

DỰ ĐOÁN LIÊN KẾT

Biên soạn: **ThS. Nguyễn Thị Anh Thư**

Email: thunta@uit.edu.vn

NỘI DUNG

1. Khái niệm
2. Định nghĩa bài toán
3. Các bước giải bài toán
4. Đồ thị có dấu (signed networks)
5. Đồ thị vô hướng có dấu
6. Đồ thị có hướng có dấu
7. Độ tương đồng cục bộ

1. KHÁI NIỆM

- Dự đoán liên kết là một trong những bài toán con cơ bản và phổ biến nhất của bài toán phân tích mạng. Nó góp phần vào quá trình **nghiên cứu sự hình thành và phát triển của mạng**.
- **Dự đoán liên kết được chia làm bốn bài toán con như sau:**

Dự đoán liên kết

Dự đoán sự tồn tại liên kết

Dự đoán trọng số liên kết

Dự đoán số lượng liên kết

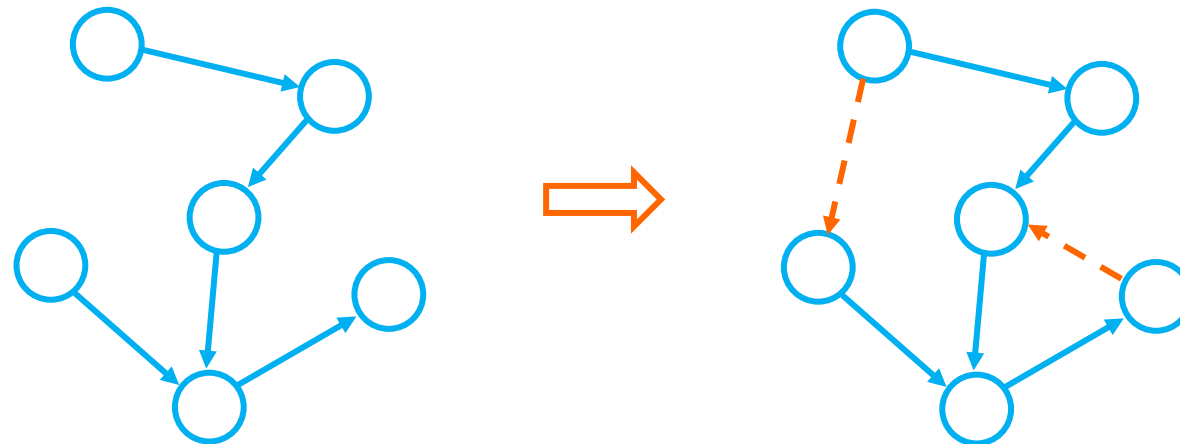
Dự đoán loại liên kết

1. KHÁI NIỆM

- Mạng được biểu diễn dưới dạng đồ thị theo lý thuyết đồ thị được gọi là đồ thị mạng.
- Bài toán dự đoán liên kết trong đồ thị mạng được định nghĩa cụ thể dựa trên nền tảng lý thuyết đồ thị.
- **Mạng được thiết kế như một đồ thị $G(V, E, W, M, S)$ trong đó:**
 - V : là tập các đỉnh.
 - E : là tập các cạnh.
 - W : là tập trọng số của các cạnh.
 - M : là tập số lượng các cạnh giữa các cặp đỉnh.
 - S : là tập dấu hiệu (tích cực (+) hoặc tiêu cực (-)) của các cạnh.

2. ĐỊNH NGHĨA BÀI TOÁN

- **Bài toán dự đoán sự tồn tại liên kết trong mạng** có input và output như sau:
 - **Input**: Thông tin về số cạnh trong đồ thị mạng tại một thời điểm cụ thể $G[t_0, t'_0]$.
 - **Output**: Các cạnh còn thiếu trong đồ thị đó trong tương lai $G[t_1, t'_1]$.
- **Đánh giá**:
 - $n = |E_{new}|$: Các cạnh mới xuất hiện trong thời gian thực nghiệm $[t_1, t'_1]$.
 - Lấy n phần tử trên cùng của L và đếm các cạnh đúng (L là một danh sách xếp hạng các cạnh mới xuất hiện trong $[t_1, t'_1]$).



Mô hình bài toán dự đoán sự tồn tại của liên kết.

2. ĐỊNH NGHĨA BÀI TOÁN

- **Bài toán dự đoán trọng số liên kết trong mạng** có input và output như sau:
 - **Input:** Thông tin về trọng số tương ứng của từng cạnh trong đồ thị mạng tại một thời điểm cụ thể $G[t_0, t'_0]$.
 - **Output:** Trọng số các cạnh còn thiếu trong đồ thị đó trong tương lai $G[t_1, t'_1]$.



Mô hình bài toán dự đoán trọng số của liên kết.

2. ĐỊNH NGHĨA BÀI TOÁN

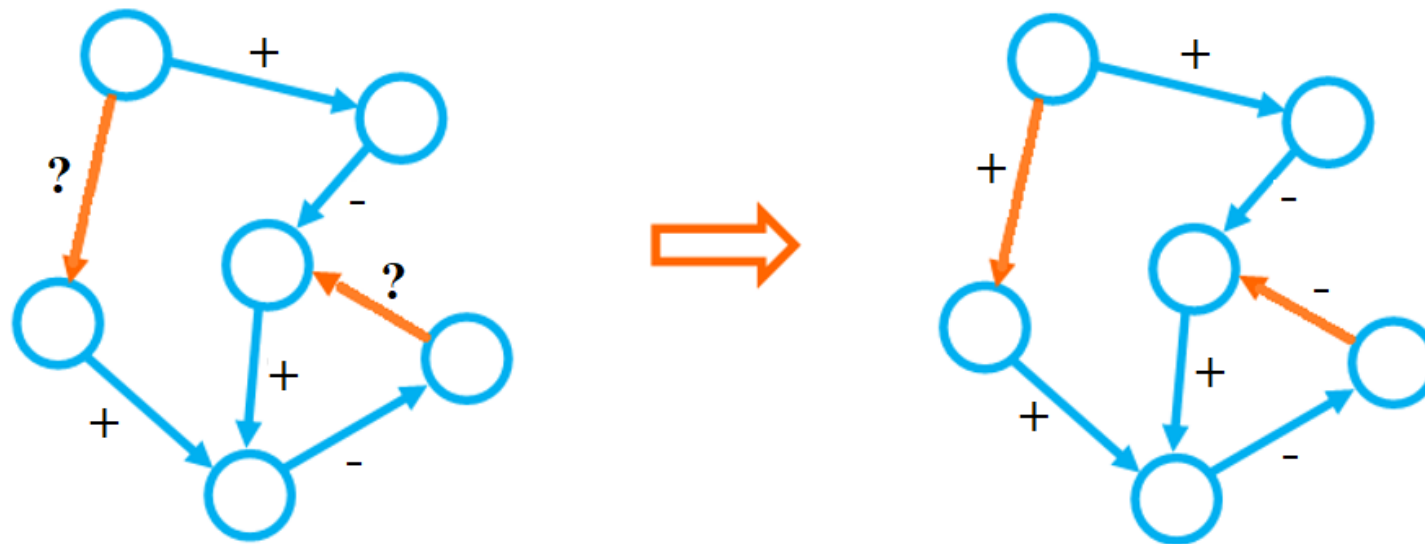
- **Bài toán dự đoán trọng số liên kết trong mạng** có input và output như sau:
 - **Input:** Thông tin về trọng số tương ứng của từng cạnh trong đồ thị mạng tại một thời điểm cụ thể $G[t_0, t'_0]$.
 - **Output:** Trọng số các cạnh còn thiếu trong đồ thị đó trong tương lai $G[t_1, t'_1]$.
- **Đánh giá:**
 - $w = |W_{new}|$: Trọng số của các cạnh mới xuất hiện trong thời gian thực nghiệm $[t_1, t'_1]$.
 - Lấy w phần tử trên cùng của L và đếm các cạnh đúng (L là một danh sách xếp hạng trọng số của các cạnh mới xuất hiện trong $[t_1, t'_1]$).
 - Trọng số của một cạnh dự đoán được tính là đúng với sai số α chấp nhận được.

