DATA SCIENCE & TRADITIONAL MEDICINE

네트워크 분석으로 보는 한의학 처방과 본초

이론편

오준호 지음

한국한의학연구원
KOREA INSTITUTE OF ORIENTAL MEDICINE

일러두기

- 이 책은 네트워크 분석에 대해 소개하고, 그 내용을 바탕으로 한의학 처방 데이터를 분석한 결과이다.

- 이 책은 전자 출판물로서 본문 검색이 가능하므로 따로 색인을 두지 않았다.

Contents

머리말 1

1 네트워크 분석과 한의학 5

1.1 한의학 처방 5

1.2 관계의 표현 7

2 네트워크 모델링 11

2.1 데이터 정규화 11

2.2 네트워크 표현 14

3 탐색적 데이터 분석 19

3.1 처방의 크기 19

3.2 본초 출현 빈도 22

3.3 본초 배합 빈도 26

3.4 희소 데이터 처리 29

4 엣지 가중치 35

4.1 주요 지표 35

4.2 지표 비교 40

4.3 분석 목적과 지표 선택 48

5 네트워크 구조 51

5.1 무작위 모델 53
5.2 좁은 세상 구조 56
5.3 모티프 분석 60
5.4 중심부-주변부 계층 구조 62

6 중심성과 강건성 71
6.1 중심성 지표 72
6.2 중심성 분석 77
6.3 강건성 시뮬레이션 85

7 네트워크 시각화 91
7.1 힘 지향 레이아웃 92
7.2 차원 축소 및 최적화 레이아웃 99
7.3 레이아웃 알고리즘 비교 102

8 커뮤니티와 다의성 107
8.1 커뮤니티 탐지와 모듈성 107
8.2 본초의 다의성 118

9 서브네트워크 133
9.1 서브네트워크와 에고 네트워크 133
9.2 서브네트워크로서의 처방 135
9.3 전체 네트워크와 서브네트워크의 비교 137
9.4 처방 시각화 모형 143

맺음말 151

참고문헌 153

머리말

동아시아에서는 수천 년에 걸쳐 독자적인 의학이 발전했다. 사람들은 고유한 의학 사상과 치료 체계를 세워 질병에 맞섰다. 동아시아라는 공간과 수천 년이라는 시간은 거대한 역사의 실험실이었다. 오늘날 임상 현장에서 쓰이는 침과 뜸, 본초와 처방은 그 실험실에서 만들어지고 검증을 거쳐 전해진 결과물이다.

동아시아 전통의학(이하 한의학)의 치료 방법은 오늘날에도 임상 현장에서 여전히 힘을 발휘하고 있다. 그 속에는 오랜 임상 경험이 응축된 암묵지가 담겨 있다. 한의학의 방대한 이론은 이를 언어로 설명한 결과다. 다만 이 이론과 설명은 전통 사회의 사상과 개념을 바탕으로 구성되어 있어 오늘날 온전히 이해하기에는 한계가 있다. 현대 과학의 이론이나 법칙과 성격이 다르다는 점도 걸림돌이다.

그렇다면 이 암묵지를 어떻게 더 파고들 수 있을까. AI의 발전 과정에서 암묵지를 명시적 규칙으로 설명하는 데 한계가 있다는 점이 드러났다(Autor 2014). 기계 번역의 발전 과정은 이를 잘 보여준다. 연구 초기에는 언어에 내재된 문법을 정교하게 정리하면 기계 번역이 가능할 것이라 여겼다. 하지만 언어마다 존재하는 문법과 수많은 예외를 모두 명시적으로 정리할 수는 없었다. 상당 부분을 정리한 뒤에도 기계 번역의 품질은 한계에 부딪혔다. 이후 대량의 말뭉치 데이터로 통계적 모델을 학습시킨 뒤에야 비로소 기계 번역의 돌파구가 마련되었다(Brown et al. 1990). ’

규칙 기반'에서 '데이터 기반'으로 패러다임을 전환했기에 가능한 일이었다(Halevy et al. 2009; Sutton 2019).

인간이 발전시켜 온 의학도 질병(또는 증상)이라는 문제를 치료 방법이라는 해결책으로 대응시키는 과정으로 볼 수 있다. 따라서 규칙 대신 데이터에서 돌파구를 찾았던 통찰은 의학 분야에도 유효하다. 문제는 정형화된 임상 데이터의 부족이다. 데이터 기반으로 모델을 학습시키려면 질병과 치료 경과가 짝을 이루는 임상 데이터가 풍부해야 한다. 하지만 한의학에는 모델을 온전히 학습시킬 만큼 정돈된 임상 데이터가 아직 충분히 축적되어 있지 않다.

임상 데이터가 충분하지 않은 상황에서 네트워크 분석은 데이터를 기반으로 복잡한 대상을 이해하는 효과적인 접근법 가운데 하나다. 네트워크 분석은 상호 연결된 데이터를 다루는 방법론이다. 구성 요소 사이의 관계를 중심으로 전체 구조를 파악한다. 현대 의학에서도 네트워크 분석으로 서로 얽힌 생체 기능과 다중 약물 표적의 복합 작용을 규명하려는 네트워크 의학(Network Medicine)과 네트워크 약리학(Network Pharmacology)이 활발히 연구되고 있다(Hopkins 2008; Barabási et al. 2011).

처방(방제)은 여러 본초의 조합으로 구성되는 한의학의 대표적인 치료 방법이다. 본초와 본초의 배합으로 완성되는 처방은 그 자체로 네트워크다. 군신좌사(君臣佐使)나 약대(藥對)처럼 본초 간 배합 원리를 설명하는 한의학 이론은 처방이 본초로 이루어진 네트워크라는 점을 잘 뒷받침한다. 한의학의 처방 전체를 본초 사이의 관계로 엮인 하나의 거대한 네트워크로 바라보면 익히 알려진 처방 하나하나는 이 거대한 본초 네트워크의 일부분을 이루는 서브네트워크에 해당한다.

이제 네트워크 분석(Network Analysis)을 바탕으로 한의학 치료 방법, 그중에서도 처방 속 본초 배합 구조를 조망하고자 한다. 이를 위해 이 책에서는 네트워크 분석의 이론적 토대와 실제 분석 방법론을 다룬다. 한의학 처방 데이터를 수치화하고 네트워크라는 수학적 구조로 변환하여 그

안에 숨겨진 배합 원리를 분석하고 시각화하는 과정을 단계별로 살펴본다.

책의 구성은 크게 두 부분으로 나뉜다. ‹제1~4장›은 한의학 처방 데이터를 네트워크 형태로 어떻게 구축할 것인가를 다룬다. ‹제5~9장›은 구축된 네트워크를 어떻게 분석할 것인가를 다룬다. 구체적으로 ‹제1~2장›에서는 한의학 처방 데이터를 그래프 구조로 모델링하고 정규화하는 과정을 다룬다. 이어 ‹제3~4장›에서는 모델링에 앞서 데이터의 특성을 탐색적 데이터 분석(EDA)으로 확인하고 엣지 가중치 산출 방법을 고찰한다. 마지막으로 ‹제5~9장›에서는 본초 네트워크의 구조적 특성을 다각도로 분석한다. 무작위 모델 기반의 좁은 세상 구조와 k-core 계층 구조, 중심성·강건성 시뮬레이션, 다차원 레이아웃 시각화, 커뮤니티 탐지와 본초 다의성 분석, 개별 처방 서브네트워크 분석을 차례로 살펴본다.

i 분석에 사용한 데이터에 관하여

이 책에 사용한 처방 데이터는 ‘ETCM 2.0 (Encyclopedia of Traditional Chinese Medicine)’ (Zhang et al. 2023) 데이터베이스를 바탕으로 구축되었다. ETCM 2.0은 동아시아 전통의약 데이터를 집대성한 통합 데이터베이스로 공식 웹사이트(URL: www.tcmip.cn/ETCM2/front)에서 열람할 수 있다. ETCM 2.0 데이터베이스를 구축하고 공개해 준 연구진에 깊은 감사를 전한다.
이 책에서 분석에 활용한 데이터는 다음 과정을 거쳐 가공했다.
우선 ETCM 2.0 데이터베이스 항목 가운데 ‘中醫古籍方劑’에서 ’處方組成’을 추출하였다. 이 데이터는 한의학 고문헌에 수록된 처방과 각 처방의 본초 구성을 담고 있다(이하 ’원시 데이터’).
원시 데이터를 바탕으로 다음과 같이 정제 작업을 진행하였다(이하 ‘정제 데이터’).

- 오류 수정: 원시 데이터에서 확인된 명백한 오류를 바로잡았다. 다만 미처 발견하지 못한 오류가 남아 있을 수 있다.
- 번체자 변환: 간체자(簡體字)로 작성된 원시 데이터를 번체자(繁體字)로 변환하였다. 예시: 人参 → 人蔘, 干姜 → 乾薑, 当归 → 當歸
- 본초명 표준화: 동의어(이명)·유의어를 정리하여 통합하였다.[1] 예시: 山栀·山栀子 → 栀子, 香附米 → 香附, 赤芍·白芍·赤芍藥·白芍藥 → 芍藥
- 한글 독음 추가: 각 본초명에 한글 독음을 병기하였다. 예시: 黃芪 → 황기(黃芪), 甘草 → 감초(甘草)

[1] 이와 관련하여 2장 내용을 참고하라.

제 1 장

네트워크 분석과 한의학

1.1 한의학 처방

처방은 한의학에서 치료 목적에 맞게 몇 가지 본초를 조합하여 만든 것이다. 질병을 치료하는 한의학 이론은 본초의 조합 방식으로 표현되며 본초의 배합 자체가 곧 의학 이론을 대변한다. 이동원(李東垣)이 중기(中氣)를 끌어올리고자 승마(升麻)와 시호(柴胡)를 쓴 것이나 주단계(朱丹溪)가 보음(補陰)을 위해 숙지황(熟地黃), 지모(知母), 황백(黃柏) 등을 배합한 것이 대표적인 예다.

한의학 처방의 효능은 개별 본초의 기미(氣味)와 약성뿐만 아니라 본초가 함께 어우러져 나타내는 상승효과까지 아우른다. 한의학에서 처방을 2~3가지 본초 조합인 '약대(藥對)'로 나누어 풀이하거나 본초의 역할을 정치 체계에 비유해 군신좌사(君臣佐使)로 해설한 것은 이러한 이유 때문이다. 예를 들어 아래 보중익기탕(補中益氣湯) 해설을 살펴보자. 황기(黃芪)·인삼(人蔘)·자감초(炙甘草) 조합으로 습열(濕熱)과 번열(煩熱)을 치료하고, 승마(升麻)·시호(柴胡) 조합으로 위(胃)의 청기(淸氣)를 올리며, 황기(黃芪)·감초(甘草)로 위기(衛氣)를 보한다고 풀이하였다(한국한의학연구원 2024). 이 설명은 처방이 단순히 본초를 물리적으로 합쳐 놓은 결과가 아니며 본초 배합에 따라

치료 효과가 달라질 수 있음을 보여준다.

> 비위(脾胃)가 허해 폐기(肺氣)가 끊기면 황기(黃芪)로 피모의 기를 더해 주리(腠理)를 막아 자한(自汗)을 막는다. 숨이 차고 짧으면 원기(元氣)를 상하므로 인삼(人蔘)으로 보한다. 심화(心火)가 비(脾)를 누르면 자감초(炙甘草)의 달고 따뜻한 성질로 화열(火熱)을 끄고 위(胃)의 원기를 보한다. 비위가 갑자기 아프고 뱃속이 당길 때 많이 쓴다. 이 세 약은 습열(濕熱)과 번열(煩熱)을 없애는 성약(聖藥)이다. 백출(白朮)은 쓰고 달고 따뜻해 위(胃) 속의 열을 없애고 혈을 잘 통하게 한다. 승마(升麻)·시호(柴胡)는 쓰고 평(平)하며 담박해 위(胃) 속의 청기(淸氣)를 올린다. 또한 황기·감초를 이끌어 상승하여 흩어진 위기(衛氣)를 보하고 겉을 실하게 하며 대맥(帶脈)의 당김을 푼다. 당귀(當歸)는 혈맥을 고르게 한다. 귤홍(橘紅)은 가슴속 기를 다스리고 양기 상승을 도와 체기(滯氣)를 흩어준다. 이것이 처방을 세운 참뜻이다.

한의학 처방 속 본초는 함께 사용되는 본초에 따라 역할과 효능이 달라지므로 개별 본초의 기미(氣味)·귀경(歸經)·효능과 같은 단편적인 규칙만으로는 온전히 설명하기 어렵다. 그렇다면 한의학 처방과 본초는 어떻게 파악할 수 있을까? 개별 대상을 독립적으로 관찰하기보다는 관계 속에서 대상을 종합적으로 이해할 필요가 있다. 이처럼 관계 속에서 대상을 파악하고자 고안된 대표적인 방법론이 네트워크 분석(Network Analysis)이다.

이제 네트워크 분석을 한의학 처방 데이터에 적용하여 관계와 맥락 속에서 처방과 본초를 이해해 보자.

한의학 처방의 수학적 표현

한의학 처방은 전체 본초 집합에서 특정 치료 목적에 따라 선택된 유의미한 부분집합이라고 수학적으로 정의할 수 있다.

$$\mathcal{H} = \{h_1, h_2, \dots, h_M\}, \quad F_k = \{h_{k,1}, h_{k,2}, \dots, h_{k,m_k}\} \subset \mathcal{H}$$

($\mathcal{H}$: 전체 본초(전체집합). M: 전체 본초 종류 수. F_k: k번째 처방. m_k: 해당 처방 F_k에 포함된 본초 수)

1.2 관계의 표현

네트워크 분석은 대상 자체보다 '관계'라는 맥락에 주목한다. 사회 연결망, 인터넷의 라우터망, 단백질 상호작용망처럼 복잡한 시스템(복잡계)은 점과 선이라는 모델로 단순하게 표현할 수 있다. 이렇게 복잡계를 점과 선으로 모델링해 의미를 탐색하는 방법을 네트워크 분석이라 한다(Newman 2018). 한의학 처방 역시 네트워크 구조로 치환하면 매우 직관적인 모델이 된다. 하나의 처방이 본초의 '집합'이라면 한의학 처방 전체는 이러한 집합이 모인 거대한 네트워크가 되기 때문이다.

분석에 앞서 데이터를 네트워크로 모델링하는 기본 개념을 정리해 보자. 네트워크는 '점'인 노드(Node)와 '선'인 엣지(Edge)로 이루어진다.

◎ **노드(Node)**

먼저 노드(Node)는 네트워크를 이루는 '점'으로 독립된 개별 구성 개체(Entity)에 해당한다. 한의학 처방에서 마황(麻黃)·계지(桂枝)·감초(甘草) 같은 개별 '본초'가 각각 하나의 고유한 노드가 된다. 네트워크 안의 노드가 동질적인지에 따라 단일 모드 네트워크(One-mode network)와 다중 모드 네트워크(Multimodal network)로 나뉜다. 본초만으로 노드를 구성하면 단일 모드 네트워크가 되고, 본초와 병증을 함께 노드로 구성하면 다중 모드 네트워크가 된다.

◎ **엣지(Edge)**

다음으로 엣지(Edge)는 노드와 노드 사이의 '관계'로 상호작용(Interaction)이나 연결 상태에 해당한다. 점과 점 사이의 상호작용을 '관계'로 추상화한 것이다. 한의학 처방에서 같은 처방

에 함께 쓰인 본초는 서로 연관되므로 엣지로 연결할 수 있다. 관계에 방향이 있다면 엣지에도 방향(Direction)을 부여하며 이러한 연결로 이루어진 네트워크를 방향성 네트워크(Directed network)라고 한다. 소셜 네트워크에서 다른 계정을 팔로우하거나 논문이 다른 논문을 인용하는 관계가 대표적인 예다. 반면 처방 내 본초의 동시 출현처럼 방향성이 없는 상호 대칭 관계는 무방향 엣지(Undirected edge)로 표현하며 모든 엣지에 방향성이 없는 네트워크를 무방향 네트워크(Undirected network)라고 부른다.

여기에 관계의 강도를 나타내는 가중치(Weight)를 더할 수 있다. 예컨대 두 본초가 여러 처방에서 자주 함께 쓰였다면 엣지 가중치를 높게 부여한다. 가중치를 반영하면 처방 내 본초 간 배합 관계를 훨씬 정밀하게 나타낼 수 있다. 노드 하나에 직접 연결된 엣지의 개수를 차수(Degree)라고 부른다. 차수는 한 노드가 네트워크 내에서 얼마나 많은 이웃과 직접 관계를 맺고 있는지를 나타내는 가장 기초적인 척도다. 엣지에 가중치가 부여된 네트워크라면 연결된 엣지 개수뿐만 아니라 가중치 총합을 함께 고려한다. 이를 노드 강도(Strength) 또는 가중 차수(Weighted degree)라고 부른다.

앞서 노드를 '본초'로, 엣지를 '본초 배합'으로 설명했지만 노드와 엣지의 정의는 고정되어 있지 않다. 네트워크 모델은 분석 목적에 따라 달라질 수 있다. 한의학 처방을 예로 들어보자. 본초 사이의 관계를 중심으로 처방을 이해하고자 한다면 본초를 노드로 삼고 본초 간 배합 관계를 엣지로 두는 단일 모드 무방향 네트워크가 가장 직관적이다. 이때 두 본초가 처방에 함께 쓰인 빈도를 엣지 가중치로 설정할 수 있다('예시 모델 1'). 한편 본초와 병증 사이의 관계를 파악하고 싶다면 병증을 노드로 추가하여 다중 모드 네트워크를 만들고 필요에 따라 치료 방향성이나 가중치를 설정할 수 있다. 또한 본초가 아니라 처방 사이의 관계를 규명하고자 할 때는 전혀 다른 방식으로 모델링해야 한다. 처방을 노드로 정의하고 '두 처방 간 공통 본초의 존재 여부'를 엣지로 연결하며 '공통 본초의 개수'를 가중치로 설정할 수 있다('예시 모델 2'). 이처럼 같은 데이터라도 분석 목적에 따라 다양하게 모델링할 수 있다.

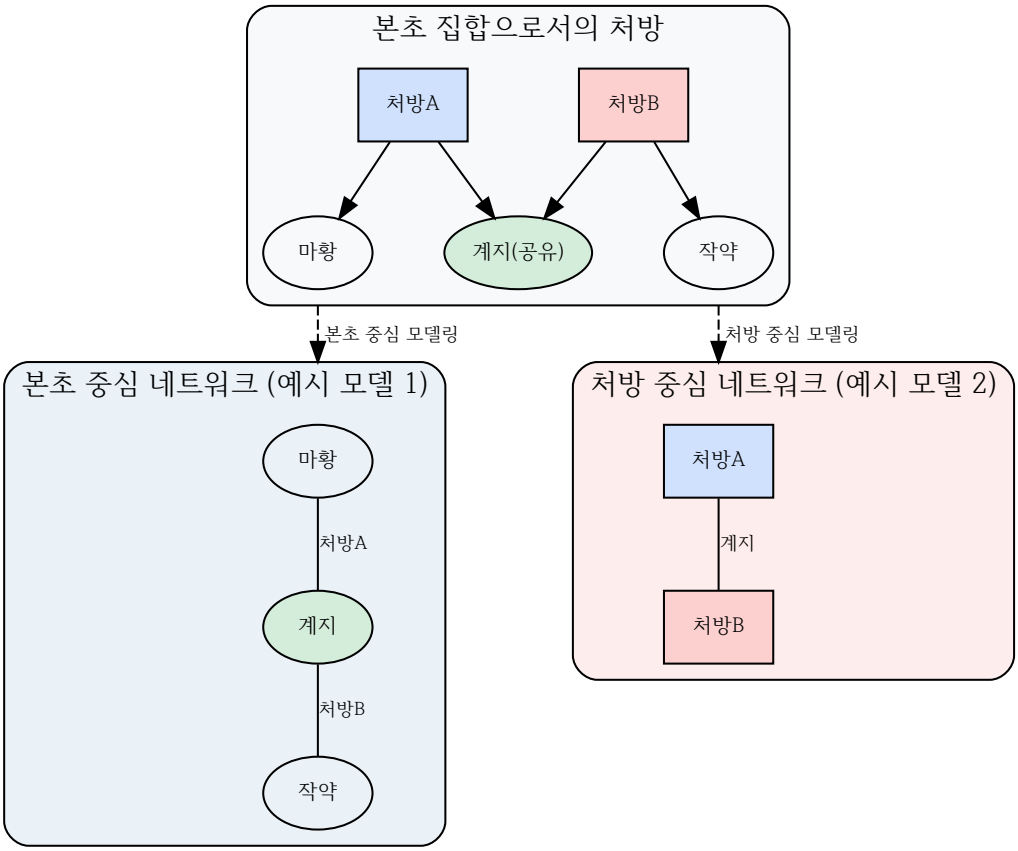

Figure 1.1 [개념도] 분석 목적에 따른 네트워크 모델링 예시

네트워크, 그래프, 위상 구조

'네트워크(Network)'는 실제 세계에 존재하는 대상과 대상 사이의 상호작용을 점과 선으로 나타낸 복잡계 시스템을 가리킨다. 수학에서는 이러한 대상을 추상적인 수리 모델 $G = (V, E)$로 다루며 이를 '그래프(Graph)'라고 부른다. 여기서 V(Vertex)는 노드 집합, E(Edge)는 엣지 집합이다. 네트워크가 현실 데이터와 현상의 맥락을 포함한다면 그래프는 수학적 구조와 연산 대상에 초점을 둔다. 한편 '위상 구조(Topology)'는 노드가 서로 어떻게 연결되어 있는지를 나타내는 구조적 패턴을 뜻한다. 이는 노드의 구체적인 기하학적 좌표나 시각적 배치 위치를 뜻하지 않는다. 따라서 노드의 시각적 배치가 바뀌더라도 노드 간 연결 관계가 유지된다면 네트워크의 위상 구조는 변하지 않는다.

제 2 장

네트워크 모델링

현실의 사건이나 현상을 수학적으로 분석하려면 먼저 대상을 계산 가능한 수학적 형태로 표현해야 한다. 이 과정을 '모델링'이라 하고, 그 결과물을 '모델'이라 부른다. 모델링은 현실의 모든 정보를 온전히 담아낼 수 없다. 문제를 추상화(Abstraction)하여 수학적 모델로 표현하는 과정에서 정보의 손실과 단순화가 필연적으로 발생한다. 이 장에서는 처방 데이터를 네트워크로 모델링하는 과정과 그 과정에서 마주하는 현실적인 난관을 살펴본다.

2.1 데이터 정규화

데이터를 모델링하려면 데이터를 먼저 정돈해야 한다. 이를 데이터 정규화(Normalization)라고 한다. 한의학 처방을 수학적으로 분석하려면 한의학 고전에서 처방을 추출하고, 이를 데이터로 정돈한 뒤 모델링 방식을 결정하여 모델을 구축해야 한다.

한의학 고전에서 처방을 추출하는 일부터 쉽지 않다. 옛 의서를 실물로 구하고 글자를 판독해 디지털로 입력하며 비정형 자연어를 구조화된 형태로 정리하는 복잡한 절차를 거쳐야 한다. 모두 중요한 과정이지만 이 책의 범위를 벗어나므로 다루지 않는다. 여기서는 한의학 고전 속 처방을

디지털화하여 구조화된 데이터로 정리한 뒤의 과정을 살펴본다.

네트워크 모델링에서는 모델의 바탕이 되는 노드와 엣지를 규정하는 작업이 매우 중요하다. 같은 의미를 지닌 노드를 여럿으로 분리하면 핵심 연결이 여러 노드로 분산된다. 반대로 다른 의미의 노드를 하나로 합치면 마땅히 구분해야 할 노드가 뭉치게 된다. 두 경우 모두 네트워크 모델을 왜곡한다. 따라서 데이터를 올바르게 정규화하여 노드와 엣지를 확정해야 한다. 데이터 정규화 관점에서 한의학 처방 데이터는 크게 세 가지 문제를 안고 있다.

◎ **이명(Synonym)**

같은 본초라도 문헌이나 저자에 따라 이름이 다르게 기록되곤 한다. 예를 들어 '감초(甘草)'는 '국로(國老)', '검실(芡實)'은 '계두(鷄頭)'로 불린다. 한자 표기 차이도 넓게 보면 같은 문제다. 나복(蘿蔔)은 '蘿菖'·'蘿蔔'·'萊菔' 등으로 쓰이고, 인삼(人蔘) 역시 '蔘'·'參'·'人參'·'仁蔘'·'仁參'처럼 표기가 다양하다. 전문적인 판단이 필요한 때도 있다. 궁궁(芎藭)·천궁(川芎)·호궁(胡芎)처럼 명칭의 시대적 변천이나 산지에 따른 이명을 검토해야 하는 사례도 있고, 백작약(白芍藥)·적작약(赤芍藥)처럼 본초의 기원 식물이나 한의학 이론을 고려해야 할 때도 있다. 상지(桑枝)·상엽(桑葉)·상백피(桑白皮)처럼 기원 식물이 같더라도 약용 부위에 따라 별개 본초로 취급하는 경우도 있다.

◎ **수치(修治)와 포제(炮製)**

같은 본초라도 찌거나 굽고 술에 담그는 가공 방식에 따라 약성이 달라진다. '생지황(生地黃)'과 '숙지황(熟地黃)'이 대표적인 예다. 이 둘은 가공 방식에 따라 약성과 주치가 크게 달라져 고전 의서에서도 구분하여 쓰였다. 그러므로 이 둘을 '지황(地黃)'이라는 단일 노드로 합치면 처방의 본래 의도가 왜곡될 수 있다. 그렇다고 모든 본초의 포제를 분리하기도 쉽지 않다. 가령 감초(甘草)를 자감초(炙甘草)와 생감초(生甘草)로 구분하려 하면 원문에 단순히 '감초'로만 적힌 본초는

어떻게 분류할지 모호해지기 때문이다.

◎ **용량(Dose)**

한의학 처방에서 용량 변화는 처방의 성격 자체를 바꾸기도 한다. 예를 들어 계지탕(桂枝湯)에서 계지(桂枝)의 용량을 늘리면 계지가계탕(桂枝加桂湯)이 되어 처방의 적응증이 달라진다. 네트워크에 용량 정보를 온전히 반영하면 이상적이겠지만 현실 데이터는 이를 뒷받침하지 못한다. 시대마다 도량형(돈錢, 냥兩, 푼分 등) 기준이 다르고 문헌마다 용량 기록이 빠진 사례도 흔하다. 또한 '몇 개'나 '한 줌'처럼 본초의 형태나 개수로 표기되어 실제 무게를 가늠하기 어려운 경우도 많다. 이런 맥락을 고려하지 않고 현대 도량형 단위로 단순 환산하면 심각한 노이즈가 발생한다.

1회 복용량도 처방마다 차이가 크다. 대부분의 문헌은 1회 조제 분량을 기준으로 용량을 적어 두었지만 한 번 조제한 약을 복용하는 기간은 처방마다 제각각이다. 하루 한 번 달여 여러 번 나누어 마시는 처방이 있는가 하면, 하루에 두세 차례 새로 달여 복용하는 처방도 있다. 환제나 산제처럼 한 번 만들어 며칠 동안 복용하는 약도 있다. 따라서 고문헌에 기재된 용량을 그대로 쓰기는 어려우며 정밀한 분석을 위해서는 1회 복용량을 기준으로 정규화해야 한다.

데이터 정규화 문제에는 명쾌한 단 하나의 해법이 없다. 이 책에서는 데이터의 현실적 한계를 인정하고 불확실하거나 편차가 심한 정보를 모델에서 과감히 배제하는 방식을 택했다. 먼저 수치와 포제 정보는 문헌 간 편차와 신뢰도 문제를 고려해 원칙적으로 일반화하여 통합했다. 다만 생지황(生地黃)과 숙지황(熟地黃)처럼 가공에 따른 약성과 주치의 차이가 현격하여 임상적 구분이 확립된 핵심 본초는 독립된 노드로 유지했다. 용량 정보 역시 객관적인 정규화가 어렵다고 보아 배제했다. 그 대신 본초가 처방에 포함되어 있는지만을 '존재함(1)'과 '존재하지 않음(0)'으로 구분하는 이진(Binary) 데이터를 사용했다.

다만 이명과 표기 문제는 분석 결과에 직접적인 영향을 미치므로 최대한 정돈했다. 이명이나 이

체자를 지나치게 넓게 통합하면 데이터 차원이 줄어 분석은 편해지지만 고전 문헌에 담긴 고유한 처방 구성이나 본초의 세부 특성이 사라질 수 있다. 반대로 기준을 너무 엄격하게 적용해 통합을 꺼리면 같은 개념이 여러 노드로 분산되어 네트워크 밀도와 중심성이 왜곡된다. 따라서 정밀한 유의어 사전을 구축해 체계적으로 정리해야 한다. 이 책에서는 처방 데이터에서 유의어 사전을 구축한 뒤 이를 바탕으로 유의어를 단계적으로 병합하는 방식을 적용했다(오준호 2022).

데이터 분석에서 흔히 깔끔하고 선명한 결과를 기대하지만 이는 잘 정제된 이상적인 소규모 데이터에서나 가능한 일이다. 현실 데이터는 수량과 품질 모두에 한계가 있어 적절한 타협이 불가피하다. 따라서 불완전한 데이터의 한계를 직시하면서도 그 안에서 가장 신뢰할 만한 결과를 도출하는 방법도 함께 고민해야 한다.

2.2 네트워크 표현

정규화된 데이터를 바탕으로 네트워크를 구성해 보자. 네트워크를 정의하려면 기본 단위인 노드(node)와 이들을 잇는 엣지(edge)가 필요하다. 이 책에서는 한의학 처방 데이터를 바탕으로 만든 본초 네트워크를 살펴본다. 특별한 언급이 없다면 노드는 '본초', 엣지는 '본초 사이의 관계', 가중치는 '관계의 강도'를 뜻한다. 엣지 가중치(edge weight)는 다양하게 정의할 수 있지만, 이 장에서는 두 본초가 한 처방에 함께 등장한 횟수, 즉 '동시 출현(Co-occurrence) 빈도'로 한다. 가중치 산출 방식의 세부 특성과 대안은 뒤에서 별도로 다룬다.

네트워크를 표현하는 방식은 크게 '엣지 리스트'와 '인접 행렬'로 나뉜다.

◎ **엣지 리스트(Edge List)** 형식

노드 쌍과 그 사이의 엣지 가중치를 목록(list) 형태로 나열하는 방식이다. 방향성이 있는 네트워크라면 (`Source-Node: A, Target-Node: B, Edge Weight`)과 (`Source-Node: B, Target-Node:`

`A, Edge Weight`)는 서로 다른 연결을 뜻한다. 반면 엣지의 방향이 없는 무방향 네트워크에서는 두 데이터가 동일한 의미를 갖는다. 따라서 중복을 피하기 위해 둘 중 하나만 기록한다.

Table 2.1 엣지 리스트(Edge List) 기반 네트워크 표현 예시 (가중치 상위 10개 엣지)

No	Node (Source)	Node (Target)	Edge Weight (Co-occurrence)
1	감초(甘草)	인삼(人蔘)	5632
2	감초(甘草)	복령(茯苓)	5333
3	복령(茯苓)	인삼(人蔘)	4909
4	감초(甘草)	당귀(當歸)	4442
5	당귀(當歸)	작약(芍藥)	4433
6	감초(甘草)	작약(芍藥)	4334
7	당귀(當歸)	천궁(川芎)	3759
8	감초(甘草)	백출(白朮)	3693
9	감초(甘草)	진피(陳皮)	3613
10	인삼(人蔘)	백출(白朮)	3490

◎ **인접 행렬(Adjacency Matrix)** 형식

노드 수에 맞춰 행(row)과 열(column)을 구성하고, 교차 지점에 두 노드 사이의 엣지 가중치를 채워 넣는 방식이다.

Table 2.2 인접 행렬(Adjacency Matrix) 기반 네트워크 표현 예시 (빈도 상위 10개 노드)

	감초 (甘草)	복령 (茯苓)	당귀 (當歸)	인삼 (人蔘)	작약 (芍藥)	백출 (白朮)	천궁 (川芎)	진피 (陳皮)	황금 (黃芩)	방풍 (防風)
감초 (甘草)	-	5333	4442	5632	4334	3693	3213	3613	3236	2771
복령 (茯苓)	5333	-	2682	4909	2434	3459	1707	2619	1322	1517
당귀 (當歸)	4442	2682	-	3068	4433	2290	3759	1602	1468	1748
인삼 (人蔘)	5632	4909	3068	-	2278	3490	1878	2289	1196	1485
작약 (芍藥)	4334	2434	4433	2278	-	1757	2596	1250	1850	1362
백출 (白朮)	3693	3459	2290	3490	1757	-	1438	2251	730	914

	감초 (甘草)	복령 (茯苓)	당귀 (當歸)	인삼 (人蔘)	작약 (芍藥)	백출 (白朮)	천궁 (川芎)	진피 (陳皮)	황금 (黃芩)	방풍 (防風)
천궁 (川芎)	3213	1707	3759	1878	2596	1438	-	1047	990	1935
진피 (陳皮)	3613	2619	1602	2289	1250	2251	1047	-	655	456
황금 (黃芩)	3236	1322	1468	1196	1850	730	990	655	-	1057
방풍 (防風)	2771	1517	1748	1485	1362	914	1935	456	1057	-

인접 행렬은 수학적으로 간단히 계산할 수 있다. 먼저 처방을 행으로 두고 본초를 열로 배치하여 포함 여부를 1과 0으로 채운 '처방-본초 소속 행렬(B)'을 구성한다. 이어서 행렬 B의 전치 행렬인 B^T에 원 행렬 B를 곱하면 본초 네트워크의 인접 행렬이 곧바로 도출된다. 만약 B의 크기가 (처방 수 × 본초 수)라면, B^T의 크기는 (본초 수 × 처방 수)가 된다. 따라서 두 행렬의 곱 $\mathrm{B}^T\mathrm{B}$는 (본초 수 × 본초 수) 크기의 정방 행렬을 이룬다. 이 정방 행렬의 (i, j) 성분은 두 본초의 출현 벡터 간 내적(Dot Product)인 $\sum_k B_{ki}B_{kj}$에 해당하며, 두 본초가 동시에 쓰인 처방의 총 횟수를 뜻한다. 이처럼 행렬 곱셈 연산으로 모든 본초 쌍의 동시 출현 빈도를 한 번에 일괄 계산할 수 있다.

$$\mathrm{A} = \mathrm{B}^T\mathrm{B}$$

*(*B*: 처방-본초 이진 소속 행렬.* B^T*:* B*의 전치 행렬.* A*: 본초 동시 출현 빈도를 담은 인접 행렬)*

이때 행렬의 대각선 성분(마황-마황, 감초-감초 등)은 각 본초가 전체 처방에 등장한 '총 출현 빈도'를 나타낸다. 네트워크 분석에서 자기 자신으로 돌아오는 연결(Self-loop)은 불필요한 노이즈가

되므로 대각선 값을 0으로 변환해 사용한다. 또한 엣지에 방향성이 없을 때 인접 행렬은 주대각선을 기준으로 상하가 대칭을 이룬다. 분석 목적에 따라 행렬 전체를 그대로 활용하기도 하며, 메모리 효율을 높이기 위해 대칭성을 활용해 대각선 기준 한쪽 절반(상삼각행렬 또는 하삼각행렬)만 취하기도 한다.

