

**ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**

Nguyễn Sỹ Quân

**DỰ ĐOÁN LIÊN KẾT ÂM, LIÊN KẾT DƯƠNG
TRONG MẠNG XÃ HỘI**

KHOÁ LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành: Công nghệ thông tin

HÀ NỘI - 2012

**ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**

Nguyễn Sỹ Quân

**DỰ ĐOÁN LIÊN KẾT ÂM, LIÊN KẾT DƯƠNG
TRONG MẠNG XÃ HỘI**

KHOÁ LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành: Công nghệ thông tin

Cán bộ hướng dẫn: PGS. TS. Hà Quang Thụy

Cán bộ đồng hướng dẫn: ThS. Nguyễn Tuấn Quang

HÀ NỘI - 2012

LỜI CẢM ƠN

Trước tiên, tôi xin được gửi lời cảm ơn chân thành và lòng biết ơn sâu sắc nhất tới PGS. TS. Hà Quang Thụy, Th.S.Nguyễn Tuấn Quang, những người đã hướng dẫn và chỉ bảo tận tình cho tôi trong suốt quá trình thực hiện khóa luận tốt nghiệp.

Tôi xin được chân thành cảm ơn các thầy cô, cán bộ trường Đại học Công nghệ - Đại học Quốc gia Hà Nội đã tạo cho tôi những điều kiện thuận lợi nhất trong suốt quá trình học tập và nghiên cứu.

Tôi cũng xin gửi lời cảm ơn tới các thầy cô, các anh chị và các bạn trong Phòng thí nghiệm KT-LAB đã chỉ bảo và giúp đỡ tôi rất nhiều về kiến thức chuyên môn và kỹ năng nghiên cứu để tôi hoàn thành tốt khóa luận tốt nghiệp.

Cuối cùng, tôi muốn gửi lời cảm ơn tới gia đình, bạn bè, người thân những người luôn bên cạnh động viên tôi trong suốt quá trình học tập, nghiên cứu cũng như thực hiện khóa luận tốt nghiệp.

Tôi xin chân thành cảm ơn!

Hà Nội, ngày 23 tháng 5 năm 2012

Sinh viên

Nguyễn Sỹ Quân

DỰ ĐOÁN LIÊN KẾT ÂM, LIÊN KẾT DƯƠNG TRONG MẠNG XÃ HỘI

Nguyễn Sỹ Quân

Khóa QHI-2008-I/CQ , ngành Công nghệ Thông tin

Tóm tắt Khóa luận tốt nghiệp:

Bài toán dự đoán liên kết trong các mạng có cấu trúc phức tạp nhận được nhiều sự quan tâm của các nhà khoa học trong lĩnh vực vật lý, khoa học máy tính và truyền thông... Đặc biệt trong lĩnh vực khoa học máy tính và truyền thông , bài toán dự đoán liên kết là một bài toán quan trọng và có ý nghĩa thực tiễn. Nó giúp cho việc xác định các thông tin bị thiếu, bị mất, xác định các tương tác giả mạo hay giúp chúng ta đánh giá các cơ chế mở rộng của mạng. Bài toán dự đoán liên kết cũng là một bài toán con trong bài toán phân tích mạng xã hội.

Trong khóa luận này, chúng tôi tập trung nghiên cứu các phương pháp dự đoán các mối quan hệ trong mạng xã hội và trình bày một mô hình một mô hình dự đoán liên kết âm dương kết hợp với việc sử dụng các đặc trưng về tính cá nhân để nâng cao kết quả cho quá trình dự đoán.

Từ khóa: predict links, ties strength, social network.

LỜI CAM ĐOAN

Tất cả các bài báo, khóa luận, tài liệu, công cụ phần mềm của các tác giả khác được sử dụng lại trong khóa luận này đều được chỉ dẫn tường minh về tác giả và đều có trong danh sách tài liệu tham khảo.

Tất cả các bài báo, khóa luận, tài liệu, công cụ phần mềm của các tác giả khác được sử dụng lại trong khóa luận này đều được chỉ dẫn tường minh về tác giả và đều có trong danh sách tài liệu tham khảo

Hà Nội, ngày 23 tháng 5 năm 2012

Sinh viên

Nguyễn Sỹ Quân

Mục lục

Mục lục	4
CHƯƠNG 1: BÀI TOÁN DỰ ĐOÁN LIÊN KẾT ÂM, LIÊN KẾT DƯƠNG TRONG MẠNG XÃ HỘI	2
1.1 Bài toán dự đoán liên kết trong mạng xã hội.....	2
1.1.1 Dự đoán liên kết trong mạng xã hội.....	2
1.1.2 Liên kết trong mạng xã hội.....	4
1.2 Bài toán dự đoán liên kết âm liên kết dương trong mạng xã hội.....	6
1.3 Kết luận chương 1	7
CHƯƠNG 2: CÁC PHƯƠNG PHÁP DỰ ĐOÁN LIÊN KẾT TRONG MẠNG XÃ HỘI	8
2.1 Phát biểu bài toán dự đoán liên kết âm, liên kết dương	8
2.2 Các thuật toán dự đoán liên kết dựa vào độ tương đồng.....	8
2.2.1 Các độ tương đồng cục bộ	9
2.2.2 Các độ tương đồng toàn cục	12
2.3 Các mô hình xác suất.....	15
2.3.1 Mô hình quan hệ xác suất	15
2.3.2 Mô hình quan hệ thực thể xác suất.....	17
2.4 Kết luận chương 2	18
CHƯƠNG 3: MÔ HÌNH DỰ ĐOÁN LIÊN KẾT ÂM, LIÊN KẾT DƯƠNG TRONG MẠNG XÃ HỘI	19
3.1 Lý thuyết cân bằng cấu trúc	19
3.1.1 Cân bằng cấu trúc.....	19
3.1.2 Đặc điểm về cấu trúc của mạng cân bằng	22
3.2 Lý thuyết trạng thái.....	25
3.3. Tính cá nhân trong mạng xã hội	26
3.4 Mô hình dự đoán liên kết âm, liên kết dương trong mạng xã hội.....	27
3.4.1 Đặc trưng của mô hình.	27
3.4.2 Phương pháp	28
3.5 Kết luận chương 3	29
CHƯƠNG 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ	30

4.1 Dữ liệu thực nghiệm	30
4.2 Môi trường thực nghiệm	31
4.3 Các công cụ phần mềm.....	31
4.4 Kết quả và đánh giá	32
4.5 Kết luận chương 4	34
KẾT LUẬN VÀ PHƯƠNG HƯỚNG	35

DANH SÁCH CÁC HÌNH VẼ

Hình 1. Ví dụ về mạng xã hội	2
Hình 2. Mạng xã hội Epinions	5
Hình 3. Mạng liên minh châu Âu thời kỳ 1872 - 1907	6
Hình 4. Bài toán dự đoán đầu của cung trên đồ thị.....	7
Hình 5. Khả năng hình thành các mối quan hệ khi có bạn chung.....	10
Hình 6. Cân bằng cấu trúc và không cân bằng cấu trúc	20
Hình 7. Ví dụ về cấu trúc cân bằng	22
Hình 8. Tính chất của cân bằng cấu trúc	23
Hình 9. Mô hình cân bằng cấu trúc và mô hình trạng thái	26
Hình 10. Các tam giác quan hệ trong đồ thị vô hướng.....	27
Hình 11. Các tam giác quan hệ trong đồ thị có hướng	28
Hình 12. Dự đoán liên kết dựa vào các đặc trưng.....	29
Hình 13. Minh họa dữ liệu đồ thị đã được gán nhãn các cung	31
Hình 14. Biểu đồ kết quả thực nghiệm.....	34

DANH SÁCH CÁC BẢNG

Bảng 1. Môi trường thực nghiệm	31
Bảng 2. Công cụ phần mềm	32
Bảng 3: Độ chính xác bộ phân lớp dự đoán.....	33

DANH SÁCH CÁC TỪ VIẾT TẮT

CN	Common Neighbour
AUC	Area Under the Curve
PRM	Probability Relation Model
PERM	Probability Entity Relation Model
RBNs	Relation Bayes Networks
RDNs	Relation Dependence Networks
HPI	Hub Promoted Index
HDI	Hub Depressed Index
LHN1	Leicht-Holm-Newman Index (local)
LHN2	Leicht-Holm-Newman Index (global)
PA	Preferential Attachment
AA	Adamic-Adar Index
RA	Resource Allocation
ACT	Average Commute Time
RWR	Random Walk with Restart

LỜI MỞ ĐẦU

Từ thế kỷ 20, lý thuyết đồ thị trở nên rất phổ biến vì ứng dụng rộng rãi của nó trong rất nhiều khía cạnh của đời sống như sinh học, xã hội học, công nghệ thông tin, mạng thông tin,... Vào năm 1930 bài toán phân tích mạng xã hội ra đời và trở thành chủ đề quan trọng nhất trong xã hội học. Trong thời đại bùng nổ thông tin hiện nay, số lượng và kích thước các mạng xã hội trực tuyến tăng lên không ngừng. Vì vậy, việc dự đoán liên kết trong mạng xã hội trực tuyến là một nhu cầu bức thiết trong thời điểm hiện nay, vì ứng dụng quan trọng của cộng đồng trong các lĩnh vực của đời sống xã hội, như khoa học máy tính, sinh học, kinh tế, chính trị,....

Nội dung chính của khóa luận là nghiên cứu về bài toán dự đoán liên kết âm, liên kết dương trong mạng xã hội, các phương pháp tiếp cận được sử dụng trong thời điểm hiện tại, từ đó trình bày giải pháp dự đoán liên kết trong mạng xã hội và từ đó cài đặt thử nghiệm thuật toán dự đoán liên kết âm liên kết dương trong mạng xã hội.

Khóa luận được chia thành các phần chính như sau:.

Chương 1: Giới thiệu tổng quan về bài toán dự đoán liên kết âm liên kết dương trong mạng xã hội.

Chương 2: Trình bày các phương pháp dự đoán liên kết trong mạng nói chung và mạng xã hội nói riêng .

Chương 3: Trình bày mô hình dự đoán liên kết âm và liên kết dương dựa vào lý thuyết cân bằng cấu trúc và lý thuyết trạng thái do Leskoves đề xuất năm 2010.

Chương 4: Trình bày thực nghiệm giải quyết mô hình trình bày ở Chương 3 và đánh giá thực nghiệm.

Kết luận và phương hướng: Tổng kết các nội dung chủ yếu của khóa luận và trình bày phương hướng nghiên cứu tiếp.

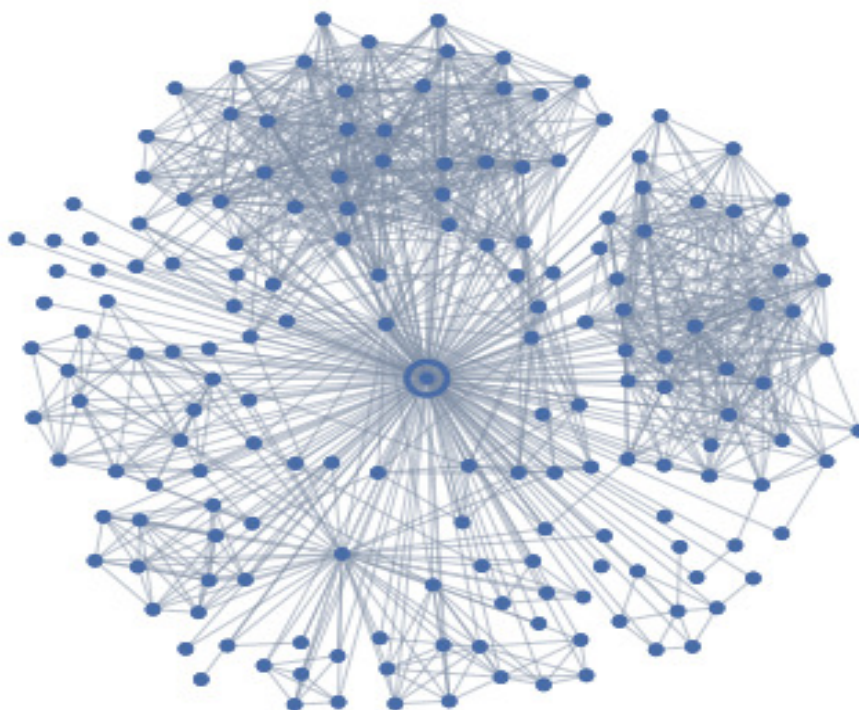
CHƯƠNG 1: BÀI TOÁN DỰ ĐOÁN LIÊN KẾT ÂM, LIÊN KẾT DƯƠNG TRONG MẠNG XÃ HỘI

1.1 Bài toán dự đoán liên kết trong mạng xã hội

1.1.1 Dự đoán liên kết trong mạng xã hội

Mạng xã hội là một mô hình mạng có tính chất xã hội được cấu tạo bởi các đỉnh và các cung, các đỉnh liên kết với nhau bởi một hoặc nhiều cung, thể hiện mối quan hệ cụ thể. Mỗi đỉnh là một thực thể trong mạng. Thực thể này có thể là một cá nhân, một tổ chức hay một quốc gia bất kỳ... Các thực thể trong mạng tương tác với nhau thông qua các liên kết. Các liên kết này có thể là quan hệ bạn bè, đồng nghiệp, cũng có thể là các quan hệ đối đầu thù địch hay các trao đổi tài chính, giao dịch...

