MẠNG XÃ HỘI

Bài 4. DỰ ĐOÁN LIÊN KẾT MẠNG XÃ HỘI

ThS. Lê Nhật Tùng

Mục lục

Nội dung

4.1.1. Giới thiệu

- "Có các quan hệ nào giữa các cá nhân trong tổ chức?".
- "Những người có khả năng tương tác trực tiếp với người này là ai?"

4.1.1. Giới thiệu

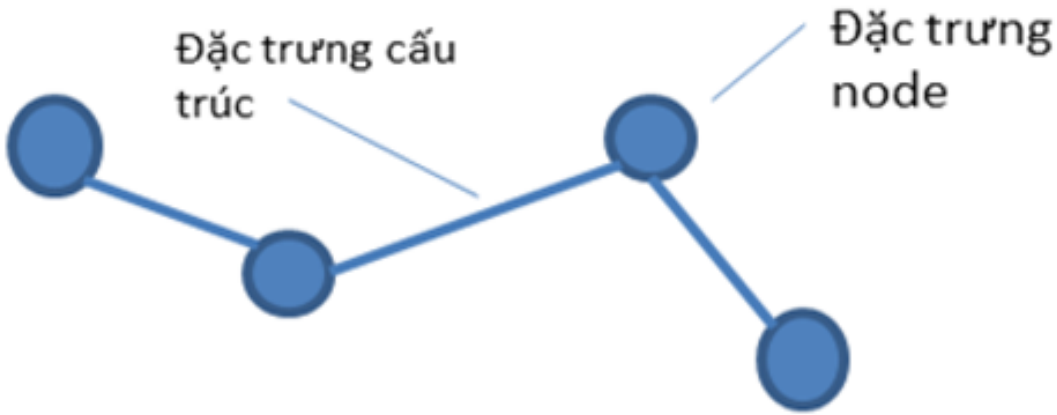
- Bài toán dự đoán liên kết nhằm mục đích nhận định có tồn tại hay không mối liên kết trong tương lai giữa hai actor trên mạng xã hội.
- Dựa trên các dữ liệu quan sát (liên kết hiện có) để suy đoán khả năng xuất hiện các liên kết mới trong tương lai.
- Ứng dụng trong business intelligence, cộng tác và dự đoán các mối quan hệ trong tương lai từ dữ liệu hiện có.

4.1.1. Giới thiệu

- Phương pháp khai phá dữ liệu truyền thống:
  - Làm việc trên một hoặc nhiều bảng dữ liệu
  - Bản ghi là các vector với giá trị thuộc tính
  - Có thể kết nối các bảng thành một bảng duy nhất
- Xu hướng mới - Hệ thống kết nối thông tin:
  - Mạng xã hội, mạng tương tác, mạng trích dẫn
  - Dữ liệu liên kết giữa các node đóng vai trò quan trọng
  - Khả năng dự đoán liên kết từ dữ liệu này

4.1.2. Nhiệm vụ dự đoán liên kết

- Bài toán dự đoán liên kết dựa trên:
  - Đặc điểm của node (node features)
  - Đặc điểm cấu trúc mạng (structural features)
- Biểu diễn mạng xã hội:
  - Cấu trúc đồ thị với node và liên kết
  - Node: đại diện cho dữ liệu
  - Liên kết: đại diện cho mối quan hệ giữa các dữ liệu
  - Mỗi node có thể liên kết với vector cấu trúc



