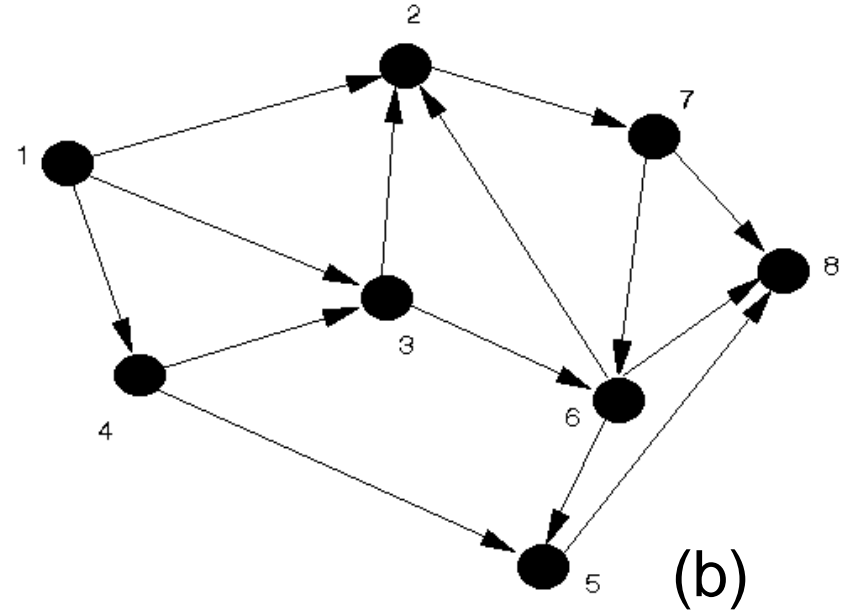
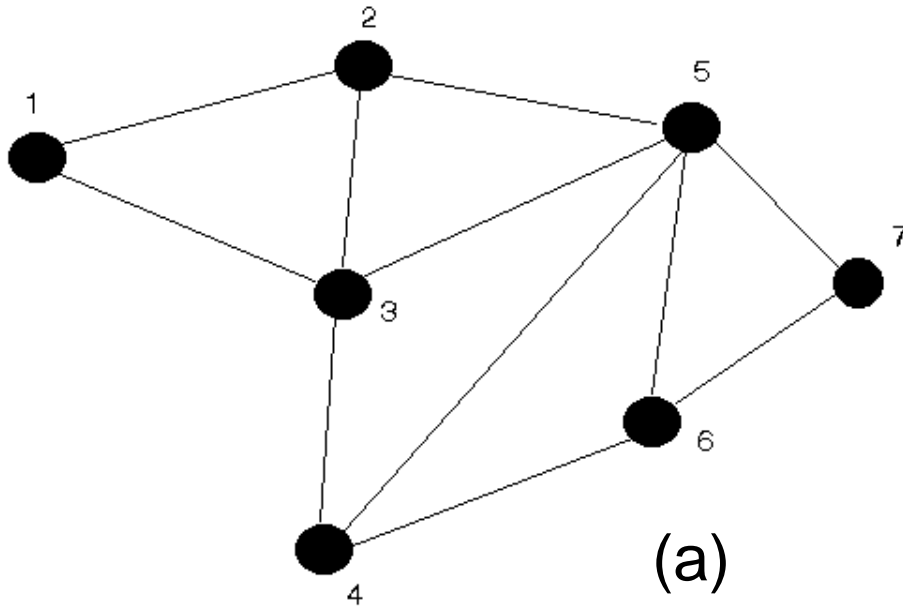
Prestige: Ví dụ



Assume that initial values of P_P for all nodes are 0.1

| i | 1 | 2 | 3 | 4 | 5 | 6 |
|----------|-----|-----|-----|-----|-----|-----|
| $P_D(i)$ | 0/5 | 1/5 | 2/5 | 1/5 | 2/5 | 1/5 |
| $P_P(i)$ | 0/0 | 1/1 | 6/4 | 3/2 | 4/3 | 6/3 |
| $P_R(i)$ | 0 | 0.1 | 0.2 | 0.1 | 0.2 | 0.1 |