2. ĐỊNH NGHĨA BÀI TOÁN

- **Bài toán dự đoán số lượng liên kết trong mạng** có input và output như sau:
 - **Input**: Thông tin về số lượng liên kết tương ứng của từng cạnh trong đồ thị mạng tại một thời điểm cụ thể $G[t_0, t'_0]$.
 - **Output**: Số lượng các liên kết xuất hiện thêm trong đồ thị đó trong tương lai $G[t_1, t'_1]$.
- **Đánh giá:**
 - $m = |M_{new}|$: Số lượng liên kết của các cạnh mới xuất hiện trong thời gian thực nghiệm $[t_1, t'_1]$.
 - Lấy m phần tử trên cùng của L và đếm số lượng liên kết của các cạnh đúng (L là một danh sách xếp hạng số lượng liên kết của các cạnh mới xuất hiện trong $[t_1, t'_1]$).

2. ĐỊNH NGHĨA BÀI TOÁN

- **Bài toán dự đoán loại liên kết trong mạng** có input và output như sau:
 - **Input:** Thông tin dấu hiệu cạnh (tích cực (+) hoặc tiêu cực (-)) tương ứng của từng cạnh trong đồ thị mạng tại một thời điểm cụ thể $G[t_0, t'_0]$.
 - **Output:** Dấu hiệu các cạnh còn thiếu trong đồ thị đó trong tương lai $G[t_1, t'_1]$.



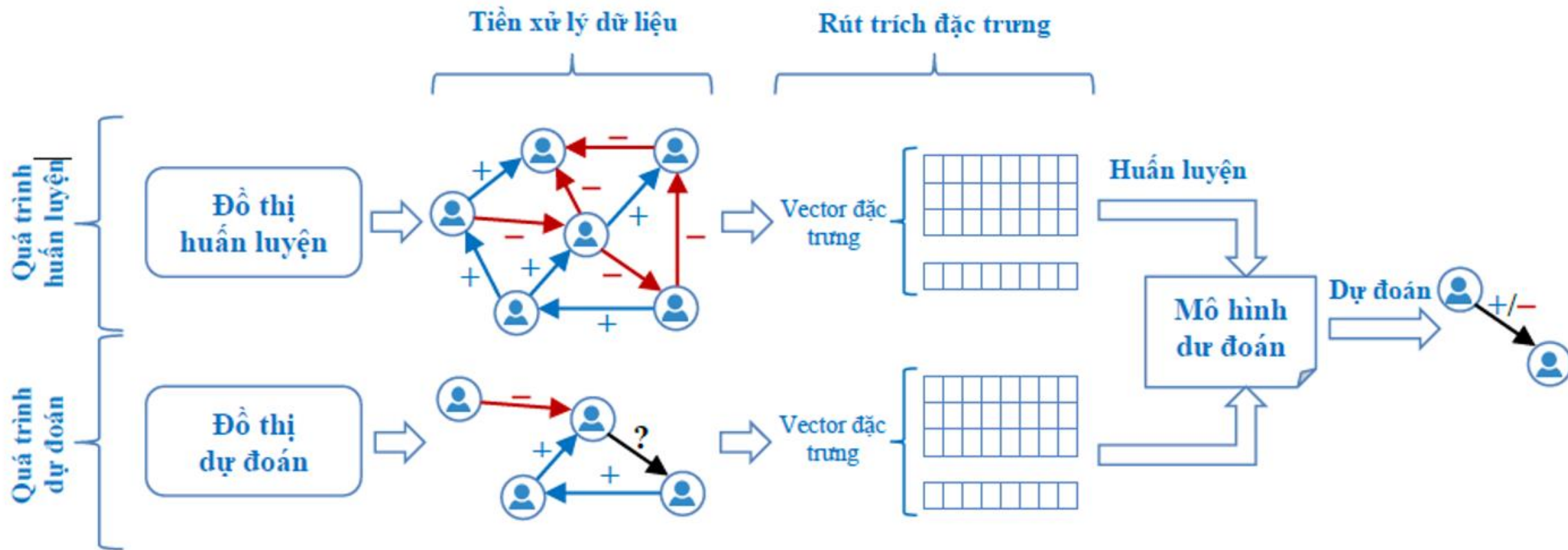
Mô hình bài toán dự đoán dấu hiệu của liên kết.

2. ĐỊNH NGHĨA BÀI TOÁN

- **Bài toán dự đoán loại liên kết trong mạng** có input và output như sau:
 - **Input**: Thông tin dấu hiệu cạnh (tích cực (+) hoặc tiêu cực (-)) tương ứng của từng cạnh trong đồ thị mạng tại một thời điểm cụ thể $G[t_0, t'_0]$.
 - **Output**: Dấu hiệu các cạnh còn thiếu trong đồ thị đó trong tương lai $G[t_1, t'_1]$.
- **Đánh giá**:
 - $s = |S_{new}|$: Dấu hiệu cạnh mới xuất hiện trong thời gian thực nghiệm $[t_1, t'_1]$.
 - Lấy s phần tử trên cùng của L và đếm dấu hiệu cạnh đúng (L là một danh sách xếp hạng dấu hiệu của các cạnh mới xuất hiện trong $[t_1, t'_1]$).

3. CÁC BƯỚC GIẢI BÀI TOÁN

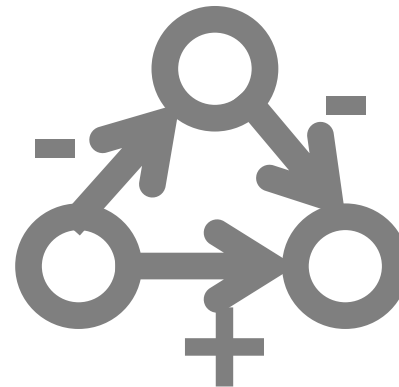
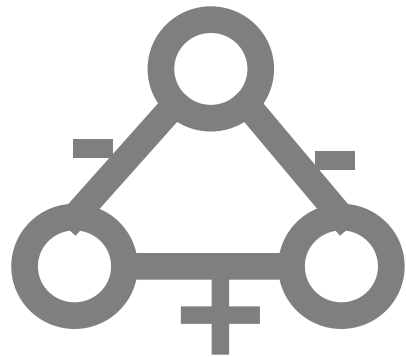
- Theo hướng máy học (học có giám sát).