Nhu cầu phân tích mạng xã hội đã được bắt đầu từ rất sớm từ những năm 1930 và ngày càng trở thành chủ đề quan trọng. Đặc biệt với sự phát triển hiện nay của mạng xã hội đã sản sinh ra một khối lượng dữ liệu khổng lồ, vì vậy bài toán phân tích mạng xã hội bài toán phân tích mạng xã hội trở thành bài toán phân tích mạng trong miền dữ liệu lớn. Đây là một bài toán khó và nhận được nhiều sự quan tâm của các nhà khoa học hiện nay.



Hình 1. Ví dụ về mạng xã hội

Các mạng xã hội trên các hệ thống trực tuyến có thể được mô tả bởi các đồ thị mạng, trong đó các đỉnh biểu diễn các thực thể và các liên kết biểu diễn các mối quan hệ hay sự tương tác giữa các đỉnh với nhau. Việc nghiên cứu các đồ thị mạng phức tạp là chủ đề thường xuyên của các nghiên cứu khoa học, đặc biệt là đồ thị mạng xã hội. Kết quả to lớn của các nghiên cứu đó là hiểu được quá trình phát triển và mở rộng của các mạng [18,20], sự tương tác giữa hình thái và chức năng của mạng [14,21], và các tính chất đặc điểm của mạng [8]. Một chủ đề khoa học quan trọng liên quan đến phân tích mạng được gọi là “*trích chọn thông tin*” [18,20] với mục đích tìm ra được những thông tin có ích và phục vụ cho các mục đích khác nhau từ nguồn dữ liệu không có cấu trúc khổng lồ như các mạng xã hội hay các nguồn thông tin trực tuyến khác.

Trong các mạng sinh học như mạng thức ăn, mạng tương tác protein-protein và mạng trao đổi vật chất, một liên kết chưa biết giữa hai đỉnh được chứng minh là tồn tại bằng kiến thức lĩnh vực đó hoặc tại các phòng nghiên cứu thường có chi phí rất cao. Với sự hiểu biết của chúng ta về các mạng là có hạn, ví dụ có đến 80% sự tương tác phân tử trong vi khuẩn nấm men và 99.7 % của con người vẫn còn chưa biết [5,15]. Thay vì việc mò mẫm kiểm tra các tương tác hay liên kết có thể tồn tại thì việc dự đoán các tương tác đó dựa trên các thông tin và các tương tác đã có rõ ràng sẽ giảm được nhiều công sức và chi phí nếu việc dự đoán đạt được một độ chính xác đủ lớn. Việc phân tích mạng xã hội cũng gặp phải nhiều khó khăn khi mà dữ liệu bị thiếu hoặc mất [9,17], khi đó các thuật toán dự đoán liên kết có thể đóng vai trò lớn cho bài toán phân tích mạng xã hội. Thêm vào đó, các dữ liệu xây dựng nên các mạng sinh học hay mạng xã hội có thể chứa các thông tin không chính xác hay các liên kết giả mạo [1,2]. Các thuật toán dự đoán liên kết có thể giúp cho việc phát hiện được các liên kết giả mạo này [19].

Ngoài việc giúp phân tích các mạng với dữ liệu bị thiếu, các thuật toán dự đoán liên kết còn giúp chúng ta có thể dự đoán được những mối quan hệ có thể xuất hiện trong tương lai trong quá trình mở rộng và phát triển của mạng. Ví dụ, trong các mạng xã hội trực tuyến, có những liên kết có thích hợp nhưng chưa được tồn tại có thể được gợi ý như một mối quan hệ triển vọng, nó có thể giúp người dùng tìm kiếm bạn mới và từ đó có thể làm tăng sự tin tưởng của người dùng đối với website đó. Các kỹ thuật tương tự cũng được đưa vào để đánh giá cơ chế tiến hóa của các mạng đã có. Ví dụ, có nhiều mô hình tiến hóa cho hình thái mạng Internet được đưa ra như mô hình sinh sản, mô hình dựa trên đặc trưng cấu trúc *k-core* ... [22]. Vì có quá nhiều các đặc trưng hình thái và chúng rất khó để đánh trọng số, chúng ta sẽ khó có thể đánh giá rằng mô hình nào tốt hơn mô hình nào. Nhận thấy rằng, mỗi mô hình về mặt lý thuyết tương ứng với một thuật toán dự đoán liên kết và do đó chúng ta có thể sử dụng các độ chính xác của các dự đoán để đánh giá hiệu suất của các mô hình khác nhau.

1.1.2 Liên kết trong mạng xã hội

Khi nhắc đến mạng xã hội, chúng ta nói đó là một mạng có tính chất xã hội. Một câu hỏi đặt ra là điều gì đã tạo ra tính chất đó. Các nghiên cứu về phương tiện xã hội đã chỉ ra rằng các mối quan hệ đã tạo ra tính chất xã hội cho các phương tiện xã hội nói chung và mạng xã hội nói riêng. Hơn nữa mỗi quan hệ trong đó có những vai trò và tính chất khác nhau[4]. Ví dụ, trong một nghiên cứu của Granovetter, khi tìm việc, người ta thường tìm được công việc thích hợp của mình thông qua những người quen biết sơ sài hơn là thông qua những người bạn thân [11], hay khi chúng ta gặp những vấn đề về sức khỏe, tình cảm thì những người thân hoặc bạn thân là những người quan tâm chăm sóc ta thường xuyên hơn [4]. Để hiểu rõ hơn chúng ta sẽ xem xét hai khía cạnh, thứ nhất liên kết giữa các nhóm với nhau trong một mạng đầy đủ và thứ hai là liên kết giữa cá nhân với cá nhân.

Đầu tiên chúng ta sẽ làm quen với một số khái niệm *độ mạnh liên kết*, *liên kết mạnh*, *liên kết yếu* [4,7,11]. Như đã biết, mỗi liên kết có một vai trò mà các liên kết khác nhau có thể có các vai trò khác nhau; để thuận tiện cho việc nghiên cứu và tính toán người ta đưa ra các khái niệm độ mạnh liên kết và dựa vào đó để chia các mối quan hệ thành hai loại: *liên kết mạnh* và *liên kết yếu*. Những mối quan hệ với bạn thân, người thân trong gia đình được gọi là liên kết mạnh. Ngược lại với liên kết mạnh đó là liên kết yếu, đây là những mối quan hệ với những người không thân thiết hoặc mới quen biết. Tuy nhiên vai trò của những liên kết yếu lại vô cùng quan trọng trong các mạng xã hội. Các liên kết yếu thường chia mạng ra thành những nhóm hay những cộng đồng riêng biệt có dựa vào các đặc điểm chung hay sở thích chung [7].

Đã có rất nhiều nghiên cứu tập trung vào chủ đề *độ mạnh liên kết* để xây dựng các ứng dụng hay các kế hoạch kinh doanh cho cá nhân hoặc tổ chức (Có tới hơn 7000 bài báo khoa học đã trích dẫn bài viết **"The Strength of Weak Ties"**[11], Google Scholar). Ví dụ các ngân hàng thường tìm những sự kết hợp thích hợp giữa liên kết mạnh và liên kết yếu của ngân hàng với các công ty họ hướng tới để đem lại lợi nhuận cao nhất có thể. Theo các nghiên cứu trong lĩnh vực kinh tế xã hội thì các liên kết yếu lại đem lại hiệu quả cao hơn khi các công ty tìm kiếm các hợp đồng mới[23]. Thêm vào đó việc nghiên cứu các liên kết trong xã hội còn được ứng dụng rất rộng rãi trong lĩnh vực y tế và giáo dục. Theo một nghiên cứu trong lĩnh vực y tế cộng đồng, những cô gái tuổi teen có số lượng bạn bè ít trong mạng bạn bè thường có xu hướng tự tử nhiều hơn những người có hệ số gom cụm cao[16].

Nhìn ở một góc nhìn khác, Jure Leskovec và Jon Kleinberg đã đưa ra các khái niệm về liên kết âm và liên kết dương [6]. Trong đó các mối quan hệ bạn bè, *người thân được coi là liên kết dương*, còn các *mối quan hệ đối đầu thù địch được coi là liên kết âm*. Vai trò của liên kết âm và liên kết dương là khá rõ ràng và quan trọng giống nhau trong các mạng xã hội, tuy nhiên phần lớn các nghiên cứu liên quan đến mạng xã hội chủ yếu tập trung vào liên kết dương [12]. Một vài năm gần đây đã có một số bài báo tập trung vào liên kết âm cũng như liên kết dương trong môi trường trực tuyến. Ví

dù, người dùng Wikipedia có thể bình chọn cho một ai đó hay bỏ phiếu chống lại một ai đó cho việc người đó ứng cử vào vị trí người quản trị (admin); với mạng Slashdot, một mạng xã hội chuyên bình chọn các sản phẩm công nghệ, người dùng có thể định nghĩa một người khác là *bạn* hay *kẻ thù* của mình hay mạng đánh giá sản phẩm khác như Epinions cho phép người dùng có thể đánh dấu rằng họ tin tưởng ai hay không tin tưởng ai. Việc đưa liên kết âm, liên kết dương vào trong các mạng xã hội giúp người dùng dễ dàng phân biệt có thể quan sát những người cần quan tâm đơn giản hơn.

Email Alerts

New review alert
Notify me when this member writes a new review

Web of Trust

bablondie25 trusts:

1. gothicdreams
2. sleeper54
3. Freak369
4. sojourseeker
5. becky2259

▶ View all 47 members whom bablondie25 trusts

bablondie25 is trusted by:

1. monkey_kju
2. bargainhunter
3. dannigirl5173
4. mommy1usa
5. gothicdreams


▶ View all 47 members who trust bablondie25

Web of Trust

You Trust bablondie25

Remove bablondie25 from your Trust list.

bablondie25's Profile



About bablondie25

TOP REVIEWER in Wellness & Beauty, Restaurants & Gourmet

[POPULAR AUTHOR] - Top 1000

Member: **bablondie25**

Epinions.com ID: **bablondie25**

Location: **Seattle Area**

Member Since: **Aug 25 '02**

Email Address: **bablondie25@hotmail.com**

Homepage: **My Funny T-Shirts**

Activity Summary

Reviews Written: **922**

Member Visits: **15,496**

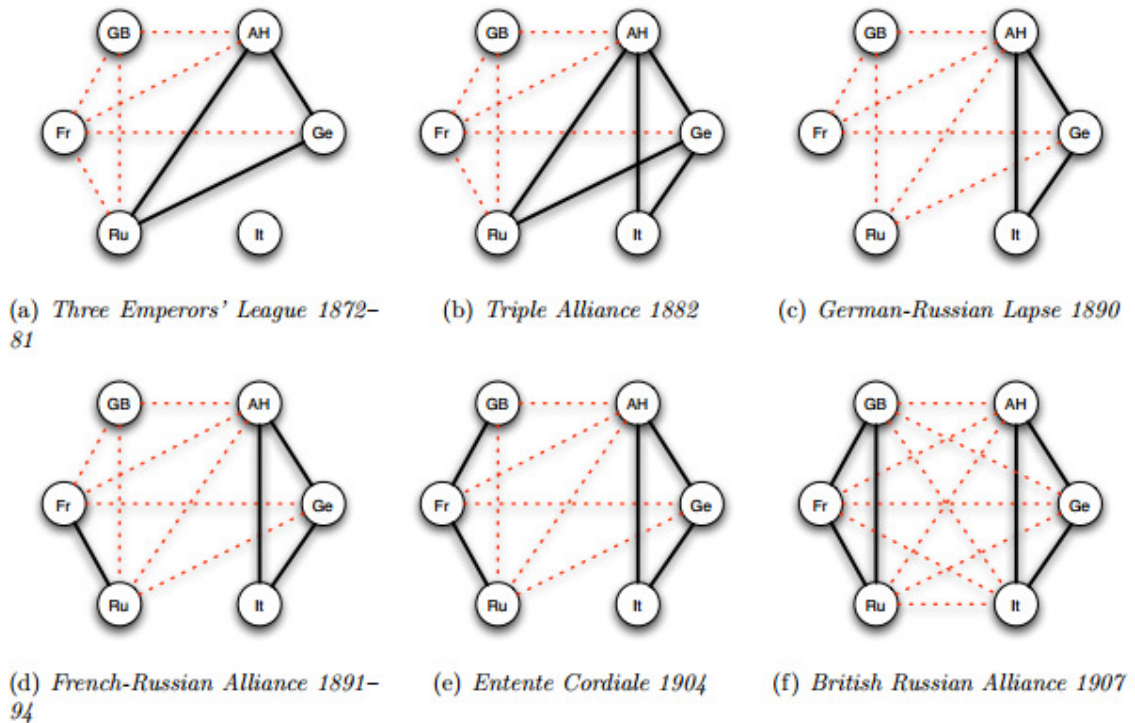
Total Visits: **198,121**

Favorite Websites: [My 2nd Funny Tee Shop](#)
[My Hubby's Photography](#)
[Cash Back 4 Shopping](#)

I stay at home and work online designing t-shirts, writing and managing my household. [more](#)

Hình 2. Mạng xã hội Epinions

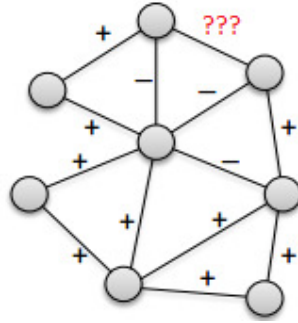
Việc nghiên cứu các liên kết âm, liên kết dương có rất nhiều ứng dụng trong thực tế, một ví dụ đơn giản là các hệ thống đánh giá sản phẩm trực tuyến **trust/distrust** như Epinions hay Slashdots. Một ứng dụng quan trọng khác được ứng dụng trong các hoạt động quan hệ quốc tế. Các mối quan hệ chính trị quốc tế được biểu diễn thông qua một mạng quan hệ quốc tế, mỗi nước là một đỉnh và các quan hệ là các cung. Mỗi cung thể hiện sự liên minh hay thù địch nhau bằng các liên kết dương và liên kết âm tương ứng. Hình 3 dưới đây thể hiện một đồ thị mạng quan hệ quốc tế khu vực châu Âu giai đoạn 1872 – 1907. Trong đó các cung đứt màu đỏ thể hiện mối quan hệ thù địch, các cung liền màu đen thể hiện mối quan hệ liên minh với nhau.