Bài tập 1: Centrality và Prestige



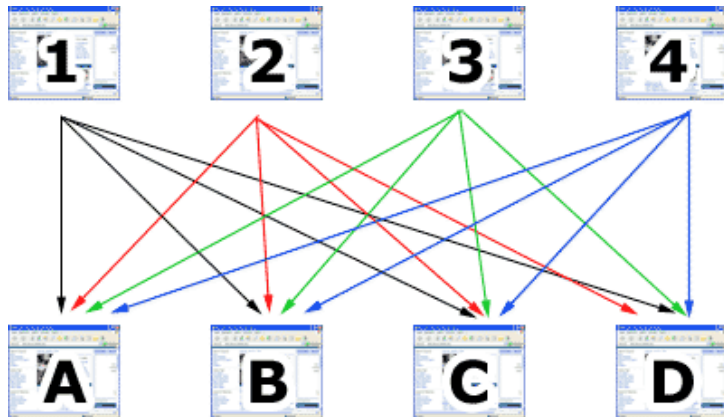
- Xác định giá trị degree centrality, closeness centrality và betweenness centrality cho cả hai đồ thị (a) and (b)
- Xác định giá trị degree prestige, proximity prestige và rank prestige (giá trị khởi đầu 0.1) cho đồ thị (b).



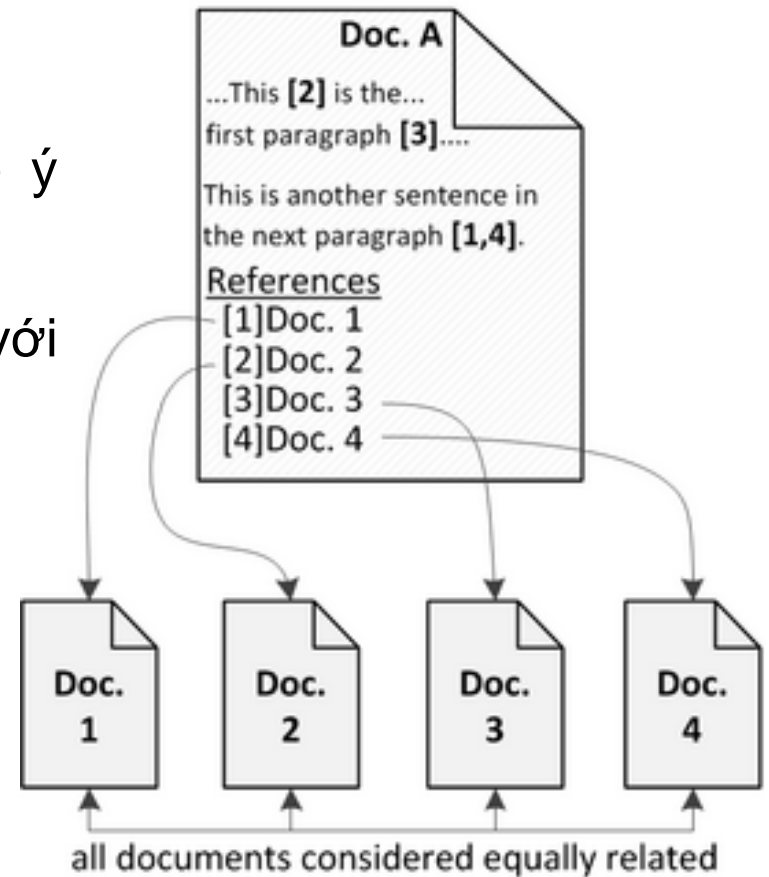
Phân tích trích dẫn

Trích dẫn (Citation)

- Một công bố học thuật thường trích dẫn những công trình trước đó.
 - Thừa nhận nguồn gốc của một số ý tưởng trong bài công bố hiện hành
 - So sánh ý tưởng được đề xuất với công trình đã có trước đó