Ví dụ : Framework bài toán dự đoán dấu hiệu của liên kết.

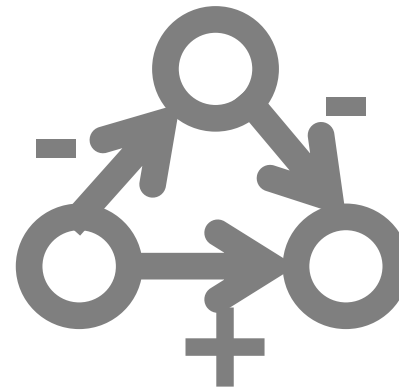
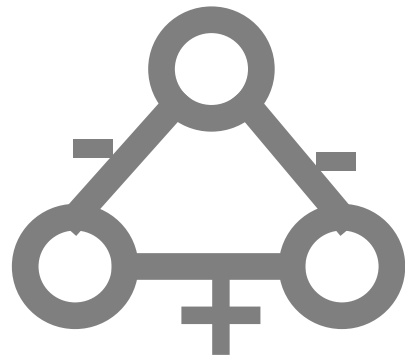
4. ĐỒ THỊ CÓ DẤU (SIGNED NETWORKS)

- **Mạng có mối quan hệ tích cực (+) và tiêu cực (-).**
- Đơn vị cơ bản để tạo nên mạng này là các hình tam giác có dấu.
- **Ứng dụng:** Dự đoán sự tồn tại của liên kết dựa vào dấu hiệu liên kết tích cực (+) và tiêu cực (-).



5. ĐỒ THỊ VÔ HƯỚNG CÓ DẤU

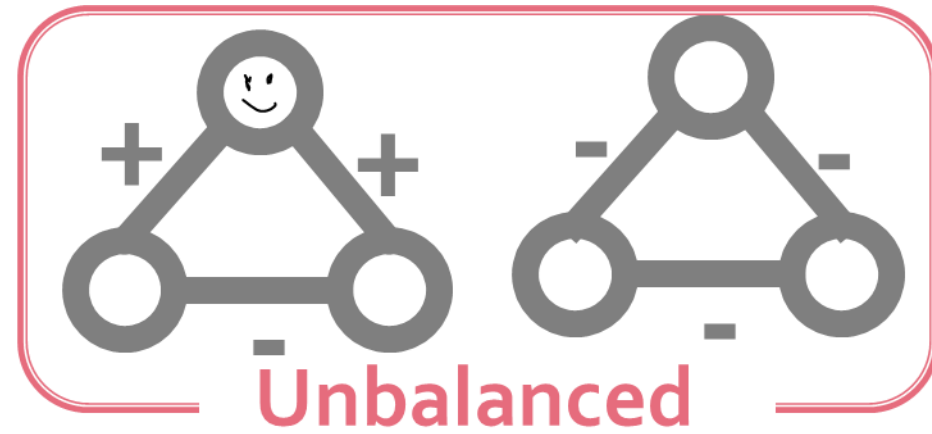
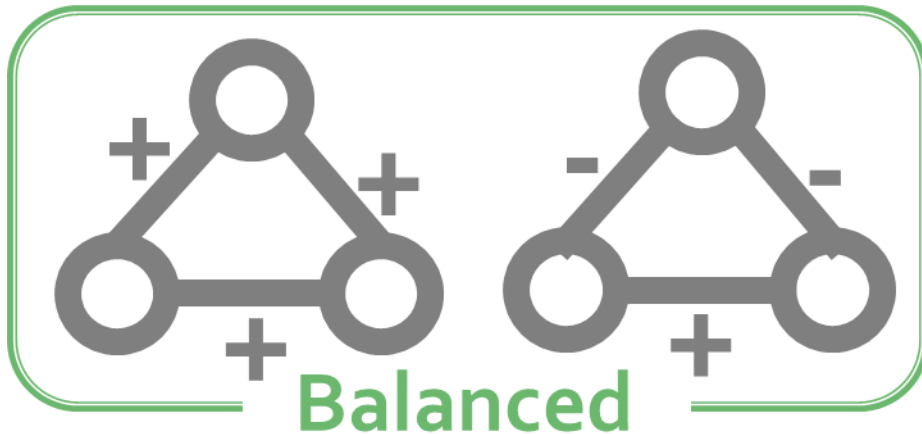
- **Mạng có mối quan hệ tích cực (+) và tiêu cực (-).**
- Xem xét một biểu đồ hoàn chỉnh vô hướng:
 - Gắn nhãn mỗi cạnh là:
 - **Tích cực:** tình bạn, sự tin tưởng, tình cảm tích cực, ...
 - **Tiêu cực:** kẻ thù, sự ngờ vực, tình cảm tiêu cực, ...
 - Kiểm tra bộ ba của các nút được kết nối A, B, C.



5. ĐỒ THỊ VÔ HƯỚNG CÓ DẤU

LÝ THUYẾT CÂN BẰNG CẤU TRÚC

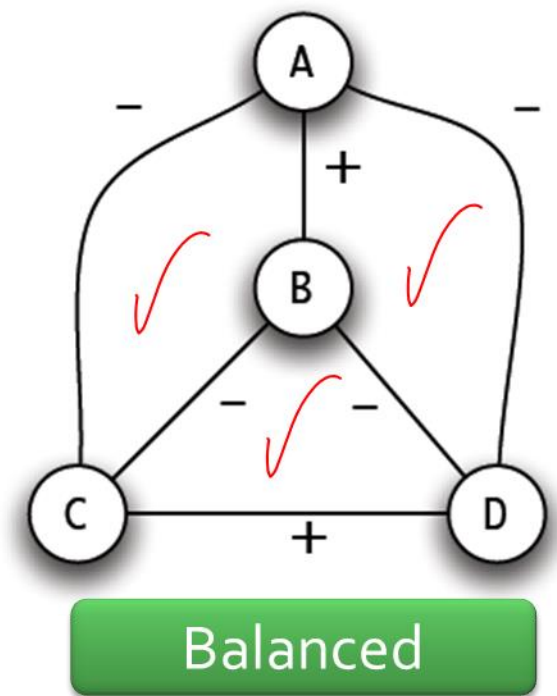
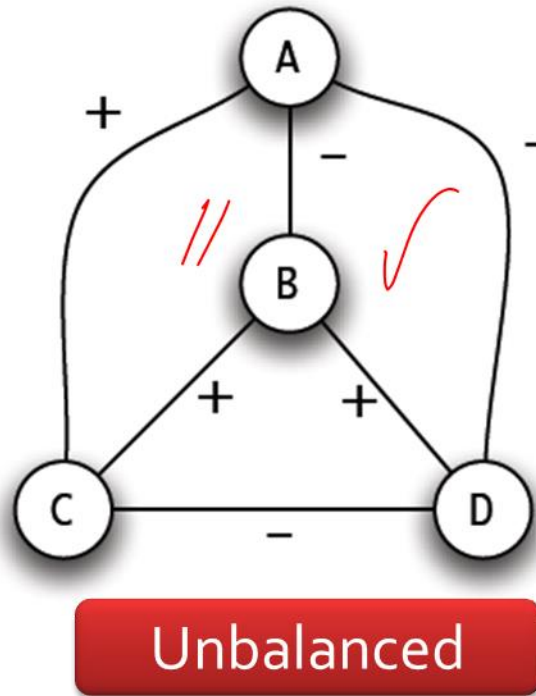
- Dựa vào lý thuyết tâm lý xã hội [Heider '46]:
 - **Bạn của bạn tôi là bạn của tôi.**
 - **Kẻ thù của kẻ thù là bạn của tôi.**
 - **Kẻ thù của bạn bè là kẻ thù của tôi.**
- Nhìn vào bộ ba nút được kết nối:



5. ĐỒ THỊ VÔ HƯỚNG CÓ DẤU

LÝ THUYẾT CÂN BẰNG CẤU TRÚC

- Đồ thị là cân bằng nếu mọi bộ ba nút được kết nối có:
 - Cả 3 cạnh có nhãn + , hoặc
 - Chính xác 1 cạnh có nhãn +

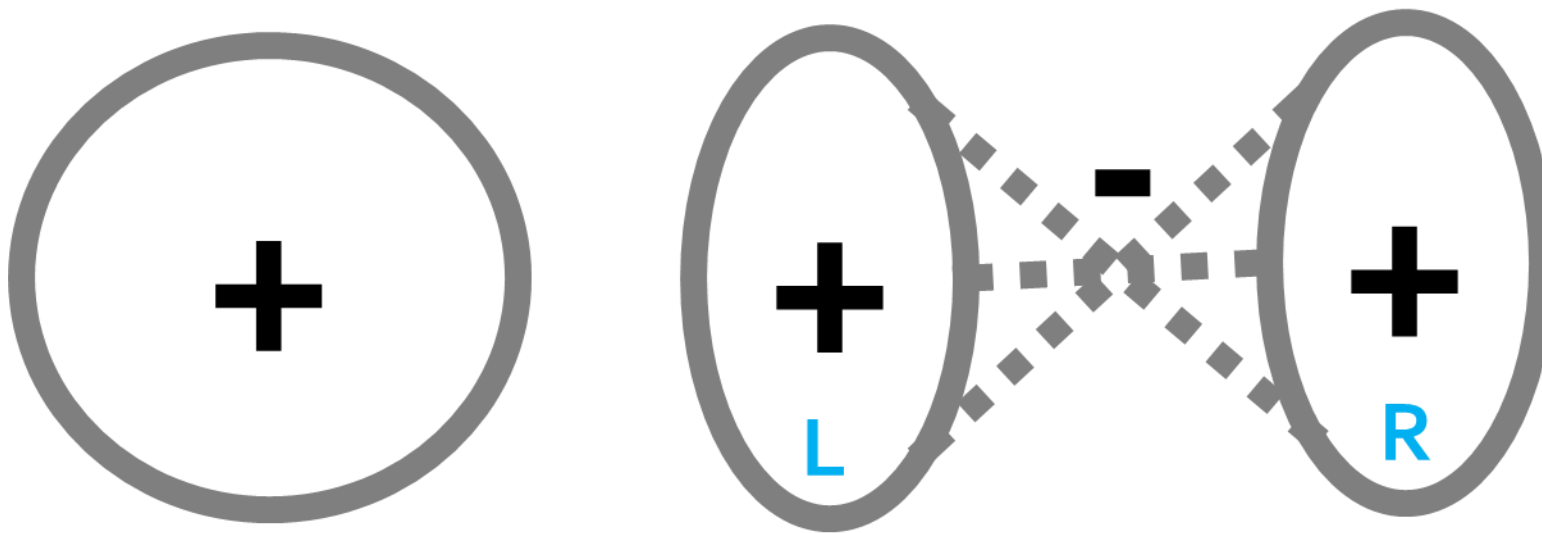


5. ĐỒ THỊ VÔ HƯỚNG CÓ DẤU

LÝ THUYẾT CÂN BẰNG CẤU TRÚC

Cân bằng cục bộ

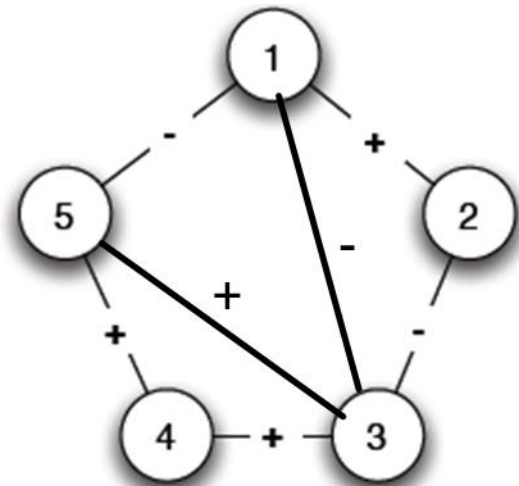
- Cân bằng ngụ ý các liên minh toàn cầu [Cartwright-Harary]
- Nếu tất cả các tam giác đều cân bằng, thì:
 - Mạng chỉ chứa các cạnh dương, hoặc
 - Các nút có thể được chia thành 2 bộ trong đó các cạnh âm chỉ trở giữa các bộ



5. ĐỒ THỊ VÔ HƯỚNG CÓ DẤU

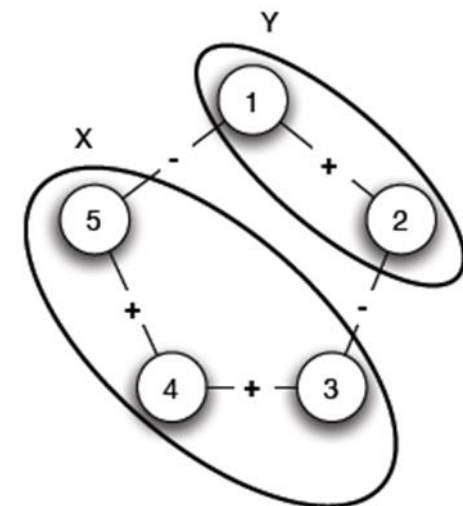
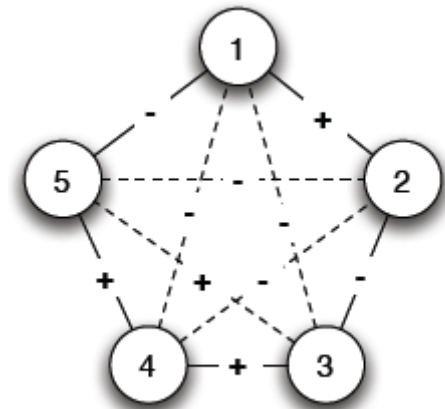
LÝ THUYẾT CÂN BẰNG CẤU TRÚC

Cân bằng trong mạng chung



Balanced?

- **Định nghĩa 1: Chế độ xem cục bộ**
 - Điền vào các cạnh còn thiếu để đạt được sự cân bằng
- **Định nghĩa 2: Chế độ xem toàn cầu**
 - Chia biểu đồ thành hai liên minh
- Hai định nghĩa này là tương đương.