Hình 3. Mạng liên minh châu Âu thời kỳ 1872 - 1907

1.2 Bài toán dự đoán liên kết âm liên kết dương trong mạng xã hội

Trong phần này, chúng ta sẽ đi định nghĩa bài toán dự đoán liên kết âm, liên kết dương trong mạng xã hội. Chúng ta qui định rằng với liên kết dương chúng ta sẽ biểu diễn bằng một cung có dấu “+” trên đồ thị mạng và ngược lại với liên kết âm chúng ta sẽ biểu diễn bằng một cung có dấu “-” trên đồ thị mạng. Khi đó bài toán sẽ trở thành bài toán dự đoán dấu của cung trên đồ thị mạng xã hội. Bài toán dự đoán dấu của cung được định nghĩa như sau: Giả sử chúng ta có một mạng xã hội cho trước với tất cả các cung đều có dấu $+/-$, nhưng vì một lý do nào đó mà cung nối từ đỉnh u tới đỉnh v , ký hiệu là $s(u,v)$, bị ẩn đi mất. Làm thế nào chúng ta có thể suy ra được dấu của $s(u,v)$ dựa vào các thông tin có được từ đồ thị mạng cho sẵn? Giải quyết bài toán này chúng ta sẽ trả lời được câu hỏi các mẫu điển hình của dấu liên kết tương tác với nhau như thế nào, và cũng đưa ra các hướng tiếp cận cho những ứng dụng gợi ý quan điểm hay bạn bè trên mạng xã hội. Đây cũng là một bài toán dự đoán liên kết trong mạng xã hội [3] .



Hình 4. Bài toán dự đoán dấu của cung trên đồ thị

1.3 Kết luận chương 1

Chương 1 của khóa luận này đã đưa ra một số định nghĩa về bài toán dự đoán liên kết trong mạng và cụ thể hơn là bài toán dự đoán liên kết âm, liên kết dương trong mạng xã hội. Tiếp đó chương 1 cũng đã trình bày một số khái niệm liên quan đến bài toán dự đoán liên kết trong mạng xã hội như độ mạnh liên kết, liên kết mạnh, liên kết yếu hay liên kết âm liên kết dương... Qua đó chúng ta thấy được các ứng dụng thực tế của bài toán dự đoán liên kết âm, liên kết dương trong mạng xã hội.

Trong chương tiếp theo, khóa luận sẽ trình bày một số thuật toán để giải quyết bài toán dự đoán liên kết trong mạng xã hội và đặc biệt là mô hình để lý thuyết cho bài toán dự đoán liên kết âm, liên kết dương trong mạng xã hội.

CHƯƠNG 2: CÁC PHƯƠNG PHÁP DỰ ĐOÁN LIÊN KẾT TRONG MẠNG XÃ HỘI

Hiện nay có nhiều phương pháp để giải quyết bài toán dự đoán liên kết trong mạng xã hội. Trong chương này, khóa luận sẽ giới thiệu một số phương pháp và thuật toán để giải quyết bài toán dự đoán liên kết dựa vào độ tương đồng và mô hình xác suất. Đây là những phương pháp tiếp cận đơn giản và phổ biến và cho kết quả tương đối khả quan.

2.1 Phát biểu bài toán dự đoán liên kết âm, liên kết dương

Trước khi đi vào tìm hiểu các thuật toán chúng ta sẽ phát biểu lại bài toán dự đoán liên kết âm, liên kết dương dựa vào lý thuyết đồ thị. Bài toán được phát biểu như sau:

Đầu vào của bài toán:

- Cho một đồ thị có hướng hoặc không có hướng $G = (V, E)$ với V là tập các đỉnh của đồ thị, E là tập các cung của đồ thị.
- Mỗi cung $s(x,y)$ thuộc đồ thị biểu diễn cho một cung nối hai đỉnh x và y của đồ thị, các cung này đều có dấu dương hoặc âm.
- Nếu $s(x,y) = 1$, khi đó dấu của cung (x,y) là dương, $s(x,y) = -1$ thì dấu của cung đó là âm, $s(x,y) = 0$ khi không tồn tại cung (x,y) trên đồ thị.
- Đối với đồ thị có hướng khi viết $\bar{s}(x,y) = 1$ có nghĩa là cả hai hướng (x,y) và (y,x) đều mang dấu dương. Tương tự vậy với $\bar{s}(x,y) = -1$ thì cả hai hướng đều mang dấu âm. Và khi viết $\bar{s}(x,y) = 0$ thì tương ứng với các trường hợp còn lại.
- Giả sử rằng chúng ta có một cung (u,v) và dấu của nó $s(u,v)$ bị ẩn đi.

Đầu ra của bài toán:

- Dấu của cung $s(u,v)$ là dương hay là âm.

2.2 Các thuật toán dự đoán liên kết dựa vào độ tương đồng

Cơ chế đơn giản nhất của các phương pháp dự đoán liên kết là sử dụng các thuật toán dựa trên độ tương đồng, trong đó mỗi cặp các đỉnh x và y , được gán cho một điểm số s_{xy} , điểm số này được tính toán trực tiếp từ độ tương đồng giữa x và y . Tất cả các liên kết không nhìn thấy được xếp hạng dựa vào số điểm của chúng, và các liên kết giữa các đỉnh có độ tương đồng cao hơn thường có khả năng tồn tại cao hơn. Mặc dù các thuật toán dựa vào độ tương đồng rất đơn giản nhưng nó lại thuật toán được áp dụng rất nhiều. Các chỉ số tương đồng có thể đơn giản hay phức tạp và nó cũng có thể thích hợp hoặc không thích hợp với một số loại mạng khác nhau. Thêm vào đó, độ tương tự có thể được sử dụng bằng nhiều cách, ví dụ như các tích hợp

cục bộ dựa trên cơ chế của bộ lọc cộng tác (quá trình lọc thông tin sử dụng kỹ thuật kết hợp nhiều tác nhân, quan điểm, tài nguyên dữ liệu ...).

Độ tương đồng của các đỉnh có thể được xác định bằng các tính chất cơ bản của những đỉnh đó: hai đỉnh được coi là tương tự khi chúng có nhiều đặc điểm chung với nhau (các đặc trưng giống nhau) [10]. Tuy nhiên, nói chung thì các đặc tính của các đỉnh thường bị ẩn đi, vì thế chúng ta sẽ tập chung vào một số các hệ số tương đồng khác, chúng được gọi là sự tương đồng cấu trúc, chúng được phân loại theo các cách khác nhau như: cục bộ và toàn cục, độc lập tham số và phụ thuộc tham số, phụ thuộc đỉnh hay phụ thuộc cung,... Các hệ số tương đồng cũng được phân loại thành các loại phức tạp như tương đồng cấu trúc và tương đồng đều (tương đồng thường).

Trong phần này chúng ta sẽ tập trung vào các phương pháp đơn giản nhất, đó là 17 độ đo tương đồng chia làm 2 lớp chính: 10 độ tương đồng cục bộ và 7 độ tương đồng toàn cục.

2.2.1 Các độ tương đồng cục bộ

Trong phần này, khóa luận sẽ trình bày các độ tương đồng được dùng trong bài toán dự đoán liên kết trong mạng. Các độ đo này là các độ đo cục bộ, chúng chỉ ra sự tương đồng giữa hai đỉnh của đồ thị mạng dựa vào các tính chất chung của hai đỉnh đó mà chưa có sự tương đồng của các đỉnh liên kết với chúng [10].

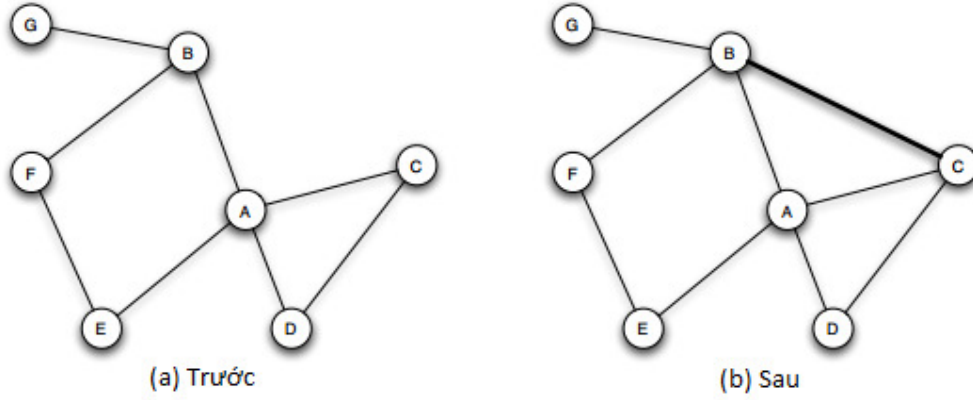
(1) *Hàng xóm, láng giềng hay bạn chung (CN – Common Neighbours):*

Cho một đỉnh x , và $\Gamma(x)$ là số đỉnh láng giềng của đỉnh x . Theo [3] thì với hai đỉnh x và y , khả năng hình thành một liên kết giữa chúng sẽ xảy ra nếu chúng có một hoặc nhiều bạn chung. Độ đo đơn giản nhất là hệ số trùng lặp hàng xóm được tính trực tiếp bằng các đếm các bạn chung và đánh dấu.

$$S_{xy}^{CN} = |\Gamma(x) \cap \Gamma(y)|. \quad (1)$$

trong đó, $|\mathbf{Q}|$ là lực lượng của tập \mathbf{Q} .

Khi đó dễ thấy rằng $s_{xy} = (A^2)_{xy}$, A là ma trận kề với $A_{xy} = 1$ nếu x và y có kết nối trực tiếp và $A_{xy} = 0$ với các trường hợp khác. Chú ý rằng $(A^2)_{xy}$ cũng là số đường đi khác của x và y với độ dài bằng 2. Newman [13, 7] đã sử dụng số lượng này để nghiên cứu về mạng cộng tác, và chỉ ra rằng các cộng tác có ích giữa các hàng xóm chung và xác suất cộng tác của hai nhà khoa học trong tương lai. Kossinets và Watts cũng đã phân tích các mạng xã hội với dữ liệu lớn và chỉ ra rằng với hai sinh viên có nhiều bạn chung với nhau thì có khả năng là bạn của nhau cao. Theo Jon Kleinberg và David Easley[7] thì hai người có khả năng trở thành của bạn của nhau khi họ có một người bạn thân là chung của nhau.



Hình 5. Khả năng hình thành các mối quan hệ khi có bạn chung [7]

(2) Độ đo Salton

Độ đo Salton được định nghĩa bằng công thức sau:

$$S_{xy}^{\text{Salton}} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{k_x \times k_y}}, \quad (2)$$

Trong đó k_x là bậc của đỉnh x . Độ đo Salton còn được gọi là độ tương tự cosine trong một số tài liệu khác.

(3) Độ đo Jaccard

Độ đo này được đề xuất bởi Jaccard cách đây hơn một trăm năm. Công thức của nó như sau:

$$S_{xy}^{\text{Jaccard}} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}, \quad (3)$$

(4) Độ đo Sørensen

Độ đo này được sử dụng chính trong việc nghiên cứu các mạng cộng đồng sinh học [26]:

$$S_{xy}^{\text{Sørensen}} = \frac{|\Gamma(x) \cap \Gamma(y)|}{k_x + k_y}, \quad (4)$$

(5) Độ đo Hub Promoted Index (HPI)

Độ đo này [26] được đề xuất cho việc xác định các hình thái trùng lặp trong các cặp chất gốc trong mạng trao đổi chất, và nó được định nghĩa như sau:

$$S_{xy}^{HPI} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\min(k_x, k_y)}, \quad (5)$$

Trong độ đo này, các liên kết gần với trung tâm có thể được gán điểm số cao vì mẫu số chỉ là bậc thấp hơn trong các hai bậc (cận trên).

(6) Độ đo Hub Depressed Index (HDI)

Giống với độ đo (5) chúng ta có một độ đo ngược lại với nó là độ đo HDI [26]:

$$S_{xy}^{HDI} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\max(k_x, k_y)}, \quad (6)$$

(7) Độ đo Leicht-Holme-Newman (LHN1).

Độ đo này tính độ tương tự cao cho cặp đỉnh bằng việc so sánh số bạn chung so với kỳ vọng số bạn chung có thể. Công thức như sau:

$$S_{xy}^{LHN1} = \frac{|\Gamma(x) \cap \Gamma(y)|}{k_x \times k_y}, \quad (7)$$

$k_x \times k_y$ là kỳ vọng của số bạn chung giữa đỉnh x và y .

