



HỆ HỖ TRỢ QUYẾT ĐỊNH

Bài 11: Khai phá phương tiện truyền
thông xã hội

Nội dung

- ❶ Giới thiệu
- ❷ Biểu diễn đồ thị
- ❸ Số đo mạng
- ❹ Mô hình mạng
- ❺ Khai phá dữ liệu



1. Giới thiệu

Facebook

The screenshot shows Mark Zuckerberg's Facebook profile. At the top, there is a search bar and navigation links for Home, Profile, and Account. Below the header, there is a large photo of Mark Zuckerberg, followed by a section titled "Mark Zuckerberg". It includes a bio: "Has worked at Facebook Studied Computer Science at Harvard University Lives in Palo Alto, California From Dobbs Ferry, New York Born on May 14, 1984". There are five small thumbnail images below his name. A "Send Message" and "Poke" button are located on the right. On the left sidebar, there are links for Wall, Info (which is selected), Photos (826), Questions, Family, and a list of his parents and sisters. The main content area shows his education and work history, including his employer (Facebook) from February 2004 to present in Palo Alto, California, and his college (Harvard University) where he studied Computer Science and Psychology, taking courses like CS182 and CS121. It also lists his high school (Ardsley High School) and philosophy (Phillips Exeter Academy). A "Favorite Quotes" section contains a quote from毕加索: "All children are artists. The problem is how to remain an artist once he grows up." On the right side, there are several sponsored ads: "You and Mark" (3 mutual friends), "Police Auctions" (gsaauctions.gov), "SF Bucket List" (partners.livingsocial.com), "Stay close to your team" (AT&T), and "Craft Beer Attorney" (Craft Beer Attorney).

- Facebook sử dụng dữ liệu của bạn như thế nào?
- Bạn nghĩ Facebook sử dụng dữ liệu của bạn ở đâu?

Phương tiện truyền thông xã hội

Định nghĩa

Phương tiện truyền thông xã hội là việc sử dụng các công cụ điện tử và internet để chia sẻ và thảo luận thông tin, kinh nghiệm với người khác theo cách hiệu quả hơn.

Social Media Landscape 2015



FredCavazza.net

Thí dụ

- Một bài báo wiki
- Trang web đánh giá và xếp hạng 1 điểm bán pizza nổi tiếng ở thành phố của bạn
 - Thí dụ: Yelp.com
- Mạng xã hội trực tuyến với các đối tác, bạn bè của bạn
 - Thí dụ: Facebook.com, LinkedIn.com
- Một ứng dụng iPhone thông báo cho bạn biết vị trí bạn có thể đỗ xe
 - FasPark

Các dạng phương tiện truyền thông xã hội

- Online Social Networking
- Publishing
 - Blogging
 - Wiki
- Micro blogging
- Social News
- Social Bookmarking
- Media Sharing
 - Video Sharing
 - Photo Sharing
 - Podcast Sharing
- Opinion, Review, and Ratings Websites
- Answers
- Entertainment



Mạng xã hội trực tuyến

Mạng xã hội trực tuyến là dịch vụ web cho phép các cá nhân và tổ chức kết nối với bạn bè và người quen trong thế giới thực

- Tương tác
 - Tương tác bạn bè
 - Friends, like, comments, ...
 - Media Sharing
 - Gửi và nhận thông báo

- Thí dụ
 - Facebook.com
 - MySpace.com
 - Bebo.com
 - Orkut.com

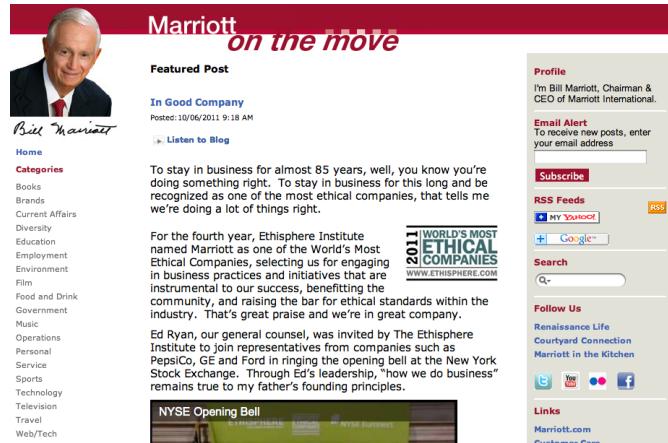
The image contains three separate screenshots of social media interfaces:

- MySpace:** A screenshot of a MySpace profile page for "Pei Pei". It shows a profile picture, basic information (Seattle, United States), and links to "Contacting Pei Pei" (Send Message, Add to Friends, IM / call, Add to Group) and "MySpace URL". Below this is a "General Info" section with details like Member Since (3/1/2010), Band Members (Sayuri Wijaya Gould), Influences (too many to list them all), and Type of Label (Unsigned). A "Music" section displays a track by "Gremlake" and a blog entry by Pei Pei.
- Facebook:** A screenshot of a Facebook profile for "Barack Obama". It shows a large profile picture, a timeline with posts from August 2010, and sections for "Wall", "Info", "OFA Store", "Photos", "Join OFA", and "Video". A sidebar on the right shows news items and a "Statement by the President on the Occasion of Ramadan".
- Mobile Device:** A screenshot of a mobile phone displaying a social media feed. The top status bar shows "Facebook" and signal strength. The main screen shows a news feed with various posts and photos, similar to the Facebook desktop version.

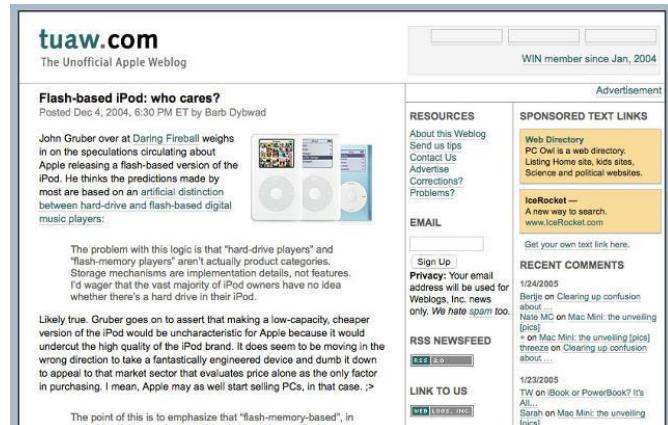
Viết blog

Blog là một website giống như tạp chí dành cho người dùng, còn gọi là blogger, đóng góp nội dung văn bản và đa phương tiện, được sắp xếp theo thứ tự thời gian ngược

- Được duy trì bởi cá nhân hay cộng đồng
 - Xem hướng dẫn ở KDD
http://videolectures.net/kdd08_liu_briat/
- Sử dụng:
 - Chia sẻ thông tin và các ý tưởng với bạn bè hay người lạ
 - Phổ biến nội dung theo một chủ đề cụ thể
 - Ai là người có ảnh hưởng
http://videolectures.net/wsdm08_agarwal_iib/



The screenshot shows a blog post by Bill Marriott. The post is titled "In Good Company" and was posted on 10/06/2011 at 9:18 AM. It includes a link to "Listen to Blog". The post discusses Marriott's commitment to ethical business practices over 85 years. Below the post, there's a sidebar with links to "RSS Feeds" (Yahoo!, Google), a search bar, and social media links for LinkedIn, YouTube, and Facebook. A banner on the right says "WORLD'S MOST ETHICAL COMPANIES" with a link to WWW.ETHISPHERE.COM.



The screenshot shows a news article from tuaw.com titled "Flash-based iPod: who cares?". The article was posted on Dec 4, 2004, at 6:30 PM ET by Barb Dwydaw. It discusses John Gruber's opinion on the classification of iPods as either "hard-drive players" or "flash-memory players". The article notes that Apple's implementation details are different from what Gruber expected. The sidebar includes links to "RESOURCES" and "SPONSORED TEXT LINKS", and a "RECENT COMMENTS" section with several entries.

Viết tiêu blog

Viết tiêu blog có thể được xem như viết blog nhưng với nội dung hạn chế

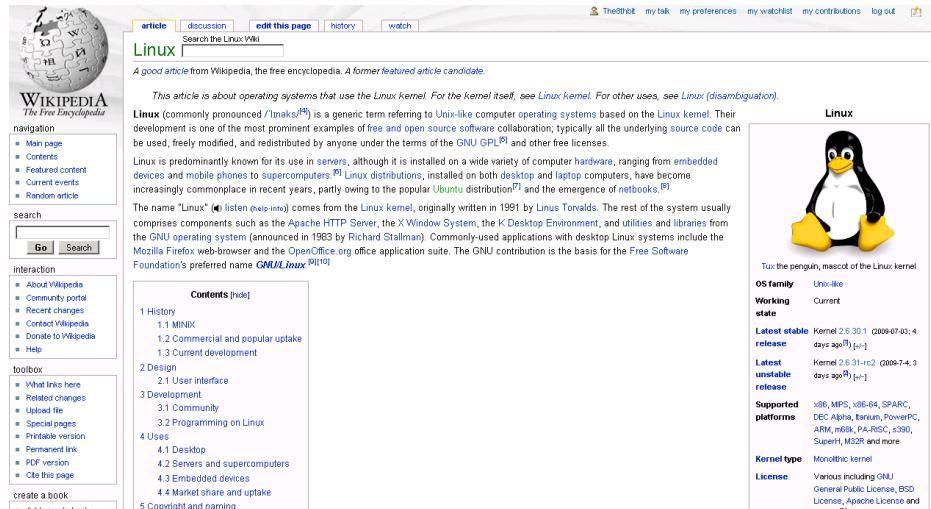
- Sử dụng
 - communication medium
 - social interaction
 - citizen journalism
- Nhà cung cấp dịch vụ:
 - Twitter
 - Google buzz



Wiki

Wiki là một môi trường biên tập cộng tác cho phép người dùng phát triển các trang web sử dụng ngôn ngữ đánh dấu đơn giản

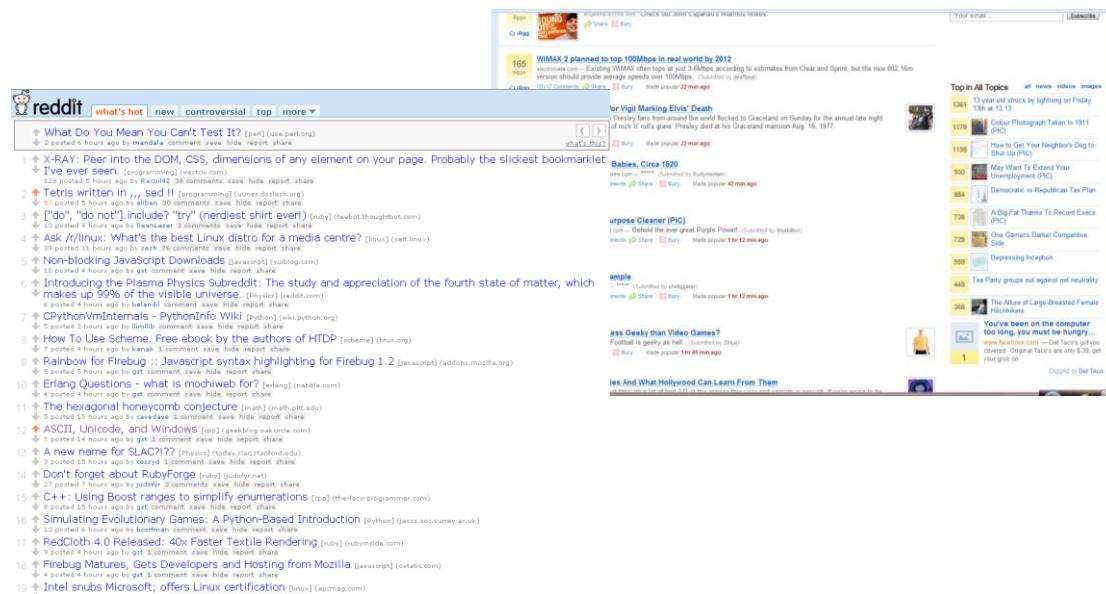
- Wikipedia cho phép các cá nhân cộng tác viết các bài báo về nhiều chủ đề.
- Sử dụng trí tuệ của tập thể hiệu quả, nó đã trở thành một kho thông tin toàn diện và hữu ích cho nhiều các nhân



Social News

Tin tức xã hội đề cập tới việc chia sẻ và lựa chọn các tin bài và các bài báo của một cộng đồng người sử dụng.