Hình 4.1. Đặc trưng cấu trúc mạng

4.1.2. Nhiệm vụ dự đoán liên kết

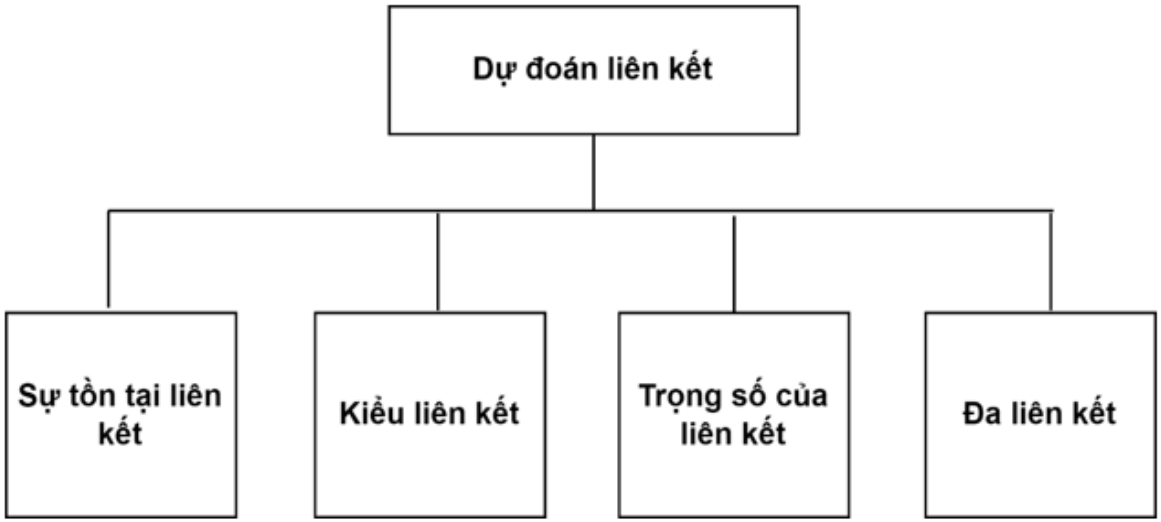
- Học bộ phân loại nhị phân (Binary Classification):
  - Dự đoán sự tồn tại của liên kết giữa cặp node
  - Dựa trên nghiên cứu của Hassan et al. (2006)
- Các thuật toán học có giám sát được sử dụng:
  - Cây quyết định (Decision Trees)
  - K láng giềng gần nhất (K-Nearest Neighbors)
  - Máy vector hỗ trợ (Support Vector Machines - SVM)

4.1.2. Nhiệm vụ dự đoán liên kết

- Phân loại dựa trên hai nhóm thuộc tính chính:
  - Thuộc tính của node (Node properties):
    - Số láng giềng (Number of neighbors)
    - Sở thích (Preferences)
    - Mô hình chủ đề (Topic models)
    - Cộng đồng (Communities)
    - Dữ liệu nhân khẩu học - vị trí địa lý (Demographic data - geographic location)
  - Thuộc tính dựa trên đồ thị (Graph-based properties):
    - Chiều dài đường đi ngắn nhất (Shortest path length)
    - Chồng chéo vùng lân cận (Neighborhood overlap)
    - Tầm quan trọng (Importance)
    - Thời điểm liên kết (Link timestamp)
- Mô hình đồ thị (Graph models):
  - Đồ thị có hướng vs vô hướng
  - Mạng Bayesian và PRMs:
    - Dễ dàng nắm bắt sự phụ thuộc của liên kết
    - Hạn chế xác suất phụ thuộc đồ thị có hướng

4.1.2. Nhiệm vụ dự đoán liên kết

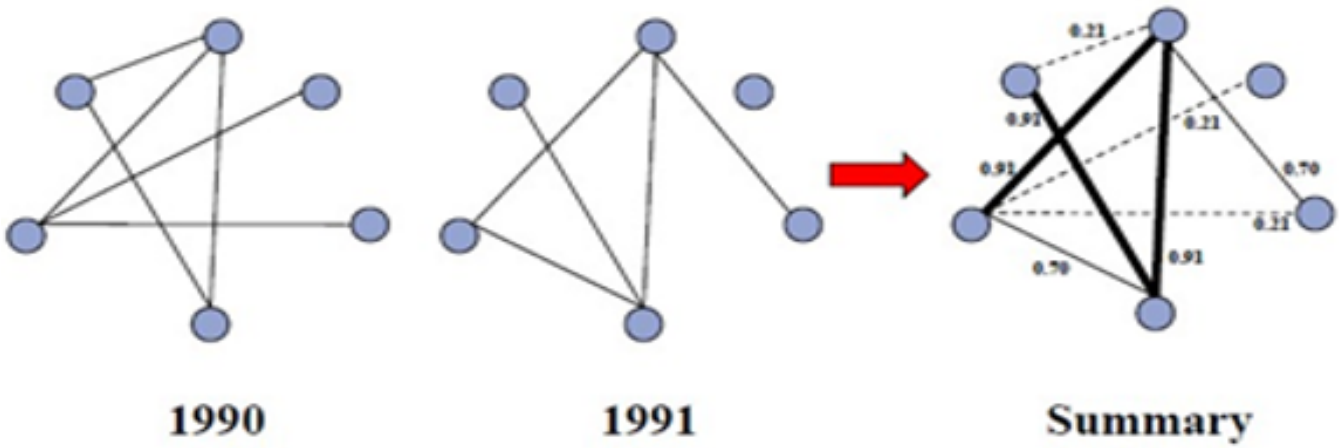
- Hầu hết các nghiên cứu về dự đoán liên kết đều tập trung vào vấn đề có tồn tại liên kết:
  - Dự đoán trong tương lai có xuất hiện liên kết giữa hai node trong mạng xã hội hay không
- Mở rộng bài toán:
  - Liên kết có trọng số
  - Đa liên kết (nhiều hơn một liên kết giữa cùng cặp nút)
  - Dự đoán kiểu liên kết và vai trò của liên kết giữa hai actor



