

Khai Thác Dữ Liệu Đồ Thị

DỰ ĐOÁN LIÊN KẾT

Giảng viên: Lê Ngọc Thành

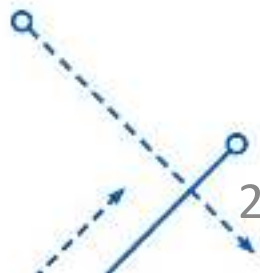
Email: lnthanh@fit.hcmus.edu.vn



fit@hcmus

Nội dung

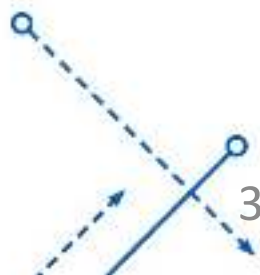
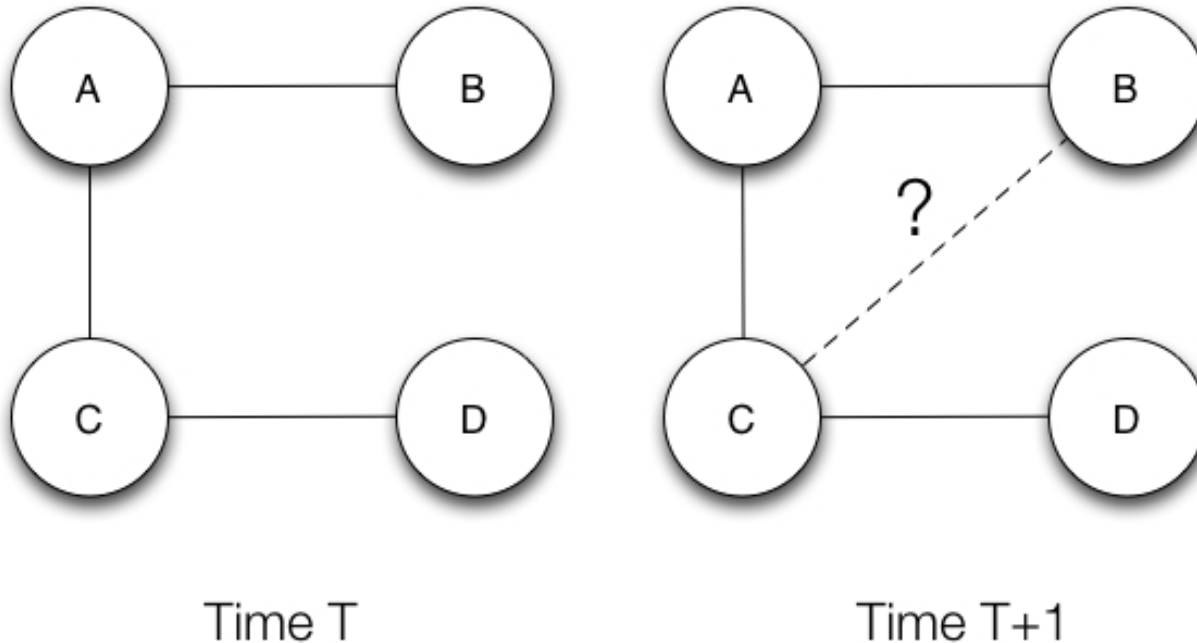
- Dự đoán liên kết
- Học trình dự đoán liên kết
 - Học không giám sát
 - Học có giám sát



Dự đoán liên kết

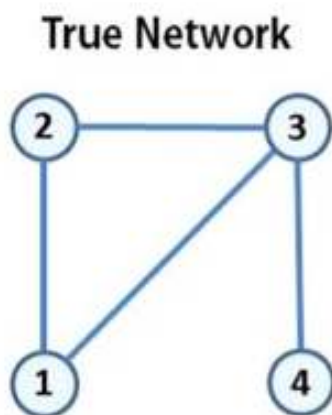
- Bài toán dự đoán liên kết:

- Cho trước các liên kết trong đồ thị ở thời điểm t , dự đoán các cạnh sẽ được thêm vào đồ thị trong khoảng thời gian từ t đến thời điểm t' trong tương lai.

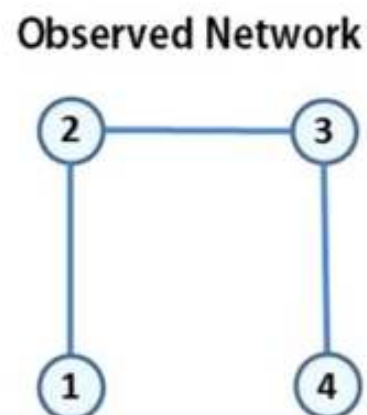


Suy luận liên kết bị thiếu

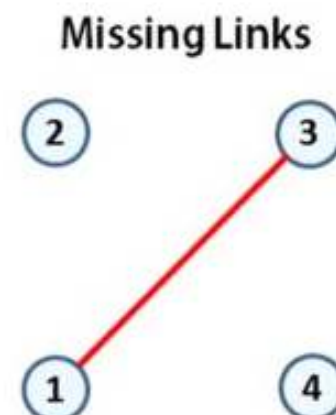
- Dự đoán liên kết khác với suy luận liên kết bị thiếu (ẩn) ở chỗ:
 - Dự đoán liên kết để tìm các liên kết xuất hiện theo thời gian
 - Suy luận liên kết bị thiếu để tìm ra các liên kết thêm trong đồ thị tĩnh.



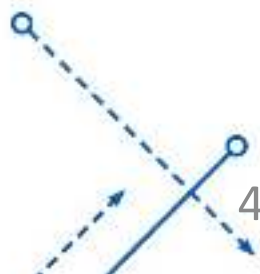
Network



Training Set



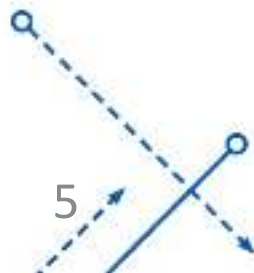
Probe Set



Suy luận liên kết bị thiếu

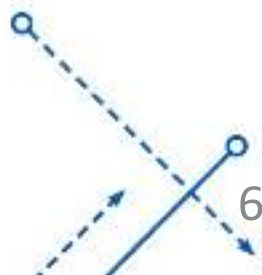
Ứng dụng

- Gợi ý kết bạn trên mạng xã hội
- Dự đoán kết nối giữa các thành viên trong tổ chức khủng bố
- Đề xuất hợp tác giữa các nhà nghiên cứu
- Đề xuất bài báo nên được tham chiếu đến
- Dự đoán sản phẩm được mua
- Dự đoán tương tác của các protein

The Facebook logo, consisting of the word "facebook" in white lowercase letters on a blue rectangular background.The Tinder logo, featuring a red flame icon above the word "tinder" in a red, lowercase, sans-serif font.The Amazon logo, with the word "amazon" in black lowercase letters and a curved orange arrow underneath it.