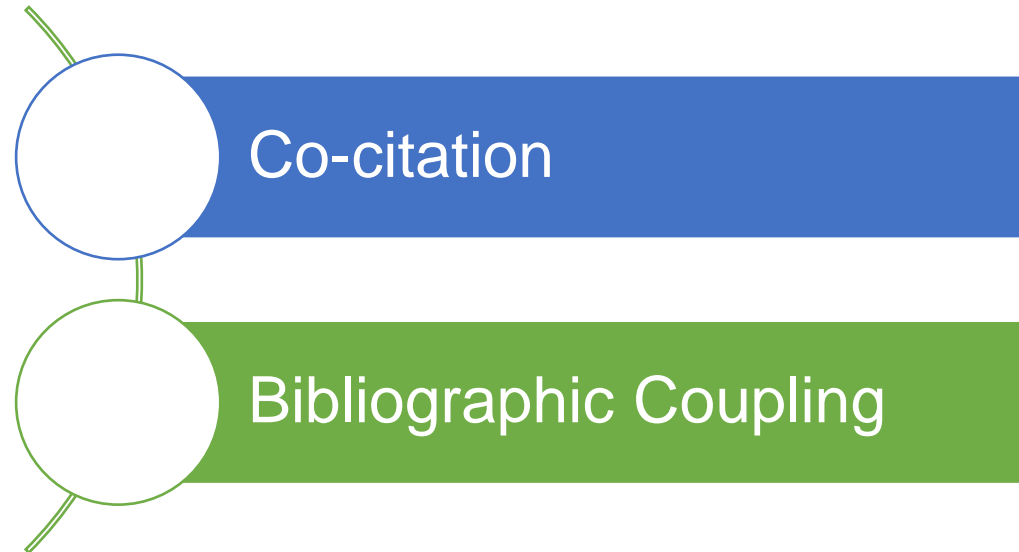


Traditional Co-Citation



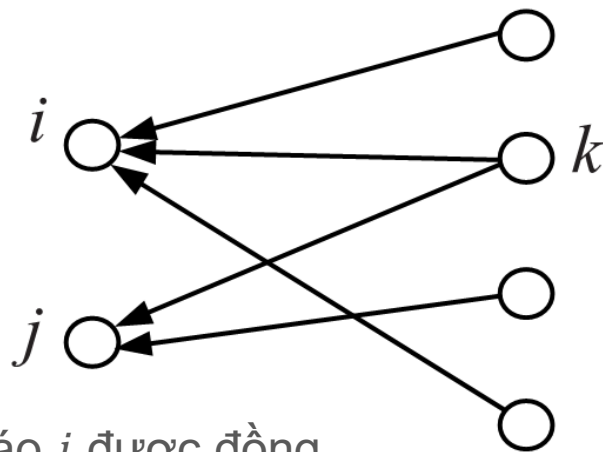
Phân tích trích dẫn

- Thuộc lĩnh vực nghiên cứu thư mục ([bibliometric research](#))
- Nghiên cứu các trích dẫn để hình thành mối quan hệ giữa tác giả và công trình tương ứng của họ

