Khai Thác Dữ Liệu Đồ Thị

GIỚI THIỆU MÔN HỌC

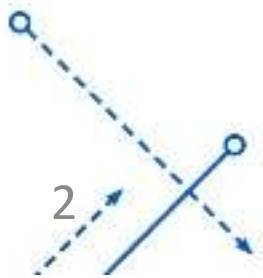
Giảng viên: Lê Ngọc Thành

Email: lnthanh@fit.hcmus.edu.vn

fit@hcmus

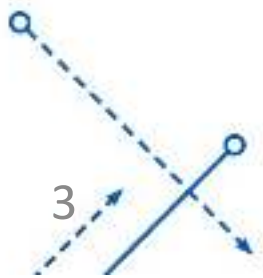
Nội Dung

- **Giới thiệu môn học và các chủ đề**
- Qui định
- Mạng và đồ thị
- Khai thác đồ thị
- Bài toán và ứng dụng



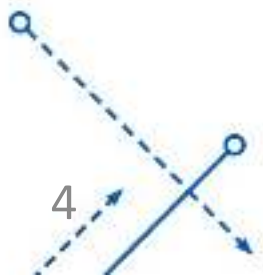
Về môn học

- Tên môn: Khai thác dữ liệu đồ thị
  (Mining Graph Data)
- LT/TH: 45/30 tiết.
- Tỉ lệ nghe giảng và tự học: 40/60.
- Tài liệu tham khảo:
  - Slide bài giảng
  - Aggarwal, Charu C., and Haixun Wang, eds. Managing and mining graph data. Vol. 40. New York: Springer, 2010.
  - Easley, David, and Jon Kleinberg. "Networks, crowds, and markets: Reasoning about a highly connected world." Significance 9 (2012): 43-44.
  - Ketmaneechairat, Hathairat. "Graph Mining Laws, Tools and Case Studies." Journal of Digital Information Management 12, no. 6 (2014): 446
  - …



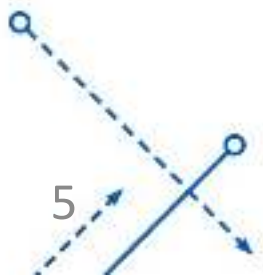
Chủ đề môn học

Tuần	Tên chủ đề
1	Giới thiệu Khai thác dữ liệu đồ thị, các thuật toán và ứng dụng
2	Ôn lại các kiến thức liên quan đến đồ thị và khai thác dữ liệu
3	Mẫu trong đồ thị tĩnh và động; phát sinh đồ thị.
4	Đánh chỉ mục đồ thị và xếp hạng
5	Khai thác mẫu đồ thị
6	Phân lớp đồ thị



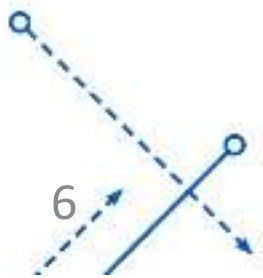
Chủ đề môn học

Tuần	Tên chủ đề
7	Gom nhóm và phát hiện cộng đồng
8	Dự đoán liên kết
9	Nhúng đồ thị
10-15	Chủ đề seminar bao gồm: học sâu cho đồ thị, tóm tắt đồ thị, hệ thống tư vấn, phát hiện bất thường, đồ thị kích thước lớn.



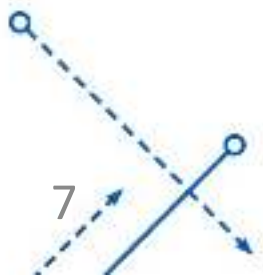
Nội Dung

- Giới thiệu môn học và các chủ đề
- **Qui định**
- Mạng và đồ thị
- Khai thác đồ thị
- Bài toán và ứng dụng



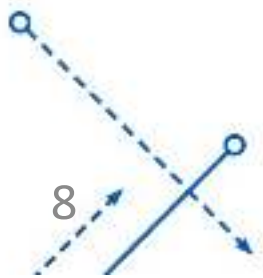
Về quy định và đánh giá học tập

- Lí thuyết: 40%
  - Thi cuối kì: 40%
- Thực hành: 30%
- Đồ án: 30%
- Các quy định:
  - Gian lận, copy bài, cho bạn chép → miễn thi LT
  - Các hành động gây phá hoại lớp học → miễn học và thi LT



Nội Dung

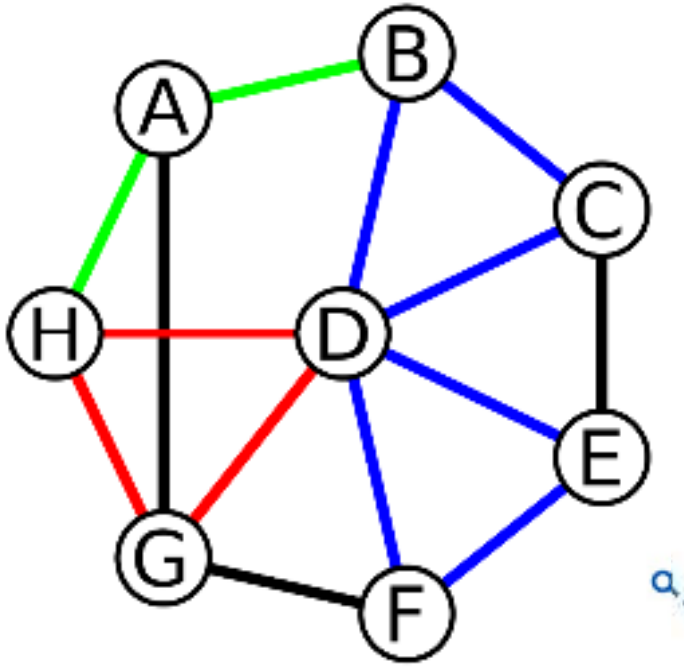
- Giới thiệu môn học và các chủ đề
- Qui định
- **Mạng và đồ thị**
- Khai thác đồ thị
- Bài toán và ứng dụng



Mạng và đồ thị

- Mạng (network) thường được sử dụng để thể hiện mối quan hệ tự nhiên của đối tượng trong thế giới thật.
- Trong khi đó, đồ thị (graph) thể hiện mối quan hệ được phát sinh qua tiến trình tự động.