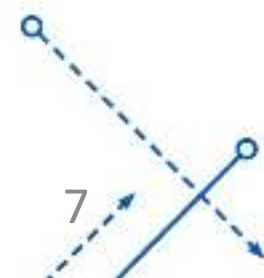
Dự đoán liên kết

- Quá trình dự đoán liên kết là đi tính điểm số $score(u, v)$ cho mỗi cặp $\langle u, v \rangle$, sau đó xếp hạng để chọn ra v có độ đo cao nhất để nối với u
- Thường được chia làm ba nhóm:
 - Phương pháp ước lượng dựa trên đỉnh láng giềng: số láng giềng chung, hệ số Jaccard, Adamic-Adar, đường đi, ...
 - Phương pháp dựa trên tất cả đường đi: PageRank, SimRank
 - Từ phương pháp khác: bigram, gộp nhóm

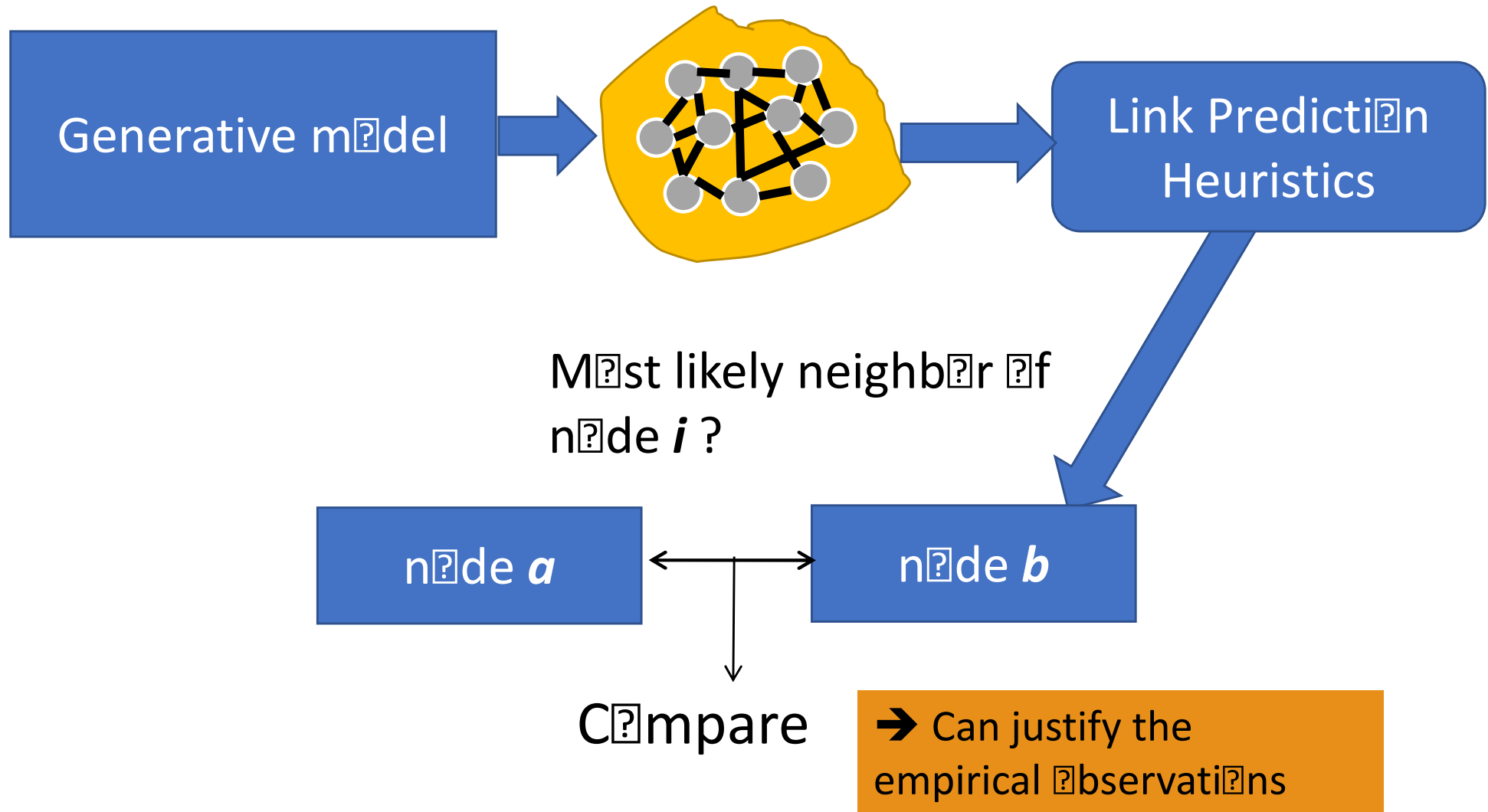


Mô hình chung

- Bước 1: tính toán khoảng cách đỉnh dựa trên một số phương pháp đo như Jaccard, đường đi ngắn nhất, ...
- Bước 2: chọn một số lượng các cặp đỉnh có khoảng cách gần nhất
- Bước 3: dự đoán cạnh mới từ các cặp đã chọn
- Bước 4: đánh giá đồ thị được dự đoán với đồ thị gốc



Mô hình chung



Dự đoán liên kết

graph distance	(negated) length of shortest path between x and y
common neighbors	$ \Gamma(x) \cap \Gamma(y) $
Jaccard's coefficient	$\frac{ \Gamma(x) \cap \Gamma(y) }{ \Gamma(x) \cup \Gamma(y) }$
Adamic/Adar	$\sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log \Gamma(z) }$
preferential attachment	$ \Gamma(x) \cdot \Gamma(y) $
Katz $_{\beta}$	$\sum_{\ell=1}^{\infty} \beta^{\ell} \cdot \text{paths}_{x,y}^{(\ell)} $

where $\text{paths}_{x,y}^{(\ell)} := \{\text{paths of length exactly } \ell \text{ from } x \text{ to } y\}$

weighted: $\text{paths}_{x,y}^{(1)} :=$ number of collaborations between x, y .

unweighted: $\text{paths}_{x,y}^{(1)} := 1$ iff x and y collaborate.

hitting time	$-H_{x,y}$
stationary-normed	$-H_{x,y} \cdot \pi_y$
commute time	$-(H_{x,y} + H_{y,x})$
stationary-normed	$-(H_{x,y} \cdot \pi_y + H_{y,x} \cdot \pi_x)$

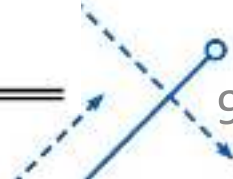
where $H_{x,y} :=$ expected time for random walk from x to reach y