(8) Độ đo Ràng buộc ưu tiên (PA).

Cơ chế ràng buộc ưu tiên có thể được sử dụng để sinh ra các mạng tự phát triển, trong đó xác suất cho một liên kết mới kết nối tới đỉnh x tỷ lệ với bậc của nó k_x . Một cơ chế tương tự cũng được sử dụng trong các mạng tự do nhưng không phát triển, trong đó tại mỗi thời điểm thì một liên kết cũ mất đi thì một liên kết mới hình thành. Xác suất của liên kết mới hình thành là tỷ lệ $k_x \times k_y$. Công thức của độ đo này được phát biểu như sau:

$$S_{xy}^S = k_x \times k_y \quad (8)$$

(9) Độ đo Adamic-Adar (AA).

Độ đo này làm tinh hơn so với độ đo CN và được xác định như sau:

$$S_{xy}^{AA} = \sum_{z \in |\Gamma(x) \cap \Gamma(y)|} \frac{1}{\log k_z} \quad (9)$$

(10) Độ đo phân phối tài nguyên (RA).

Độ đo này được thúc đẩy bởi cơ chế phân phối và cấp phát tài nguyên động trên các mạng phức tạp. Giả sử hai đỉnh x và y không có liên kết với nhau trực tiếp. Tuy nhiên đỉnh x có thể gửi một vài tài nguyên cho đỉnh y thông qua các bạn chung của họ với vai trò là những người vận chuyển. Trong trường hợp đơn giản nhất chúng ta giả sử người vận chuyển có một đơn vị tài nguyên, và sẽ phân phối bằng nhau cho tất cả các bạn chung của người đó. Khi đó độ tương tự giữa x và y được tính bằng số lượng tài nguyên được trao đổi giữa y và x , như sau:

$$S_{xy}^{RA} = \sum_{z \in |\Gamma(x) \cap \Gamma(y)|} \frac{1}{k_z} \quad (10)$$

Rõ ràng là độ đo này là đối xứng $S_{xy} = S_{yx}$. Mặc dù kết quả là khác nhau từ những cách tiếp cận khác nhau nhưng hai độ đo RA và AA rất giống nhau.

2.2.2 Các độ tương đồng toàn cục

Trong phần này chúng ta tiếp tục tìm hiểu về một số độ đo sự tương đồng toàn cục. Các độ đo này thường được tính toán dựa trên toàn bộ các đường đi có thể có giữa hai đỉnh [10].

(11) Độ đo Katz.

Độ đo này được tính toán dựa vào toàn bộ số đường đi, nó là kết quả tổng trực tiếp các đường đi và hàm mũ giảm dần theo chiều giảm của độ dài các đường đi. Công thức toán học của độ đo này như sau:

$$S_{xy}^{Katz} = \sum_{l=1}^{\infty} \beta^l \cdot |paths_{xy}^{<l>}| = \beta A_{xy} + \beta^2 (A^2)_{xy} + \beta^3 (A^3)_{xy} + \dots \quad (11)$$

Trong đó $paths_{xy}^{<l>}$ là tập tất cả các đường đi với độ dài là l giữa đỉnh x và đỉnh y . β là một tham số tự do (ví dụ là hệ số hãm) để điều khiển trọng số của các đường đi. Rõ ràng là với tham số β rất nhỏ thì độ đo này gần sát với độ đo CN, vì các đường đi dài thường là rất ít. Ma trận tương đồng có thể được viết như sau:

$$S^{Katz} = (I - \beta A)^{-1} - I. \quad (12)$$

Chú ý rằng tham số β phải thấp hơn nghịch đảo lớn nhất của trị số ma trận A để chắc chắn rằng công thức (11) là hội tụ.

(12) Độ đo Leicht-Holme-Newman (LHN2).

Độ đo này khác một chút so với độ đo Katz. Nó dựa trên nguyên lý rằng hai đỉnh là tương tự nhau nếu các hàng xóm trực tiếp của chúng là tương tự nhau, công thức được biểu diễn dưới dạng:

$$S = \phi AS + \psi I = \psi(I - \phi A)^{-1} = \psi(I + \phi A + \phi^2 A^2 + \dots), (13)$$

trong đó ϕ và ψ là các tham số tự do để điều khiển sự cân bằng giữa hai thành phần tương đồng nhau. Nếu đặt $\psi = 1$ thì nó rất giống với độ đo Katz. $(A^l)_{xy}$ bằng số đường đi có độ dài là l từ x đến y . Kỳ vọng của $(A^l)_{xy}$ ký hiệu là $E[(A^l)_{xy}]$, được tính bằng $(k_x k_y / 2M) \lambda_1^{l-1}$ trong đó λ_1 là trị số lớn nhất của A và M là tổng số cung trong mạng. Thay $(A^l)_{xy}$ trong công thức (13) bằng $(A^l)_{xy} / E[(A^l)_{xy}]$ chúng ta có công thức:

$$S_{xy}^{LHN2} = \delta_{xy} + \frac{2M}{k_x k_y} \sum_{l=0}^{\infty} \phi^l \lambda_1^{1-l} (A^l)_{xy} = \left[1 - \frac{2M\lambda_1}{k_x k_y} \right] \delta_{xy} + \frac{2M\lambda_1}{k_x k_y} \left[\left(I - \frac{\phi}{\lambda_1} \right)^{-1} \right]_{xy} (14)$$

Trong đó δ_{xy} là hàm *Kronnecker*.

(13) Độ đo thời gian trao đổi lẫn nhau trung bình (ACT).

Cho $m(x,y)$ là số bước trung bình để di chuyển bắt đầu từ đỉnh x đến đỉnh y , khi đó thời gian trao đổi trung bình giữa hai đỉnh x và y là :

$$n(x,y) = m(x,y) + m(y,x) (15)$$

Áp dụng ma trận Laplace giả nghịch đảo (pseudoinverse), $L^+(L = D - A)$ ta có:

$$n(x,y) = M(l_{xx}^+ + l_{yy}^+ - 2l_{xy}^+) (16)$$

trong đó l_{xy}^+ là ký hiệu cho đầu vào của L^+ . Giả thiết là nếu hai đỉnh tương đồng nhau nhiều hơn nếu chúng có một thời gian trao đổi lẫn nhau nhỏ hơn, từ đó độ tương đồng giữa hai đỉnh x và y có thể được bằng nghịch đảo của $n(x,y)$, cụ thể như sau:

$$S_{xy}^{ACT} = \frac{1}{l_{xx}^+ + l_{yy}^+ - 2l_{xy}^+} (17)$$

(14) Độ đo Cosine dựa vào L^+

Đây là độ đo dựa vào tích vô hướng. Trong không gian Euclidean $v_x = \Lambda^{\frac{1}{2}} U^T \vec{e}_x$, trong đó U là ma trận trực chuẩn được tạo bởi vector đặc trưng của L^+ được sắp xếp giảm dần theo thứ tự tương ứng của các giá trị đặc trưng λ_x , $\Lambda = \text{diag}(\lambda_x)$, \vec{e}_x là vector $N \times 1$ với phần tử thứ x có giá trị bằng 1 và các phần tử khác là 0, và T là ma trận chuyển vị, giả nghịch đảo của ma trận Laplace là tích vô hướng của các vector đỉnh, $l_{xy}^+ = v_x^T v_y$. Do đó, độ tương đồng cosine được xác định bằng cosin của các vectors đỉnh như sau:

$$S_{xy}^{\cos+} = \cos(x, y)^+ = \frac{v_x^T v_y}{|v_x| \cdot |v_y|} = \frac{l_{xy}^+}{\sqrt{l_{xx}^+ \cdot l_{yy}^+}} \quad (18)$$

(15) Độ đo Random Walk with Restart (RWR).

Độ đo này là một ứng dụng có hướng của thuật toán PageRank. Giả sử một người đi bộ ngẫu nhiên từ đỉnh x , người sẽ di chuyển lặp lại đến một hàng xóm ngẫu nhiên với xác suất là c và trở lại x với xác suất là $1-c$. Kí hiệu q_{xy} là xác suất ngẫu nhiên người đi bộ xác định đỉnh y là vị trí dừng lại, khi đó chúng ta có:

$$\vec{q}_x = cP^T \vec{q}_x + (1-c) \vec{e}_x \quad (19)$$

Trong đó P là ma trận chuyển tiếp với $P_{xy} = 1/k_x$ nếu x và y có kết nối, và $P_{xy} = 0$ với các trường hợp còn lại. Dễ dàng biến đổi công thức trên thành:

$$\vec{q}_x = (1-c) (I - cP^T)^{-1} \vec{e}_x \quad (20)$$

Khi đó độ đo RWR được tính như sau:

$$S_{xy}^{RWR} = q_{xy} + q_{yx} \quad (21)$$

(16) Độ đo SimRank.

Tương tự như độ đo LHN2, SimRank được định nghĩa theo giả thiết rằng hai đỉnh là tương đồng với nhau nếu chúng được kết nối tới các đỉnh tương đồng nhau:

$$S_{xy}^{SimRank} = C \cdot \frac{\sum_{z \in \Gamma(x)} \sum_{z' \in \Gamma(y)} S_{zz'}^{SimRank}}{k_x \cdot k_y} \quad (22)$$

Trong đó $S_{xx} = 1$ và $C \in [0, 1]$ là hệ số phân rã. Độ đo SimRank có thể giải thích cho quá trình đi bộ ngẫu nhiên (random-walk).

(17) Chỉ số Matrix Forest Index (MFI)

Chỉ số này được định nghĩa bằng công thức:

$$S = (I + L)^{-1} \quad (23)$$

Nhận xét: Khi so sánh các độ tương đồng cục bộ và toàn cục toàn bộ thông tin tô-pô, cho dù các độ tương đồng toàn cục có thể đạt được độ chính xác cho việc dự đoán cao hơn các độ tương đồng cục bộ, nhưng nó lại có 2 nhược điểm lớn:

- (i) Việc tính toán các độ tương đồng toàn cục thường mất nhiều thời gian và nó thường không thể làm được với các mạng có dữ liệu lớn, đặc biệt là mạng xã hội.
- (ii) Đôi khi các thông tin tô-pô toàn cục lại không thể sử dụng được, đặc biệt nếu chúng ta muốn thực thi một thuật toán theo phương pháp phân tán.

2.3 Các mô hình xác suất

Chúng ta đã tìm hiểu các thuật toán sử dụng độ tương đồng và một số độ tương đồng. Trong phần này chúng ta sẽ tìm hiểu một số mô hình xác suất để giải quyết bài toán dự đoán liên kết trong các mạng phức tạp nói chung và mạng xã hội nói riêng. Các mô hình xác suất tập trung vào việc trừu tượng cấu trúc cơ bản của mạng mà chúng ta quan sát, và sau đó dự đoán các liên kết bị thiếu bằng các mô hình học máy. Cho một đồ thị mạng đích $G = (V, E)$, mô hình xác suất sẽ xây dựng một hàm mục tiêu tối ưu để thiết lập một mô hình tĩnh với một nhóm các tham số Θ , chúng có thể là các tham số thích hợp nhất cho dữ liệu quan sát của mạng mục tiêu. Sau đó xác suất tồn tại của một liên kết chưa tồn tại (i, j) sẽ được xác định bởi xác suất $P(A_{ij} = 1 | \Theta)$. Trong phần này của khóa luận sẽ giới thiệu hai mô hình xác suất chính là mô hình *Quan hệ xác suất (PRM)* và mô hình *Quan hệ thực thể xác suất (PREM)*. Trong một số tài liệu, thuật ngữ PRM chỉ được cho là một mạng quan hệ Bayes.[10]

2.3.1 Mô hình quan hệ xác suất

RPM biểu diễn một phân phối xác suất thông qua các thuộc tính của tập dữ liệu quan hệ. Chúng cho phép các thuộc tính của một đối tượng phụ thuộc vào xác suất của cả các thuộc tính khác của đối tượng đó và các thuộc tính của các đối tượng liên quan. Khác với các mô hình đồ thị truyền thống sử dụng một đồ thị để mô hình các mối quan hệ giữa các thuộc tính và các thực thể đồng nhất, RPM chứa ba đồ thị: Đồ thị dữ liệu G_D , đồ thị mô hình G_M và đồ thị suy luận $G_I[xx]$.

Cho đồ thị $G_D = (V_D, E_D)$ biểu diễn cho mạng đầu vào, trong đó các đỉnh là các đối tượng trong miền dữ liệu và các cung biểu diễn cho các mối quan hệ giữa các đối tượng đó. Mỗi đỉnh $v_i \in V_D$ và $e_j \in E_D$ được kết hợp với một kiểu $T(v_i) = t_{v_i}$, $T(e_j) = t_{e_j}$. Mỗi mục $t \in T$ có một số các thuộc tính X^t . Vì thế, mỗi đối tượng v_i và liên kết e_j đều liên quan đến một tập thuộc tính, $x_{v_i}^{t_{v_i}}$ và $x_{e_j}^{t_{e_j}}$, được xác định bởi kiểu của chúng. Một mô hình PRM biểu diễn một phân phối xác suất thông qua tất cả các giá trị thuộc tính trong đồ thị dữ liệu, $x = \{x_{v_i}^{t_{v_i}} : v_i \in V_D, T(v_i) = t_{v_i}\} \cup \{x_{e_j}^{t_{e_j}} : e_j \in E_D, T(e_j) = t_{e_j}\}$. Ví dụ hệ thống đăng ký môn học của sinh viên, sinh viên và các môn học là các đỉnh, các cung thể hiện mối quan hệ lựa chọn môn học giữa sinh viên và môn học đó. Rõ ràng là ở đây có hai kiểu đỉnh, cụ thể là sinh viên và môn học. Và kiểu sinh viên thì có

bốn thuộc tính là: lớp, tuổi, giới tính và chuyên ngành đang theo, trong khi các môn học có năm thuộc tính là: giáo viên, năm học, thời gian, mô tả và môn học thuộc nhóm nào.