대부분의 프로그래밍 언어에서 행렬 연산은 벡터 연산으로 고속 처리되므로 네트워크 분석에서도 인접 행렬 구조를 널리 채택한다.

처방 데이터에서 인접 행렬을 구축하고 네트워크로 변환하는 과정을 정리하면 다음과 같다.

엣지 리스트와 인접 행렬은 각각 뚜렷한 장단점을 지닌다. 인접 행렬은 모든 노드 간의 관계를 한눈에 조망할 수 있어 직관적이지만, 연결이 없는 '0'인 위치까지 모두 저장하므로 메모리 낭비가 심하다. 반면 엣지 리스트는 실제 연결이 존재하는 데이터만 기록하므로 저장 공간을 크게 아낄 수 있으나, 전체 네트워크 구조를 한눈에 파악하기는 어렵다.

i 네트워크 그래프의 수학적 표현

수학적으로 가중 네트워크(Weighted Network)는 노드 집합, 엣지 집합, 그리고 엣지의 연결 강도를 나타내는 가중치 함수로 정의된다. 노드 간의 전체 연결 관계는 정방 행렬인 인접 행렬로 나타낸다.

$$G = (V, E, W), \quad \mathbf{A} = (A_{ij})_{N \times N}$$

($V = \{v_1, \dots, v_N\}$: 노드 집합. $E \subseteq \binom{V}{2}$: 무방향 엣지 집합. $W : E \to \mathbb{R}^+$: 엣지 가중치 함수. $\mathbf{A}$: 노드 v_i, v_j 간 연결 가중치 A_{ij}를 담은 인접 행렬)

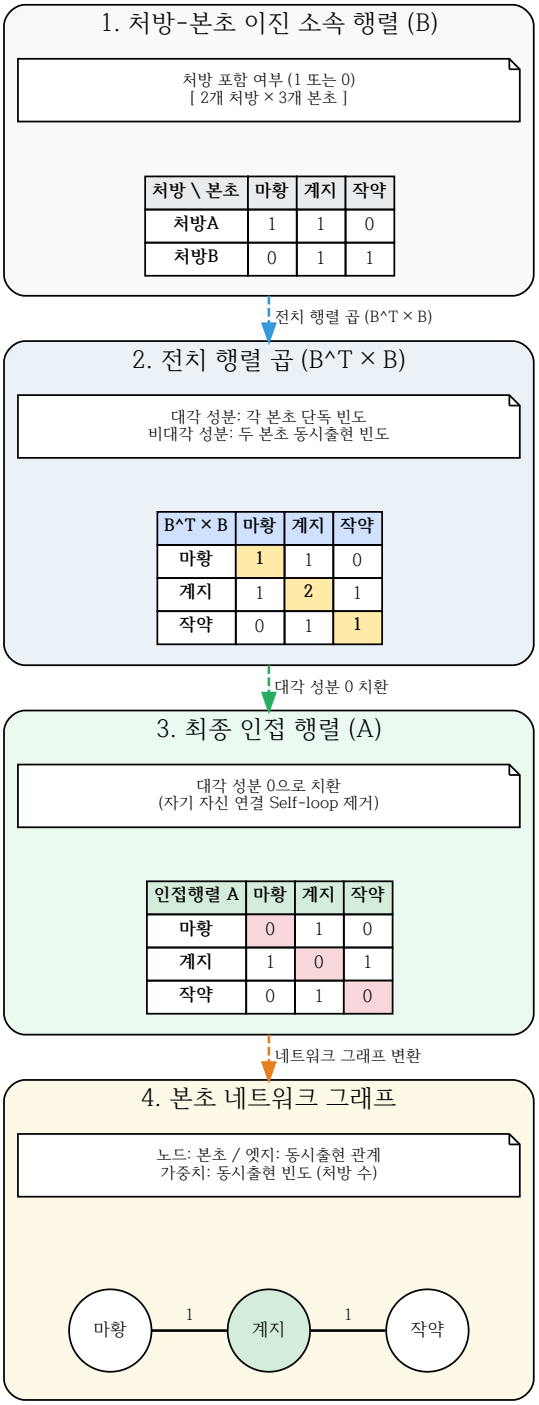

Figure 2.1 [개념도] 한의학 처방 데이터를 인접 행렬 기반 네트워크로 표현하는 과정

제 3 장

탐색적 데이터 분석

데이터는 저마다 고유한 특성을 띤다. 이 특성을 이해하지 못한 채 분석 알고리즘을 적용하면 원하는 결과를 얻기 어렵다. 설령 결과를 얻더라도 잘못된 결론에 이를 수 있다. 본격적인 분석에 앞서 기초 통계를 점검해 데이터의 형태를 파악하고 특징을 미리 살피면 어떤 방식으로 모델링하고 어떤 알고리즘을 적용할지 단서를 얻을 수 있다. **탐색적 데이터 분석(Exploratory Data Analysis, EDA)**은 이를 위해 빈도와 분포 등 데이터의 기초 통계를 확인하는 과정이다.

3.1 처방의 크기

Table 3.1 한의학 처방 데이터 기초 통계

지표	수치
총 처방 수 (N of Formulas)	48,442
고유 본초 수 (N of Unique Herbs)	3,196
이론적 최대 연결 수 (Potential Connections)	5,105,610
실제 관찰 연결 수 (Actual Connections)	106,145
네트워크 밀도 (Density)	2.08%
처방 당 본초 수 평균 (Mean Formula Size)	7.4
처방 당 본초 수 중앙값 (Median Formula Size)	7

먼저 분석에 활용할 전체 데이터셋을 살펴보자. 이 책에서 다루는 처방 데이터는 총 48,442개이며 3,196종의 본초가 조합되어 처방을 이룬다. 3,196종의 본초가 둘씩 짝을 지어 만들 수 있는 이론적 조합은 5,105,610개에 달한다. 하지만 실제 데이터에서 관찰된 본초 간 연결(엣지)은 106,145개로 이론적 최대치의 2.08%(네트워크 밀도)에 불과하다. 48,442개 처방에서 생성된 동시 출현 엣지의 총 토큰 수가 1,651,234개라는 점을 감안하더라도 적은 수치다. 이는 한의학 처방이 본초의 무작위 결합이 아니라 의도된 배합 원리에 따라 구성되었음을 보여준다.

다음으로 처방 하나를 구성하는 본초의 개수, 즉 '처방 크기(Formula Size)'를 살펴보자. 데이터셋의 처방 48,442개를 분석한 결과 하나의 처방은 평균 7.4개의 본초로 이루어졌으며 중앙값은 7개였다. 처방 크기를 히스토그램으로 나타내면 최빈값을 중심으로 양 극단으로 갈수록 빈도가 줄어드는 형태를 띤다. 처방 대다수는 대략 4개에서 10개 사이에 집중되어 있다. 다만 데이터 전체 범위는 최솟값 1개부터 최댓값 91개까지 비교적 넓게 분포한다. 평균과 중앙값을 고려하면 오른쪽으로 꼬리가 긴 비대칭(Right-skewness) 형태다.

처방에 들어가는 본초가 너무 적으면 복합적인 증상을 다스리기 어렵고 너무 많으면 본초 고유의 기미(氣味)가 상쇄되거나 예기치 않은 부작용이 생길 수 있다. 따라서 '평균 7.4개', '중앙값 7개'라는 수치는 오랜 임상 경험을 거치며 효능과 안전성의 균형을 맞춘 결과로 해석할 수 있다.

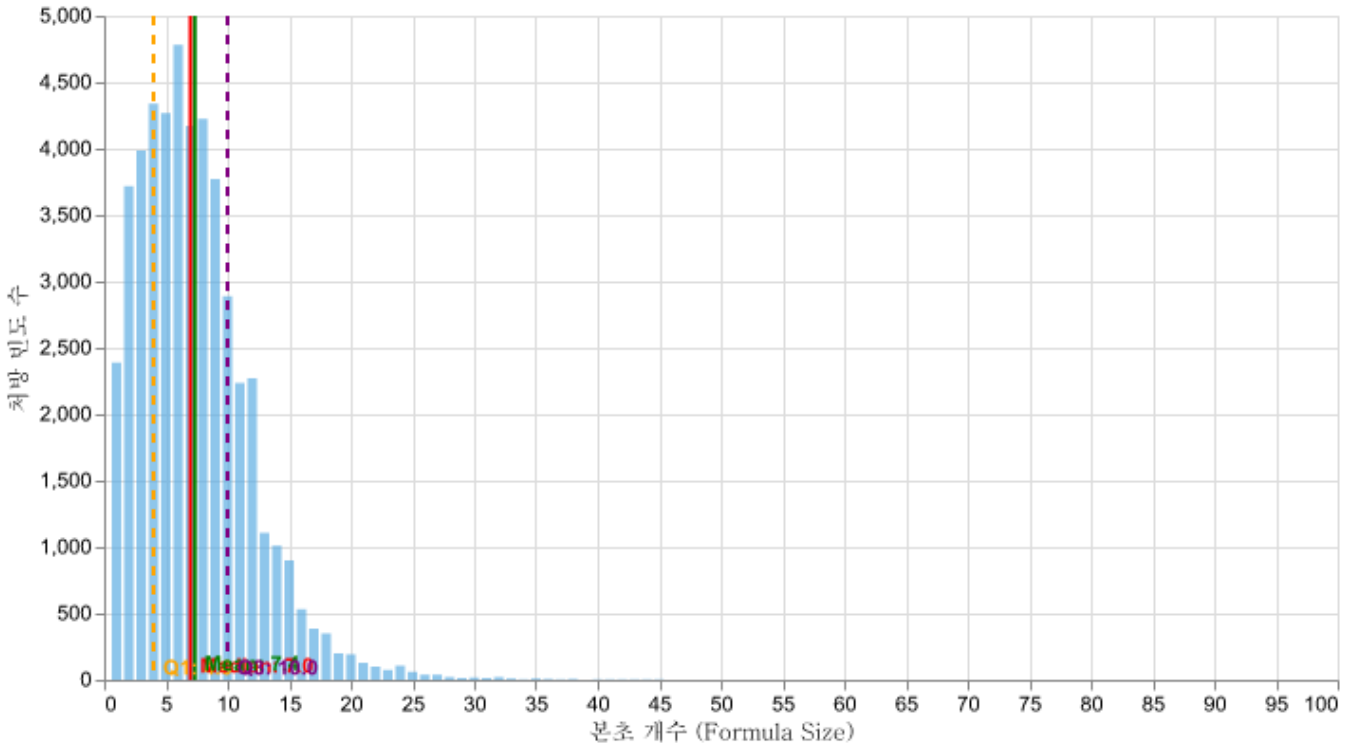

Figure 3.1 처방 크기(Formula Size) 히스토그램

3.2 본초 출현 빈도

이제 개별 본초의 출현 빈도를 살펴보자. 본초 출현 빈도는 해당 본초가 전체 처방에서 몇 번 등장했는지를 나타낸다. 처방 하나에 같은 본초가 중복되어 들어가지 않으므로 출현 빈도는 곧 해당 본초를 포함하는 처방 수와 일치한다. 가장 많이 쓰인 본초는 '감초(甘草)'다. 전체 처방의 34.2%에 해당하는 16,581개 처방에 포함되었다. 복령(茯苓)(9,940회), 당귀(當歸)(9,922회), 인삼(人蔘)(9,582회), 작약(芍藥)(7,869회), 백출(白朮)(6,598회), 천궁(川芎)(6,269회), 진피(陳皮)(6,265회), 황금(黃芩)(5,302회), 방풍(防風)(5,171회) 등 익숙한 본초가 그 뒤를 이었다.

Table 3.2 출현 빈도 상위 본초 (상위 20개)

순위	본초명	출현 빈도(회)	비중(전체 처방 대비, %)	비중(전체 본초 사용 빈도 대비, %)
1	감초(甘草)	16,581	34.2	4.7
2	복령(茯苓)	9,940	20.5	2.8
3	당귀(當歸)	9,922	20.5	2.8
4	인삼(人蔘)	9,582	19.8	2.7
5	작약(芍藥)	7,869	16.2	2.2
6	백출(白朮)	6,598	13.6	1.9
7	천궁(川芎)	6,269	12.9	1.8
8	진피(陳皮)	6,265	12.9	1.8
9	황금(黃芩)	5,302	10.9	1.5
10	방풍(防風)	5,171	10.7	1.5
11	육계(肉桂)	4,972	10.3	1.4
12	목향(木香)	4,737	9.8	1.3
13	대황(大黃)	4,551	9.4	1.3
14	반하(半夏)	4,475	9.2	1.3
15	건강(乾薑)	4,413	9.1	1.2
16	황련(黃連)	4,358	9.0	1.2
17	부자(附子)	4,295	8.9	1.2
18	생지황(生地黃)	4,086	8.4	1.1
19	맥문동(麥門冬)	3,813	7.9	1.1
20	황기(黃芪)	3,729	7.7	1.0

자주 쓰인 본초가 있는 반면 데이터 전체에서 몇 개 처방에만 쓰인 본초도 많다. 전체 본초 3,196종 가운데 출현 빈도가 5회 이하인 본초는 2,316종에 달한다. 전체 본초의 72.5%가 특정 문헌이나

특수한 질환에만 제한적으로 쓰였다는 의미다.

> **희소 본초 기준**
>
> 드물게 등장하는 본초를 '희소 본초(Rare Herbs)'라고 부르자. 희소 본초는 단순한 입력 오류나 노이즈일 수도 있지만 특정 희귀 질환에 쓰이는 중요한 본초일 수도 있다. 따라서 무분별하게 삭제하면 귀중한 임상 정보가 손실될 위험이 있으므로 정제 기준을 신중히 세워야 한다. 여기서는 희소 본초의 분포 규모를 탐색적으로 살펴보고자 '5회'라는 대략적인 기준을 적용했다. 제5장 이후 이어지는 네트워크 분석에서는 뒤에서 설명할 '누적 빈도 커버리지' 방식으로 도출한 기준을 사용한다.

개별 본초의 출현 빈도를 그래프로 나타내면 한쪽 끝이 매우 높고 반대쪽은 긴 꼬리처럼 늘어지는 형태를 보인다. 이는 소수 본초가 대다수 처방에 두루 쓰인 반면 대다수 본초는 소수 처방에만 쓰였음을 뜻한다. 평균 주변에 데이터가 집중되었던 처방 크기 분포와는 뚜렷하게 대비된다. 이러한 비대칭 분포 양상을 통계학에서는 '헤비테일(Heavy-tailed)' 또는 '두터운 꼬리 분포'라고 부른다.[1]

[1] 실습 데이터의 본초 출현 빈도는 우측 꼬리가 두터운 비대칭 헤비테일 분포에 해당한다. 통계학적으로 멱법칙 적합성을 엄밀하게 검증할 때는 최우추정법(MLE)으로 파라미터를 추정하고 콜모고로프-스미르노프(KS) 적합도 검정 및 로그정규분포 등 대안 헤비테일 모델과의 우도비 검정(Likelihood Ratio Test)을 수행한다.

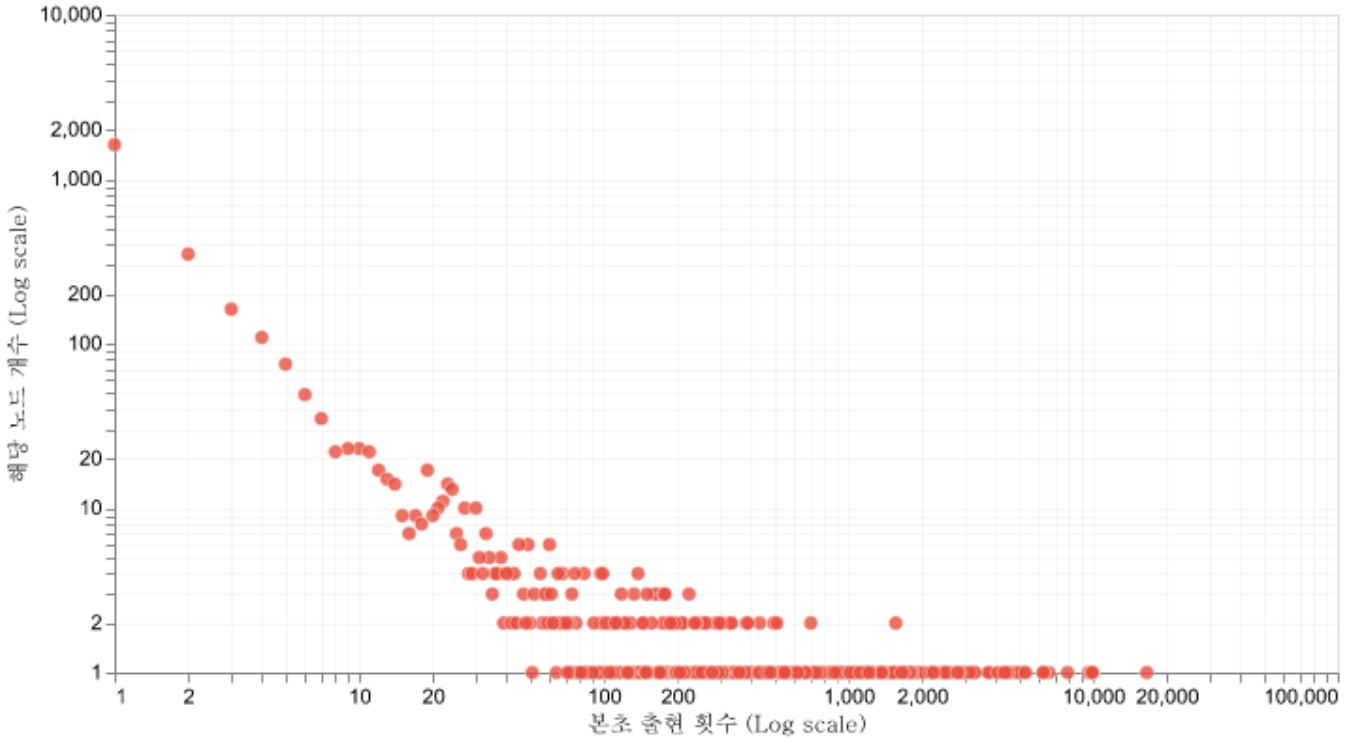

Figure 3.2 본초 출현 빈도 분포

이러한 헤비테일 현상을 설명하는 대표적인 이론적 수리 모델이 **멱법칙 분포(Power-law Distribution)**다(Clauset et al. 2009). 멱법칙은 한 수량이 다른 수량의 거듭제곱(Power)에 비례하는 확률분포로, 극단적으로 큰 값을 지닌 소수 사건과 매우 작은 값을 갖는 다수 사건이 공존하는 구조를 띤다. 멱법칙은 측정 단위나 관찰 배율(스케일)을 바꾸어도 특유의 비대칭 분포 구조가 동일하게 유지되는 척도 불변성(Scale Invariance)을 지닌다.[2] 평균을 중심으로 대칭을 이루는 정규분포와 달리 멱법칙 분포에서는 연결 수가 압도적으로 많은 중심 거점(Hub) 노드가 자연스럽게 나타난다. 이러한 비대칭 분포 양상을 현상적 관점에서는 흔히 '롱테일 현상'이라 부른다.

$$P(X = x) \propto x^{-\gamma}$$

(X: 관측 대상 확률변수. x: 관측값. γ: 멱지수(Power-law Exponent, 통상 $2 < \gamma < 3$))

롱테일(Long-tail) 현상은 상위 소수(Head)에 전체 비중이 집중되고 하위 다수(Tail)가 긴 꼬리 형태로 늘어나는 데이터의 비대칭 양상을 가리킨다. 파레토 법칙(80/20 법칙)과도 맥락을 같이한다. 한의학 처방 데이터에서도 사용 빈도가 높은 핵심 본초가 머리(Head) 영역을 차지하고 출현 빈도가 낮은 다수 본초가 긴 꼬리(Tail)를 형성한다. 이처럼 롱테일 현상이 나타날 때 항목의 순위(Rank)와 빈도(Frequency) 사이에 성립하는 경험적 규칙을 '지프의 법칙'이라 부른다.

지프의 법칙(Zipf's Law)은 멱법칙을 데이터 항목의 순위(r)와 출현 빈도(f)의 관계에 적용한 법칙이다(Zipf 1949). 순위가 r번째인 본초의 출현 빈도는 순위의 거듭제곱에 반비례한다. 스케일 지수 $k \approx 1$일 때 순위와 빈도의 곱이 일정한 상수($f(r) \cdot r \approx$ const)를 유지한다. 즉, 1위 본초의 빈도를 기준으로 할 때 2위 본초는 약 1/2, 3위 본초는 약 1/3 수준으로 빈도가 급격히 줄어든다.

[2] 척도 불변성(Scale Invariance)은 관찰 단위를 c배 확대하거나 축소해도 확률분포의 함수 형태가 보존되는 수학적 성질이다. 임의의 양수 c에 대해 $P(cx) = c^{-\gamma}P(x)$ 관계가 성립한다.

양쪽 축에 모두 로그를 취한 로그-로그(Log-Log) 스케일 좌표계로 시각화하면 기울기 $-k$를 지닌 우하향 직선으로 나타난다.

$$f(r) \propto \frac{1}{r^{k}} \quad (k \approx 1)$$

(r: 본초의 출현 빈도 순위(Rank). $f(r)$: 해당 순위 본초의 출현 빈도. k: 멱지수(스케일 지수))

3.3 본초 배합 빈도

이제 본초 사이의 관계, 즉 엣지(Edge) 관점에서 데이터를 살펴보자. 두 본초가 한 처방에 함께 등장하는 동시 출현(Co-occurrence) 빈도를 측정하면 본초 간 배합 관계를 파악할 수 있다. 가장 자주 함께 쓰인 본초 쌍은 '감초(甘草) - 인삼(人蔘)'다. 전체 처방의 11.6%에 해당하는 5,632개 처방에서 함께 쓰였다. 감초(甘草) - 복령(茯苓)(5,333회), 복령(茯苓) - 인삼(人蔘)(4,909회), 감초(甘草) - 당귀(當歸)(4,442회), 당귀(當歸) - 작약(芍藥)(4,433회), 감초(甘草) - 작약(芍藥)(4,334회), 당귀(當歸) - 천궁(川芎)(3,759회), 감초(甘草) - 백출(白朮)(3,693회), 감초(甘草) - 진피(陳皮)(3,613회), 인삼(人蔘) - 백출(白朮)(3,490회) 등이 뒤를 이으며 높은 동시 출현 빈도를 보였다.

Table 3.3 동시 출현 빈도 상위 본초 배합 (상위 20개)

순위	본초 쌍	동시 출현 빈도 (회)	비중(전체 처방 대비, %)	비중(전체 동시 출현 빈도 대비, %)
1	감초(甘草) - 인삼(人蔘)	5,632	11.6	0.3
2	감초(甘草) - 복령(茯苓)	5,333	11.0	0.3
3	복령(茯苓) - 인삼(人蔘)	4,909	10.1	0.3
4	감초(甘草) - 당귀(當歸)	4,442	9.2	0.3

순위	본초 쌍	동시 출현 빈도 (회)	비중(전체 처방 대비, %)	비중(전체 동시 출현 빈도 대비, %)
5	당귀(當歸) - 작약 (芍藥)	4,433	9.2	0.3
6	감초(甘草) - 작약 (芍藥)	4,334	8.9	0.3
7	당귀(當歸) - 천궁 (川芎)	3,759	7.8	0.2
8	감초(甘草) - 백출 (白朮)	3,693	7.6	0.2
9	감초(甘草) - 진피 (陳皮)	3,613	7.5	0.2
10	인삼(人蔘) - 백출 (白朮)	3,490	7.2	0.2
11	복령(茯苓) - 백출 (白朮)	3,459	7.1	0.2
12	감초(甘草) - 황금 (黃芩)	3,236	6.7	0.2
13	감초(甘草) - 천궁 (川芎)	3,213	6.6	0.2
14	당귀(當歸) - 인삼 (人蔘)	3,068	6.3	0.2
15	감초(甘草) - 방풍 (防風)	2,771	5.7	0.2
16	복령(茯苓) - 당귀 (當歸)	2,682	5.5	0.2
17	복령(茯苓) - 진피 (陳皮)	2,619	5.4	0.2
18	작약(芍藥) - 천궁 (川芎)	2,596	5.4	0.2
19	복령(茯苓) - 작약 (芍藥)	2,434	5.0	0.1
20	감초(甘草) - 반하 (半夏)	2,385	4.9	0.1

앞서 살펴본 노드 분포와 마찬가지로 엣지 빈도를 산점도로 나타내면 동시 출현 빈도 역시 두터운 꼬리를 갖는 헤비테일(Heavy-tailed) 분포 양상을 보인다.

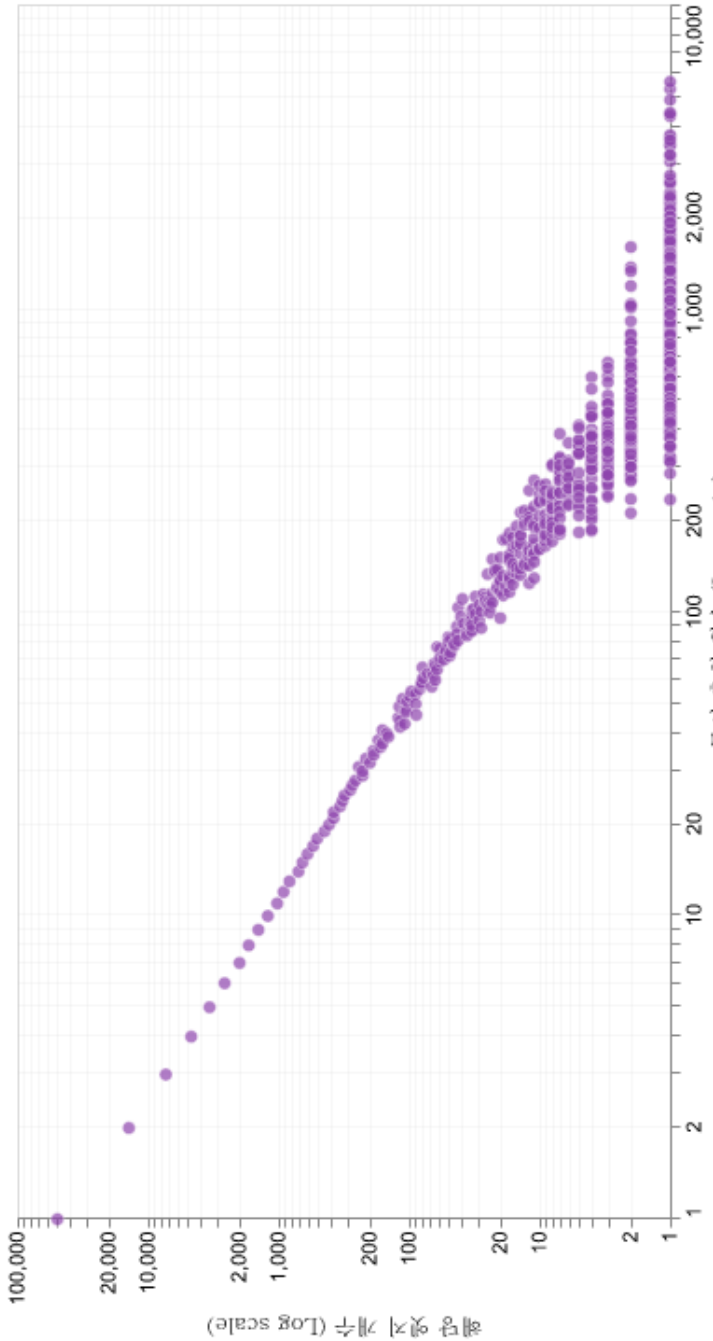

Figure 3.3 본초 간 동시 출현 빈도 분포

3.4 희소 데이터 처리

처방 데이터에서 '본초 출현 빈도'와 '본초 배합 빈도'는 모두 전형적인 헤비테일(Heavy-tailed) 양상의 비대칭 분포를 보인다. 이는 빈도가 낮은 본초나 배합 관계가 전체의 대다수를 차지함을 뜻한다. 이렇게 출현 횟수가 적은 구간을 희소 데이터 또는 롱테일 영역이라 부른다. 희소 데이터는 전체 데이터 양에서 큰 비중을 차지하지만 개별 항목이 담고 있는 정보량은 적다. 예컨대 등장 횟수가 1~2회에 불과한 본초는 임상적 의도를 해석하기 어렵고 단순한 기록 오류일 가능성도 배제할 수 없다. 두 본초의 배합 관계인 동시 출현(Co-occurrence) 빈도도 마찬가지다.

헤비테일 분포와 희소 데이터는 실제 복잡계 네트워크에서 보편적으로 관찰된다. 희소 데이터는 종류가 매우 방대하지만 개별 빈도는 낮으므로 분석 목적에 맞게 정제해야 한다. 출현 빈도가 극히 낮은 노드와 엣지를 모두 포함하면 네트워크가 지나치게 복잡해져 구조적 특징을 파악하기 어렵다. 노이즈에 가까운 희소 노드나 엣지를 걸러내면 네트워크의 핵심 골격을 명확히 드러낼 수 있다. 일반적으로 특정 임계값을 설정하여 이에 미치지 못하는 데이터를 걸러내는 필터링 기법을 활용한다.

희소 데이터를 정제하려면 어떤 값을 기준으로 삼을지 결정해야 한다. 도메인 지식을 바탕으로 기준을 정할 수도 있지만 객관적이고 재현 가능한 분석을 위해서는 데이터 자체의 분포 특성에서 기준을 도출하는 편이 바람직하다. 여기서는 '누적 빈도 커버리지' 방식을 활용해 희소 데이터 필터링 기준을 도출해보자.

누적 빈도 커버리지(Cumulative Coverage)는 빈도를 내림차순으로 정렬한 뒤 전체 총량 대비 누적 비율을 계산해 필터링 기준을 도출하는 방법이다. 수학적으로는 경험적 누적분포함수

(ECDF)[3]의 원리를 적용하여 전체 출현 빈도 총합(확률 질량) 가운데 상위 k개 본초가 차지하는 누적 비율을 계산한다. 출현 빈도가 높은 순서대로 누적해 전체 본초 사용량의 90%, 95%, 99%를 포괄하는 경계 순위를 찾는다. 이 경계에 위치한 본초의 빈도가 '최소 임계값'이 되며 이 값에 미치지 못하는 희소 본초를 노이즈로 분류해 걸러낸다.

내림차순으로 정렬된 본초 빈도 f_i를 바탕으로 누적 빈도 커버리지 함수는 다음과 같이 정의된다.

$$C(k) = \frac{\sum_{i=1}^{k} f_i}{\sum_{i=1}^{N} f_i}$$

(f_i: 내림차순 정렬된 i번째 본초의 출현 빈도. k: 누적 계산 대상 상위 본초 수(임계 순위). N: 전체 본초 종류 수)

Table 3.4 누적 빈도 커버리지 기준별 필터링 결과 (노드 기준)

커버리지(%)	최소 임계값 (빈도)	포함 (전체 종류 대비)	제외 (전체 종류 대비)
99%	5	892종 (27.9%)	2,304종 (72.1%)
98%	23	580종 (18.1%)	2,616종 (81.9%)
97%	43	462종 (14.5%)	2,734종 (85.5%)
96%	68	397종 (12.4%)	2,799종 (87.6%)
95%	98	354종 (11.1%)	2,842종 (88.9%)
94%	127	321종 (10.0%)	2,875종 (90.0%)
93%	150	296종 (9.3%)	2,900종 (90.7%)
92%	178	274종 (8.6%)	2,922종 (91.4%)
91%	202	256종 (8.0%)	2,940종 (92.0%)
90%	223	239종 (7.5%)	2,957종 (92.5%)

위 표를 살펴보면 커버리지를 98%로 설정할 경우 전체 본초의 대다수인 2,616종(전체 본초 종류 대비 81.9%)를 제외하더라도 손실되는 실제 데이터 총량은 2%에 그친다. 이를 그래프로 시각화하면 다음과 같다.

[3] 경험적 누적분포함수(ECDF, Empirical Cumulative Distribution Function): 표본 데이터를 크기순으로 정렬했을 때 특정 값 이하인 관측치의 누적 비율을 계산하는 계단식 통계 함수이다.