$\pi_y :=$ stationary distribution weight of y

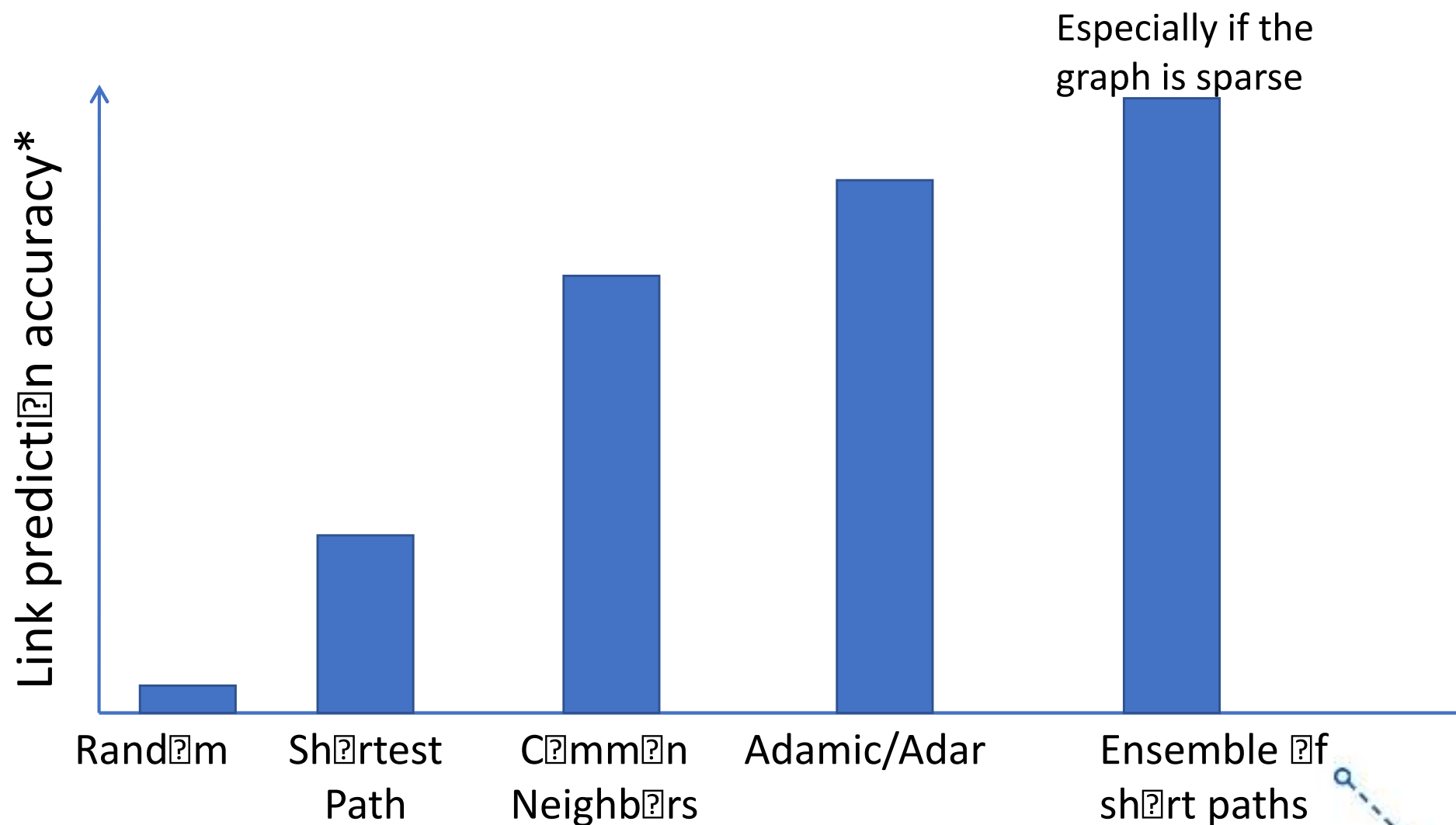
(proportion of time the random walk is at node y)

rooted PageRank $_{\alpha}$	stationary distribution weight of y under the following random walk: with probability α , jump to x . with probability $1 - \alpha$, go to random neighbor of current node.
-----------------------------	---

SimRank $_{\gamma}$	$\begin{cases} 1 & \text{if } x = y \\ \gamma \cdot \frac{\sum_{a \in \Gamma(x)} \sum_{b \in \Gamma(y)} \text{score}(a,b)}{ \Gamma(x) \cdot \Gamma(y) } & \text{otherwise} \end{cases}$
---------------------	---



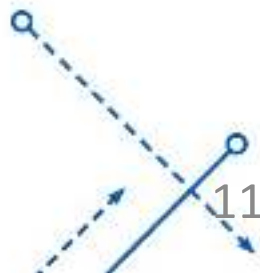
So sánh giữa các phương pháp



*Liben-Nowell & Kleinberg, 2003; Brand, 2005; Sarkar & Moore, 2007

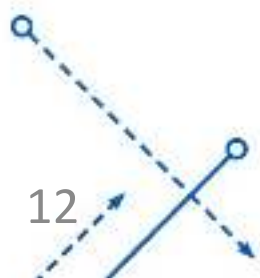
Nội dung

- Dự đoán liên kết
- **Học trong dự đoán liên kết**
 - Học không giám sát
 - Học có giám sát



Học trong dự đoán liên kết

- Học trong dự đoán liên kết được chia làm hai loại:
 - Có giám sát (supervised): cung cấp tập các đỉnh để huấn luyện mô hình
 - Không giám sát (unsupervised): không cần tập huấn luyện
 - Dựa trên độ tương tự (similarity-based): các đỉnh có độ tương tự được kết nối với nhau
 - Dựa trên gộp nhóm (cluster-based): các đỉnh từ cùng một nhóm chứng tỏ các mẫu kết nối tương tự.



Dự đoán liên kết dựa trên độ tương tự

- **Độ tương tự** (học không giám sát) nhằm để khám phá cách giữa các đỉnh trong đồ thị.
- **Phân loại**:
 - **1-step**: các đỉnh lân cận được kết nối
 - **Laplacian**: các đỉnh không tương tự được kết nối
 - **Degree**: các đỉnh có bậc tương tự nhau
 - **A^2** : các đỉnh chia sẻ cùng láng giềng
 - **Closeness**: các đỉnh có trung tâm gần giống nhau
 - **Betweenness**: các đỉnh có trung tâm trung gian giống nhau

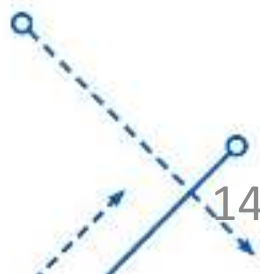


Dựa trên đỉnh láng giềng

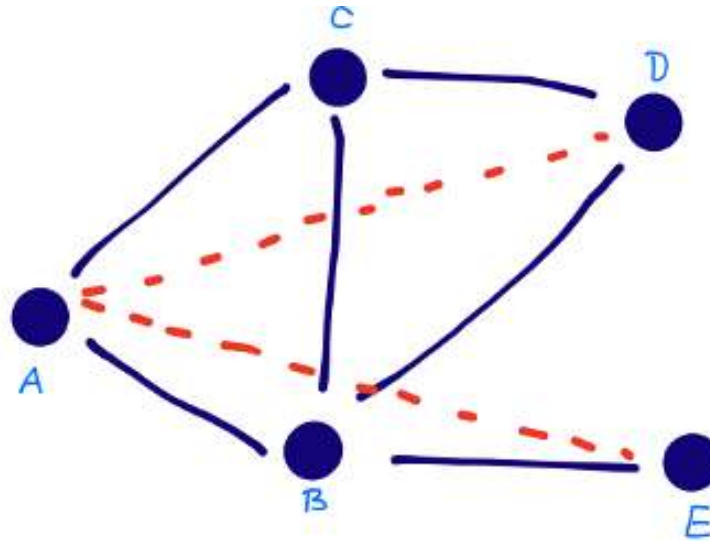
- Láng giềng chung (common neighbors) được đề xuất bởi Ahmad Sadrei là một độ đo đơn giản nhất.
 - Được xem là **hiệu ứng của đóng tam giác** (closing a triangle)
- Đầu tiên là tìm giao giữa hai láng giềng của hai đỉnh, độ đo tương đồng giữa hai đỉnh chính là **số phần tử trong tập này**.