Đồ thị tiếp theo là đồ thị mô hình $G_M = (V_M, E_M)$ biểu diễn sự phụ thuộc giữa thuộc tính với cấp độ của các kiểu mục. Thuộc tính của một mục có thể phụ thuộc xác suất vào các thuộc tính khác của cùng mục đó. Như vậy G_M có hai phần: cấu trúc phụ thuộc giữa các kiểu thuộc tính và phân phối xác suất có điều kiện (CPD) liên quan đến các đỉnh trong đồ thị G_M . Đồ thị mô hình có thể được xác định dựa vào đồ thị dữ liệu thông qua các thuật toán học máy.

Cuối cùng là đồ thị suy luận $G_I = (V_I, E_I)$ biểu diễn các phụ thuộc xác suất giữa tất cả các biến trong một tập test đơn. Nó có thể là kết quả của sự kết hợp giữa đồ thị dữ liệu và đồ thị mô hình. Vì vậy cấu trúc của đồ thị suy luận được xác định dựa vào hai đồ thị đã cho là đồ thị dữ liệu và đồ thị mô hình.

Với các phương pháp biểu diễn đồ thị mô hình G_M khác nhau, các cách thức học và suy luận chúng ta có thể chia mô hình quan hệ xác suất PRM ra làm ba nhóm như sau: Các mạng quan hệ Bayes (RBNs), các mạng quan hệ Markov (RMNs), và các mạng quan hệ phụ thuộc (RDNs). Tiếp theo khóa luận sẽ tìm hiểu ba nhóm mạng quan hệ này.

- Mạng quan hệ Bayes (RBNs): là mô hình đồ thị thể hiện xác suất. Các đỉnh của đồ thị biểu diễn các biến. Các cung biểu diễn các quan hệ phụ thuộc giữa các biến và cha của nó. Nếu có một cung từ đỉnh A tới đỉnh B, thì biến B phụ thuộc trực tiếp vào biến A và A được gọi là cha của đỉnh B. Nếu với mỗi biến $X_i, i \in \{1, \dots, N\}$, tập hợp các biến cha được ký hiệu bởi $parents(X_i)$ thì phân phối có điều kiện phụ thuộc của các biến là tích của các phân phối địa phương :

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i \mid parents(X_i)) \quad (24)$$

Nếu X_i không có cha, ta nói rằng phân phối xác suất địa phương của nó là không có điều kiện, nếu không, nó là có điều kiện. Nếu biến được biểu diễn bởi một đỉnh được quan sát, thì ta nói rằng đỉnh đó là một đỉnh hiển nhiên (evidence node).

Một ưu điểm của mạng Bayes là, về mặt trực quan, con người có thể hiểu các quan hệ phụ thuộc trực tiếp và các phân phối địa phương dễ dàng hơn là phân phối có điều kiện phụ thuộc hoàn chỉnh.

- Mạng quan hệ Markov (RMNs): Một mạng quan hệ Markov sử dụng một đồ thị vô hướng, và một tập các hàm tiềm năng Φ để biểu diễn phân phối thông qua các thuộc tính của các kiểu mục. Khi đó xác suất của x được tính như sau:

$$p(x) = \frac{1}{Z} \prod_{c \in C} \Phi_c(x_c), \quad (25)$$

Trong đó $c \in C$ là các thuộc tính. Z là hằng số chuẩn hóa.

- Mạng quan hệ phụ thuộc (RDNs): Mạng quan hệ phụ thuộc là mô hình mạng hai chiều với một tập hợp các phân phối xác suất có điều kiện, chúng được biểu diễn bằng các quan hệ phụ thuộc lẫn nhau. Các mô hình RDNs thường sử dụng phương pháp học pseudo-likelihood để lấy xấp xỉ gần đúng cho các phân phối xác suất của các giá trị thuộc tính trong một tập dữ liệu quan hệ. Công thức tính pseudo-likelihood cho đồ thị dữ liệu G_D với kiểu mục t , và các thuộc tính của kiểu t X^t , và đỉnh v với cung e như sau:

$$PL(G_D; \Theta) = \prod_{t \in T} \prod_{X^t} \prod_{T(v)=t} p(x_{v_i}^t | pa_{v_i}; \Theta) \prod_{T(e)=t} p(x_{e_i}^t | pa_{e_i}; \Theta) \quad (26)$$

2.3.2 Mô hình quan hệ thực thể xác suất

Một dạng đặc biệt của mô hình quan hệ thực thể xác suất là đồ thị có hướng không có chu trình PERM (hay còn gọi là DAPER), nó sử dụng các cung có hướng để biểu diễn các mối quan hệ giữa các thuộc tính [24]. DAPER tạo ra các mối với các lớp đối tượng trong mô hình ngôn ngữ, và giúp cho việc biểu diễn các phân phối xác suất dễ dàng hơn. Mô hình DAPER gồm sáu lớp như sau:

- Lớp thực thể: Là các lớp đối tượng trong thể giới thực.
- Lớp thuộc tính: Mô tả các thuộc tính của các lớp thực thể hoặc các mối quan hệ.
- Lớp quan hệ: Biểu diễn sự tương tác giữa các thực thể với nhau.
- Lớp cung: Biểu diễn xác suất phụ thuộc giữa các thuộc tính tương ứng.
- Lớp phân phối cục bộ: xây dựng cấu trúc phân phối cục bộ cho các thuộc tính tương ứng với các lớp thuộc tính.
- Lớp ràng buộc: chỉ ra đồ thị được suy luận như thế nào cho mô hình DAPER

Mô hình DAPER có thể được sử dụng cho trường hợp mà cấu trúc quan hệ không chắc chắn. Mô hình DAPER là mô hình có chi phí đắt hơn các mô hình PRMs.

2.4 Kết luận chương 2

Trong chương này chúng ta đã tìm hiểu một số phương pháp và mô hình dự đoán liên kết trong mạng. Các phương pháp dựa trên độ tương đồng khá đơn giản và cũng đạt được hiệu quả khá tốt. Các mô hình xác suất cũng được đánh giá là phương pháp tốt phù hợp với nhiều loại mô hình mạng.

Trong chương tiếp theo của khóa luận chúng ta sẽ xem xét một mô hình cụ thể.

CHƯƠNG 3: MÔ HÌNH DỰ ĐOÁN LIÊN KẾT ÂM, LIÊN KẾT DƯƠNG TRONG MẠNG XÃ HỘI

Trong chương này chúng ta sẽ tìm hiểu một mô hình dự đoán liên kết âm liên kết dương trong mạng xã hội được Jure Leskoves và các cộng sự đề[1] đề xuất năm 2010. Mô hình là sự kết hợp của mô hình xác suất với kỹ thuật học máy hồi quy, giữa lý thuyết cân bằng cấu trúc và lý thuyết trạng thái. Với mô hình này kết quả thực nghiệm cho được kết quả độ chính xác là 80%, một kết quả tương đối cao.

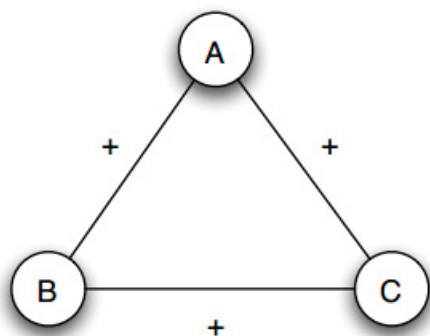
3.1 Lý thuyết cân bằng cấu trúc

3.1.1 Cân bằng cấu trúc

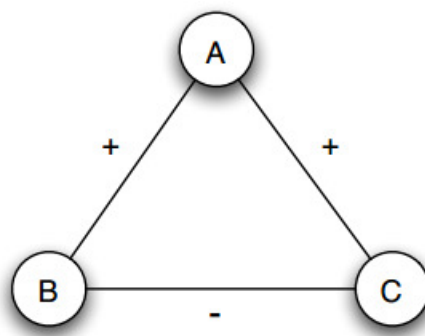
Chúng ta tập trung vào vấn đề này có lẽ là mô hình cơ bản nhất của liên kết âm và liên kết dương, vì nó nắm bắt được ý tưởng thiết yếu. Giả sử rằng chúng ta có một mạng xã hội trên một tập người, trong đó mỗi người đều biết đến tất cả những người khác – vì vậy chúng ta có một cung nối giữa các cặp của các đỉnh. Một mạng lưới được gọi là một nhóm (clique), hoặc một đồ thị đầy đủ. Sau đó chúng ta gán nhãn cho mỗi cung một trong hai nhãn + hoặc –; nhãn + chỉ ra rằng hai thiết bị đầu cuối là bạn bè, trong khi nhãn – chỉ ra rằng hai thiết bị đầu cuối là kẻ thù.

Chú ý rằng, kể từ khi có một cung kết nối vào mỗi cặp, chúng ta giả định rằng mỗi cặp là bạn hoặc thù – không thể có hai người không quan tâm đến một người khác hoặc không biết về nhau. Như vậy, mô hình chúng tôi đang xem xét làm cho một nhóm người nhỏ đủ để có mức độ nhận thức lẫn nhau (ví dụ một phòng học, một công ty nhỏ, một đội thể thao, một tình huynh đệ hoặc hội phụ nữ), hoặc cho một thiết lập như quan hệ quốc tế mà trong đó các đỉnh là các quốc gia và mỗi quốc gia có một vị trí ngoại giao chính thức đối với mỗi quốc gia khác.

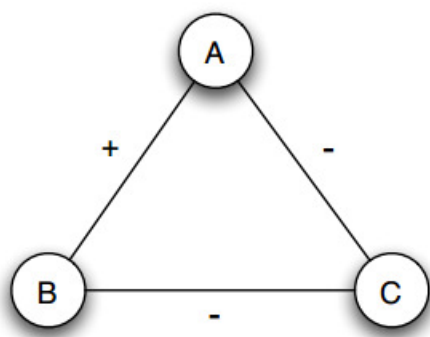
Các nguyên tắc cơ bản cân bằng cấu trúc dựa trên lý thuyết về tâm lý xã hội học quay lại thời kỳ làm việc của Heider trong những năm 1940, và tổng quát hóa và mở rộng đến ngôn ngữ của đồ họa bắt đầu với công việc của Cartwright và Harary trong những năm 1950. Ý tưởng quan trọng là ý tưởng sau đây. Nếu chúng ta nhìn vào bất kỳ hai người trong nhóm, cung giữa chúng có thể được gán nhãn + hoặc –, có nghĩa là họ là bạn hay thù. Nhưng khi chúng ta nhìn vào bộ ba người tại một thời điểm, một số cấu hình của + và – là xã hội chung và tâm lý chung của hầu hết những người khác. Đặc biệt, có bốn cách khác nhau (đối xứng) để gán nhãn ba cung giữa ba người với nhãn + và – (nhìn hình 5.1).



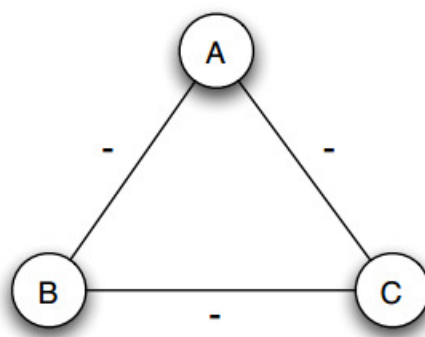
(a) *A, B, and C are mutual friends: balanced.*



(b) *A is friends with B and C, but they don't get along with each other: not balanced.*



(c) *A and B are friends with C as a mutual enemy: balanced.*



(d) *A, B, and C are mutual enemies: not balanced.*

Hình 6. Cân bằng cấu trúc và không cân bằng cấu trúc[7]

Chúng ta có thể phân biệt giữa bốn khả năng như sau:

Với tập hợp những người A, B và C, có ba dấu cộng giữa chúng (như hình 6(a)): nó tương ứng với ba người cùng là bạn bè.

Có một dấu cộng duy nhất và hai dấu trừ trong quan hệ giữa ba người: nó có nghĩa rằng hai trong ba người là bạn, và họ có một kẻ thù chung thứ ba. (Hình 6(c)).

Hai nhãn khác có thể của tam giác A, B và C giới thiệu một vài số lượng của tâm lý căng thẳng hoặc không ổn định trong mối quan hệ. Một tam giác với hai dấu cộng và một dấu trừ tương ứng (như hình 6(b)) để một người A là bạn với mỗi người B và C, nhưng B và C không là bạn của nhau. Trong kiểu vị trí này, có thể tiềm ẩn một tác dụng thúc đẩy A để có B và C trở thành bạn bè (vì vậy biến nhãn cung B-C thành +); hoặc ngược lại để A cung một trong hai B và C để chống lại người kia (biến một nhãn của cung ngoài của A thành -).

Tương tự như vậy, có một sự không ổn định trong cấu hình nơi mà mỗi A, B và C là kẻ thù của nhau (như trong hình 6(d)). Trong trường hợp này, có một tác dụng thúc đẩy hai trong ba người trở thành “team up” chống lại người thứ ba (biến một trong nhân của các cung thành –).