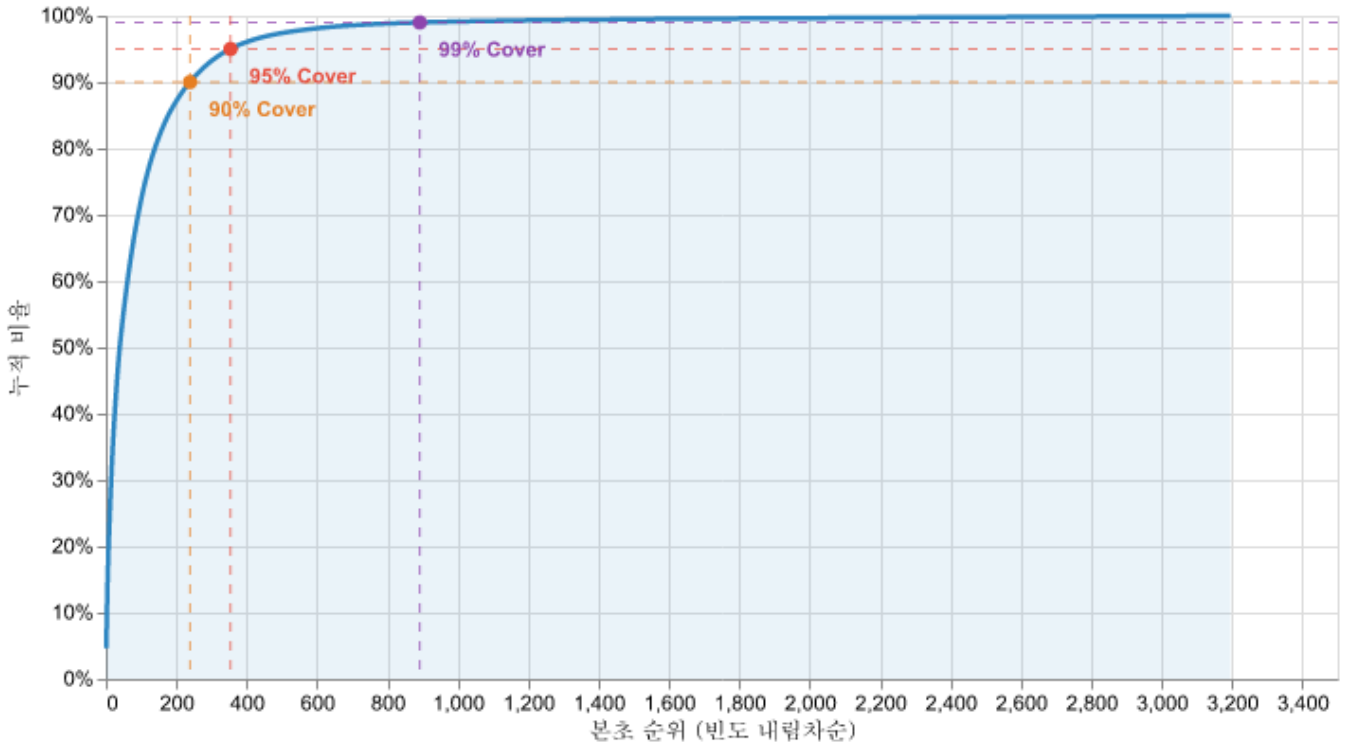

Figure 3.4 누적 빈도 커버리지 분포 (노드 기준)

이처럼 누적 빈도 커버리지로 임계 기준을 정해 롱테일 영역의 희소 데이터를 걸러내면 노이즈가 줄어들면서 네트워크의 핵심 구조가 한층 선명하게 드러난다.

본초 배합 빈도에도 동일한 누적 빈도 커버리지 방식을 적용할 수 있다.

Table 3.5 누적 빈도 커버리지 기준별 필터링 결과 (엣지 기준)

커버리지(%)	최소 임계값 (동시출현 빈도)	포함 (전체 종류 대비)	제외 (전체 종류 대비)
99%	1	89,633개 (84.4%)	16,512개 (15.6%)
98%	1	73,121개 (68.9%)	33,024개 (31.1%)
97%	2	56,642개 (53.4%)	49,503개 (46.6%)
96%	2	48,386개 (45.6%)	57,759개 (54.4%)
95%	3	40,974개 (38.6%)	65,171개 (61.4%)
94%	3	35,469개 (33.4%)	70,676개 (66.6%)
93%	4	31,314개 (29.5%)	74,831개 (70.5%)
92%	5	27,888개 (26.3%)	78,257개 (73.7%)
91%	6	25,041개 (23.6%)	81,104개 (76.4%)
90%	8	22,650개 (21.3%)	83,495개 (78.7%)

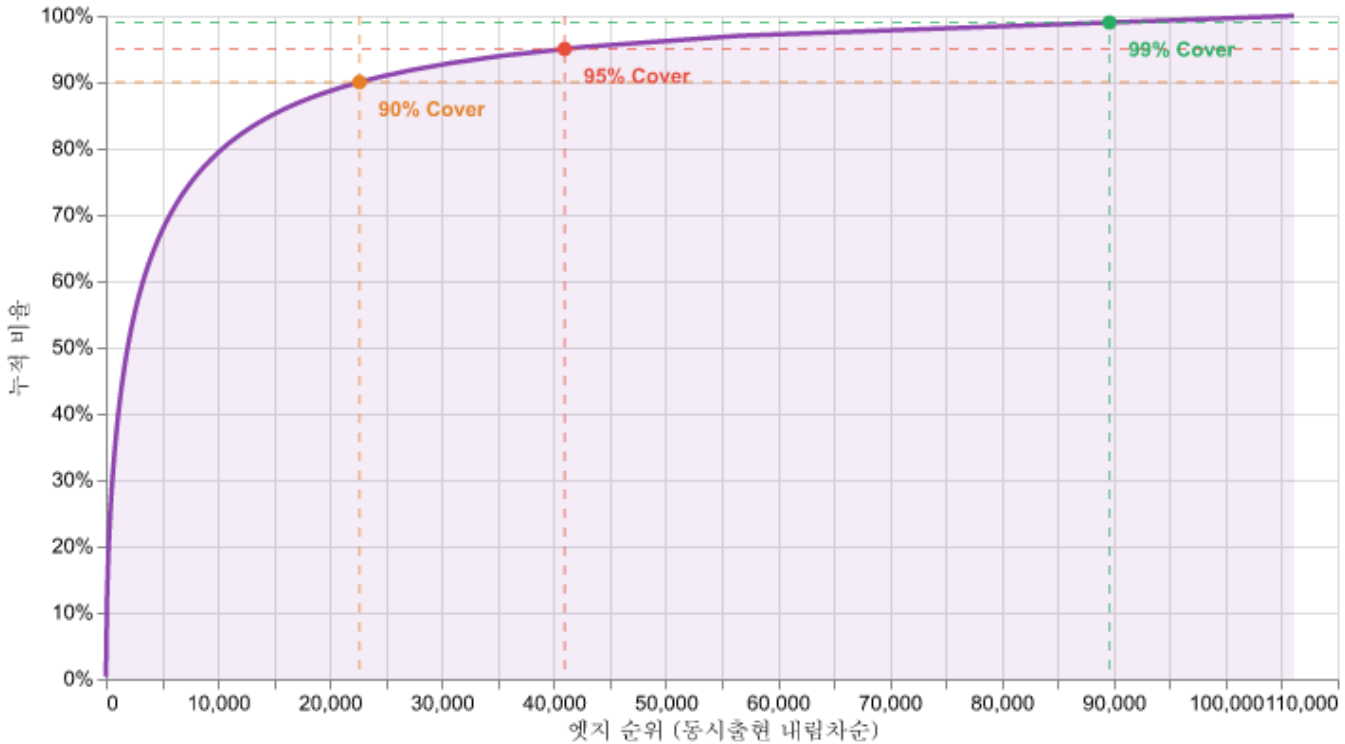

Figure 3.5 누적 빈도 커버리지 분포 (엣지 기준)

표와 그래프에서 확인되듯 동시 출현 빈도가 매우 낮은 엣지를 제외하더라도 본초 네트워크의 핵심 배합 골격은 대부분 온전하게 보존된다.

지금까지 처방 및 본초 데이터의 크기와 분포 특성, 롱테일 구조를 살피고 누적 빈도 커버리지를 활용한 데이터 필터링 임계값을 도출했다.

단순 빈도 기반 필터링은 노드와 엣지를 걸러내는 양적 기준을 제시하지만 개별 연결의 질적인 배합 강도까지 설명하지는 못한다. 제4장에서는 이러한 한계를 보완하기 위해 엣지 가중치를 산출하는 다양한 평가 척도를 살펴본다.

> **i** 이후 분석에 사용되는 임계값
>
> 이 책에서는 노드(본초 빈도)의 필터링 기준으로는 **98% 커버리지**를 채택한다. 현재 데이터셋 기준 임계값은 **23회**다. 엣지(본초 간 동시 출현 빈도)는 **90% 커버리지**를 기준으로 설정하며 이에 해당하는 동시 출현 빈도는 **8회**다. 제5장 이후 네트워크 분석과 배합 패턴 추출 과정에서 롱테일 노이즈를 제거하는 기준으로 이 두 수치를 사용한다.

제 4 장

엣지 가중치

네트워크의 형태, 즉 위상 구조는 노드와 노드를 잇는 엣지에 따라 달라진다. 그리고 엣지는 엣지 가중치(Edge Weight)를 어떻게 정의하느냐에 따라 결정된다. '본초'를 노드로, '본초와 본초의 관계'를 엣지로 모델링할 때 두 본초가 얼마나 자주 함께 쓰였는지를 나타내는 '동시 출현 빈도'를 직관적인 엣지 가중치로 활용할 수 있다. 하지만 앞서 탐색적 데이터 분석(EDA)에서 살펴보았듯 동시 출현 빈도는 헤비테일(Heavy-tailed) 분포를 띤다. 이 수치를 그대로 사용하면 엣지 가중치 사이의 편차가 지나치게 커져 비교와 연산에서 왜곡이 발생할 수 있다. 따라서 롱테일 영역을 적절히 정돈하는 작업과 더불어 엣지 가중치를 도출하는 방식도 함께 고려해야 한다.

이 장에서는 직관적이면서 널리 쓰이는 네 가지 수학적 지표로 본초 네트워크의 엣지 가중치를 도출하고 그 결과를 비교해 본다.

4.1 주요 지표

먼저 각 지표를 살펴보자.

◎ **동시 출현 빈도 (Co-occurrence Frequency)**

가장 직관적인 지표다. 두 노드가 함께 등장한 횟수를 그대로 가중치로 쓴다. 본초 네트워크에 적용했을 때 계산이 간단하고 임상에서 흔히 쓰이는 조합을 직관적으로 보여준다. 하지만 출현 빈도가 높은 노드일수록 가중치가 높게 산출될 가능성이 크다. 감초나 생강처럼 그 자체로 많이 쓰인 본초일수록 동시 출현 빈도 역시 커지기 때문이다. 이처럼 동시 출현 빈도에는 '고빈도 편향(High-frequency Bias)'이 나타날 수 있다.

◎ **자카드 지수 (Jaccard Index)**

두 집합 사이의 유사도를 측정하는 대표적인 지표이다(Jaccard 1901). 수학적으로는 두 집합의 교집합 크기를 합집합 크기로 나눈 비율로 정의된다. 지숫값은 0과 1 사이이며 두 집합이 완전히 일치하면 1, 공통 원소가 전혀 없으면 0이 된다. 본초 네트워크에 적용하면 처방 데이터에서 두 본초 A와 B의 자카드 지수를 구할 수 있다. 이는 두 본초가 함께 등장한 처방 수(교집합)를 두 본초 중 적어도 하나가 포함된 전체 처방 수(합집합)로 나눈 비율이다.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{\text{cooc}(A,B)}{\text{freq}(A) + \text{freq}(B) - \text{cooc}(A,B)}$$

(A, B: 비교 대상 본초. $freq(A), freq(B)$: 각 본초의 개별 출현 처방 수. $cooc(A,B)$: 두 본초의 동시 출현 처방 수.)

자카드 지수는 동시 출현 빈도에서 나타나는 '고빈도 편향'을 완화하는 효과가 있다. 특정 본초의 출현 빈도가 매우 높아도 각 본초의 개별 출현 빈도가 분모에 반영되므로 함께 쓰이지 않는다면 합집합이 커져 지숫값이 작아진다. 하지만 출현 빈도가 낮은 희소 데이터(롱테일 영역)에서는 점수가 과대평가되는 '저빈도 편향(Low-frequency Bias)'이 나타난다. 전체 처방에서 단 한

번씩만 등장한 두 희소 본초가 우연히 한 처방에 함께 사용된 경우에도 지수는 최댓값인 1.0이 된다.

◎ **상호 정보량 (Pointwise Mutual Information, PMI)**

정보 이론에서 유래하여 자연어 처리 및 단어 연관성 분석에 널리 활용되는 지표이다(Church and Hanks 1990). 두 사건이 서로 독립이라고 가정했을 때의 '기대 확률'과 실제 '관측 확률'의 비율을 로그(Log)로 나타낸 값이다. 두 본초가 독립적으로 결합하는 귀무 상태($P(x,y) = P(x)P(y)$)에서는 $\log_2(1) = 0$으로 기준선이 된다. 두 본초가 우연한 기대보다 자주 함께 처방되면 양수(> 0)가 되고, 반대로 덜 처방되면 음수(< 0)가 된다. 로그의 밑으로 2를 사용하므로 우연 대비 결합 확률의 증가 비율을 비트(bit) 단위 정보량으로 정량화한다. 이 방법 역시 분모의 개별 확률이 커지는 고빈도 본초의 점수를 낮추어 '고빈도 편향'을 완화할 수 있다. 하지만 출현 빈도가 극히 낮은 본초 조합에서는 점수가 비정상적으로 과대평가되는 '저빈도 편향'이 발생한다. 관측 확률($P(x,y) \sim O(1/N)$)을 나타내는 분자보다 독립 기대 확률의 곱($P(x)P(y) \sim O(1/N^2)$)인 분모가 훨씬 빠르게 0으로 수렴하기 때문이다.

$$\text{PMI}(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

($P(x,y)$: 두 본초 x, y의 동시 출현 확률. $P(x), P(y)$: 각 본초의 개별 출현 확률)

◎ **G-score**

통계학에서는 두 범주형 변수가 서로 독립인지 유의미한 연관성이 있는지를 판단하기 위해 **우도비 검정(Likelihood Ratio Test)**[1]을 사용한다(Dunning 1993). 2×2 교차표(Contingency

[1] 우도비 검정(Likelihood Ratio Test, G-test): 관측 빈도와 독립 가정하의 기대 빈도 사이의 비율을 로그 우도로

Table)[2]를 만들어 4개 셀(모든 경우의 수)의 관측 빈도(O)와 기댓값인 기대 빈도(E)의 차이를 합산한다. 이를 '표준 G-test 통계량(G^2_{standard})'이라 한다. 수학적으로는 독립을 가정한 귀무 모형과 관측 데이터의 대립 모형 사이의 우도비(Likelihood Ratio)에 로그를 취한 통계량이다.[3] 이 통계량은 표본 크기가 충분할 때 점근적으로 자유도 1인 카이제곱 분포(χ^2)를 따르며 두 사건의 결합이 단순한 우연이 아니라는 통계적 증거를 제공한다. 희소 데이터가 많은 환경에서도 정규분포를 가정하는 Z-점수나 표본 크기에 민감한 피어슨 카이제곱 검정보다 안정적인 결과를 보인다.

$$G^2_{\text{standard}} = 2\sum_{i,j} O_{ij} \ln\left(\frac{O_{ij}}{E_{ij}}\right)$$

(O_{ij}: 2×2 교차표 각 셀의 관측 빈도. E_{ij}: 두 사건이 독립일 때의 기대 빈도)

앞에서 살펴보았듯 한의학 처방 데이터에서는 자연어 텍스트 데이터와 마찬가지로 지프의 법칙(Zipf's Law)이 관찰된다. 전체 처방 수(N)가 매우 크기 때문에 어떤 두 본초가 한 처방에서 만날 확률(p)은 극도로 낮다($N \to \infty, p \to 0$). 이런 환경에서는 4개 셀을 모두 합산하는 표준 G-test가 계산 면에서 비효율적이며 기대 빈도가 극히 낮은 구간에서 검정 통계량이 과대평가될 우려도 있다.

통계 언어학자 Stefan Evert는 텍스트 연어(Collocation) 분석에서 이항 분포 대신 희소 데이터에 강건한 포아송 분포(Poisson Distribution)를 가정하여 표준 G-test 통계량의 근사 공식을 제안했다(Evert 2005, 2008). 4개 셀을 합산하는 복잡한 연산 대신 아래 수식처럼 두

변환하여 두 범주형 변수의 통계적 연관성을 검정하는 방법이다.

[2] 분할표 또는 교차표(Contingency Table): 두 변수의 범주 조합에 따른 출현 빈도를 행과 열로 교차 집계한 행렬이다.

[3] 정보이론 관점에서 표준 G-test 통계량은 실제 관측 분포와 독립 가정 귀무 분포 사이의 정보 차이를 측정하는 쿨백-라이블러 발산(Kullback-Leibler Divergence)을 바탕으로 한다.

단어가 함께 등장하는 관측값(O)과 기댓값(E)만으로 근삿값을 도출한다. 수식에서 $O = E$일 때 $O\ln(1) - 0 = 0$이 되어 점수의 기준점이 된다. 이 책에서는 편의상 이 지표를 **G-score**라고 부른다.

$$G \approx 2\left(O\ln\left(\frac{O}{E}\right) - (O - E)\right)$$

(O: 관찰 빈도. E: 기대 빈도. $O\ln(O/E)$: 정보 이론의 PMI($\ln(O/E)$)에 관측 빈도(O)를 가중치로 곱한 형태. $-(O-E)$: 이항 분포를 포아송 분포로 근사하는 과정에서 도출된 항으로, 관측 빈도와 기대 빈도가 일치할 때 점수가 0이 되도록 보정.)

G-score는 기본적으로 '독립성에서 얼마나 벗어났는가'를 측정하는 지표다. 본초를 예로 들면 이 '벗어남'은 대개 '두 본초가 우연보다 훨씬 자주 함께 쓰이는 경우'를 뜻한다. 하지만 수학적으로는 '두 본초가 우연보다 훨씬 드물게 함께 쓰이는 경우' 역시 이에 해당한다. 즉, 우연보다 훨씬 자주 사용된 조합뿐만 아니라 반대로 우연보다 훨씬 드물게 쓰인 조합도 점수가 높아진다.

따라서 분석 목적에 따라 관측값이 기댓값보다 큰 경우($O > E$)와 작은 경우($O < E$)를 구분해야 한다. 본초 사이의 긍정적인 배합 관계($O > E$)를 탐색할 때는 배합 기피 관계($O < E$)에 해당하는 조합을 제거하거나 0 이하의 값을 부여하여 보정해야 한다. 만약 상오(相惡)·상반(相反)과 같은 배합 기피 관계를 밝히고자 한다면 기준을 반대로 적용하면 된다. 이 책에서는 긍정적인 배합 관계에 집중하기 위해 관측값이 기댓값보다 작으면 점수를 0으로 처리하여 분석에서 제외했다.

> **임베딩 기반 엣지 가중치**
>
> 노드를 고차원 벡터 공간에 투영하는 임베딩 기법으로 노드 사이의 관계를 정의할 수도 있다. TF-IDF와 같은 선형 통계 방식이나 Word2Vec(Mikolov et al. 2013), Node2Vec(Grover and Leskovec 2016) 같은 비선형 기계학습 방식을 활용하여 각 노드를 수치 벡터로 변환한다. 이렇게 생성된 벡터 사이의 유사도를 코사인 유사도(Cosine Similarity) 등으로 측정한다. 계산된 유사도 점수는 네트워크의 엣지 가중치로 직접 활용할 수 있다. 예를 들어 처방-본초 데이터에 이를 적용할 때는 각 처방을 하나의 문서로 보고 본초를 단어로 간주할 수 있다. 이렇게 본초 임베딩 벡터를 추출한 뒤 본초 사이의 코사인 유사도를 계산하면 배합 가중치로 활용할 수 있다(오준호 2019). 노드 임베딩과 유사도 측정을 바탕으로 한 접근은 데이터 분석 및 인공지능 분야에서 널리 쓰이는 중요한 방법론이다. 다만 이 책에서는 그래프 위상 구조 분석에 초점을 맞추므로 이러한 임베딩 기법을 따로 다루지 않는다.

4.2 지표 비교

네 가지 지표를 한의학 처방 데이터로 구축한 본초 네트워크에 적용해 보고 상위 조합을 비교해 본다.

◎ 동시 출현 빈도

동시 출현 빈도 기준 상위 10개 엣지는 다음과 같다. 예상대로 출현 빈도가 높은 본초가 높은 점수를 얻었다. 한의학 처방에서 임상적으로 널리 쓰이는 본초 배합을 보여준다는 의미는 있지만, 결합 특이성이 높은 배합을 가려내기는 어렵다.

Table 4.1 동시 출현 빈도 상위 조합 (상위 10개)

순위	본초 Source	본초 Target	동시 출현 빈도	자카드 지수	PMI	G-score
1	감초(甘草)	인삼(人蔘)	5632	0.274	0.78	1385.87
2	감초(甘草)	복령(茯苓)	5333	0.252	0.65	932.54
3	복령(茯苓)	인삼(人蔘)	4909	0.336	1.32	3097.65
4	감초(甘草)	당귀(當歸)	4442	0.201	0.39	293.31
5	당귀(當歸)	작약(芍藥)	4433	0.332	1.46	3327.74
6	감초(甘草)	작약(芍藥)	4334	0.215	0.69	842
7	당귀(當歸)	천궁(川芎)	3759	0.302	1.55	3125.51
8	감초(甘草)	백출(白朮)	3693	0.19	0.71	763.11
9	감초(甘草)	진피(陳皮)	3613	0.188	0.75	832.42
10	인삼(人蔘)	백출(白朮)	3490	0.275	1.42	2495.86

◎ 자카드 지수

자카드 지수 기준 상위 10개 엣지는 다음과 같다. 전체 데이터에서 출현 빈도가 매우 낮은 희소 본초 조합이 상위권을 차지했다. 각 본초의 개별 출현 빈도와 동시 출현 빈도가 일치하여 자카드 지수가 최댓값인 1.0으로 산출되었기 때문이다. 저빈도 구간에서 유사도가 과대평가되는 저빈도 편향을 뚜렷하게 관찰할 수 있다.

Table 4.2 자카드 지수 상위 조합 (상위 10개)

순위	본초 Source	본초 Target	자카드 지수	동시 출현 빈도	PMI	G-score
1	철마편(鐵馬鞭)	주담근(酒罈根)	1	2	14.56	36.38
2	배백(排帛)	도극자(倒棘刺)	1	2	14.56	36.38
3	왜장엽(矮樟葉)	서강삼편 (西江杉片)	1	1	15.56	19.58
4	형화(螢火)	단조회철 (鍛灶灰鐵)	1	1	15.56	19.58
5	보탑초(寶塔草)	왜금둔엽 (矮金屯葉)	1	1	15.56	19.58
6	보탑초(寶塔草)	추면등엽 (皺面藤葉)	1	1	15.56	19.58
7	왜금둔엽(矮金屯葉)	추면등엽 (皺面藤葉)	1	1	15.56	19.58
8	진극우담성 (陳極牛膽星)	향서우각 (香犀牛角)	1	1	15.56	19.58
9	석회석(石灰石)	가과회(茄科灰)	1	1	15.56	19.58
10	용규(龍葵)	건지황엽 (乾地黃葉)	1	1	15.56	19.58

◎ 상호 정보량(PMI)

PMI 기준 상위 10개 엣지는 다음과 같다. 출현 빈도가 매우 낮은 희소 본초 사이의 배합에서 높은 점수가 나타났다. 이는 앞서 살펴보았듯 저빈도 구간에서 분모의 기대 확률 곱($P(x)P(y)$)이 급격히 0으로 수렴하여 지숫값이 과대평가되는 저빈도 편향의 결과다.

Table 4.3 PMI 상위 조합 (상위 10개)

순위	본초 Source	본초 Target	PMI	자카드 지수	G-score	동시 출현 빈도
1	황니(黃泥)	적반두(赤飯豆)	15.56	1	19.58	1
2	구룡등(九龍藤)	두귤수두 (頭橘樹頭)	15.56	1	19.58	1
3	지고루(地骷髏)	선인사의 (仙人蓑衣)	15.56	1	19.58	1
4	자배초(紫背草)	마제향(馬蹄香)	15.56	1	19.58	1
5	금성여석 (金星礜石)	악설여석 (握雪礜石)	15.56	1	19.58	1
6	사향각(麝香殼)	화지주(花蜘蛛)	15.56	1	19.58	1
7	오대초(五臺草)	황화근(黃花根)	15.56	1	19.58	1
8	춘동(春冬)	하추(夏秋)	15.56	1	19.58	1
9	엽자청(葉子靑)	마도석(磨刀石)	15.56	1	19.58	1
10	자제구교 (自製龜膠)	별갑교(鱉甲膠)	15.56	1	19.58	1

◎ G-score

G-score 기준 상위 10개 엣지는 다음과 같다. 앞선 지표들보다 균형 잡힌 결과를 보인다. 자주 쓰이면서도 두 본초 간 특이성이 높은 배합이 상위권에 자리 잡았다.

Table 4.4 G-score 상위 조합 (상위 10개)

순위	본초 Source	본초 Target	G-score	동시 출현 빈도	자카드 지수	PMI
1	유향(乳香)	몰약(沒藥)	3372.14	825	0.403	4.32
2	당귀(當歸)	작약(芍藥)	3327.74	4433	0.332	1.46
3	당귀(當歸)	천궁(川芎)	3125.51	3759	0.302	1.55
4	복령(茯苓)	인삼(人蔘)	3097.65	4909	0.336	1.32
5	방풍(防風)	강활(羌活)	2697.78	1514	0.243	2.47
6	사향(麝香)	주사(硃砂)	2566.37	1100	0.263	2.94
7	삼릉(三稜)	봉아출(蓬莪朮)	2526.95	532	0.386	4.82
8	인삼(人蔘)	백출(白朮)	2495.86	3490	0.275	1.42
9	신곡(神麴)	맥아(麥芽)	2285.28	477	0.351	4.85
10	복령(茯苓)	백출(白朮)	2278.91	3459	0.264	1.35

표에서 보듯 지표에 따라 상위권 본초 배합 양상은 확연히 다르다. 이를 직관적으로 비교하기 위해 각 지표의 상위 100개 엣지를 추출하여 벤 다이어그램으로 그려보았다.

크게 보아 동시 출현 빈도 상위 100개 엣지는 자카드 지수와 PMI의 상위 100개 엣지와 전혀

겹치지 않았다(교집합 0개). 자카드 지수와 PMI는 저빈도 편향 문제가 있고 동시 출현 빈도는 고빈도 편향 문제가 있으므로 예상할 수 있는 결과다. 자카드 지수와 PMI는 상위 100개 중 36개의 엣지를 공유했지만, 동시 출현 빈도와 겹치는 영역은 없었다. 이로써 자카드 지수와 PMI의 공통적 특성을 확인할 수 있다. 두 지표 모두 단순히 자주 등장한다고 해서 점수가 높아지지는 않지만, 출현 횟수가 적은 희소 본초 조합에서 점수가 과대평가되는 저빈도 편향을 보인다. 한편 G-score는 자카드 지수 및 PMI와 교집합이 없었으며, 동시 출현 빈도와는 30개의 엣지를 공유했다. 이는 G-score가 저빈도 편향에서 비교적 자유롭고, 단순 우연에 의한 결합을 배제하여 통계적 유의성이 검증된 조합을 선별하면서도 출현 빈도를 완전히 무시하지는 않음을 보여준다.

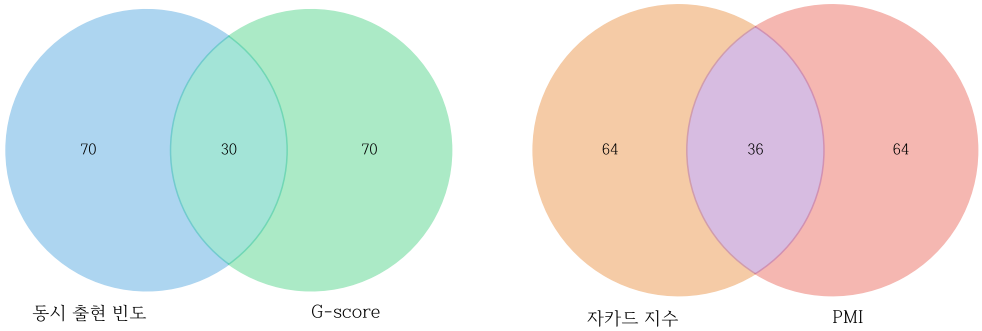

Figure 4.1 척도별 주요 엣지 비교

동시 출현 빈도와 자카드 지수, PMI, G-score가 어떤 관계를 나타내는지 산점도로 다시 살펴보자. 자카드 지수는 동시 출현 빈도가 높아질수록 완만하게 높아지는 경향을 보인다. 하지만 동시 출현 빈도가 낮은 구간에서 점들이 넓게 퍼지며, 빈도가 1인 지점에서도 최댓값인 1.0이 나타난다. PMI는 독특하게 동시 출현 빈도의 영향을 거의 받지 않으며, 고빈도 영역에서는 값이 일정하게 수렴하는 경향까지 보인다. 그러나 저빈도 영역에서는 값이 매우 넓게 퍼진다. 이는 자카드 지수와 PMI 모두 저빈도 편향이 뚜렷함을 보여준다(오준호 2016). 롱테일 영역이 넓은 데이터에서 저빈도

편향은 간과하기 어려운 문제다. 다만 이들 지표는 일정 빈도 이상에서는 고빈도 편향 없이 매우 안정적인 모습을 나타내므로, 희소 데이터 전처리를 적절히 수행하여 롱테일 영역을 억제해야 좋은 결과를 얻을 수 있다.

마지막으로 G-score는 동시 출현 빈도가 높아질수록 커지는 경향을 보인다. 아울러 고빈도나 저빈도 영역에서 값이 특정 구간에 집중되거나 비정상적으로 넓게 퍼지는 편향 현상도 관찰되지 않는다. 한편 지표값이 0에 가까운 엣지가 고빈도 영역에서도 관찰되는데, 이 역시 고빈도 편향이 효과적으로 억제되고 있음을 보여준다.