5. ĐỒ THỊ VÔ HƯỚNG CÓ DẤU

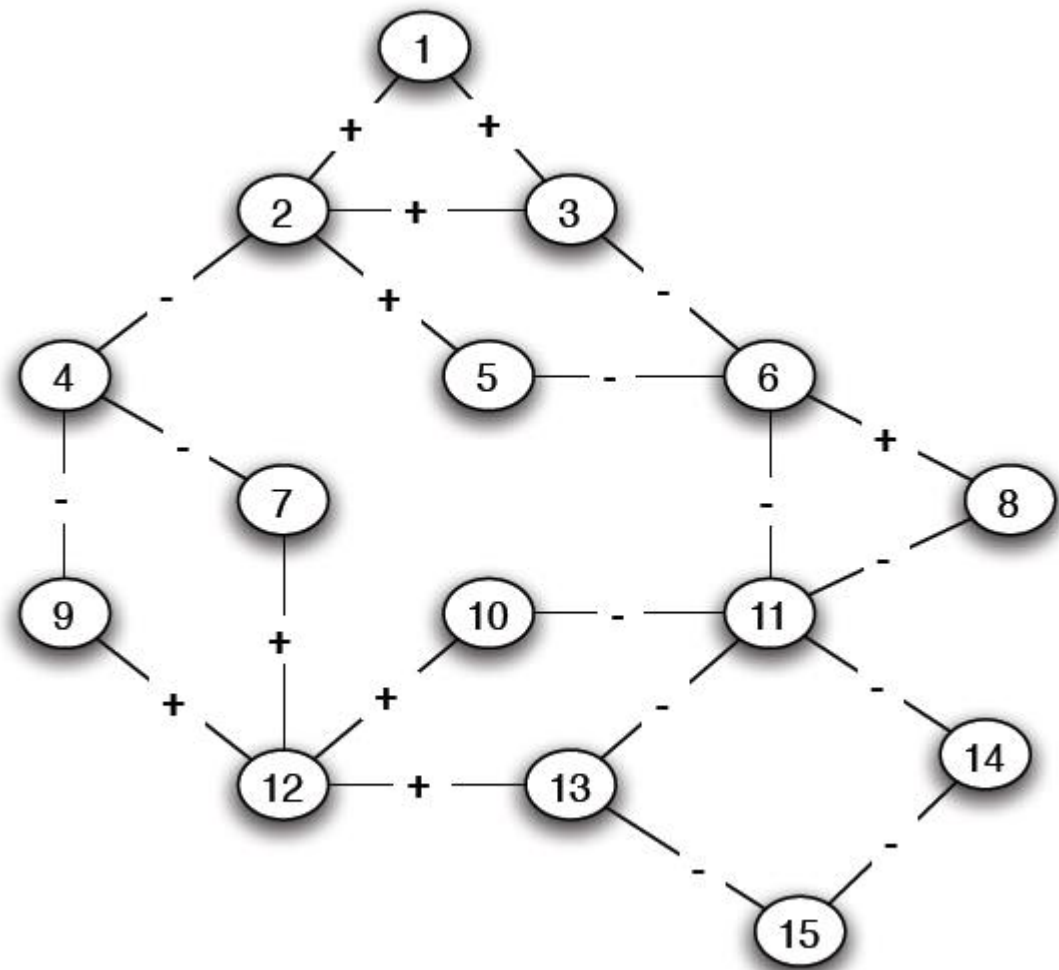
LÝ THUYẾT CÂN BẰNG CẤU TRÚC

- Đồ thị là cân bằng khi và chỉ khi nó không chứa chu trình với số cạnh âm là số lẻ.
- **Cách kiểm tra tính cân bằng của đồ thị có dấu:**
 1. Tìm các thành phần được kết nối trên cạnh +.
 2. Đối với mỗi thành phần, hãy tạo một siêu nút.
 3. Kết nối các thành phần A và B nếu có một cạnh tiêu cực giữa các thành viên.
 4. Gán các siêu nút cho các bên bằng BFS.

5. ĐỒ THỊ VÔ HƯỚNG CÓ DẤU

LÝ THUYẾT CÂN BẰNG CẤU TRÚC

- **Ví dụ:** Kiểm tra tính cân bằng của đồ thị bên.



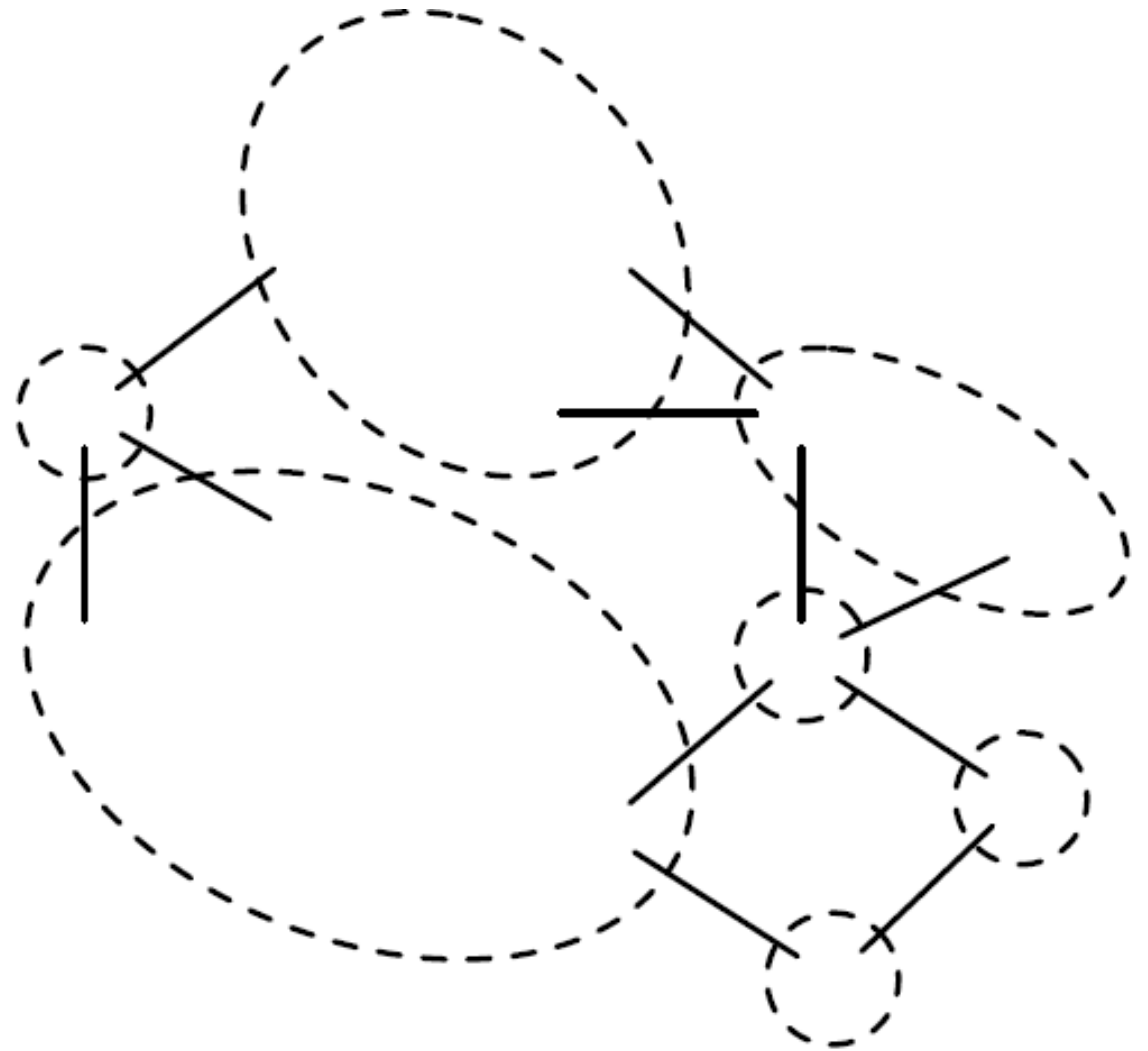
LÝ THUYẾT CÂN BẰNG CẤU TRÚC

-
- The graph consists of 15 nodes and 18 edges. The nodes are numbered 1 to 15. The edges are labeled with '+' or '-' signs. The graph is partitioned into four clusters by dashed lines: Cluster 1 (nodes 1, 2, 3, 4, 5), Cluster 2 (nodes 6, 7, 8, 9, 10, 11, 12, 13), Cluster 3 (nodes 14, 15), and Cluster 4 (node 16). The edges are labeled with '+' or '-' signs.