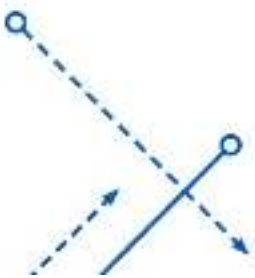
Mạng xã hội



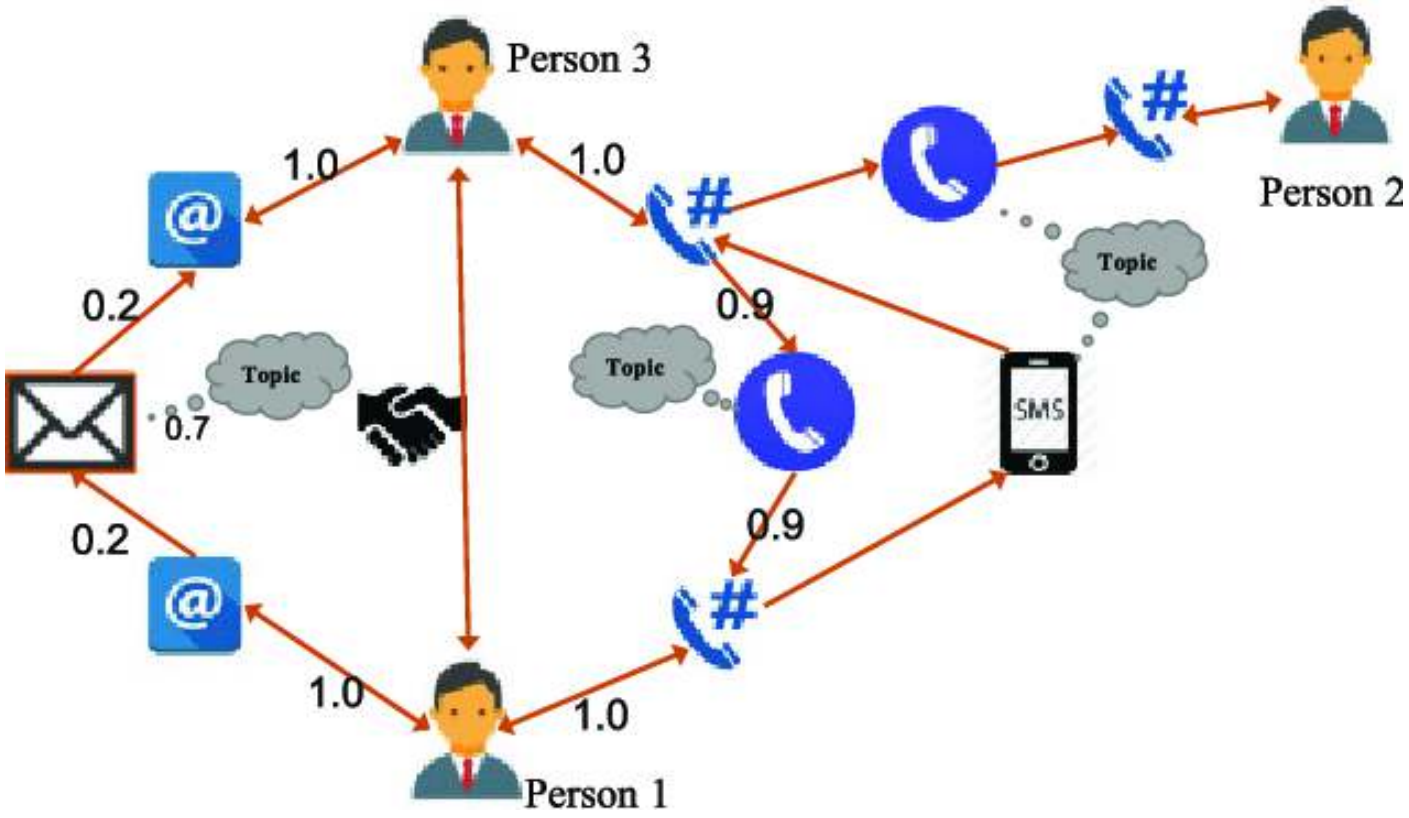
Đồ thị

Mạng xã hội

- Đỉnh: People
- Cạnh: Friendships

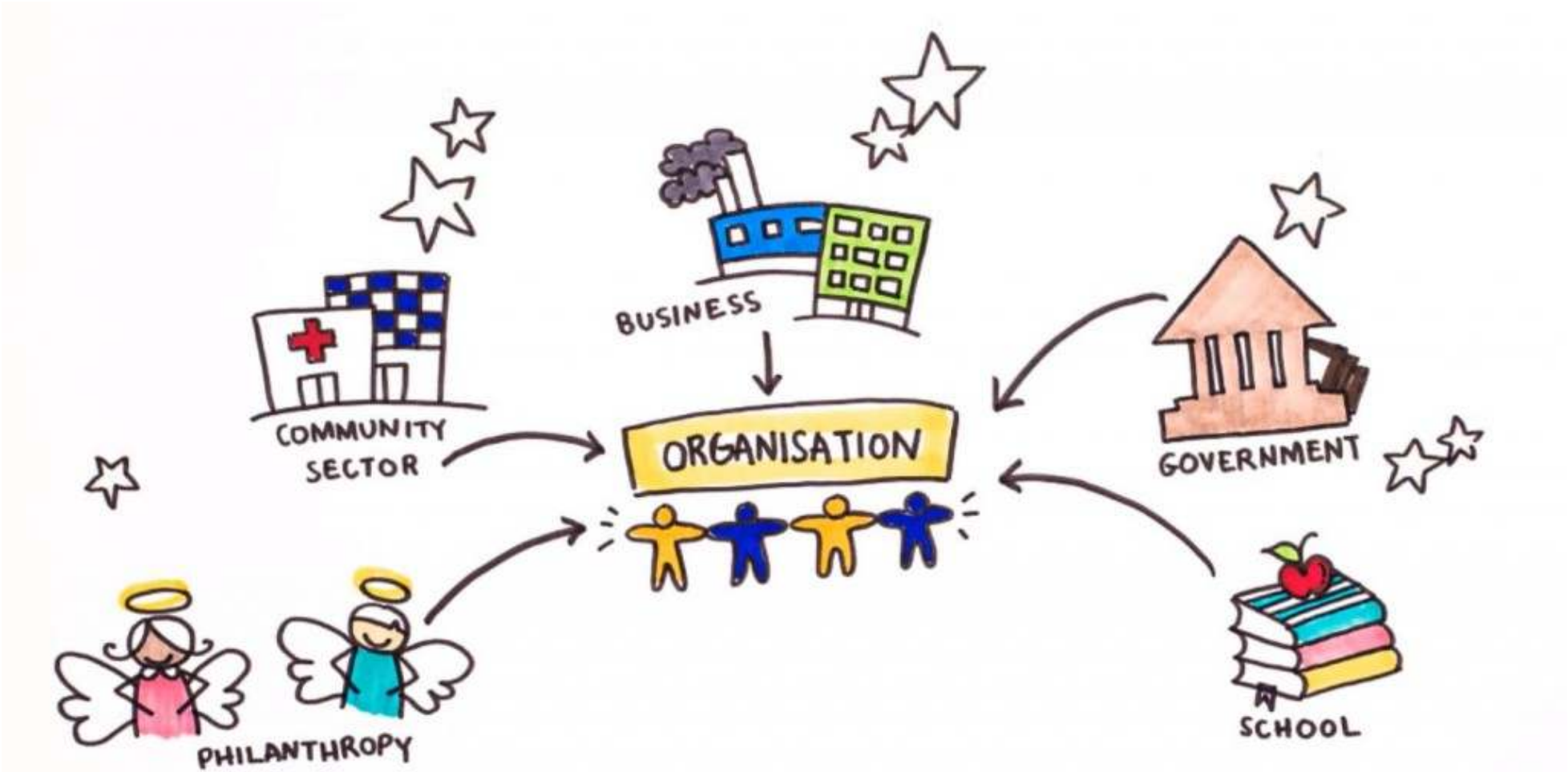


Mạng giao tiếp



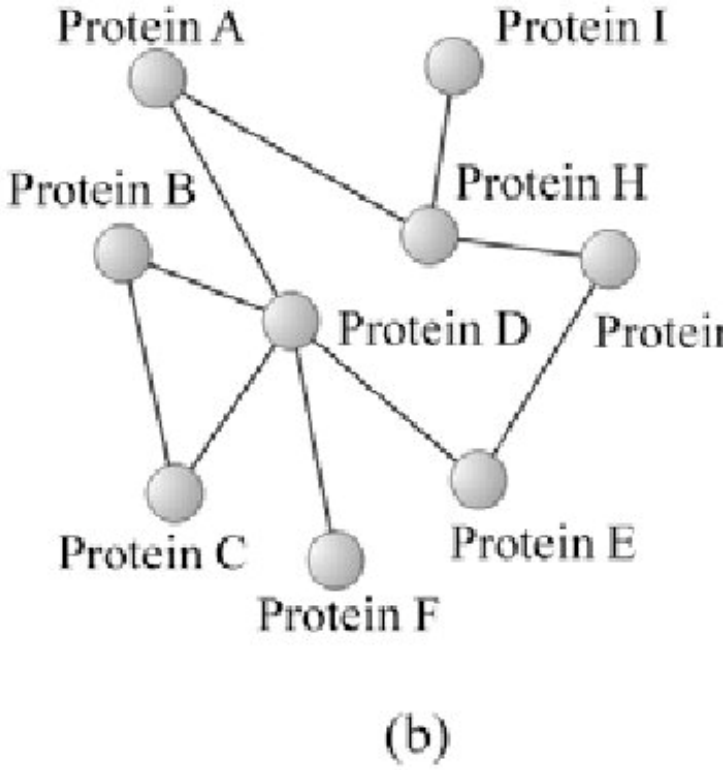
- Đỉnh: People
- Cạnh: email exchange

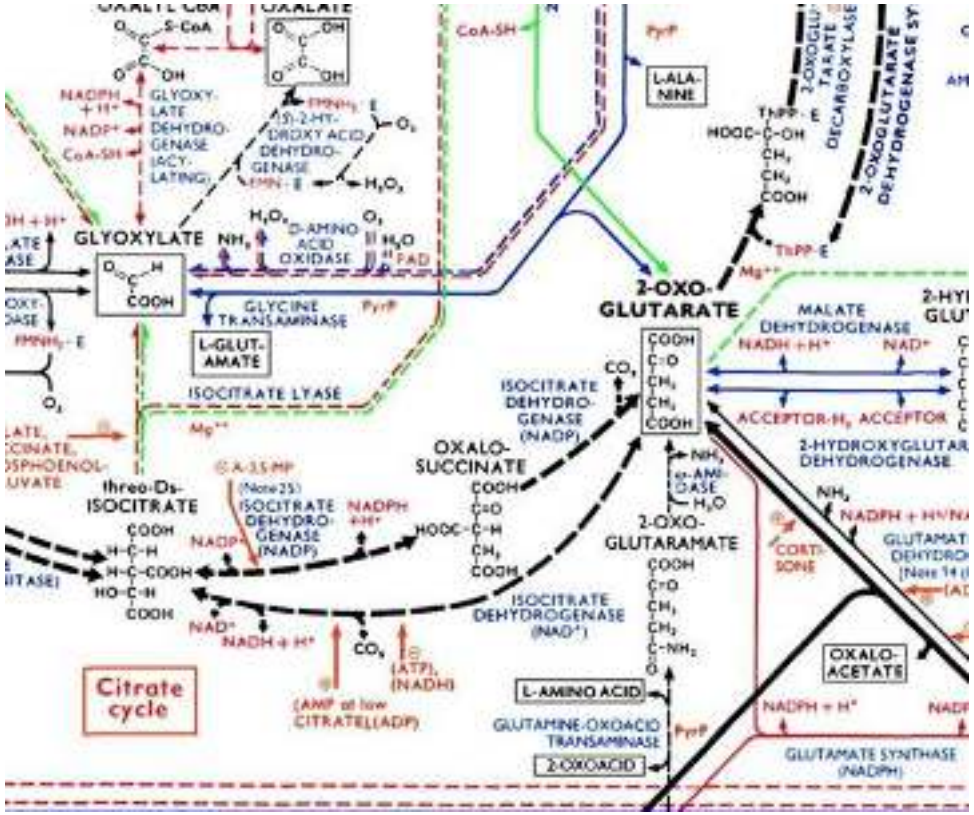
Mạng tài chính



- Đỉnh: Companies
- Cạnh: relationships (financial, collaboration)