$$score(u, v) = |N(u) \cap N(v)|$$

- Nếu độ đo này lớn hơn ngưỡng cho trước thì liên kết được tạo ra



Dựa trên đỉnh láng giềng



$$|\Gamma(A) \cap \Gamma(D)|$$

↓ ↓

$\{B, C\}$ $\{B, C\}$

$S = 2$ ✓

$$|\Gamma(A) \cap \Gamma(E)|$$

↓ ↓

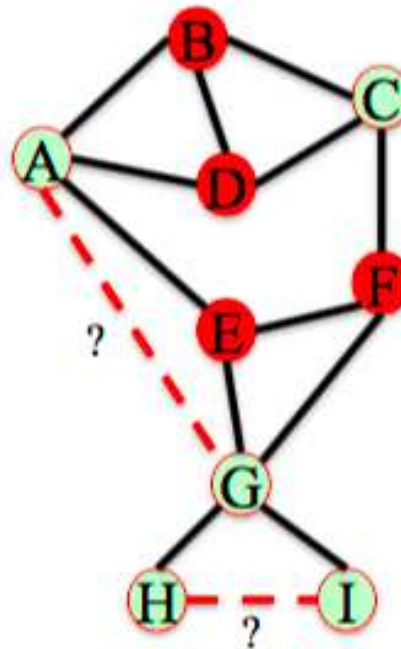
$\{B, C\}$ $\{B\}$

$S = 1$

Hệ số Jaccard

- Hệ số Jaccard chuẩn hóa số láng giềng chung bằng tổng số láng giềng.

$$\text{score}(u, v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$$



$$\text{jacc_coeff}(A, C) = \frac{|\{B, D\}|}{|\{B, D, E, F\}|} = \frac{2}{4} = \frac{1}{2}$$

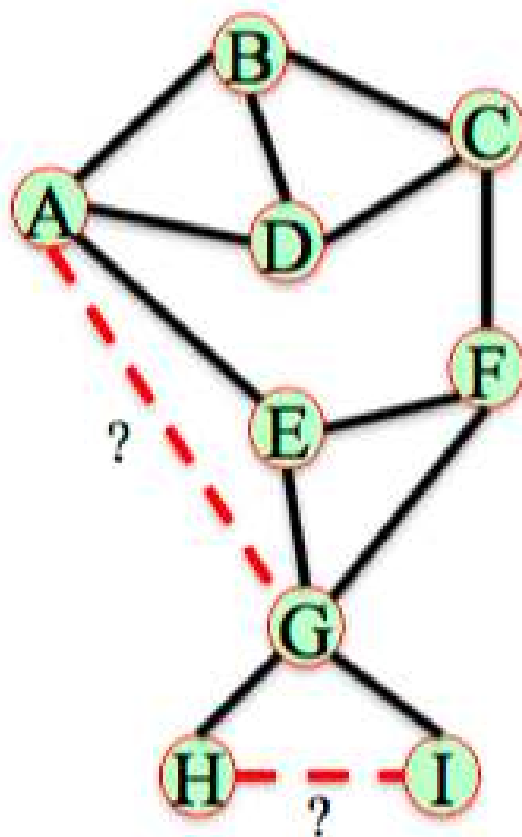
Độ đo Adamic Adar

- Độ đo được đề xuất bởi Lada Adamic và Eytan Adar (2003)
- Ngoài đếm số láng giềng chung, độ đo Adamic Adar còn tính **tổng các log nghịch đảo của bậc các láng giềng**.
 - Hiệu ứng đóng tam giác bị ảnh hưởng nhiều bởi các đỉnh có bậc thấp.

$$score(u, v) = \sum_{z \in N(u) \cap N(v)} \frac{1}{\log(N(z))}$$



Độ đo Adamic Adar

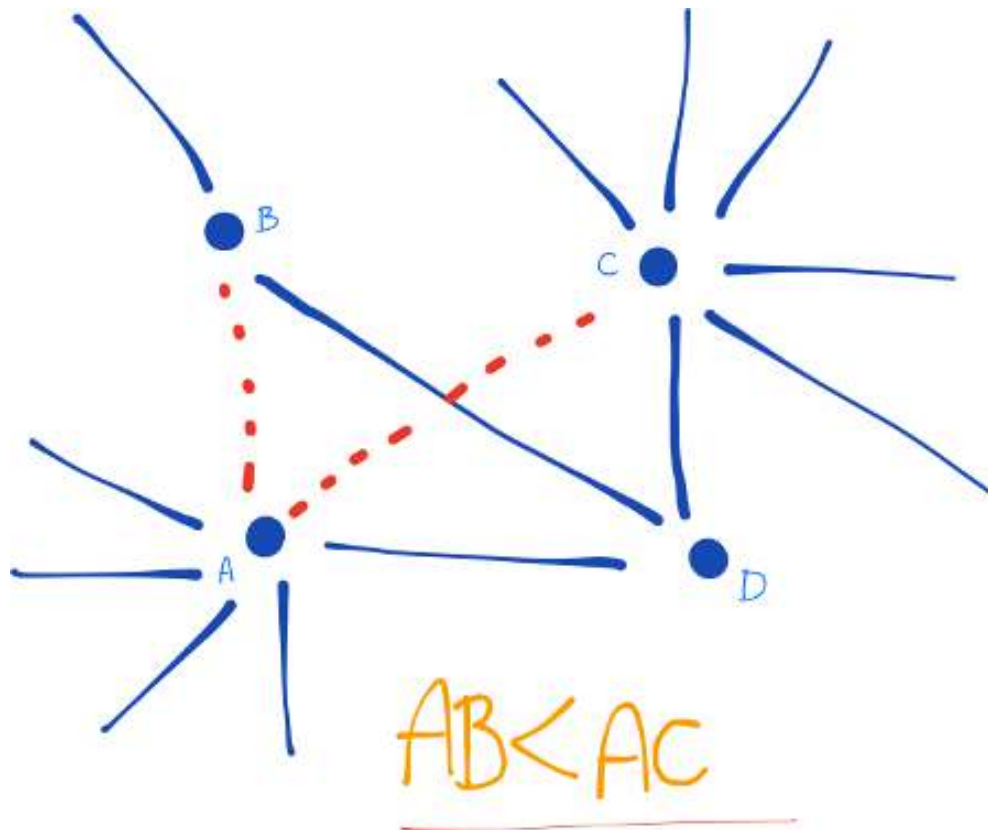


$$\text{adamic_adar}(A, C) = \frac{1}{\log(3)} + \frac{1}{\log(3)} = 1.82$$

Kết tương đồng

- **Kết tương đồng** (preferential attachment) đề xuất bởi Albert-László Barabási và Réka Albert để mô tả hiện tượng các đỉnh có nhiều mối quan hệ có xu hướng liên kết với nhau.