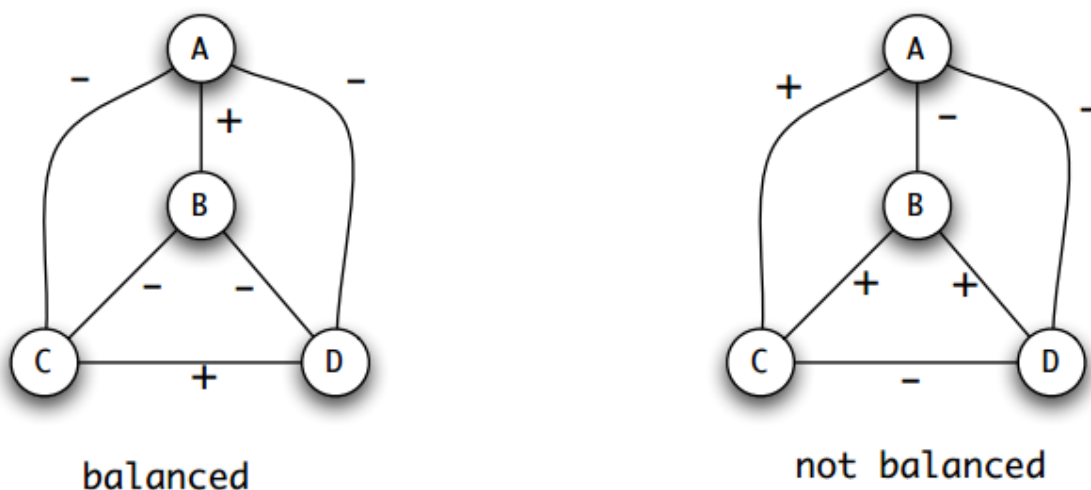
Dựa trên lý luận này, chúng ta sẽ quy định một tam giác với một trong ba nhân là nhân + là sự cân bằng, vì chúng là tự do trong nguồn không ổn định, và chúng ta quy định tam giác với không hoặc hai dấu + là không cân bằng. Sự tranh luận của các nhà khoa học về *sự cân bằng cấu trúc vì tam giác không cân bằng là nguồn gốc của sự căng thẳng hoặc sự bất hòa tâm lý, người ta cố gắng để giảm thiểu chúng trong quan hệ cá nhân của họ.*

Định nghĩa cân bằng cấu trúc cho mạng. Chúng ta đã nói về sự cân bằng cấu trúc cho các nhóm ba đỉnh. Nhưng nó dễ dàng để tạo ra một định nghĩa tổng quát cho các đồ thị đầy đủ trên một số tùy ý các đỉnh với các cung được gán nhãn là + và – .

Cụ thể, chúng ta nói rằng một đồ thị đầy đủ có nhân là cân bằng nếu một trong các tam giác của nó cân bằng – có nghĩa là, nếu nó tuân theo những điều sau đây:

Tính chất cân bằng cấu trúc: với mỗi tập ba đỉnh, nếu chúng ta xem xét ba cung kết nối giữa chúng, hoặc là tất cả ba cung đó được gán nhãn +, hoặc là chính xác một trong số chúng được gán nhãn +.

Ví dụ: xem xét hai nhân được gán cho bốn đỉnh mạng trong Hình 7. Đồ thị bên trái là cân bằng, vì chúng ta có thể kiểm tra mỗi tập ba đỉnh thỏa mãn tính chất cân bằng cấu trúc bên trên. Đồ thị bên phải là không cân bằng, vì giữa ba đỉnh A, B, C có chính xác hai cung được gán nhãn +, vi phạm tính chất cân bằng cấu trúc. (Tam giác B, C, D cũng vi phạm điều kiện.)



Hình 7. Ví dụ về cấu trúc cân bằng[7]

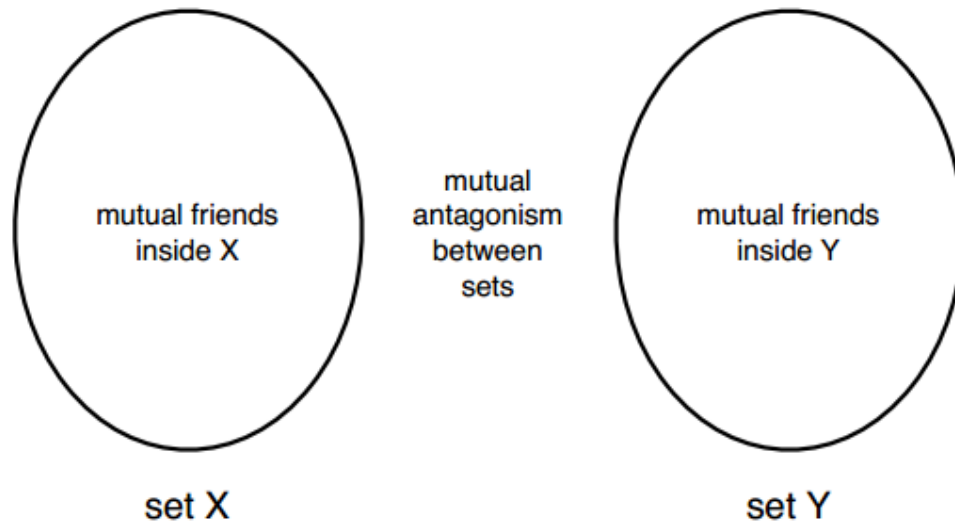
Định nghĩa của chúng ta về cân bằng mạng ở đây biểu diễn giới hạn của một hệ thống xã hội mà đã loại bỏ tất cả các tam giác không cân bằng. Như vậy, nó là một định nghĩa khá cực đoan – ví dụ, thay vào đó có thể đề xuất một định nghĩa mà chỉ yêu cầu có *ít nhất một số tỷ lệ phần trăm* của các tam giác cân bằng, *cho phép một vài tam giác có thể không cân bằng*. Tuy nhiên, phiên bản với tất cả các tam giác cân bằng là bước cơ bản đầu tiên trong việc suy nghĩ về các khái niệm này; và chúng ta sẽ nhìn thấy tiếp theo, thì ra có cấu trúc toán học rất thú vị mà trong thực tế giúp cho việc đưa ra các kết luận của nhiều mô hình phức tạp.

3.1.2 Đặc điểm về cấu trúc của mạng cân bằng

Ở một mức độ chung, một mạng cân bằng (tức là đồ thị đầy đủ được gán nhãn cân bằng) trông giống gì? Với bất kỳ ví dụ cụ thể, chúng ta có thể kiểm tra tất cả các tam giác để đảm bảo rằng chúng tuân theo các điều kiện cân bằng; nhưng nó sẽ tốt hơn để có một mô tả khái niệm đơn giản “ of what a balanced network looks like in general”.

Một cách cho một mạng cân bằng là nếu mọi người thích những người khác; trong trường hợp này, tất cả các tam giác có ba cung được gán nhãn +. Đồ thị bên trái trong hình 5.2 đưa ra một cách hơi phức tạp cho một mạng để được cân bằng: nó bao gồm hai nhóm bạn (A, B và C, D), với liên kết dương giữa những người trong các nhóm khác nhau. Điều này là đúng trong phần nói chung: giả sử chúng ta có một đồ thị đầy đủ được gán nhãn trong đó các đỉnh có thể được chia thành hai nhóm X và Y, như vậy mỗi cặp của các đỉnh trong X thích nhau, mỗi cặp của các đỉnh trong Y thích

nhau, và mỗi người trong X là kẻ thù của mỗi người trong Y. (Xem sơ đồ minh họa trong hình 5.3).



Hình 8. Tính chất của cân bằng cấu trúc[7]

Bạn có thể kiểm tra rằng một mạng như vậy là cân bằng: vì một tam giác bao gồm toàn bộ những trong một nhóm đều có ba nhãn +, và một tam giác với hai người trong một nhóm và một người ở nhóm còn lại có chính xác một nhãn +.

Vì vậy, ở đây mô tả hai cách cơ bản để đạt được cấu trúc cân bằng: hoặc là tất cả mọi người thích nhau; hoặc là trên thế giới bao gồm hai nhóm bạn bè với sự đối lập hoàn toàn giữa các nhóm. Thực tế đáng ngạc nhiên là: đây là các cách duy nhất để có một mạng cân bằng. Chúng ta trình bày rõ ràng thực tế này trong định lý cân bằng, được cung cấp bởi Frank Harary trong năm 1953:

Định lý cân bằng: *nếu một đồ thị đầy đủ có gán nhãn được cân bằng, hoặc là tất cả các cặp của các đỉnh là bạn, hoặc là các đỉnh có thể phân thành hai nhóm X và Y, như vậy mỗi cặp của các đỉnh trong X thích nhau, mỗi cặp của các đỉnh trong Y thích nhau, và mỗi người trong X là kẻ thù của mỗi người trong Y.*

Định lý cân bằng không phải là tất cả thực tế rõ ràng, “nor should it be initially clear why it is true”. Về cơ bản, chúng ta đang nói về tính chất hoàn toàn cục bộ, là tính chất cân bằng cấu trúc, để áp dụng chỉ cho ba đỉnh trong một thời điểm, và cho thấy rằng nó bao hàm một tính chất toàn cầu mạnh mẽ: hoặc là tất cả mọi người trên thế giới ở trong cùng một nhóm hoặc là thế giới bị chia thành hai phe phái đối lập.

Chúng ta sẽ cho thấy nhận định này trong thực tế là đúng.

Chứng minh định lý cân bằng: chứng minh định lý yêu cầu một bằng chứng: giả sử chúng ta có một đồ thị đầy đủ được gán nhãn tùy ý, giả sử rằng nó cân bằng, và kết luận rằng hoặc tất cả mọi người là bạn, hoặc có một tập X và Y như được mô tả trong yêu cầu. Nhớ rằng, chúng ta làm đã việc thông qua một bằng chứng ở chương 3 là tốt, khi chúng ta sử dụng các giả định đơn giản về “triadic closure” trong một mạng xã hội để kết luận tất cả các bridge cục bộ trong mạng phải là mối quan hệ yếu. Bằng chứng của chúng ta ở đây sẽ dài hơn một chút, nhưng vẫn rất tự nhiên và không phức tạp – chúng ta sử dụng trực tiếp định nghĩa cân bằng để có được kết luận của định lý.

Để bắt đầu, giả sử chúng ta có một đồ thị đầy đủ gán nhãn, và tất cả chúng ta biết rằng nó cân bằng. Chúng ta phải hiển thị rằng nó có một cấu trúc như trong định lý. Nếu nó không có các cung tiêu cực, thì mọi người là bạn và “we’re all set”. Nếu không, có ít nhất một cung tiêu cực, và bằng cách nào đó chúng ta cần phải đi đến với một bộ phận của các đỉnh trong tập X và Y, với sự đối lập hoàn toàn giữa chúng. Khó khăn là, biết rất ít về đồ thị của chính nó hơn là nó cân bằng, nó không rõ ràng để giả thiết làm thế nào chúng ta xác định X và Y.

Chọn bất kỳ một đỉnh trong mạng – chúng ta gọi nó là A – và xem xét mọi thứ từ quan điểm của A. Mọi đỉnh khác hoặc là bạn của A hoặc là kẻ thù của A. Do đó, một cách tự nhiên, các ứng viên vào trong các tập X và Y: xác định trong tập X là A và tất cả bạn của nó, và xác định trong Y là tất cả kẻ thù của A. Vì vậy, mỗi đỉnh hoặc là bạn của A hoặc là kẻ thù của A.

Nhớ lại những gì chúng ta cần hiển thị trong hai tập X và Y để thỏa mãn các điều kiện yêu cầu:

- (i) Mỗi hai đỉnh trong X là bạn
- (ii) Mỗi hai đỉnh trong Y là bạn
- (iii) Mỗi đỉnh trong X là kẻ thù của mỗi đỉnh trong Y

Lập luận rằng mỗi điều kiện trên là thực tế đúng cho sự lựa chọn của X và Y. Điều này có nghĩa là X và Y thỏa mã các điều kiện yêu cầu và sẽ hoàn thành việc chứng minh. Phần còn lại của đối số, chứng minh (i), (ii) và (iii), được minh họa trong sơ đồ hình 5.4.

Đối với (i), chúng ta biết rằng A là bạn với tất cả các đỉnh khác trong X. Làm thế nào để hai đỉnh khác trong X (gọi chúng là B và C) – phải là bạn? Chúng ta biết rằng A là bạn với cả B và C, vì vậy nếu B và C là kẻ thù của nhau thì A, B và C tạo thành một tam giác với hai nhãn + => vi phạm điều kiện cân bằng. Từ khi chúng ta biết về mạng cân bằng, điều này là không thể xảy ra, vì vậy trên thực tế B và C phải là

bạn. B và C là tên của bất kỳ hai đỉnh trong X, chúng ta kết luận rằng, mỗi hai đỉnh trong X đều là bạn.

Chúng ta thử đối số cùng loại trong (ii). Xem xét bất kỳ hai đỉnh trong Y (gọi chúng là D và E) – chúng phải là bạn. Chúng ta biết rằng A là kẻ thù của cả hai D và E, vì vậy nếu D và E là kẻ thù của nhau, thì A, D và E tạo thành một tam giác không có nhãn + \Rightarrow điều này vi phạm điều kiện cân bằng. Do vậy, trên thực tế, D và E phải là bạn. Vì D và E là tên của hai đỉnh bất kỳ trong Y, chúng ta kết luận rằng mỗi hai đỉnh trong Y đều là bạn.