Mạng sinh học

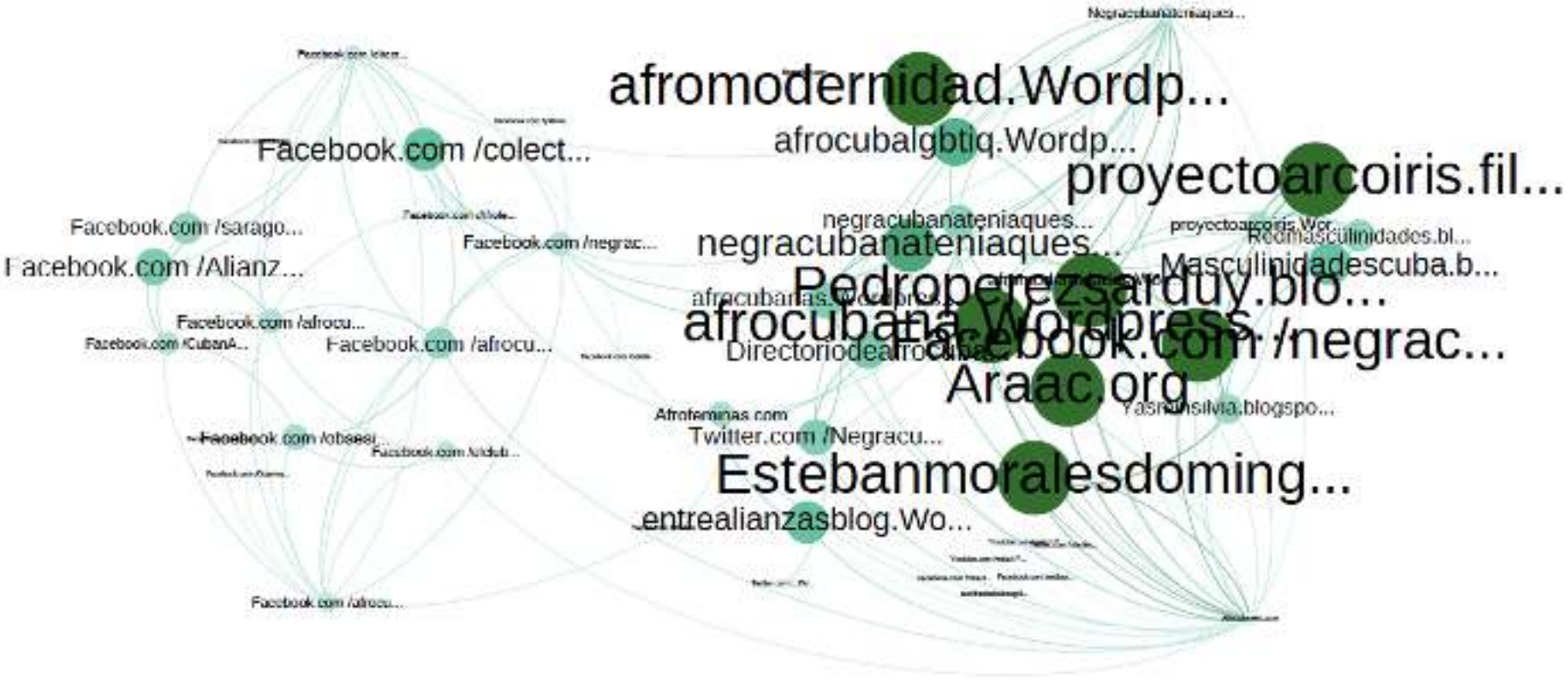




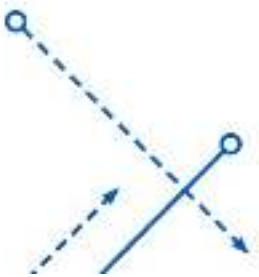
- Đỉnh: Proteins
- Cạnh: interactions

- Đỉnh: metabolites, enzymes
- Cạnh: chemical reactions

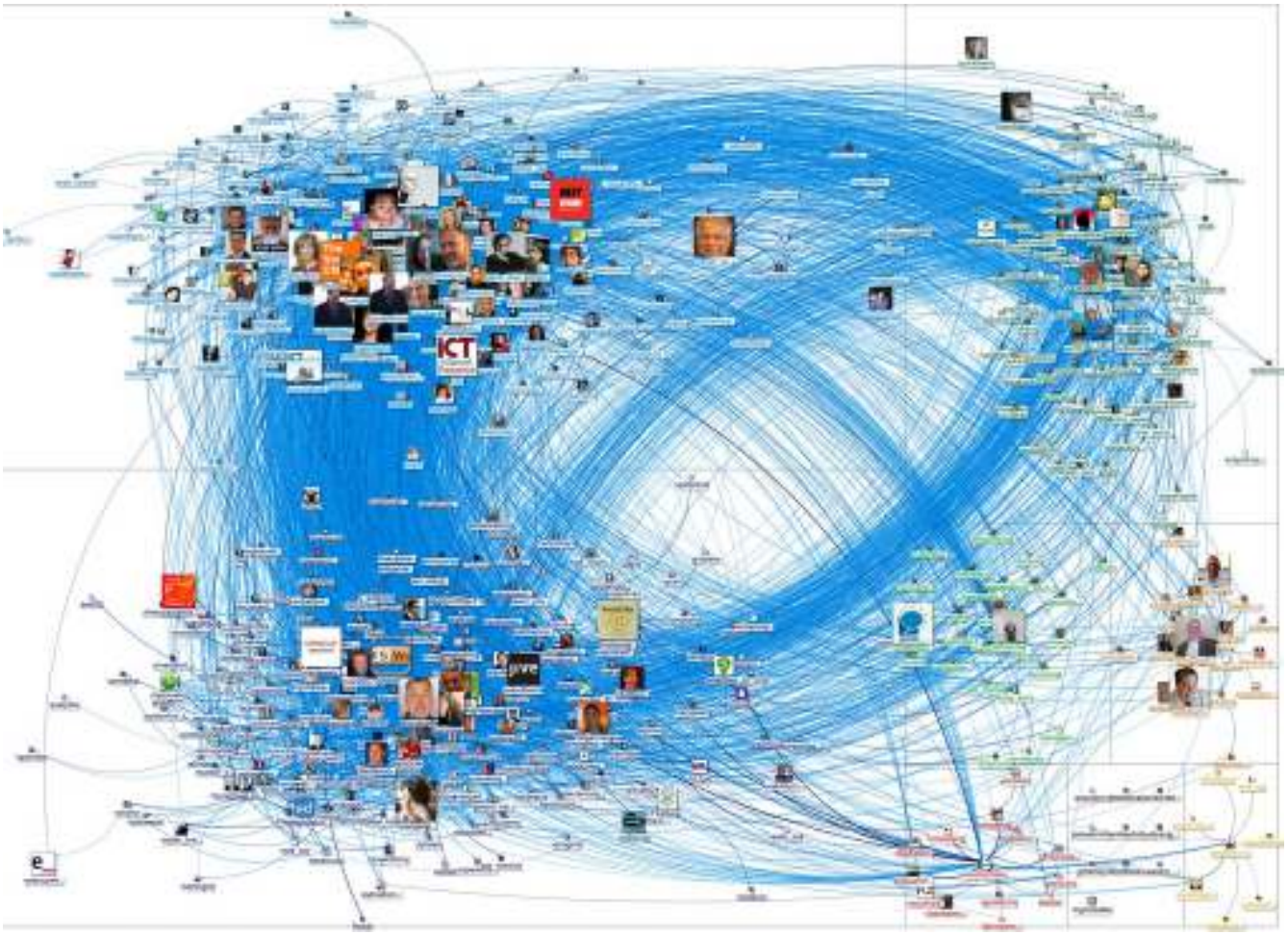
Mạng thông tin (WWW)



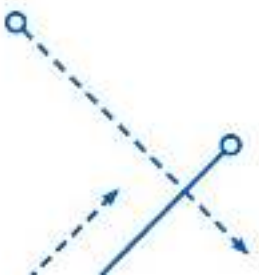
- Đỉnh: Web Pages
- Cạnh: Links



Mạng truyền thông

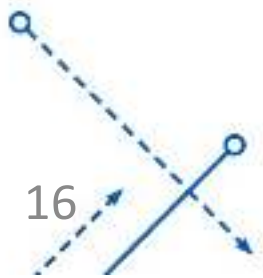


- Đỉnh: Twitter users
- Cạnh: Follows/conversations



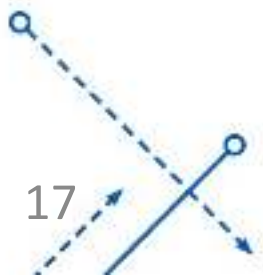
Nội Dung

- Giới thiệu môn học và các chủ đề
- Qui định
- Mạng và đồ thị
- **Khai thác đồ thị**
- Bài toán và ứng dụng



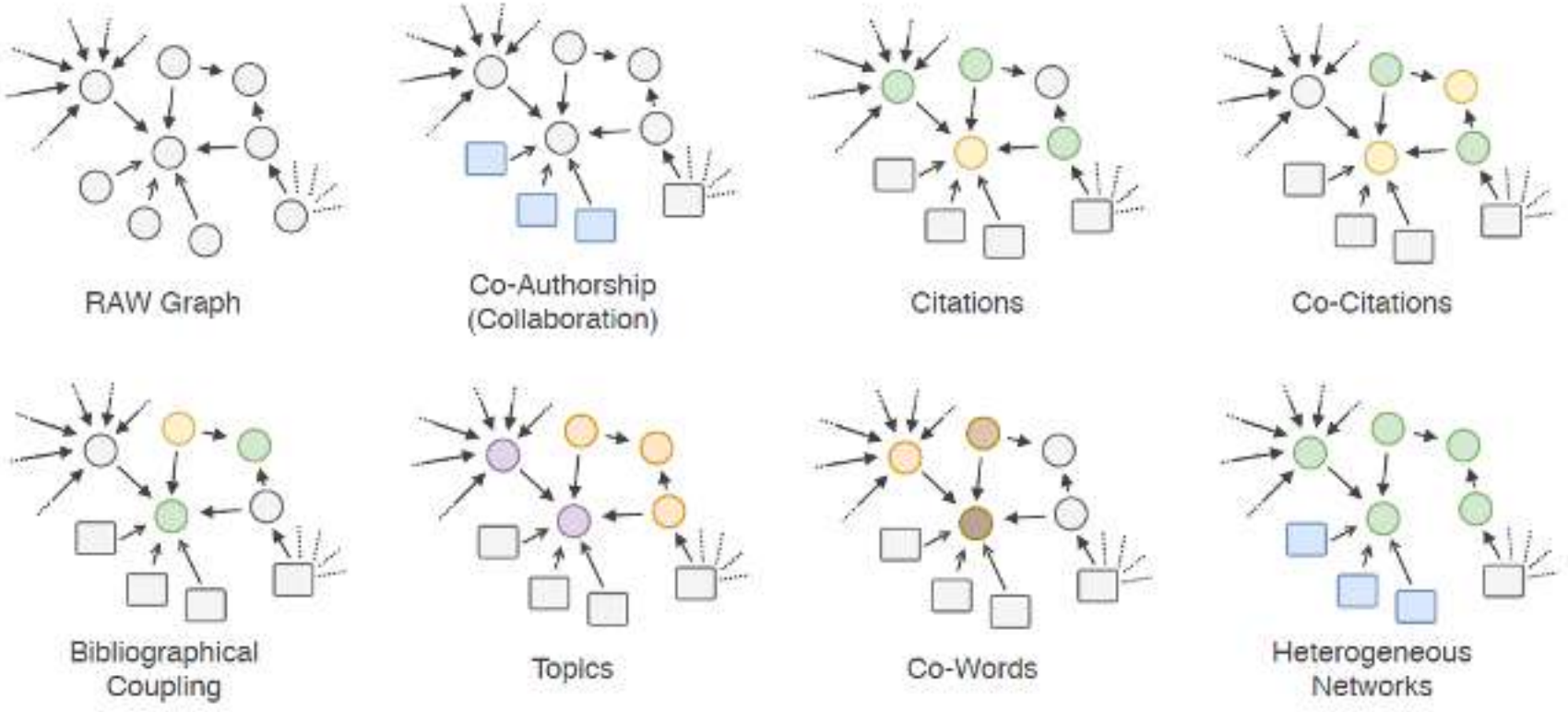
Tại sao phân tích mạng lại quan trọng?