- Người dùng có thể chia sẻ các bài báo mà họ cho rằng sẽ khiến cộng đồng quan tâm
- Thí dụ:
 - Digg.com
 - Slashdot
 - Fark
 - Reddit



Social Bookmarking

Các website đánh dấu trang xã hội cho phép người dùng đánh dấu nội dung web để lưu trữ, tổ chức và chia sẻ.

- Các đánh dấu trang này có thể được gắn với siêu dữ liệu để phân loại và cung cấp ngũ cảnh cho nội dung được chia sẻ, cho phép người dùng sắp xếp thông tin giúp dễ dàng tìm kiếm và xác định thông tin liên quan.
- Thí dụ:
 - Delicious.com
 - StumbleUpon.com

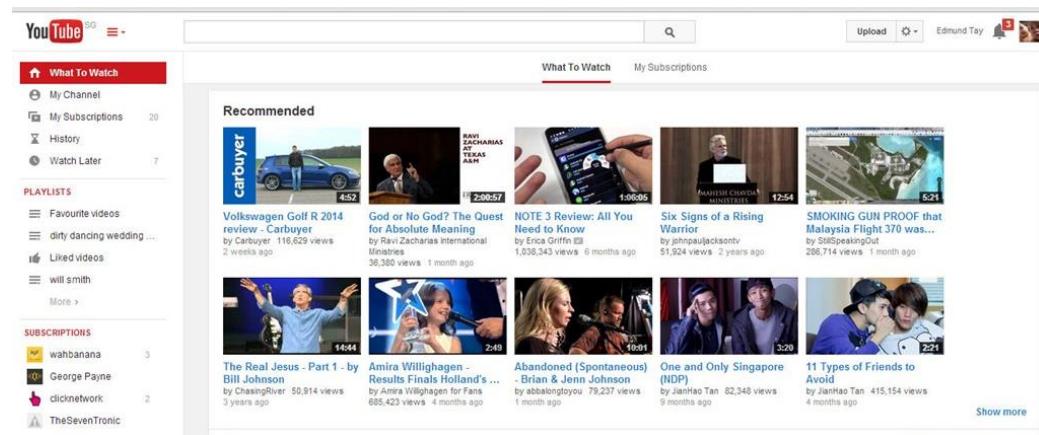
The screenshot shows the Delicious.com homepage. At the top, there's a banner with the text "The tastiest bookmarks on the web. Save your own or see what's fresh now!" and a "Learn More" button. Below the banner is a search bar with the placeholder "Search the biggest collection of bookmarks in the universe...". To the right of the search bar are buttons for "Search Delicious" and "Search". A navigation bar below the search bar includes "Popular Bookmarks" and "Explore Tags". The main content area displays a list of popular bookmarks with their titles, save counts, and tags. To the right of the bookmark list is a sidebar titled "Popular Tags" with a scrollable list of tags like "design", "blog", "video", etc. At the bottom of the page, there's a footer with links to "delicious", "about", "blog", "terms of service", "privacy policy", "copyright policy", "forums", "support", and "What's new?".

Chia sẻ phương tiện

Chia sẻ phương tiện (Media sharing) là khái niệm bao trùm để chỉ dẫn tới việc chia sẻ của nhiều phương tiện trên web.

Người dùng chia sẻ nội dung đa phương tiện có thể được người khác quan tâm

- Thí dụ:
 - Video Sharing:
 - YouTube.com
 - Photo Sharing:
 - Flickr.com, picasa.com
 - Document Sharing:
 - Scribd.com, Slideshare.com
 - Livecasting:
 - Justin.tv, Ustream.com



Website ý tưởng, đánh giá, và xếp hạng

Website ý tưởng, đánh giá và xếp hạng là những website có chức năng chính là thu thập và công bố nội dung người dùng gửi ở dạng nhận xét về sản phẩm, dịch vụ, giải trí, kinh doanh, vị trí, ... Một số website thương mại có thể phục vụ mục đích phụ là đánh giá website bằng cách công bố các đánh giá sản phẩm được gửi bởi khách hàng.

- Thí dụ
 - Cnet.com
 - Epinions.com
 - yelp.com
 - tripadvisor.com

The screenshot shows a Yelp search results page for 'Tartine Bakery' in San Francisco, California. The main focus is a product page for 'Croissant'. At the top, there's a search bar with 'sf, ca' entered. Below it, a navigation bar includes 'Welcome', 'About Me', 'Write a Review', 'Find Friends', 'Messaging', 'Talk', and 'Events'. A 'Search' button is also present. The main content area shows a large image of a croissant next to a cup of coffee. To the right, the price '\$3.85' is listed. Below the image, there are three smaller photos of different pastries. A 'See more photos' link is visible. A review from 'Stephanie S.' is shown, dated 10/23/2012, with a rating of 5 stars. The review text reads: 'This was our first stop from the airport and we were starving! The line was long, but it went pretty fast. This was our first time here and we couldn't decide what to order. We tried the morning bun, chocolate and almond croissant, bread pudding, & the chocolate eclair. Everything was delicious, but the morning bun was soo amazing. I loved the hints of citrus and the flakiness of the bun. I made my hubby go back & buy me another one to save for later. Oh, Tartine! I wish you were also located in So. Cal.' Below the review are buttons for 'Write a Review', 'Add a photo', 'Complement', 'Send Message', and 'Follow This Reviewer'. At the bottom of the page, there are links for 'Bookmark', 'Send to a Friend', 'Link to This Review', and 'Flag this review'.

Trả lời câu hỏi bởi cộng đồng

Trong các website này, người dùng yêu cầu chỉ dẫn, lời khuyên hay tri thức có thể đặt câu hỏi. Người sử dụng khác trong cộng đồng có thể trả lời những câu hỏi này dựa trên tri thức có từ các kinh nghiệm trước, ý kiến cá nhân hay từ các nghiên cứu liên quan.

- Không giống các website đánh giá và ý kiến, chứa các đóng góp ý kiến tự thúc đẩy, website trả lời câu hỏi chứa tri thức được chia sẻ trong trả lời câu hỏi cụ thể.
- Thí dụ:
 - WikiAnswers, Yahoo Answers, Quora

Search Google Analytics Questions and Topics Add Question

Question added to topic Google Analytics:
What percentage of visits would Omniture / Google Analytics / Coremetrics etc miss?
Assuming client-side integration, compared with the numbers from the web servers and proxy logs.
Follow · Repost · 0 Answers · 5:55pm

Answer added in topic Google Analytics:
How can I track Pinterest in Google Analytics?
1 Ross Allen, Front End Engineer at Airbnb
Their Javascript pinit.js file (<http://assets.pinterest.com/js/pinit.js>) doesn't seem to add any callbacks, so the best you can do is track clicks on the 'Pin It' button in Goo... (more)
Upvote · Repost · 2 Answers · 5:17pm

Answer added in topic Google Analytics:
Google Analytics: Why would someone from an email marketing company tell me that Google analytics does not track visits from Mac users?
2 Anon User
The person was seeing if you were gullible enough to be a good fit with their product.
Sales 101.
Upvote · Repost · 4 Answers · 3:52pm

Share Topic · Invite People
Twitter Facebook Quora

Top Answerers
Mike Sullivan 20 Answers
Ozberk Olcer 20 Answers Director of Web Analytics in SEM AS. (Google Analytics Certified Partner)
Shay Sharon 22 Answers
AJ Kohn 17 Answers
Christopher O'Donnell 11 Answers



*Khai phá phương tiện truyền thông xã hội
(Social Media Mining) là tiến trình thể hiện,
phân tích và trích mẫu có ý nghĩa từ dữ liệu
phương tiện truyền thông xã hội*

Các đặc điểm chính

- **Sự tham gia**
 - Phương tiện truyền thông xã hội khuyến khích sự đóng góp và phản hồi của mọi người quan tâm. Nó làm mờ ranh giới giữa truyền thông và độc giả.
- **Tính mở**
 - Hầu hết các dịch vụ phương tiện truyền thông xã hội mở có phản hồi và tham gia. Họ khuyến khích bỏ phiếu, bình luận và chia sẻ thông tin. Hiếm khi có rào cản đối với việc truy cập và sử dụng nội dung – nội dung bảo vệ bằng mật khẩu là không chấp nhận được.
- **Hội thoại/tương tác**
 - Trong khi phương tiện truyền thông truyền thống ở dạng “phát sóng” (nội dung được truyền hay phân phối tới độc giả) phương tiện truyền thông xã hội là hội thoại 2 chiều.
- **Tính cộng đồng**
 - Phương tiện truyền thông xã hội cho phép nhanh chóng hình thành các cộng đồng và trao đổi thông tin hiệu quả với nhau. Các cộng đồng chia sẻ các sở thích chung, như yêu thích nhiếp ảnh, quan tâm tới một vấn đề chính trị hay một chương trình truyền hình.
- **Tính kết nối**
 - Hầu hết các dạng phương tiện truyền thông xã hội phát triển mạnh nhờ tính kết nối của chúng, tận dụng các liên kết web, tài nguyên và con người.

Các thách thức của khai phá phương tiện truyền thông xã hội

1. Dữ liệu lớn

- Dữ liệu phương tiện truyền thông xã hội lớn, không phân phối đồng đều.
- Thường ít dữ liệu đối với một cá nhân cụ thể

2. Khó lấy mẫu đại diện

- Các mẫu của chúng ta có đại diện cho dữ liệu đầy đủ (quần thể) không?

3. Khó giảm nhiễu

- Loại bỏ quá nhiều nhiễu làm cho dữ liệu trở nên thưa
- Xác định đâu là nhiễu là tương đối và phức tạp, phụ thuộc vào công việc cụ thể

4. Khó đánh giá

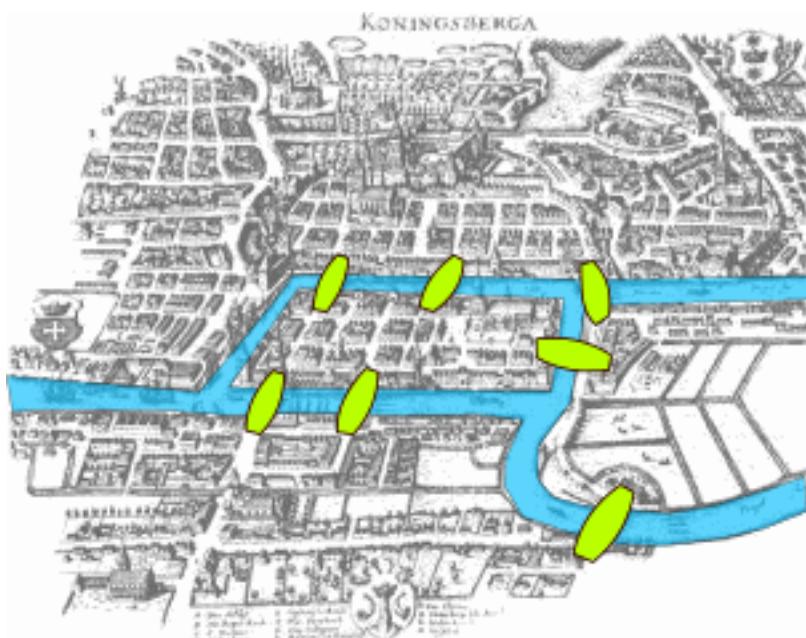
- Khi không có nhãn (ground truth), làm sao ta có thể đánh giá được?