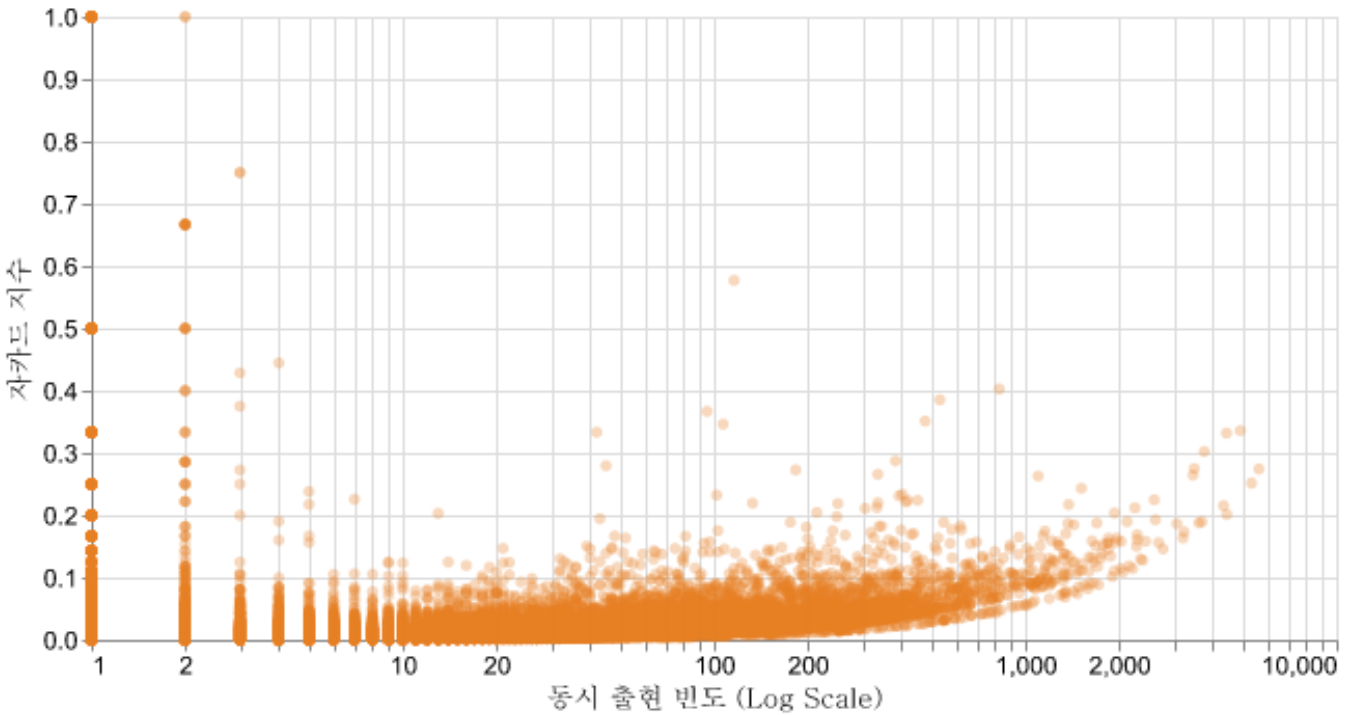

Figure 4.2 동시 출현 빈도 vs 자카드 지수 분포

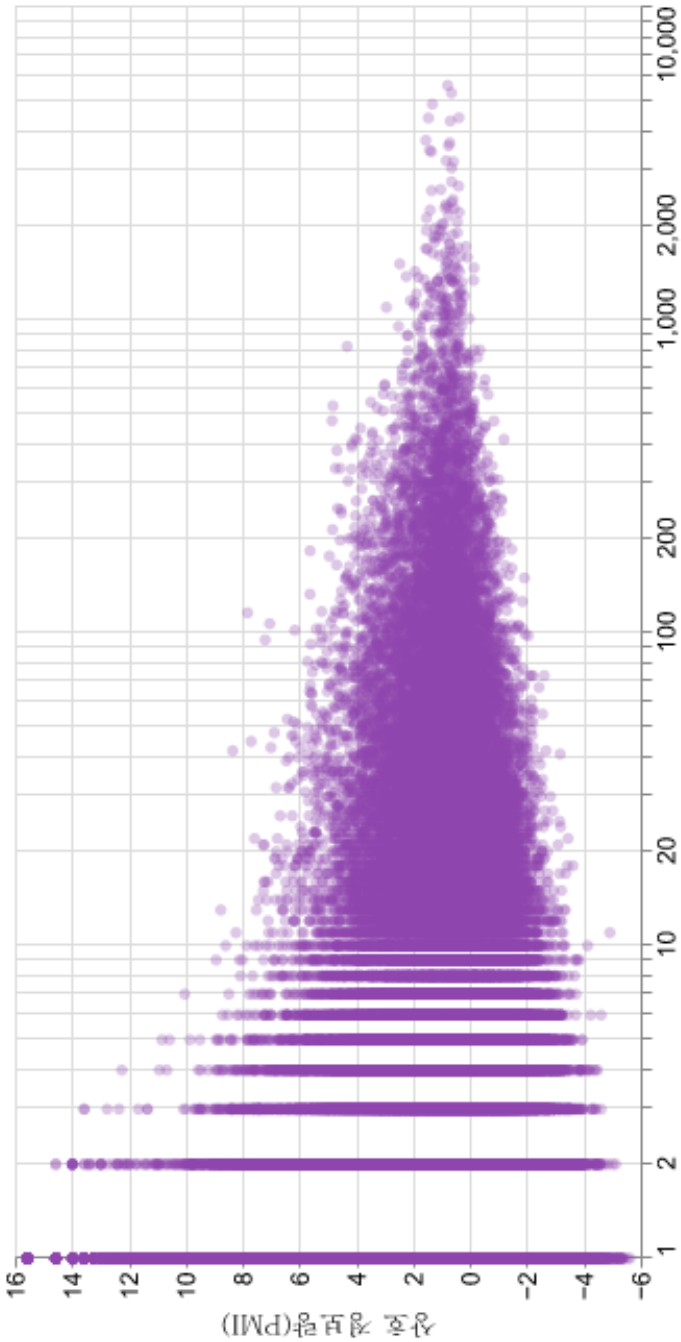

Figure 4.3 동시 출현 빈도 vs PMI 분포

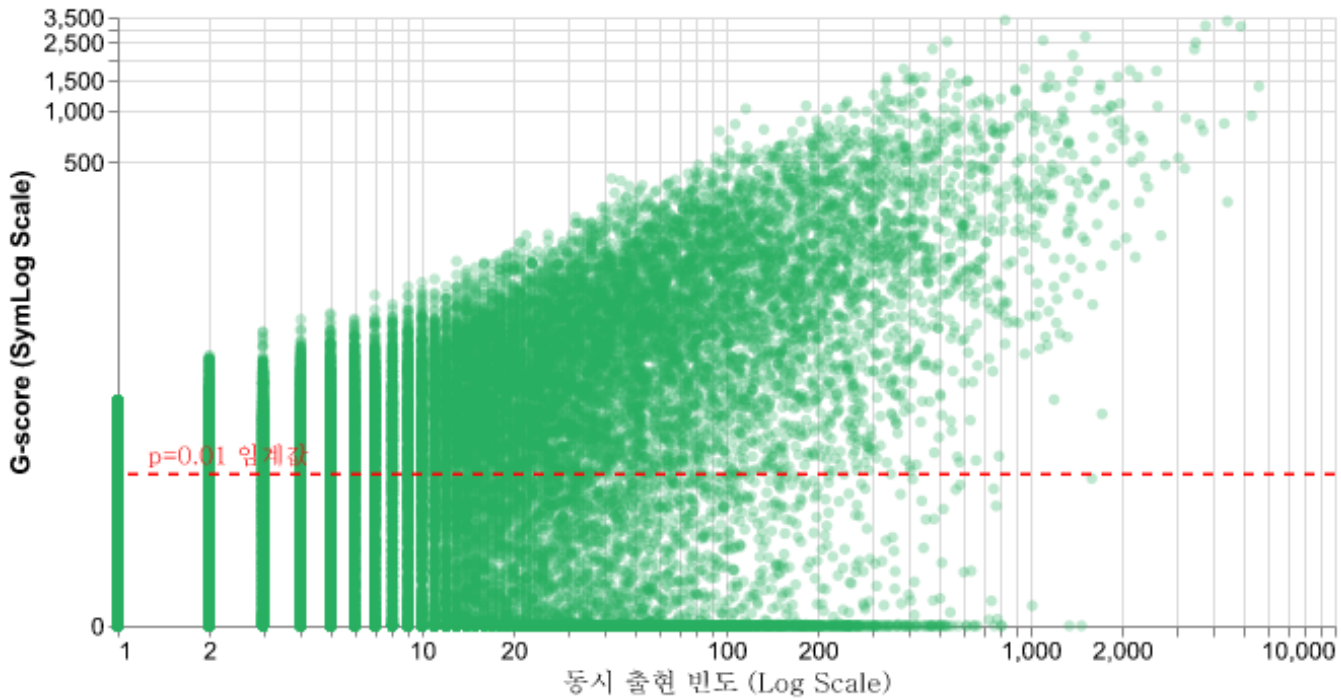

Figure 4.4 동시 출현 빈도 vs G-score 분포

4.3 분석 목적과 지표 선택

지금까지 엣지 가중치(Edge Weight)에 사용할 수 있는 몇 가지 지표를 살펴보았다. 엣지 가중치는 네트워크 위상을 결정하는 중요한 요소이다. 분석 목적에 따라 알맞은 지표를 선택해야 한다. 본초 네트워크에서는 양적으로 함께 사용된 빈도가 중요하다면 가장 직관적인 동시 출현 빈도를 쓸 수 있다. 빈도의 영향력을 제한하려면 자카드 지수나 PMI를 고려할 수 있다. 다만 두 지표 모두 저빈도 희소 데이터가 과대평가되는 한계가 있으므로 롱테일 영역을 충분히 걸러내야 한다. 자카드 지수는 상대적인 교집합 비율을 직관적으로 확인하고자 할 때 유용하며, 값이 0과 1 사이로 정규화된다는 장점도 있다. PMI는 기댓값 대비 관측값이라는 의미를 가지며 일정 빈도 이상에서 빈도의 영향을 거의 받지 않는다는 독특한 특성을 띤다. 분석 목적에 맞춰 활용한다면 두 지표 모두 좋은 선택이 될 수 있다. 한편 G-score는 빈도의 영향을 적절히 받으면서도 고빈도 편향과 저빈도 편향에서 상대적으로 자유롭다. 따라서 빈도 편향이 우려되는 경우에 적용하면 좋은 결과를 기대할 수 있다.

G-score를 조금 더 살펴보자. 이 지표의 최대 장점은 단순 점수가 아닌 통계량(Statistic)이라는 점이다. G-score는 근사적으로 자유도가 1인 카이제곱 분포(χ^2)를 따른다. 이를 활용하면 엣지 가중치로부터 유의확률(p-value)을 직접 계산할 수 있다.

Table 4.5 유의수준(p-value)에 따른 G-score 임계값 (자유도=1)

유의수준 (p-value)	신뢰수준 (%)	G-score 임계값	해석
0.1	90.00%	2.71	약한 증거
0.05	95.00%	3.84	일반적인 통계적 유의성 기준
0.01	99.00%	6.63	강한 증거
0.001	99.90%	10.83	매우 강한 증거
0.0001	99.99%	15.14	극도로 강한 증거

제3장에서는 엣지의 롱테일 영역을 덜어내고자 누적 빈도 커버리지 방식으로 '동시 출현 빈도'의 임계값을 도출했다. 그러나 G-score를 활용하면 동시 출현 빈도의 임계값을 "10 이상"과 같이

임의로 설정하는 대신, “유의수준 0.01 미만($p < 0.01$)”처럼 통계적 유의성을 바탕으로 엣지를 선택할 수 있다.

이러한 방식으로 본초 네트워크 전체를 다시 정리해 보자. 한 번이라도 같이 등장한 적이 있는 엣지 조합은 총 106,145개이다. 이 중 통계적 유의수준 0.01 ($p < 0.01$) 기준을 통과한 조합, 즉 독립 귀무가설 하에서 우연히 관측될 확률이 희박한 조합은 26,181개뿐이다. 전체의 24.7%만이 독립 결합의 귀무가설을 기각하는 핵심 배합이라 볼 수 있다. 이들은 독립 귀무가설 하에서 단순한 우연으로 관측되었을 가능성이 극히 희박한, 통계적으로 유의미한 결합이다. 다만 기준을 통과하지 못한 연결이 임상적으로 무의미하다는 뜻은 아니다. 통계적으로 단순 우연에 의해 결합되었을 가능성을 배제하기 어렵다는 의미다.

G-score로 엣지 가중치를 계산하고 통계적 유의성을 기준으로 필터링한 결과를 시각적으로 확인해 보자. 아래는 본초 네트워크에서 '방풍(防風)'과 연결된 본초만을 추출하여 구성한 에고 네트워크(Ego Network)이다. 필터링 전후로 엣지와 노드가 정돈되는 양상을 확인할 수 있다.

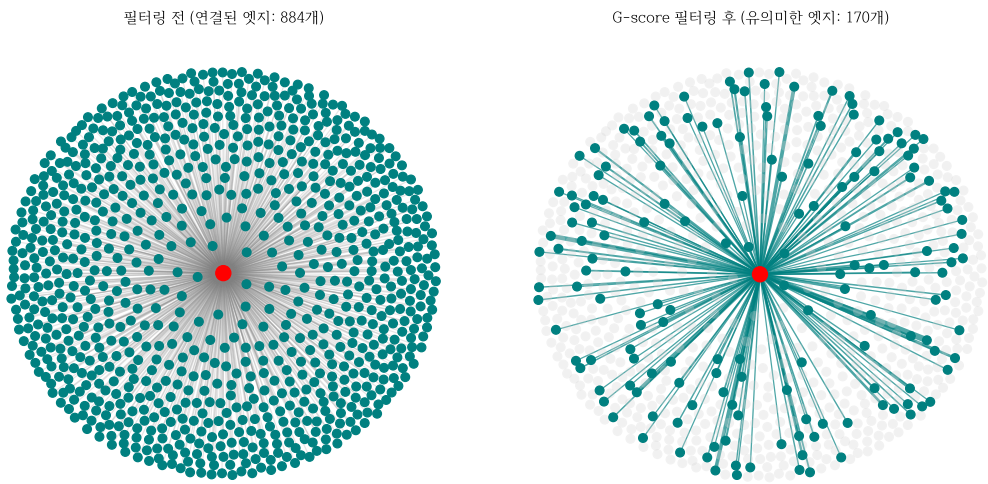

Figure 4.5 G-score 필터링 전후 비교 예시: 본초 '방풍(防風)'과 연결된 본초 네트워크

(필터링 전 왼쪽 그래프는 본초 방풍(防風)과 연결된 전체 동시 출현 노드를 청록색으로 표시. 필터링 후 오른쪽 그래프에서는 G-score 통계적 유의성 기준($p < 0.01$)을 통과한 핵심 배합 본초만 청록색으로 유지되고 통계적 유의성 기준을 통과하지 못한 배합 본초는 연회색으로 표시.)

다음 제5장에서는 이를 바탕으로 한의학 처방을 더 깊이 탐구해 보자.

제 5 장

네트워크 구조

제4장에서 엣지 가중치를 살펴보았다. 여기서 근본적인 의문이 생긴다. 분석하려는 네트워크가 아무런 원리 없이 본초를 무작위로 섞어도 나타날 수 있는 형태는 아닐까? 만약 그렇다면 오랜 기간 축적된 한의학 처방의 배합 원리는 데이터 과학의 시선에서 단순한 확률적 우연의 산물에 불과할 수 있다. 이번 장에서는 ‹단순 빈도 네트워크›와 ‹G-score 네트워크›를 무작위 네트워크(Null Model)와 대조하며 한의학 처방 데이터가 지닌 통계적 유의성과 고유한 구조적 특성을 살펴본다.

제5장부터는 실습에 쓸 기준 네트워크를 새로 설정한다. 대표적인 두 가지 형태의 네트워크를 생성하여 비교 기준으로 삼는다. 먼저 희소 데이터(롱테일 영역)가 만드는 노이즈를 줄이고자 임계값을 기준으로 노드를 필터링했다. 노드 누적 빈도 커버리지 비율 98%를 만족하는 최소 임계값 23회를 적용했다.[1]

이어서 '본초'를 노드로 삼고 처방에서 함께 등장하는 '본초 간의 관계'를 엣지로 삼아 아래와 같이 두 네트워크를 구축했다.

[1] 자세한 산출 과정은 제3장 '희소 데이터 처리'를 참고하라.

◎ ‹단순 빈도 네트워크›

데이터 가공을 최소화한 네트워크이다. 엣지 가중치는 '동시 출현 빈도'이다. 엣지 누적 빈도 커버리지 비율 90%를 기준으로 동시 출현 빈도 8회 미만인 엣지는 제거했다.[2] 단, 노드 필터링 이후에 엣지를 필터링했으므로 전체 엣지 비율에는 약간의 차이가 있다. 엣지 가중치(동시 출현 빈도)는 그대로 사용하지 않고 $\log_2$ 스케일을 적용했다.

> **i** 동시 출현 빈도의 $\log_2$ 스케일 변환
>
> 자연어 텍스트와 같은 복잡계 데이터는 롱테일 분포인 '지프의 법칙(Zipf's Law)'을 따른다. 한의학 처방 데이터 역시 제3장에서 살펴보았듯이 본초 배합(엣지)에서 이러한 현상을 관찰할 수 있었다. 동시 출현 빈도를 그대로 가중치로 사용하면 일부 상위 배합의 수치가 지나치게 커서 전체 네트워크 계산이나 레이아웃 시각화 결과를 크게 왜곡할 수 있다. 이러한 편향을 줄이고자 ‹단순 빈도 네트워크›의 엣지 가중치로 동시 출현 빈도를 사용할 때 $\log_2$ 변환을 적용했다. 밑이 2인 로그 변환은 수치의 크기 순위를 보존하면서도 고빈도 편향을 완화해 준다.

◎ ‹G-score 네트워크›

통계적으로 가공한 네트워크이다. 엣지 가중치는 'G-score'이다. G-score를 바탕으로 통계적 유의수준($p < 0.01$) 검증을 통과하지 못한 엣지를 제거했다.

두 네트워크의 모델링 방식을 정리하면 아래 표와 같다.

Table 5.1 두 네트워크의 기초 지표 비교

명칭	노드	엣지	엣지 가중치	필터링: 노드	필터링: 엣지 가중치	노드 수 (본초)	엣지 수 (배합)	네트워크 밀도
‹단순 빈도 네트워크›	본초	본초 배합	동시 출현 빈도	23회 이상 출현	8회 이상	563 종	22,550 개	0.1425
‹G-score 네트워크›	본초	본초 배합	G-score	23회 이상 출현	6.635 이상	580 종	14,778 개	0.088

[2] 엣지 가중치 임계값 산출 과정 역시 제3장의 '희소 데이터 처리'를 참고하라.

두 네트워크 사이에서 '네트워크 밀도' 차이가 두드러진다. **네트워크 밀도(Density)**는 가능한 모든 연결 경우의 수 대비 실제 연결된 엣지의 비율을 의미한다. ‹단순 빈도 네트워크›의 밀도가 0.1425이고 ‹G-score 네트워크›의 밀도가 0.0880으로 나타났다. 전자에 비해 후자가 더 성글다는 의미이다.

네트워크 밀도 값은 구조적 차이를 나타낼 뿐 네트워크의 품질을 뜻하지는 않는다. 밀도가 낮으면 네트워크의 핵심 골격을 잘 유지했을 수도 있지만 현실 네트워크를 과도하게 단순화했을 가능성도 있다. 반대로 밀도가 높으면 현실을 풍부하게 반영했을 수 있으나 분석에 불필요한 노이즈가 섞였을 수 있다. 네트워크를 구성하는 방법은 다양하며 분석 목적에 알맞은 방식을 선택할 따름이다. 여기서 두 네트워크를 함께 분석하고 비교하는 이유도 이런 구조적 차이를 확인하기 위해서다.

5.1 무작위 모델

실제 네트워크의 위상 구조가 통계적으로 유의미한지 판별하기 위해 무작위 모델을 대조군으로 활용한다. **무작위 모델(Null Model)**은 실제 네트워크에서 관찰된 특정 조건을 고정한 채 나머지 연결 관계를 확률적으로 무작위화한 네트워크로 '귀무 모델'이나 '널(null) 모델'이라고도 부른다. 무작위 모델과 비교하여 실제 네트워크에서 관찰된 패턴이 단순한 확률적 우연이 아님을 통계적으로 검증할 수 있다.

무작위 모델은 대상 네트워크의 정보를 어디까지 보존할 것인가에 따라 차수 기반 계층(dK-series)으로 분류할 수 있다(Orsini et al. 2015). 여기서 차수(Degree)란 한 노드에 직접 연결된 엣지의 수를 뜻한다. 차수 기반 계층(dK-series)은 차수 통계의 세밀한 정도를 단계화한 것으로 계수 d가 높아질수록 노드 개별 차수를 넘어 노드 쌍 사이의 차수 상관관계나 삼각 모티프 같은 더 고차원적인 연결 구조 정보를 보존한다.

◎ 0차 무작위 모델 ($0K$: 전역 규모 보존)

전체 노드 수와 엣지 수(평균 연결 차수)만 고정하고 모든 연결을 무작위로 생성하여 전역(Global) 규모를 보존하는 모델이다. 에르되시-레니(Erdős-Rényi, ER) 모델이 가장 대표적이다(Erdős and Rényi 1959). ER 모델은 고정된 엣지 수 m개를 N개의 노드 사이에 무작위로 균등 배치하는 $G(N, m)$ 모델과 가능한 모든 $\frac{N(N-1)}{2}$개의 노드 쌍이 각각 독립적인 연결 확률 p를 갖는 $G(N, p)$ 모델로 구분된다.

$$G(N, m), \quad G(N, p)$$

(N: 전체 노드 수. m: 고정된 엣지 수. p: 임의의 두 노드 쌍이 독립적으로 연결될 확률)

현실 네트워크에는 차수 불균형, 즉 빈도 편차가 심하여 중심 거점(Hub) 역할을 하는 노드가 존재한다. 그러나 0차 무작위 모델은 모든 노드가 균등한 연결 확률을 갖는다고 가정하므로 이러한 특성을 반영하지 못한다.

◎ 1차 무작위 모델 ($1K$: 개별 차수 보존)

네트워크 전역 규모뿐 아니라 개별 노드의 연결 차수(k_i)와 전체 차수 분포 $P(k)$를 실제 데이터와 동일하게 보존하는 모델이다. 노드 간 차수 차이로 생기는 통계적 착시를 보정한 모델이다.

1차 무작위 모델을 구현하는 대표적인 알고리즘은 두 가지다. 첫 번째는 배치 모델(Configuration Model)이다(Molloy and Reed 1995). 각 노드에 실제 연결 차수만큼 가상의 연결 고리(스텁, Stub)를 부여한 뒤 스텁 쌍을 무작위로 결합하는 생성형 방식이다. 수학적 처리가 단순하지만 자가 루프(Self-loop)와 다중 엣지(Multi-edge)가 발생할 수 있다. 두 번째는 이중 엣지 교환(Double Edge Swap) 알고리즘이다(Maslov and Sneppen 2002). 기존 네트워크에서 임의의

두 엣지 (u, v)와 (x, y)를 선택하여 연결 대상을 (u, y)와 (x, v)로 맞바꾸는 과정을 반복하는 재배선 방식이다. 단순 그래프 조건을 만족하면서 각 노드의 차수를 정확하게 보존하므로 실증 네트워크 분석에서 표준적으로 사용한다.

◎ **고차 무작위 모델 ($2K$ 이상: 차수 상관관계 및 군집 계수 보존)**

2차 무작위 모델($2K$)은 개별 차수 분포에 더해 노드 쌍 사이의 차수 상관관계(Degree Correlation)[3]까지 보존하여 차수 간 연결 성향을 실제 네트워크와 동일하게 통제하고자 할 때 사용한다. 3차 무작위 모델($3K$)은 삼각 모티프(닫힌 삼각형 구조) 개수까지 보존하여 국소(Local) 군집 계수를 통제한다.

보존하는 조건의 차수가 높아지면 무작위 생성 네트워크의 자유도가 급격히 줄어들어 실제 네트워크와 거의 같아지는 한계가 있다. 따라서 유의성을 평가하는 귀무 모델로는 개별 차수만 보존하는 1차 무작위 모델(1K)을 가장 많이 활용한다.

이 장에서는 1차 무작위 모델 중 **이중 엣지 교환(Double Edge Swap)** 알고리즘을 사용했다. 제3장에서 살펴보았듯이 본초 네트워크에서 노드와 엣지는 모두 두터운 꼬리를 갖는 헤비테일(Heavy-tailed) 분포 양상을 보인다. 따라서 개별 본초의 엣지 수를 고정해야 출현 빈도 차이로 생기는 통계적 착시에서 벗어날 수 있다.

실제 네트워크(‹단순 빈도 네트워크›와 ‹G-score 네트워크›) 각각에 대해 본초별 출현 빈도 편차를 통제한 무작위 모델을 100회 반복 생성하고 산출된 위상 지표의 평균값을 실제 네트워크와 비교하여 처방 배합 구조의 통계적 유의성을 검증해 본다.

---

[3] 차수 상관관계(Degree Correlation, $P(k, k')$): 엣지로 연결된 두 노드의 차수 사이에서 나타나는 통계적 연관성을 뜻한다. 차수가 높은 중심 거점(Hub) 노드가 다른 중심 노드와 주로 연결되는 경향(동류성)이나 차수가 낮은 노드와 결합하는 경향(이지류성)을 가리킨다.

5.2 좁은 세상 구조

복잡계 시스템은 공통으로 '좁은 세상 구조'라는 구조적 특성을 나타낸다. **좁은 세상 구조(Small-world Structure)**는 높은 군집 계수와 짧은 평균 최단 경로 길이를 동시에 만족하는 위상 구조를 가리킨다(Watts and Strogatz 1998; Newman 2018).

◎ 높은 군집 계수

군집 계수(Clustering Coefficient)는 한 노드의 이웃 노드들이 서로 연결되어 있을 확률을 나타내며 노드들이 국소적(Local)으로 얼마나 조밀하게 무리를 이루는지 보여주는 지표다. 국소 군집 계수는 대상 노드의 이웃들 사이에서 형성될 수 있는 이론적 최대 엣지 수($\frac{k_i(k_i-1)}{2}$) 대비 실제 연결된 엣지 수(E_i)의 비율을 뜻한다. 네트워크 분석에서는 개별 노드의 국소 군집 계수와 이를 전체 노드에 걸쳐 평균한 평균 군집 계수를 함께 활용한다. 한의학 처방에서는 특정 효능을 위해 본초들이 약대(藥對)나 기본방의 핵심 조합 형태로 소그룹을 형성하므로 군집 계수가 높게 나타날 가능성이 크다.

$$C_i = \frac{2E_i}{k_i(k_i-1)}, \quad C = \frac{1}{N}\sum_{i=1}^{N} C_i$$

(C_i: 노드 v_i의 국소 군집 계수. E_i: 노드 v_i의 이웃 노드들 사이에 존재하는 실제 엣지 수. k_i: 노드 v_i의 연결 차수. C: 전체 노드의 평균 군집 계수. N: 전체 노드 수)

◎ 짧은 평균 경로 길이

평균 최단 경로 길이(Average Path Length)는 네트워크 내 임의의 두 노드를 잇는 최단 경로 단계의 평균값이다. 수학적으로는 네트워크에 존재하는 모든 서로 다른 노드 순서쌍 사이의 최단

경로 거리 $d(v_i, v_j)$의 총합을 전체 순서쌍의 수($N(N-1)$)로 나누어 계산한다. 네트워크 전체의 정보 전달과 연결 효율을 평가하는 전역(Global) 지표다. 네트워크 내 노드 간 상호작용이 신속하고 유기적으로 이루어질수록 이 값은 작아진다.

$$L = \frac{1}{N(N-1)} \sum_{i \neq j} d(v_i, v_j)$$

(N: 전체 노드 수. $d(v_i, v_j)$: 노드 v_i와 v_j 사이의 최단 경로 거리)

두 지표는 동시에 만족하기 어려운 특성을 지닌다. 규칙적인 격자 형태의 '정규 네트워크(Regular Network)'와 임의로 선을 연결한 '무작위 네트워크(Random Network)'라는 양극단의 두 네트워크를 떠올려 보자. 전자는 이웃끼리 잘 뭉쳐있어 군집 계수가 크지만 멀리 있는 노드와 소통하기 어려워 평균 경로 길이가 길다. 반면 후자는 소그룹을 형성하지 않아 군집 계수는 작지만 지름길이 있어 평균 경로 길이가 짧다. 좁은 세상 구조는 노드끼리 조밀하게 소그룹을 형성하면서도 소그룹 사이를 잇는 지름길(Shortcut) 노드가 존재해 높은 군집 계수와 짧은 평균 경로 길이를 동시에 나타낸다.

무작위 모델은 주로 무작위 네트워크를 사용한다. 따라서 실제 네트워크를 무작위 모델과 비교했을 때 실제 네트워크가 좁은 세상 구조를 띤다면 평균 경로 길이는 양자 모두 짧게 나타나지만 군집 계수는 실제 모델에서 훨씬 크게 나타날 것으로 예상할 수 있다. 만약 한의학 본초 네트워크가 이 좁은 세상 구조를 띤다면 이는 한의학 처방이 본초를 임의로 뒤섞은 무작위 배합이 아니라는 점을 통계적·위상학적으로 뒷받침하는 셈이다.

실제 처방 데이터가 이러한 좁은 세상 구조를 갖추었는지 수치로 확인해 본다.

> **위상 지표와 엣지 가중치**
>
> 이 장에서 다루는 군집 계수, 평균 경로 길이, 좁은 세상 지수, 삼각 모티프 등은 엣지의 가중치 크기를 쓰지 않고 엣지가 이어져 있는지 여부만으로 계산하는 지표다. 동시 출현 빈도나 G-score 같은 엣지 가중치는 어떤 엣지를 분석 대상에 포함할지 결정하는 기준으로만 쓰였다. 일단 네트워크가 구축되어 위상 구조가 결정되면 실제 네트워크와 무작위 모델 모두 엣지 가중치 수치가 계산에 반영되지 않는다. 모든 엣지를 동등한 단순 엣지로 다루어 네트워크의 순수한 형태와 결합 구조를 계산하고 비교한다.

Table 5.2 실제 네트워크와 무작위 모델의 군집 계수 비교

구분	실제 군집 계수	무작위 모델 군집 계수	비율
‹단순 빈도 네트워크›	0.771	0.75	1.027
‹G-score 네트워크›	0.382	0.179	2.136

Table 5.3 실제 네트워크와 무작위 모델의 평균 경로 길이 비교

구분	실제 평균 경로 길이	무작위 모델 평균 경로 길이	비율
‹단순 빈도 네트워크›	1.901	1.92	0.99
‹G-score 네트워크›	2.112	2.011	1.05

두 네트워크 모두 평균 경로 길이는 무작위 모델과 비교했을 때 큰 차이 없이 짧았다. ‹단순 빈도 네트워크›의 평균 경로 길이는 1.90이며, ‹G-score 네트워크›는 2.11이다. 임의의 한 본초에서 다른 본초에 도달하기까지 평균 2단계 안팎만 거치면 연결된다는 의미이다.

군집 계수는 두 네트워크에서 다른 양상을 보였다. ‹단순 빈도 네트워크›의 군집 계수는 0.771로 무작위 모델(0.750)과 비슷한 수준이다. 이에 비해 ‹G-score 네트워크›의 군집 계수는 0.382로 무작위 모델(0.179)보다 2.1배 높게 나타났다.

네트워크가 나타내는 좁은 세상 특성은 군집 계수와 평균 경로 길이를 바탕으로 단일 수치로 정량화할 수 있다. 무작위 모델 대비 군집 계수의 비율인 γ와 평균 경로 길이의 비율인 λ를 결합하여 **좁은 세상 지수(Small-Worldness Index, σ)**를 계산해 보자(Humphries and Gurney 2008).

γ는 실제 군집도가 무작위 상태보다 몇 배 증가되었는지를 나타내며, λ는 평균 최단 경로 길이가 무작위 상태보다 몇 배 늘어났는지를 나타낸다. 따라서 σ는 경로 길이 증가율 대비 군집도 증가 비율을 의미한다.

$$\gamma = \frac{C}{C_{\text{null}}}, \quad \lambda = \frac{L}{L_{\text{null}}}, \quad \sigma = \frac{\gamma}{\lambda} = \frac{C/C_{\text{null}}}{L/L_{\text{null}}}$$

(C, L: 실제 네트워크의 평균 군집 계수 및 평균 경로 길이. C_{null}, L_{null}: 동일 규모 무작위 모델의 기댓값)

수학적으로 $\sigma > 1$을 만족할 때 해당 네트워크는 유의미한 좁은 세상 구조를 지닌 것으로 판정한다. 비율 γ가 1.0보다 매우 크면서 λ가 1.0에 근접할수록 σ는 급격히 커진다.

Table 5.4 실제 네트워크와 무작위 모델의 좁은 세상 지표 비교

구분	군집 계수 비율 (γ)	경로 길이 비율 (λ)	좁은 세상 지수 (σ)
‹단순 빈도 네트워크›	1.027 배	0.990 배	1.038
‹G-score 네트워크›	2.136 배	1.050 배	2.035

각 본초의 연결 차수를 고정한 무작위 모델을 기준으로 비교했을 때, ‹단순 빈도 네트워크›의 좁은 세상 지수 σ는 1.038이며, ‹G-score 네트워크›는 2.035로 나타났다.

‹G-score 네트워크›의 좁은 세상 지수(σ)는 2.035로 판정 기준치인 1.0을 크게 웃돌며 좁은 세상 구조를 뚜렷하게 나타냈다. 이 결과는 한의학 처방의 구조적 특성으로 풀이할 수 있다. 한의학 처방에서 본초들이 이루는 기본방 구성이나 약대(藥對) 배합은 국소 군집을 형성하고(높은 군집 계수), 폭넓은 배합 관계를 갖는 다빈도 핵심 거점(Hub) 본초들이 이들 군집 사이를 연결하여 전역 연결성을 높이는(짧은 평균 경로 길이) 구조로 이해할 수 있다.

그러나 ‹단순 빈도 네트워크›의 좁은 세상 지수(σ)는 1.038로 무작위 모델과 차이가 거의 없었다. 이는 개별 본초의 연결 차수를 보존한 상태에서 연결을 바꾸었을 때 네트워크 위상에 유의미한 차이가 없다는 뜻이다. 다시 말해 ‹단순 빈도 네트워크›에서 나타난 높은 군집 계수는 배합의 고유한 규칙이라기보다 자주 사용되는 본초들이 많은 노드와 연결되면서 자연스럽게 형성된 고빈도 효과에 가깝다.

5.3 모티프 분석

거시적인 구조를 확인했으므로 네트워크 안에 반복적으로 나타나는 미세 패턴을 분석하기 위해 '모티프'를 탐색해 보자. **모티프(Motif)**는 복잡한 네트워크를 이루는 국소적 기본 단위 서브네트워크(Subnetwork)를 의미한다(Milo et al. 2002). 세 노드가 서로 모두 연결된 닫힌 삼각형 구조인 '삼각 모티프(3-Node Triad Motif)'가 대표적이다. A와 B, B와 C가 함께 쓰일 때 A와 C도 함께 배합되는 닫힌 결합 구조는 무작위 상황에서는 흔히 발생하지 않기 때문이다.

네트워크 분석에서는 무작위 모델(Null Model) 대비 특정 모티프의 출현 빈도가 통계적으로 얼마나 유의미한지를 Z-점수(Z-score)로 정량화한다. Z-점수는 실제 관찰된 모티프 출현 횟수 $N_{\text{obs},i}$가 무작위 대조군의 평균 기댓값 $\langle N_{\text{null},i} \rangle$으로부터 표준편차($\sigma_{\text{null},i}$)의 몇 배만큼 벗어나 있는지를 나타내는 표준화 통계량이다. $Z_i > 0$이면 실제 네트워크에서 무작위 기댓값보다 빈번하게 나타남을 뜻하며, 통상 $Z > 2$ 또는 $Z > 3$ 이상일 때 통계적으로 유의미하다고 판정한다.

$$Z_i = \frac{N_{\text{obs},i} - \langle N_{\text{null},i} \rangle}{\sigma_{\text{null},i}}$$

($N_{obs,i}$: 실제 네트워크의 모티프 관측값. $\langle N_{null,i} \rangle$: 무작위 대조군의 모티프 평균 기댓값. $\sigma_{null,i}$:

무작위 대조군의 표준편차)

Table 5.5 무작위 모델 대비 삼각 모티프

구분	실제 네트워크 (관측값)	무작위 모델 (기댓값)	Z-점수 (Z-score)
‹단순 빈도 네트워크›	647,857 개	601,350.7 개	46.7
‹G-score 네트워크›	121,928 개	60,106.1 개	176.4

무작위 모델과의 비교 결과, ‹단순 빈도 네트워크›의 Z-점수는 46.7이며, ‹G-score 네트워크›의 Z-점수는 176.4로 나타났다. 두 네트워크 모두 통계적 유의성 판정 기준($Z > 3$)을 크게 초과하므로 삼각 모티프가 무작위 우연이 아니라 통계적으로 유의미한 배합 관계에 따라 형성되었음을 보여준다.