- Hệ thống được kết nối bởi rất nhiều thành phần, nếu chúng ta chỉ tập trung hiểu cá nhân đơn lẻ thì không thể nắm bắt được cả hệ thống.
- Có 2 câu hỏi lớn:
  - Thuộc tính cấu trúc của mạng là gì?
  - Tiến trình tương tác gì đang xảy ra trong mạng?



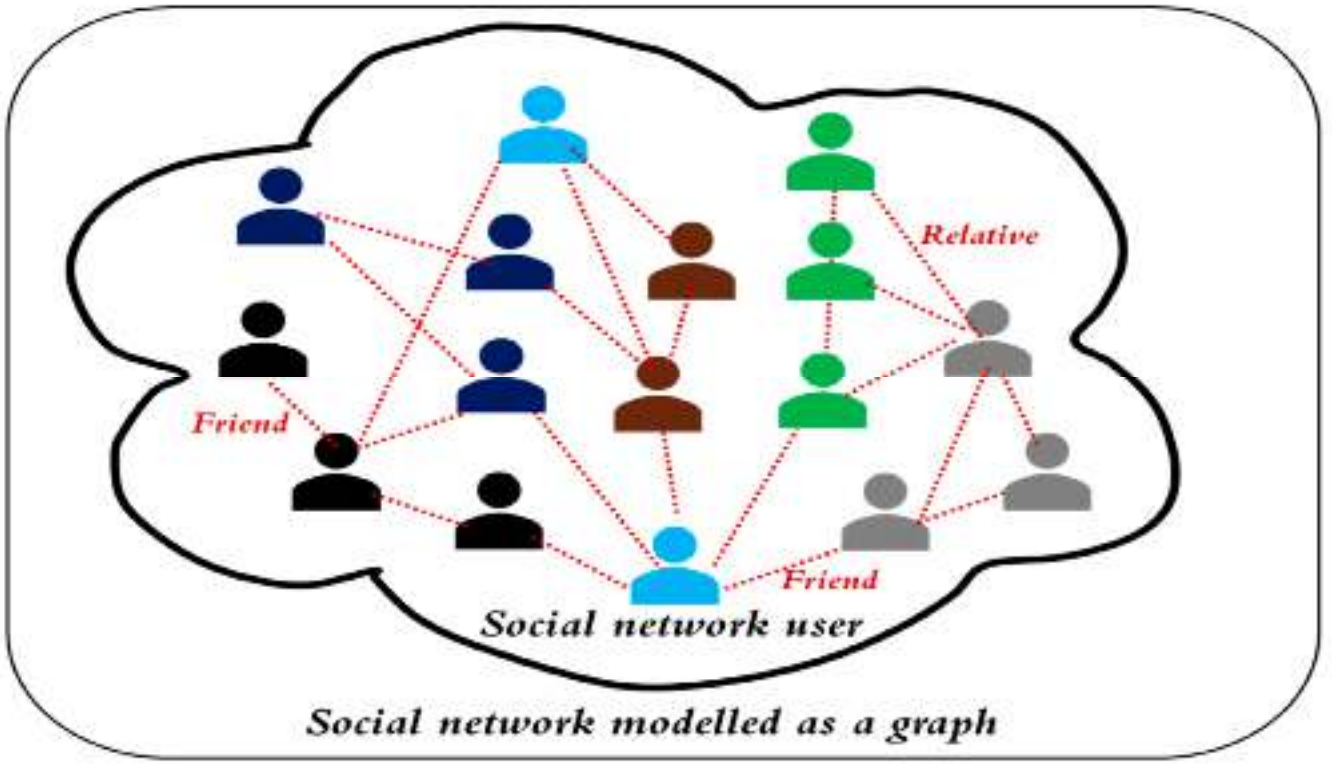
Nghiên cứu trong mạng

- Trong lĩnh vực phân tích mạng, người ta tập trung nghiên cứu các hành xử trong mạng như hành xử giữa người với người trong mạng xã hội.
  - Dự đoán các cư xử dựa trên các thuộc tính có thể đo của nó.



Mô hình hóa mạng bằng đồ thị

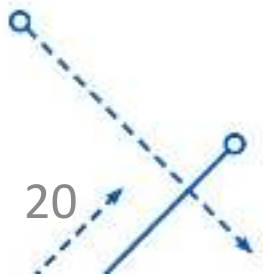
- Mạng không nằm tách rời với đồ thị mà chúng có thể mô hình hóa lại dưới dạng đồ thị và tận dụng các nền tảng lý thuyết của nó.



Social network modelled as a graph

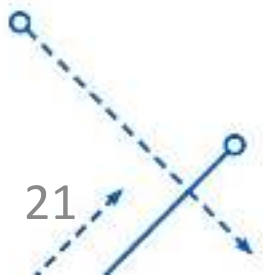
Các thách thức khai thác mạng

- Đồ thị thường:
  - Kích thước lớn, rất rất lớn (massive)
  - Quá thưa (sparsity)/ quá dày đặc (density)
  - Đường kính nhỏ (small diameter)
  - Động (dynamic)
- Đòi hỏi các thuật toán hiệu quả về lưu trữ và tính toán.



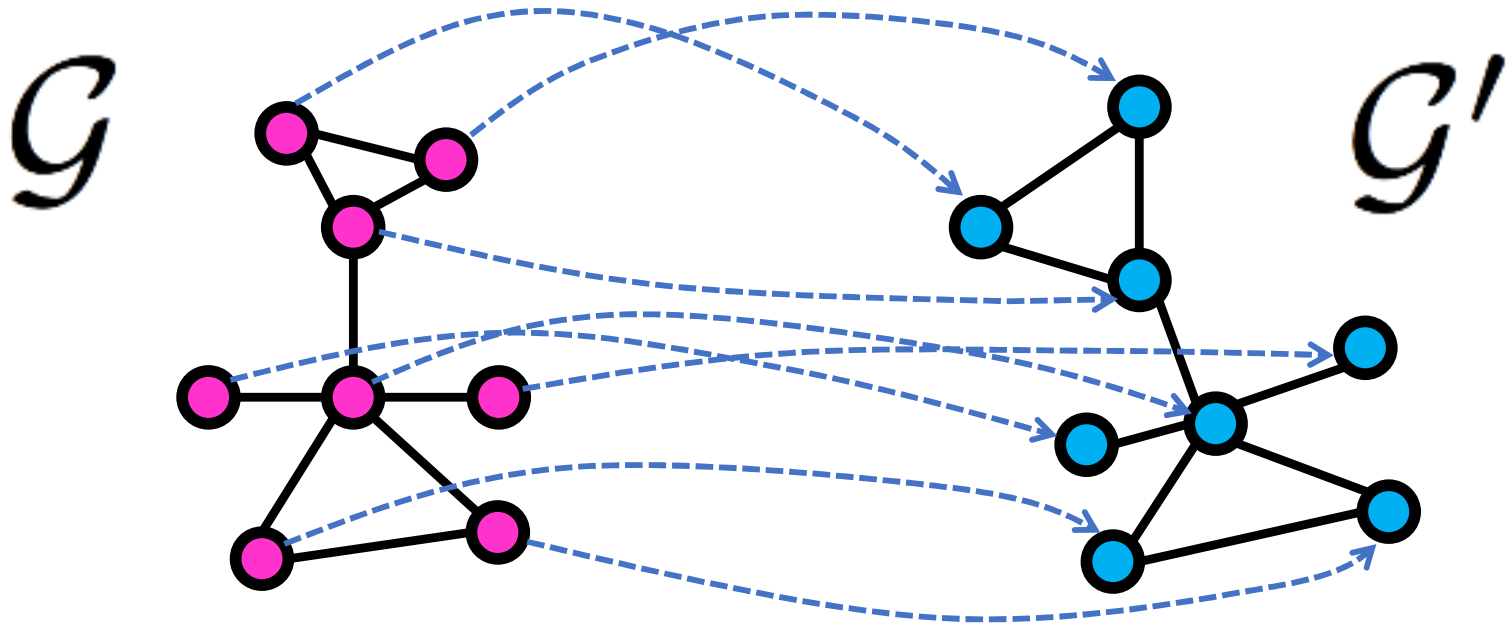
Nội Dung

- Giới thiệu môn học và các chủ đề
- Qui định
- Mạng và đồ thị
- Khai thác đồ thị
- **Bài toán và ứng dụng**



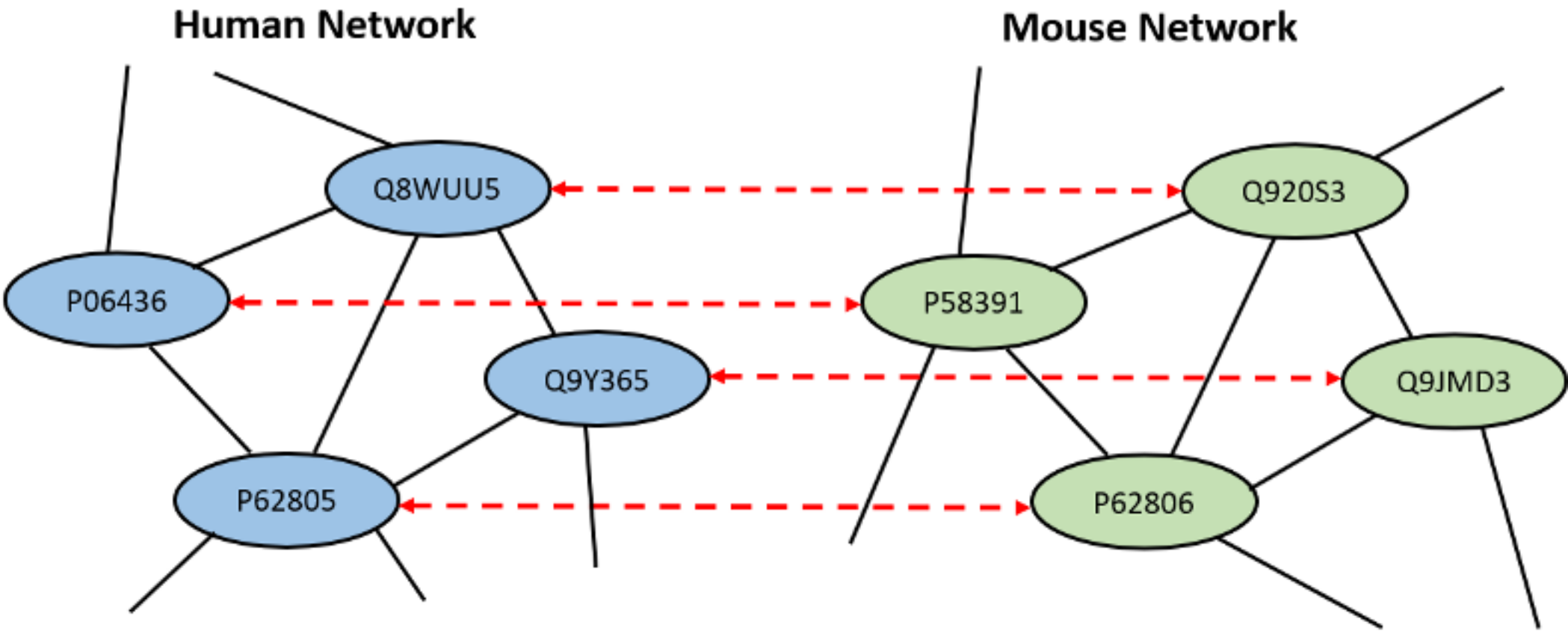
Các bài toán và ứng dụng

- **So khớp đồ thị** (graph matching)