5. ĐỒ THỊ VÔ HƯỚNG CÓ DẦU

LÝ THUYẾT CÂN BẰNG CẤU TRÚC

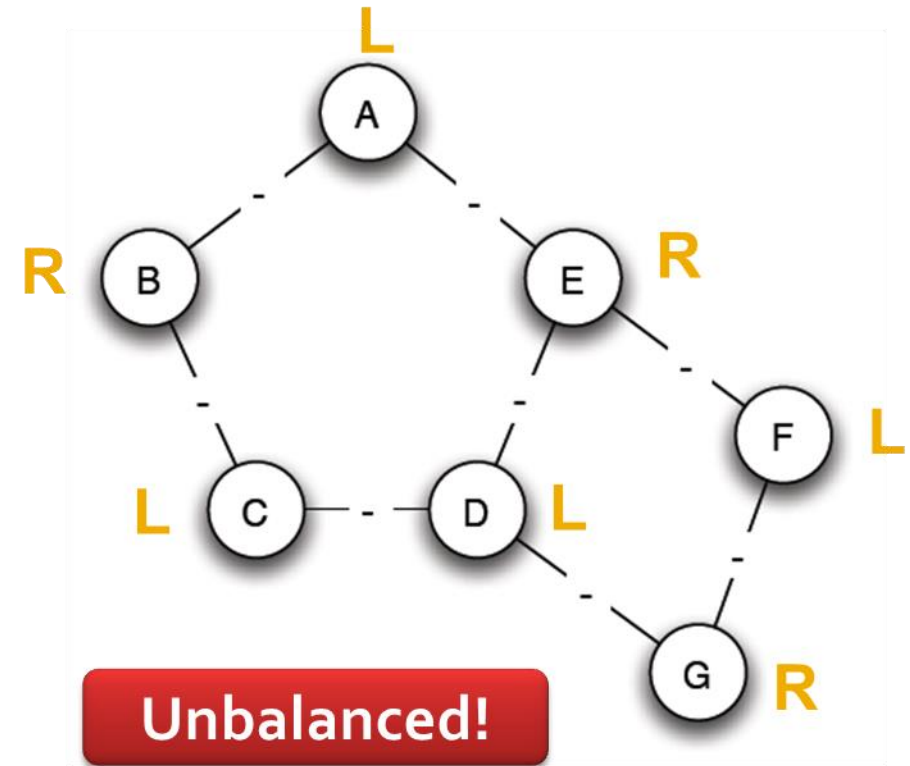
- **Ví dụ:** Kiểm tra tính cân bằng của đồ thị bên.
- 2. Đối với mỗi thành phần, hãy tạo một siêu nút.
- 3. Kết nối các thành phần A và B nếu có một cạnh tiêu cực giữa các thành viên.



5. ĐỒ THỊ VÔ HƯỚNG CÓ DẪU

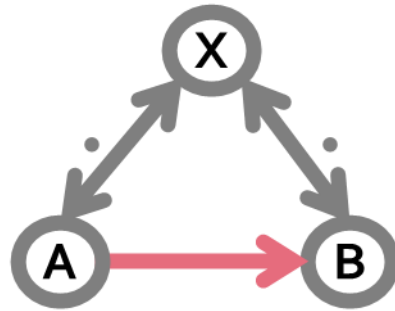
LÝ THUYẾT CÂN BẰNG CẤU TRÚC

- Đồ thị là cân bằng khi và chỉ khi nó không chứa chu trình với số cạnh âm là số lẻ.
- **Ví dụ:** Kiểm tra tính cân bằng của đồ thị bên.
- 4. Gán các siêu nút cho các bên bằng BFS.
- Đồ thị không cân bằng nếu bất kỳ hai siêu nút nào được gán cùng một phía.

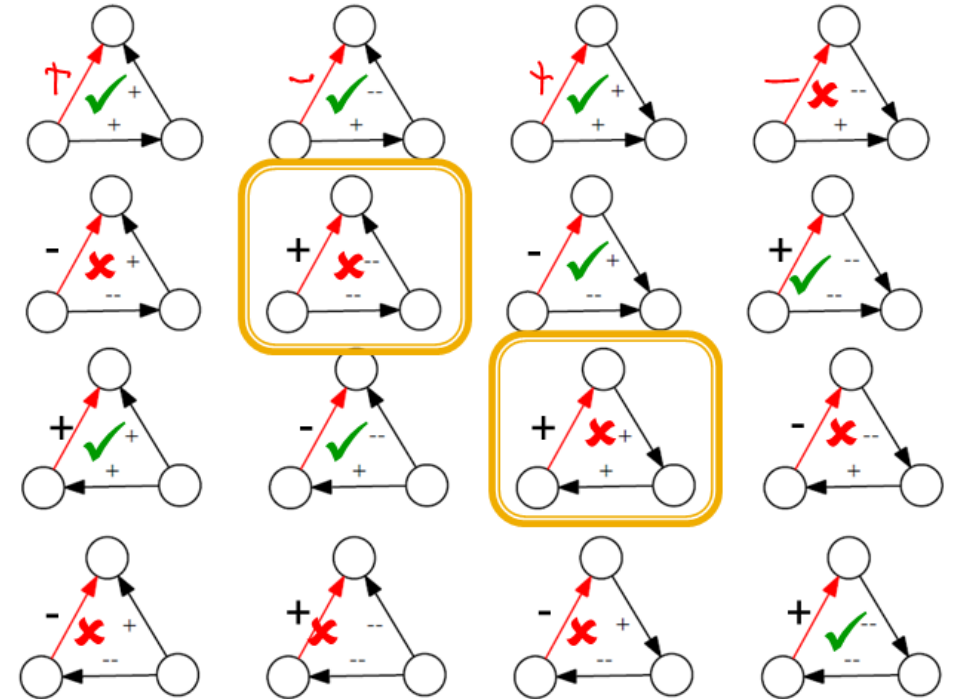


6. ĐỒ THỊ CÓ HƯỚNG CÓ DẤU

- Các liên kết được định hướng và tạo ra theo thời gian.



- Chỉ có 1/2 số tam giác được xem là cân bằng.