Hình 4.2. Bốn nhiệm vụ dự đoán liên kết

4.1.2. Nhiệm vụ dự đoán liên kết

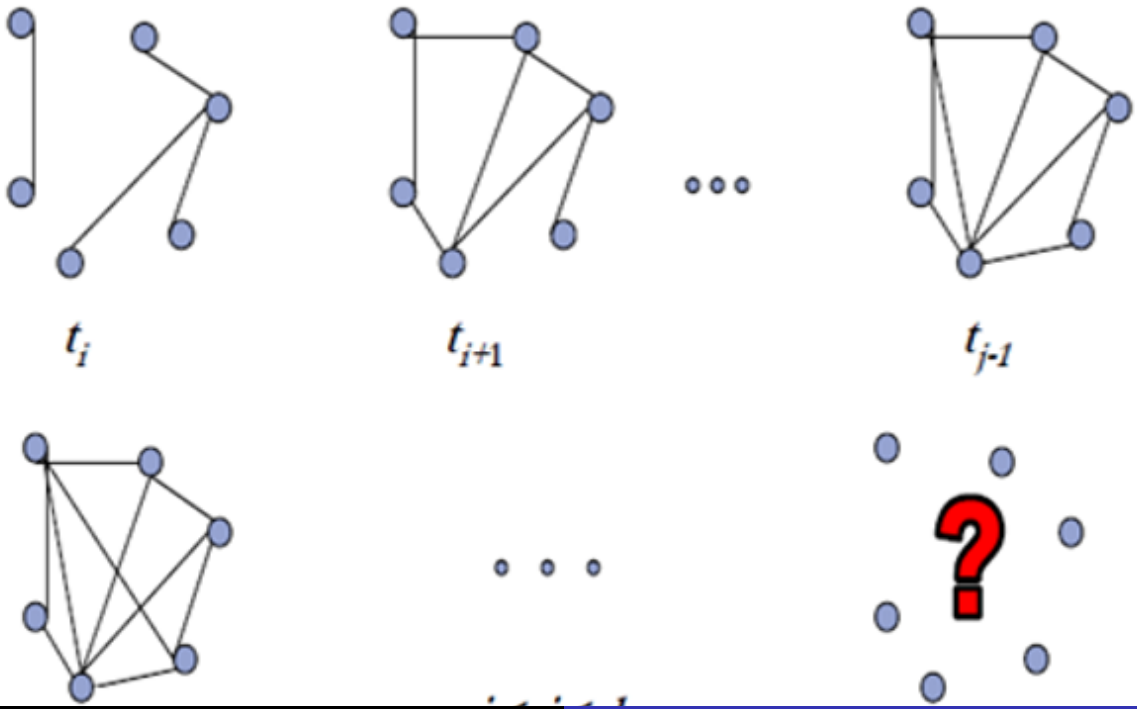
- Khi dự báo liên kết, có thể áp dụng mô hình đồ thị xác suất:
  - Mỗi cung sẽ có trọng số là xác suất
  - Giúp đánh giá khả năng xuất hiện liên kết



Hình 4.3. Áp dụng mô hình xác suất để dự đoán liên kết

4.1.3. Mô tả bài toán dự đoán liên kết

- Biểu diễn dữ liệu:
  - Dạng đồ thị với các đỉnh đại diện cho các thực thể
  - Cung đại diện cho các liên kết hay tương tác giữa các node
- Các loại mối quan hệ:
  - Giao tiếp trực tiếp
  - Đồng xảy ra
  - Chia sẻ các thuộc tính chung



4.1.3. Mô tả bài toán dự đoán liên kết

- Phát biểu toán học:
  - Cho tập dữ liệu $V = \{v_i\}_{i=1}^{n}$
  - Mạng xã hội $G = (V, E)$, với E là tập hợp các liên kết
  - Dự đoán liên kết chưa thấy $e_{ij} \in E$ giữa cặp node $\langle v_i, v_j \rangle$
- Mô hình hóa đồ thị:
  - Tại thời điểm t(i): $G_{t(i)} = (V, E)$
  - Đồ thị đang phát triển:

$$EG = \{G_{t(i)} | i = 1...p, t(i) < t(i+1)\}$$

4.1.3. Mô tả bài toán dự đoán liên kết

- Đặc điểm mô hình dự đoán:
  - Kết quả dự đoán là đồ thị $G_{t(p+1)}$
  - Số node không đổi trong quá trình phát triển
  - Chỉ dự đoán sự tồn tại của liên kết
  - Trọng số cạnh = tổng số lần liên lạc giữa hai node
- Giới hạn bài toán:
  - Không dự đoán trọng số của các liên kết
  - Chỉ dự đoán liên kết tồn tại hoặc không tồn tại