Các bài toán và ứng dụng

- **So khớp đồ thị** (graph matching)



Tìm protein cùng chức năng trên các loài khác nhau dựa trên mạng tương tác giữa chúng

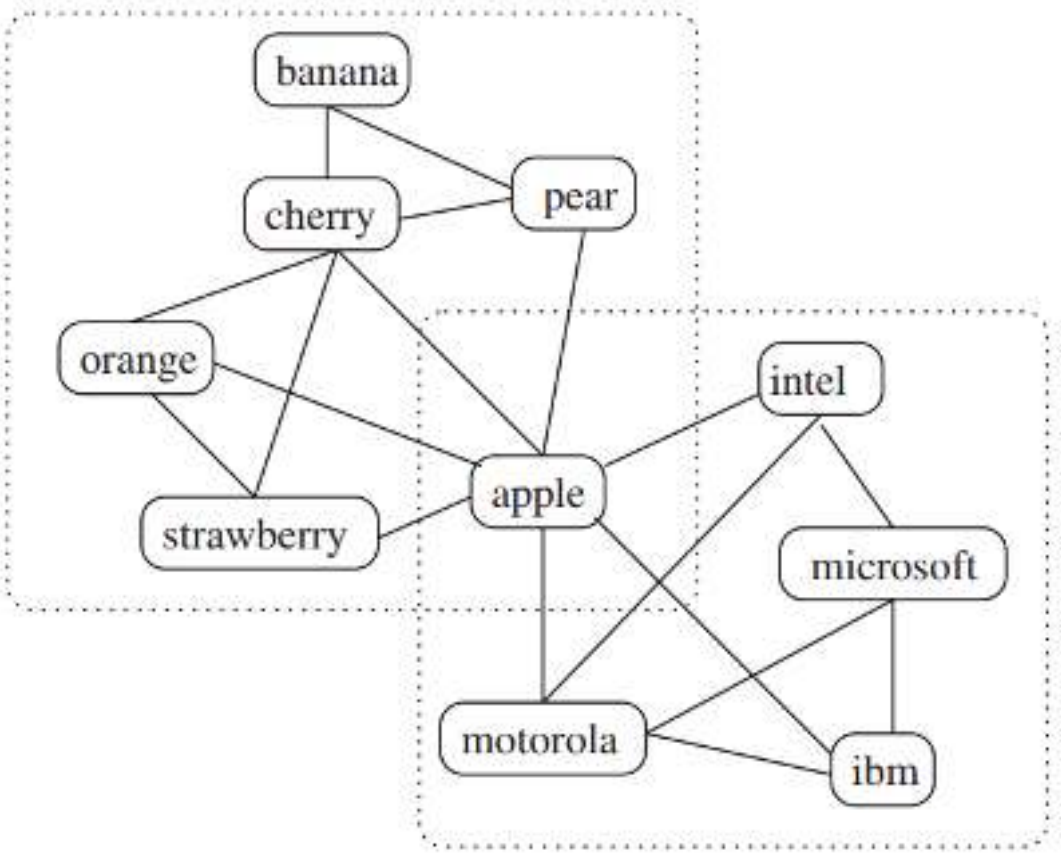
Các bài toán và ứng dụng

- **So khớp đồ thị** (graph matching)

Sử dụng đồ hình trong một mạng để xác định danh tính bị ẩn trong một mạng xã hội

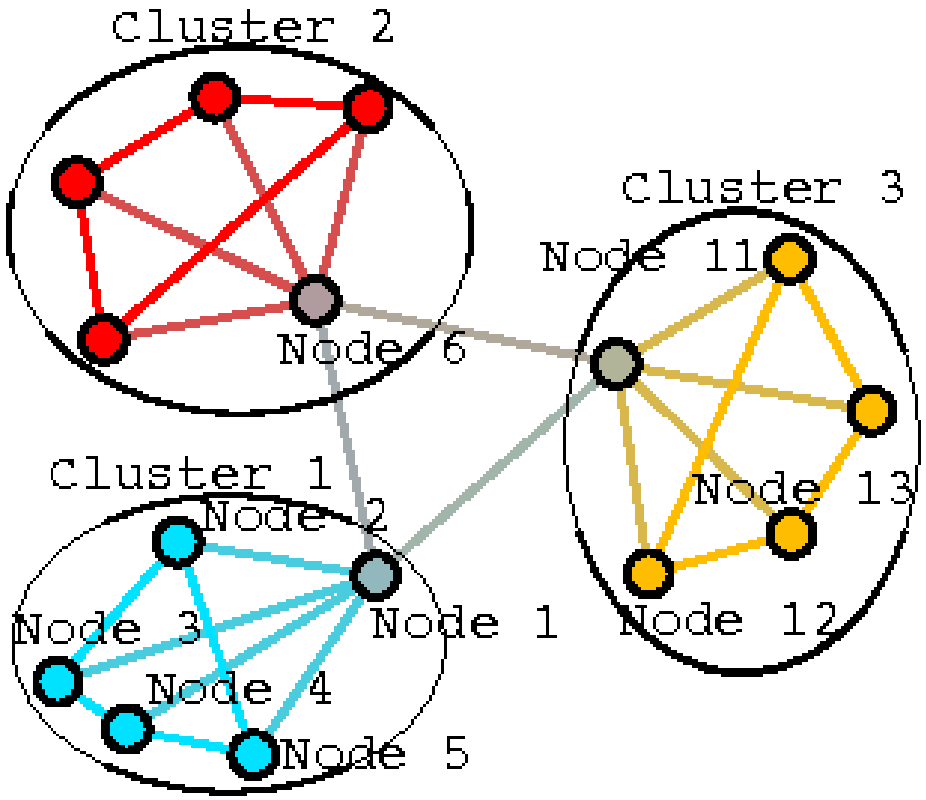
Các bài toán và ứng dụng

- **Xác định ngữ nghĩa** (semantic class)



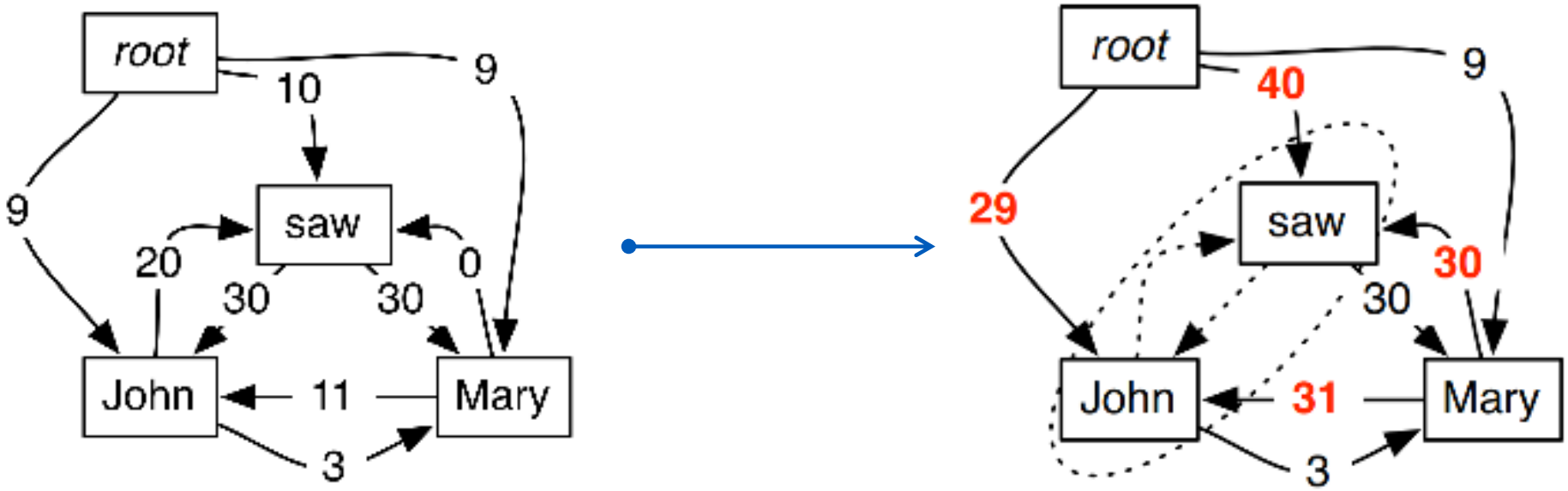
Các bài toán và ứng dụng

- **Gom nhóm đồ thị** (graph clustering)