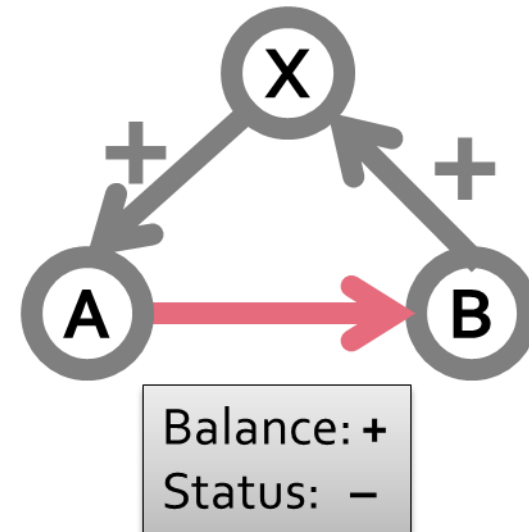
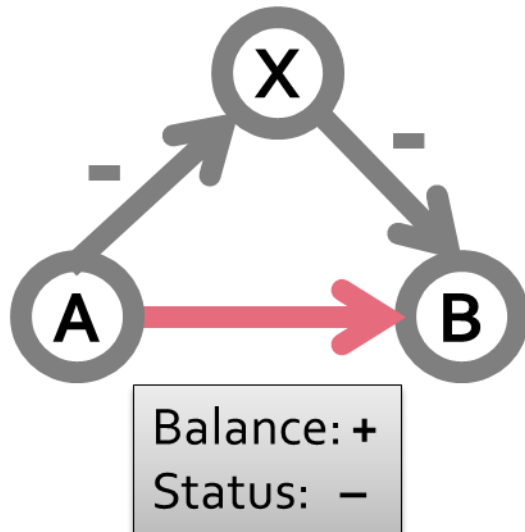
16 * 2 signed directed triads

6. ĐỒ THỊ CÓ HƯỚNG CÓ DẤU

- Các liên kết được định hướng và tạo ra theo thời gian.

LÝ THUYẾT TRẠNG THÁI [Davis-Leinhardt '68, Guha et al. '04, Leskovec et al. '10]

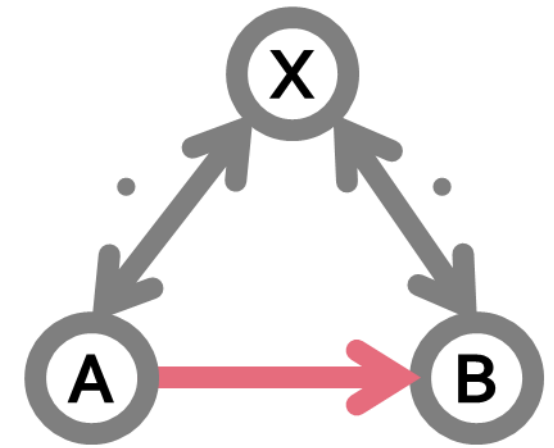
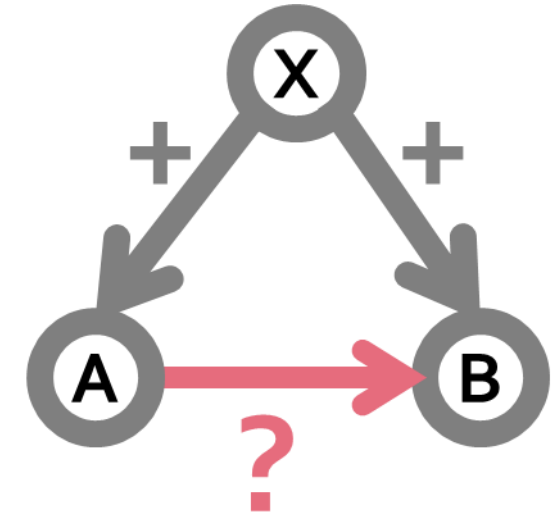
- Liên kết $A \xrightarrow{+} B$ nghĩa là: B có địa vị cao hơn A.
- Liên kết $A \xrightarrow{-} B$ nghĩa là: B có địa vị thấp hơn A.
- Lý thuyết cân bằng cấu trúc và lý thuyết trạng thái đưa ra các dự đoán khác nhau.



6. ĐỒ THỊ CÓ HƯỚNG CÓ DẤU

LÝ THUYẾT TRẠNG THÁI

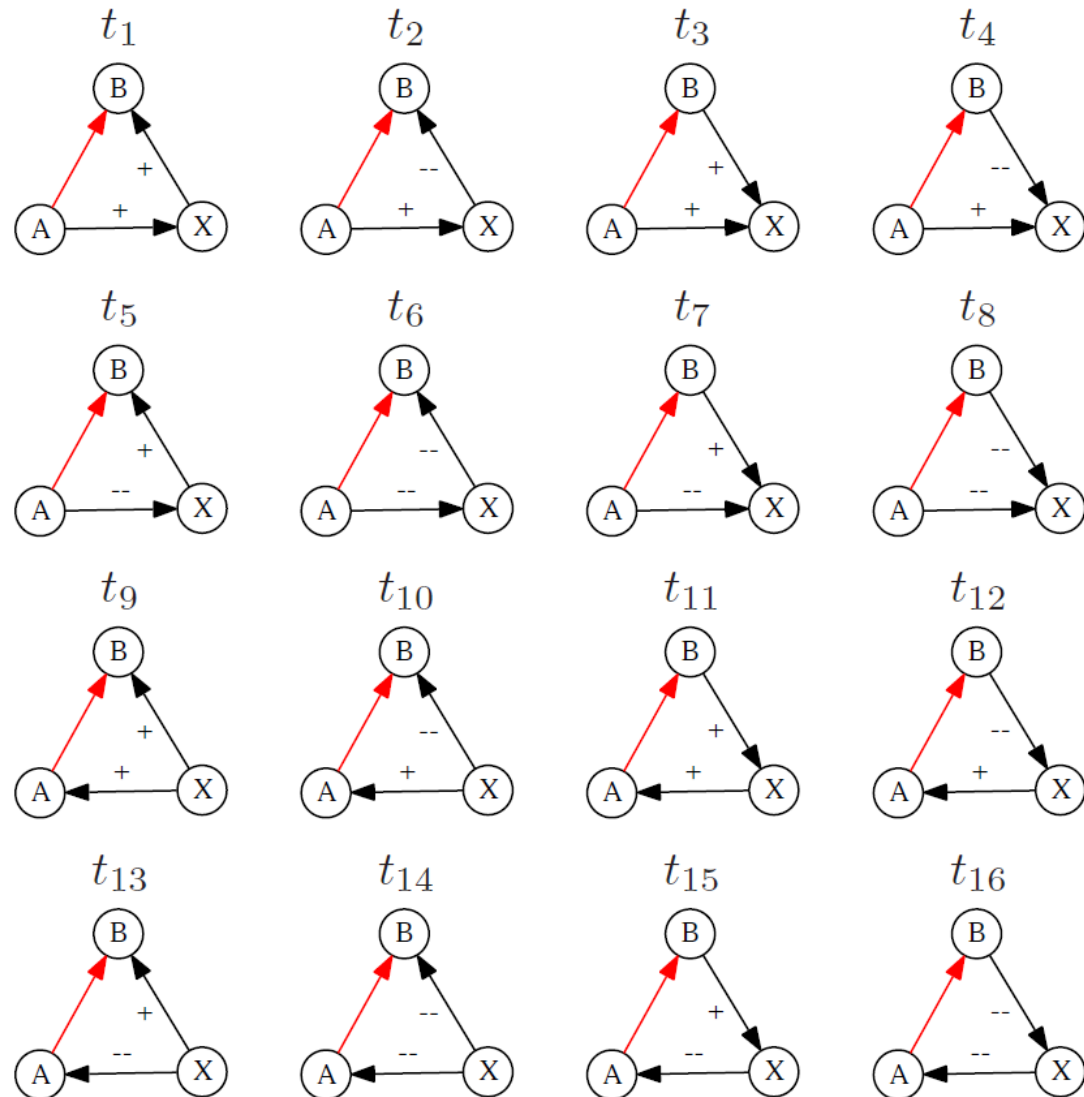
- Các cạnh **có hướng**.
- Các cạnh được **tạo ra theo thời gian**.
 - X có liên kết với A và B.
 - Bây giờ, A liên kết với B (bộ ba A-B-X).
 - Dấu hiệu của A-B phụ thuộc dấu hiệu của X như thế nào?
- Hình thức chung:
 - Các liên kết được nhúng trong bộ ba:
 - Cung cấp ngữ cảnh cho các dấu hiệu.
 - Người dùng không đồng nhất trong hành vi liên kết của họ.