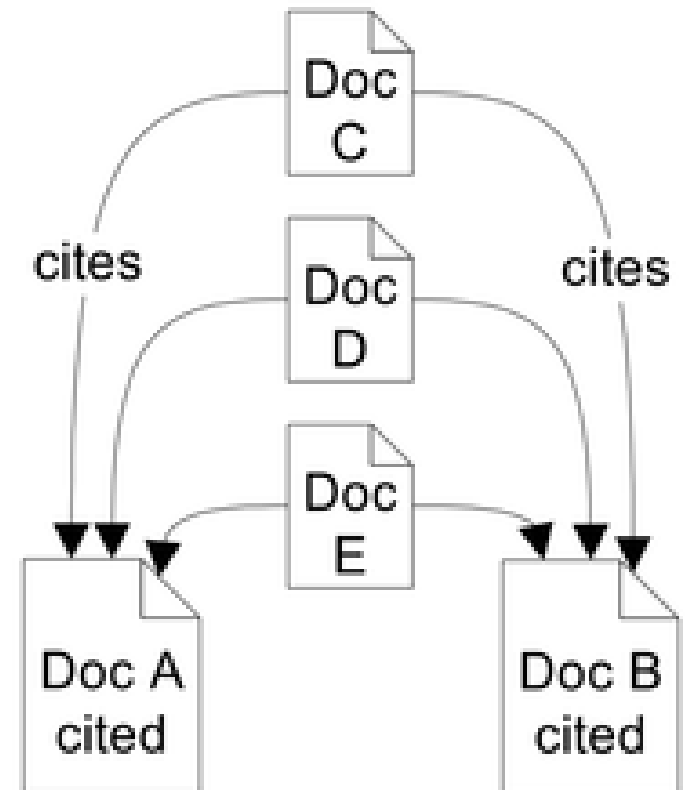


Co-citation

- Nếu hai bài báo, *i* và *j*, cùng được trích dẫn bởi bài báo *k*, chúng được xem là liên hệ với nhau theo một cách nào đó, mặc dù chúng không trích dẫn trực tiếp lẫn nhau.
- Mỗi quan hệ giữa *i* và *j* càng mạnh khi càng có nhiều bài báo đồng thời trích dẫn đến chúng



Bài báo *i* và bài báo *j* được đồng trích dẫn bởi bài báo *k* $\rightarrow C_{ij} = 1$



Co-citation

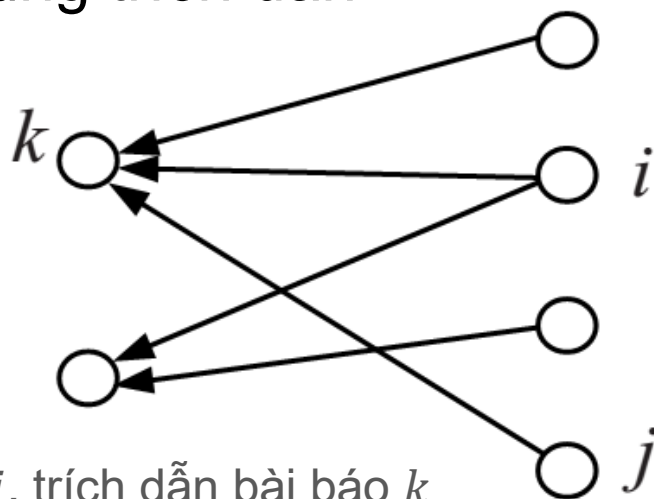
- Xét tập hợp gồm có n bài báo.
- Gọi L là ma trận trích dẫn.
 - $L_{ij} = 1$ nếu bài báo i trích dẫn bài báo j , và ngược lại, $L_{ij} = 0$
- **Co-citation** C_{ij} là thước đo độ tương tự, được định nghĩa là số bài báo đồng trích dẫn cả i và j .

$$C_{ij} = \sum_{k=1}^n L_{ki} L_{kj}$$

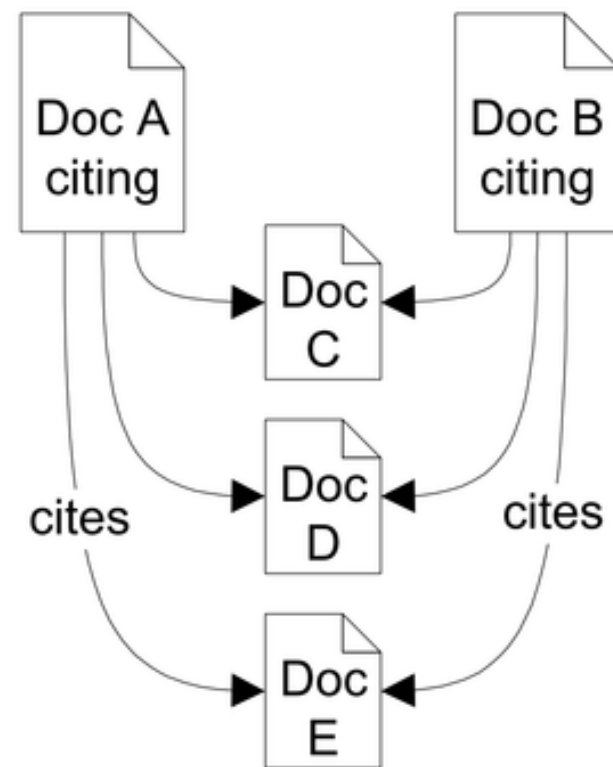
- C là ma trận vuông và đối xứng.
- Thường được dùng để đánh giá độ tương tự giữa hai bài báo \rightarrow có thể áp dụng để gom nhóm bài báo có cùng chủ đề

