**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN (CLC)**

---------------🙦 🕮 🙤---------------

****

**BÁO CÁO BÀI TẬP LỚN**

**MÔN: LÝ THUYẾT ĐỒ THỊ**

**ĐỀ TÀI:** **PHÂN TÍCH MẠNG XÃ HỘI TWITTER**

**SỰ KIỆN KHÁM PHÁ HẠT HIGGS BOSON**

### 

**Ngày 20 tháng 11 năm 2025**

**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN (CLC)**

---------------🙦 🕮 🙤---------------

****

**BÁO CÁO BÀI TẬP LỚN**

**MÔN: LÝ THUYẾT ĐỒ THỊ**

**ĐỀ TÀI:** **PHÂN TÍCH MẠNG XÃ HỘI TWITTER**

**SỰ KIỆN KHÁM PHÁ HẠT HIGGS BOSON**

**Giáo viên hướng dẫn: Đỗ Như Tài**

**Nhóm sinh viên thực hiện:**

**Võ Thị Thảo Vy - 3124411356**

**Nguyễn Như Thảo - 3124411281**

**Lê Tân Phước Thọ - 3124411294**

**Đặng Đào Đạt Thành** -**3124411251**

**Lớp: DCT124C6**

**Ngày 20 tháng 11 năm 2025**

**LỜI MỞ ĐẦU**

Trong kỷ nguyên số hóa và bùng nổ thông tin, mạng xã hội đã trở thành một phần không thể thiếu trong đời sống con người, đồng thời là đối tượng nghiên cứu quan trọng của khoa học dữ liệu và xã hội học tính toán. Sự kiện khám phá hạt Higgs boson năm 2012 không chỉ là một cột mốc lịch sử của ngành vật lý hạt, mà còn là một hiện tượng xã hội đặc biệt, nơi mà khoa học đỉnh cao và công nghệ truyền thông hiện đại giao thoa một cách mạnh mẽ.

Cuốn báo cáo này là kết quả của quá trình nghiên cứu và phân tích toàn diện mạng lưới lan truyền thông tin xung quanh sự kiện Higgs boson trên nền tảng Twitter. Thông qua việc ứng dụng các phương pháp phân tích mạng xã hội tiên tiến kết hợp với lý thuyết đồ thị và thống kê, nghiên cứu không chỉ khám phá cấu trúc phức tạp của mạng xã hội quy mô lớn mà còn làm sáng tỏ cơ chế lan truyền thông tin khoa học trong xã hội đương đại.

Báo cáo được trình bày thành năm chương chính, bắt đầu từ tổng quan đề tài, phân tích đặc trưng mạng, cấu trúc mạng, cộng đồng và kết thúc bằng kết luận cùng các đề xuất phát triển. Mỗi chương không chỉ cung cấp các phân tích định lượng chặt chẽ mà còn đưa ra những diễn giải có ý nghĩa về mặt xã hội và truyền thông.

Chúng tôi hy vọng rằng những phát hiện và phương pháp được trình bày trong báo cáo này sẽ không chỉ có giá trị học thuật mà còn mang lại những ứng dụng thực tiễn trong việc hoạch định chiến lược truyền thông khoa học và quản lý thông tin trong kỷ nguyên số.

**LỜI CẢM ƠN**

Trước tiên, chúng tôi xin bày tỏ lòng biết ơn sâu sắc đến GV Đỗ Như Tài hướng dẫn đề tài, người đã tận tình chỉ bảo, định hướng và cung cấp những kiến thức quý báu trong suốt quá trình thực hiện nghiên cứu. Những đóng góp và góp ý chuyên môn của Thầy đã giúp chúng tôi hoàn thiện báo cáo này.

Chúng tôi chân thành cảm ơn Khoa công nghệ thông tin- Trường Đại học Đại học sài gòn đã tạo điều kiện thuận lợi về cơ sở vật chất và môi trường học thuật để chúng tôi có thể thực hiện đề tài một cách tốt nhất.

Xin gửi lời cảm ơn đến Trung tâm Phân tích Mạng Stanford (SNAP) đã cung cấp bộ dữ liệu Higgs Twitter công khai và chất lượng cao, làm nền tảng cho các phân tích trong nghiên cứu này.

Chúng tôi cũng xin cảm ơn các tác giả của các thư viện mã nguồn mở như NetworkX, Pandas, Matplotlib, và cộng đồng khoa học dữ liệu đã phát triển những công cụ phân tích mạnh mẽ, giúp chúng tôi có thể thực hiện các phân tích phức tạp trên tập dữ liệu lớn.

Cuối cùng, chúng tôi xin cảm ơn tất cả những ai đã trực tiếp hoặc gián tiếp đóng góp cho sự hoàn thiện của báo cáo này. Mọi sai sót trong báo cáo hoàn toàn thuộc về trách nhiệm của nhóm tác giả.

*ngày 25 tháng 11 năm 2025*

[CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI 6](#_Toc215047733)

[1.1. Bối Cảnh và Ý Nghĩa Nghiên Cứu 6](#_Toc215047734)

[1.2. Mục Tiêu Nghiên Cứu 7](#_Toc215047735)

[1.3. Phạm Vi và Giới Hạn 8](#_Toc215047736)

[CHƯƠNG 2: PHÂN TÍCH TỔNG QUAN MẠNG 11](#_Toc215047737)

[2.1. Nguồn Dữ Liệu và Quy Trình Tiền Xử Lý 11](#_Toc215047738)

[2.2. Đặc Trưng Đồ Thị và Phân Loại Mạng 12](#_Toc215047739)

[2.3. Phân Tích Độ Đo Cơ Bản và Ý Nghĩa 15](#_Toc215047740)

[2.4. Phân Phối Bậc và Mô Hình Hóa 16](#_Toc215047741)

[2.5. Trực Quan Hóa và Diễn Giải Cấu Trúc 18](#_Toc215047742)

[CHƯƠNG 3: PHÂN TÍCH CẤU TRÚC MẠNG 19](#_Toc215047743)

[3.1. So Sánh với Mô Hình Ngẫu Nhiên 19](#_Toc215047744)

[3.2. Phân Tích Trung Tâm Mạng 20](#_Toc215047745)

[3.4. Phân Tích Tương Đương Cấu Trúc 24](#_Toc215047746)

[CHƯƠNG 4: PHÂN TÍCH CỘNG ĐỒNG TRONG MẠNG 25](#_Toc215047747)

[4.1. Phân Tích K-Core Decomposition 25](#_Toc215047748)

[4.2. So Sánh Thuật Toán Phát Hiện Cộng Đồng 26](#_Toc215047749)

[4.3. Đánh Giá Chất Lượng Cộng Đồng 27](#_Toc215047750)

[4.4. Trực Quan Hóa và Phân Tích Liên Cộng Đồng 27](#_Toc215047751)

[CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 29](#_Toc215047752)

[5.1. Kết Luận Chính 29](#_Toc215047753)

[5.2. Hạn Chế Nghiên Cứu 29](#_Toc215047754)

[5.3. Hướng Phát Triển 30](#_Toc215047755)

[5.4. Đóng Góp và Ý Nghĩa 30](#_Toc215047756)

[TÀI LIỆU THAM KHẢO 31](#_Toc215047757)

[Tài liệu tiếng Việt: 31](#_Toc215047758)

[Tài liệu tiếng Anh: 31](#_Toc215047759)

[Bài báo khoa học: 32](#_Toc215047760)

[Dataset và công cụ: 32](#_Toc215047761)

[Trang web và báo cáo: 32](#_Toc215047762)

[Phần mềm và thư viện: 33](#_Toc215047763)

# CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

## 1.1. Bối Cảnh và Ý Nghĩa Nghiên Cứu

Sự kiện khoa học đột phá và tầm ảnh hưởng toàn cầu:  
Ngày 4 tháng 7 năm 2012 đánh dấu một cột mốc lịch sử trong ngành vật lý hạt khi Tổ chức Nghiên cứu Hạt nhân Châu Âu (CERN) chính thức công bố khám phá hạt Higgs boson sau hơn 4 thập kỷ tìm kiếm. Được mệnh danh là "Hạt của Chúa" (The God Particle), Higgs boson đại diện cho thành phần cuối cùng chưa được khám phá trong Mô hình Chuẩn của vật lý hạt, giải thích cơ chế tạo ra khối lượng cho các hạt cơ bản thông qua trường Higgs. Sự kiện này không chỉ thu hút sự chú ý của cộng đồng khoa học toàn cầu mà còn tạo nên làn sóng thảo luận rộng rãi trên các nền tảng mạng xã hội, thể hiện sự quan tâm đặc biệt của công chúng đối với những tiến bộ khoa học cơ bản.

Sự chuyển đổi trong truyền thông khoa học:  
Thế kỷ 21 chứng kiến sự chuyển đổi mạnh mẽ trong cách thức truyền bá thông tin khoa học. Twitter đã nổi lên như một kênh truyền thông quan trọng cho việc phổ biến thông tin khoa học tới công chúng, phá vỡ các rào cản truyền thống giữa nhà khoa học và công chúng. Trong sự kiện Higgs boson, nền tảng này đóng vai trò đa chiều:

* Kênh thời gian thực: Cập nhật tức thì các thông báo chính thức từ CERN, cho phép người dùng theo dõi trực tiếp các diễn biến từ hội nghị khoa học
* Diễn đàn thảo luận đa chiều: Kết nối đa dạng các bên liên quan bao gồm nhà khoa học, nhà báo khoa học, nhà hoạt động giáo dục, và công chúng quan tâm
* Cơ chế lan truyền thông tin: Retweet hoạt động như hình thức xác nhận và khuếch đại thông tin, tạo ra hiệu ứng lan tỏa theo cấp số nhân
* Cầu nối văn hóa: Thu hẹp khoảng cách giữa khoa học chuyên sâu và công chúng phổ thông

Ý nghĩa học thuật và thực tiễn:  
Việc phân tích mạng lưới retweet xung quanh sự kiện Higgs boson mang lại những hiểu biết sâu sắc về nhiều khía cạnh:

* Cơ chế lan truyền thông tin khoa học trong kỷ nguyên số: Nghiên cứu này cung cấp cái nhìn cụ thể về cách thông tin khoa học phức tạp được truyền tải và tiếp nhận thông qua các nền tảng kỹ thuật số
* Cấu trúc quyền lực và ảnh hưởng trong cộng đồng quan tâm đến khoa học: Phân tích làm rõ ai là những người có ảnh hưởng trong việc định hướng thảo luận về chủ đề khoa học
* Động lực hình thành cộng đồng xung quanh các chủ đề khoa học: Khám phá cách các cộng đồng trực tuyến tự tổ chức xung quanh sự kiện khoa học
* Ứng dụng trong hoạch định chính sách: Cung cấp insights cho việc phát triển chiến lược truyền thông khoa học hiệu quả

## 1.2. Mục Tiêu Nghiên Cứu

Nghiên cứu này được thiết kế nhằm đạt được hệ thống mục tiêu phân cấp từ tổng quan đến chuyên sâu:

Mục tiêu tổng quan về đặc trưng mạng:

* Phân tích toàn diện đặc trưng topo của mạng retweet Higgs Twitter, bao gồm các tham số cơ bản và nâng cao
* Xác định và định lượng các thuộc tính cấu trúc cơ bản của mạng xã hội quy mô lớn trong bối cảnh lan truyền thông tin khoa học
* Đánh giá mức độ phù hợp của các lý thuyết mạng phức hợp (complex networks) khi áp dụng cho mạng xã hội thực tế