6. ĐỒ THỊ CÓ HƯỚNG CÓ DẤU

LÝ THUYẾT TRẠNG THÁI

- 16 loại bối cảnh
 - Liên kết (A, B) xuất hiện trong ngữ cảnh (A, B; X).
- 16 liên kết được ngữ cảnh hóa khác nhau.



6. ĐỒ THỊ CÓ HƯỚNG CÓ DẦU

LÝ THUYẾT TRẠNG THÁI

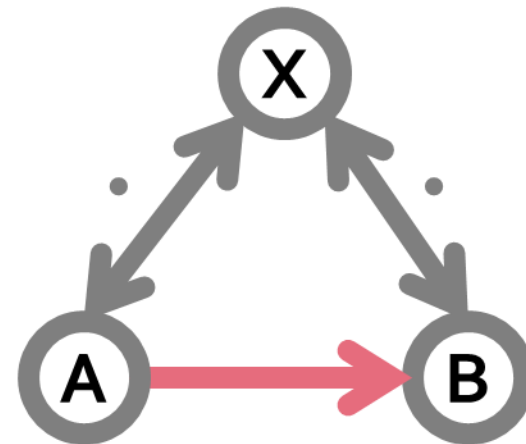
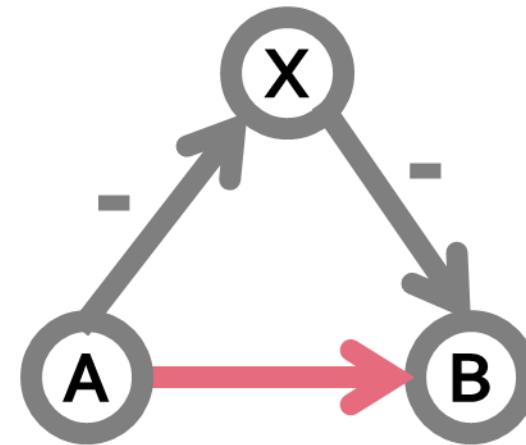
- **Surprise:** Hành vi của người dùng sai lệch bao nhiêu so với đường cơ sở trong ngữ cảnh t :
 - $(A_1, B_1; X_1), \dots, (A_n, B_n; X_n)$ các trường hợp của liên kết được ngữ cảnh hóa t .
 - k trong số trường hợp đã kết thúc bằng một dấu cộng.
 - $p_g(A_i) \dots$ đường cơ sở tổng quát của A_i
- **Generative surprise** của loại bộ ba t :

$$s_g(t) = \frac{k - \sum_{i=1}^n p_g(A_i)}{\sqrt{\sum_i^n p_g(A_i)(1 - p_g(A_i))}}$$

6. ĐỒ THỊ CÓ HƯỚNG CÓ DẦU

LÝ THUYẾT TRẠNG THÁI

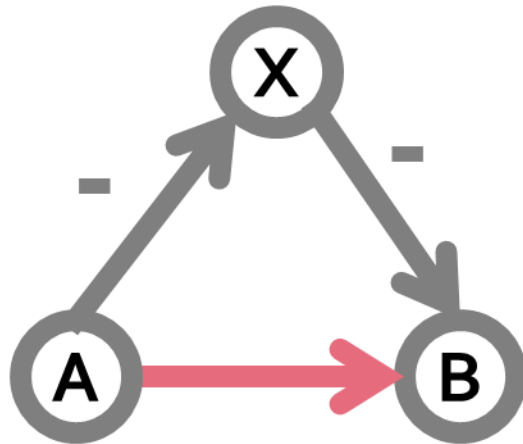
- Giải thích rõ hơn:
 - 1) Đối với mỗi nút tính toán đường cơ sở
 - 2) Xác định tất cả các cạnh đóng cùng một loại bộ ba
 - 3) Tính toán **Surprise**



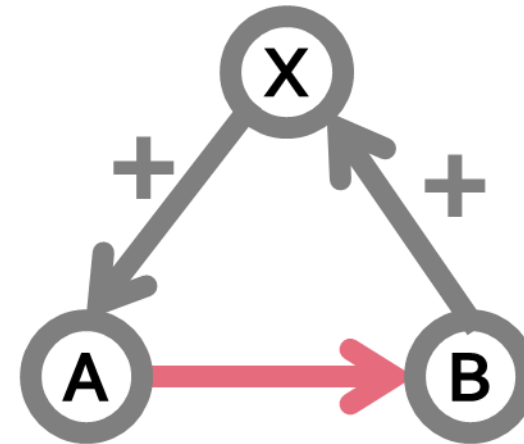
6. ĐỒ THỊ CÓ HƯỚNG CÓ DẦU

LÝ THUYẾT TRẠNG THÁI

- Hai ví dụ cơ bản:



Gen. surprise of A: —
Rec. surprise of B: —



Gen. surprise of A: —
Rec. surprise of B: —

7. ĐỘ TƯƠNG ĐỒNG CỤC BỘ

- **Different scoring functions** $c(x, y) =$
 - **Graph distance:** (negated) Shortest path length
 - **Common neighbors:** $|\Gamma(x) \cap \Gamma(y)|$
 - **Jaccard's coefficient:** $|\Gamma(x) \cap \Gamma(y)| / |\Gamma(x) \cup \Gamma(y)|$
 - **Adamic/Adar:** $\sum_{z \in \Gamma(x) \cap \Gamma(y)} 1 / \log |\Gamma(z)|$
 - **Preferential attachment:** $|\Gamma(x)| \cdot |\Gamma(y)|$ $\Gamma(x) \dots$ neighbors of node x
 - **PageRank:** $r_x(y) + r_y(x)$
 - $r_x(y)$... stationary distribution score of y under the random walk:
 - with prob. 0.15, jump to x
 - with prob. 0.85, go to random neighbor of current node
- **Then, for a particular choice of $c(\cdot)$**
 - For every pair of nodes (x, y) compute $c(x, y)$
 - Sort pairs (x, y) by the decreasing score $c(x, y)$
 - **Predict top n pairs as new links**



Q & A