Bibliographic coupling

- Nếu hai bài báo, i và j , cùng trích dẫn bài báo k , chúng được xem là liên hệ với nhau theo một cách nào đó, mặc dù chúng không trích dẫn trực tiếp lẫn nhau.
- Mỗi quan hệ giữa i và j càng mạnh khi có càng nhiều bài báo được chúng cùng trích dẫn



Cả hai bài báo, i và j , trích dẫn bài báo k
 $\rightarrow B_{ij} = 1$



Bibliographic coupling

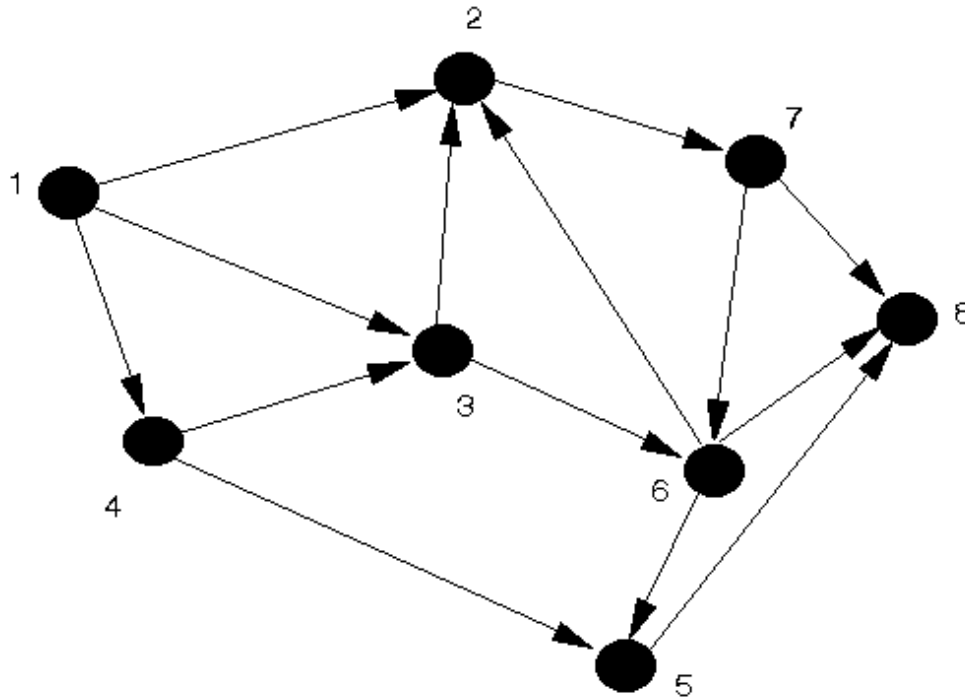
- **Bibliographic coupling** B_{ij} là thước đo độ tương tự, được định nghĩa là số bài báo được trích dẫn bởi cả i và j .

$$B_{ij} = \sum_{k=1}^n L_{ik} L_{jk}$$

- Ma trận B cũng vuông và đối xứng.

Bài tập 2: Phân tích trích dẫn

- Xây dựng ma trận co-citation và ma trận bibliographic coupling cho đồ thị bên dưới.



Tài liệu tham khảo



- Bing Liu. 2007. *Web Data Mining-Exploring Hyperlinks, Contents, and Usage Data*. Springer Series on Data-Centric Systems and Applications. **Chapter 7**.