앞서 살펴본 좁은 세상 지수(σ)에서 ‹단순 빈도 네트워크›는 무작위 모델과 뚜렷이 구별되지 않았지만, 삼각 모티프 분석에서는 두 네트워크 모두 무작위 모델(기댓값)을 크게 웃도는 높은 Z-점수를 나타냈다.

‹단순 빈도 네트워크›의 경우 고빈도 본초의 엣지가 워낙 많아 무작위로 재배선하더라도 무작위 모델 자체에서 약 60만 개의 삼각형이 생성되므로 군집 계수 평균 비율($\gamma \approx 1.03$)상으로는 큰 차이가 드러나지 않았다. 그러나 삼각 모티프의 총량을 비교하면 실제 네트워크가 무작위 모델 기댓값보다 약 4.7만 개(7.7%) 더 많은 삼각형을 형성하고 있으며, 시뮬레이션의 작은 편차로 인해 통계적으로 유의미한 수준($Z = 46.7$)임이 확인된다. 나아가 ‹G-score 네트워크›에서는 삼각 모티프 수가 무작위 기댓값의 2배 이상($Z = 176.4$)에 달해, 빈도 편향을 통제했을 때 처방 고유의 삼각 배합 구조가 더욱 뚜렷하게 보존되었다.

이처럼 1차 무작위 모델과 대조해 보았을 때 한의학 처방 네트워크에 나타나는 삼각 배합 구조가 단순한 차수 효과나 우연한 연결로 형성되었을 가능성이 매우 낮다는 점을 통계적으로 확인할 수 있다.

i 창발성(Emergence)

창발성은 하위 구성 요소들이 상호작용하는 과정에서 시스템 전체의 새로운 특성이 나타나는 현상을 뜻한다. 이는 개별 요소의 속성이나 단순한 합산만으로는 설명하거나 예측하기 어렵다. 요소들이 서로 관계를 맺으면서 개별 차원에는 없던 거시적 패턴과 기능이 자발적으로 형성된다.
한의학 처방 체계 역시 이러한 창발적 특성을 잘 보여준다. 한의학 처방에서 의가들은 본초 조합의 거대한 구조(전역 네트워크)를 미리 설계하고 이를 전제로 처방을 만들어 사용한 것이 아니다. 환자의 증상에 맞춰 몇 가지 본초를 조합하는 국소 배합 규칙을 개발했을 뿐이다. 그러나 이러한 국소 배합들이 오랜 기간 처방 문헌을 통해 축적되면서 전체 네트워크 차원에서는 무작위 연결과 구별되는 고유한 위상 패턴을 형성했다.

5.4 중심부-주변부 계층 구조

마지막으로 네트워크 전체의 형상을 보여주는 거시적 분해 기법인 'k-core 분해' 알고리즘을 적용해 보자. **k-core 분해(k-core decomposition)** 알고리즘은 연결 수가 적은 주변부 노드부터 차례대로 제거하여 가장 깊은 곳에서 뭉쳐 있는 핵심 덩어리(Maximum k-core)를 찾아내는 기법이다(Seidman 1983).

구체적으로 k-core는 모든 노드가 해당 서브네트워크 내에서 최소 k개의 연결(차수, Degree)을 유지하는 최대 유도 서브네트워크(Maximal Induced Subnetwork)를 뜻한다. k-core 분해는 연결 차수가 1 미만인 고립 노드(차수 0)들을 먼저 제거하고, 잇따라 $k = 2, k = 3$으로 기준을 높여가며 서브네트워크 내 차수가 k 미만으로 떨어지는 노드들을 반복적으로 가지치기(Pruning)하는 방식으로 진행된다. 노드 v가 제거되지 않고 끝까지 남는 최고 k 값을 해당 노드의 코어니스(Coreness, $k(v)$)라고 한다.

이러한 과정을 반복하여 도달하는 가장 큰 k 값을 $k_{\max}$라고 한다. 이 $k_{\max}$ 기준을 만족하며 끝까지 남아 있는 노드 집합이 곧 네트워크의 가장 안쪽 중심부인 '최대 k-core(Maximum k-core)'이다.

이 기법을 활용하면 본초 네트워크에서 중심부와 주변부가 얼마나 뚜렷하게 계층화되는지, 서로 촘촘하게 얽히며 처방의 중심축을 이루는 '핵심 본초군'이 무엇인지 직관적으로 가려낼 수 있다.

Table 5.6 중심부-주변부 계층 구조 요약

구분	계층 수(k max)	최대 k-core 본초 수
‹단순 빈도 네트워크›	91 겹	136 종
‹G-score 네트워크›	37 겹	172 종

분석 결과, 두 네트워크 모두 계층 구조를 보였고 가장 깊은 중심부에는 비교적 많은 본초가 모여 있었다. 중심부와 주변부가 실제로 어떻게 분포되어 있는지, 각 네트워크의 계층별 본초 개수를 집계해 살펴보자.

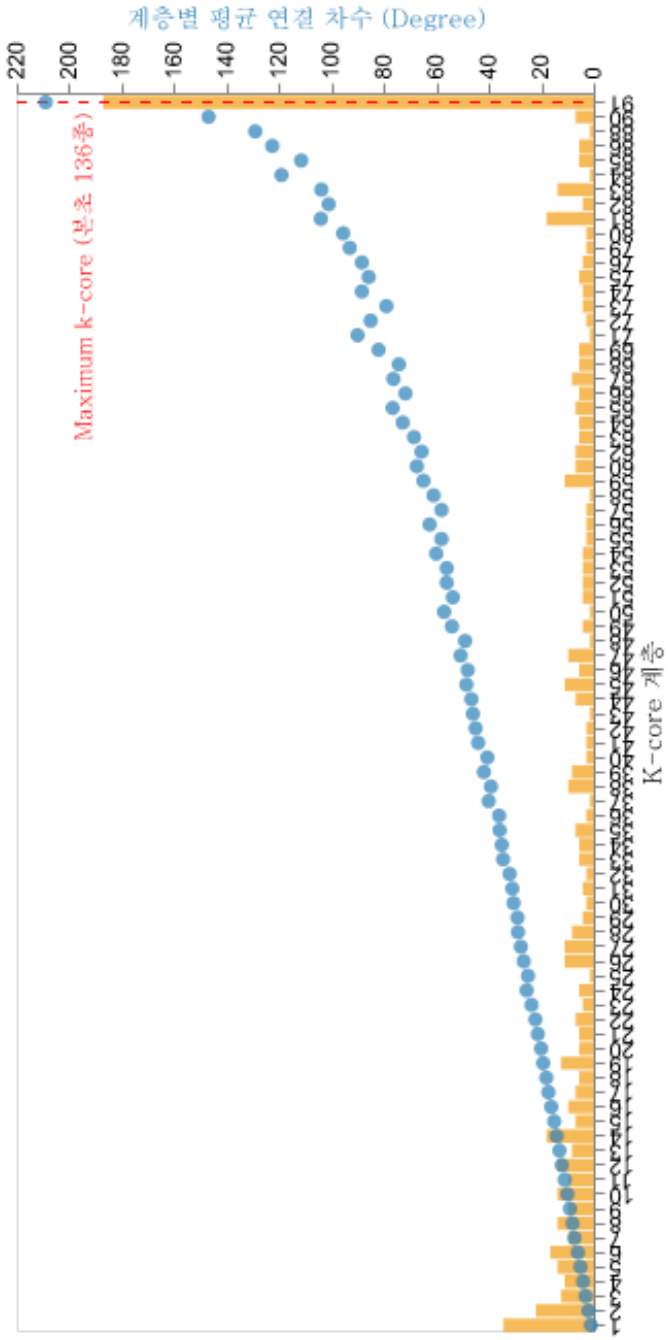

Figure 5.1 계층별 본초 개수 및 평균 연결 차수 분포 | ‹단순 빈도 네트워크›

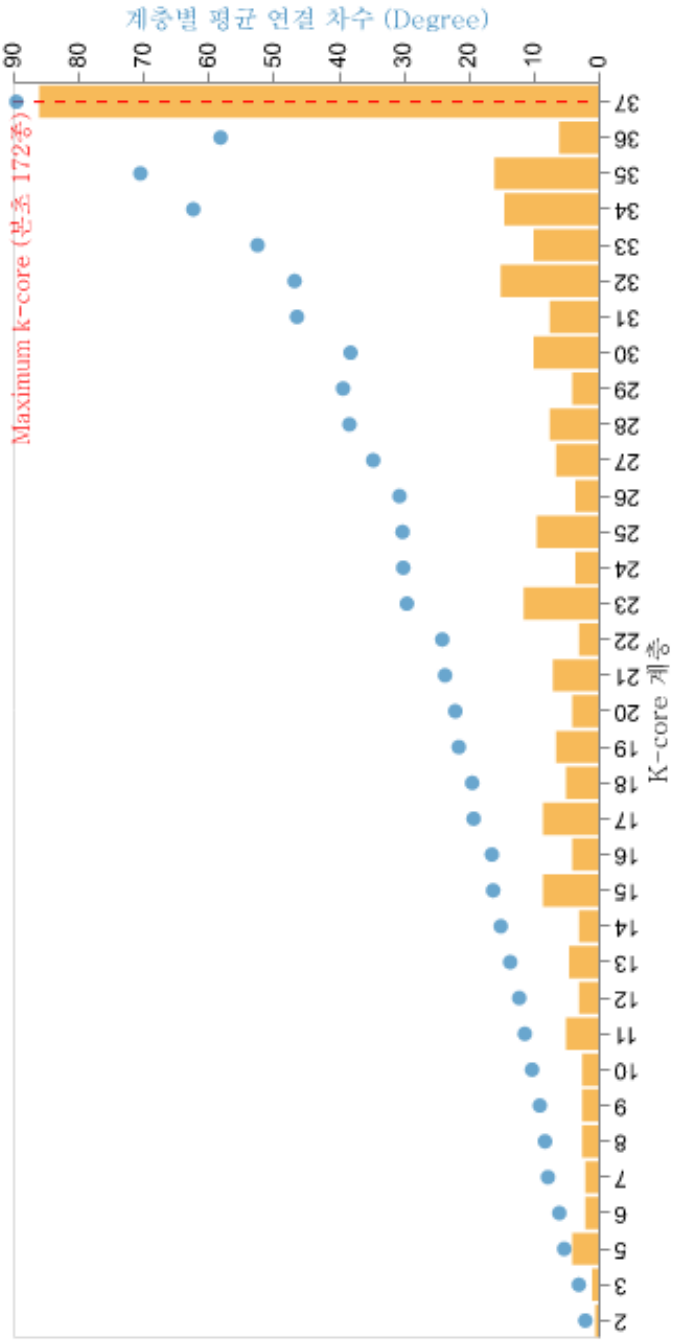

Figure 5.2 계층별 본초 개수 및 평균 연결 차수 분포 | ‹G-score 네트워크›

(그래프의 주황색 막대는 각 계층에 속한 본초의 개수(왼쪽 축)를 나타내고, 파란색 점은 해당 계층 본초들의 평균 연결 차수(오른쪽 축)를 나타낸다.)

결과를 보면 두 네트워크 모두에서 유사한 형태의 계층 구조를 관찰할 수 있다. 바깥쪽 주변부에는 계층마다 적은 수의 본초가 위치하며, 마지막 가장 안쪽 중심부에는 다수의 핵심 본초가 강하게 결속되어 있다. 또한 계층이 깊어질수록 본초들의 평균 연결 차수 역시 상승하여 최대 k-core에 도달했을 때 연결 차수가 급격히 높아지는 양상을 보인다. 다만 ‹G-score 네트워크›의 계층 수가 ‹단순 빈도 네트워크›보다 적은 이유는, ‹단순 빈도 네트워크›의 엣지 수가 더 많아 고차수 연결이 상대적으로 많이 유지되기 때문이다.

이러한 모습은 어떻게 이해할 수 있을까? 가장 안쪽 중심부 본초는 여러 처방에서 기본방(基本方)을 구성하는 핵심 약재로서 처방의 기본 골격을 형성하고, 여러 계층으로 나뉜 주변부 본초는 병증에 따라 가감(加減)되며 처방의 효능과 성격을 변주하는 기능을 맡는 것으로 해석할 수 있다.

가장 안쪽 중심부와 가장 바깥쪽 주변부를 차지하는 본초가 무엇인지 구체적으로 비교해 보자.

Table 5.7 중심부와 주변부 본초 (상위 20개) | ‹단순 빈도 네트워크›

k=1	k=2	k=3	…	k=89	k=90	k=91
석회(石灰)	염(鹽)	백면(白麵)	…	—	빙편(冰片)	감초(甘草)
대산(大蒜)	운태자(芸苔子)	이어(鯉魚)	…	—	파두(巴豆)	복령(茯苓)
마치현(馬齒莧)	묵한련(墨旱蓮)	능소화(凌霄花)	…	—	보골지(補骨脂)	당귀(當歸)
다엽(茶葉)	마란자(馬蘭子)	장미(薔薇)	…	—	초오(草烏)	인삼(人蔘)
백매화(白梅花)	절패모(浙貝母)	면분(麵粉)	…	—	위령선(威靈仙)	작약(芍藥)
초(醋)	희렴초(豨薟草)	저담(豬膽)	…	—	—	백출(白朮)
내복(萊菔)	납(蠟)	인유(人乳)	…	—	—	천궁(川芎)
다(茶)	연지화(胭脂花)	산인진(山茵陳)	…	—	—	진피(陳皮)
문합(文蛤)	호동루(胡桐淚)	합등자(榼藤子)	…	—	—	황금(黃芩)
정천석(井泉石)	인중황(人中黃)	—	…	—	—	방풍(防風)

k=1	k=2	k=3	…	k=89	k=90	k=91
창이초 (蒼耳草)	도교(桃膠)	—	…	—	—	육계(肉桂)
백백피 (柏白皮)	홍설(紅雪)	—	…	—	—	목향(木香)
홍곡(紅麴)	발제(荸薺)	—	…	—	—	대황(大黃)
구채(韭菜)	삼칠근 (三七根)	—	…	—	—	반하(半夏)
만청(蔓菁)	출미(秫米)	—	…	—	—	건강(乾薑)
도피(桃皮)	황반(黃礬)	—	…	—	—	황련(黃連)
화예석 (花蕊石)	—	—	…	—	—	부자(附子)
종려피 (棕櫚皮)	—	—	…	—	—	생지황 (生地黃)
비(砒)	—	—	…	—	—	맥문동 (麥門冬)
불회목 (不灰木)	—	—	…	—	—	황기(黃芪)

Table 5.8 중심부와 주변부 본초 (상위 20개) | ‹G-score 네트워크›

k=1	k=2	k=3	…	k=35	k=36	k=37
—	백염(白鹽)	양육(羊肉)	…	사향(麝香)	상백피(桑白皮)	감초(甘草)
—	—	호로(葫蘆)	…	주사(硃砂)	패모(貝母)	복령(茯苓)
—	—	—	…	유향(乳香)	고량강(高良薑)	당귀(當歸)
—	—	—	…	웅황(雄黃)	용담초(龍膽草)	인삼(人蔘)
—	—	—	…	천남성(天南星)	선복화(旋覆花)	작약(芍藥)
—	—	—	…	파두(巴豆)	자소자(紫蘇子)	백출(白朮)
—	—	—	…	아교(阿膠)	초과(草果)	천궁(川芎)
—	—	—	…	우황(牛黃)	목적(木賊)	진피(陳皮)
—	—	—	…	견우자(牽牛子)	결명자(決明子)	황금(黃芩)
—	—	—	…	조각(皂角)	청상자(靑葙子)	방풍(防風)
—	—	—	…	유황(硫黃)	검실(芡實)	육계(肉桂)
—	—	—	…	노사(磠砂)	지부자(地膚子)	목향(木香)
—	—	—	…	저령(豬苓)	—	대황(大黃)
—	—	—	…	오령지(五靈脂)	—	반하(半夏)
—	—	—	…	호박(琥珀)	—	건강(乾薑)
—	—	—	…	혈갈(血竭)	—	황련(黃連)
—	—	—	…	호초(胡椒)	—	부자(附子)
—	—	—	…	지룡(地龍)	—	생지황(生地黃)
—	—	—	…	내복자(萊菔子)	—	맥문동(麥門冬)
—	—	—	…	아위(阿魏)	—	황기(黃芪)

두 네트워크는 서로 다른 방식으로 모델링되었으므로 바깥쪽 계층 본초의 종류에서 차이를 보이는 것은 자연스럽다. 하지만 모델링 방식이 달라도 처방의 뼈대를 이루는 핵심 본초군은 두 네트워크 모두에서 일관되게 확인된다. 게다가 이 중심부 본초의 구성은 매우 비슷했다. 이는 한의학 처방

기반의 본초 네트워크가 처방의 골격을 이루는 중심부 핵심 본초군과 특정 병증에 맞춰 결합하는 주변부 본초군이 계층별로 분화된 위상 구조를 지니고 있음을 보여준다.

두 네트워크의 중심부 본초의 구성을 구체적으로 비교하면 다음과 같다.

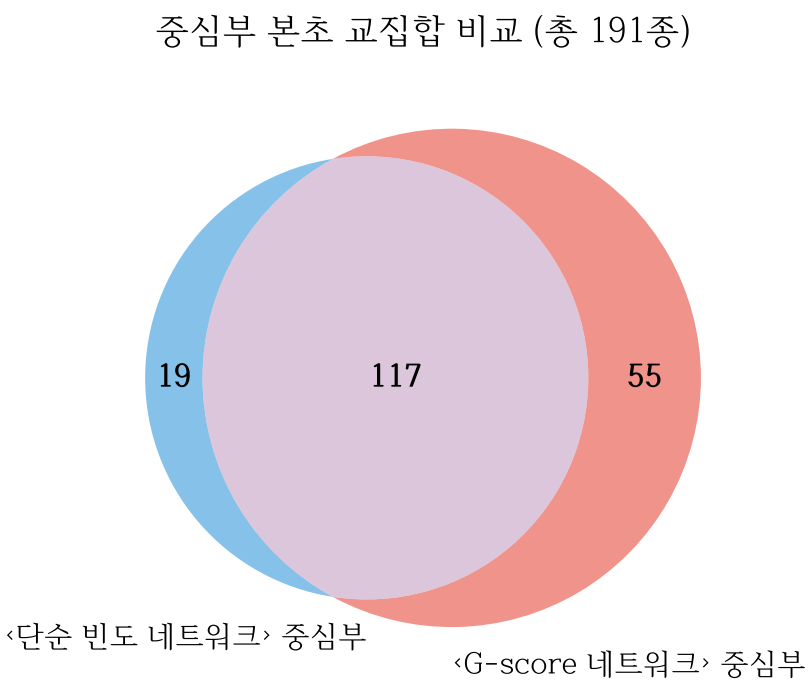

Figure 5.3 중심부(Maximum k-core) 본초 비교 벤 다이어그램

공통 본초: 117종

가자(訶子), 갈근(葛根), 감국(甘菊), 감초(甘草), 강활(羌活), 건강(乾薑), 계지(桂枝), 고본(藁本), 고삼(苦蔘), 곽향(藿香), 구기자(枸杞子), 귤홍(橘紅), 길경(桔梗), 단삼(丹蔘), 당귀(當歸), 대황(大黃), 도인(桃仁), 독활(獨活), 두중(杜仲), 마황(麻黃), 만형자(蔓荊子), 맥문동(麥門冬), 모과(木瓜), 모려(牡蠣), 목단피(牡丹皮), 목통(木通), 목향(木香), 몰약(沒藥), 박하(薄荷), 반하(半夏), 방기(防己), 방풍(防風), 백강잠(白僵蠶), 백부자(白附子), 백자인(柏子

仁), 백지(白芷), 백질려(白蒺藜), 백출(白朮), 별갑(鱉甲), 복령(茯苓), 봉아출(蓬莪朮), 부자(附子), 비해(萆薢), 빈랑(檳榔), 사인(砂仁), 산수유(山茱萸), 산약(山藥), 산조인(酸棗仁), 삼릉(三稜), 생지황(生地黃), 서우각(犀牛角), 석고(石膏), 석곡(石斛), 석창포(石菖蒲), 선태(蟬蛻), 세신(細辛), 소회향(小茴香), 속단(續斷), 숙지황(熟地黃), 승마(升麻), 시호(柴胡), 신곡(神麴), 연교(連翹), 연호색(延胡索), 영양각(羚羊角), 오가피(五加皮), 오두(烏頭), 오미자(五味子), 오수유(吳茱萸), 오약(烏藥), 용골(龍骨), 우방자(牛蒡子), 우슬(牛膝), 원지(遠志), 육계(肉桂), 육두구(肉豆蔻), 육종용(肉蓯蓉), 의이인(薏苡仁), 인삼(人蔘), 자소엽(紫蘇葉), 자완(紫菀), 작약(芍藥), 적석지(赤石脂), 전갈(全蠍), 전호(前胡), 정향(丁香), 지각(枳殼), 지골피(地骨皮), 지모(知母), 지실(枳實), 진교(秦艽), 진피(陳皮), 차전자(車前子), 창출(蒼朮), 천궁(川芎), 천동(天冬), 천마(天麻), 천웅(天雄), 천초(川椒), 천화분(天花粉), 청피(靑皮), 치자(梔子), 침향(沉香), 택사(澤瀉), 토사자(菟絲子), 파극천(巴戟天), 하수오(何首烏), 행인(杏仁), 향부(香附), 현삼(玄蔘), 형개(荊芥), 홍화(紅花), 황금(黃芩), 황기(黃芪), 황련(黃連), 황백(黃柏), 후박(厚朴)

‹단순 빈도 네트워크›에만 해당하는 본초: 19종

견우자(牽牛子), 대조(大棗), 망초(芒硝), 백반(白礬), 사향(麝香), 상백피(桑白皮), 생강(生薑), 아교(阿膠), 애엽(艾葉), 오매(烏梅), 용담초(龍膽草), 우황(牛黃), 웅황(雄黃), 유향(乳香), 조각(皂角), 주사(硃砂), 천남성(天南星), 패모(貝母), 활석(滑石)

‹G-score 네트워크›에만 해당하는 본초: 55종

강황(薑黃), 건칠(乾漆), 골쇄보(骨碎補), 구채자(韭菜子), 구척(狗脊), 권백(卷柏), 금은화(金銀花), 녹용(鹿茸), 대복피(大腹皮), 망초(莽草), 맥아(麥芽), 모형자(牡荊子), 무이(蕪荑), 백두구(白豆蔻), 백렴(白斂), 백미(白薇), 백석영(白石英), 백선피(白鮮皮), 백화사(白花蛇), 보골지(補骨脂), 복분자(覆盆子), 사삼(沙蔘), 사상자(蛇床子), 산사(山楂), 상표소(桑螵蛸),

서여(薯蕷), 석남엽(石楠葉), 석룡예(石龍芮), 소목(蘇木), 양척촉(羊躑躅), 연자(蓮子), 오초사(烏梢蛇), 우여량(禹餘糧), 위령선(威靈仙), 위유(葳蕤), 음양곽(淫羊藿), 익지인(益智仁), 인우(茵芋), 자석(磁石), 자석영(紫石英), 저실자(楮實子), 종유석(鍾乳石), 창이자(蒼耳子), 천련자(川楝子), 천우슬(川牛膝), 초두구(草豆蔻), 초오(草烏), 측자(側子), 택란(澤蘭), 팔각회향(八角茴香), 필징가(蓽澄茄), 해동피(海桐皮), 해풍등(海風藤), 호로파(葫蘆巴), 호마(胡麻)

제6장에서는 네트워크에서 영향력이 큰 핵심 노드가 무엇인지 살펴본다.

제 6 장

중심성과 강건성

제5장에서는 한의학 처방을 바탕으로 본초 네트워크를 구축하고 무작위 모델과 비교하였다. 본초 네트워크는 견고한 중심부와 여러 계층의 주변부로 이루어졌으며 그 속에는 우연으로 보기 어려운 구조적 규칙이 존재함을 확인하였다. 이제 시선을 네트워크 전체 위상에서 개별 노드로 옮겨보자.

네트워크 안에서 노드의 역할은 동등하지 않다. 주변 노드와 어떤 관계를 얼마나 맺고 있는가에 따라 각 노드는 네트워크에 저마다 다른 영향력을 미친다. 이를 파악하기 위해 '중심성'이라는 개념을 사용한다. 네트워크 분석에서 **중심성(Centrality)**이란 그래프 안에서 특정 노드가 전체 네트워크에서 행사하는 영향력을 정량적으로 평가하는 지표이다(Freeman 1978; Newman 2018). 중심성을 측정하면 정보를 통제하는 병목 지점, 집단을 이끄는 중심 노드, 외곽에 머무는 주변 노드 등을 구분할 수 있다.

이 장에서는 네트워크 분석의 핵심 지표인 중심성(Centrality)으로 노드의 영향력을 정량적으로 측정한다. 아울러 이러한 중심 노드가 네트워크 구조 유지에 미치는 영향을 '강건성(Robustness)' 시뮬레이션으로 검증한다.

6.1 중심성 지표

노드의 중심성은 분석 목적에 따라 다양하게 정의된다. 가장 기본적이고 널리 쓰이는 방법은 다음과 같다.

◎ **차수 중심성(Degree Centrality)**

한 노드가 다른 노드와 직접 맺은 엣지의 개수를 측정한다. 소셜 네트워크에서 '친구 수'가 가장 많은 대상을 찾는 것과 유사하다. 이 지표가 높은 노드는 네트워크에서 직접적인 교류 범위가 가장 넓다. 수학적으로는 대상 노드가 가질 수 있는 이론상 최대 연결 수($N-1$) 대비 실제 연결 차수(k_v)의 비율을 뜻하며 0과 1 사이의 값을 가진다.

$$C_D(v) = \frac{k_v}{N-1}$$

(k_v: 노드 v의 연결 차수(엣지 수). N: 전체 노드 수)

엣지 가중치가 있는 네트워크에서는 단순히 연결된 이웃 수를 세는 대신 자신과 연결된 모든 엣지의 가중치를 합산한 **노드 강도(Node Strength)**를 사용할 수 있다. 연결된 이웃 수가 적더라도 각 엣지의 가중치가 크다면 더 높은 강도를 갖는다.

$$s_v = \sum_{u \in \mathcal{N}(v)} w_{vu}$$

(s_v: 노드 v의 노드 강도. $\mathcal{N}(v)$: 노드 v의 이웃 노드 집합. w_{vu}: 노드 v와 u 사이의 엣지 가중치)

◎ **근접 중심성(Closeness Centrality)**

한 노드가 네트워크의 다른 모든 노드까지 얼마나 가까운지를 측정한다. 소셜 네트워크에서 '모든 사람에게 가장 빠르게 연락이 닿는 대상'을 찾는 것과 같다. 이 지표가 높은 노드는 물류 센터, 환승 터미널, 정보 전파 허브(Hub)처럼 정보나 자원을 네트워크 전체에 가장 신속하게 전달하는 핵심 중계 거점 역할을 한다. 수학적으로는 대상 노드에서 다른 모든 노드에 이르는 평균 최단 경로 거리의 역수이다. 모든 노드와 1단계 거리로 직접 연결될 때 최댓값 1을 나타내며 평균 도달 거리가 멀어질수록 0에 수렴한다.

$$C_C(v) = \frac{N-1}{\sum_{u \neq v} d(v,u)}$$

(N: 전체 노드 수. $d(v,u)$: 노드 v와 u 사이의 최단 경로 거리)

엣지 가중치가 있는 경우에는 최단 경로 $d(v,u)$를 계산할 때 엣지 가중치를 '경로 비용(거리)'으로 반영한다. 엣지 가중치가 결합 강도나 빈도(유사도)를 나타낼 때는 이를 거리로 변환한 후 최단 경로 알고리즘을 적용하여 거리를 합산한다.

◎ **매개 중심성(Betweenness Centrality)**

네트워크의 다른 노드 쌍 사이 최단 경로에 해당 노드가 얼마나 자주 놓이는지 측정한다. 소셜 네트워크에서 '서로 다른 집단을 이어 주는 연결 다리(Bridge)' 역할을 하는 대상을 찾는 것과 유사하다. 이 지표가 높은 노드는 집단 사이에 오가는 정보나 자원을 통제하고 중개하는 게이트웨이 역할을 한다. 수학적으로는 임의의 두 노드 사이 최단 경로가 노드 v를 경유할 비율($\sigma_{st}(v)/\sigma_{st}$)을 무방향 네트워크의 모든 비순서 노드 쌍($s < t$)에 걸쳐 누적 합산한 값이다. 비순서 노드 쌍($s < t$)만을 합산함으로써 무방향 네트워크에서 동일한 경로가 양방향으로 중복 계산되는 것을 방지한다.

$$C_B(v) = \sum_{s<t,\ s\neq v\neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

(σ_{st}: 노드 s와 t 사이의 전체 최단 경로 수. $\sigma_{st}(v)$: 그중 노드 v를 거치는 최단 경로 수)

엣지 가중치가 주어졌을 때는 단순히 거쳐 가는 엣지의 개수 대신 거리로 변환된 엣지 가중치의 합이 최소가 되는 가중 최단 경로를 탐색한다. 노드 v가 이러한 가중 최단 경로상에 자주 포함될수록 매개 중심성이 더 높아진다.

◎ **고유벡터 중심성(Eigenvector Centrality)**

자신과 연결된 이웃 노드의 중요도까지 고려하여 노드의 영향력을 측정한다(Bonacich 1987). 소셜 네트워크에서 영향력 있는 다수와 긴밀히 연결된 대상을 찾는 원리와 같다. 이 지표가 높은 노드는 네트워크의 핵심 노드와 연결되어 있어 파급력이 크다. 수학적으로는 인접 행렬 $\mathbf{A}$의 최대 고유값 λ에 대응하는 고유벡터 $\mathbf{x}$의 성분 x_v로 정의된다. 즉 노드 v의 중심성은 자신과 연결된 이웃의 중심성 점수 합을 고유값 λ로 정규화한 값이다.

$$C_E(v) = x_v = \frac{1}{\lambda}\sum_{u} A_{vu}x_u \quad (\text{행렬 표현: } \mathbf{Ax} = \lambda\mathbf{x})$$

($\mathbf{A}$는 네트워크 연결 관계를 담은 인접 행렬(Adjacency Matrix), $\mathbf{x}$는 각 노드의 고유벡터 중심성 점수 $C_E(v) = x_v$를 담은 고유벡터, λ는 행렬의 최대 고유값(Eigenvalue)이다. 즉 어떤 노드의 중심성 점수는 자신과 연결된 이웃의 중심성 점수 합에 비례하며 비례상수 λ로 전체 점수 척도가 정규화된다.)

엣지 가중치가 있는 네트워크에서는 0과 1로 구성된 단순 인접 행렬 $\mathbf{A}$ 대신 엣지 가중치를 원소로

갖는 가중 인접 행렬 **W**를 적용한다. 영향력 있는 이웃과 연결되어 있더라도 엣지 가중치가 큰 경로로 연결될수록 중심성이 더 높아진다.

◎ **페이지랭크 중심성(PageRank Centrality)**

연결 수라는 양적 한계를 극복하기 위해 제안된 질적 지표이다. 구글(Google)의 초기 웹페이지 검색 알고리즘으로 널리 알려졌으며(Page et al. 1999), 고유벡터 중심성을 방향성과 가중치가 있는 네트워크에 맞게 보완한 형태이다. 원래 방향성 네트워크(웹페이지 링크망 등)에서는 주변부 다수와 연결된 노드보다 소수의 중심 거점(Hub) 노드로부터 지지를 받는 노드가 더 높은 점수를 얻도록 고안되었다. 그러나 본초 네트워크와 같은 대칭적 무방향 네트워크에서는 연결이 양방향으로 작용하므로 랜덤 워크의 정상분포 체류 확률이 노드의 차수(가중 네트워크에서는 노드 강도)에 비례하여 수렴하는 수학적 성질을 보인다. 직관적으로는 가상의 탐색자가 본초 사이 연결 강도에 따라 네트워크를 계속 이동할 때 각 본초 노드에 머무르게 되는 장기적 체류 확률을 나타낸다.[1] 탐색자는 일정 확률(d)로 현재 본초와 연결된 이웃 본초로 이동하고, 나머지 확률($1-d$)로는 전체 본초 중 하나를 무작위 선택하여 순간이동한다. 거듭제곱 반복법(Power Iteration)으로 이 확률값이 균형 상태로 수렴할 때까지 계산하여 최종 중심성을 도출한다.

$$\mathrm{PR}(v) = d \sum_{u \in \mathcal{N}(v)} \frac{w_{uv}}{s_u} \mathrm{PR}(u) + (1-d)\frac{1}{N}$$

(d: 무작위 이동을 고려한 감쇠 계수(Damping Factor, 보통 0.85). $\mathcal{N}(v)$: 노드 v와 연결된 이웃 노드 집합. w_{uv}: 노드 u와 v 사이의 엣지 가중치. $s_u = \sum_k w_{uk}$: 노드 u의 노드 강도(가중 차수 합). 결합 가중치가 큰 경로일수록 더 높은 중심성 점수가 전달된다. 가중치가 없는

[1] 수학적으로 페이지랭크는 마르코프 연쇄(Markov Chain)의 정상 상태 확률분포(Stationary Distribution)를 모델링한 것이며, 상태 전이 확률 행렬의 최대 고유벡터에 해당한다.