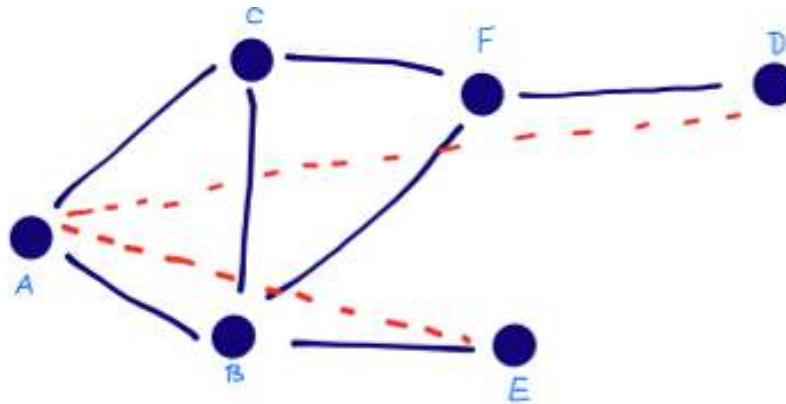
$$score(u, v) = degree(u) \times degree(v)$$



Đường đi ngắn nhất

- Cặp đỉnh nào có đường đi ngắn nhất sẽ có xu hướng kết nối

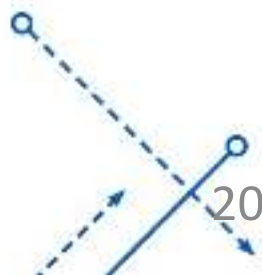
$$\text{score}(u, v) = -\text{shortestPath}(u, v)$$



$$\text{Score}(A, E) = -2 \checkmark$$

$$\text{Score}(A, D) = -3$$

↓ desc order

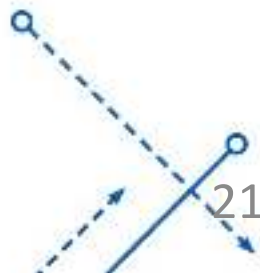


Độ đo Katz

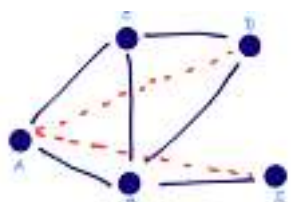
- Độ đo Katz không chỉ xem xét đường đi ngắn nhất giữa cặp đỉnh mà còn xét tất cả các đường đi giữa chúng.
 - Đường đi càng ngắn càng có trọng số cao hơn

$$score(u, v) = \sum_{l=1}^{\infty} \beta^l |paths_{u,v}^{<l>}|$$

với $paths_{u,v}^{<l>}$ là tập các đường đi chiều dài l giữa u và v , β là hằng số rất nhỏ để khiến cho đường đi càng dài càng có ít đóng góp vào phép tính tổng.



Độ đo Katz



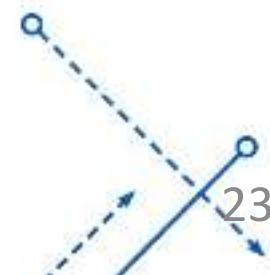
$$\begin{array}{l}
 \text{Độ đo Katz} \\
 P_{AB}^2 = 2 \quad P_{AC}^2 = 2 \\
 S = \frac{1}{2} \cdot 2 + \frac{1}{2} \cdot 2 + \dots \\
 \text{Độ đo Katz}
 \end{array}
 \quad
 \begin{array}{l}
 P_{AD}^2 = 1 \quad P_{AE}^2 = 1 \\
 S = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 1 + \dots
 \end{array}$$

SimRank

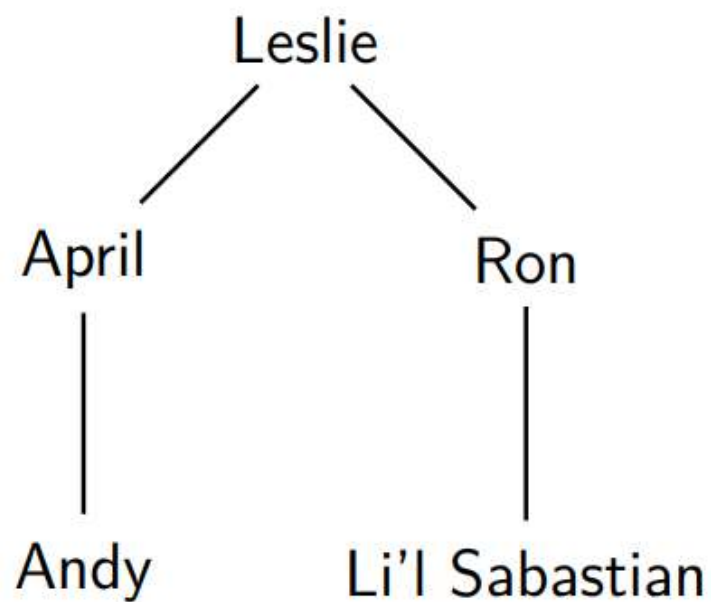
- **SimRank** được tính dựa trên độ đồng nhất của các láng giềng, nghĩa là láng giềng càng tương tự nhau thì hai đỉnh đó có xu hướng kết nối với nhau.

$$score(u, v) = \frac{C}{|N(u)| \cdot |N(v)|} \sum_{z \in N(u)} \sum_{z' \in N(v)} score(z, z')$$

với C là hằng số trọng số nằm trong $[0, 1]$



SimRank



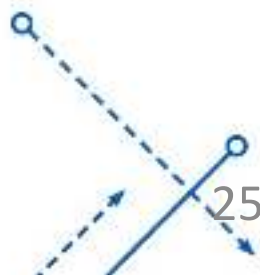
$$s(April, Ron) = 0.4$$

↓

$$s(Andy, Li'l Sebastian) = 0.33$$

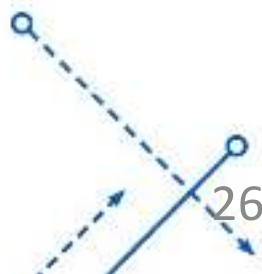
Tính trung tâm gần

- *Tính trung tâm gần* (closeness centrality): một tác nhân i là trung tâm gần nếu nó có thể tương tác **dễ dàng** với tất cả các tác nhân khác.
- Hay, khoảng cách của i đến tất cả tác nhân khác đều ngắn.



Tính trung tâm gần (tt)

- *Khoảng cách ngắn nhất* từ tác nhân i đến tác nhân j (kí hiệu $d(i,j)$) được đ \square bằng số liên kết trên đường đi ngắn nhất.
- Tính trung tâm gần của tác nhân i được kí hiệu là $C_c(i)$ và được chuẩn hóa với $n-1$ là tổng các kh \square ảng cách ngắn nhất từ i đến tất cả các tác nhân khác.