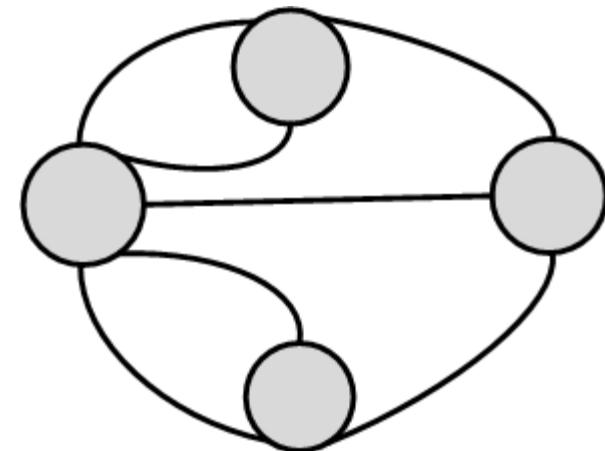
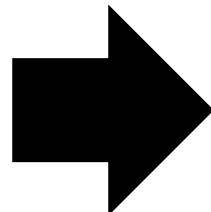
2. Thể hiện đồ thị

Các cây cầu ở Koenigsberg

- Có 2 đảo (cù lao) và 7 cây cầu kết nối chúng và đất liền
- Tìm đường đi mà chỉ qua mỗi cầu chính xác 1 lần



City Map (From Wikipedia)



Thể hiện đồ thị

Mạng đường cao tốc liên bang Mỹ

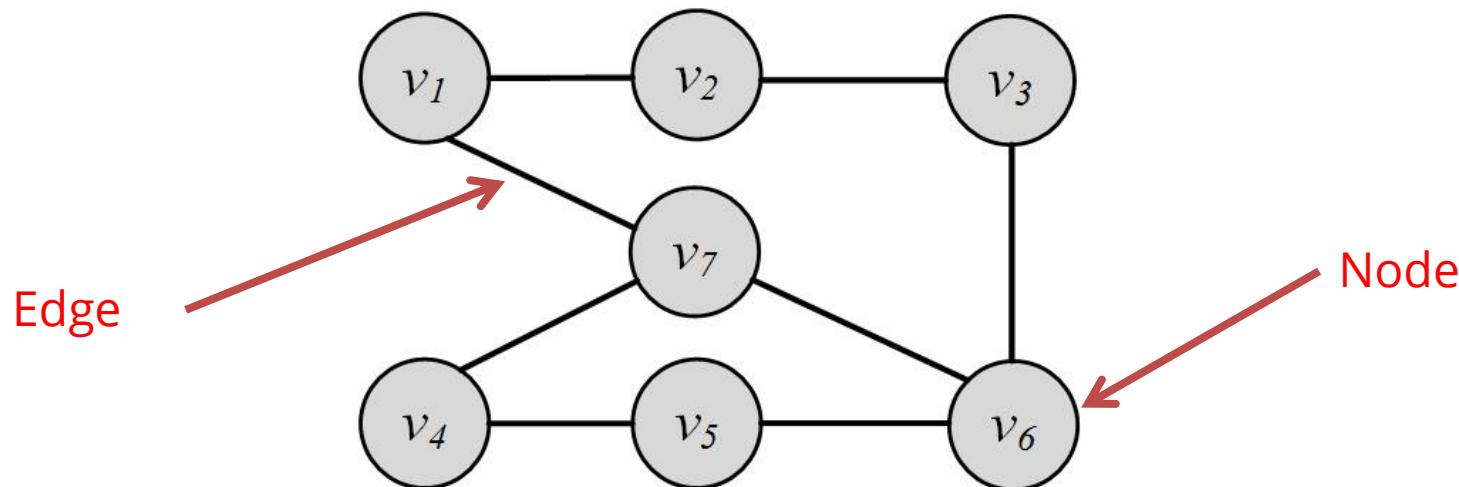
A network of interstates



Mạng và lý thuyết đồ thị

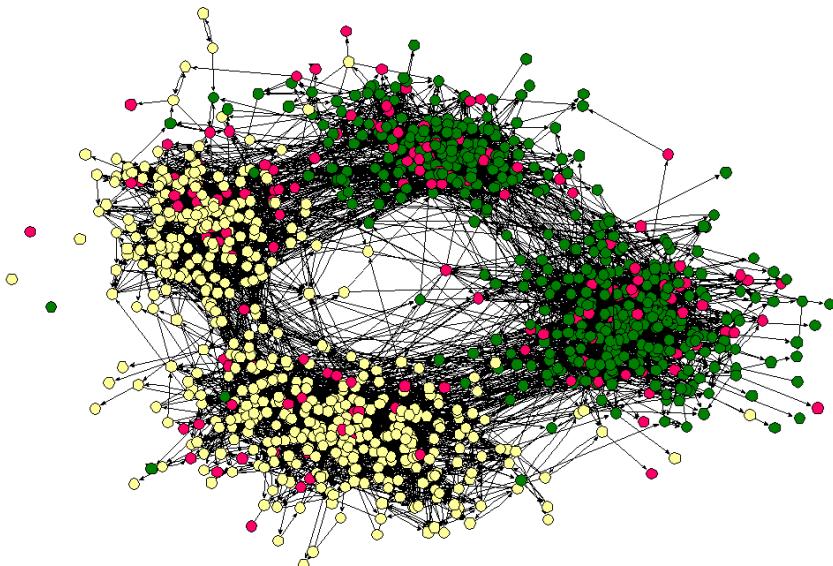
Mạng là một đồ thị hay tập các điểm được kết nối bởi các đường

- Các điểm được gọi là **nút** (node), **tác nhân** (actor), hay **đỉnh** (vertice)
- Các kết nối được gọi là các **cạnh** (edge) hay **quan hệ** (tie)

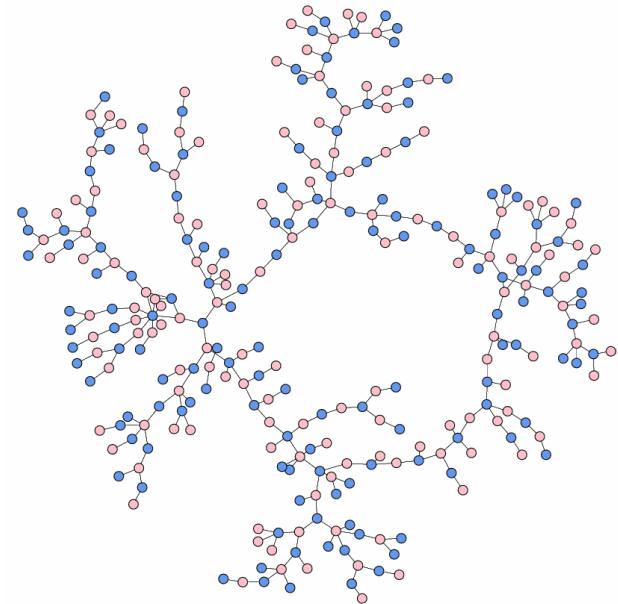


Mạng xã hội

Mạng hẹn hò



Mạng quan hệ bạn bè



Nút

- Trong đồ thị quan hệ bạn bè xã hội, nút thể hiện các cá nhân có tham gia quan hệ bạn bè với các cá nhân (nút) khác
- Tùy thuộc vào ngũ cảnh, các nút cũng có thể được gọi là các tác nhân
 - Trong đồ thị web, “*nút*” thể hiện các website và kết nối giữa các nút thể hiện các web-link giữa chúng
 - Trong môi trường xã hội, các nút được gọi là các tác nhân

$$V = \{v_1, v_2, \dots, v_n\}$$

- Kích thước của đồ thị $|V| = n$

Cạnh

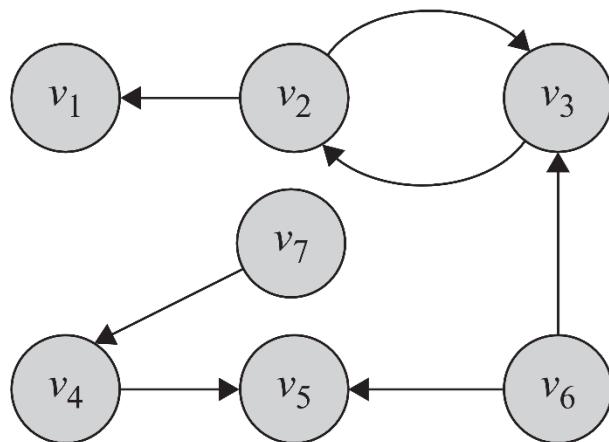
- Cạnh kết nối các nút và được gọi là các **quan hệ**
- Trong môi trường xã hội, ở đó các nút thể hiện các thực thể xã hội như các cá nhân, các cạnh thể hiện quan hệ giữa các nút và được gọi là quan hệ xã hội

$$E = \{e_1, e_2, \dots, e_m\}$$

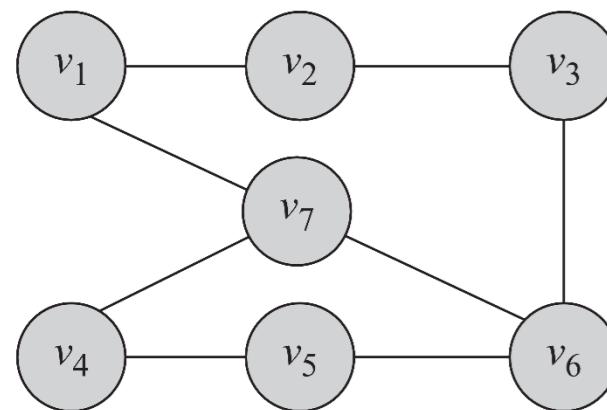
- Số các cạnh được ký hiệu là $|E| = m$

Đồ thị có hướng

- Các cạnh có thể có hướng. Cạnh có hướng thỉnh thoảng được gọi là **cung** (arc)



(a) Directed Graph



(b) Undirected Graph

- Các cạnh được thể hiện qua các điểm cuối của nó $e(v_2, v_1)$.
- Trong đồ thị vô hướng, cạnh thể hiện cả 2 hướng.

Hàng xóm và bậc của nút (bậc vào, bậc ra)

Cho nút v , trong đồ thị vô hướng, tập tất cả các nút kết nối tới nó qua một cạnh nào đó được gọi là hàng xóm (neighborhood) của nó và được ký hiệu là $N(v)$

- Trong đồ thị có hướng ta có hàng xóm vào (incoming neighbor) $N_{in}(v)$ (các nút kết nối tới v) và hàng xóm ra (outgoing neighbor) $N_{out}(v)$.

Số cạnh kết nối tới một nút được gọi là bậc/độ (degree) của nút đó (kích thước của hàng xóm của nó)

- Bậc của nút i được ký hiệu là d_i

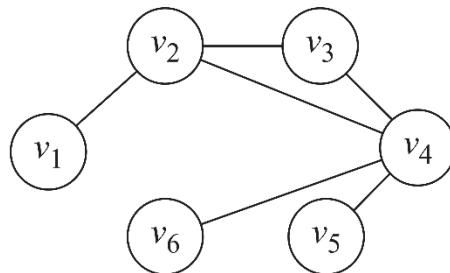
Trong đồ thị có hướng:

d_i^{in} – Bậc vào là số cạnh hướng tới nút

d_i^{out} – Bậc ra là số cạnh hướng ra khỏi nút

Thể hiện đồ thị

- Ma trận kề



(a) Graph

	v ₁	v ₂	v ₃	v ₄	v ₅	v ₆
v ₁	0	1	0	0	0	0
v ₂	1	0	1	1	0	0
v ₃	0	1	0	1	0	0
v ₄	0	1	1	0	1	1
v ₅	0	0	0	1	0	0
v ₆	0	0	0	1	0	0

(b) Adjacency Matrix

- Danh sách kề

- Danh sách cạnh

Node	Connected To
v ₁	v ₂
v ₂	v ₁ , v ₃ , v ₄
v ₃	v ₂ , v ₄
v ₄	v ₂ , v ₃ , v ₅ , v ₆
v ₅	v ₄
v ₆	v ₄

(v₁,v₂)

(v₂,v₃)

(v₂,v₄)

(v₃,v₄)

(v₄,v₅)

(v₄,v₆)



3. Số đo mạng

Klout

Barack Obama
ADD +

This account is run by #Obama2012 campaign staff. Tweets from the President are signed -bo.