Các bài toán và ứng dụng

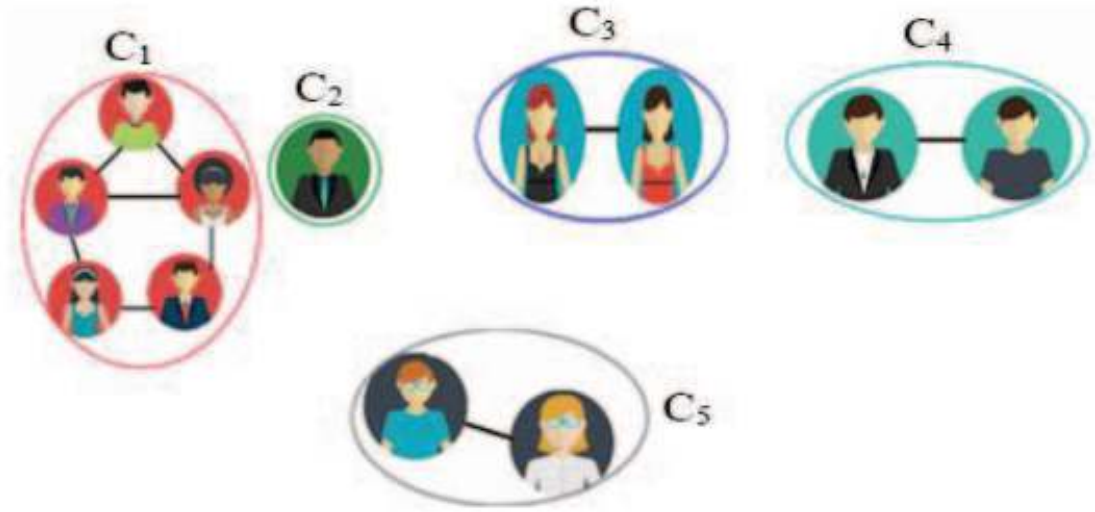
- **Gom nhóm đồ thị** (graph clustering)



Gom nhóm để giảm độ phức tạp của đồ thị

Các bài toán và ứng dụng

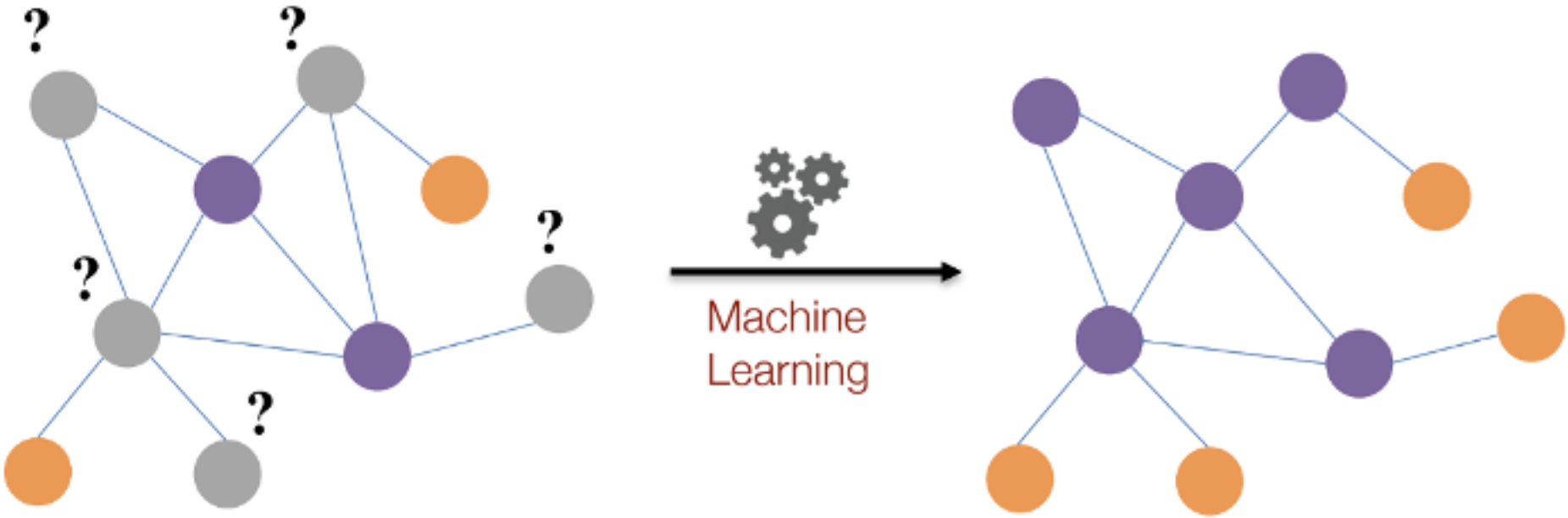
- **Gom nhóm đồ thị** (graph clustering)



Phát hiện cộng đồng

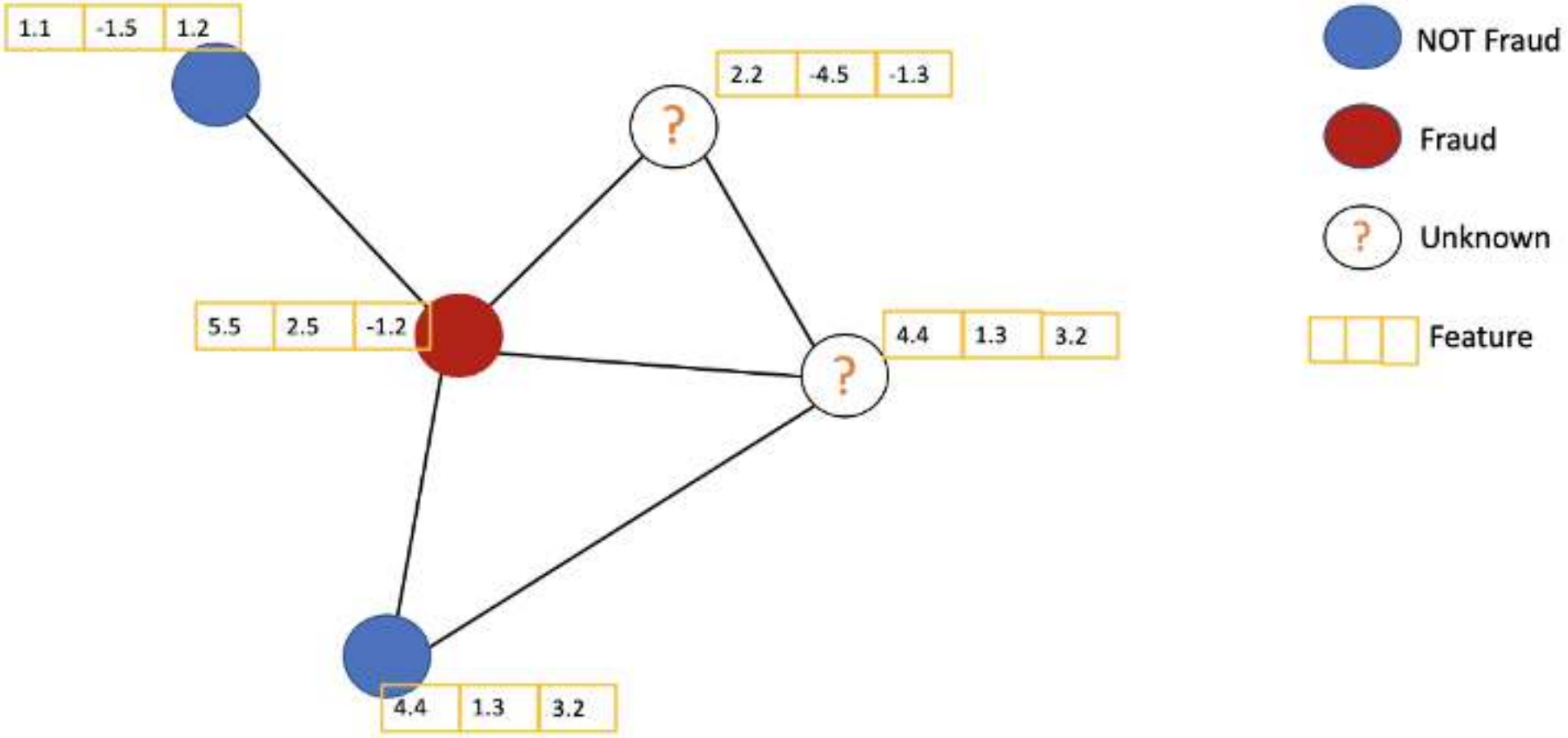
Các bài toán và ứng dụng

- **Phân lớp đồ thị** (graph classification)
  - Đánh nhãn đỉnh
  - Đánh nhãn liên kết
  - Đánh nhãn đồ thị/đồ thị con



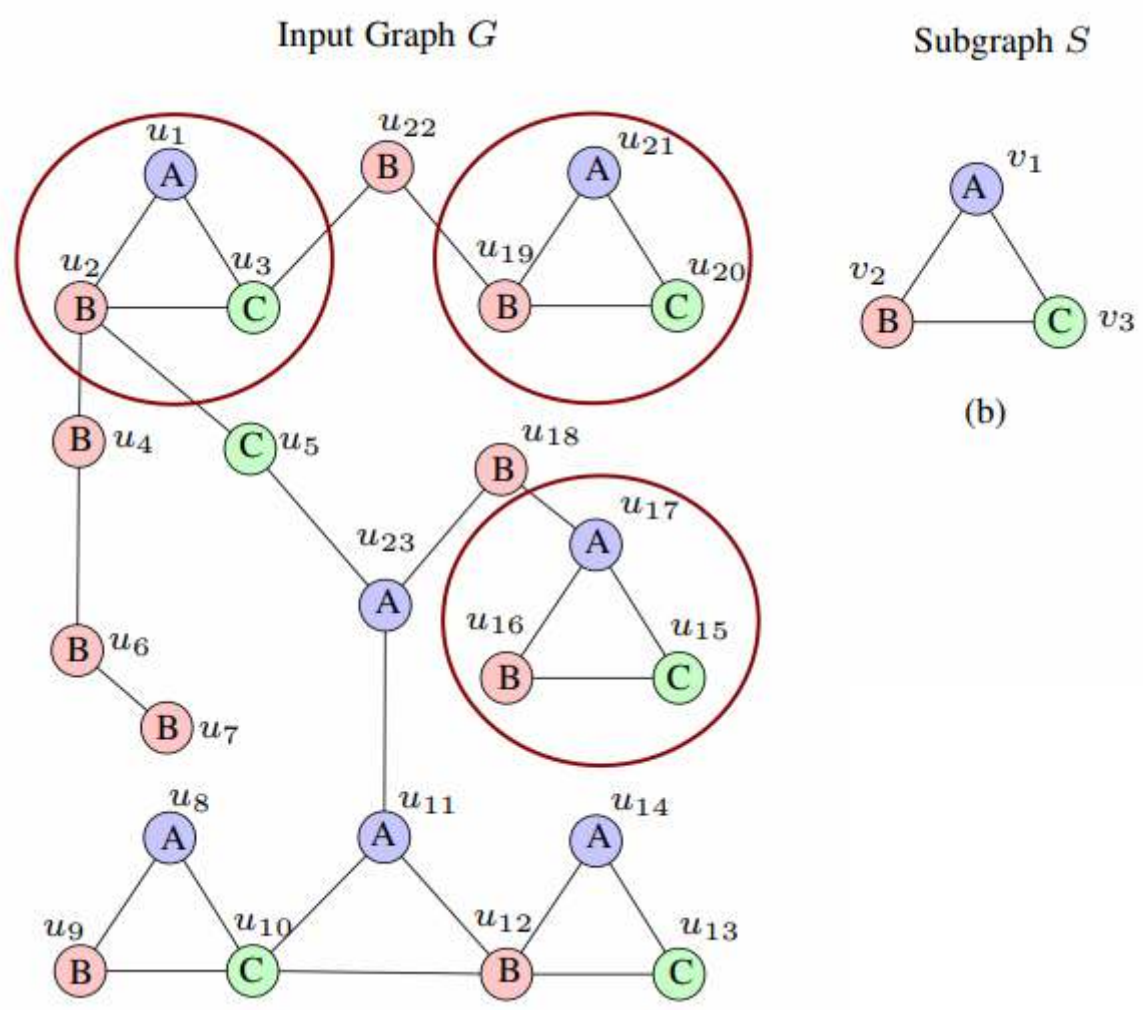
Các bài toán và ứng dụng

- **Phân lớp đồ thị** (graph classification)