Tính trung tâm gần (tt)

- Đối với đồ thị vô hướng: trung tâm gần $C_c(i)$ của tác nhân i được định nghĩa như:

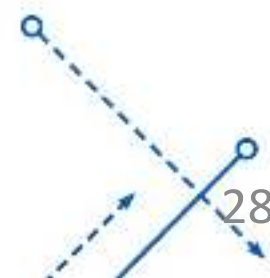
$$C_c(i) = \frac{n - 1}{\sum_{j=1}^n d(i, j)}$$

Lưu ý: biểu thức này chỉ thực hiện được trong trường hợp đồ thị liên thông



Tính trung tâm trung gian

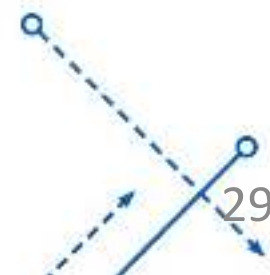
- Tính *trung tâm trung gian* (betweenness centrality): Nếu hai tác nhân j và k không kề nhau muốn tương tác và tác nhân i nằm giữa j và k thì i có thể có một số kiểm soát lên các tương tác của chúng.
- Nếu i ở trên đường đi của rất nhiều các tương tác khác nhau thì i là một tác nhân quan trọng.



Tính trung tâm trung gian (tt)

- Đối với đồ thị vô hướng, tính trung gian của một tác nhân i được định nghĩa bằng số lượng đường đi ngắn nhất qua i (kí hiệu $p_{jk}(i)$, $j \neq i$ và $k \neq i$) và được chuẩn hóa bởi tổng số lượng đường đi ngắn nhất của tất cả các cặp tác nhân ngoại trừ i :

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

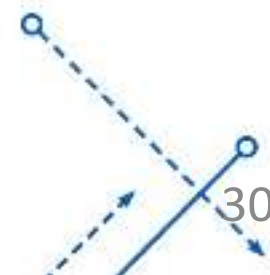


Tính trung tâm trung gian (tt)

- Để đảm bảo giá trị nằm giữa 0 và 1, $C_B(i)$ được chuẩn hóa với $(n-1)(n-2)/2$, đó là giá trị cực đại của $C_B(i)$:

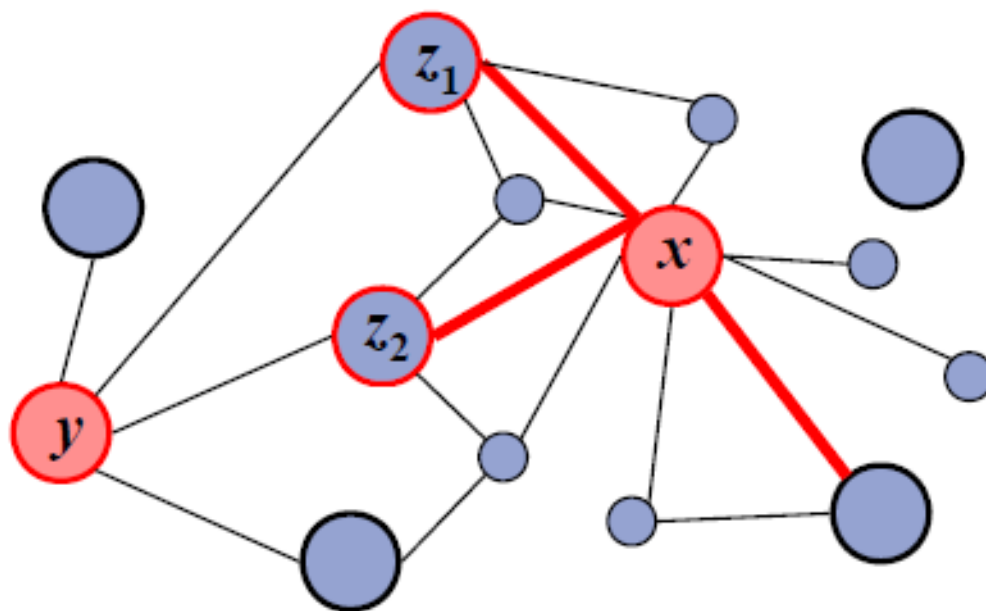
$$C_B(i) = \frac{2 \sum_{j < k} \frac{p_{jk}(i)}{p_{jk}}}{(n-1)(n-2)}$$

- Không giống như độ dài tính gần, tính trung gian có thể được tính thậm chí nếu đồ thị không liên thông.



Unseen Bigram

- **Bigram** (N-gram) thể hiện hai ký tự/từ đứng cạnh nhau trong câu ngôn ngữ tự nhiên.
- Nếu bigram không xuất hiện trong tập huấn luyện mà xuất hiện trong tập kiểm thử, người ta gọi đó là bigram ẩn (unseen bigram).
- Sử dụng một phương pháp trong mô hình ngôn ngữ để tính.

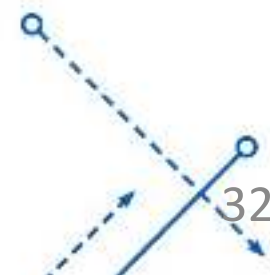


Unseen Bigram

- Đặt $S_u^{<\delta>}$ là tập δ đỉnh có độ tương tự cao với u thông qua độ dài gần đó:

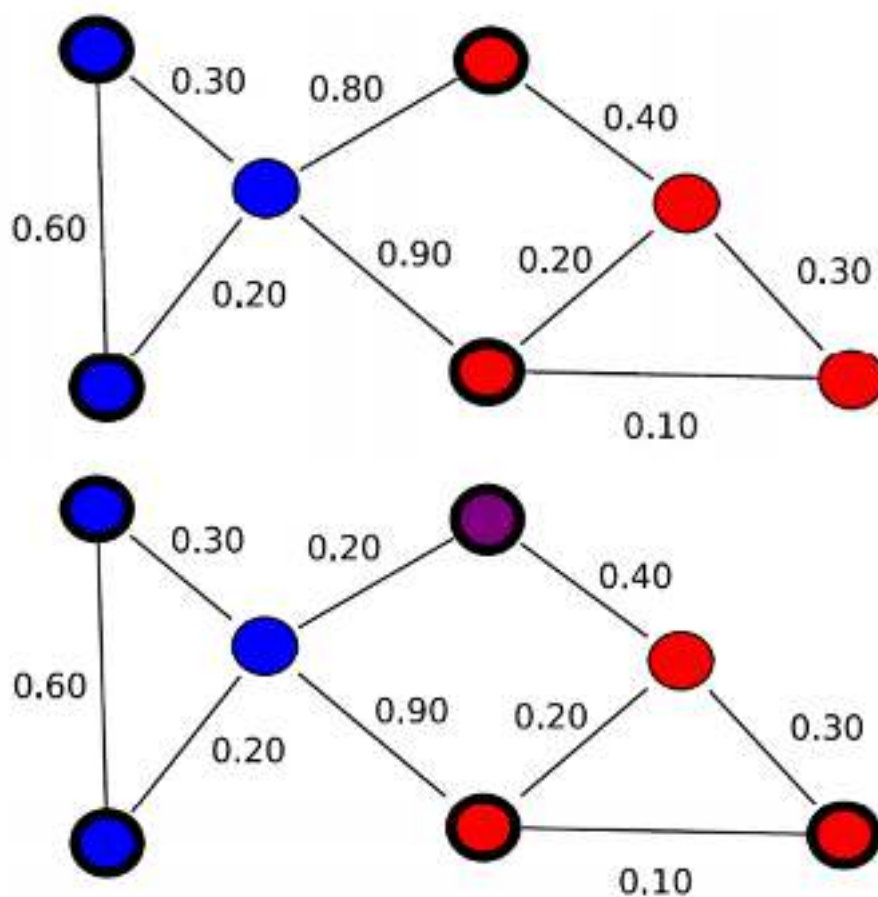
$$score_{unweighted}(u, v) = |\{z: z \in N(v) \cap S_u^{<\delta>}\}|$$