Influences 2M others

Influential about 20 topics

Government, Politics, Media

99

CELEBRITY

OBAMA BIDEN

1F tweet • 1 share • see more...

1F tweet • 1 share • see all

Rất khó để đo
lường ảnh
hưởng!

Justin Bieber
ADD +

#BELIEVE is MUCH LOVING and I will always be here for you.

Influences 10M others

KLOUT the Standard for Influence

92

CELEBRITY

1F tweet • 1 share • see more...

Klout Summary for Warren Buffett

Score Analysis

Warren Buffett
Investor, Philanthropist
Omaha, Nebraska

36
klout score

Tại sao ta cần các số đo?

- Nhân vật trung tâm (những cá nhân có ảnh hưởng) trong mạng lưới là ai?
 - Độ trung tâm - Centrality
- Những kiểu tương tác nào thường gặp ở bạn bè?
 - Có đi có lại (Reciprocity) và BẮC CẦU (Transitivity)
 - CÂN BẰNG (Balance) và ĐỊA VỊ (Status)
- Ai là những người có cùng chí hướng và làm sao để tìm được những cá nhân tương đồng này?
 - SỰ TƯƠNG ĐỒNG - Similarity
- Để trả lời các câu hỏi này và các câu hỏi tương tự, cần xác định các số đo định lượng về mức/độ trung tâm, mức độ tương tác và độ tương đồng.

Độ trung tâm

Độ trung tâm xác định mức độ quan trọng của một nút
trong mạng



Độ trung tâm theo những người mà ta kết nối

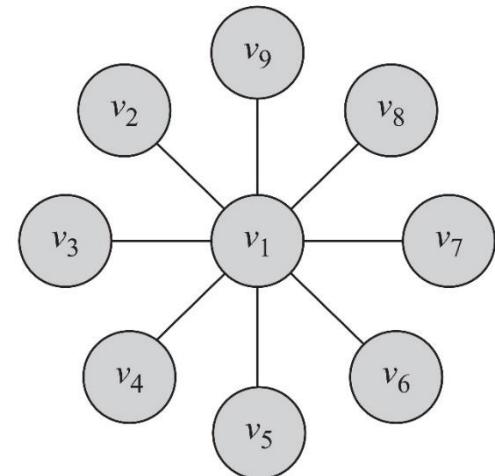
Độ trung tâm- Degree Centrality

- **Độ trung tâm:** xếp hạng các nút nhiều kết nối thì cao hơn về độ trung tâm

$$C_d(v_i) = d_i$$

- d_i là bậc (số bạn) của nút v_i
 - Nghĩa là số các đường kết nối độ dài 1 (length-1 paths) – cũng có thể tổng quát hóa

Trong đồ thị này, độ trung tâm của nút v_1 là $d_1=8$ và cho tất cả các nút khác $d_j = 1, j \neq 1$



Độ trung tâm trong đồ thị có hướng

- Trong đồ thị có hướng, ta có thể sử dụng hoặc độ vào (in-degree), độ ra (out-degree), hay kết hợp của chúng như giá trị của độ trung tâm:
- Trong thực tế, thường dùng in-degree.

$$C_d(v_i) = d_i^{\text{in}} \quad (\textit{prestige})$$

$$C_d(v_i) = d_i^{\text{out}} \quad (\textit{gregariousness})$$

$$C_d(v_i) = d_i^{\text{in}} + d_i^{\text{out}}$$

d_i^{out} là số liên kết đi ra của nút v_i

Độ trung tâm chuẩn hóa

- Chuẩn hóa bởi độ lớn nhất có thể

$$C_d^{\text{norm}}(v_i) = \frac{d_i}{n-1}$$

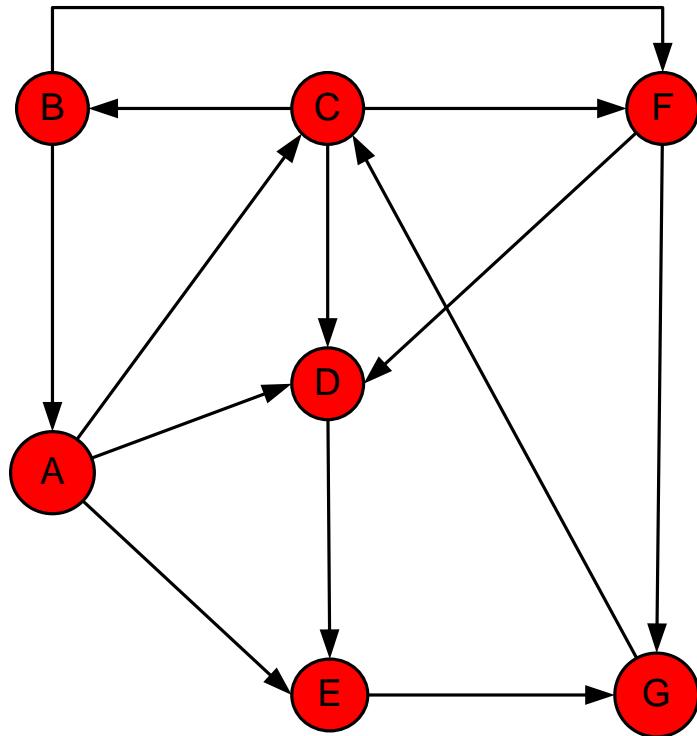
- Chuẩn hóa bởi độ lớn nhất

$$C_d^{\max}(v_i) = \frac{d_i}{\max_j d_j}$$

- Chuẩn hóa bởi tổng

$$C_d^{\text{sum}}(v_i) = \frac{d_i}{\sum_j d_j} = \frac{d_i}{2|E|} = \frac{d_i}{2m}$$

Thí dụ độ trung tâm trên đồ thị có hướng

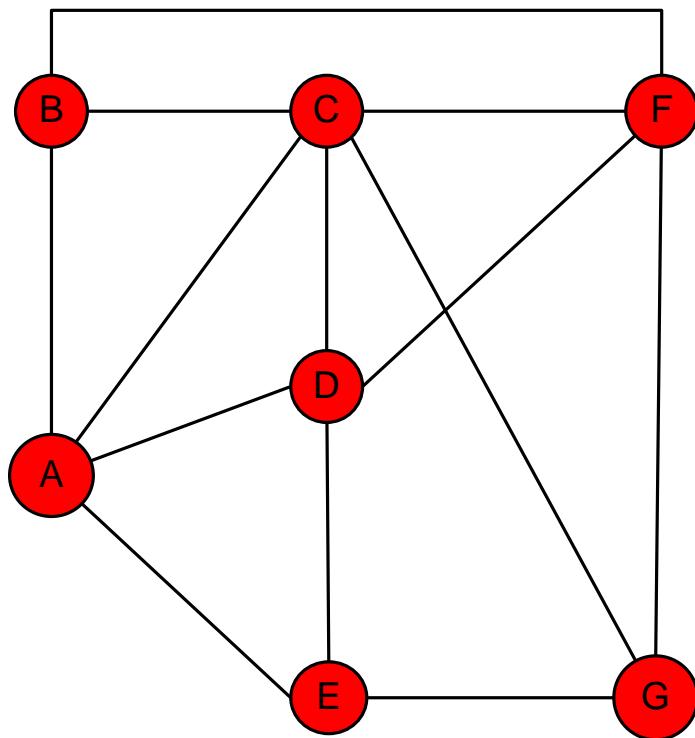


Node	In-Degree	Out-Degree	Centrality	Rank
A	1	3	1/2	1
B	1	2	1/3	3
C	2	3	1/2	1
D	3	1	1/6	5
E	2	1	1/6	5
F	2	2	1/3	3
G	2	1	1/6	5

Được chuẩn hóa bởi độ lớn nhất có thể

$$C_d^{\text{norm}}(v_i) = \frac{d_i}{n-1}$$

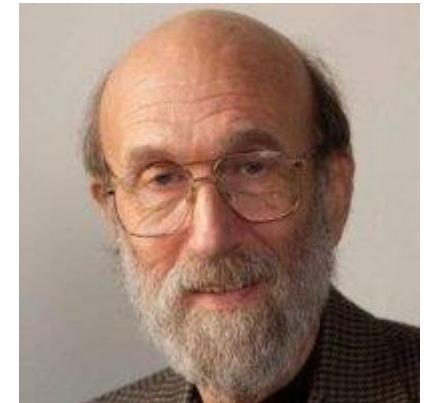
Thí dụ độ trung tâm trên đồ thị vô hướng



Node	Degree	Centrality	Rank
A	4	2/3	2
B	3	1/2	5
C	5	5/6	1
D	4	2/3	2
E	3	1/2	5
F	4	2/3	2
G	3	1/2	5

Véc tơ riêng trung tâm

- Có nhiều bạn bè hơn không đảm bảo cá nhân đó quan trọng hơn người khác
 - Có nhiều **bạn bè quan trọng** đưa ra tín hiệu mạnh hơn về mức độ quan trọng
- Véc tơ riêng trung tâm \vec{t} tổng quát hóa độ trung tâm bằng cách tích hợp sự quan trọng của các hàng xóm (vô hướng)
- Đối với đồ thị có hướng, ta có thể sử dụng chỉ các cạnh vào hay ra



Phillip Bonacich

Công thức

- Giả sử véc tơ riêng trung tâm của một nút v_i là $c_e(v_i)$ (**không biết**)
- Ta muốn $c_e(v_i)$ sẽ cao hơn khi các hàng xóm quan trọng (nút v_j với $c_e(v_j)$ cao hơn) kết nối với nó.
 - Hàng xóm vào hay ra?
 - Đối với hàng xóm vào $A_{j,i} = 1$
- Chúng ta có thể giả thiết rằng mức trung tâm của v_i là tổng các mức trung tâm của các hàng xóm của nó
$$c_e(v_i) = \sum_{j=1}^n A_{j,i} c_e(v_j)$$
- Tổng này có bị chặn không?
 - Chúng ta phải chuẩn hóa nó! $c_e(v_i) = \frac{1}{\lambda} \sum_{j=1}^n A_{j,i} c_e(v_j)$
 λ : là **một hằng số nào đó**

Véc tơ riêng trung tâm (công thức ma trận)

- Đặt $\mathbf{C}_e = (C_e(v_1), C_e(v_2), \dots, C_e(v_n))^T$

$$\rightarrow \lambda \mathbf{C}_e = A^T \mathbf{C}_e$$

- Nghĩa là \mathbf{C}_e là véc tơ riêng của ma trận kè A^T (hay A đối với đồ thị vô hướng) và λ là giá trị riêng tương ứng
- Cặp giá trị riêng – véc tơ riêng nào nên được chọn?
- Cho mục đích so sánh ta thường muốn tất cả các giá trị mức trung tâm dương, do đó ta sẽ chọn giá trị riêng nào mà véc tơ riêng của nó có tất cả các phần tử là dương.

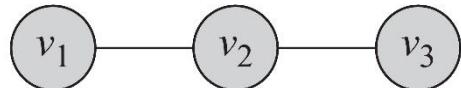
Véc tơ riêng trung tâm

Theorem 1 (Perron-Frobenius Theorem). *Let $A \in \mathbb{R}^{n \times n}$ represent the adjacency matrix for a [strongly] connected graph or $A : A_{i,j} > 0$ (i.e. a positive n by n matrix). There exists a positive real number (Perron-Frobenius eigenvalue) λ_{\max} , such that λ_{\max} is an eigenvalue of A and any other eigenvalue of A is strictly smaller than λ_{\max} . Furthermore, there exists a corresponding eigenvector $\mathbf{v} = (v_1, v_2, \dots, v_n)$ of A with eigenvalue λ_{\max} such that $\forall v_i > 0$.*

Như vậy, để tính véc tơ riêng trung tâm của A ,

1. Tính các giá trị riêng của A
2. Chọn giá trị riêng lớn nhất λ
3. Véc tơ riêng tương ứng của λ là \mathbf{C}_e .
4. Theo định lý Perron-Frobenius, tất cả các thành phần của \mathbf{C}_e sẽ dương
5. Các thành phần của \mathbf{C}_e là véc tơ riêng độ trung tâm của đồ thị.