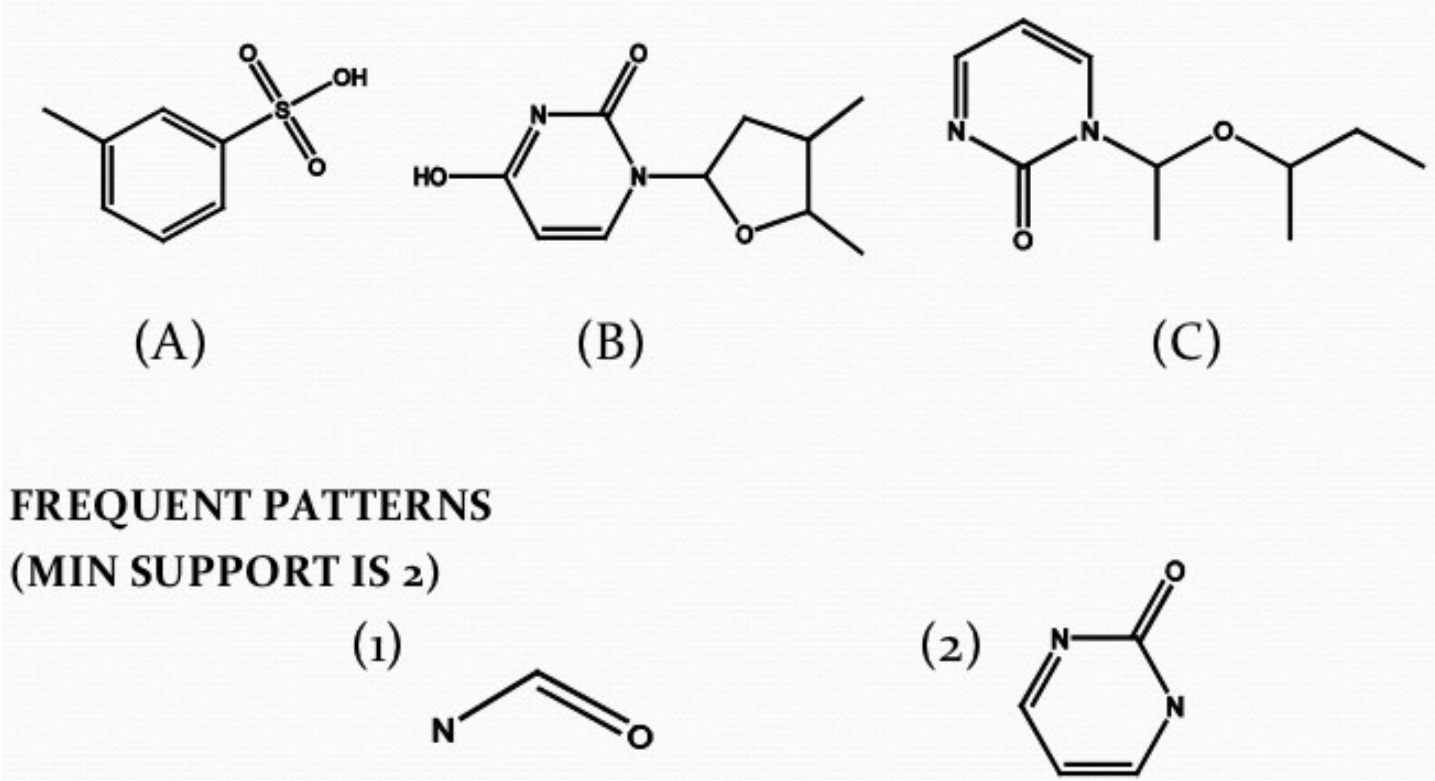
Các bài toán và ứng dụng

- **Khai thác mẫu phổ biến trong đồ thị** (frequent pattern mining in graph)



Các bài toán và ứng dụng

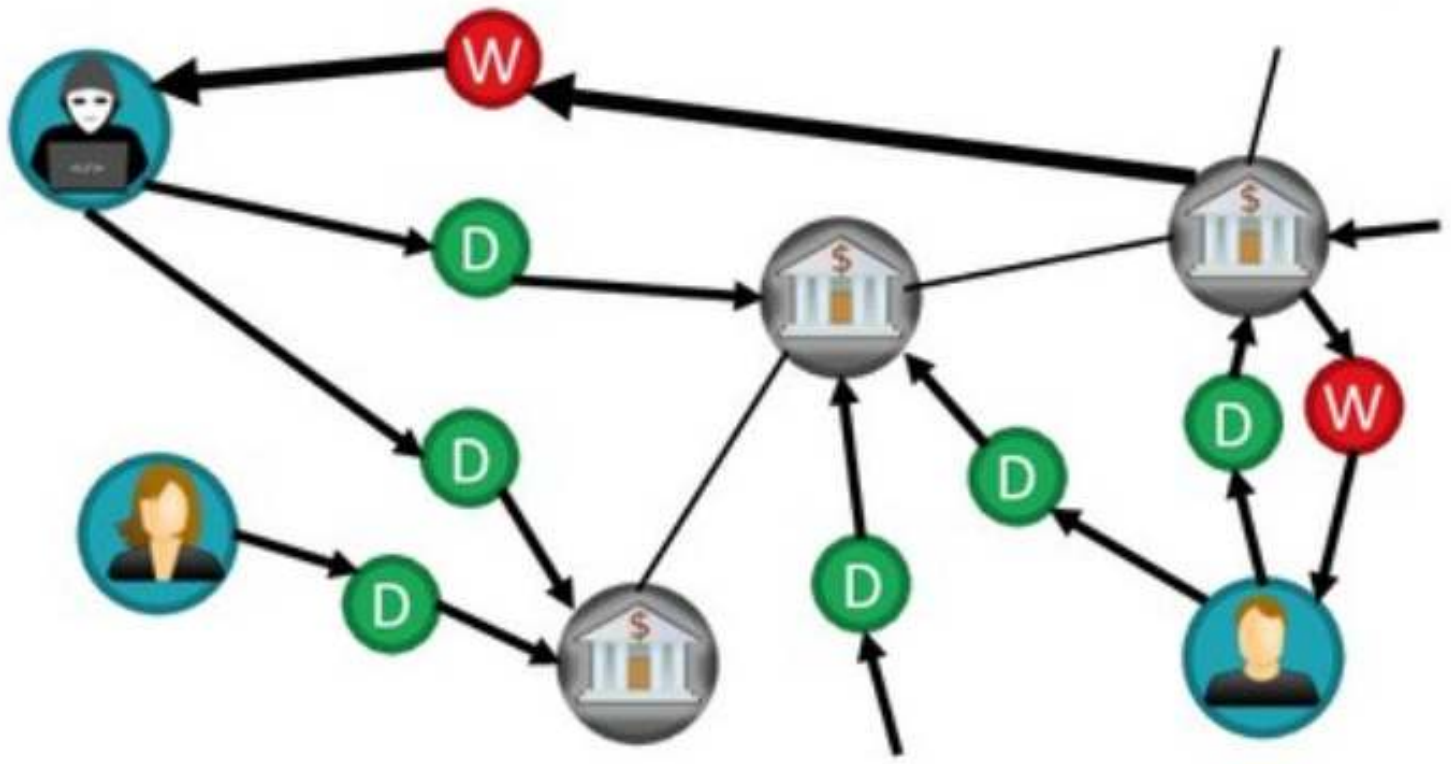
- **Khai thác mẫu phổ biến trong đồ thị** (frequent pattern mining in graph)



Mẫu chuỗi kết nối hóa học phổ biến

Các bài toán và ứng dụng

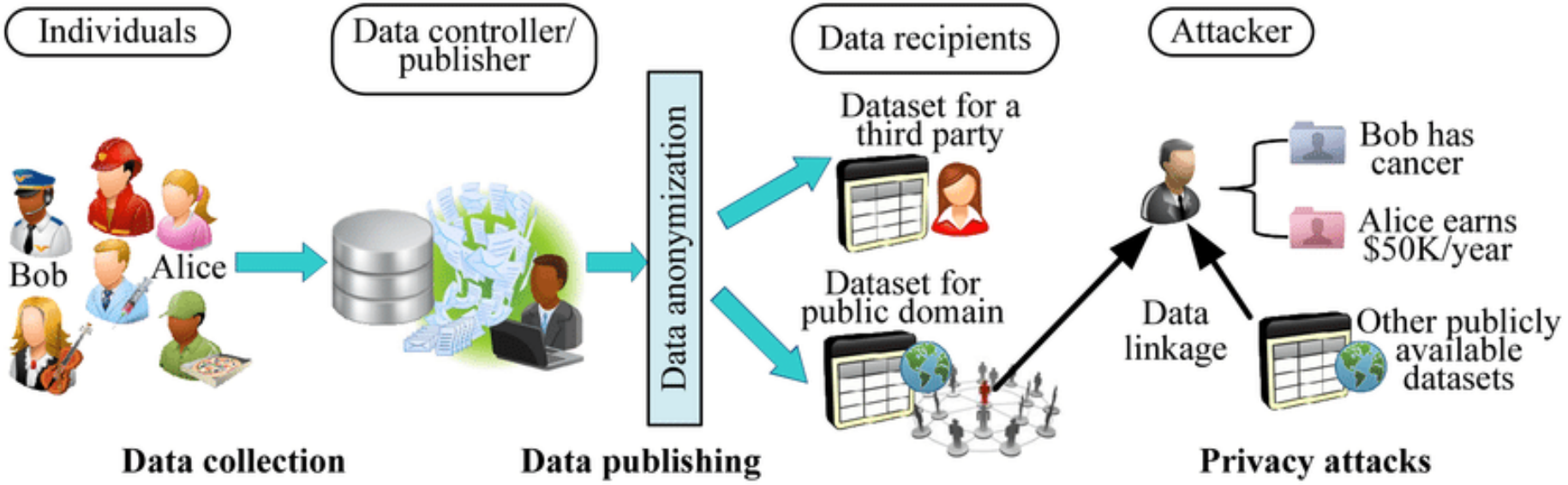
- **Khai thác mẫu phổ biến trong đồ thị** (frequent pattern mining in graph)