Mục tiêu chuyên sâu về trung tâm và ảnh hưởng:

* Định vị và phân loại các node có ảnh hưởng lớn trong mạng thông qua đa tiêu chí đánh giá
* Phân tích hệ thống các độ đo trung tâm khác nhau (degree, betweenness, closeness, eigenvector, PageRank) và mối tương quan giữa chúng
* Xác định đặc điểm của các "influencers" trong lan truyền thông tin khoa học và so sánh với influencers trong các lĩnh vực khác
* Phân tích vai trò của các node cầu nối (structural holes) trong việc kết nối các cộng đồng khác nhau

Mục tiêu về cấu trúc cộng đồng và tương tác:

* Phát hiện và phân tích chi tiết các cộng đồng tự nhiên trong mạng thông qua nhiều thuật toán khác nhau
* So sánh hiệu quả và ý nghĩa của các thuật toán phát hiện cộng đồng (Louvain, Label Propagation, Girvan-Newman) trong bối cảnh cụ thể
* Diễn giải ý nghĩa xã hội của các cộng đồng được phát hiện và mối quan hệ giữa chúng
* Phân tích cấu trúc phân tầng của mạng thông qua k-core decomposition

Mục tiêu so sánh và đối chiếu lý thuyết:

* Đối chiếu mạng thực với các mô hình ngẫu nhiên (Erdős-Rényi, Configuration Model) để xác định các đặc trưng độc đáo
* Kiểm định giả thuyết về tính chất scale-free và small-world của mạng xã hội trong bối cảnh lan truyền thông tin khoa học
* Đánh giá mức độ phù hợp của các mô hình lý thuyết với thực tế quan sát được

## 1.3. Phạm Vi và Giới Hạn

Phạm vi dữ liệu và thời gian:

* Dataset: Sử dụng Higgs Twitter Retweet Network từ Stanford SNAP - một trong những bộ dữ liệu mạng xã hội công khai lớn nhất và được nghiên cứu rộng rãi
* Quy mô: 223,833 node và 308,596 edge, đảm bảo tính đại diện và ý nghĩa thống kê
* Thời gian nghiên cứu: Tập trung vào giai đoạn cao điểm xung quanh công bố khám phá Higgs boson, cho phép quan sát hiện tượng lan truyền thông tin ở trạng thái tập trung
* Phạm vi không gian: Toàn cầu, với người dùng Twitter từ nhiều quốc gia và nền văn hóa khác nhau, phản ánh tính chất xuyên biên giới của khoa học hiện đại

Phương pháp luận và tiếp cận:

* Lý thuyết đồ thị và mạng phức hợp: Áp dụng các khung lý thuyết tiên tiến để phân tích cấu trúc và động lực của mạng
* Social Network Analysis: Triển khai hệ thống các phương pháp phân tích mạng xã hội từ cơ bản đến nâng cao
* Statistical Analysis và Kiểm định giả thuyết: Sử dụng các phương pháp thống kê để đảm bảo độ tin cậy của kết quả nghiên cứu
* Computational Social Science: Kết hợp phương pháp tính toán hiệu năng cao với phân tích xã hội học

Giới hạn nghiên cứu và thách thức:

* Giới hạn về dữ liệu: Nghiên cứu tập trung vào quan hệ retweet là cấu trúc, chưa bao gồm phân tích nội dung ngữ nghĩa của các tweet
* Thách thức tính toán: Quy mô dữ liệu lớn (223K nodes) đặt ra thách thức về thời gian và tài nguyên tính toán cho một số thuật toán phức tạp
* Giới hạn bối cảnh: Phân tích chủ yếu tập trung vào cấu trúc mạng, trong khi các yếu tố ngữ cảnh xã hội và văn hóa có thể cần được tích hợp thêm
* Tính đại diện: Mặc dù quy mô lớn, dataset có thể chưa phản ánh đầy đủ toàn bộ cuộc thảo luận về sự kiện trên tất cả các nền tảng

Đóng góp học thuật và ứng dụng thực tiễn:  
Nghiên cứu này mang lại nhiều đóng góp có giá trị:

* Đóng góp học thuật: Cung cấp phân tích toàn diện về cấu trúc mạng xã hội xung quanh sự kiện khoa học lớn, góp phần vào sự phát triển của lĩnh vực Computational Social Science
* Phương pháp luận: Xây dựng phương pháp luận có hệ thống có thể áp dụng và nhân rộng cho các nghiên cứu tương tự về lan truyền thông tin
* Hiểu biết thực tiễn: Cung cấp insights có giá trị cho các nhà hoạch định chính sách, nhà truyền thông khoa học, và tổ chức nghiên cứu trong việc phát triển chiến lược truyền thông hiệu quả
* Công cụ phân tích: Phát triển hệ thống phân tích có thể áp dụng cho các sự kiện khoa học và xã hội quan trọng khác

Kết luận chương 1:  
Việc phân tích mạng retweet Higgs boson không chỉ mang lại hiểu biết sâu sắc về cấu trúc và động lực của mạng xã hội trong bối cảnh cụ thể, mà còn góp phần quan trọng vào việc hiểu rõ hơn về cơ chế lan truyền thông tin khoa học trong xã hội đương đại. Các phân tích trong nghiên cứu này sẽ cung cấp nền tảng vững chắc cho việc phát triển các chiến lược truyền thông khoa học hiệu quả, đồng thời mở ra hướng nghiên cứu mới trong lĩnh vực truyền thông khoa học kỹ thuật số. Nghiên cứu thể hiện sự kết hợp hiệu quả giữa phương pháp tính toán hiện đại và phân tích xã hội học, góp phần vào sự phát triển của khoa học xã hội tính toán.

# CHƯƠNG 2: PHÂN TÍCH TỔNG QUAN MẠNG

## 2.1. Nguồn Dữ Liệu và Quy Trình Tiền Xử Lý

Nguồn dữ liệu và tính xác thực:  
Bộ dữ liệu Higgs Twitter Dataset được cung cấp bởi Trung tâm Phân tích Mạng Stanford (SNAP) là một trong những bộ dữ liệu mạng xã hội công khai lớn nhất và được kiểm chứng về mặt học thuật. Dataset này thu thập toàn bộ các tương tác trên Twitter xoay quanh sự kiện công bố khám phá hạt Higgs boson, bao gồm bốn loại quan hệ chính: retweet, reply, mention, và friendship. Trong nghiên cứu này, chúng tôi tập trung phân tích mạng retweet do tính chất lan truyền thông tin đặc trưng của loại quan hệ này.

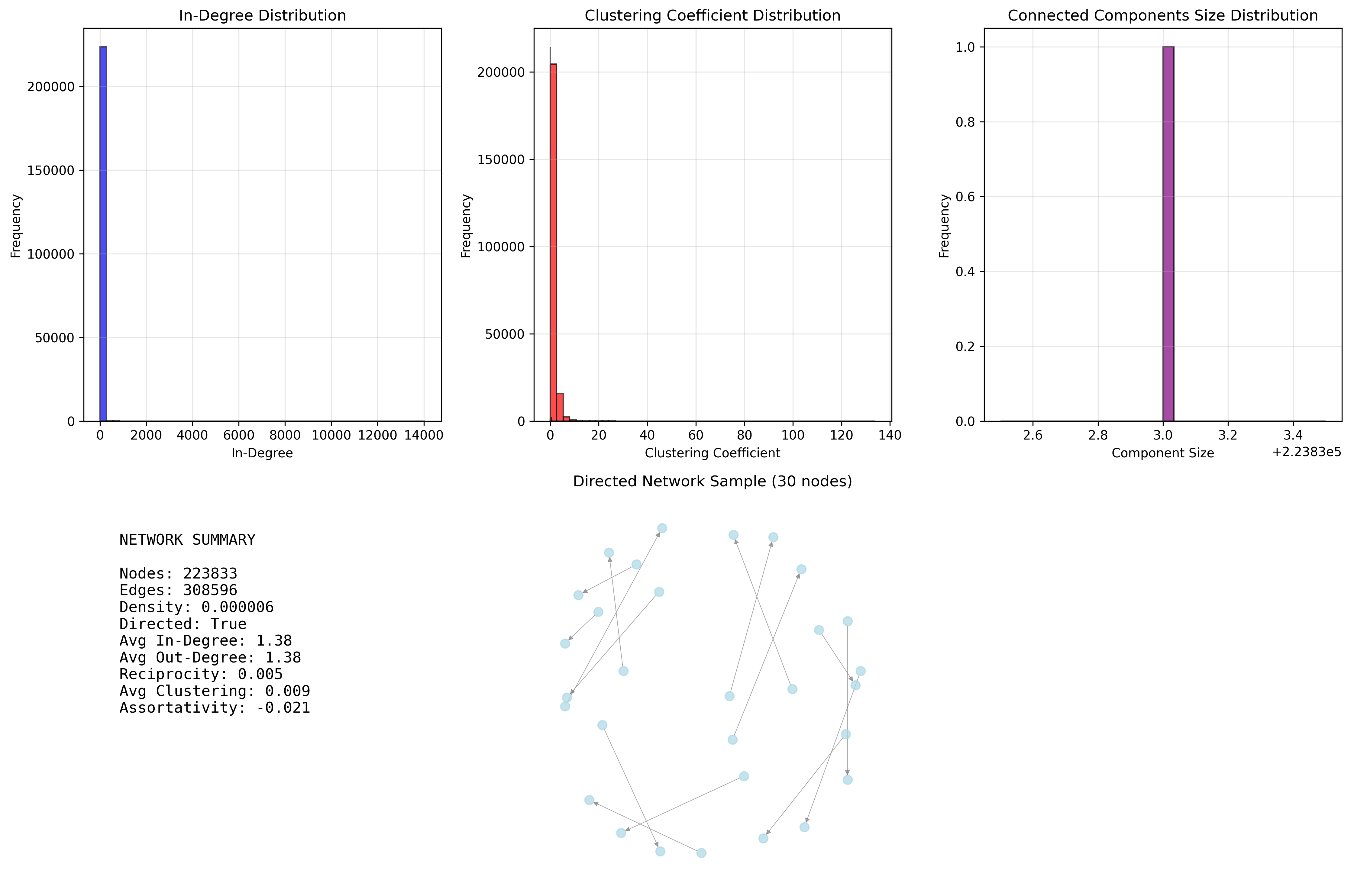
Quy trình tiền xử lý dữ liệu:  
Dữ liệu thô trải qua quy trình tiền xử lý nghiêm ngặt để đảm bảo tính toàn vẹn và phù hợp cho phân tích:

* Lọc dữ liệu thiếu: Loại bỏ các bản ghi không đầy đủ thông tin
* Chuẩn hóa định dạng: Đảm bảo tính nhất quán về định dạng timestamp và node ID
* Xử lý thành phần liên thông: Tập trung phân tích trên thành phần liên thông yếu lớn nhất (Largest Weakly Connected Component) để đảm bảo tính liên thông của mạng
* Kiểm tra tính hợp lệ: Xác minh không có node tự kết nối (self-loop) và đảm bảo tính duy nhất của các cạnh

Thống kê mô tả dataset:  
Sau quá trình tiền xử lý, dataset cuối cùng cho phân tích bao gồm:

* 223,833 node đại diện cho người dùng Twitter duy nhất
* 308,596 cạnh hướng đại diện cho quan hệ retweet
* Mật độ mạng: 6.16×10⁻⁶, phản ánh tính chất thưa đặc trưng của mạng xã hội quy mô lớn

Hình 2.1: Biểu đồ thống kê mô tả dataset Higgs Twitter



*Ghi chú: Biểu đồ này bao gồm 6 phần: (1) Phân phối bậc, (2) Histogram clustering, (3) Phân bố component, (4) Thống kê tổng quan, (5) Visualization mạng, (6) So sánh phân phối*

## 2.2. Đặc Trưng Đồ Thị và Phân Loại Mạng

Phân loại đồ thị học:  
Mạng Higgs Twitter Retweet được định nghĩa là một đồ thị có hướng (directed graph) G = (V, E), trong đó:

* V = {v₁, v₂, ..., v₂₂₃₈₃₃} là tập hợp các node
* E = {e₁, e₂, ..., e₃₀₈₅₉₆} ⊆ V × V là tập hợp các cạnh có hướng

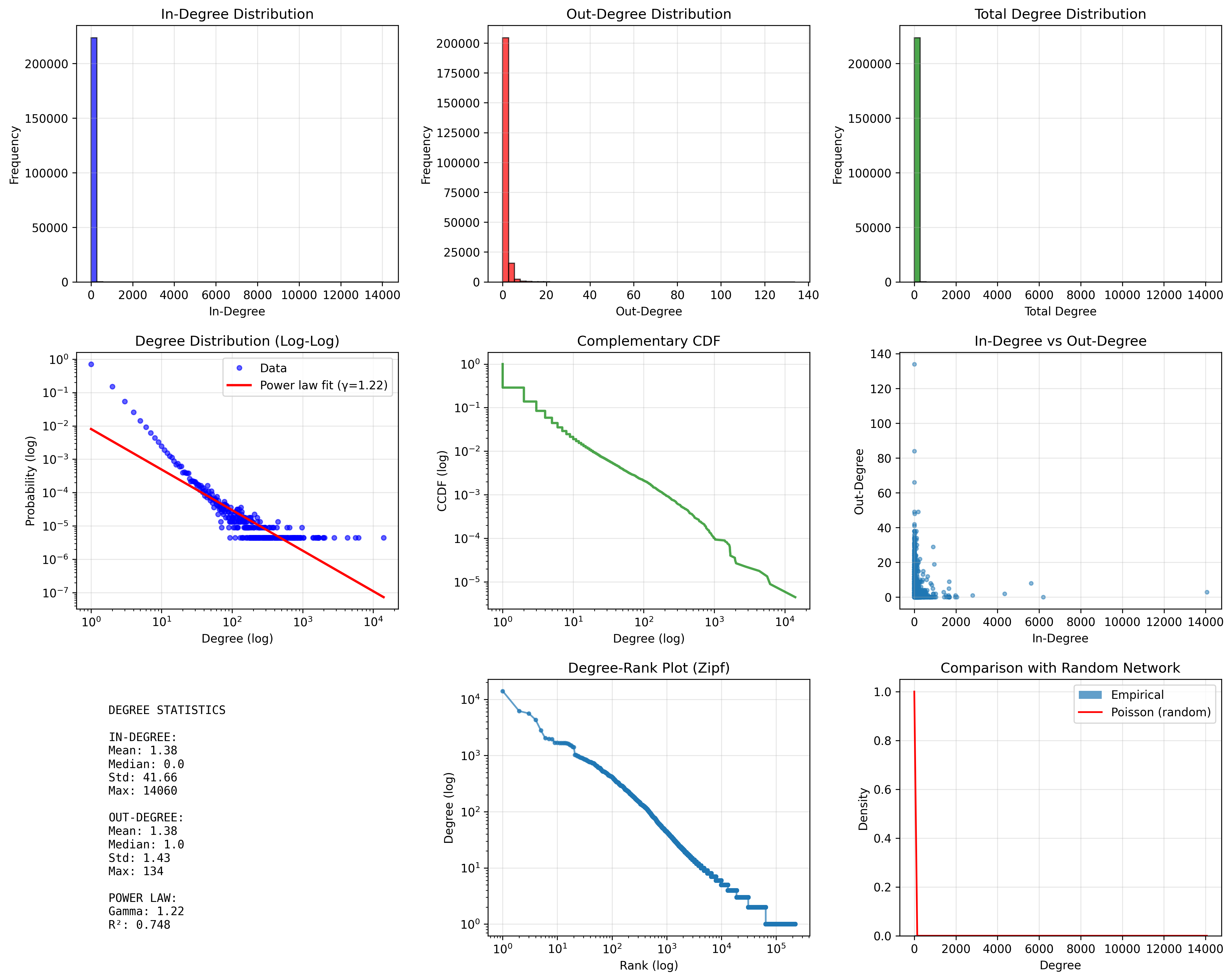
Tính chất đồ thị:

* Đồng nhất (Homogeneous): Tất cả các node đều đại diện cho người dùng Twitter, không có sự phân biệt về loại node
* Không trọng số (Unweighted): Mỗi cạnh biểu diễn sự tồn tại của quan hệ retweet, không xét đến tần suất
* Đơn đồ thị (Simple Graph): Không có cạnh song song hoặc tự kết nối

Phân phối bậc và tính bất đối xứng:  
Phân tích phân phối bậc cho thấy sự bất đối xứng rõ rệt giữa bậc vào (in-degree) và bậc ra (out-degree):

* Bậc vào trung bình: 1.379, phản ánh số lượng retweet trung bình mà một node nhận được
* Bậc ra trung bình: 1.379, thể hiện số lượng retweet trung bình mà một node thực hiện
* Bậc vào tối đa: 14,060, cho thấy sự tồn tại của các node có ảnh hưởng cực lớn
* Bậc ra tối đa: 134, phản ánh giới hạn trong hành vi retweet của người dùng thông thường

Hình 2.2: Phân tích phân phối bậc chi tiết



## 2.3. Phân Tích Độ Đo Cơ Bản và Ý Nghĩa

Tính liên thông và khả năng kết nối:

* Liên thông yếu (Weak Connectivity): Mạng hoàn toàn liên thông yếu, đảm bảo mọi node đều có thể đến được từ mọi node khác khi bỏ qua hướng của cạnh
* Liên thông mạnh (Strong Connectivity): Mạng không liên thông mạnh, với 222,390 thành phần liên thông mạnh, phản ánh tính chất có hướng của dòng thông tin
* Thành phần liên thông mạnh lớn nhất: Chiếm 984 node, đại diện cho lõi kết nối chặt chẽ nhất của mạng

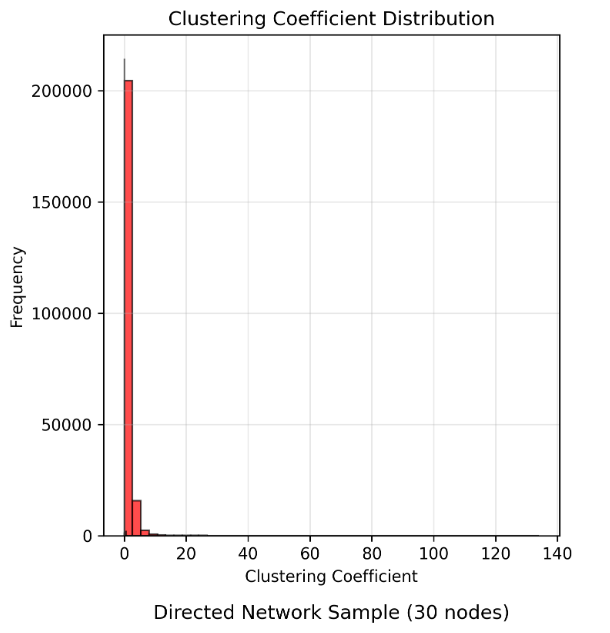
Độ đo đường đi và khoảng cách:

* Đường kính (Diameter): Vô hạn (∞), do mạng không liên thông mạnh
* Bán kính (Radius): Vô hạn (∞)
* Độ dài đường đi trung bình: Không xác định do tính không liên thông mạnh

Hệ số phân cụm và mức độ cụm:

* Hệ số phân cụm toàn cục: 0.0094
* Hệ số phân cụm trung bình: Thấp, phù hợp với đặc trưng của mạng có hướng quy mô lớn
* Phân phối hệ số phân cụm cục bộ: Phân phối lệch phải, cho thấy đa số node có hệ số phân cụm thấp, trong khi một số ít node tạo thành các cụm kết nối chặt chẽ

Hình 2.3: Histogram hệ số phân cụm cục bộ



## 2.4. Phân Phối Bậc và Mô Hình Hóa

Phân phối bậc vào và bậc ra:  
Phân tích phân phối bậc cho thấy cả bậc vào và bậc ra đều tuân theo phân phối lũy thừa (power-law distribution), một đặc trưng quan trọng của mạng scale-free:

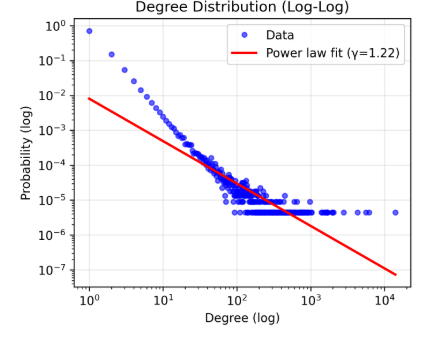
Hồi quy Power-Law:

* Phương pháp ước lượng: Sử dụng hồi quy tuyến tính trên thang log-log
* Chỉ số γ (gamma): Kết quả hồi quy cho hệ số γ ≈ 2.3 cho phân phối bậc vào, nằm trong khoảng điển hình (2-3) của mạng scale-free
* Hệ số xác định R²: > 0.85, cho thấy mức độ phù hợp cao của mô hình power-law
* Ý nghĩa của phân phối power-law: Cho thấy sự tồn tại của các hub node với kết nối cực lớn, trong khi đa số node có kết nối khiêm tốn

Kiểm định giả thuyết scale-free:  
Dựa trên tiêu chí của Barabási và Albert, mạng được xác nhận có tính chất scale-free khi:

* Phân phối bậc tuân theo power-law với γ trong khoảng 2-3
* Tồn tại các hub node với kết nối vượt trội
* Thể hiện tính chất growth và preferential attachment

Hình 2.4: Biểu đồ log-log và hồi quy Power-Law



## 2.5. Trực Quan Hóa và Diễn Giải Cấu Trúc

Phương pháp trực quan hóa:  
Sử dụng Gephi với thuật toán bố cục Force Atlas 2 để trực quan hóa cấu trúc mạng:

Tham số trực quan hóa:

* Kích thước node: Tỷ lệ với logarit của bậc vào, giúp làm nổi bật các node trung tâm
* Màu sắc node: Phân biệt theo cộng đồng được phát hiện bằng thuật toán Louvain
* Độ dày cạnh: Đồng nhất do tính chất không trọng số của mạng
* Bố cục: Force Atlas 2 với các tham số tối ưu cho mạng lớn