Thí dụ 1: véc tơ riêng trung tâm



$$\lambda \mathbf{C}_e = A \mathbf{C}_e \quad (A - \lambda I) \mathbf{C}_e = 0 \quad \mathbf{C}_e = [u_1 \ u_2 \ u_3]^T$$

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 - \lambda & 1 & 0 \\ 1 & 0 - \lambda & 1 \\ 0 & 1 & 0 - \lambda \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\det(A - \lambda I) = \begin{vmatrix} 0 - \lambda & 1 & 0 \\ 1 & 0 - \lambda & 1 \\ 0 & 1 & 0 - \lambda \end{vmatrix} = 0$$

$$(-\lambda)(\lambda^2 - 1) - 1(-\lambda) = 2\lambda - \lambda^3 = \lambda(2 - \lambda^2) = 0$$

Các giá trị riêng là

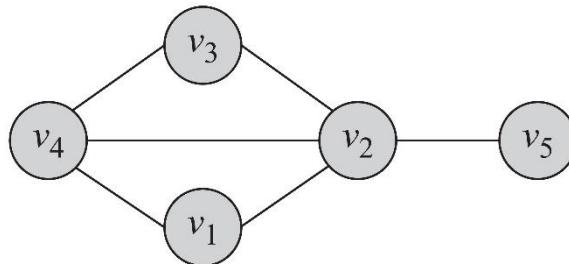
$$(-\sqrt{2}, 0, +\sqrt{2})$$

Giá trị riêng lớn nhất

Véc tơ riêng tương ứng (giả sử \mathbf{C}_e có chuẩn 1)

$$\begin{bmatrix} 0 - \sqrt{2} & 1 & 0 \\ 1 & 0 - \sqrt{2} & 1 \\ 0 & 1 & 0 - \sqrt{2} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad \mathbf{C}_e = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} 1/2 \\ \sqrt{2}/2 \\ 1/2 \end{bmatrix}$$

Thí dụ 2: véc tơ riêng trung tâm



$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \rightarrow \lambda = (2.68, -1.74, -1.27, 0.33, 0.00)$$

↑ Véc tơ các giá trị riêng

$$\lambda_{\max} = 2.68 \quad \rightarrow$$

$$C_e = \begin{bmatrix} 0.4119 \\ 0.5825 \\ 0.4119 \\ 0.5237 \\ 0.2169 \end{bmatrix}$$

Độ trung tâm Katz

- Vấn đề chính với véc tơ riêng trung tâm này sinh khi làm việc với đồ thị có hướng
- Độ trung tâm chỉ qua các cạnh ra và trong trường hợp đặc biệt khi một nút nằm trong một đồ thị không có chu trình có hướng, độ trung tâm của nó trở thành 0
 - Nút có thể có nhiều cạnh kết nối tới nó
- Để giải quyết vấn đề này, ta thêm độ lệch β vào các giá trị độ trung tâm của tất cả các nút



Elihu Katz

Véc tơ riêng trung tâm

$$C_{\text{Katz}}(v_i) = \boxed{\alpha \sum_{j=1}^n A_{j,i} C_{\text{Katz}}(v_j)} + \beta$$

Độ trung tâm Katz

Viết lại phương trình ở dạng véc tơ

$$\mathbf{C}_{\text{Katz}} = \alpha A^T \mathbf{C}_{\text{Katz}} + \beta \mathbf{1} \quad \begin{matrix} \nearrow \\ \text{vector với tất cả giá trị 1} \end{matrix}$$

Độ trung tâm Katz:

$$\mathbf{C}_{\text{Katz}} = \beta(\mathbf{I} - \alpha A^T)^{-1} \cdot \mathbf{1}$$

Độ trung tâm Katz

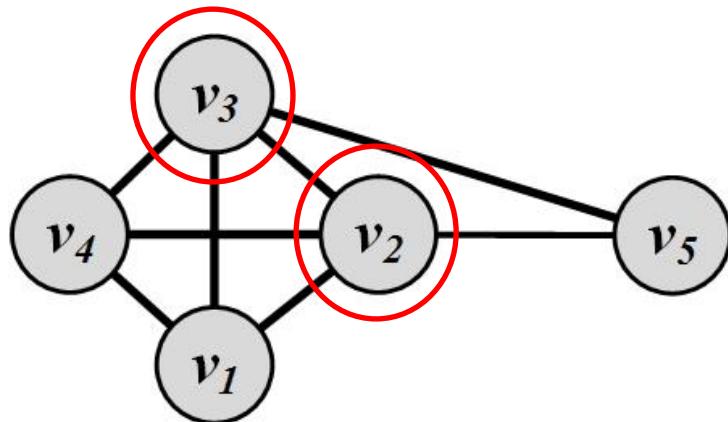
$$C_{\text{Katz}}(v_i) = \alpha \sum_{j=1}^n A_{j,i} C_{\text{Katz}}(v_j) + \beta$$

- Khi $\alpha=0$, véc tơ riêng độ trung tâm bị bỏ đi và tất cả các nút sẽ có cùng một giá trị độ trung tâm β
 - Khi α lớn lên, ảnh hưởng của β sẽ giảm
- Để ma trận $(I - \alpha A^T)$ khả nghịch, cần có
 - $\det(I - \alpha A^T) \neq 0$
 - Tổ chức lại, ta có $\det(A^T - \alpha^{-1} I) \neq 0$
 - Đây là phương trình đặc trưng cơ bản,
 - Phương trình đặt trung **trước hết** trở thành 0 khi giá trị riêng lớn nhất là α^{-1}

Giá trị riêng lớn nhất để tính (phương pháp power)

Trong thực hành, ta chọn $\alpha < 1/\lambda$, với λ là giá trị riêng lớn nhất của A^T

Thí dụ: độ trung tâm Katz



$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix} = A^T$$

- Các giá trị riêng là: -1.68, -1.0, -1.0, 0.35, 3.32
- Giả thuyết $\alpha=0.25 < \boxed{1/3.32}$ and $\beta = 0.2$

$$\mathbf{C}_{Katz} = \beta(\mathbf{I} - \alpha A^T)^{-1} \cdot \mathbf{1} = \begin{bmatrix} 1.14 \\ 1.31 \\ 1.31 \\ 1.14 \\ 0.85 \end{bmatrix}$$

Các nút quan trọng nhất!

PageRank

- Vấn đề với độ trung tâm Katz:
 - Trong đồ thị có hướng, một khi một nút trở thành một người có thẩm quyền (độ trung tâm cao), nó chuyển tất cả độ tập trung của nó tới tất cả các nút nằm trên liên kết ra của nó
- Điều này không chuẩn vì không phải tất cả những người được biết đến bởi một người nổi tiếng là người nổi tiếng
- Giải pháp?
 - Có thể chia giá trị của độ trung tâm chuyển qua bởi số các liên kết ra, nghĩa là độ ra của nút đó
 - Mỗi hàng xóm kết nối nhận một phần độ trung tâm của nút nguồn

PageRank

$$C_p(v_i) = \alpha \sum_{j=1}^n A_{j,i} \frac{C_p(v_j)}{d_j^{\text{out}}} + \beta$$

Điều gì xảy ra
nếu độ trung
tâm là 0?

$$\begin{cases} d_j^{\text{out}} > 0 \\ D = \text{diag}(d_1^{\text{out}}, d_2^{\text{out}}, \dots, d_n^{\text{out}}) \end{cases} \rightarrow \mathbf{C}_p = \alpha A^T D^{-1} \mathbf{C}_p + \beta \mathbf{1}$$

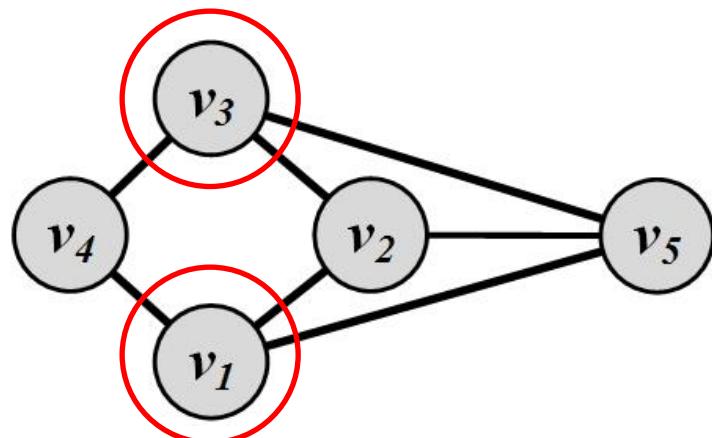


$$\mathbf{C}_p = \beta(\mathbf{I} - \alpha A^T D^{-1})^{-1} \cdot \mathbf{1}$$

Tương tự độ trung tâm Katz, trong thực tế, $\alpha < 1/\lambda$, với λ là giá trị riêng lớn nhất của $A^T D^{-1}$. Trong đồ thị vô hướng, giá trị riêng lớn nhất của $A^T D^{-1}$ là $\lambda = 1$; do đó, $\alpha < 1$.

Thí dụ PageRank

- Giả sử $\alpha=0.95 < 1$ và $\beta = 0.1$



$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix}$$

$$\mathbf{C}_p = \beta(\mathbf{I} - \alpha A^T D^{-1})^{-1} \cdot \mathbf{1} =$$

$$\begin{bmatrix} 2.14 \\ 2.13 \\ 2.14 \\ 1.45 \\ 2.13 \end{bmatrix}$$

Một số số đo mức trung tâm khác

- Mức trung tâm theo cách ta kết nối những người khác với nhau
 - Thí dụ: số đường kết nối ngắn nhất giữa các nút khác đi qua nút đang xét
- Mức trung tâm theo độ nhanh có thể tiến tới những người khác
 - Thí dụ: trung bình độ dài ngắn nhất tới các nút khác



Độ trung tâm cho một nhóm các nút

Mức trung tâm nhóm

- Tất cả các số đo trung tâm được xác định tới nay là số đo trung tâm cho mỗi nút đơn. Các số đo này có thể được tổng quát hóa cho nhóm các nút.
- Cách tiếp cận đơn giản là thay thế tất cả các nút trong nhóm bởi một nút duy nhất
 - Bỏ qua cấu trúc của nhóm.
- Ký hiệu S là tập các nút trong nhóm và $V - S$ là tập các nút ngoài nhóm

Mức trung tâm nhóm

I. Độ trung tâm nhóm

$$C_d^{\text{group}}(S) = |\{v_i \in V - S | v_i \text{ is connected to } v_j \in S\}|$$

– Normalization: Chia cho $|V - S|$

II. Độ trung tâm kết nối nhóm

$$C_b^{\text{group}}(S) = \sum_{s \neq t, s \notin S, t \notin S} \frac{\sigma_{st}(S)}{\sigma_{st}}$$

– Normalization: Chia cho $2 \binom{|V - S|}{2}$

Mức trung tâm nhóm

III. Mức trung tâm gần nhất nhóm

$$C_c^{\text{group}}(S) = \frac{1}{\bar{l}_S^{\text{group}}}$$

- Là trung bình khoảng cách từ các phần tử không là thành viên tới nhóm

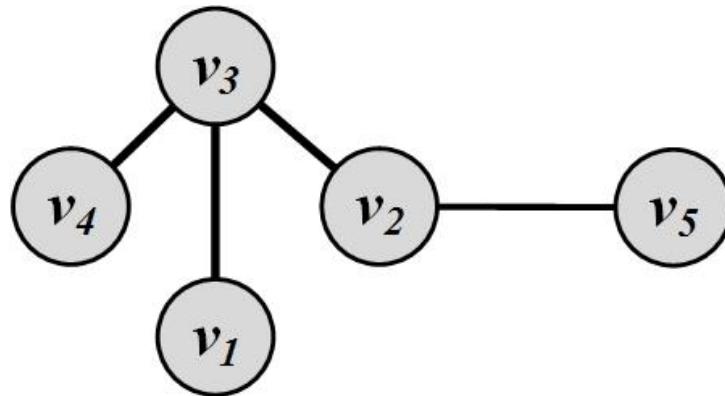
$$\bar{l}_S^{\text{group}} = \frac{1}{|V-S|} \sum_{v_j \notin S} l_{S,v_j}$$

$$l_{S,v_j} = \min_{v_i \in S} l_{v_i,v_j}$$

- Có thể sử dụng khoảng cách lớn nhất hay khoảng cách trung bình

Thí dụ độ trung tâm nhóm

- Xét nhóm $S = \{v_2, v_3\}$



- Độ trung tâm nhóm = **3**
- Độ trung tâm kết nối nhóm = **3**
- Độ trung tâm gần nhất nhóm = **1**



Các mẫu quan hệ bạn bè

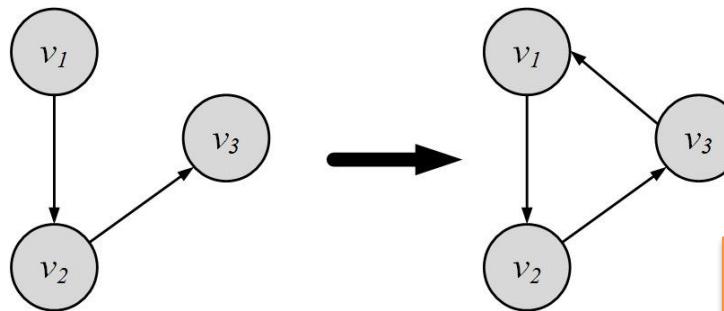
Transitivity/Reciprocity
Status/Balance



I. Transitivity và Reciprocity

Bắc cầu

- Trong toán học:
 - Quan hệ R có tính bắc cầu nếu: $aRb \wedge bRc \rightarrow aRc$



cRa hay aRc ?