$$score_{weighted}(u, v) = \sum_{z \in N(v) \cap S_u^{<\delta>}} score(u, z)$$



Dựa trên gom nhóm

- Áp dụng một số phương pháp gom nhóm (ví dụ DB-Scan) để thực hiện gom nhóm đồ thị (phát hiện cộng đồng).
- Kết nối các đỉnh trong một nhóm dựa trên tiêu chí xác định.



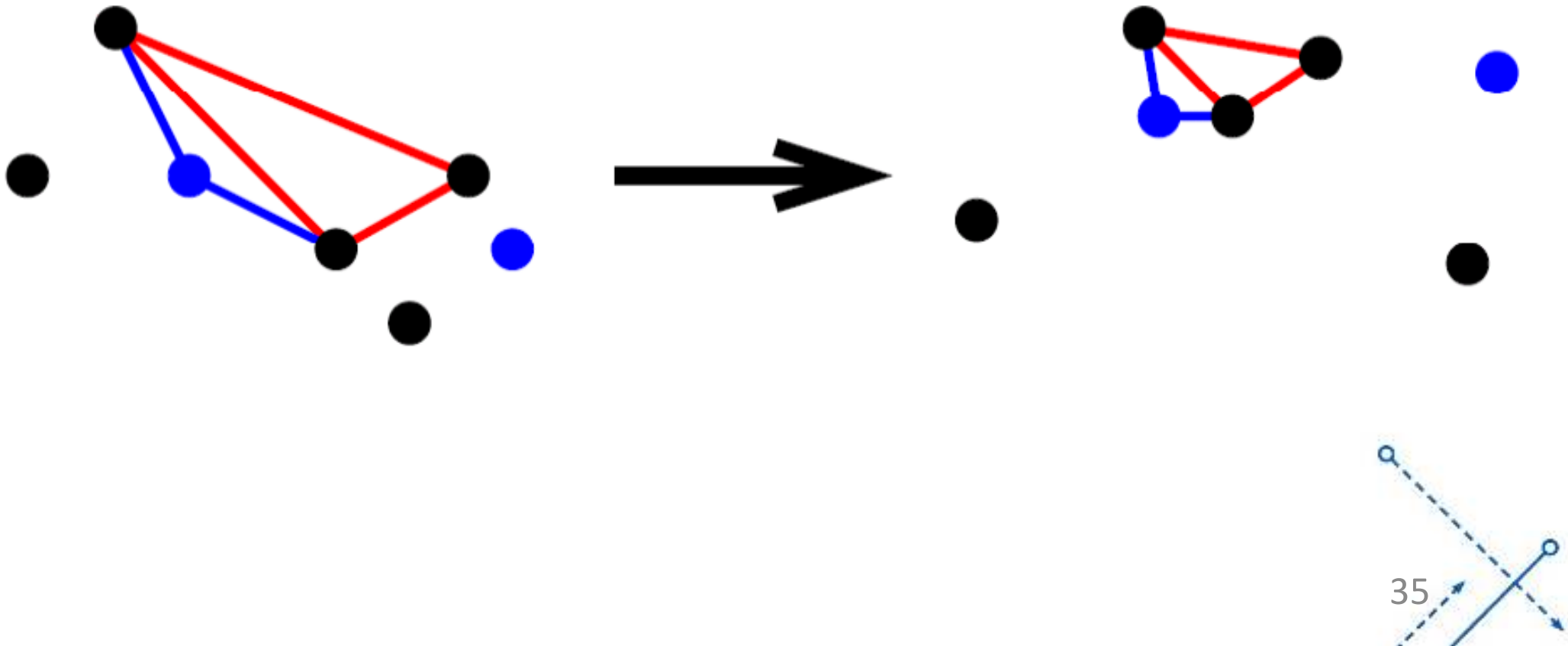
Nội dung

- Dự đoán liên kết
- **Học trong dự đoán liên kết**
 - Học không giám sát
 - **Học có giám sát**