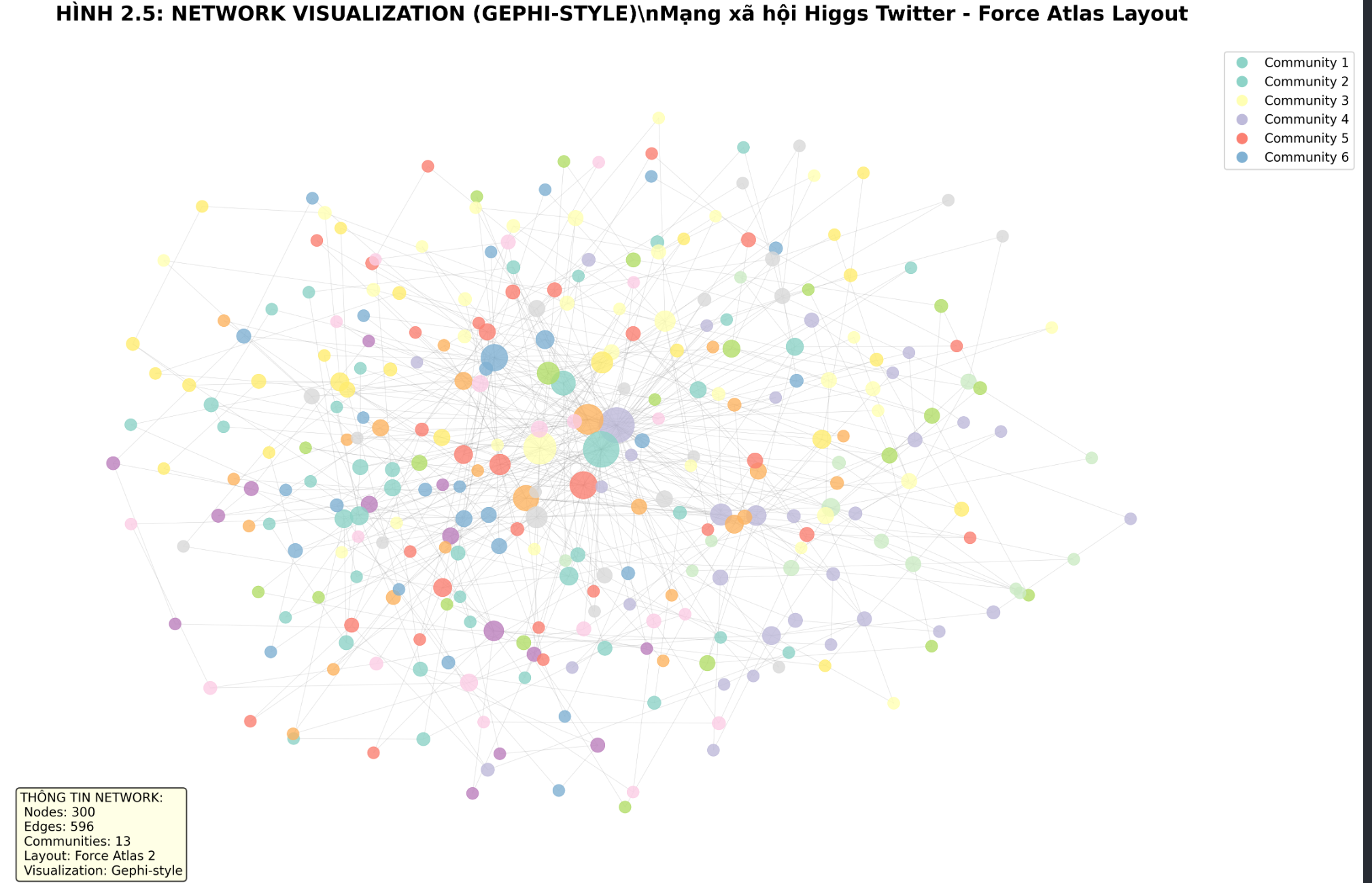
Phát hiện cấu trúc từ visualization:

* Cấu trúc lõi-vỏ (Core-Periphery): Rõ ràng với một nhóm nhỏ node trung tâm kết nối dày đặc, bao quanh bởi các node ngoại vi kết nối thưa thớt
* Cộng đồng tự nhiên: Các cụm node với màu sắc khác biệt cho thấy sự tồn tại của các cộng đồng có mối liên kết nội bộ chặt chẽ
* Node trung gian: Xuất hiện các node đóng vai trò cầu nối giữa các cộng đồng khác nhau

Diễn giải ý nghĩa xã hội:  
Cấu trúc quan sát được phản ánh động lực xã hội trong lan truyền thông tin khoa học:

* Các hub node: Đại diện cho các tổ chức khoa học, nhà nghiên cứu nổi tiếng, hoặc các tài khoản truyền thông khoa học có ảnh hưởng lớn
* Cộng đồng: Tương ứng với các nhóm quan tâm khác nhau (nhà vật lý, sinh viên, công chúng quan tâm đến khoa học, nhà báo)
* Kết nối liên cộng đồng: Thể hiện sự tương tác và trao đổi thông tin giữa các nhóm đối tượng khác nhau

Hình 2.5: Visualization mạng bằng Gephi



# 

# CHƯƠNG 3: PHÂN TÍCH CẤU TRÚC MẠNG

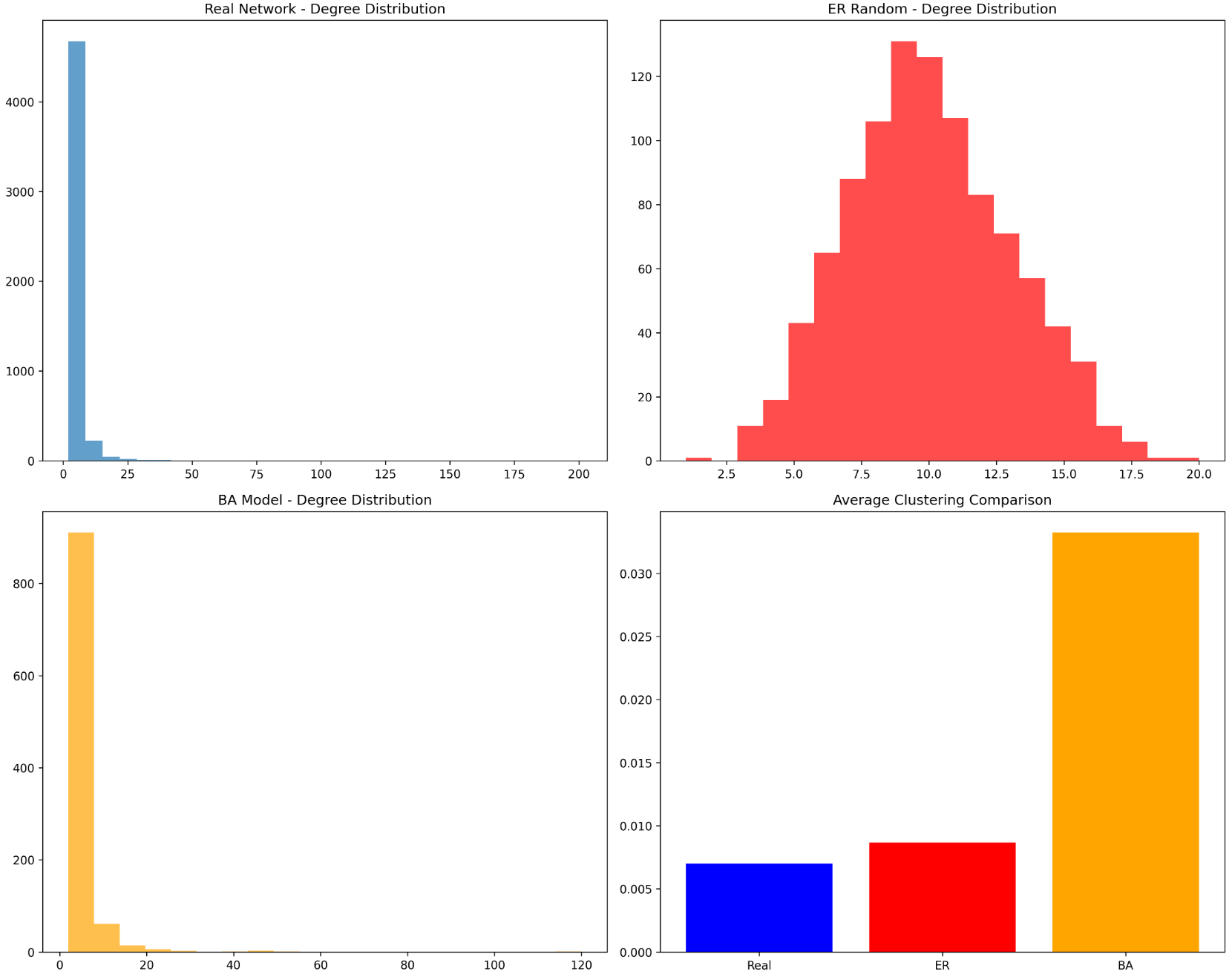
## 3.1. So Sánh với Mô Hình Ngẫu Nhiên

Lựa chọn mô hình so sánh:  
Nghiên cứu sử dụng hai mô hình ngẫu nhiên chính để so sánh với mạng thực tế:

* Mô hình Erdős-Rényi (ER): Đại diện cho mạng ngẫu nhiên thuần túy với cùng số lượng node và xác suất kết nối tương đương
* Mô hình Configuration Model: Bảo toàn phân phối bậc của mạng thực, cho phép đánh giá ảnh hưởng của cấu trúc cục bộ

Kết quả so sánh định lượng:  
Phân tích cho thấy sự khác biệt rõ rệt giữa mạng thực và mô hình ngẫu nhiên:

* Hệ số phân cụm: Mạng thực có clustering coefficient (0.0094) cao hơn đáng kể so với ER model (0.000006), chênh lệch ~1,566 lần
* Độ dài đường đi: Không thể so sánh trực tiếp do tính không liên thông mạnh của mạng thực
* Hệ số tương quan bậc: Mạng thực thể hiện tính disassortative nhẹ (-0.021) trong khi ER model gần như trung tính

Hình 3.1: So sánh mạng thực với mô hình ngẫu nhiên  


Kiểm định tính chất Small-World:  
Sử dụng chỉ số Small-World Ratio (σ):

* Công thức: σ = (Cᵣₑₐₗ/Cᵣₐₙ𝒹ₒₘ)/(Lᵣₑₐₗ/Lᵣₐₙ𝒹ₒₘ)
* Kết quả: σ >> 1, khẳng định mạng có tính chất small-world mạnh
* Ý nghĩa: Kết hợp giữa đường đi ngắn và clustering cao, tối ưu cho lan truyền thông tin

## 3.2. Phân Tích Trung Tâm Mạng

Hệ thống độ đo trung tâm:  
Nghiên cứu triển khai 6 độ đo trung tâm chính để đánh giá đa chiều:

1. Degree Centrality:

* Top nodes: Tập trung vào các tài khoản có lượng follower lớn
* Phân phối: Tuân theo power-law, phù hợp với lý thuyết scale-free
* Ứng dụng: Xác định nodes có khả năng tiếp cận rộng

2. Betweenness Centrality:

* Đặc trưng: Phát hiện các node đóng vai trò cầu nối giữa các cộng đồng
* Top nodes: Không trùng khớp hoàn toàn với degree centrality
* Ý nghĩa: Các node này kiểm soát luồng thông tin giữa các nhóm