Cuối cùng, chúng ta chứng minh điều kiện (iii). Theo kiểu đối số trong (i) và (ii), coi một đỉnh trong X (gọi là B) và một đỉnh trong Y (gọi là D) – phải là kẻ thù của nhau. Chúng ta biết rằng A là bạn của B và là kẻ thù của D, vì vậy nếu B và D là bạn thì A, B và D tạo thành một tam giác có hai nhãn + \Rightarrow vi phạm điều kiện cân bằng. Vì vậy, trên thực tế B và D phải là kẻ thù của nhau. Vì B và D là tên của bất kỳ một đỉnh trong X và bất kỳ một đỉnh Y nên chúng ta kết luận rằng mỗi cặp như vậy là một cặp kẻ thù.

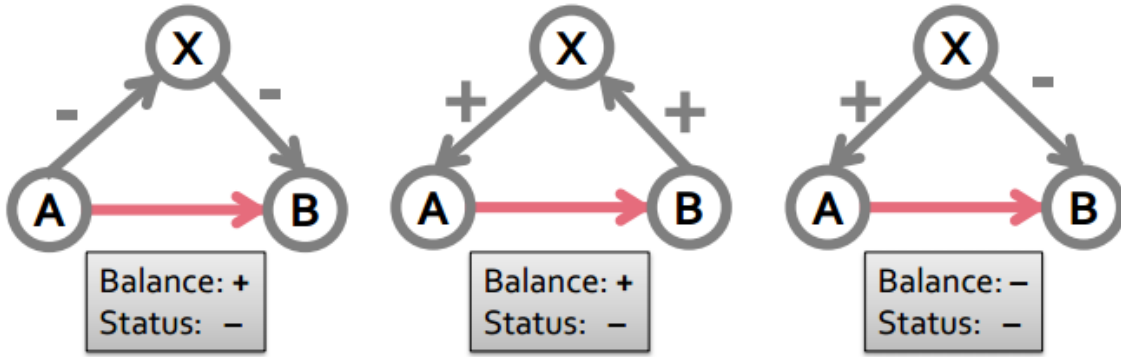
Vì vậy, trong kết luận, giả sử rằng chỉ mạng như thế là cân bằng, chúng ta mô tả một bộ phận của các đỉnh trong hai tập X và Y, và kiểm tra các điều kiện (i), (ii) và (iii). Điều này làm cho định lý cân bằng được chứng minh.

3.2 Lý thuyết trạng thái

Lý thuyết trạng thái là một cách tiếp cận khác để dự đoán liên kết trong mạng xã hội. Lý thuyết trạng thái được phát biểu như sau:

“Một liên kết dương từ A đến B có thể có nghĩa là B là bạn của A, nhưng nó cũng có thể là B có trạng thái cao hơn A. Tương tự như vậy một liên kết âm từ A đến B có nghĩa là B là kẻ thù của A hoặc có nghĩa là B có trạng thái thấp hơn A.”[6].

Với lý thuyết cân bằng cấu trúc và lý thuyết trạng thái, việc dự đoán cung của liên kết sẽ cho ra các kết quả khác nhau (Hình 9).



Hình 9. Mô hình cân bằng cấu trúc và mô hình trạng thái

3.3. Tính cá nhân trong mạng xã hội

Để dự đoán liên kết âm liên kết dương trong mạng xã hội chúng ta cũng có thể dùng các độ đo về tính cá nhân. Tính cá nhân của mỗi người thể hiện tính cách, cách suy nghĩ, sự tương tác với người khác trong mạng xã hội. Dựa vào các tính cách này chúng ta có thể nhìn thấy khả năng kết nối của mỗi người, sự tương tác giữa người và người hay người và nhóm người trong mạng.

Tính cá nhân trong mạng xã hội được biểu diễn thông qua mô hình “Big Five” – năm tính cách cá nhân tiêu biểu sau[26]:

- Tính mở (sự cởi mở, thẳng thắn, chân thật): Tò mò, thông minh, giàu trí tưởng tượng. Những người có chỉ số "tính mở" cao thường là những người có xu hướng nghệ thuật đạt được nhiều kỹ năng trong việc đánh giá, đưa ra các ý tưởng.
- Tính lương tâm (ngay thẳng): Có trách nhiệm, có tổ chức và kiên trì trong công việc. Những người có chỉ số điểm lương tâm cao thường đáng tin cậy, và có thành tích cao, chăm chỉ và hay đưa ra các kế hoạch.
- Tính hướng ngoại: Hướng ngoại, năng động, quyết đoán. Thân thiện và năng động, những người mang tính cách hướng ngoại thường lấy cảm hứng làm việc từ các hoạt động xã hội.
- Tính hòa đồng (tính dễ chịu, dễ thích ứng với môi trường): Hòa đồng với mọi người, dễ thích ứng và cộng tác với người xung quanh. Những người mang tính cách này thích hợp với công việc hòa giải, đem lại sự lạc quan tin tưởng cho người khác.
- Tính nhạy cảm (nhạy cảm về thân kinh, dễ kích động hay nổi nóng ...): Nhạy cảm, dễ lo lắng, buồn rầu, căng thẳng, dễ mắc vào tiêu cực.

Ứng dụng của "Big Five":

- Trong giáo dục (Phát huy các ưu điểm của sinh viên, ví dụ: Chia nhóm có sự hỗ trợ giữa nhiều tính cách)
- Quảng cáo: Trong bối cảnh của quảng cáo tiếp thị, Big Five đã được dự đoán chính xác một sở thích người tiêu dùng cho các thương hiệu quốc gia hoặc thương hiệu độc lập. Các nghiên cứu như thế này cho thấy một tương lai đầy hứa hẹn cho sự tích hợp của phân tích cá nhân và hồ sơ của người tiêu dùng.
- Trong tương tác người máy: Mỗi giao diện được thiết kế phù hợp cho mỗi tính cách để đạt hiệu quả cao trong công việc (môi trường chuyên nghiệp)

3.4 Mô hình dự đoán liên kết âm, liên kết dương trong mạng xã hội

Phần này chúng ta sẽ xây dựng một mô hình dự đoán liên kết dựa trên các tiếp cận bằng học máy. Để xây dựng được mô hình chúng ta cần xác định các đặc trưng của mô hình và phương pháp học của mô hình. Tiếp theo chúng ta sẽ đi vào chi tiết của từng phần của mô hình.

3.4.1 Đặc trưng của mô hình.

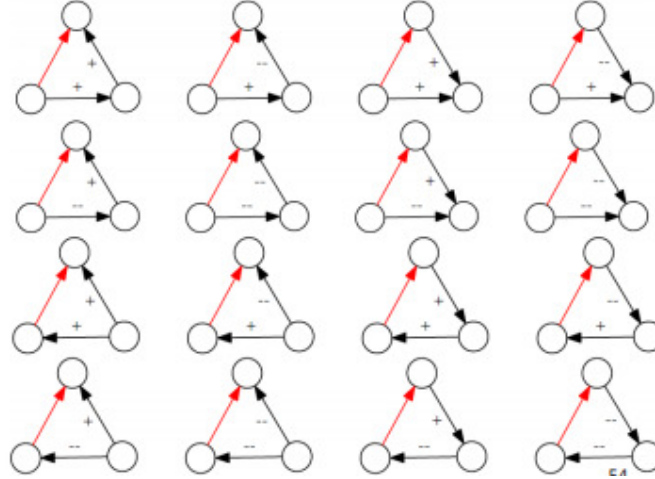
Dựa vào các tiếp cận học máy, chúng ta cần định nghĩa các đặc trưng cho mô hình. Các đặc trưng được chia ra làm hai lớp. Lớp thứ nhất dựa vào nhân của các cung, đây là đặc trưng cơ bản mô tả quan hệ của một đỉnh trong đồ thị với các đỉnh khác. Lớp thứ 2 dựa trên nguyên lý xã hội, chúng ta có thể hiểu được một mối quan hệ giữa u và v dựa vào một đối tượng trung gian w . Ví dụ, nếu một ai đó có liên kết dương đến cả u và v thì có khả năng liên kết giữa u và v hình thành là rất cao.

Đặc trưng của lớp thứ nhất được dựa trên số liên kết xuất phát từ đỉnh v và số liên kết đến đỉnh v . Chúng ta kí hiệu $d_{in}^+(v)$ và $d_{in}^-(v)$, cho liên kết dương và âm đi vào v . Tương tự ta có $d_{out}^+(v)$ và $d_{out}^-(v)$ cho các liên kết dương và âm từ đỉnh v đi ra.

Lớp đặc trưng thứ hai được mô tả bởi các tam giác quan hệ dựa vào lý thuyết cân bằng cấu trúc hoặc lý thuyết trạng thái. Với đồ thị vô hướng thì chúng ta có 4 loại tam giác (Hình 10). Với đồ thị có hướng số tam giác sẽ là 16 (Hình 11).



Hình 10. Các tam giác quan hệ trong đồ thị vô hướng



Hình 11. Các tam giác quan hệ trong đồ thị có hướng[6]

Đặc trưng tiếp theo được sử dụng vào trong mô hình dự đoán này là đặc trưng về độ gắn kết của 2 đỉnh u và v được ký hiệu là $C(u, v)$. Đặc trưng này biểu diễn cho số bạn chung của hai đỉnh u và v .

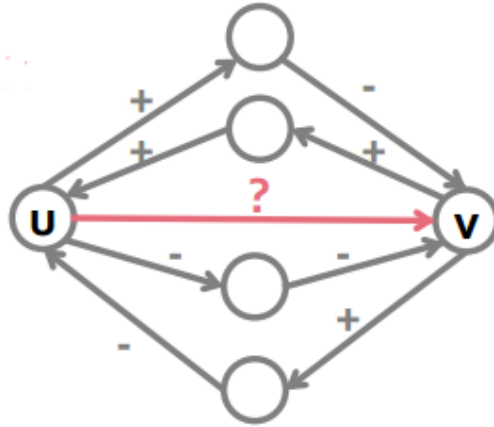
Đặc trưng tiếp theo được sử dụng trong mô hình là đặc trưng về tính cá nhân của hai người với nhau. Nếu hai người có liên kết với nhau và các đặc điểm cá nhân của họ giống nhau thì khả năng. Các đặc trưng về tính cá nhân dựa vào mô hình “Big Five” gồm năm tính cách cá nhân khác nhau đã được trình bày ở phần 3.3[26].

3.4.2 Phương pháp

Trong phần này chúng ta sẽ sử dụng phương pháp học máy hồi quy kết đối với các đặc trưng mà chúng ta đã mô tả ở phần trên cho bộ phân lớp dự đoán liên kết âm liên kết dương. Công thức hồi quy được thể hiện như sau[6]:

$$P(+|x) = \frac{1}{1+e^{-(b_0+\sum_1^n b_i x_i)}} \quad (27)$$

Trong đó x là vector đặc trưng (x_1, x_2, \dots, x_n) , b_0, \dots, b_n là các tham số được xác định từ dữ liệu huấn luyện.



Hình 12. Dự đoán liên kết dựa vào các đặc trưng

3.5 Kết luận chương 3

Trong chương này khóa luận đã trình bày mô hình cho việc dự đoán liên kết âm liên kết dương dựa vào lý thuyết cân bằng cấu trúc kết hợp với phương pháp học máy hồi quy.

Trong chương tiếp theo, chúng ta sẽ tiến hành thực nghiệm và đánh giá kết quả mô hình đã trình bày ở đây để thấy được tính chính xác của mô hình.

CHƯƠNG 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ

Trong chương này khóa luận sẽ trình bày một số thực nghiệm để chứng minh tính đúng đắn và tính thực tiễn của mô hình. Mô hình được thực nghiệm trên ba mạng xã hội là Epinions, Slashdot Zoo và Wikipedia. Dữ liệu của các mạng đã được chuyển thành đồ thị và kết quả đạt tương đối khả quan.

4.1 Dữ liệu thực nghiệm

Dữ liệu trong phần thực nghiệm này là ba mạng xã hội lớn với các liên kết trong mạng đã được gán nhãn âm hoặc dương. Đó là ba mạng Epinions, Slashdot Zoo và Wikipedia.

Mạng Epinions: đây là một Website đánh giá sản phẩm với cộng đồng người sử dụng rất tích cực. Người dùng với nhau thông qua sự tin cậy hay không tin cậy vào nhau trong mạng, nó là sự kết hợp giữa việc đánh giá sản phẩm của người dùng đưa lên cũng như đánh giá chính người dùng đó. Dữ liệu của website được thu thập từ năm 1999 đến tháng Tám năm 2003. Mạng Epinions chứa 119,217 đỉnh và 841,000 cung, với 85% là các liên kết dương. Trong đó có 80,668 người dùng có ít nhất một liên kết âm hoặc dương.

Mạng Slashdot Zoo: đây là mạng tin tức công nghệ. Năm 2002 Slashdot Zoo được giới thiệu và cho phép người dùng có thể gán nhãn các người dùng khác là *bạn* hay *kẻ thù* của mình. Cách thức gán nhãn cho các cung tương tự với mạng Epinions, mỗi quan hệ bạn bè có nghĩa là người dùng này thích hay có những bình luận tốt về người dùng khác, trong khi đó các quan hệ kẻ thù có nghĩa là người dùng không thích hoặc có bình luận không tốt về người dùng khác. Dữ liệu của mạng Slashdot được thu thập từ tháng Hai năm 2009 với 82,144 người dùng và 549,202 các cung với 77.4% là các liên kết dương.