네트워크에서는 w_{uv}/s_u가 단순히 $1/k_u$로 치환된다.)

유사도(Similarity)와 거리(Distance)

네트워크 데이터에 부여된 엣지 가중치나 분석 알고리즘이 입력값으로 요구하는 파라미터는 모두 숫자로 표현되는 양적 수치이다. 그러나 모든 양적 수치가 같은 의미를 갖지는 않는다. 해당 수치가 '유사도(Similarity)'인지 '거리(Distance)'인지 그 수학적 성격을 먼저 파악하고 알고리즘의 요구 조건과 일치시켜야 한다. 유사도와 거리(비유사도)는 지향하는 수치의 방향이 정반대이기 때문이다. 유사도는 수치가 클수록 두 노드가 긴밀하고 가깝다는 의미이지만, 거리는 수치가 작을수록 가깝고 수치가 클수록 서로 멀리 떨어져 있음을 뜻한다.

본초 네트워크에서 다루는 엣지 가중치는 두 본초가 처방에서 얼마나 긴밀하게 결합했는지를 나타내는 전형적인 유사도 수치이다. 따라서 분석 알고리즘이 유사도를 요구하는지 거리를 요구하는지에 따라 가중치 처리 방식이 달라진다.

'차수 중심성', '페이지랭크 중심성', '고유벡터 중심성' 같은 중심성 지표나 이후에 살펴볼 '프루크터만-라인골드', 'UMAP'와 같은 레이아웃 알고리즘, 그리고 '라이덴 커뮤니티 탐지' 알고리즘에서는 엣지의 결합 강도나 연결 확률이 파라미터로 사용된다. 이들은 모두 유사도 계통의 수치이므로 본초 네트워크의 엣지 가중치를 변환 없이 그대로 사용할 수 있다.

반면 '근접 중심성', '매개 중심성'처럼 최단 경로를 탐색하는 중심성 지표나 이후에 살펴볼 '카마다-카와이(KK)', '다차원 척도법(MDS)'과 같은 레이아웃 알고리즘, 그리고 '계층적 군집화' 등에서는 공간적 간격이나 경로 비용이 파라미터로 사용된다. 이들은 모두 거리 계통의 수치이다. 그러므로 유사도 가중치를 거리 자리에 그대로 투입하면 결합 강도가 가장 강한 본초 쌍이 가장 멀리 떨어진 관계로 왜곡 계산되는 치명적인 오류가 발생한다. 따라서 이러한 알고리즘을 적용할 때는 반드시 유사도 가중치를 거리로 변환해야 한다.

유사도를 거리로 바꾸는 방법에는 여러 가지가 있다. 가장 널리 쓰이는 방식은 가중치의 역수를 취하는 역수 변환($d = 1/w$)이다. 계산이 단순하여 널리 쓰이지만, 가중치가 0에 가까우면 거리가 무한대로 발산하고 비선형 왜곡이 커지는 문제가 생긴다. 거리를 지수 함수로 감쇠시키는 변환($d = \exp(-w)$)이나 가중치 최댓값으로 나눈 값을 차감하는 정규화 변환($d = 1 - w/w_{\max}$)도 활용된다. 다만 정규화 차감 방식은 최대 결합 쌍의 거리가 0이 되어 거리 기반 분모 계산에서 오류를 유발할 수 있다.

이 책에서는 수치 범위를 안정적으로 유지하고 극단적인 발산이나 거리가 0이 되는 결함을 방지하기 위해 **선형 반전(Linear Inversion)** 공식을 표준으로 적용한다. 아울러 계층적 군집화나 거리 기반 분석에서 엣지가 존재하지 않는 비연결 노드 쌍($(i,j) \notin E$)에는 패널티 거리 승수($\beta = 1.5$)를 부여하여 최대 관찰 거리보다 충분히 먼 거리로 설정한다.

$$d_{ij} = \begin{cases} 0, & \text{if } i = j \\ (w_{\max} + w_{\min}) - w_{ij}, & \text{if } (i,j) \in E \\ (w_{\max} + w_{\min}) \cdot \beta, & \text{if } (i,j) \notin E \end{cases}$$

(d_{ij}: 본초 i, j 사이의 비유사도 거리, w_{ij}: 두 본초 간 유사도 엣지 가중치, $w_{\max}, w_{\min}$: 네트워크 내 유효 가중치의 최댓값과 최솟값, β: 엣지가 결여된 비연결 노드 쌍에 부여하는 패널티 승수 ($\beta = 1.5$))
이 공식은 가중치 최댓값과 최솟값의 합을 기준으로 각 엣지의 가중치를 차감한다. 가장 강하게 결합된 본초 쌍($w_{ij} = w_{\max}$)이라도 최소 거리 $d_{ij} = w_{\min} > 0$을 보장받아 거리가 0이 되는 오류를 방지하며 결합 강도가 약할수록 거리가 선형적으로 늘어난다. 연결 엣지가 없는 쌍에는 패널티 승수(β)를 곱해 군집 병합이나 거리 최적화 시 가장 나중에 연결되도록 유도한다. 이 선형 반전 변환은 이후 설명에서 본초 네트워크 엣지 가중치를 경로나 거리의 의미로 변환할 때 공통적으로 사용한다.

6.2 중심성 분석

여기서는 여러 중심성 지표 가운데 양적 지표인 '차수 중심성'과 영향력의 질적 가치를 평가하는 '페이지랭크 중심성' 두 지표를 살펴본다. ‹단순 빈도 네트워크›와 ‹G-score 네트워크›에 이 지표를 각각 적용하고 네트워크에 따라 '중요한 노드'가 어떻게 변하는지 비교해 본다. 차수 중심성을 계산할 때 엣지 가중치를 적용하지 않고 이웃 노드 수만 측정하는 기본 공식을 사용하였다. 페이지랭크 중심성을 계산할 때는 엣지 가중치를 반영하였는데 ‹단순 빈도 네트워크›의 엣지 가중치는 '동시 출현 빈도'이며 ‹G-score 네트워크›에서는 'G-score'이다.

Table 6.1 중심성 분석 상위 본초 (상위 20개) | ‹단순 빈도 네트워크›

순위	본초 (차수)	차수 중심성	본초 (페이지랭크)	페이지랭크 중심성
1	감초(甘草)	0.826	감초(甘草)	0.0155
2	당귀(當歸)	0.783	당귀(當歸)	0.0133
3	복령(茯苓)	0.71	복령(茯苓)	0.0112
4	작약(芍藥)	0.701	인삼(人蔘)	0.0103
5	인삼(人蔘)	0.685	작약(芍藥)	0.01
6	천궁(川芎)	0.673	천궁(川芎)	0.0093
7	육계(肉桂)	0.616	황련(黃連)	0.0086
8	방풍(防風)	0.607	방풍(防風)	0.0085
9	대황(大黃)	0.605	육계(肉桂)	0.0084
10	생지황(生地黃)	0.6	대황(大黃)	0.0081
11	황금(黃芩)	0.591	백출(白朮)	0.008
12	백출(白朮)	0.587	생지황(生地黃)	0.0079
13	목향(木香)	0.584	황금(黃芩)	0.0077
14	황련(黃連)	0.58	사향(麝香)	0.0077
15	사향(麝香)	0.555	목향(木香)	0.0076
16	부자(附子)	0.553	진피(陳皮)	0.0071

순위	본초 (차수)	차수 중심성	본초 (페이지랭크)	페이지랭크 중심성
17	황기(黃芪)	0.543	부자(附子)	0.0071
18	백지(白芷)	0.541	건강(乾薑)	0.0068
19	세신(細辛)	0.537	황기(黃芪)	0.0068
20	건강(乾薑)	0.536	세신(細辛)	0.0067

‹단순 빈도 네트워크›에서는 출현 빈도가 가장 높은 감초(甘草)가 두 중심성 모두에서 1위를 차지했다. 당귀(當歸), 복령(茯苓) 등의 본초가 그 뒤를 이었다.

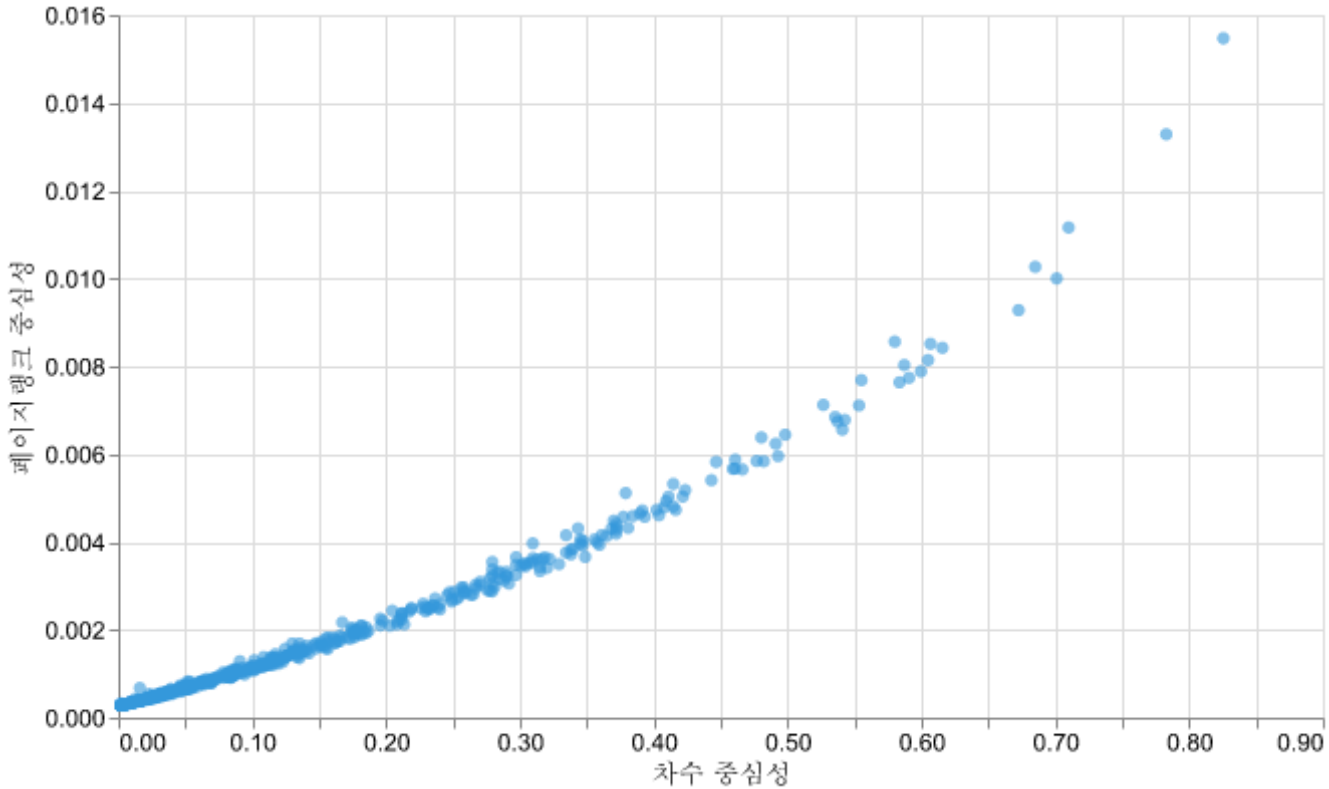

Figure 6.1 중심성 비교 (차수 vs 페이지랭크) | ‹단순 빈도 네트워크›

산점도로 두 중심성 지표를 비교해 보면 둘 사이에 강한 양의 상관관계가 나타난다. 또 상위 100개 본초 중 96개가 일치하여 높은 정합성을 보였다. 이러한 높은 일치도는 데이터의 우연한 특성이 아니라 무방향 네트워크에서 랜덤 워크의 정상분포가 노드 차수(가중 네트워크에서는 노드 강도)에 비례하는 수학적 성질에 기인한다. 따라서 무방향 단순 빈도 네트워크에서는 연결된 이웃이 많은 노드일수록 페이지랭크 점수 역시 비례하여 높아진다.

Table 6.2 중심성 분석 상위 본초 (상위 20개) | ‹G-score 네트워크›

순위	본초 (차수)	차수 중심성	본초 (페이지랭크)	페이지랭크 중심성
1	방풍(防風)	0.276	방풍(防風)	0.0105
2	당귀(當歸)	0.263	사향(麝香)	0.0098
3	생지황(生地黃)	0.259	당귀(當歸)	0.009
4	육계(肉桂)	0.256	주사(硃砂)	0.0088
5	우슬(牛膝)	0.256	복령(茯苓)	0.0086
6	복령(茯苓)	0.254	천궁(川芎)	0.0083
7	백지(白芷)	0.25	우슬(牛膝)	0.0076
8	세신(細辛)	0.24	인삼(人蔘)	0.0074
9	몰약(沒藥)	0.24	진피(陳皮)	0.0074
10	천궁(川芎)	0.235	육종용(肉蓯蓉)	0.0073
11	부자(附子)	0.233	목향(木香)	0.0073
12	인삼(人蔘)	0.231	숙지황(熟地黃)	0.007
13	사향(麝香)	0.231	유향(乳香)	0.0069
14	천초(川椒)	0.23	작약(芍藥)	0.0066
15	작약(芍藥)	0.214	백출(白朮)	0.0066
16	유향(乳香)	0.214	황금(黃芩)	0.0066
17	감초(甘草)	0.212	감초(甘草)	0.0065
18	웅황(雄黃)	0.211	강활(羌活)	0.0064
19	강활(羌活)	0.209	오미자(五味子)	0.0063
20	진교(秦艽)	0.207	빙편(冰片)	0.0063

‹G-score 네트워크›에서는 특정 효능군을 대표하는 방풍(防風), 당귀(當歸), 생지황(生地黃) 같은 본초가 상위권에 올랐다. 이들 본초는 다른 본초와 배합되어 처방의 치료 방향을 규정하는 핵심 본초이다.

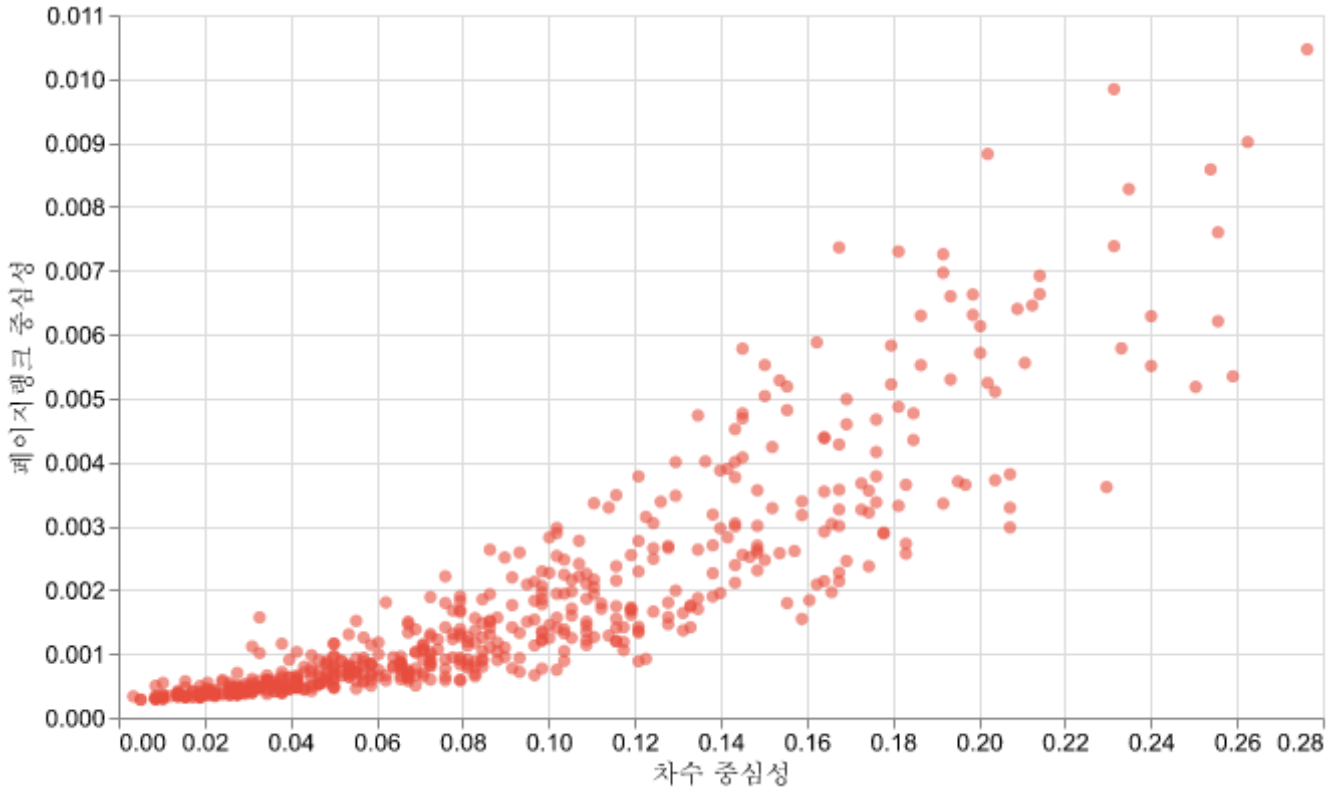

Figure 6.2 중심성 비교 (차수 vs 페이지랭크) | ‹G-score 네트워크›

‹G-score 네트워크›의 두 중심성 사이에서도 양의 상관관계가 나타났다. 상위 100개 중 77개 본초가 일치하여 지표 사이의 전반적인 정합성도 유지되었다. 이 역시 무방향 네트워크에서 랜덤 워크의 정상분포가 연결 구조에 비례하는 수학적 성질이 작용한 결과다. 그러나 ‹단순 빈도 네트워크›에 비해 산점도의 분포 폭은 훨씬 넓었다. 특히 우상향 영역에서 차수 중심성이 비슷한 상위권 본초 사이에서도 페이지랭크 점수가 위아래로 넓게 분산되는 양상을 보였다. 이는 차수 중심성이 엣지 유무만을 센 비가중 지표인 반면 페이지랭크는 통계적 연관 강도(G-score)를 가중치로 반영한 지표이기 때문이다. 무방향 가중 네트워크에서 랜덤 워크의 정상분포는 노드 차수가 아니라 각 엣지 가중치의 합인 **노드 강도(Node Strength)**에 비례한다. 따라서 단순 연결 차수가 비슷하더라도 특정 처방군에서 매우 강하게 결합한 본초는 노드 강도가 크게 증가하여 페이지랭크 점수가 분산되는 것이다.

이제 동일한 본초가 두 네트워크에서 각각 어떤 중심성을 나타내는지 비교해 본다.

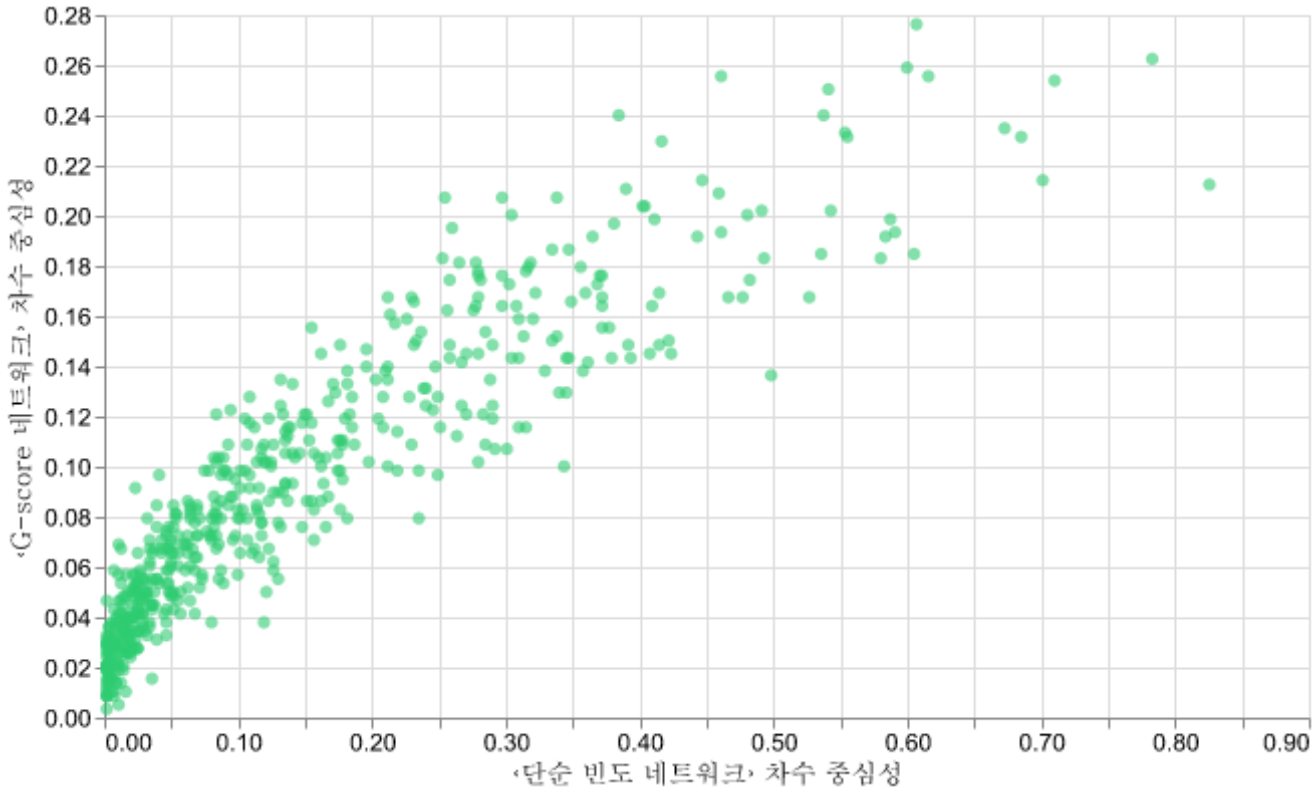

Figure 6.3 네트워크 간 차수 중심성 비교

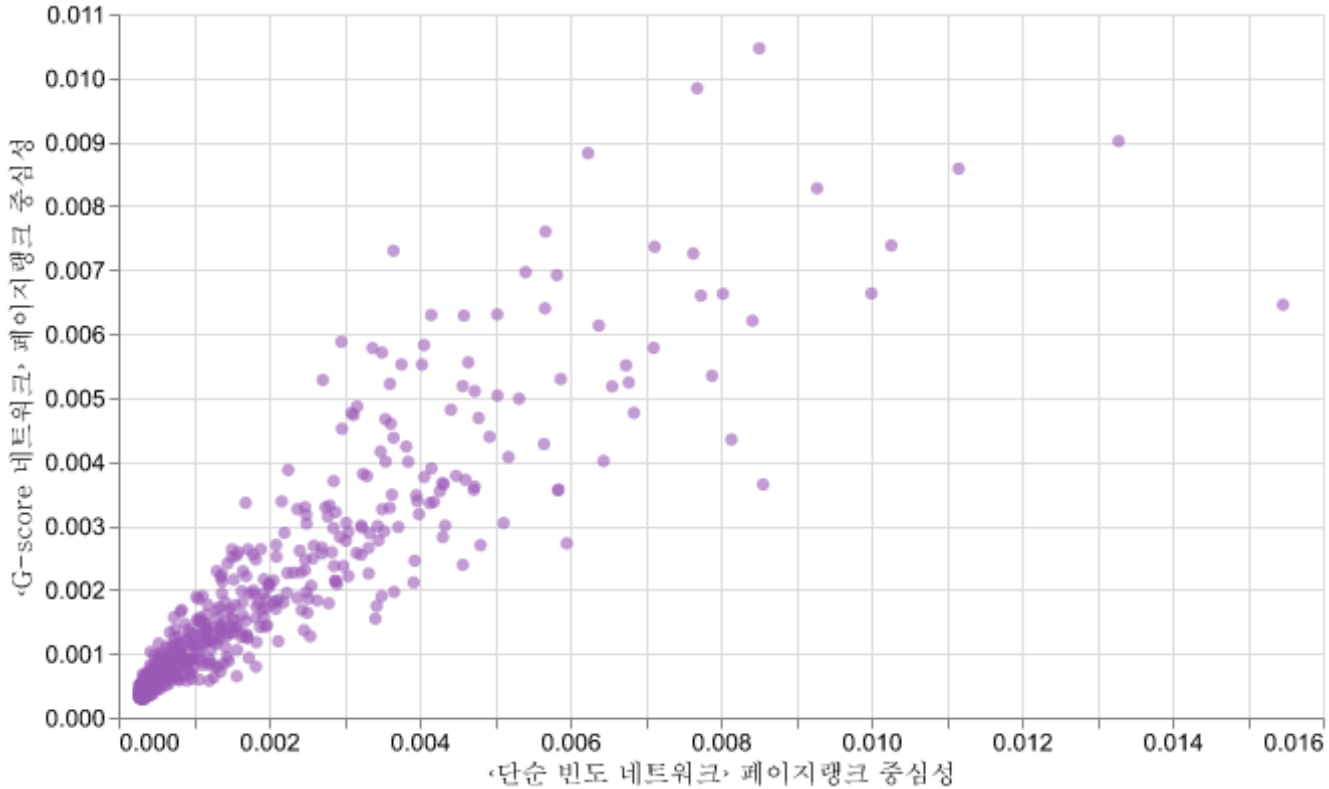

Figure 6.4 네트워크 간 페이지랭크 중심성 비교

산점도로 두 네트워크의 산점도를 비교해보면 ‹단순 빈도 네트워크›에서 중심성이 높았던 본초가 ‹G-score 네트워크›에서도 대체로 높은 중심성을 유지하는 분포를 보였다. 상위 100개 본초를 기준으로 비교했을 때에도 차수 중심성에서 76개, 페이지랭크 중심성에서 82개 본초가 공통으로 포함되었다.

차수 중심성(76개)보다 페이지랭크 중심성(82개)에서 상위 본초 일치도가 약간 더 높게 나타났다. 엣지 필터링 기준을 단순 빈도에서 G-score로 전환할 때 국소(Local) 엣지 유무가 달라지면서 차수 중심성은 상대적으로 민감하게 변했기 때문으로 보인다. 반면 페이지랭크는 이러한 변화 속에서도 더 안정적으로 유지되었다. 이는 한의학 처방의 골격을 형성하는 주요 본초가 단순한 출현 빈도 효과에 그치지 않고 네트워크 전역에서 높은 위상학적 안정성을 지니고 있음을 보여준다.

6.3 강건성 시뮬레이션

네트워크가 외부 충격이나 구성 요소의 손실에 얼마나 견딜 수 있는지를 평가하기 위해 '강건성 시뮬레이션'을 수행한다. **강건성(Robustness) 시뮬레이션**은 특정 노드가 제거되었을 때 네트워크 전체의 연결 구조가 얼마나 안정적으로 유지되는지 검증하는 방법이다.

이를 위해 전체 노드의 일정 비율을 순차적으로 제거하는 두 가지 시나리오를 비교한다. **무작위 공격(Random Attack)**은 임의의 노드를 무작위로 제거하며 네트워크 붕괴 과정을 살펴보는 방식이다. 반면 **타깃 공격(Targeted Attack)**은 페이지랭크 중심성이 높은 상위 노드부터 차례로 제거해 나가는 방식이다.

평가 지표는 노드 제거 후 남은 '거대 연결 요소'의 크기 비율이다. **거대 연결 요소(Giant Connected Component, GCC)**는 네트워크에서 서로 연결된 노드 집합 중 가장 규모가 큰 단일 서브네트워크를 가리킨다. 그 상대적 크기로 네트워크 전체가 하나의 통합된 구조를 유지하는

정도를 정량화할 수 있다.

노드의 연결 차수 분포가 멱법칙을 따르는 네트워크를 **척도 없는 네트워크(Scale-Free Network)**라고 한다(Barabási and Albert 1999). 평균적인 연결 수를 대표 척도로 삼기 어려울 만큼 소수의 중심 거점(Hub) 노드가 연결을 독점하는 비대칭 구조를 띤다. 이러한 척도 없는 네트워크는 무작위 공격에서 높은 내성을 보인다. 그러나 타깃 공격으로 핵심 노드를 소수만 제거해도 거대 연결 요소(GCC)가 급격히 와해되는 구조적 취약성을 보인다(Albert et al. 2000). 무작위 공격에서는 다수를 차지하는 주변부 노드가 먼저 제거되는 반면 타깃 공격에서는 중심성이 높은 중심 거점(Hub) 노드가 먼저 제거되어 거대 연결 요소를 유지할 수 없게 되기 때문이다.

강건성 시뮬레이션은 한의학 처방 관점에서는 “처방에서 특정 본초를 제외했을 때 본초 네트워크의 핵심 골격이 얼마나 유지되는가?”라는 질문과 맞닿아 있다. 노드를 0%에서 50%까지 점진적으로 제거했을 때 네트워크가 어떻게 변화하는지 전체적인 시뮬레이션 결과를 그래프로 살펴본다. 여기서 타깃 공격의 중심성 지표로는 페이지랭크 중심성을 적용하였다. 단순히 연결이 많은 노드보다 네트워크 전체의 흐름에 실질적 영향을 미치는 노드를 제거하는 편이 공격의 취약점을 평가하는 데 더 적합하기 때문이다.

> **i** 계층 분석(제5장)과 강건성 시뮬레이션(제6장)의 비교
>
> 제5장의 ’중심부-주변부 분석’과 제6장의 ’강건성 시뮬레이션’은 모두 네트워크에서 노드를 단계적으로 제거하며 남은 구조의 변화를 관찰한다는 공통점이 있다. 그러나 노드를 제거하는 방식과 분석 목적에는 뚜렷한 차이가 있다. 중심부-주변부 분석(k-core)은 연결 차수가 기준 이하인 노드를 순차적으로 제외하며 외곽 주변부에서 중심부에 이르는 네트워크 골격을 계층적으로 확인한다. 반면 강건성 시뮬레이션은 무작위로 노드를 선택하거나 중심성이 높은 핵심 거점(Hub) 노드를 제거한다. 이를 바탕으로 특정 노드가 결손되었을 때 전체 네트워크가 얼마나 빠르게 파편화되는지 구조적 내구성을 평가한다.

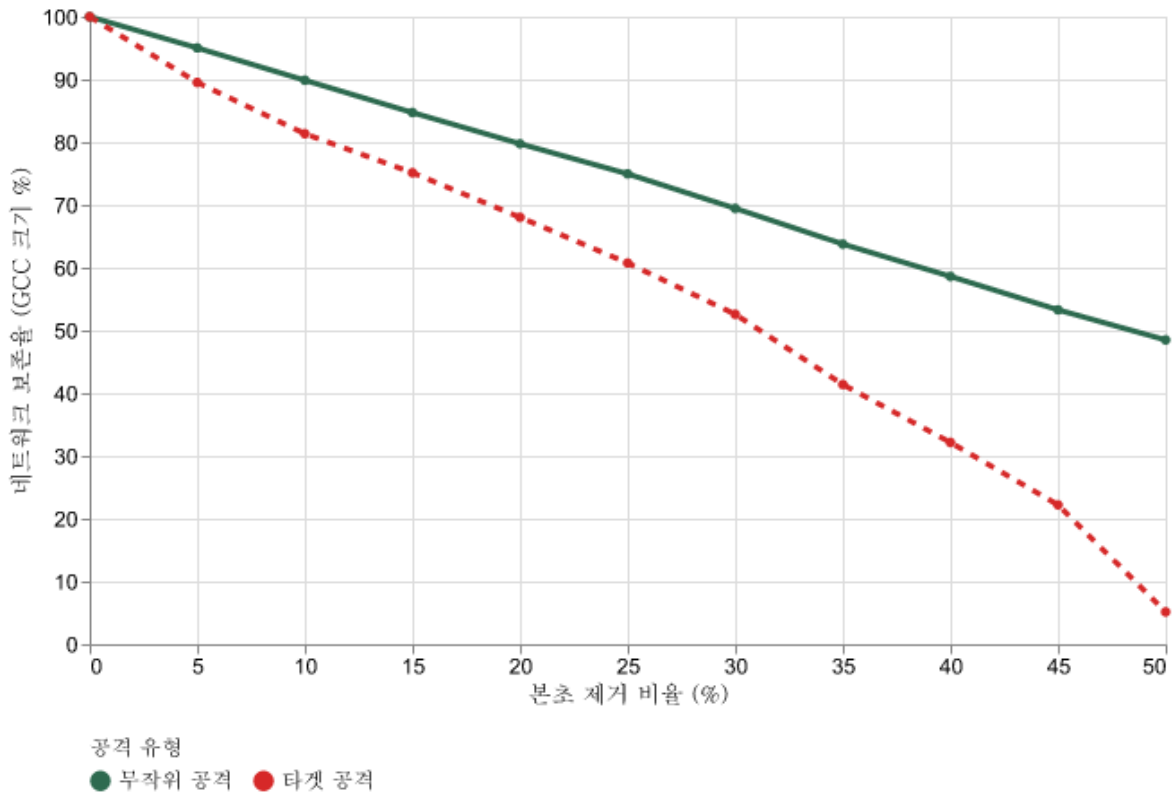

Figure 6.5 강건성 시뮬레이션 | ‹단순 빈도 네트워크›

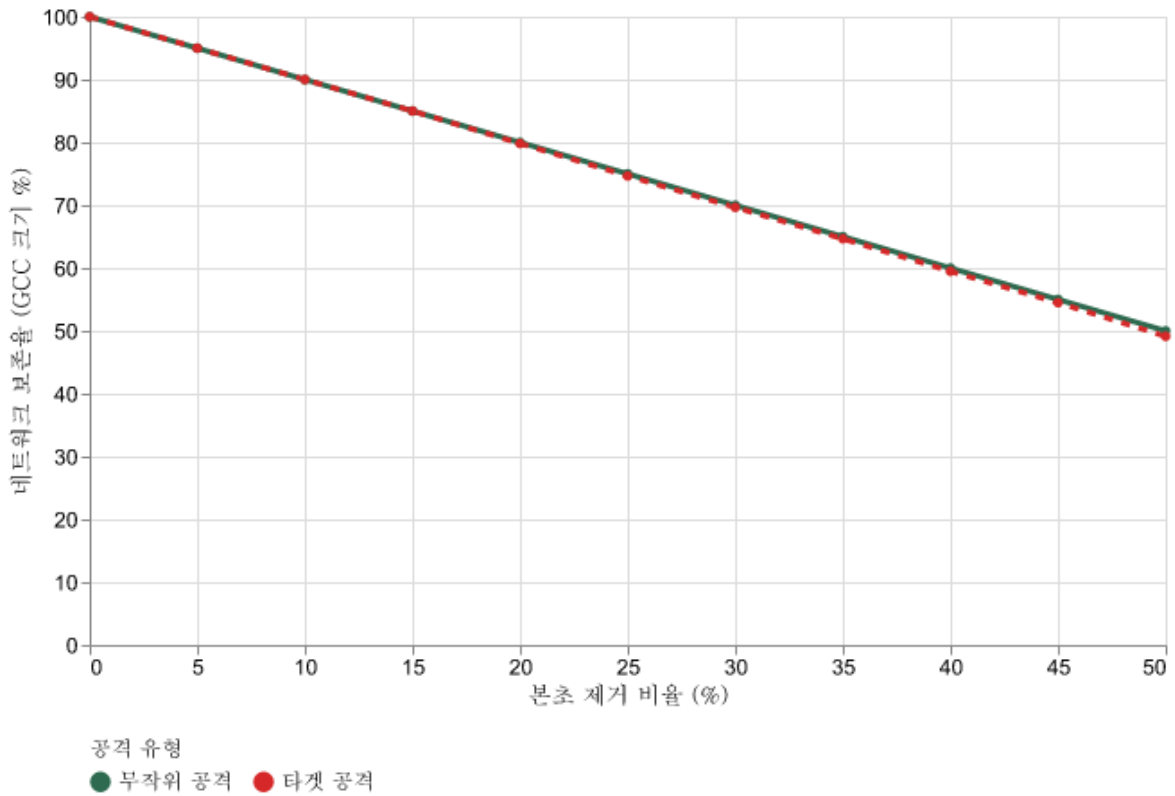

Figure 6.6 강건성 시뮬레이션 | ‹G-score 네트워크›

시뮬레이션 결과를 살펴보면 ‹단순 빈도 네트워크›에서는 무작위 공격으로 본초 절반(50%)을 차례로 제거했을 때 GCC 보존율이 48.5%까지 완만하게 감소한다. 반면 타깃 공격의 양상은 크게 달랐다. 25%의 노드만 제거해도 GCC 보존율은 60.7%로 크게 떨어졌고, 50%를 제거하자 GCC는 전체의 5.2%만 남아 네트워크가 사실상 완전히 파편화되었다.