- Trong mạng xã hội:
 - Bắc cầu là khi bạn của bạn tôi là bạn của tôi***
 - Bắc cầu trong 1 mạng xã hội dẫn đến một đồ thị dày, và tiến tới đồ thị đầy đủ
 - Ta có thể xác định cách đồ thị gần tới đồ thị đầy đủ bằng cách đo tính bắc cầu

Hệ số phân cụm toàn cục – Global Clustering coefficient

- **Hệ số phân cụm** đo tính bắc cầu của các đồ thị vô hướng
 - Đếm các đường dẫn độ dài 2 và kiểm tra xem có tồn tại cạnh thứ 3 không

$$C = \frac{|\text{Closed Paths of Length 2}|}{|\text{Paths of Length 2}|}$$

Khi đếm các tam giác, do mọi tam giác có 6 đường dẫn đóng chiều dài 2

$$C = \frac{(\text{Number of Triangles}) \times 6}{|\text{Paths of Length 2}|}$$

Hệ số phân cụm và bộ ba (triple)

Ta có thể viết

$$C = \frac{(\text{Number of Triangles}) \times 3}{\text{Number of Connected Triples of Nodes}}$$

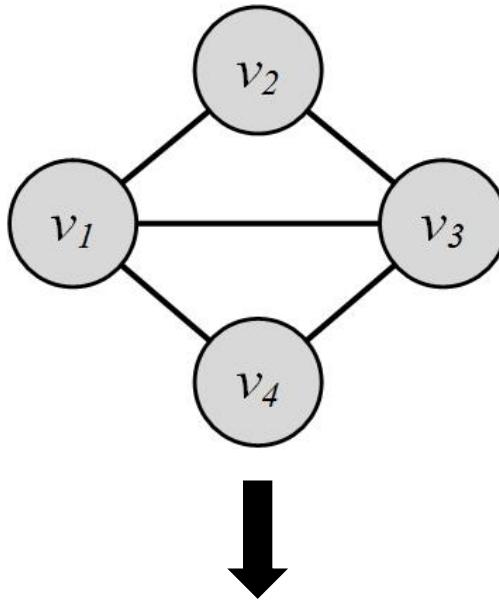
- **Bộ ba (Triple)**: tập 3 nút có thứ tự,
 - Kết nối bởi 2 (bộ 3 mở) cạnh hoặc
 - Ba cạnh (bộ 3 đóng)
- Một tam giác (triangle) có thể thiếu 1 cạnh nào trong 3 cạnh của nó
 - Một tam giác có **3 bộ ba**

$v_i v_j v_k$ và $v_j v_k v_i$ là các bộ ba khác nhau

- Cùng **các thành phần**
- Cái đầu tiên thiếu cạnh $e(v_k, v_i)$ và cái thứ 2 thiếu cạnh $e(v_i, v_j)$

$v_i v_j v_k$ và $v_k v_j v_i$ là cùng 1 bộ ba

Thí dụ: hệ số phân cụm toàn cục



$$\begin{aligned} C &= \frac{\text{(Number of Triangles)} \times 3}{\text{Number of Connected Triples of Nodes}} \\ &= \frac{2 \times 3}{2 \times 3 + \underbrace{2}_{v_2 v_1 v_4, v_2 v_3 v_4}} = 0.75. \end{aligned}$$

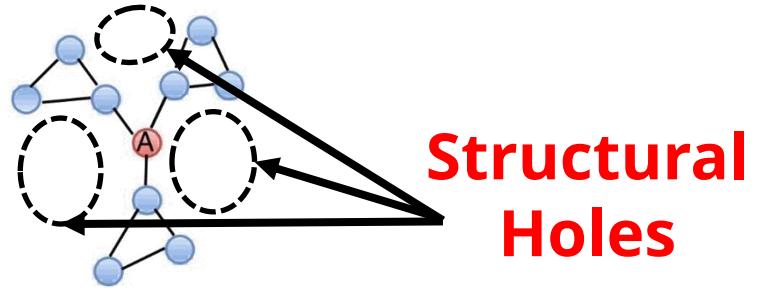
Hệ số phân cụm địa phương

- Hệ số phân cụm địa phương đo tính bắc cầu ở mức nút
 - Dùng cho các đồ thị vô hướng
 - Tính mức độ các lân cận của nút v (các nút kề cận với v) kết nối với nhau

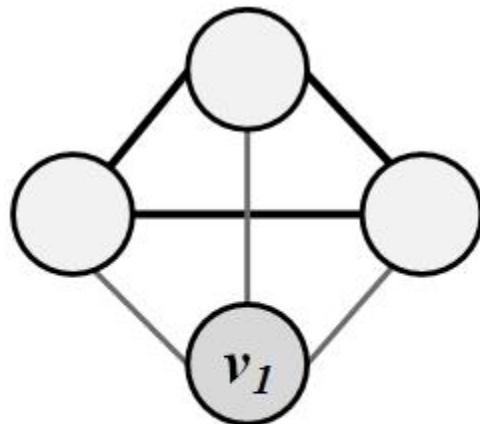
$$C(v_i) = \frac{\text{Number of Pairs of Neighbors of } v_i \text{ That Are Connected}}{\text{Number of Pairs of Neighbors of } v_i}.$$

Trong đồ thị vô hướng, mẫu số có thể viết là: $\binom{d_i}{2} = d_i(d_i - 1)/2$

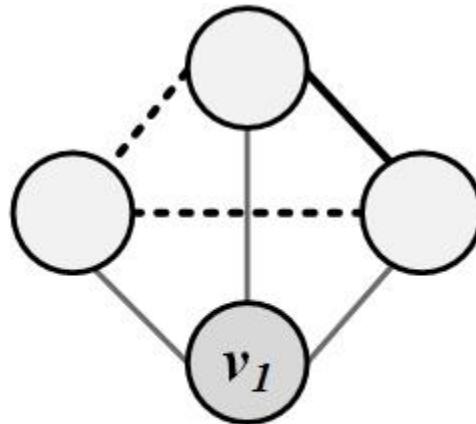
Cung cấp cách xác định các lỗ
cấu trúc



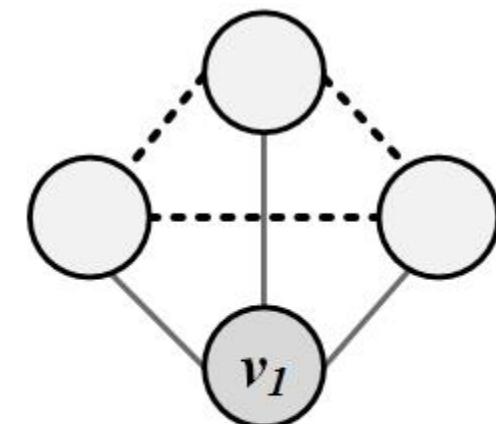
Thí dụ hệ số phân cụm địa phương



$$C(v_I) = 1$$



$$C(v_I) = 1/3$$



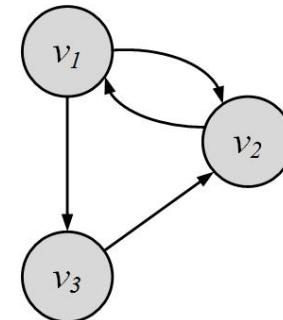
$$C(v_I) = 0$$

- Đường mỏng mô tả kết nối tới lân cận
- Đường đứt đoạn là các liên kết thiếu giữa các lân cận
- Đường đậm chỉ các kết nối của các lân cận
 - Khi không có các lân cận được kết nối $C = 0$
 - Khi tất cả các lân cận được kết nối $C = 1$

Tính có đi có lại - Reciprocity

*Nếu bạn trở thành bạn của tôi,
tôi sẽ là bạn của bạn*

- Tính có đi có lại là phiên bản đơn giản của tính bắc cầu
 - Nó xem xét các vòng lặp đóng chiều dài 2
- Nếu nút v kết nối tới nút u ,
 - u bằng kết nối tới v , thể hiện tính có đi có lại

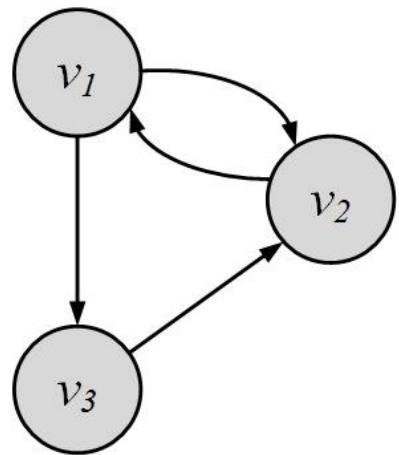


$$\begin{aligned}
 R &= \frac{\sum_{i,j,i < j} A_{i,j} A_{j,i}}{|E|/2}, \\
 &= \frac{2}{|E|} \sum_{i,j,i < j} A_{i,j} A_{j,i}, \\
 &= \frac{2}{|E|} \times \frac{1}{2} \text{Tr}(A^2), \\
 &= \frac{1}{|E|} \text{Tr}(A^2), \\
 &= \frac{1}{m} \text{Tr}(A^2).
 \end{aligned}$$

Còn $i = j$
thì sao?

$$\text{Tr}(A) = A_{1,1} + A_{2,2} + \cdots + A_{n,n} = \sum_{i=1}^n A_{i,i}$$

Thí dụ: tính có đi có lại



$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

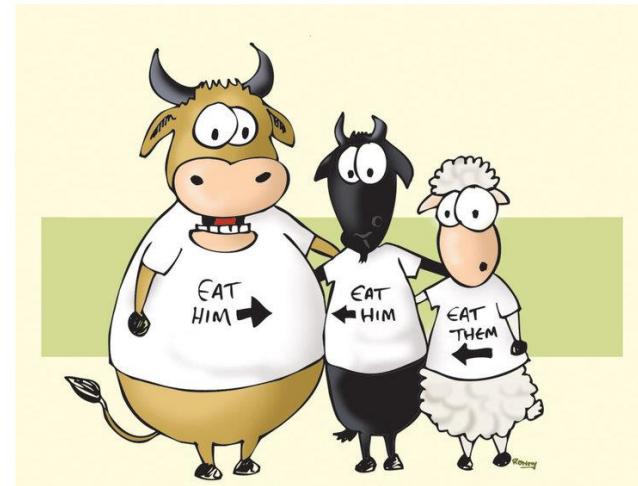


Các nút có đi có lại: v_1, v_2

$$R = \frac{1}{m} \text{Tr}(A^2) = \frac{1}{4} \text{Tr} \left(\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix} \right) = \frac{2}{4} = \frac{1}{2}.$$