4.1.3. Mô tả bài toán dự đoán liên kết

- Ba phương pháp dự đoán liên kết chính:
  - Học máy (Machine Learning):
    - Tạo mô hình phân loại nhị phân
  - Mô hình cấu trúc topo
  - Mô hình xác suất:
    - Markov
    - Bayes
    - Random

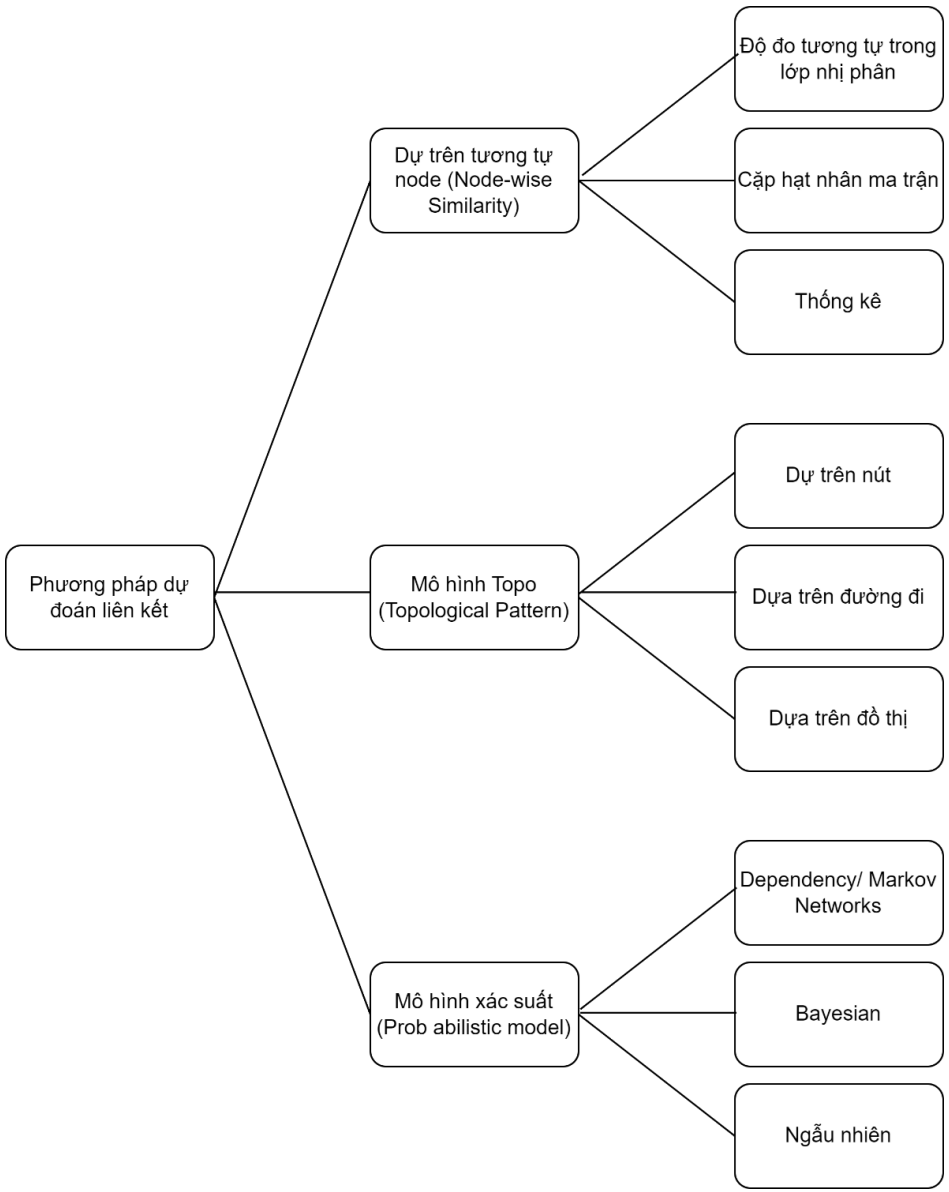
Nội dung

4.2. Các cách tiếp cận dự đoán liên kết

- Học máy có giám sát:
  - Mohammad Al Hasan, Vineet Chaoji và cộng sự
  - Sử dụng đặc trưng cấu trúc mạng và thuộc tính node
  - Thử nghiệm trên mạng BIOBASE và DBLP
- Mạng Bayes (2009):
  - Anet Potgieter, Kurt A. April và cộng sự
  - Thử nghiệm trên mạng kết bạn Pussokram
- Mô hình trích dẫn và Random Walks (2011):
  - Naoki Shibata, Yuya Kajikawa, Ichiro Sakata: Mô hình SVM
  - L. Backström và J. Leskovec: Supervised Random Walks
  - Thử nghiệm trên mạng Facebook
  - Kết hợp cấu trúc mạng với thuộc tính node