‹G-score 네트워크› 역시 무작위 공격으로 본초 절반을 제거했을 때 GCC의 50.0%가 유지되었다. 그러나 타깃 공격을 가했을 때는 반응이 사뭇 달랐다. 25%의 핵심 노드를 제거해도 GCC 보존율은 74.7%에 달했고, 50%를 제거하더라도 전체의 49.1%를 유지하였다. 이는 무작위 공격과 비교해도 큰 차이가 없는 수준이다.

일반적으로 강건성 시뮬레이션에서는 노드를 무작위로 제거할 때(무작위 공격) 연결망이 비교적 잘 유지되지만 중심성이 높은 핵심 노드를 집중 제거하면(타깃 공격) 연결성이 급격히 붕괴하는 경향을 보인다. 두 네트워크의 시뮬레이션 결과를 이 기준에 비추어 보면 ‹단순 빈도 네트워크›는 소수 거점(Hub) 노드에 연결이 집중된 네트워크가 지닌 전형적 취약성을 그대로 드러냈으나 ‹G-score 네트워크›는 예외적 반응을 보였다. 무작위 공격에서는 두 네트워크 모두 거대 연결 요소(GCC)의 크기가 완만하게 줄어들며 유사한 내구성을 나타냈다. 그러나 타깃 공격에서는 큰 차이가 발생했다. ‹단순 빈도 네트워크›는 GCC가 급감하여 네트워크가 빠르게 해체된 반면 ‹G-score 네트워크›는 핵심 노드가 제거된 뒤에도 GCC가 상당 수준 보존되며 완만하게 줄어들었다.

이 시뮬레이션 결과는 두 네트워크의 구조적 특성 차이를 잘 보여준다. 타깃 공격에서 ‹단순 빈도 네트워크›가 급격히 붕괴한 까닭은 제거 대상이 된 상위 중심성 노드가 감초(甘草)·복령(茯苓)처럼 수많은 처방에 두루 사용되는 범용 본초였기 때문이다. 이들 거점(Hub) 노드가 전역 네트워크의 중심 골격을 독점하고 있었기에 일반적인 강건성 시뮬레이션의 전형적 붕괴 양상이 재현되었다. 반면 ‹G-score 네트워크›의 중심 노드는 방풍(防風)·당귀(當歸)·사향(麝香)처럼 서로 다른 치료

영역을 대표하는 거점들이었다. 네트워크 전체가 특정 소수 중심 노드에 전적으로 의존하지 않으므로 상위 거점이 순차적으로 제거되더라도 거대 연결 구조(GCC)에 미치는 파급력이 분산되어 완만한 감소세를 유지할 수 있었다.

그렇다면 타깃 공격에서 급격히 파편화되지 않은 ‹G-score 네트워크›는 한의학 처방의 특성을 제대로 반영하지 못하는 것일까? 그렇지 않다. 두 네트워크는 한의학 처방을 바라보는 관점과 분석 층위가 다를 뿐이다. ‹단순 빈도 네트워크›는 다양한 처방에 두루 쓰이는 본초를 연결하여 한의학 임상 처방의 실제 현황을 충실하게 보여준다. 반면 ‹G-score 네트워크›는 단순 출현 빈도 효과를 통계적으로 통제하고 본초 사이의 유의미한 기능적 배합 관계에 주목한다. 즉 ‹단순 빈도 네트워크›가 본초 사용 현황을 양적으로 드러낸다면 ‹G-score 네트워크›는 본초 사이 결합을 질적으로 보여준다. 이를 통해 분석 목적에 맞추어 네트워크 모델링 방식을 선택해야 한다는 점을 다시 확인할 수 있다.

지금까지 네트워크 노드의 역할과 영향력, 그리고 강건성을 살펴보았다. 이러한 노드와 엣지의 구조적 관계를 직관적으로 파악하려면 2차원 공간 배치 기법이 필수적이다. 제7장에서는 이를 위해 고안된 다양한 레이아웃 알고리즘을 살펴본다.

제 7 장

네트워크 시각화

네트워크 시각화는 수많은 노드와 엣지가 맺는 연결 관계(위상 구조)를 한정된 2차원 평면에 배치하는 과정이다. 네트워크의 노드와 엣지를 사람이 직관적으로 인식할 수 있게 시각화하는 작업은 중요하다. 수치 지표만으로는 복잡한 네트워크의 전체적인 커뮤니티 분포, 중심축과 주변부의 분리, 병목 구조 등을 한눈에 파악하기 어렵기 때문이다. 시각화는 복잡계 데이터 이면에 숨은 패턴과 구조를 밝혀내고 새로운 가설을 발견하도록 돕는다.

하지만 적절한 배치 알고리즘이 없으면 노드가 중앙으로 뭉쳐 구조를 구분하기 어려운 현상이 일어난다.[1] 이러한 한계를 극복하고 네트워크의 위상 구조를 드러내고자 다양한 레이아웃 알고리즘이 개발되었다. 레이아웃 알고리즘은 노드와 엣지의 정보를 바탕으로 각 노드의 2차원 좌표를 계산한다. 이 장에서는 가장 널리 쓰이는 힘 지향 알고리즘과 차원 축소 알고리즘을 살펴본다.

[1] 이를 털뭉치(Hairball) 현상이라고 부른다.

> **i** 레이아웃(Layout)의 한계
>
> '레이아웃'은 노드의 공간적 배치 상태를 가리킨다. '레이아웃 알고리즘'은 그래프의 연결 구조와 가중치를 바탕으로 각 노드의 최적 좌표 (x, y)를 계산하는 수학적 절차이다.
> 네트워크 분석에서 레이아웃은 보통 2차원 평면에서 이루어진다. 본래 네트워크의 연결 구조와 경로 거리는 수십, 수백 차원에 걸친 고차원 위상 공간에 존재한다. 이를 2차원 평면에 투영하는 과정에서 필연적으로 기하학적 왜곡이 발생한다. 어떤 왜곡을 허용하느냐에 따라 서로 다른 레이아웃 알고리즘이 쓰인다. 여러 알고리즘이 고안된 이유이기도 하다. 따라서 시각화된 2차원 물리적 거리는 절대적인 관계를 뜻하지 않는다. 2차원 평면에 표시된 배치는 위상적 연결 패턴을 탐색하는 보조 수단으로 이해해야 한다.

7.1 힘 지향 레이아웃

힘 지향 레이아웃(Force-directed Layout)은 물리학의 입자 동역학 모델을 그래프 시각화에 적용한 알고리즘 계열이다. 이 방식은 노드 사이에 서로 밀어내는 척력을, 엣지로 연결된 노드 사이에는 당기는 인력을 부여한다. 모든 노드 쌍은 거리가 가까워질수록 서로 밀어내는 척력을 받는다. 반대로 엣지로 연결된 노드 쌍은 인력을 받아 서로 가까워진다. 힘 지향 알고리즘에는 표준 방법인 프루크터만-라인골드 알고리즘과 전역(Global) 최단 경로 거리를 보존하는 카마다-카와이 알고리즘 등이 있다.

알고리즘은 노드의 초기 좌표를 무작위로 설정한 뒤 시뮬레이션을 반복한다. 단계마다 각 노드에 작용하는 척력과 인력의 합력을 계산한 뒤 합력의 방향으로 노드를 조금씩 이동시킨다. 이 과정을 반복하면 노드에 작용하는 알짜힘이 줄어들어 국소(Local) 힘 평형 상태에 도달하거나('프루크터만-라인골드'), 목적 함수로 정의된 시스템 전체의 에너지가 최소화되는 방향으로 수렴한다('카마다-카와이'). 노드의 이동이 안정화되면 최종 배치가 결정된다.

다음 시각화 예제에서는 ‹단순 빈도 네트워크›와 ‹G-score 네트워크›를 함께 살펴보며 네트워크에서 가장 큰 연결 요소인 자이언트 컴포넌트만 대상으로 삼는다. 고립되거나 분리된 소규모

컴포넌트가 포함되면 컴포넌트 간 최단 경로 거리가 무한대(∞)가 되어 전역 거리를 바탕으로 작동하는 알고리즘(카마다-카와이, MDS 등)이 정상 동작하지 않기 때문이다. 아울러 동일한 연결 부분그래프를 대상으로 삼아야 레이아웃 알고리즘 간의 공간 배치 특성을 일관되게 비교할 수 있다.

◎ **프루크터만-라인골드(Fruchterman-Reingold) 알고리즘**

프루크터만-라인골드 알고리즘(Fruchterman and Reingold 1991)은 전역 에너지 함수를 직접 정의해 최적화하는 대신, 노드 간 인력과 척력의 국소 힘 평형을 찾는 휴리스틱 기반의 대표적인 힘 지향 레이아웃 방법이다. 이 알고리즘은 2차원 평면의 넓이(Area)와 전체 노드 수($|V|$)를 고려하여 노드 사이의 이상적인 균형 거리 k를 설정한다. 연결된 노드 사이에는 거리의 제곱에 비례하는 인력(F_a)이 작용하고 모든 노드 쌍 사이에는 거리에 반비례하는 척력(F_r)이 작용한다. 연결된 두 노드 사이의 거리 d가 정확히 k와 같아질 때($d = k$) 인력과 척력이 상쇄되어 알짜힘이 0이 되는 힘의 평형 상태에 도달한다. 엣지 가중치가 있을 때는 인력 계산에 가중치를 반영하여 결합 강도가 높을수록 공간에서 서로 더 가깝게 배치되도록 유도한다.

$$F_a(d) = \frac{d^2}{k}, \quad F_r(d) = -\frac{k^2}{d}, \quad k = C\sqrt{\frac{\text{Area}}{|V|}}$$

($F_a(d)$: 연결된 노드 쌍 사이에 작용하는 인력. $F_r(d)$: 모든 노드 쌍 사이에 작용하는 척력. d: 두 노드 사이의 거리. k: 화면 면적(Area)과 전체 노드 수($|V|$)로 결정되는 최적 균형 거리. C: 배치 스케일을 조절하는 상수)

프루크터만-라인골드 알고리즘은 시각적으로 균형 잡힌 그래프 배치를 빠르게 생성하며 인접한 노드 사이의 국소 연결 관계를 잘 드러낸다. 하지만 연결이 복잡한 밀집 네트워크에서는 노드가

중앙으로 쏠리는 경향이 있어 커뮤니티 사이의 경계가 불분명해지는 한계가 있다. 커뮤니티 구조는 다음 장에서 자세히 다룬다.

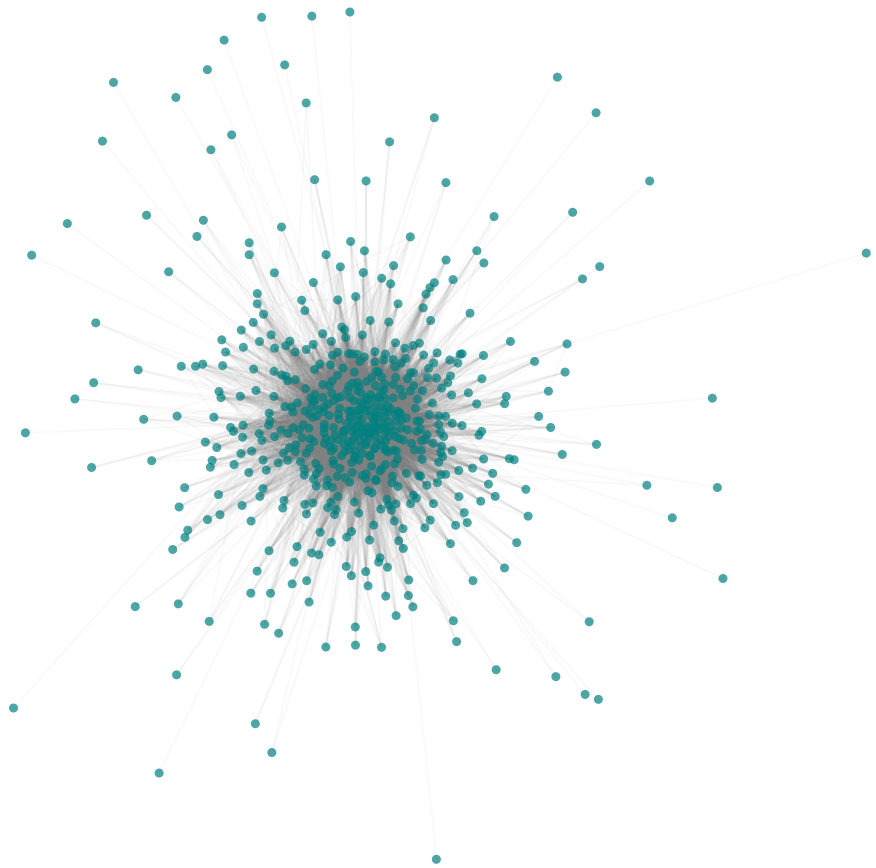

Figure 7.1 프루크터만-라인골드 레이아웃 | ‹단순 빈도 네트워크›

◎ **카마다-카와이(Kamada-Kawai) 알고리즘**

카마다-카와이 알고리즘(Kamada and Kawai 1989)은 전역 거리 보존을 목표로 하는 에너지



Figure 7.2 프루크터만-라인골드 레이아웃 | ‹G-score 네트워크›

최소화 레이아웃 방법이다. 프루크터만-라인골드 알고리즘은 국소 연결 관계를 잘 드러내지만 네트워크 전체의 거시적 거리를 보존하지는 못한다. 반면 카마다-카와이 알고리즘은 모든 노드 쌍 사이의 전역 최단 경로 거리를 기하학적 거리에 직접 대응시키는 방식을 취한다.

이 알고리즘은 모든 노드 쌍 사이의 이론적 그래프 최단 경로 거리를 계산한다. 이어 2차원 평면에서의 실제 기하학적 거리($\|p_i - p_j\|$)가 그래프 최단 경로 거리(d_{ij})에 비례하도록 노드 좌표를 배치한다. 수학적으로는 2차원 기하 거리와 이상적 목표 거리(l_{ij}) 간 편차의 제곱합을 가중 최소제곱법(Weighted Least Squares) 형태로 최적화한다. 가중치 계수 $k_{ij} = K/d_{ij}^2$는 그래프에서 가까운 노드 쌍일수록 거리 오차에 더 큰 페널티를 부과하여 국소 결합 구조와 전역 거시 구조를 균형 있게 보존하도록 작동한다.

카마다-카와이 알고리즘에서 엣지 가중치는 유사도가 아니라 거리를 뜻한다는 점에 유의해야 한다. 예를 들어 동시 출현 빈도가 높은 본초 쌍은 관계가 가까우므로 경로 거리가 짧아야 한다. 따라서 동시 출현 빈도 값을 그대로 사용하지 않고 선형 반전 변환을 적용하여 '거리 값'으로 변환해 사용한다.[2] 그래야 동시 출현 빈도가 높을수록 반전된 거리 값이 작아져 두 본초가 공간상에서 가깝게 배치된다.

이 방식은 직접 연결되지 않은 노드 쌍의 전역 관계까지 공간에 반영하므로 네트워크 전체의 거시 구조를 충실하게 보여준다. 다만 모든 노드 쌍 사이의 최단 경로를 계산해야 하므로 노드 수가 많은 대규모 네트워크에서는 계산 시간이 급격히 늘어난다.

$$E = \sum_{i<j} k_{ij}(\|p_i - p_j\| - l_{ij})^2$$

[2] 유사도 가중치를 거리로 변환하는 선형 반전 공식의 원리는 제6장의 '유사도(Similarity)와 거리(Distance)' 절을 참조한다.

(p_i, p_j: 2차원 평면에서 노드 i, j의 실제 배치 좌표. $\|p_i - p_j\|$: 두 노드 사이의 2차원 기하학적 거리. l_{ij}: 그래프 상의 최단 경로 거리(d_{ij})에 비례하는 이상적인 목표 거리. $k_{ij} = K/d_{ij}^2$: 이상적 최단 경로 거리의 제곱에 반비례하는 가중치 계수, 원래 가까운 노드 쌍일수록 거리 오차에 더 큰 페널티를 부여함.)

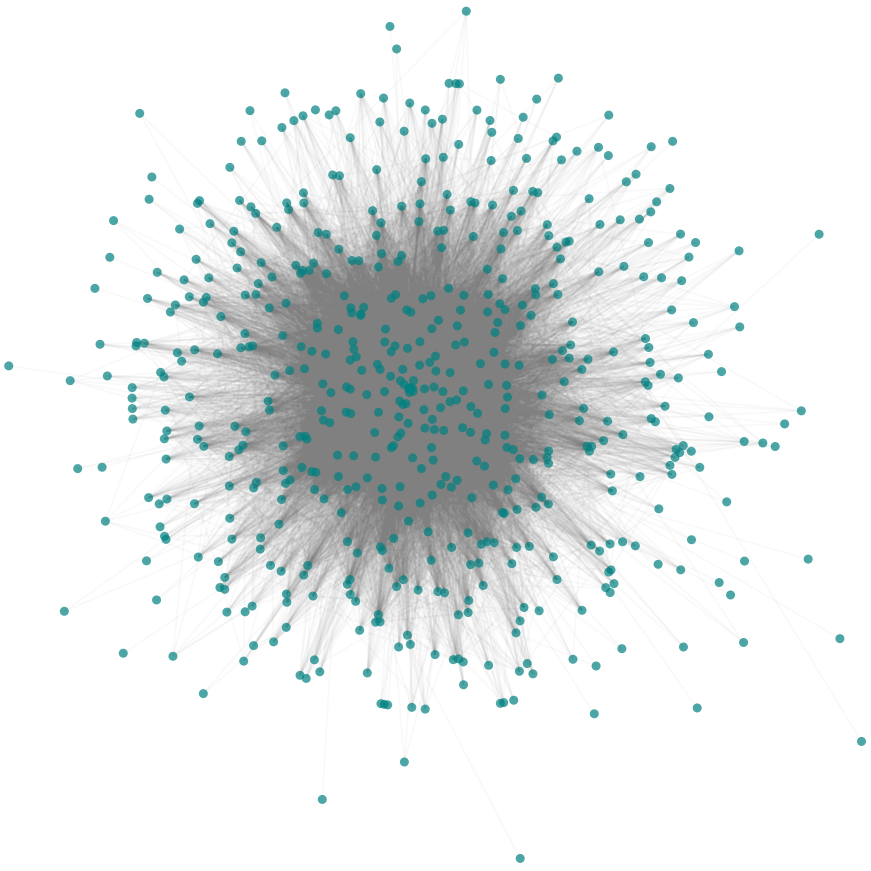

Figure 7.3 카마다-카와이 레이아웃 | ‹단순 빈도 네트워크›

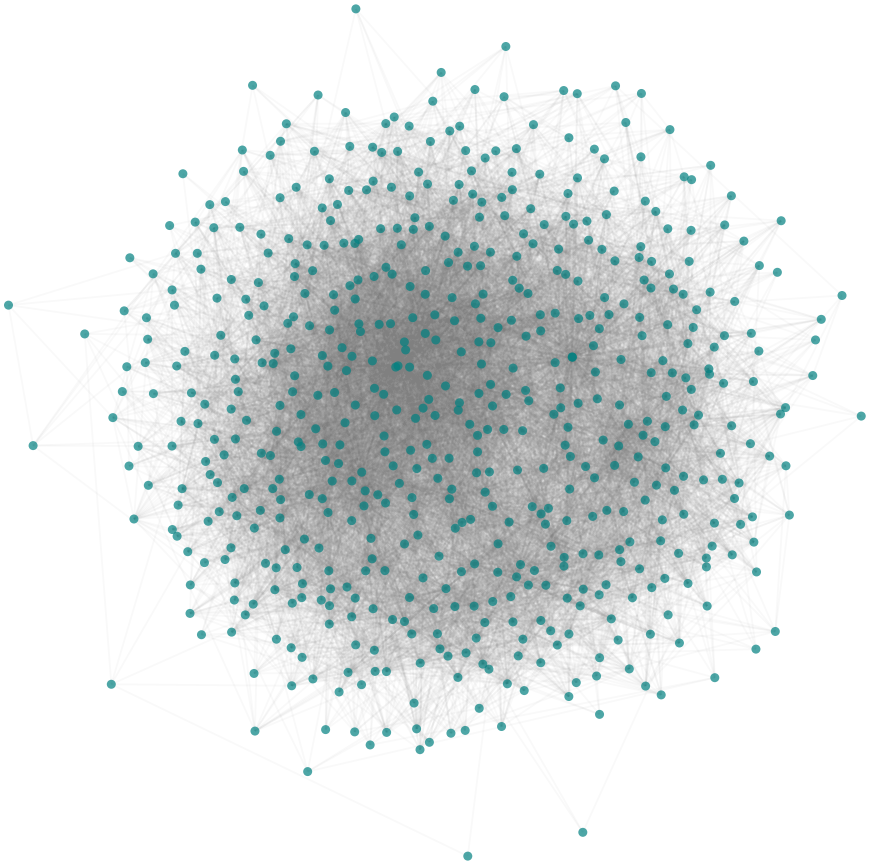

Figure 7.4 카마다-카와이 레이아웃 | ‹G-score 네트워크›

7.2 차원 축소 및 최적화 레이아웃

힘 지향 레이아웃이 가상의 물리 입자 시뮬레이션으로 노드 좌표를 찾는다면, 차원 축소 및 최적화 레이아웃은 수학과 머신러닝 기법을 활용하여 그래프의 고차원 위상 구조를 2차원 평면으로 직접 투영한다. 이 계열 알고리즘은 노드 사이의 최단 경로 거리 행렬이나 국소 연결 확률 구조를 모델링한 뒤 선형대수의 고유값 분해나 비선형 다양체 임베딩으로 좌표를 도출한다. 물리 시뮬레이션의 수렴 문제를 피하면서 네트워크의 거시적 최단 경로 구조나 비선형 커뮤니티 경계를 더 선명하게 드러내는 장점이 있다.

◎ **다차원 척도법(Multidimensional Scaling, MDS)**

다차원 척도법(MDS)은 개체 간 거리(비유사도) 정보를 보존하면서 고차원 관계망을 2차원 평면 좌표로 변환하는 기하학적 차원 축소 기법이다(Torgerson 1952; Kruskal 1964). 그래프 시각화에서는 먼저 모든 노드 쌍 사이의 가중 최단 경로 거리를 계산하여 거리 행렬을 구성한다. 그런 다음 거리 행렬을 선형대수적으로 변환(거리 제곱 행렬의 이중 중심화 및 고유값 분해)하여 네트워크의 거리 구조를 가장 잘 대변하는 고유벡터 2개를 2차원 평면의 좌표축으로 추출한다. 이로써 오차를 최소화하며 고차원 최단 경로 거리를 2차원 유클리드 평면 위에 배치한다.

이 방법은 카마다-카와이 알고리즘과 마찬가지로 전역 거리를 보존하는 데 목적이 있다. 카마다-카와이 알고리즘이 반복적 최적화로 에너지를 줄여간다면, 다차원 척도법은 단 한 번의 대수적 계산으로 좌표를 산출한다. 다차원 척도법에서도 엣지 가중치는 거리를 뜻하므로 카마다-카와이 알고리즘과 마찬가지로 엣지 가중치에 선형 반전을 적용하여 구한 거리 값을 바탕으로 최단 경로 거리 행렬을 계산한다.

다차원 척도법은 선형대수적 변환을 거치므로 계산 과정이 명확하고 무작위 초기값의 영향을 받지

않아 항상 일관된 배치를 만든다. 다만 선형 투영 방식을 취하므로 곡면처럼 휘어지거나 복잡하게 얽힌 비선형 커뮤니티 구조를 포착하는 데는 한계가 있다.



Figure 7.5 다차원 척도법 레이아웃 | ‹단순 빈도 네트워크›

◎ UMAP(Uniform Manifold Approximation and Projection)

UMAP은 복잡한 고차원 데이터의 국소 연결 형태를 저차원 공간에 효과적으로 투영하는 대표

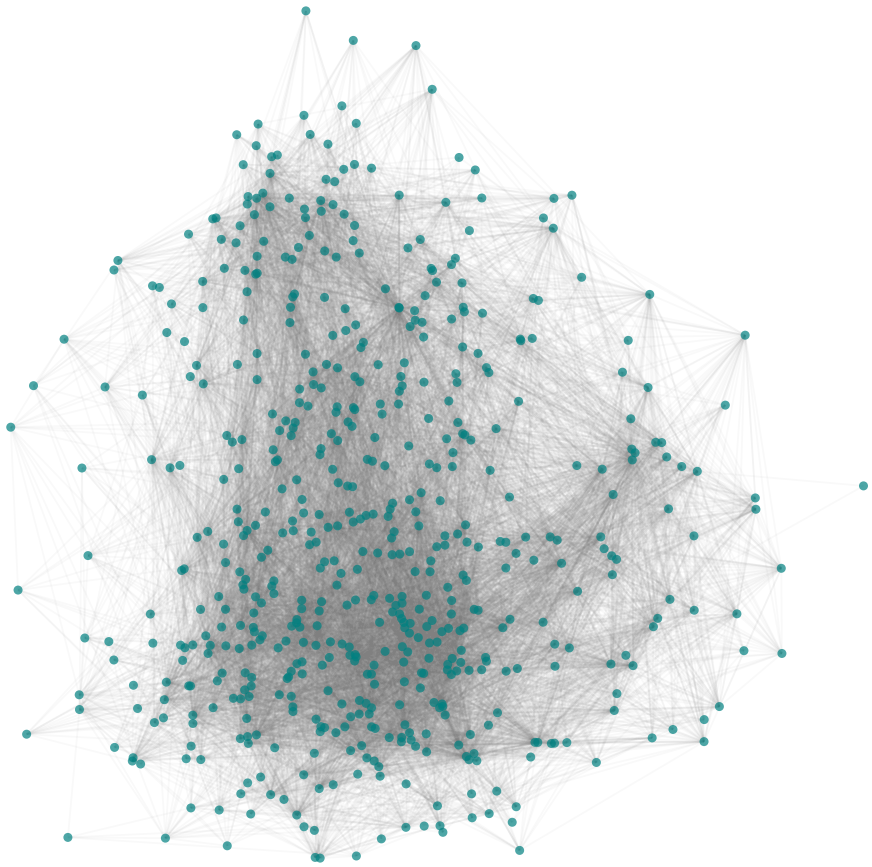

Figure 7.6 다차원 척도법 레이아웃 | ‹G-score 네트워크›

적인 비선형 차원 축소 기법이다(McInnes et al. 2018). 이 알고리즘은 위상수학(Topology)에 바탕을 둔다. 데이터가 고차원 공간에서 굽어 있거나 복잡하게 얽혀 있더라도 노드 사이의 국소 연결 관계와 위상적 형태 패턴을 최대한 보존하며 평면에 펼쳐낸다.

원리는 크게 두 단계로 이루어진다. 먼저 고차원 네트워크에서 각 노드가 이웃 노드와 얼마나 강하게 연결되어 있는지를 확률적으로 모델링하여 국소 위상 구조를 구성한다. 이어 2차원 평면에서도 이 연결 확률 구조가 유지되도록 노드 좌표를 배치하고 고차원과 저차원 구조 사이의 오차(교차 엔트로피 손실)를 최소화하도록 좌표를 점진적으로 최적화한다. 대규모 그래프에서도 연산 속도가 빠르다는 장점이 있다.

UMAP의 가장 큰 특징은 비선형 변환으로 커뮤니티 구조를 선명하게 드러낸다는 점이다. 엣지 가중치가 높은 노드는 2차원 평면에서 조밀하게 뭉치고 결합이 약한 그룹과는 분명하게 분리된다. 그 결과 다른 알고리즘보다 세부 커뮤니티 경계를 뚜렷하게 보여준다. 다만 커뮤니티 내부의 결속 관계는 신뢰할 수 있어도 멀리 떨어진 커뮤니티 사이의 물리적 거리 자체가 실제 관계의 거리와 정확히 비례하지는 않는다는 점에 유의해야 한다. 아울러 확률적 최적화 과정을 거치므로 난수 시드(Seed)에 따라 시각적 배치가 조금씩 달라질 수 있다.

7.3 레이아웃 알고리즘 비교

지금까지 네트워크 시각화에 널리 쓰이는 네 가지 레이아웃 알고리즘을 간략히 살펴보고 이를 본초 네트워크에 적용한 시각화 결과를 비교해 보았다. 각 알고리즘의 주요 특성을 정리하면 다음과 같다.