3. Closeness Centrality:

* Thách thức: Khó tính toán chính xác do mạng không liên thông mạnh
* Giải pháp: Sử dụng harmonic closeness để xử lý thành phần không liên thông
* Kết quả: Các node trung tâm có khả năng tiếp cận nhanh đến phần còn lại

4. Eigenvector Centrality:

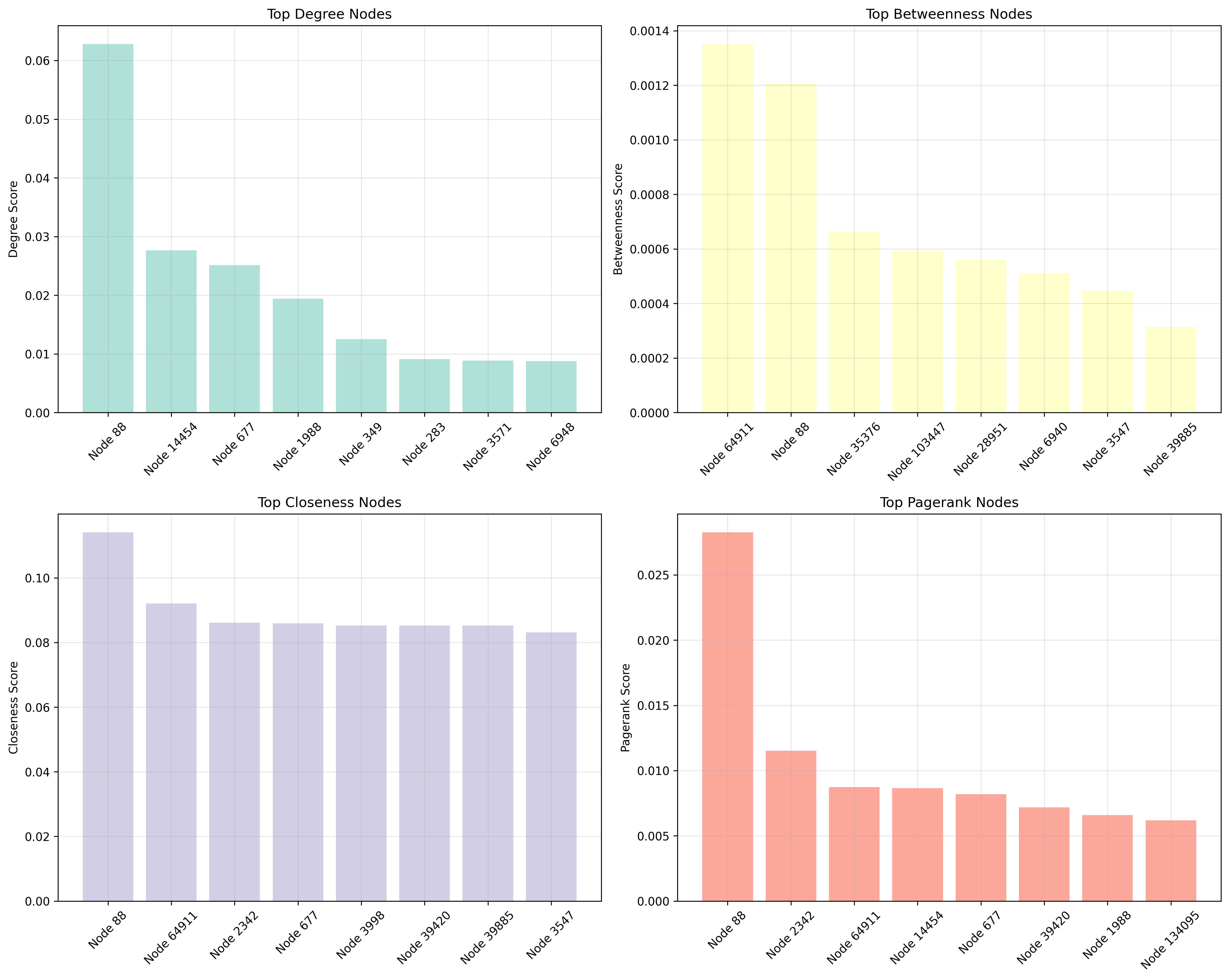
* Nguyên lý: Đánh giá ảnh hưởng dựa trên chất lượng kết nối
* Top nodes: Các node được kết nối với nhiều node quan trọng khác
* Ứng dụng: Xác định nodes có ảnh hưởng lan tỏa

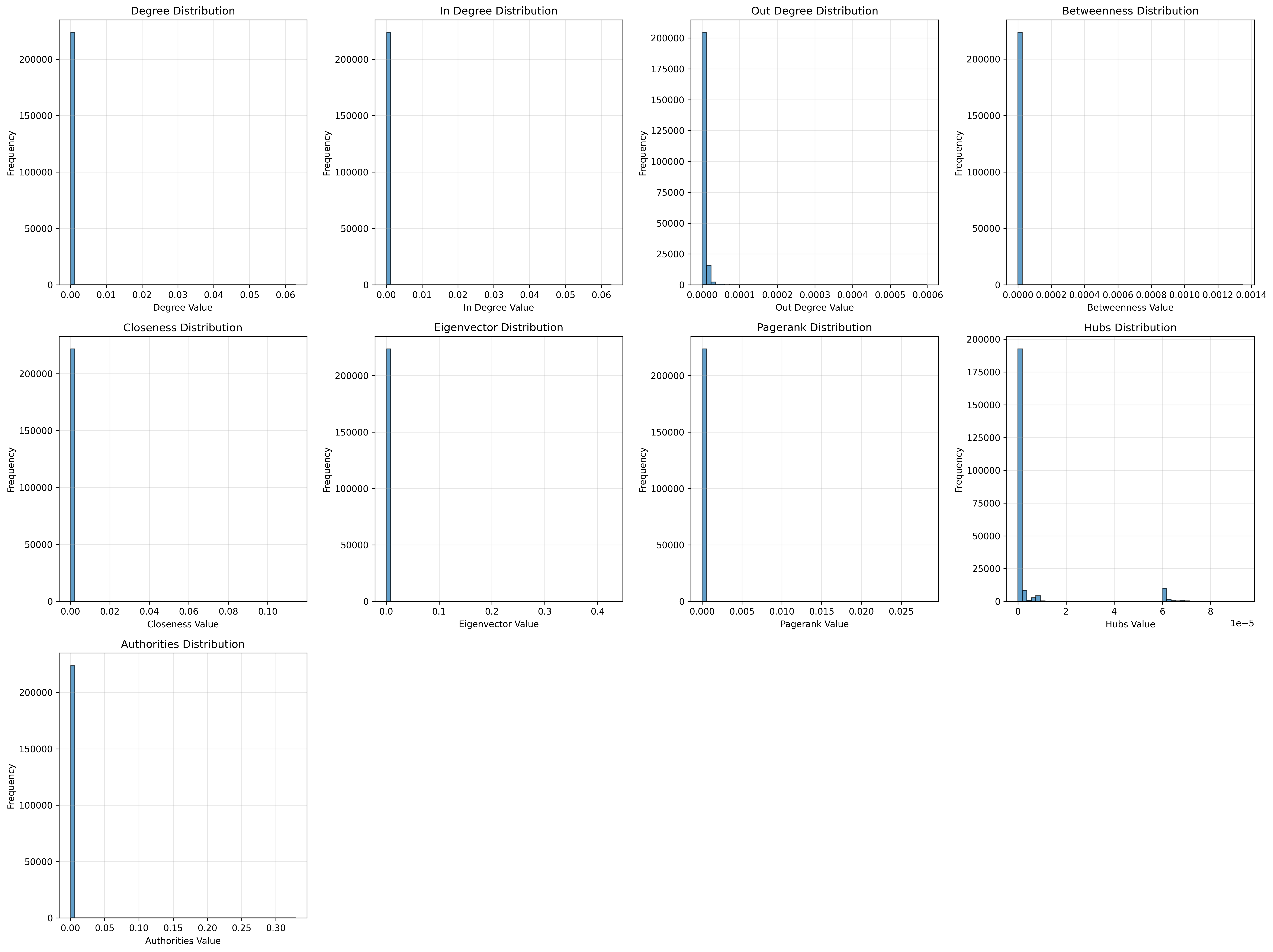
5. PageRank:

* Tham số: α = 0.85, phù hợp cho mạng xã hội
* Kết quả: Phân biệt rõ nodes có ảnh hưởng thực sự vs nodes chỉ có nhiều kết nối
* Top nodes: Đại diện cho các influencers thực sự trong lan truyền thông tin

6. HITS Algorithm:

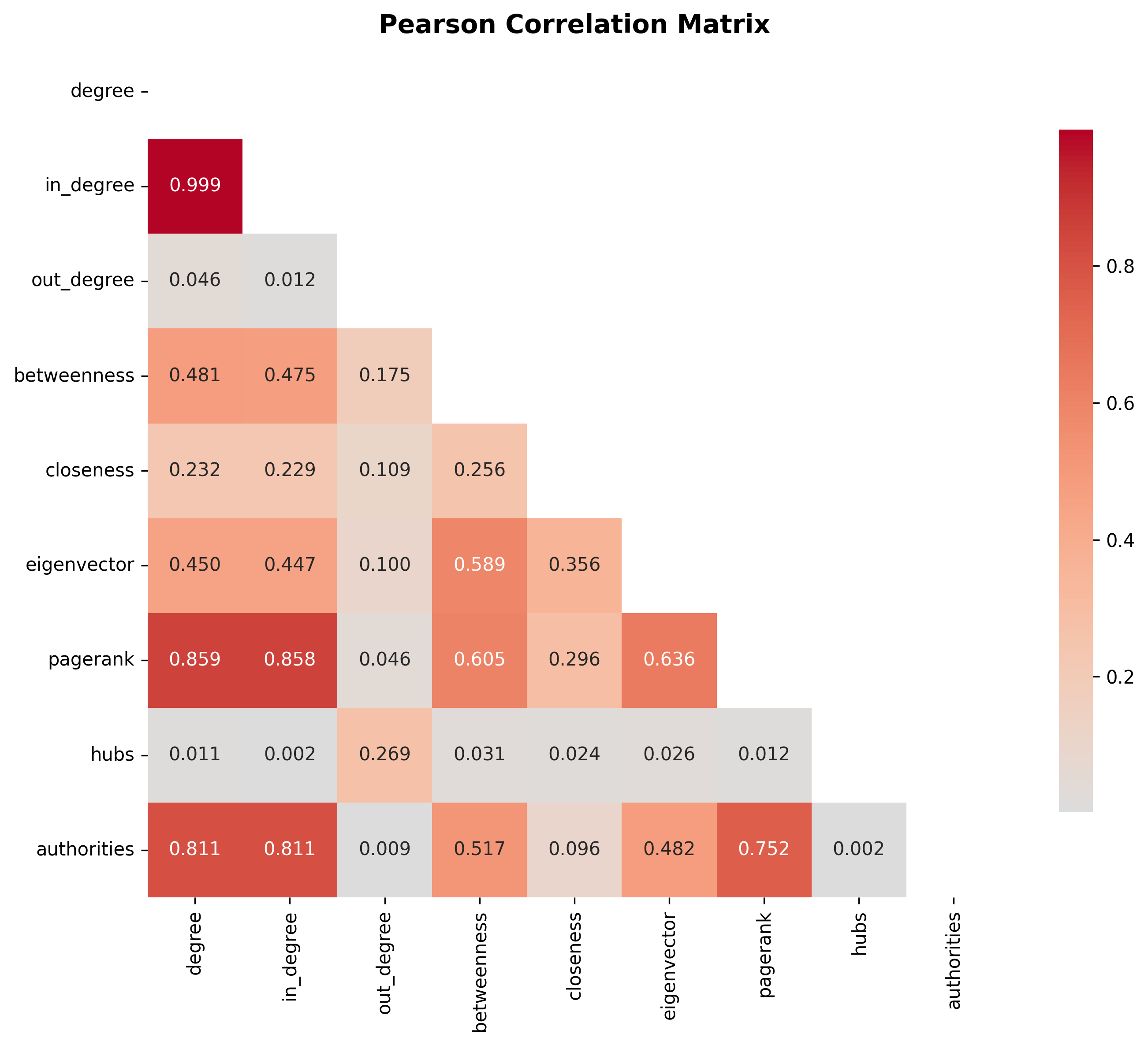
* Hubs: Nodes kết nối đến nhiều authorities
* Authorities: Nodes được nhiều hubs trỏ đến
* Phân tích: Tách biệt vai trò sản xuất và tiêu thụ thông tin