II. Tính cân bằng (Balance) và địa vị (Status)

Đo sự nhất quán trong quan hệ bạn bè



Lý thuyết cân bằng xã hội

Lý thuyết cân bằng xã hội

- Nhất quán trong quan hệ bạn/thù giữa các cá nhân
- Một cách không chính thức, quan hệ bạn/thù là nhất quán khi

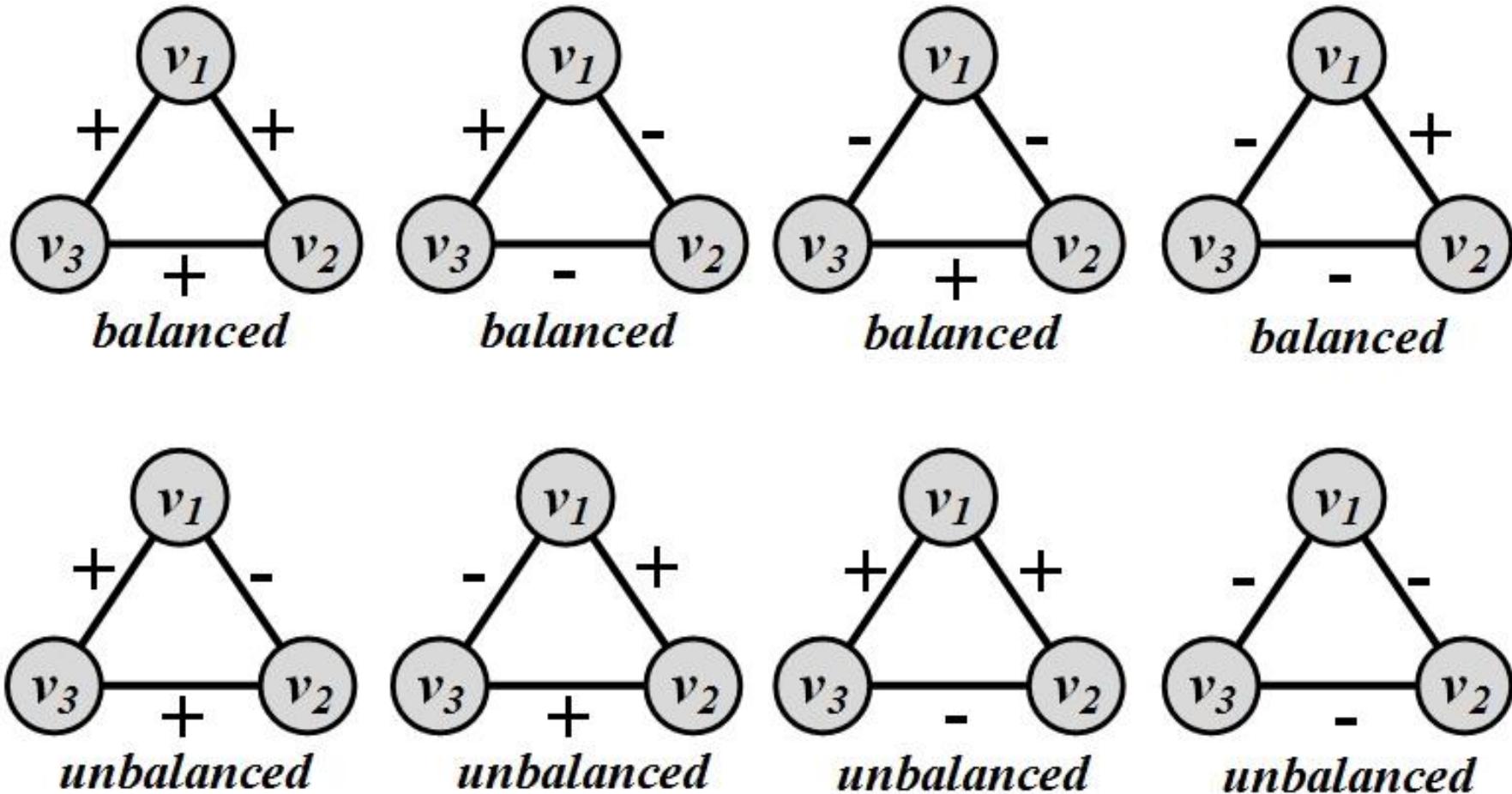
*The friend of my friend is my friend,
The friend of my enemy is my enemy,
The enemy of my enemy is my friend,
The enemy of my friend is my enemy.*

- Trong mạng
 - Các cạnh dương thể hiện quan hệ bạn ($w_{ij} = 1$)
 - Các cạnh âm thể hiện quan hệ thù ($w_{ij} = -1$)
- Tam giác với các nút i, j , và k , là cân bằng nếu và chỉ nếu

$$w_{ij}w_{jk}w_{ki} \geq 0.$$

- w_{ij} là giá trị của cạnh kết nối i và j

Lý thuyết cân bằng xã hội: các tổ hợp có thể



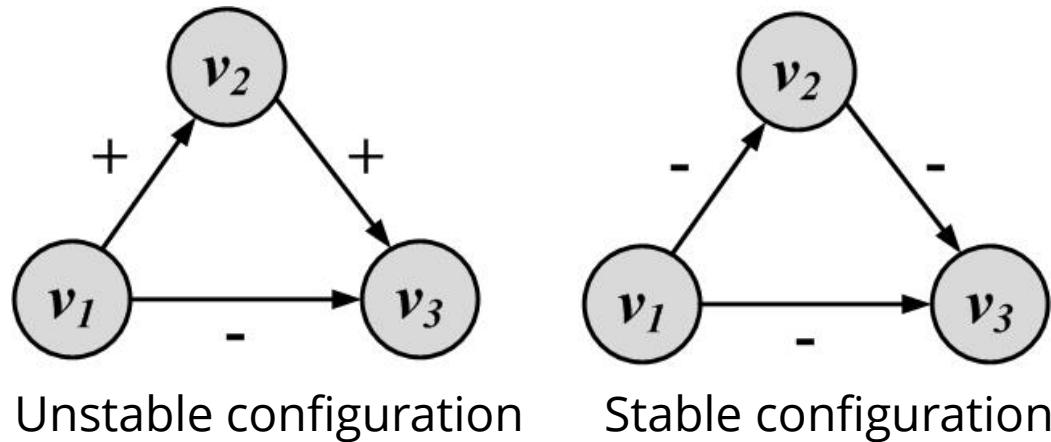
Đối với một vòng tròn, nếu tích của các giá trị cạnh dương, thì vòng tròn đó là cân bằng xã hội

Lý thuyết địa vị xã hội (Status)

- **Địa vị (Status):** cách một cá nhân được xếp hạng uy tín trong xã hội
- **Lý thuyết trạng thái xã hội:**
 - Mức độ nhất quán của các cá nhân trong việc gán địa vị tới các lân cận của nó
 - Một cách không chính thức,

If X has a higher status than Y and Y has a higher status than Z, then X should have a higher status than Z.

Thí dụ: lý thuyết địa vị xã hội



- Cạnh có hướng ‘+’ từ nút X tới nút Y thể hiện Y có địa vị cao hơn X và cạnh có hướng ‘-’ thể hiện điều ngược lại

Sự tương đồng

Hai nút trong mạng giống nhau thế nào?

Tương đương cấu trúc (Structural Equivalence)

Tương đương chính quy (Regular Equivalence)

Tương đương cấu trúc

- **Tương đương cấu trúc (Structural Equivalence):**
 - Ta xem xét hàng xóm **chung** của 2 nút;
 - Kích thước của hàng xóm chung xác định mức tương đồng giữa 2 nút.
- **Thí dụ:**
 - *Hai anh em có chung*
 - *Các anh chị, bố mẹ, ông bà...*
 - *Điều này chứng tỏ họ tương đồng,*
 - *Hai cá nhân ngẫu nhiên không có điểm chung và là không tương đồng.*

Tương đồng cấu trúc: định nghĩa

- **Tương đồng mức đỉnh (Vertex similarity):**

$$\sigma(v_i, v_j) = |N(v_i) \cap N(v_j)|$$

Chuẩn hóa?

Tương đồng Jaccard: $\sigma_{Jaccard}(v_i, v_j) = \frac{|N(v_i) \cap N(v_j)|}{|N(v_i) \cup N(v_j)|}$

Tương đồng Cosine: $\sigma_{Cosine}(v_i, v_j) = \frac{|N(v_i) \cap N(v_j)|}{\sqrt{|N(v_i)||N(v_j)|}}$

- Hàng xóm $N(v)$ thường không bao gồm chính v .

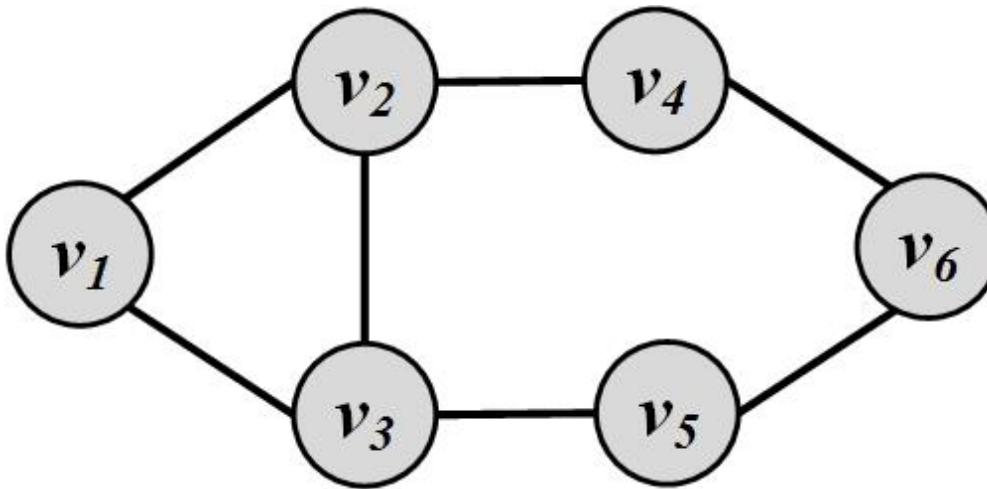
– **Có sai lầm gì đây?**

- Các nút kết nối không chia sẻ một hàng xóm và sẽ được gán mức tương đồng 0

– **Giải pháp:**

- Ta có thể giả thuyết hàng xóm chứa cả chính nút đó

Thí dụ: tính tương đồng

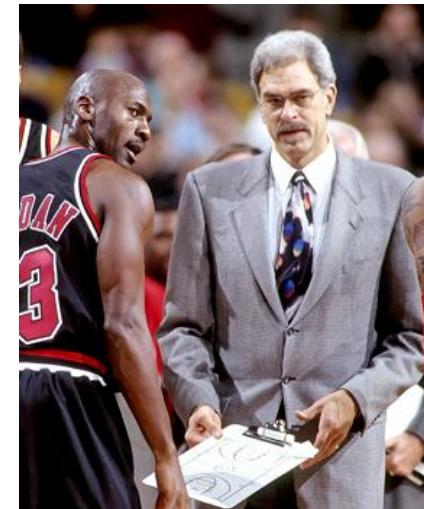


$$\sigma_{\text{Jaccard}}(v_2, v_5) = \frac{|\{v_1, v_3, v_4\} \cap \{v_3, v_6\}|}{|\{v_1, v_3, v_4, v_6\}|} = 0.25$$

$$\sigma_{\text{Cosine}}(v_2, v_5) = \frac{|\{v_1, v_3, v_4\} \cap \{v_3, v_6\}|}{\sqrt{|\{v_1, v_3, v_4\}| |\{v_3, v_6\}|}} = 0.40.$$

Tương đồng chính quy (Regular equivalence)