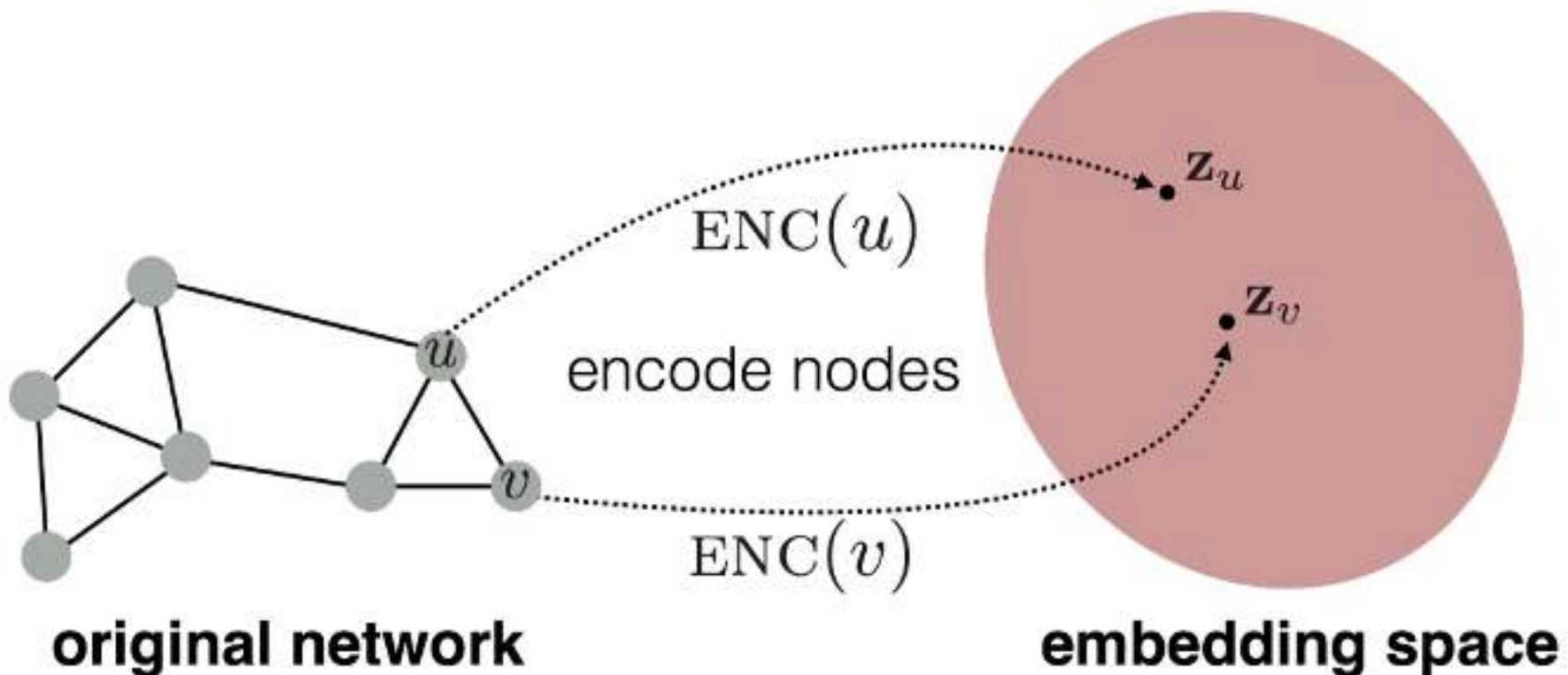
Học có giám sát cho dự đoán liên kết

- Ý tưởng:
 - Sử dụng mạng c_{in} đã biết để **điều chỉnh lại khoảng cách** trước khi áp dụng độ tương tự

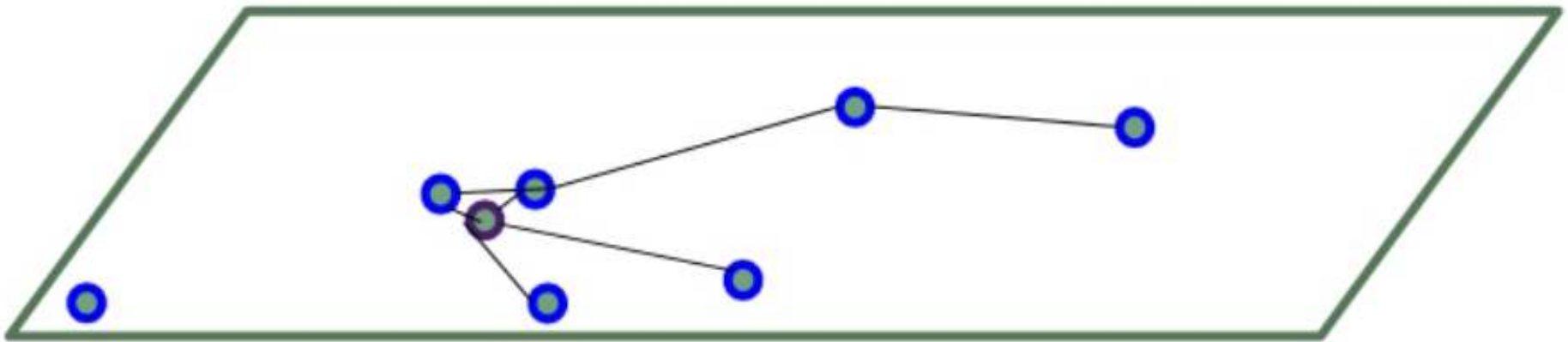
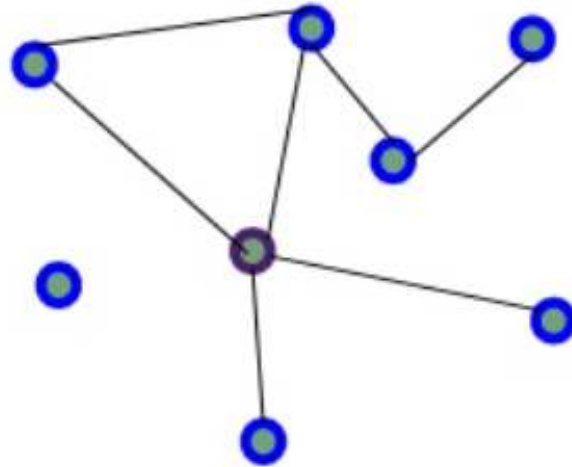


Học có giám sát cho dự đoán liên kết

- Có thể học dựa trên chuyển đổi về không gian khác để xác định độ tương tự (**nhúng đỉnh**).



Ví dụ

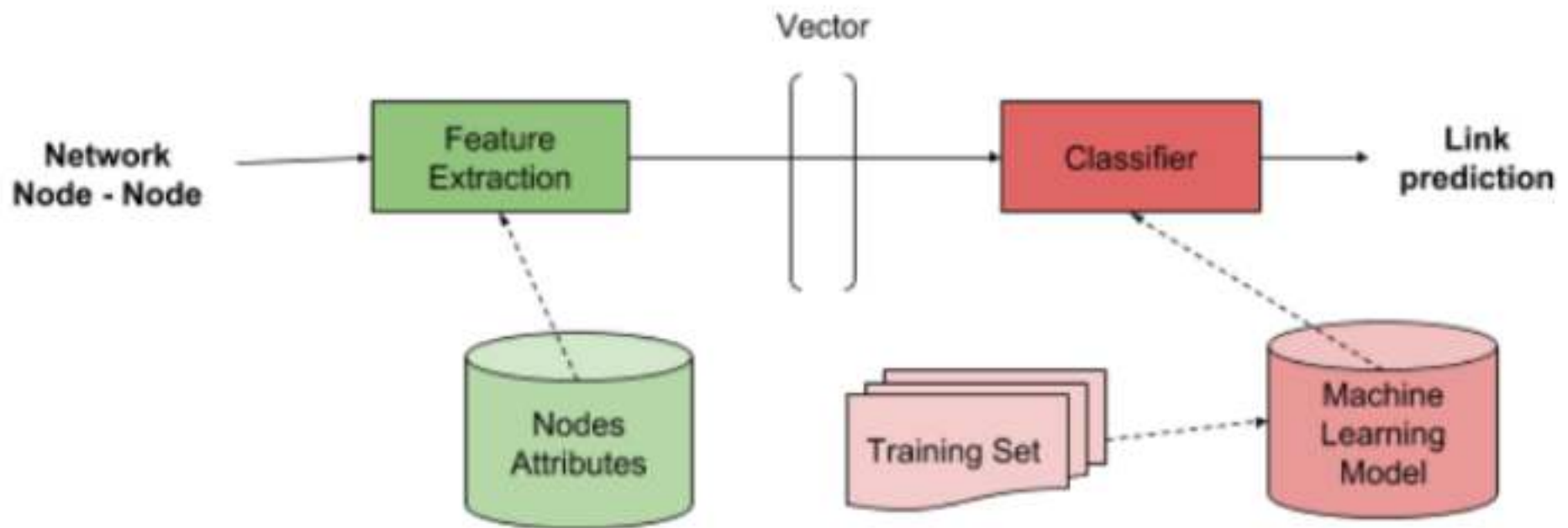


Hàm mục tiêu

$$F(\mathbf{A}_{source}) \simeq \mathbf{A}_{target}$$

$$\min ||F(\mathbf{A}_{source}) - \mathbf{A}_{target}||_F$$

Mô hình học máy



Tài liệu tham khảo

- <http://redalertproject.eu/wp-content/uploads/2019/05/D3-2-Link-prediction-models-FINAL.pdf>