Mẫu rút và nạp tiền

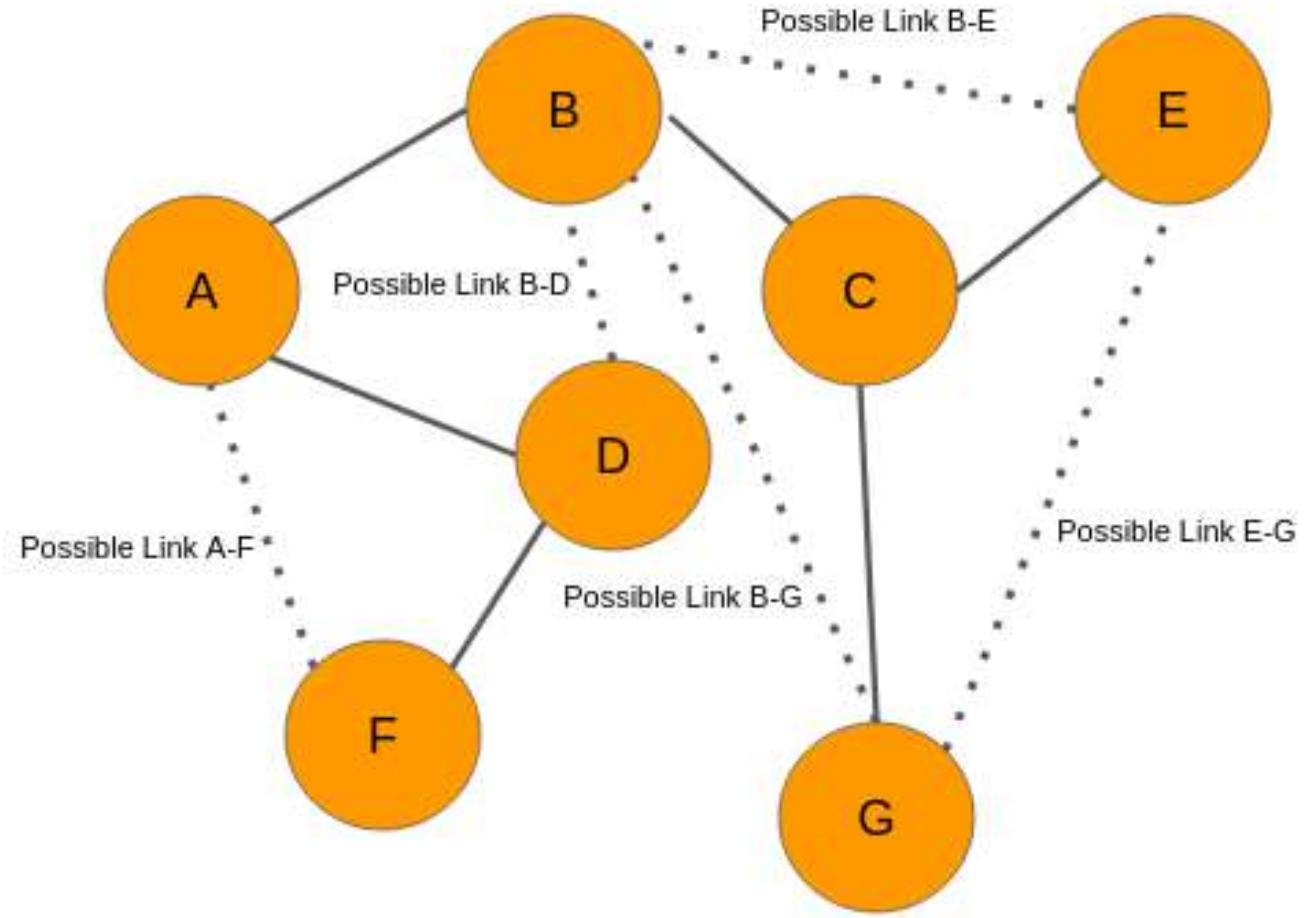
Các bài toán và ứng dụng

- **Bảo vệ quyền riêng tư trong đồ thị** (privacy-preserving in graph)
  - Bỏ đi thông tin định danh có thể chưa đủ vì có thể thông tin được nội suy từ các đỉnh đã biết.
  - Làm cách nào che đi thông tin định danh mà không phá vỡ cấu trúc tổng thể của đồ thị?



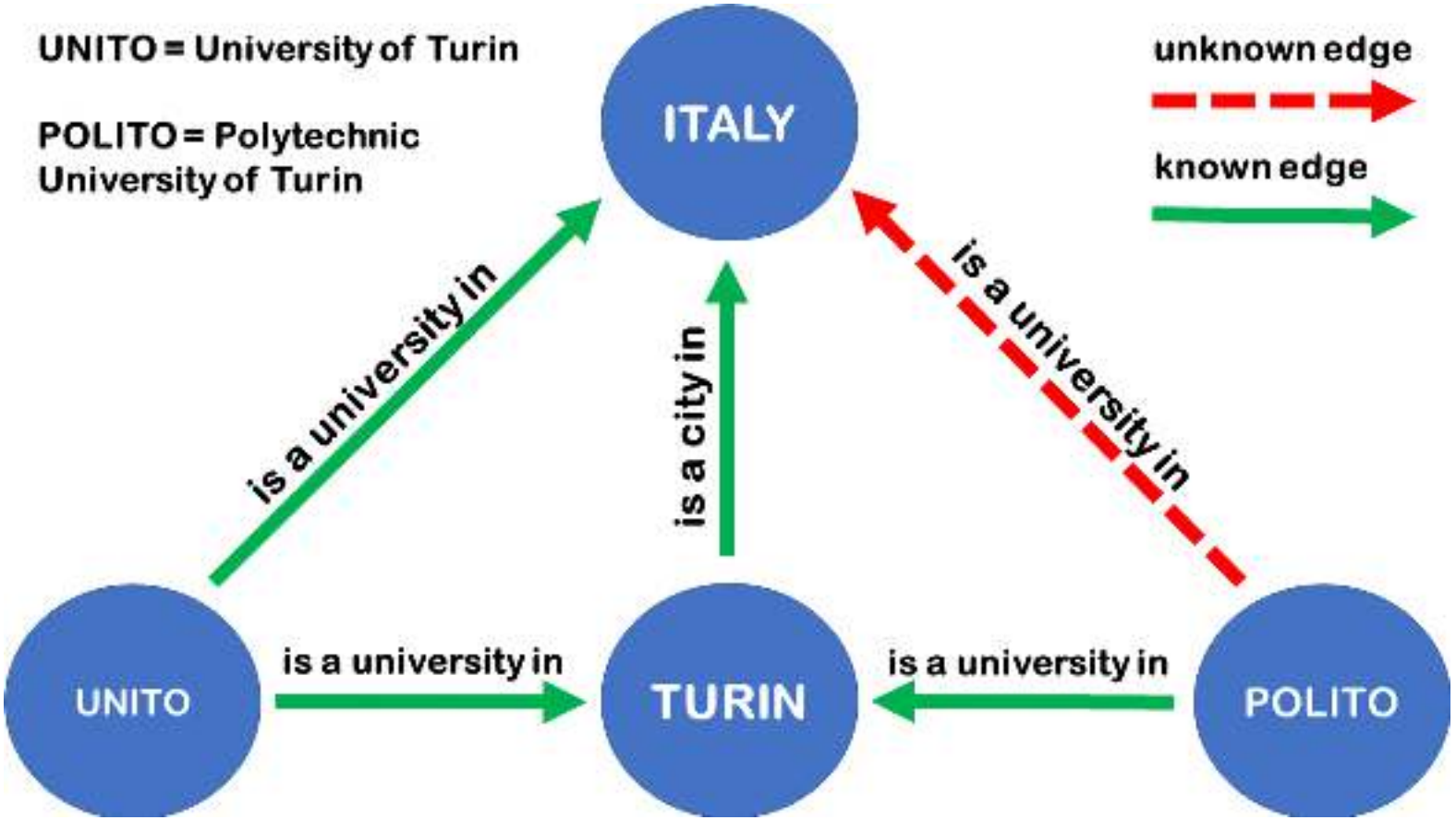
Các bài toán và ứng dụng

- **Dự đoán liên kết** (link prediction)



Các bài toán và ứng dụng

- **Dự đoán liên kết** (link prediction)



Các bài toán và ứng dụng

Graph analysis task	Application	Field
Graph clustering	Data storage Data compression	Database systems
	Popularity prediction Tag recommendation	Social network analysis
	Substructure indentification Network usage optimization	Computer networks
Graph matching	2D,3D Image analysis Face recognition Face verification Object registration/retrieval	Computer vision
	Document analysis	Language engineering
	Molecular structure study	Computational chemistry
Random walks	Enumeration	Multiple
	Volume computation	Computational geometry
	Mobile agent modelling	Distributed systems
	Web crawling	Internet computing
Anomaly detection	System intrusion detection Network attack detection	Computer security
	Financial fraud detection	Law enforcement
	Influential individual detection	Social network analysis

Tài liệu tham khảo

- Aggarwal, Charu C., and Haixun Wang, eds. Managing and mining graph data