Nội dung

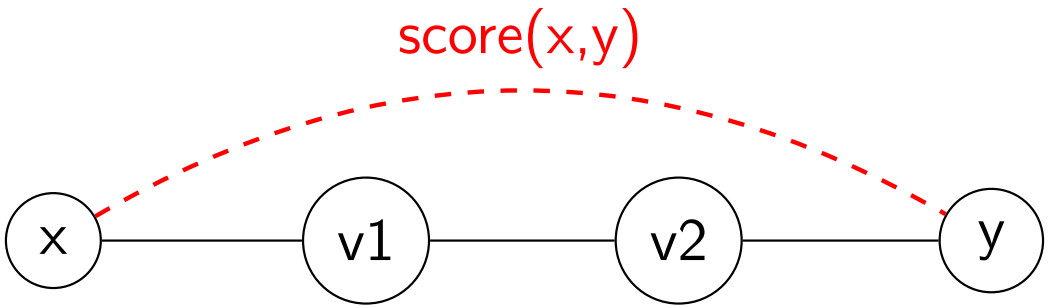
4.3. Môt số phương pháp dự đoán liên kết



Nội dung

4.4.1. Khoảng cách đồ thị

- Phương pháp dự đoán dựa trên sự tương tự của node:
  - Xác định score(x,y) cho cặp node x, y
  - Tạo danh sách node xếp theo thứ tự giảm dần của score(x,y)
- Khoảng cách đồ thị:
  - Phản ánh sự "tương tự"giữa các nút x và y
  - Là chiều dài đường đi ngắn nhất giữa 2 node



Khoảng cách đồ thị = 3

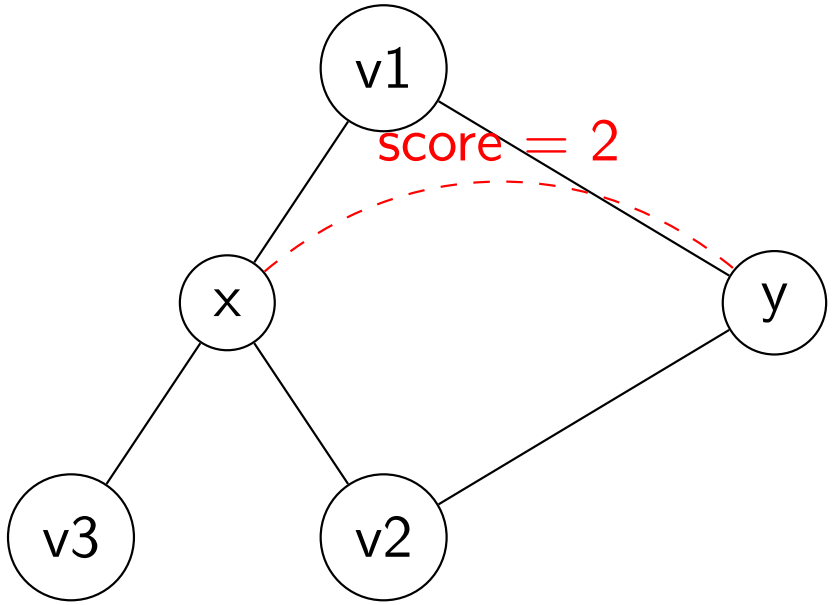
4.4.2. Láng giềng chung

- Phương pháp dự đoán dựa trên mô hình topo
- Công thức tính điểm:

$$score(x, y) = |\Gamma(x) \cap \Gamma(y)|$$

Trong đó:

  - $\Gamma(x)$: tập láng giềng của node x
  - $\Gamma(y)$: tập láng giềng của node y
- Nghiên cứu của Newman:
  - Áp dụng trong mạng cộng tác
  - Dự đoán mối quan hệ tương lai



Ví dụ:

$\Gamma(x) = \{v1, v2, v3\}$

$\Gamma(y) = \{v1, v2\}$

$|\Gamma(x) \cap \Gamma(y)| = 2$

4.4.3. Hệ số Jaccard

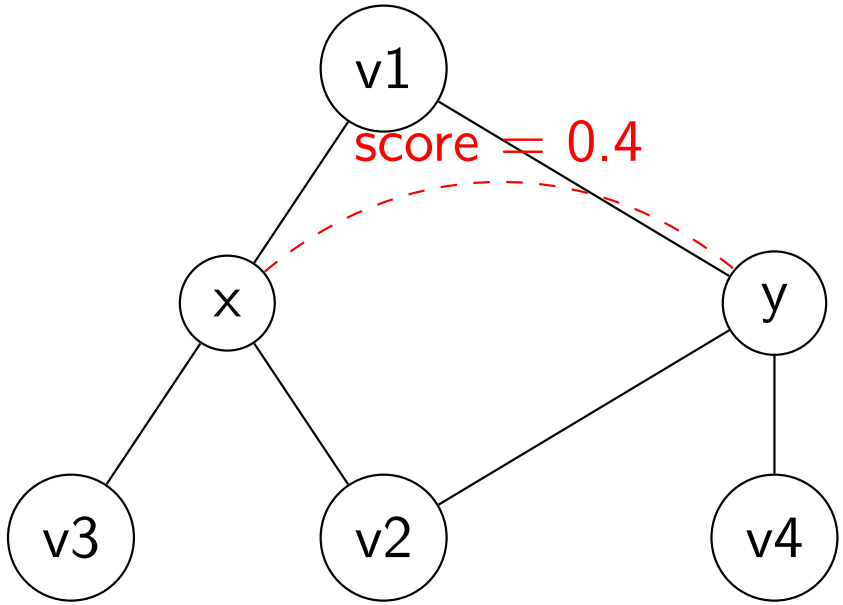
- Phương pháp dự đoán dựa trên mô hình topo
- Công thức hệ số Jaccard:

$$score(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

Trong đó:

- $\Gamma(x)$: tập láng giềng của node x
- $\Gamma(y)$: tập láng giềng của node y

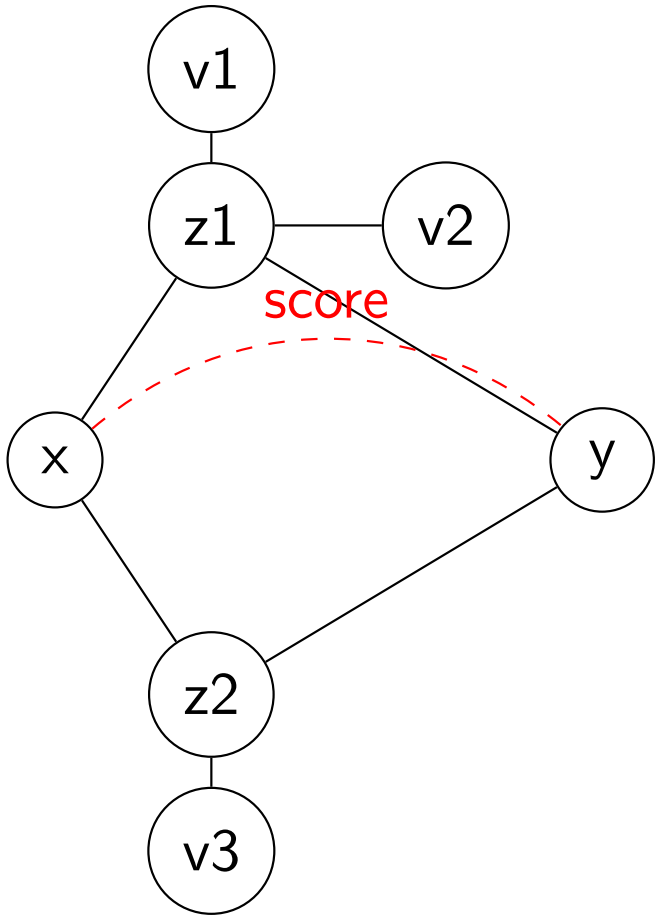
- Ứng dụng:
  - Truy vấn thông tin xác suất, So sánh sự giống nhau của láng giềng
  - Đánh giá đa dạng của



Ví dụ:

$$\Gamma(x) = \{v1, v2, v3\}$$
$$\Gamma(y) = \{v1, v2, v4\}$$
$$|\Gamma(x) \cap \Gamma(y)| = 2$$
$$|\Gamma(x) \cup \Gamma(y)| = 4$$
$$score = \frac{2}{4} = 0.5$$