Mạng Wikipedia: là một mạng từ điển bách khoa toàn thư với cộng đồng người dùng hoạt động tích cực. Mạng Wikipedia cho phép người dùng bỏ phiếu tín nhiệm với một người dùng khác cho các vị trí quản trị viên. Nếu người dùng đồng ý cho việc người dùng khác làm quản trị viên thì liên kết sẽ là liên kết dương và ngược lại với liên kết âm. Theo thống kê mạng Wikipedia cho đến 7,115 người dùng tham gia bỏ phiếu và có 103,689 phiếu tín nhiệm được đưa ra và có 78,7 % các liên kết là liên kết dương.

Dữ liệu được biểu diễn dưới dạng đồ thị trong một file text với định dạng mỗi dòng là một cung gồm đỉnh xuất phát, đỉnh tới và dấu của của cung đó. Các dòng có chữ ký tự # là các dòng ghi chú. Hình 12 dưới đây mô tả cấu trúc của file text:


```

1 # Directed graph: soc-sign-epinions
2 # Epinions signed social network
3 # Nodes: 131828 Edges: 841372
4 # FromNodeId ToNodeId Sign
5 0 1 -1
6 1 128552 -1
7 2 3 1
8 4 5 -1
9 4 155 -1
10 4 558 1
11 4 1509 -1
12 4 2282 1
13 4 2984 1
14 4 7263 1
15 4 10876 -1
16 4 48703 -1
17 4 51253 1
18 4 87217 1
19 4 98617 1
20 4 100981 1
21 4 101858 1
22 5 5 1
23 5 8 -1
24 5 20 1
25 5 50 1
26 5 52 1
27 5 155 1
28 5 183 1
29 5 197 1
30 5 245 1
31 5 264 -1
32 5 302 1

```

Hình 13. Minh họa dữ liệu đồ thị đã được gán nhãn các cung

4.2 Môi trường thực nghiệm

Thực nghiệm của chúng tôi được tiến hành trên môi trường phần cứng như sau:

Bảng 1. Môi trường thực nghiệm

Thành phần	Chỉ số
Bộ vi xử lý (CPU)	Intel Core Due 2.2 GHz
Bộ nhớ chính (RAM)	4096 MB
Ổ cứng (HDD)	320 GB
Hệ điều hành (OS)	Windows 7-64 bits

4.3 Các công cụ phần mềm

Trong phần thực nghiệm này khóa luận đã sử dụng các phần mềm và thư viện sau đây:

Bảng 2. Công cụ phần mềm

STT	Tên phần mềm	Mô tả
1	Bộ công cụ phân tích mạng xã hội SNAP	Tác giả: Jure Leskovec http://snap.stanford.edu/snap/download.html
2	Visual C++ 2010 Express	Tác giả: Microsoft http://www.microsoft.com/visualstudio/en-us/products/2010-editions/visual-cpp-express

4.4 Kết quả và đánh giá

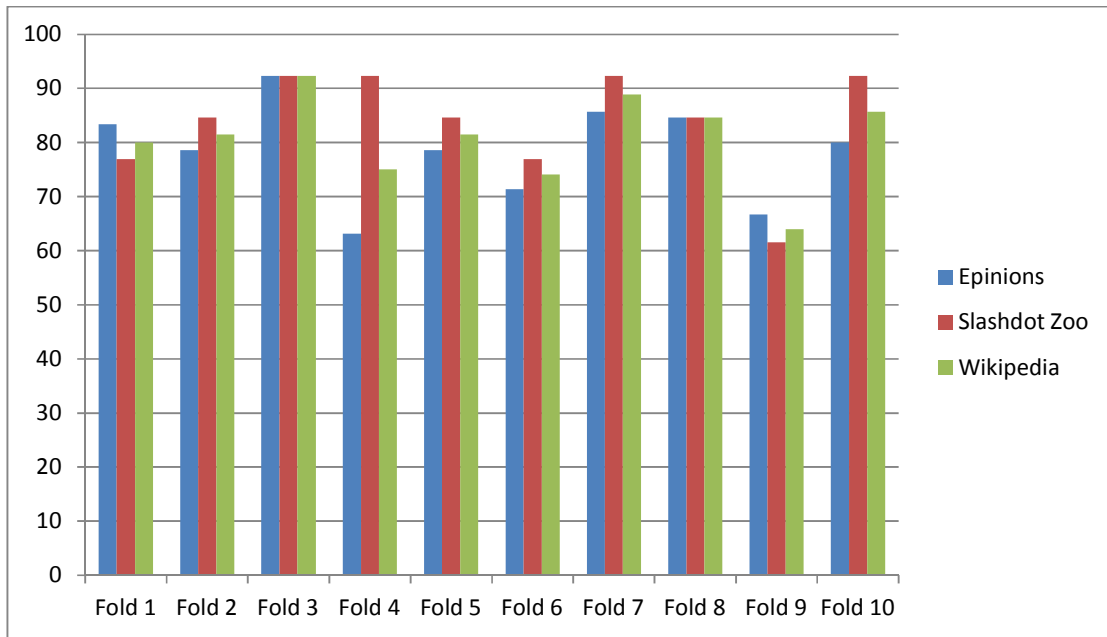
Mô hình được đánh giá thông qua độ đo AUC, độ đo AUC có thể được biểu diễn thông qua xác suất ngẫu nhiên chọn một liên kết bị mất và cho nó một số điểm cao hơn xác suất chọn một liên kết không tồn tại. Tại mỗi lần chúng ta chọn ngẫu nhiên một liên kết bị mất và một liên kết không tồn tại để so sánh điểm số, nếu có n' lần liên kết bị mất có điểm số cao hơn và n'' lần liên kết bị mất và liên kết không tồn tại có cùng điểm số, giá trị của AUC hay độ chính xác được tính như sau:

$$AUC = \frac{n' + 0.5n''}{n} \quad (28)$$

Trong thực nghiệm này chúng tôi tiến hành chia dữ liệu thành và kiểm thử chéo theo phương pháp 10-folds, trong đó 9 phần là dữ liệu training và 1 phần làm dữ liệu test. Kết quả đạt độ chính xác xấp xỉ 80%, đây là một kết quả khả quan (Bảng 3)

Bảng 3: Độ chính xác bộ phân lớp dự đoán

	Epinions	Slashdot Zoo	Wikipedia
Fold 1	83.33%	76.92%	80.00%
Fold 2	78.57%	84.61%	81.48%
Fold 3	92.30%	92.30%	92.30%
Fold 4	63.15%	92.30%	75.00%
Fold 5	78.57%	84.61%	81.48%
Fold 6	71.42%	76.92%	74.06%
Fold 7	85.71%	92.30%	88.88%
Fold 8	84.61%	84.61%	84.61%
Fold 9	66.67%	61.53%	64.00%
Fold 10	80%	92.30%	85.71%
Trung bình	78.43%	83.84%	80.75%



Hình 14. Biểu đồ kết quả thực nghiệm

4.5 Kết luận chương 4

Chương 4 đã trình bày độ đo AUC để đánh giá bộ phân lớp dự đoán liên kết âm liên kết dương cũng như việc thực nghiệm và đánh giá kết quả của bộ phân lớp dự đoán liên kết âm liên kết dương. Chương này cũng đã mô tả về môi trường thực nghiệm, dữ liệu và đưa ra kết quả đánh giá cho thực nghiệm. Kết quả đạt xấp xỉ 80%, đây là một kết quả khá cao đối với việc dự đoán liên kết âm liên kết dương.

KẾT LUẬN VÀ PHƯƠNG HƯỚNG

Mạng xã hội và các bài toán dự đoán liên kết trong mạng xã hội là những vấn đề được nhiều nhà nghiên cứu quan tâm. Các bài toán dự đoán liên kết trong mạng xã hội có thể được áp dụng trong nhiều lĩnh vực của đời sống như kinh tế, chính trị, giáo dục, y tế và khoa học công nghệ...

Khóa luận đã mô tả tổng quan về bài toán dự đoán liên kết âm liên kết dương trong mạng xã hội và các khái niệm liên quan, cũng như các phương pháp cơ bản để dự đoán liên kết trong các mạng nói chung và mạng xã hội nói riêng. Khóa luận chú trọng trình bày về lý thuyết cân bằng cấu trúc và lý thuyết trạng thái qua đó xây dựng các đặc trưng cho bài toán dự đoán liên kết âm và liên kết dương trong mạng xã hội. Tiếp đó khóa luận cũng trình bày phương pháp dự đoán liên kết âm liên kết dương trong mạng xã hội và trình bày quá trình thực nghiệm cũng như đánh giá để thấy được tính chính xác và đúng đắn của mô hình dự đoán liên kết âm liên kết dương.

Hướng nghiên cứu tiếp theo của khóa luận là tập trung mở rộng tập đặc trưng của mô hình dự đoán bằng cách kết hợp tính cá nhân trong mạng xã hội để làm tăng độ chính xác cho mô hình dự đoán liên kết âm liên kết dương trong mạng xã hội.

TÀI LIỆU THAM KHẢO

- [1] C. T. Butts, “Network inference, error, and information (in)accuracy: A Bayesian approach”, *Social Networks* 25 (2003) 103.
- [2] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Field, P. Bork, “Comparative assessment of large-scale data sets of protein-protein interactions”, *Nature* 417 (2002) 399.
- [3] D. Liben-Nowell and J. Kleinberg, “The link-prediction problem for social networks” *J. Amer. Soc. Inf. Sci. and Tech.*, 58(7):1019–1031, 2007
- [4] E. Gilbert and K. Karahalios, “Predicting Tie Strength with Social Media”, *In Proc. of CHI*, 2009.
- [5] H. Yu et al., “High-quality binary protein interaction map of the yeast interactome network”, *Science* 322 (2008) 104.
- [6] J. Leskovec, D. Huttenlocher, J. Kleinberg, “Predicting Positive and Negative Links in Online Social Networks”, *In Proceedings of WWW’2010*, ACM Press, New York, 2010.
- [7] Jon Kleinberg, David Easley “Networks, Crowds, and Markets: Reasoning about a Highly Connected World” (2010) 47-83, 119-152.
- [8] L. da F. Costa, F. A. Rodrigues, G. Travieso, P. R. U. Boas, “Characterization of complex networks: A survey of measurements”, *Adv. Phys.* 56 (2007) 167.
- [9] L. Schafer, J. W. Graham, Missing data: “Our view of the state of the art”, *Psychol. Methods* 7 (2002) 147.
- [10] Linyuan Lu, Tao Zhou, “Link Prediction in Complex Networks: A Survey”, *Physica A* 390 (2011) 1150-1170
- [11] Mark Granovetter. The strength of weak ties. *American Journal of Sociology*, 78:1360{1380, 1973}.
- [12] M. E. J. Newman. “The structure and function of complex networks”. *SIAM Review*, 45:167–256, 2003
- [13] M. E. J. Newman, “Clustering and preferential attachment in growing networks”, *Phys. Rev. E* 64 (2001) 025102.
- [14] M. E. J. Newman, “The Structure and Function of Complex Net works”, *SIAM.Rev.* 45 (2003) 167.
- [15] M. P. H. Stumpf, T. Thorne, E. de Silva, R. Stewart, H. J. An, M. Lappe, C. Wiuf, “Estimating the size of the human interactome”, *Proc. Natl. Acad. Sci. U.S.A.* 105 (2008) 6959.
- [16] Peter Bearman and James Moody, “Suicide and friendships among American adolescents” *American Journal of Public Health*, 94(1):89{95, 2004}.
- [17] J. W. Neal, “Kracking - the Missing Data Problem: Applying Krackhardt’s Cognitive Social Structures to School-Based Social Networks”, *Sociol. Educ.* 81 (2008) 140.
- [18] R. Albert, A.-L. Barabási, “Statistical mechanics of complex networks”, *Rev. Mod. Phys.* 74 (2002) 47.
- [19] R. Guimera, M. Sales-Pardo, “Missing and spurious interactions and the reconstruction of complex networks”, *Proc. Natl. Acad. Sci. U.S.A.* 106 (2009) 22073

- [20] S. N. Dorogovtsev, J. F. F. Mendes, “Evolution of networks”, *Adv. Phys.* 51 (2002) 1079.
- [21] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Huang, “Complex networks: Structure and dynamics”, *Phys. Rep.* 424 (2006) 175.
- [22] S. Zhou, R. J. Mondragón, “Accurately modeling the internet topology”, *Phys. Rev. E* 70 (2004) 066108.
- [23] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, E. Shir, “A model of Internet topology using k-shell decomposition”, *Proc. Natl. Acad. Sci. U.S.A.* 104 (2007) 11150.
- [24] Uzzi, B. 1999. “Embeddedness in the Making of Financial Capital: How Social Relations and Networks Benefit Firms Seeking Financing” . *American Sociological Review*, 64(4), 481–505.
- [25] Z. Xu, V. Tresp, K. Yu, S. Yu, H.-P. Kriegel, “Dirichlet enhanced relational learning”, *In Proceedings of the 22nd international conference on machine learning, Bonn, Germany, 2005, p. 1004.*
- [26] Golbeck, J., Robles, C., Turner, K, “Predicting Personality with Social Media”, *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*, 2011.