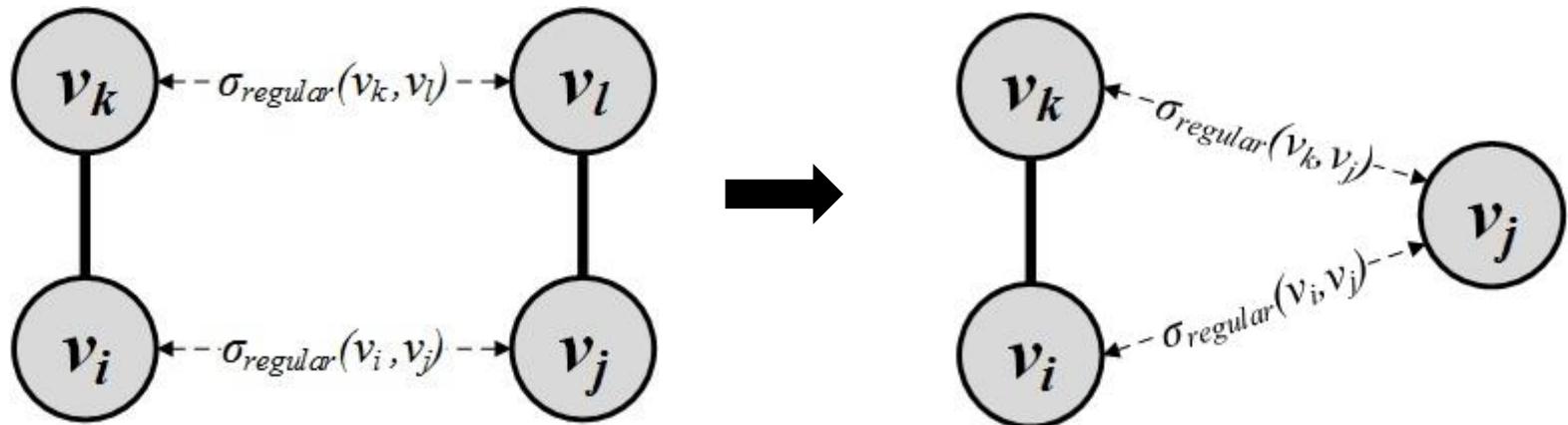
- Trong tương đồng chính quy,
 - Ta không xem xét các hàng xóm được chia sẻ giữa các cá nhân, mà
 - Mức tương đồng của chính các hàng xóm của chúng
- Thí dụ:
 - Các vận động viên tương đồng nhau không bởi vì họ biết nhau mà vì họ biết các cá nhân tương đồng như, các ông bầu, các huấn luyện viên, các người chơi khác, v.v.



Tương đồng chính quy

- v_i, v_j là tương đồng khi các hàng xóm của chúng v_k và v_l tương đồng

$$\sigma_{\text{regular}}(v_i, v_j) = \alpha \sum_{k,l} A_{i,k} A_{j,l} \sigma_{\text{regular}}(v_k, v_l)$$



- Phương trình (hình bên trái) khó giải vì nó tự tham chiếu, do đó ta giảm bớt định nghĩa bằng cách sử dụng hình bên phải

Tương đồng chính quy

- v_i và v_j tương đồng khi v_j tương đồng với hàng xóm của v_i, v_k

$$\sigma_{regular}(v_i, v_j) = \alpha \sum_k A_{i,k} \sigma_{Regular}(v_k, v_j)$$



- Viết dưới dạng véc tơ

$$\sigma_{regular} = \alpha A \sigma_{Regular}$$

Một đỉnh sẽ tương đồng cao với chính nó, để đảm bảo
điều này, ta cộng ma trận đơn vị vào phương trình



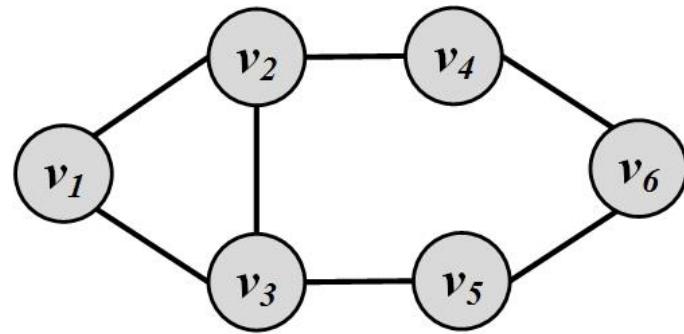
$$\sigma_{regular} = \alpha A \sigma_{Regular} + I$$



$$\sigma_{regular} = (I - \alpha A)^{-1}$$

Khi $\alpha < 1/\lambda_{max}$ ma trận khả nghịch

Thí dụ: tương đồng chính quy



$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

Giá trị riêng lớn nhất của A là 2.43

Lấy $\alpha = 0.3 < 1/2.43$

$$\sigma_{\text{regular}} = (I - 0.3A)^{-1} = \begin{bmatrix} 1.43 & 0.73 & 0.73 & 0.26 & 0.26 & 0.16 \\ 0.73 & 1.63 & 0.80 & 0.56 & 0.32 & 0.26 \\ 0.73 & 0.80 & 1.63 & 0.32 & 0.56 & 0.26 \\ 0.26 & 0.56 & 0.32 & 1.31 & 0.23 & 0.46 \\ 0.26 & 0.32 & 0.56 & 0.23 & 1.31 & 0.46 \\ 0.16 & 0.26 & 0.26 & 0.46 & 0.46 & 1.27 \end{bmatrix}$$

- Mỗi hàng/cột của ma trận này thể hiện sự tương đồng tới các đỉnh khác
- Đỉnh 1 tương đồng nhất (không kể chính nó) với đỉnh 2 và 3
- Nút 2 và 3 có mức tương đồng cao nhất (**tương đồng chính quy**)



4. Mô hình mạng

Tại sao ta nên sử dụng các mô hình mạng?



Facebook

Tháng 5/2011:

- **721 triệu** người dùng.
- Số bạn trung bình: **190**
- Tổng số **68.5 tỷ** quan hệ bạn bè

Tháng 9/2015:

- **1.35 tỷ** người dùng

1. Tiến trình bên dưới chính thức nào giúp hình thành các quan hệ bạn bè này?
2. Làm thế nào mà các quan hệ bạn bè tưởng như độc lập này hình thành mạng quan hệ bạn bè phức tạp?
3. Trong phương tiện truyền thông xã hội, có nhiều mạng có hàng triệu nút và hàng tỷ cạnh.
 - **Chúng phức tạp và khó để phân tích chúng**

Vậy, chúng ta làm gì?

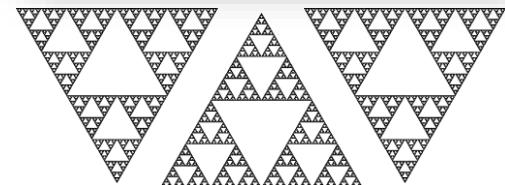
Thiết kế các mô hình xây dựng đồ thị

- Đồ thị được xây dựng nên tương tự với các mạng thế giới thực.

Nếu chúng ta có thể đảm bảo đồ thị được xây dựng tương tự mạng thế giới thực:

1. Ta có thể phân tích đồ thị mô phỏng thay vì các mạng thế giới thực (hiệu quả chi phí - **cost-efficient**)
2. Ta có thể hiểu hơn về các mạng thế giới thực nhờ cung cấp các giải thích toán học cụ thể; và
3. Ta có thể thực hiện các thử nghiệm có kiểm soát trên các mạng tổng hợp khi các mạng thế giới thực không sẵn dùng.

Các thuộc tính nào của các mạng thế giới thực nên được mô hình chính xác?



Trực giác cơ bản:

Hy vọng! Đầu ra phức tạp [mạng xã hội] được xây dựng bởi một tiến trình đơn giản

Một số mô hình mạng

- Đồ thị ngẫu nhiên (Random graph)
 - Mô hình đồ thị ngẫu nhiên giả thiết các cạnh (các quan hệ bạn bè) giữa các nút (các cá nhân) được hình thành một cách ngẫu nhiên
- Mô hình thế giới thu nhỏ (Small-World Model)
 - Mỗi cá nhân thường có một số giới hạn cố định kết nối, trong lý thuyết đồ thị, điều này nghĩa là những người dùng trong một mạng chính quy
- Mô hình đính kèm ưu tiên (Preferential Attachment Model)
 - Dựa trên giả thiết khi một người dùng mới kết nối vào mạng, xác suất kết nối tới các nút hiện tại tỷ lệ với bậc của các nút đó



5. Khai phá dữ liệu

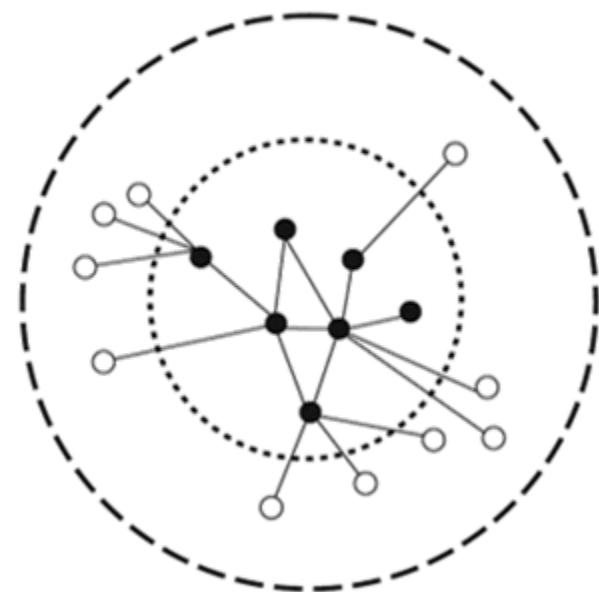
Thu thập dữ liệu phương tiện truyền thông xã hội

- Thu thập dữ liệu thô
 - Sử dụng các API
 - Flickr's: <https://www.flickr.com/services/api/>
 - Thu thập thông tin trực tiếp
- Sử dụng các kho lưu trữ (repository) được cung cấp
 - <http://socialcomputing.asu.edu>
 - <http://snap.Stanford.edu>
 - <https://github.com/caesar0301/awesome-public-datasets>

Tiền xử lý dữ liệu

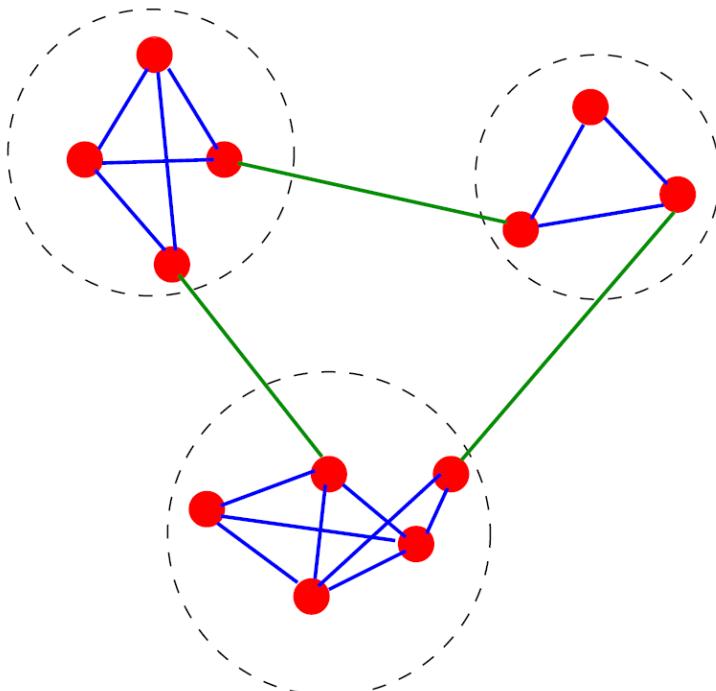
Lấy mẫu mạng xã hội:

- Bắt đầu với tập nhỏ các nút (các nút hạt nhân)
- Lấy mẫu
 - (a) các thành phần được kết nối;
 - (b) tập các nút (các cạnh) kết nối trực tiếp; hoặc
 - (c) tập các nút và các cạnh trong khoảng cách n từ chúng.



Các giải thuật khai phá dữ liệu

- Các giải thuật khai phá dữ liệu cơ bản (phân lớp, phân cụm, liên kết), khai phá dữ liệu văn bản, web
- Phát hiện, phân tích cộng đồng (community analysis)



Một đồ thị đơn trong đó 3 cộng đồng không tường minh được tìm thấy và được bọc bởi các vòng tròn nét đứt