4.4.4. Hệ số Adamic/Adar - Ví dụ minh họa



Hình: Minh họa trọng số láng giềng

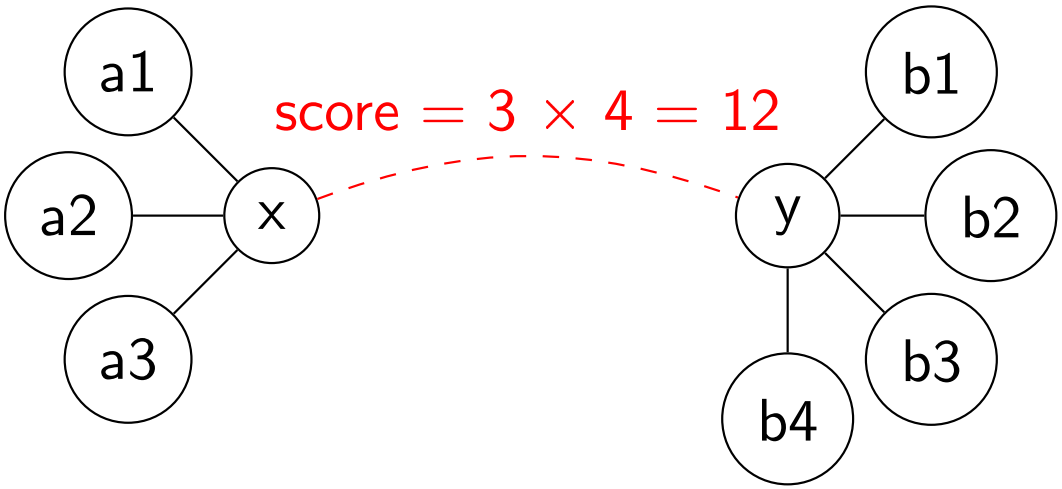
- Phân tích ví dụ:
  - z1 có 4 láng giềng (v1, v2, x, y): $\frac{1}{\log(4)}$
  - z2 có 3 láng giềng (v3, x, y): $\frac{1}{\log(3)}$
  - Score tổng = $\frac{1}{\log(4)} + \frac{1}{\log(3)}$

4.4.5. Preferential attachment

- Phương pháp dự đoán dựa trên mô hình topo
- Ý tưởng cơ bản:
  - Xác suất xuất hiện cung mới tỷ lệ thuận với số láng giềng hiện có
  - Xác suất đồng tác giả tương quan với tích số cộng tác viên
- Công thức tính điểm:

$$score(x, y) = |\Gamma(x)| \cdot |\Gamma(y)|$$

- Trong đó:
  - $|\Gamma(x)|$: số láng giềng của node x
  - $|\Gamma(y)|$: số láng giềng của



Hình: Ví dụ: x có 3 láng giềng, y có 4 láng giềng

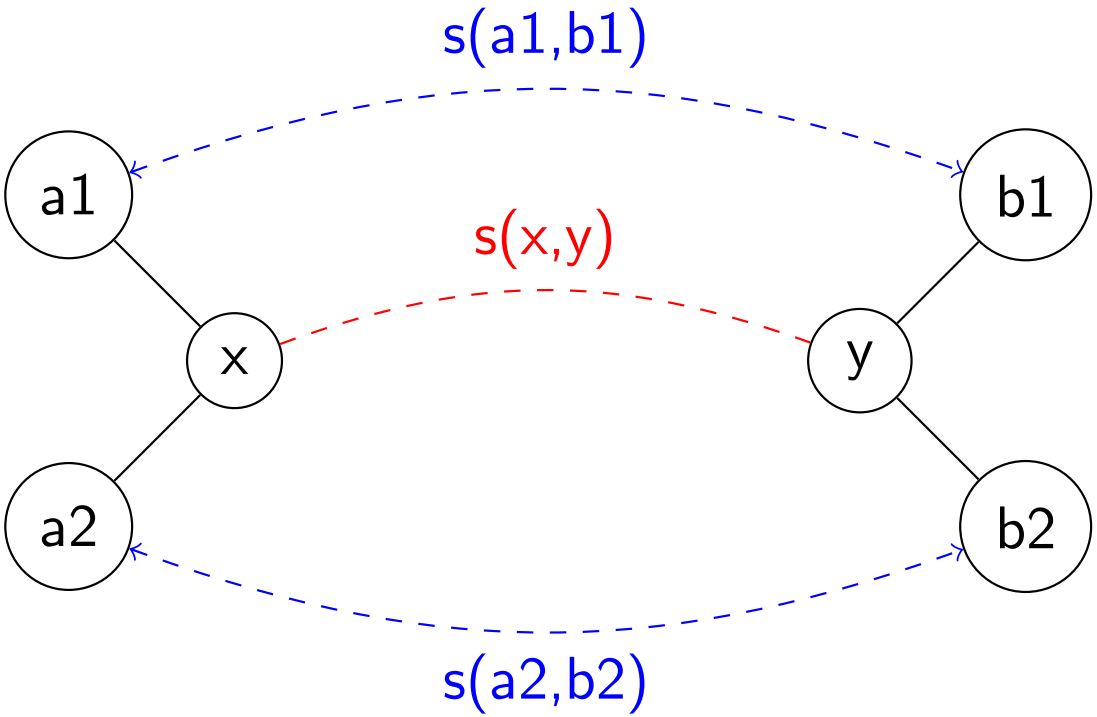
4.4.6. SimRank

- Phương pháp dự đoán dựa trên mô hình topo
- Định nghĩa:
  - Hai nút được coi là tương tự dựa trên số láng giềng tương tự
  - Mức độ tương tự phụ thuộc vào mức độ tương tự của láng giềng
- Công thức tính điểm:

$$score(x, y) = \gamma \cdot \frac{\sum_{a \in \Gamma(x)} \sum_{b \in \Gamma(y)} score(a, b)}{|\Gamma(x)| \cdot |\Gamma(y)|}$$

- Trong đó:
  - $\Gamma(x), \Gamma(y)$: tập láng giềng của x và y
  - $\gamma \in [0, 1]$: hệ số suy giảm
  - $score(a, b)$: điểm tương tự giữa láng giềng a và b

4.4.6. SimRank - Ví dụ minh họa



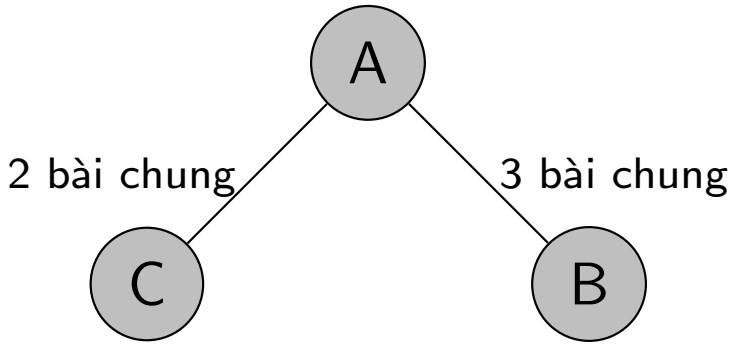
Hình: Minh họa cách tính SimRank

- Phân tích ví dụ:
  - x có 2 láng giềng: a1, a2
  - y có 2 láng giềng: b1, b2
  - Score tổng phụ thuộc vào:
    - Điểm s(a1,b1), s(a2,b2)
    - Hệ số γ

4.4.7. Dự đoán dựa trên thuộc tính của node và cung

- **Ý tưởng chính:**
  - Xem xét các đặc điểm của node và mối quan hệ
  - Các node có đặc điểm tương tự thường có xu hướng kết nối

- **Ví dụ trong mạng nghiên cứu:**
  - Thuộc tính node: lĩnh vực, chủ đề nghiên cứu
  - Thuộc tính cung: số bài báo chung, dự án chung



Các node có thuộc tính giống nhau có khả năng kết nối cao hơn

4.4.7. Dự đoán dựa trên thuộc tính - Ví dụ cụ thể

- **Xét mạng đồng tác giả:**
  1. **Thu thập thuộc tính:**
     - Lĩnh vực nghiên cứu của mỗi tác giả
     - Từ khóa trong các bài báo
     - Số lượng công bố
  2. **Phân tích mối quan hệ:**
     - Tác giả cùng lĩnh vực $\rightarrow$ khả năng hợp tác cao
     - Nhiều từ khóa chung $\rightarrow$ khả năng là đồng tác giả cao
- **Ưu điểm:**
  - Dự đoán chính xác hơn nhờ xem xét nhiều yếu tố
  - Phù hợp với thực tế xã hội

Bài thực hành: Dự đoán liên kết

- **Mục tiêu:**
  - Áp dụng các phương pháp dự đoán liên kết
  - Thực nghiệm trên dữ liệu có sẵn của NetworkX
- **Dataset:**
  - *Karate Club Network*: Mạng xã hội của câu lạc bộ karate
  - *Davis Southern Women*: Mạng xã hội phụ nữ miền Nam
  - *Florentine Families*: Mạng các gia đình ở Florence

Phân tích

1. **Phân tích mạng:**
   - Số lượng node ($|V|$) và cạnh ($|E|$)
   - Phân phối bậc của node ($P(k)$)
   - Hệ số phân cụm (C)
2. **Đánh giá mô hình:**
   - Độ chính xác (Accuracy): $\frac{TP+TN}{TP+TN+FP+FN}$
   - Độ nhạy (Recall): $\frac{TP}{TP+FN}$
   - Độ chính xác (Precision): $\frac{TP}{TP+FP}$
3. **So sánh kết quả:**
   - Hiệu quả của từng đặc trưng
   - So sánh giữa các dataset
   - Đề xuất cải tiến

Chúc các bạn học thật tốt!