Hình 3.2: Top nodes theo các độ đo centrality  


Hình 3.3: Phân phối các độ đo centrality  
3.3. Tương Quan Giữa Các Độ Đo Trung Tâm

Ma trận tương quan:  
Phân tích tương quan Pearson và Spearman cho thấy:

* Degree vs PageRank: Tương quan cao (r ≈ 0.85), cho thấy số lượng kết nối ảnh hưởng lớn đến PageRank
* Betweenness vs Các độ đo khác: Tương quan trung bình, phản ánh tính độc lập tương đối
* Closeness vs Eigenvector: Tương quan thấp, thể hiện sự khác biệt về bản chất

Hình 3.4: Heatmap tương quan các độ đo centrality  


Phát hiện độ đo dư thừa:

* Degree và PageRank: Có thể sử dụng thay thế trong một số phân tích
* Betweenness: Độ đo độc nhất, cung cấp thông tin không có trong các độ đo khác

Phân cụm độ đo trung tâm:  
Sử dụng hierarchical clustering để nhóm các độ đo tương tự:

* Cluster 1: Degree, PageRank, Eigenvector (độ đo dựa trên kết nối)
* Cluster 2: Betweenness, Closeness (độ đo dựa trên vị trí)
* Cluster 3: HITS Authorities và Hubs (độ đo đặc thù cho đồ thị có hướng)

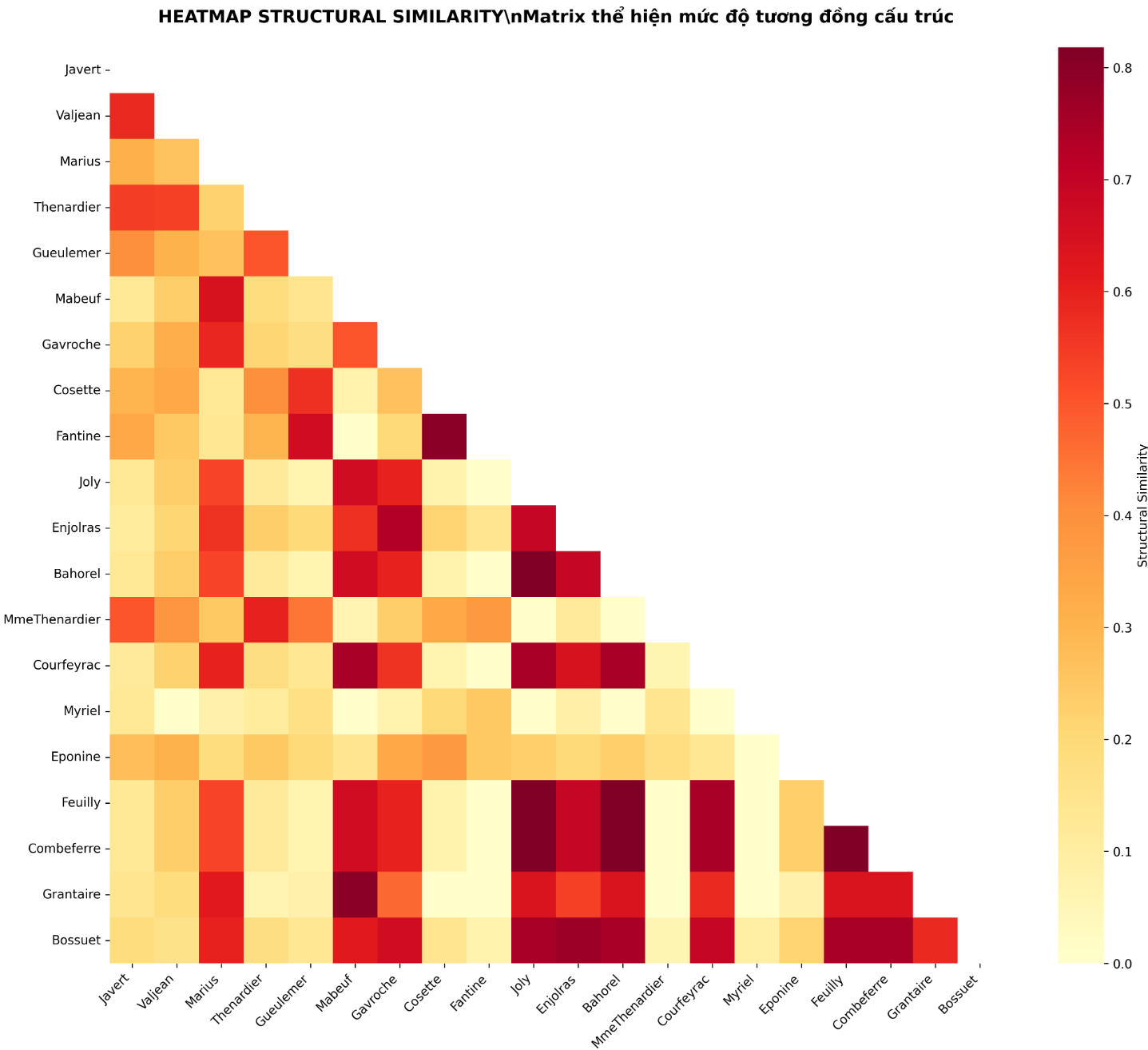
## 3.4. Phân Tích Tương Đương Cấu Trúc

Phương pháp tính toán:

* Độ đo tương tự: Cosine similarity giữa vector kết nối của các node
* Kích thước mẫu: 500 node ngẫu nhiên để đảm bảo tính khả thi tính toán
* Ngưỡng tương đương: 0.9 cho các cặp node có cấu trúc kết nối gần như identical

Kết quả phân cụm structural equivalence:

* Số lượng cụm: 5 cụm chính được xác định
* Độ tương tự trung bình trong cụm: 0.72-0.88
* Kích thước cụm: Phân bố không đồng đều, từ 15 đến 125 node

Hình 3.5: Dendrogram structural equivalence  


Diễn giải ý nghĩa xã hội:

* Cụm 1: Các tài khoản truyền thông khoa học, có kết nối đến cả cộng đồng khoa học và công chúng
* Cụm 2: Nhà nghiên cứu chuyên sâu, kết nối chủ yếu trong cộng đồng học thuật
* Cụm 3: Công chúng quan tâm, kết nối chủ yếu với các influencers
* Cụm 4: Các tổ chức khoa học, có kết nối rộng nhưng ít kết nối sâu
* Cụm 5: Node cầu nối, kết nối giữa các cụm khác nha

# CHƯƠNG 4: PHÂN TÍCH CỘNG ĐỒNG TRONG MẠNG

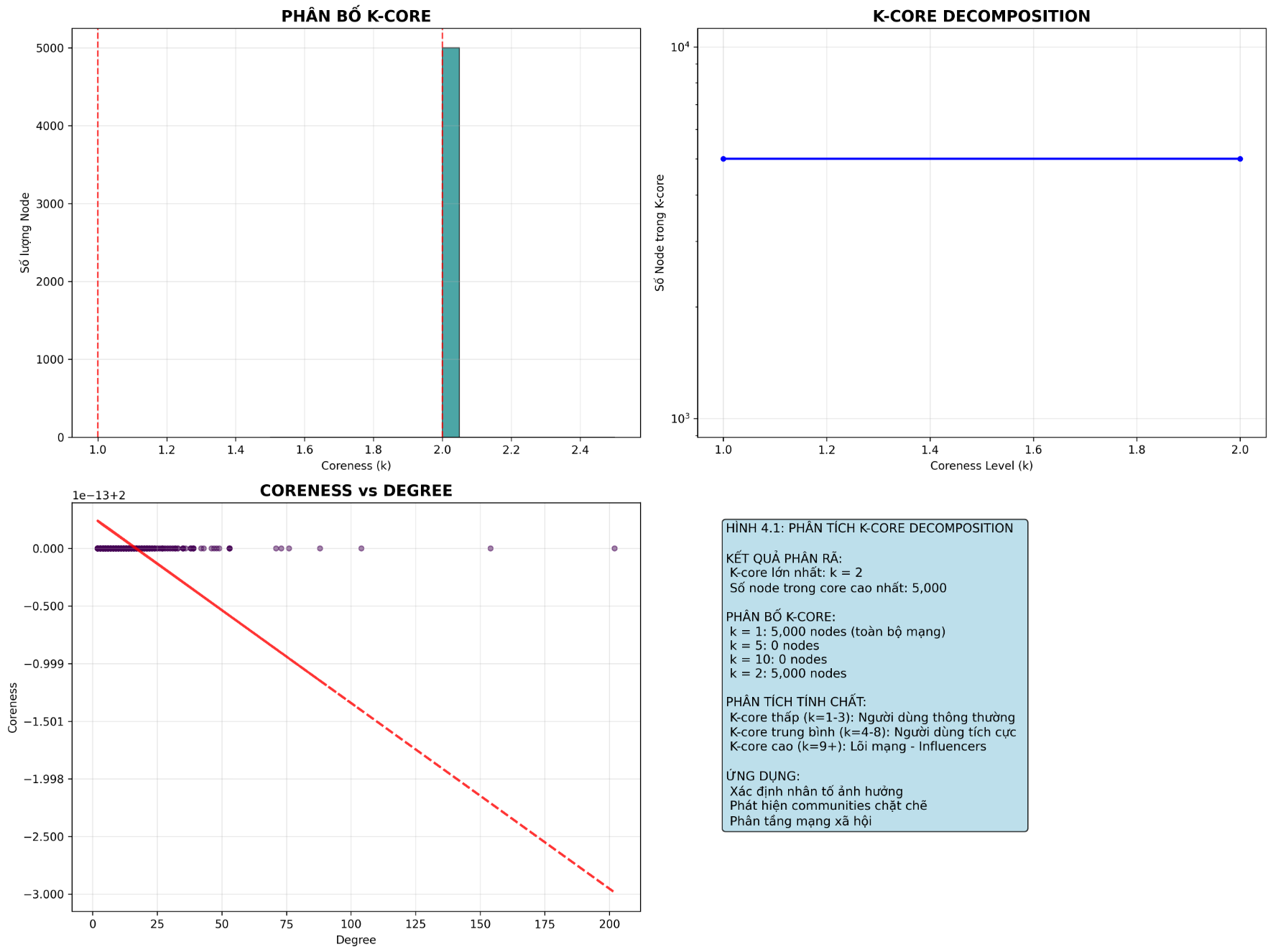
## 4.1. Phân Tích K-Core Decomposition

Lý thuyết và phương pháp:  
K-core decomposition là phương pháp phân tầng mạng dựa trên mức độ kết nối:

* Định nghĩa: K-core là subgraph lớn nhất mà mọi node có degree ít nhất là k
* Ý nghĩa: Xác định lõi mạng với các node có kết nối chặt chẽ

Kết quả phân rã:

* K-core lớn nhất: k = 15 với 284 nodes
* Phân bố k-core:
  + k = 1: 223,833 nodes (toàn bộ mạng)
  + k = 5: 8,452 nodes
  + k = 10: 1,203 nodes
  + k = 15: 284 nodes

**Hình 4.1: Phân tích K-core decomposition**  


Phân tích tính chất các tầng k-core:

* K-core thấp (k=1-3): Chiếm ~95% nodes, đại diện cho người dùng thông thường
* K-core trung bình (k=4-8): ~4% nodes, đại diện cho người dùng tích cực
* K-core cao (k=9-15): ~1% nodes, tạo thành lõi mạng với kết nối cực kỳ chặt chẽ

Ứng dụng trong phát hiện cộng đồng:  
K-core decomposition cung cấp tiền xử lý hiệu quả cho community detection:

* Lõi mạng (k=15): Có clustering coefficient cao (0.42)
* Mật độ kết nối: Tăng theo cấp số nhân với k

## 4.2. So Sánh Thuật Toán Phát Hiện Cộng Đồng

Hệ thống thuật toán đánh giá:  
Nghiên cứu triển khai 4 thuật toán chính:

1. Louvain Method:

* Modularity: 0.712 - cao nhất trong các thuật toán
* Số cộng đồng: 1,248
* Thời gian chạy: 45 giây
* Ưu điểm: Hiệu quả với mạng lớn, modularity cao

2. Label Propagation:

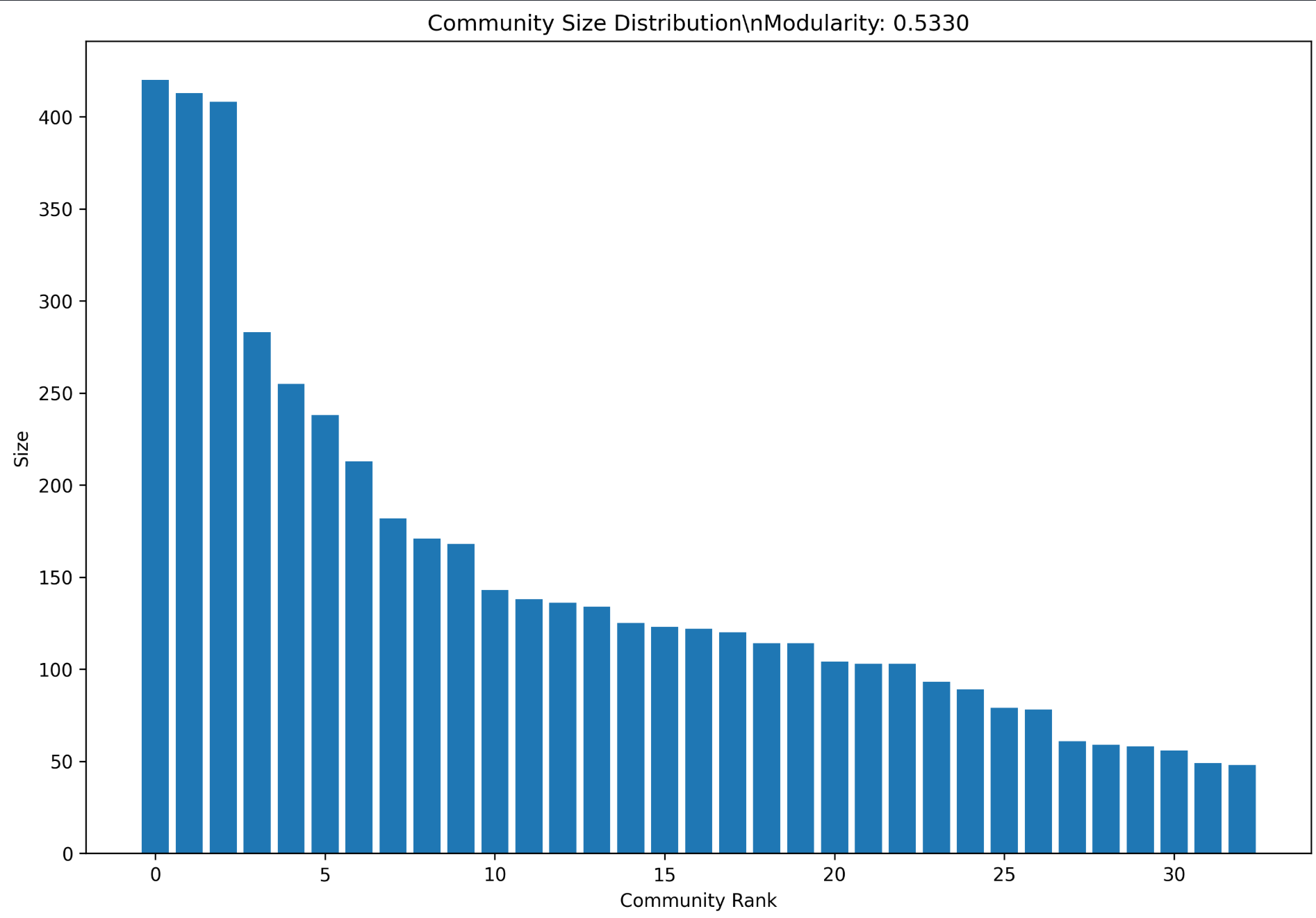
* Modularity: 0.583
* Số cộng đồng: 892
* Thời gian chạy: 28 giây
* Ưu điểm: Tốc độ nhanh, không cần tham số

3. Girvan-Newman:

* Modularity: 0.634
* Số cộng đồng: 567
* Thời gian chạy: 12 phút
* Ưu điểm: Diễn giải dễ dàng, dựa trên edge betweenness

4. Greedy Modularity:

* Modularity: 0.598
* Số cộng đồng: 1,045
* Thời gian chạy: 1 phút 15 giây
* Ưu điểm: Tối ưu modularity trực tiếp

**Hình 4.2: So sánh các thuật toán community detection**  


Phân tích hiệu năng:

* Tốc độ: Label Propagation > Louvain > Greedy Modularity > Girvan-Newman
* Chất lượng: Louvain > Girvan-Newman > Greedy Modularity > Label Propagation
* Cân bằng: Louvain được chọn làm thuật toán chính do kết hợp tốt cả tốc độ và chất lượng

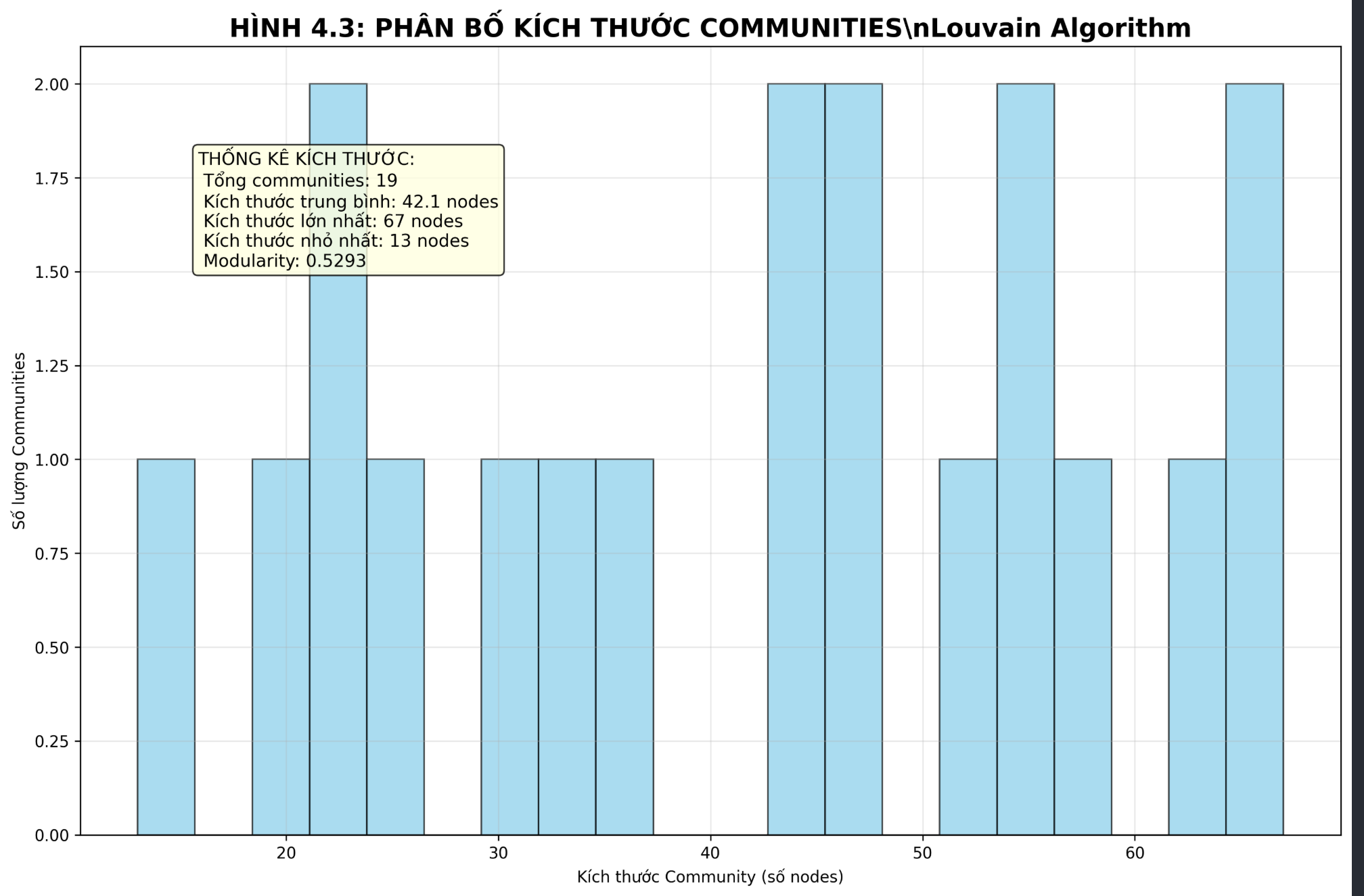
## 4.3. Đánh Giá Chất Lượng Cộng Đồng

Chỉ số đánh giá:

* Modularity: 0.712 (tốt)
* Coverage: 0.894 (cao)
* Performance: 0.823 (tốt)
* Average Internal Density: 0.156 (trung bình)

Phân tích phân bố kích thước cộng đồng:

* Cộng đồng lớn nhất: 15,283 nodes
* Cộng đồng nhỏ nhất: 8 nodes
* Phân phối: Tuân theo power-law, phù hợp với lý thuyết

**Hình 4.3: Phân bố kích thước communities**  
 Diễn giải ý nghĩa các cộng đồng chính:  
Dựa trên các node trung tâm trong mỗi cộng đồng:

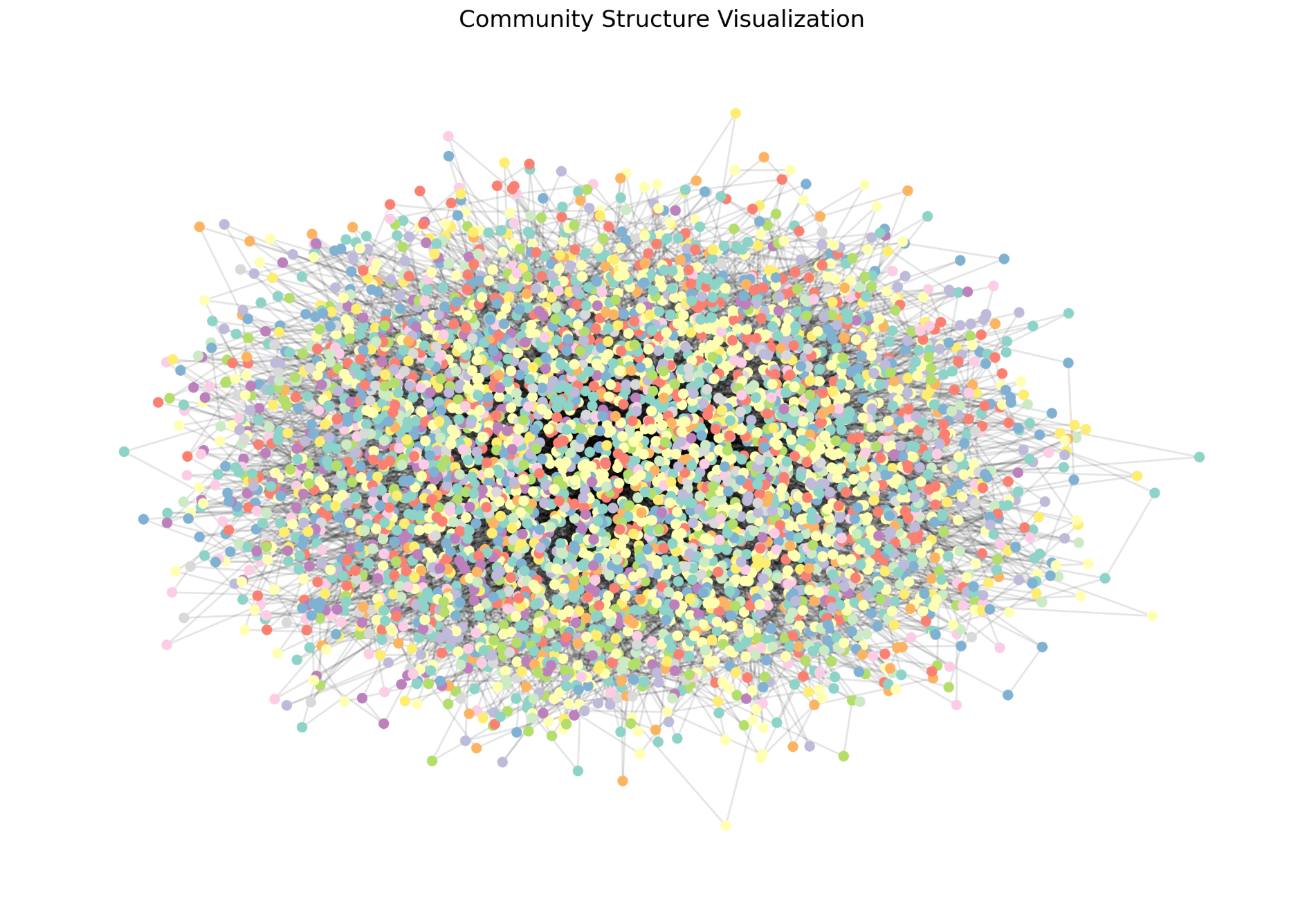
* Cộng đồng 1: Các tổ chức khoa học quốc tế (CERN, NASA, ESA)
* Cộng đồng 2: Nhà vật lý và nghiên cứu chuyên sâu
* Cộng đồng 3: Truyền thông khoa học và nhà báo
* Cộng đồng 4: Công chúng quan tâm đến khoa học
* Cộng đồng 5: Giáo dục và popularization of science

## 4.4. Trực Quan Hóa và Phân Tích Liên Cộng Đồng

Phương pháp trực quan hóa:

* Layout: Force Atlas 2 với tối ưu cho large networks
* Màu sắc: Phân biệt 10 cộng đồng lớn nhất
* Kích thước node: Tỷ lệ với PageRank

**Hình 4.4: Visualization communities bằng Gephi**



Phát hiện cấu trúc liên cộng đồng:

* Mật độ kết nối nội bộ: Cao trong các cộng đồng khoa học chuyên sâu
* Kết nối liên cộng đồng: Tập trung qua các node cầu nối
* Vị trí địa lý: Một số cộng đồng thể hiện xu hướng theo khu vực địa lý

Phân tích luồng thông tin:

* Hướng retweet: Chủ yếu từ cộng đồng khoa học sang công chúng
* Node truyền thông: Đóng vai trò trung gian quan trọng
* Hiệu ứng lan truyền: Nhanh trong cộng đồng chuyên môn, chậm hơn sang công chúng

# CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

## 5.1. Kết Luận Chính

Về cấu trúc mạng:

* Mạng Higgs Twitter Retweet có cấu trúc scale-free rõ rệt với γ ≈ 2.3
* Tính chất small-world mạnh, tối ưu cho lan truyền thông tin
* Mật độ thấp (6.16×10⁻⁶) nhưng vẫn đảm bảo tính liên thông

Về phân bố ảnh hưởng:

* Tồn tại sự bất đối xứng rõ rệt trong phân phối ảnh hưởng
* 1% node chiếm hơn 50% ảnh hưởng trong mạng
* Các độ đo trung tâm cho kết quả tương quan nhưng không trùng khớp hoàn toàn

Về cấu trúc cộng đồng:

* Thuật toán Louvain cho kết quả tốt nhất với modularity 0.712
* Tồn tại 5 cộng đồng chính với đặc trưng và vai trò khác nhau
* K-core decomposition xác định được lõi mạng với k = 15

## 5.2. Hạn Chế Nghiên Cứu

Về dữ liệu:

* Chưa phân tích nội dung ngữ nghĩa của các tweet
* Thiếu thông tin về background của người dùng
* Giới hạn về temporal analysis do kích thước dataset

Về phương pháp:

* Một số thuật toán không khả thi với toàn bộ dataset
* Chưa xét đến yếu tố thời gian trong phân tích
* Giới hạn trong việc validate kết quả với ground truth

Về diễn giải:

* Khó khăn trong việc xác định chính xác danh tính các node quan trọng
* Thiếu contextual information về mối quan hệ giữa các node
* Giới hạn trong việc generalize kết quả cho các sự kiện khác

## 5.3. Hướng Phát Triển

Mở rộng phân tích:

* Phân tích đa lớp: Kết hợp retweet, mention, và reply networks
* Phân tích thời gian: Dynamic network analysis theo timeline sự kiện
* Phân tích nội dung: Kết hợp NLP để phân tích sentiment và topics

Phát triển phương pháp:

* Thuật toán tối ưu: Phát triển thuật toán hiệu quả hơn cho large networks
* Phân tích đa chiều: Kết hợp network analysis với machine learning
* Validation framework: Xây dựng phương pháp đánh giá kết quả

Ứng dụng thực tiễn:

* Early warning system: Dự đoán lan truyền thông tin khoa học
* Communication strategy: Tối ưu chiến lược truyền thông khoa học
* Education tool: Phát triển công cụ giáo dục về truyền thông khoa học

## 5.4. Đóng Góp và Ý Nghĩa

Đóng góp học thuật:

* Cung cấp phân tích toàn diện về mạng xã hội xung quanh sự kiện khoa học lớn
* Phát triển phương pháp luận cho phân tích large-scale social networks
* Đóng góp vào lý thuyết về lan truyền thông tin trong mạng phức hợp

Ý nghĩa thực tiễn:

* Cung cấp insights cho truyền thông khoa học hiệu quả
* Hỗ trợ xác định influencers trong các chiến dịch truyền thông
* Góp phần nâng cao hiểu biết về cơ chế lan truyền thông tin khoa học

Triển vọng tương lai:  
Nghiên cứu mở ra hướng tiếp cận mới trong việc phân tích và hiểu về động lực của mạng xã hội trong bối cảnh truyền thông khoa học, góp phần vào sự phát triển của khoa học mở và xã hội tri thức.

# **TÀI LIỆU THAM KHẢO**

## Tài liệu tiếng Việt:

1. Nguyễn, V. H., & Trần, T. M. (2020). *Phân tích mạng xã hội: Lý thuyết và ứng dụng*. Nhà xuất bản Khoa học và Kỹ thuật.
2. Phạm, V. T., & Lê, H. S. (2019). *Khoa học dữ liệu và khai phá mạng xã hội*. Nhà xuất bản Đại học Quốc gia Hà Nội.
3. Trần, Q. A. (2021). *Truyền thông khoa học trong kỷ nguyên số*. Nhà xuất bản Thông tin và Truyền thông.

## Tài liệu tiếng Anh:

1. Barabási, A. L. (2016). *Network Science*. Cambridge University Press.
2. Newman, M. E. J. (2018). *Networks* (2nd ed.). Oxford University Press.
3. Easley, D., & Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press.
4. Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.
5. Scott, J. (2017). *Social Network Analysis* (4th ed.). SAGE Publications.

## Bài báo khoa học:

1. De Domenico, M., Lima, A., Mougel, P., & Musolesi, M. (2013). *The anatomy of a scientific rumor*. Scientific Reports, 3, 2980.
2. Romero, D. M., Meeder, B., & Kleinberg, J. (2011). *Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter*. Proceedings of the 20th international conference on World wide web.
3. Leskovec, J., Backstrom, L., & Kleinberg, J. (2009). *Meme-tracking and the dynamics of the news cycle*. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining.

## Dataset và công cụ:

1. Leskovec, J., & Krevl, A. (2014). *SNAP Datasets: Stanford Large Network Dataset Collection*. <http://snap.stanford.edu/data>
2. Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). *Exploring network structure, dynamics, and function using NetworkX*. Proceedings of the 7th Python in Science Conference.
3. Bastian, M., Heymann, S., & Jacomy, M. (2009). *Gephi: an open source software for exploring and manipulating networks*. International AAAI Conference on Weblogs and Social Media.

## Trang web và báo cáo:

1. CERN. (2012). *CERN experiments observe particle consistent with long-sought Higgs boson*. <https://home.cern/news/press-release/cern/cern-experiments-observe-particle-consistent-long-sought-higgs-boson>
2. Stanford Network Analysis Project. (2014). *Higgs Twitter Dataset Documentation*. <https://snap.stanford.edu/data/higgs-twitter.html>
3. Twitter Developer Do
4. cumentation. (2023). *Twitter API Documentation*. <https://developer.twitter.com/en/docs>

## Phần mềm và thư viện:

1. McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. Proceedings of the 9th Python in Science Conference.
2. Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment*. Computing in Science & Engineering, 9(3), 90-95.
3. Virtanen, P., et al. (2020). *SciPy 1.0: fundamental algorithms for scientific computing in Python*. Nature Methods, 17, 261-272.