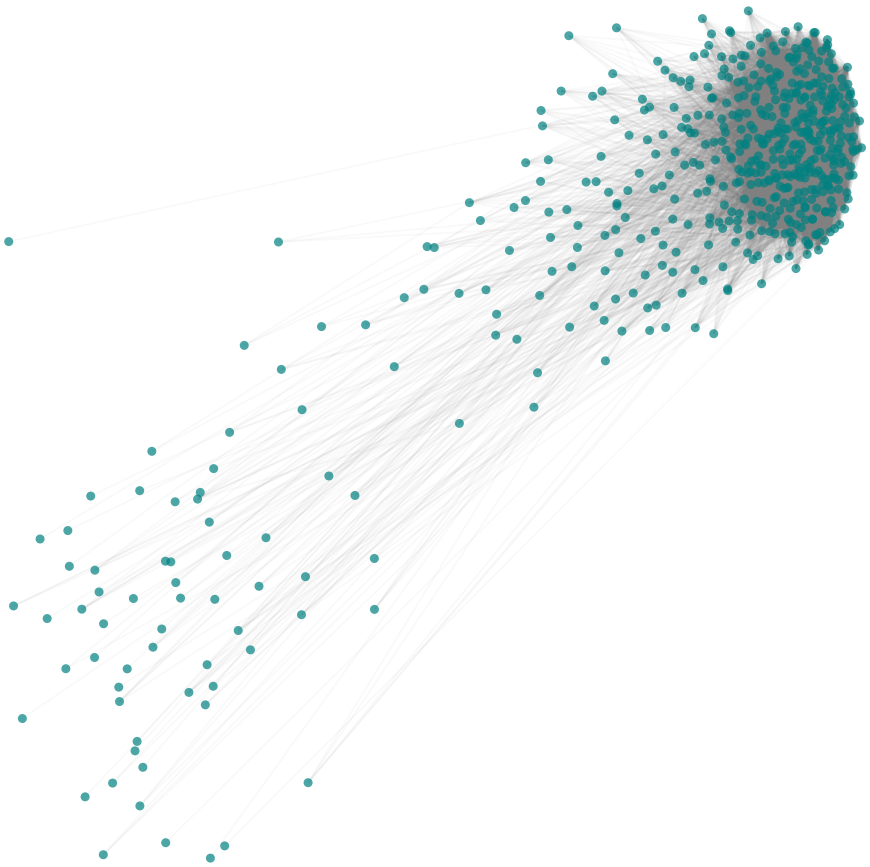

Figure 7.7 UMAP 레이아웃 | ‹단순 빈도 네트워크›

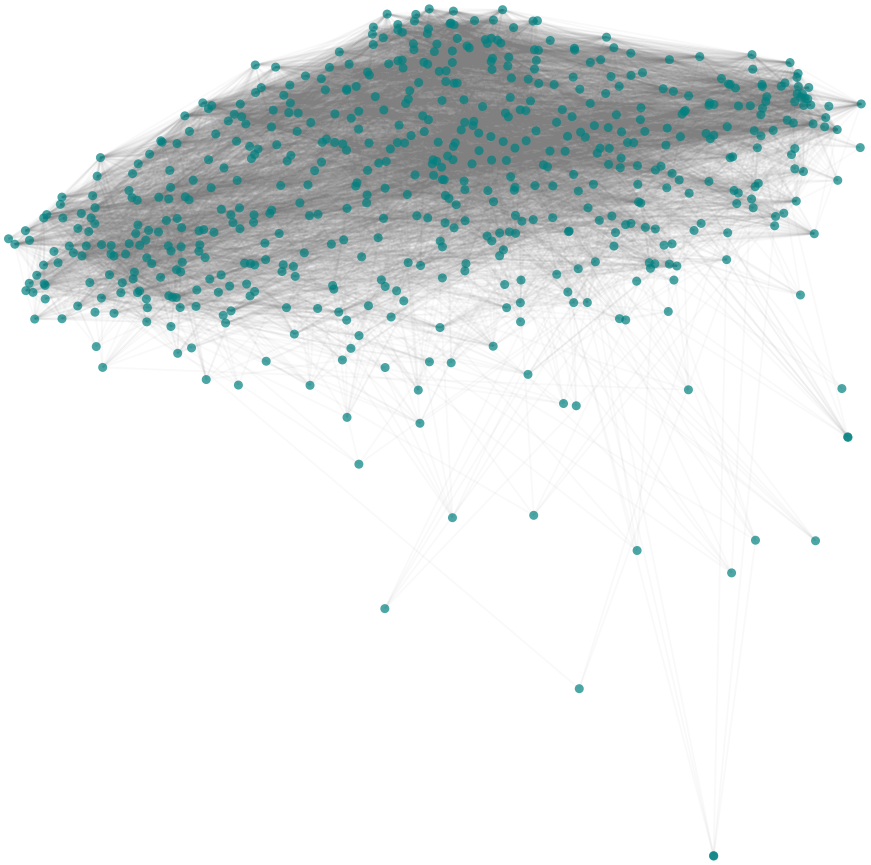

Figure 7.8 UMAP 레이아웃 | ‹G-score 네트워크›

Table 7.1 주요 네트워크 레이아웃 알고리즘 특성 비교

알고리즘	핵심 아이디어	엣지 가중치 의미	보존되는 특성	왜곡되는 특성	주요 추천 용도
Fruchterman-Reingold	국소 힘 평형 (휴리스틱)	인력 강도	이웃 노드 간 국소 연결 구조	전역 최단 경로 거리 (중앙 쏠림 발생)	일반적인 네트워크 시각화 표준
Kamada-Kawai (KK)	에너지 최소화	엣지 거리 (반전 필요)	노드 간 전역 최단 경로 거리	국소 커뮤니티 경계 (밀집 영역 뭉개짐)	소·중규모 네트워크의 거시 구조 탐색
MDS	선형대수 (고유분해)	엣지 거리 (반전 필요)	전역 기하학적 거리 및 분산	비선형 다양체 구조 및 국소 세부 관계	대수적 최단 경로 좌표화 (결정론적)
UMAP	비선형 다양체 임베딩	연결 강도	국소 이웃 관계 및 커뮤니티 분리	커뮤니티 간 전역 물리적 거리	복잡한 커뮤니티 구조의 선명한 포착

네트워크의 중심성과 시각적 레이아웃을 살펴보았으니 이제 개별 노드가 모여 형성하는 커뮤니티와 이를 응용한 본초의 다의성을 알아본다.

제 8 장

커뮤니티와 다의성

지금까지 네트워크의 전체 골격과 통계적 필연성, 시각화 방법까지 살펴보았다. 이 장에서는 네트워크 내부의 세부 기능을 파악하기 위해 커뮤니티(Community) 탐지 기법을 알아보고 본초를 기능적 배합 그룹으로 구분해 보자. 아울러 자연어 처리(NLP) 분야의 중의성 해소 개념을 응용하여 본초가 지닌 다의적(Polysemic) 특성을 검토해 보자.

8.1 커뮤니티 탐지와 모듈성

복잡한 네트워크 구조 내부에는 노드 간 연결이 밀집된 서브네트워크가 존재한다. 네트워크 분석에서는 이를 '커뮤니티'라고 한다. **커뮤니티(Community)**는 외부 노드와의 연결 밀도에 비해 내부 노드 간 연결 가중치 합이 훨씬 크게 나타나는 노드 집합을 뜻한다. 상황에 따라 '모듈(Module)'이나 '군집(Cluster)'이라고도 부른다.

네트워크가 얼마나 뚜렷한 집단으로 나뉘는지 측정하는 통계 지표를 모듈성이라고 한다. **모듈성(Modularity, Q)**은 전체 네트워크를 여러 그룹으로 나누었을 때, 우연히 발생할 법한 연결보다 실제 같은 그룹 내에 존재하는 연결이 얼마나 더 많은지를 나타낸다(Newman 2018). 수학적으로

는 각 노드의 연결 강도(가중치 차수)를 보존하는 배치 모델(Configuration Model)을 바탕으로 임의의 두 노드 i, j가 우연히 연결될 기댓값($\frac{k_i k_j}{2m}$)과 실제 연결 가중치(A_{ij})의 편차를 계산한다. 동일한 커뮤니티에 속한 노드 쌍($c_i = c_j$)의 편차만 누적 합산하고 이를 전체 엣지 가중치 합의 2배이자 모든 노드 강도의 총합인 $2m$으로 정규화한다. 모듈성 점수(Q)가 0.3을 넘으면 무작위 기댓값을 유의미하게 웃도는 뚜렷한 커뮤니티 구조가 형성된 것으로 판정한다.

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

(m: 전체 엣지 가중치 합, A_{ij}: 실제 연결 가중치, $\frac{k_i k_j}{2m}$: 무작위 상태의 연결 기댓값, k_i, k_j: 노드 i, j의 연결 가중치 합, $\delta(c_i, c_j)$: 두 노드가 같은 커뮤니티에 속하면 1, 아니면 0)

전체 네트워크에서 커뮤니티를 찾는 작업을 '커뮤니티 탐지'라고 한다. **커뮤니티 탐지(Community Detection)**는 네트워크의 위상 구조와 연결 강도를 바탕으로 서로 밀접하게 연관된 노드 부분집합을 식별하는 과정이다.

초기 커뮤니티 탐지 기법은 커뮤니티 사이를 잇는 가교 엣지를 찾아 하나씩 제거해 나가는 분할(Divisive) 방식에서 출발했다. 거번-뉴먼(Girvan-Newman) 알고리즘(Girvan and Newman 2002)이 대표적이다. 개념은 직관적이지만 전체 경로를 반복해서 다시 계산해야 하므로 네트워크가 커지면 속도가 급격히 느려지는 한계가 있었다. 이후 연구는 그룹 내부의 결속 정도를 수학적으로 극대화하는 방식과 노드 사이를 오가는 상호작용 경로를 추적하는 방식으로 발전했다.

오늘날 주요 커뮤니티 탐지 기법은 모듈성 최적화, 정보 흐름 분석, 이웃 기반 전파의 세 가지 축으로 나눌 수 있다. 첫째, 모듈성 최적화 기법은 네트워크가 뚜렷하게 분할되도록 모듈성(Q)을 최대화하는 방향으로 노드를 묶어 나간다. 루방(Louvain)(Blondel et al. 2008)과 라이덴

(Leiden)(Traag et al. 2019) 알고리즘이 여기에 속한다. 계산 속도가 빠르고 커뮤니티 구조를 효과적으로 찾아낸다. 둘째, 정보 흐름 기법은 네트워크 위를 움직이는 신호나 이동 경로를 추적한다. 한 그룹 안에 머무는 체류 패턴을 포착하는 인포맵(Infomap) 알고리즘(Rosvall and Bergstrom 2008)[1]이 대표적이다. 엣지에 방향이 있거나 세부적인 작은 커뮤니티를 정밀하게 탐색할 때 유리하다. 셋째, 이웃 기반 전파 기법은 각 노드가 주변 이웃이 가장 많이 속한 그룹을 따라 소속을 갱신하는 방식이다. 라벨 전파(LPA) 알고리즘(Raghavan et al. 2007)이 여기에 해당하며 계산량이 매우 적어 수백만 개 이상의 거대한 네트워크도 빠르게 처리할 수 있다.

[1] 인포맵(Infomap): 네트워크 위를 움직이는 무작위 보행자의 이동 경로를 부호화할 때 정보 압축률(Map Equation)을 최대화하는 방식으로 최적의 커뮤니티 경계를 탐색하는 흐름 기반 알고리즘이다.

Table 8.1 주요 커뮤니티 탐지 알고리즘 비교

알고리즘	최적화 목표	계산 속도 및 확장성	가중치 지원	주요 특징 및 적합 도메인
거번-뉴먼 (Girvan-Newman)	엣지 매개 중심성	느림 (소규모 전용)	지원	가교 연결 순차 제거를 통한 하향식 분할, 계층적 덴드로그램 시각화에 적합
루방 (Louvain)	모듈성(Q) 최대화	매우 빠름 (대규모 처리)	지원	탐욕적 계층 병합 방식의 고속 표준 기법, 불연속 커뮤니티 결함 및 해상도 한계 존재
라이덴 (Leiden)	모듈성(Q) 또는 CPM[2] 최대화	매우 빠름 (대규모 처리)	지원	서브 커뮤니티 정제(Refinement)로 불연속 커뮤니티 결함 해결, 구조적 연결성 엄밀 보장
인포맵 (Infomap)	경로 설명 길이 (Map Equation) 최소화	빠름 (대규모 처리)	지원	네트워크 내 흐름과 정보 전파 추적, 방향성 네트워크 및 미세 커뮤니티 탐색에 최적
라벨 전파 (LPA)	이웃 노드 소속 라벨의 다수결 일치	극도로 빠름 (초대규모 처리)	지원	목적함수 없는 초고속 이웃 전파 기법, 실행 시 무작위성으로 결과 변동 가능

[2] CPM(Constant Potts Model)은 해상도 한계(Resolution Limit) 문제 없이 작은 규모의 커뮤니티까지 안정적으로 탐지할 수 있도록 고안된 품질 함수(Quality Function)이다.

이 책에서는 커뮤니티 탐지를 위해 **라이덴 알고리즘(Leiden Algorithm)**(Traag et al. 2019)을 사용한다. 루방 알고리즘은 빠르지만 서로 연결되지 않은 노드까지 한 커뮤니티로 묶는 결함이 있다. 라이덴 알고리즘은 커뮤니티 내부를 한 번 더 점검(정제)하여 실제로 긴밀하게 이어진 노드만 남도록 개선한 기법이다. 계산 속도가 빠르고 엣지 가중치도 정확히 반영한다.

라이덴 알고리즘으로 본초 네트워크의 커뮤니티를 탐지해 보자. 모듈성 점수를 살펴보면 ‹단순 빈도 네트워크›는 0.157로 비교적 낮지만 ‹G-score 네트워크›는 0.489로 뚜렷한 커뮤니티 구조를 나타냈다.

Table 8.2 네트워크별 커뮤니티 탐지 및 모듈성 비교

구분	커뮤니티 수	모듈성(Modularity, Q) 점수
‹단순 빈도 네트워크›	5 개	0.1572
‹G-score 네트워크›	6 개	0.4889

알고리즘으로 탐지한 커뮤니티의 세부 구성은 다음과 같다. 커뮤니티는 포함된 노드 수 기준 내림차순으로 정렬하고 각 커뮤니티 내부에서 출현 빈도가 높은 상위 20개 본초를 표시하였다.

Table 8.3 전역 커뮤니티 탐지 결과 (출현 빈도 상위 20개) | ‹단순 빈도 네트워크›

커뮤니티 구분	노드 개수	출현 빈도 상위 본초
커뮤니티 A	152개	대황(大黃), 황련(黃連), 사향(麝香), 주사(硃砂), 망초(芒硝), 황백(黃柏), 유향(乳香), 백반(白礬), 서우각(犀牛角), 웅황(雄黃), 오두(烏頭), 천남성(天南星), 빙편(冰片), 파두(巴豆), 전갈(全蠍), 천마(天麻), 몰약(沒藥), 경분(輕粉), 백강잠(白僵蠶), 우황(牛黃)
커뮤니티 B	144개	감초(甘草), 복령(茯苓), 인삼(人蔘), 작약(芍藥), 황금(黃芩), 반하(半夏), 맥문동(麥門冬), 시호(柴胡), 길경(桔梗), 행인(杏仁), 승마(升麻), 치자(梔子), 마황(麻黃), 석고(石膏), 생강(生薑), 목통(木通), 지모(知母), 상백피(桑白皮), 갈근(葛根), 현삼(玄蔘)
커뮤니티 C	116개	당귀(當歸), 육계(肉桂), 부자(附子), 생지황(生地黃), 황기(黃芪), 숙지황(熟地黃), 오미자(五味子), 우슬(牛膝), 목단피(牡丹皮), 택사(澤瀉), 천초(川椒), 원지(遠志), 아교(阿膠), 산약(山藥), 용골(龍骨), 산수유(山茱萸), 육종용(肉蓯蓉), 진교(秦艽), 차전자(車前子), 석창포(石菖蒲)
커뮤니티 D	88개	백출(白朮), 진피(陳皮), 목향(木香), 건강(乾薑), 지각(枳殼), 계지(桂枝), 빈랑(檳榔), 후박(厚朴), 청피(靑皮), 창출(蒼朮), 가자(訶子), 정향(丁香), 향부(香附), 오수유(吳茱萸), 도인(桃仁), 사인(砂仁), 침향(沉香), 지실(枳實), 신곡(神麴), 육두구(肉豆蔻)
커뮤니티 E	63개	천궁(川芎), 방풍(防風), 세신(細辛), 강활(羌活), 백지(白芷), 형개(荊芥), 독활(獨活), 영양각(羚羊角), 박하(薄荷), 연교(連翹), 감국(甘菊), 백질려(白蒺藜), 고삼(苦蔘), 우방자(牛蒡子), 선태(蟬蛻), 용담초(龍膽草), 만형자(蔓荊子), 고본(藁本), 금은화(金銀花), 위령선(威靈仙)

Table 8.4 전역 커뮤니티 탐지 결과 (출현 빈도 상위 20개) | ‹G-score 네트워크›

커뮤니티 구분	노드 개수	출현 빈도 상위 본초
커뮤니티 a	176개	사향(麝香), 주사(硃砂), 유향(乳香), 백반(白礬), 웅황(雄黃), 오두(烏頭), 천남성(天南星), 빙편(冰片), 파두(巴豆), 전갈(全蠍), 천마(天麻), 몰약(沒藥), 경분(輕粉), 백강잠(白僵蠶), 우황(牛黃), 조각(皂角), 백부자(白附子), 청대(靑黛), 유황(硫黃), 연단(鉛丹)
커뮤니티 b	129개	감초(甘草), 황금(黃芩), 대황(大黃), 황련(黃連), 맥문동(麥門冬), 시호(柴胡), 길경(桔梗), 행인(杏仁), 승마(升麻), 치자(梔子), 석고(石膏), 망초(芒硝), 생강(生薑), 황백(黃柏), 목통(木通), 지모(知母), 상백피(桑白皮), 서우각(犀牛角), 갈근(葛根), 현삼(玄蔘)
커뮤니티 c	77개	복령(茯苓), 인삼(人蔘), 육계(肉桂), 부자(附子), 황기(黃芪), 숙지황(熟地黃), 오미자(五味子), 우슬(牛膝), 택사(澤瀉), 천초(川椒), 원지(遠志), 산약(山藥), 용골(龍骨), 산수유(山茱萸), 육종용(肉蓯蓉), 석창포(石菖蒲), 두중(杜仲), 석곡(石斛), 모려(牡蠣), 토사자(菟絲子)
커뮤니티 d	73개	천궁(川芎), 방풍(防風), 세신(細辛), 강활(羌活), 마황(麻黃), 백지(白芷), 형개(荊芥), 독활(獨活), 영양각(羚羊角), 박하(薄荷), 연교(連翹), 감국(甘菊), 진교(秦艽), 차전자(車前子), 방기(防己), 백질려(白蒺藜), 고삼(苦蔘), 우방자(牛蒡子), 선태(蟬蛻), 만형자(蔓荊子)
커뮤니티 e	68개	백출(白朮), 진피(陳皮), 목향(木香), 반하(半夏), 건강(乾薑), 지각(枳殼), 계지(桂枝), 빈랑(檳榔), 후박(厚朴), 청피(靑皮), 창출(蒼朮), 가자(訶子), 정향(丁香), 향부(香附), 오수유(吳茱萸), 사인(砂仁), 침향(沉香), 지실(枳實), 신곡(神麴), 육두구(肉豆蔻)
커뮤니티 f	57개	당귀(當歸), 작약(芍藥), 생지황(生地黃), 목단피(牡丹皮), 도인(桃仁), 아교(阿膠), 애엽(艾葉), 연호색(延胡索), 홍화(紅花), 지유(地楡), 포황(蒲黃), 건칠(乾漆), 측백엽(側柏葉), 택란(澤蘭), 백미(白薇), 괴각(槐角), 소목(蘇木), 포강(炮薑), 복룡간(伏龍肝), 괴화(槐花)

Table 8.5 두 네트워크의 전역 커뮤니티 중첩 비교 (공통 노드 수 및 자카드 지수)

공통 노드	커뮤니티 a (176종)	커뮤니티 b (129종)	커뮤니티 c (77종)	커뮤니티 d (73종)	커뮤니티 e (68종)	커뮤니티 f (57종)
커뮤니티 A (152종)	138 (0.73)	8 (0.03)	0 (0.00)	3 (0.01)	1 (0.00)	2 (0.01)
커뮤니티 B (144종)	7 (0.02)	109 (0.66)	8 (0.04)	10 (0.05)	6 (0.03)	4 (0.02)
커뮤니티 C (116종)	4 (0.01)	3 (0.01)	69 (0.56)	6 (0.03)	0 (0.00)	34 (0.24)
커뮤니티 D (88종)	11 (0.04)	4 (0.02)	0 (0.00)	1 (0.01)	61 (0.64)	11 (0.08)
커뮤니티 E (63종)	6 (0.03)	2 (0.01)	0 (0.00)	52 (0.62)	0 (0.00)	3 (0.03)

두 네트워크의 커뮤니티 탐지 결과를 비교하면 공통점과 차이점이 확인된다. ‹단순 빈도 네트워크›와 ‹G-score 네트워크›는 엣지 형성 기준이 서로 다르지만, 동일한 처방 데이터를 바탕으로 구축했으므로 유사한 커뮤니티가 포착되었다. ‹단순 빈도 네트워크›의 커뮤니티 A 및 ‹G-score 네트워크›의 커뮤니티 a(사향(麝香), 주사(硃砂) 등 138종 공유, 자카드 지수 0.73), ‹단순 빈도 네트워크›의 커뮤니티 B 및 ‹G-score 네트워크›의 커뮤니티 b(감초(甘草), 황금(黃芩) 등 109종 공유, 자카드 지수 0.66), ‹단순 빈도 네트워크›의 커뮤니티 D 및 ‹G-score 네트워크›의 커뮤니티 e(백출(白朮), 진피(陳皮) 등 61종 공유, 자카드 지수 0.64) 등이 대표적이다. 반면 세부적으로는 엣지 필터링 기준의 차이로 관련성이 거의 없는 서로 다른 커뮤니티도 함께 나타났다.

커뮤니티 탐지 결과를 시각화해 보자.

(각 레이아웃은 엣지 가중치를 사용해 노드 위치를 결정했으며, 커뮤니티 소속은 색상으로만 표시했다. 레이아웃 알고리즘 특성에 따라 커뮤니티 경계의 가시성은 달라지지만 색상으로 구분된 커뮤니티 구성은 동일하다.)

‹단순 빈도 네트워크›에서는 커뮤니티 경계가 상대적으로 뚜렷하지 않았다. 이는 ‹단순 빈도 네트워크›의 모듈성 점수(Q)가 0.157로 낮게 측정된 결과와 부합한다. 반면 ‹G-score 네트워크›에서는 커뮤니티 경계가 더 선명하게 구분되었다. 상대적으로 높은 모듈성 점수 0.489 역시 이러한 구조를 뒷받침한다. 이는 전자에서는 다빈도 본초들이 여러 군집 간에 광범위한 배경 연결을 형성하여 경계가 흐려진 반면, 후자에서는 통계적 유의성 검정을 통해 이러한 비특이적 결합이 억제되고 기능적 군집 경계가 뚜렷해졌기 때문으로 풀이할 수 있다.

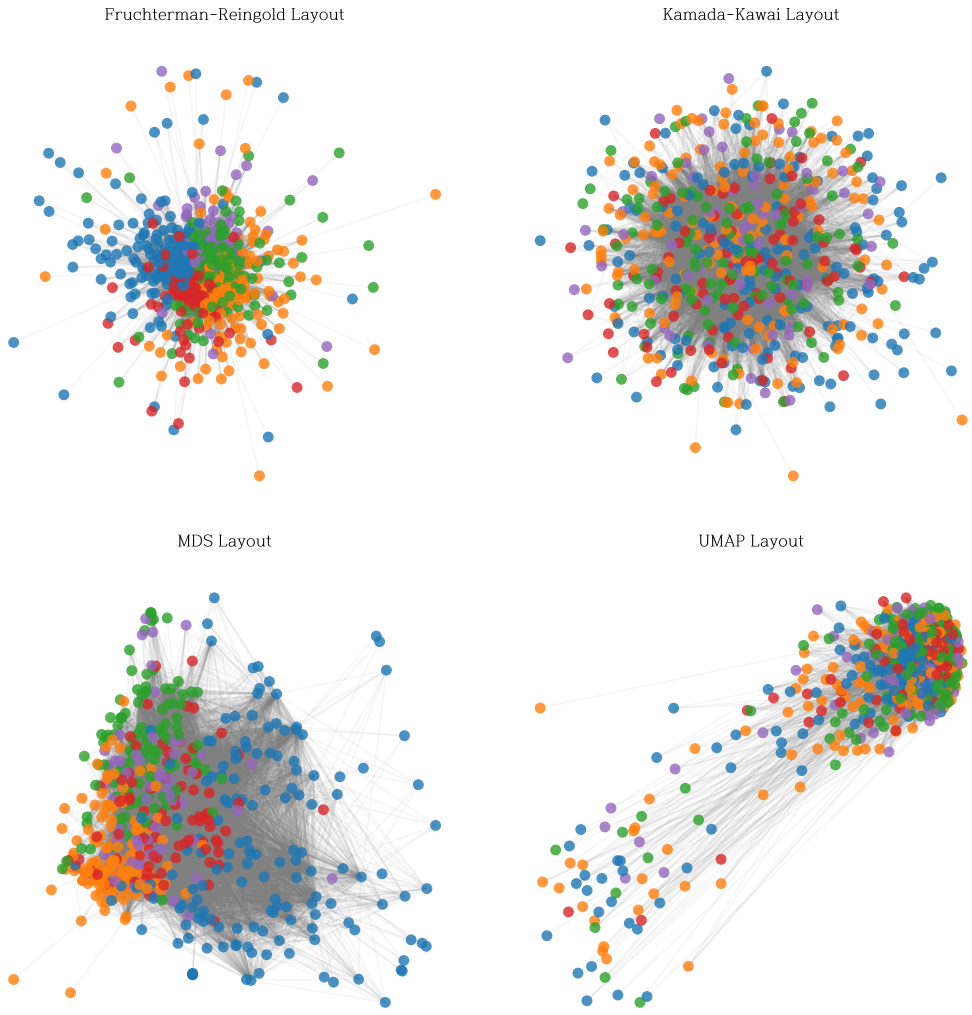

Figure 8.1 전역 커뮤니티 탐지 결과 | ‹단순 빈도 네트워크›

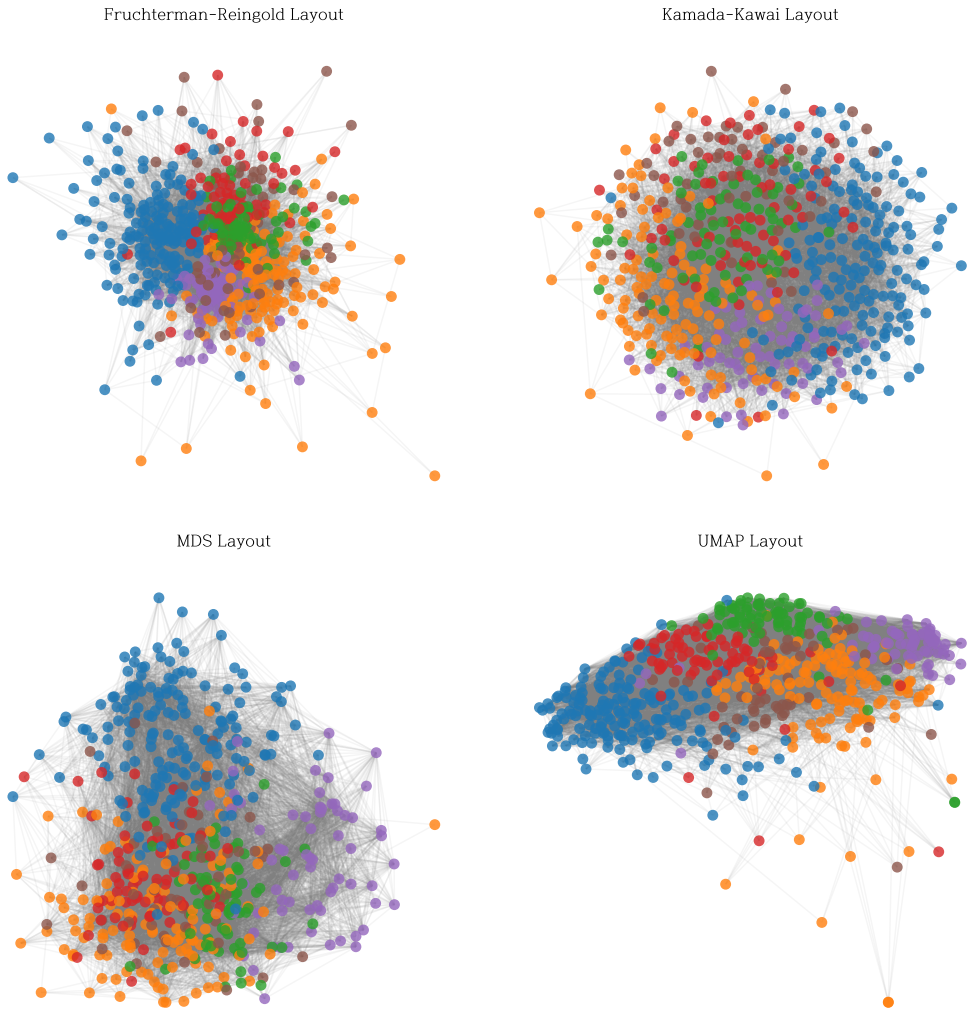

Figure 8.2 전역 커뮤니티 탐지 결과 | ‹G-score 네트워크›

8.2 본초의 다의성

네트워크 분석에서는 개체의 속성이 고정되어 따로 존재하는 것이 아니라 주변과의 관계 속에서 규정된다고 본다. 이는 단어의 의미가 서술 문맥과 이웃 단어의 관계에 따라 결정된다는 언어학의 구조주의적 관점과 맞닿아 있다.

단어의 의미가 문맥에 따라 달라지듯, 한의학 본초 역시 함께 배합하는 본초 조합에 따라 기대 효능이 달라질 수 있다. 자연어 처리(NLP) 분야에서는 사전 정의에 의존하지 않고 데이터 문맥으로부터 단어의 다양한 의미를 추론하는 단어 의미 귀납(Word Sense Induction, WSI) 기술을 연구해 왔다(Di Marco and Navigli 2013). 이 책에서는 이 개념을 본초의 다의성 분석에 응용하여, 주변 본초 배합에 따른 특정 본초의 역할을 도출하는 과제를 **본초 의미 귀납(Herb Sense Induction, HSI)**이라고 부르기로 한다.

본초 의미 귀납은 여러 방식으로 접근할 수 있을 것이다. 그러나 네트워크 분석에 대해 살펴보고 있는 만큼 여기서는 커뮤니티 탐지를 바탕으로 이 문제를 살펴보고자 한다. 핵심 원리는 단순하다. 대상 본초와 연결된 주변 본초를 커뮤니티로 구분하는 것이다. 구분된 각 커뮤니티가 곧 해당 본초가 활용된 세부 배합 문맥을 나타내기 때문이다.

구체적인 분석 절차는 다음과 같다. 먼저 대상 본초 v와 직접 연결된 1차 이웃 노드로 에고 네트워크(Ego Network)[3]를 구축한다. 다음으로 중심 노드 v를 일시적으로 제외하고, 남은 이웃 노드를 대상으로 커뮤니티 탐지를 수행한다. 대상 본초가 여러 치료 맥락으로 활용되었다면 이웃 노드는 기능에 따라 별개의 커뮤니티를 이루며 분리될 것이다. 이때 도출된 각 하위 커뮤니티는 대상 본초 v가 사용된 구체적인 한의학적 문맥이라고 할 수 있다.

[3] 에고 네트워크에 대해서는 제9장에서 살펴본다.

본초의 다의성을 정량화하기 위해 다음 두 가지 지표를 살펴보자.

◎ **모듈성(Q, Modularity)**

모듈성은, 대상 본초를 제외하고 남은 에고 네트워크가 얼마나 뚜렷하게 분리되어 하위 커뮤니티를 형성하는지 측정한다. 에고 네트워크에서 2개 이상의 커뮤니티가 탐지되고 모듈성 Q가 높다는 것은, 이웃 본초가 임의로 뒤섞이지 않고 기능별로 명확히 나뉜다는 뜻이다. 따라서 대상 본초가 주변 본초에 따라 서로 다른 맥락에서 사용될수록 모듈성 Q가 커진다.

◎ **섀넌 엔트로피(H, Shannon Entropy)**

섀넌 엔트로피는 정보이론에서 확률 분포의 불확실성과 다양성을 정량화하는 척도이다(Shannon 1948). 이를 활용하면 대상 본초가 세부 문맥 사이에서 얼마나 균등하게 사용되는지 평가할 수 있다. 대상 본초 v의 에고 네트워크가 K개의 세부 배합 문맥으로 분할되고 각 문맥 k로 분배되는 엣지 연결 가중치 비율을 p_k라 할 때 다음과 같이 정의된다. 수학적으로는 이산 확률 질량 분포 p_k에 대한 정보량의 기댓값($-\sum p_k \log_2 p_k$)에 해당하며, 문맥 분포의 균등성을 비트(bit) 단위로 정량화한다.

$$H(v) = -\sum_{k=1}^{K} p_k \log_2 p_k$$

(p_k: 타깃 본초의 전체 연결 엣지 가중치 중 서브 커뮤니티 k에 분배된 비중. K: 분할된 에고 서브 커뮤니티 총개수)

섀넌 엔트로피의 이론적 범위는 $0 \leq H(v) \leq \log_2 K$이다. 최솟값인 0은 대상 본초의 배합 연결이 특정 단일 문맥에 전부 편중되었을 때 나타난다($p_1 = 1$, 나머지 0). 반대로 최댓값인 $\log_2 K$는 K개의 모든 배합 문맥에 가중치가 정확히 균등하게 분산될 때($p_k = 1/K$) 도달한다.

즉, 엔트로피 값이 클수록 대상 본초가 여러 배합 문맥에 걸쳐 한쪽으로 치우치지 않고 고르게 활용되었음을 뜻한다.[4]

Q와 H는 서로 보완적인 성격을 가지고 있다. 모듈성(Q)이 서로 다른 하위 문맥의 '존재'를 보여준다면 엔트로피(H)는 해당 문맥의 '사용 비중'을 평가한다. 아무리 하위 커뮤니티가 잘 분리되어 Q가 높더라도 특정 한 문맥에 편중되어 사용되었다면 실질적으로는 단일 용도로 쓰인 것과 다름없다. 반면 여러 문맥에 걸쳐 고르게 사용되었다면 다의성이 크다고 볼 수 있다.

제5장에서 규명한 중심부 공통 본초 117종을 대상으로 각 네트워크에서 에고 네트워크를 분할할 때 나타나는 모듈성(Q)과 섀넌 엔트로피(H)의 전반적인 분포를 살펴보자.

아래 두 산점도는 각 네트워크에서 중심부 공통 본초의 에고 네트워크 모듈성(Q, X축)과 섀넌 엔트로피(H, Y축)를 비교하여 나타낸 것이다.

[4] 엔트로피 값을 이론상 최댓값으로 나누어 0.0과 1.0 사이로 정규화하는 균등도 지수(Pielou's Evenness Index, $H/\log_2 K$)의 적용도 고려할 수 있다. 그러나 균등도 지수는 커뮤니티 개수(K)의 효과를 완전히 배제하고 오직 비율의 균등함만 측정한다. 이를 적용하면 2개의 커뮤니티에 절반씩 쓰인 본초가, 여러 처방 문맥에 걸쳐 광범위하게 쓰인 핵심 본초보다 더 높은 값을 갖는 왜곡이 발생한다. 다의성은 활용 문맥의 개수와 배분 균등성이 함께 결합될 때 비로소 커지므로, 본 연구에서는 문맥 수의 규모를 보존하는 절대 섀넌 엔트로피를 분석 지표로 채택한다.

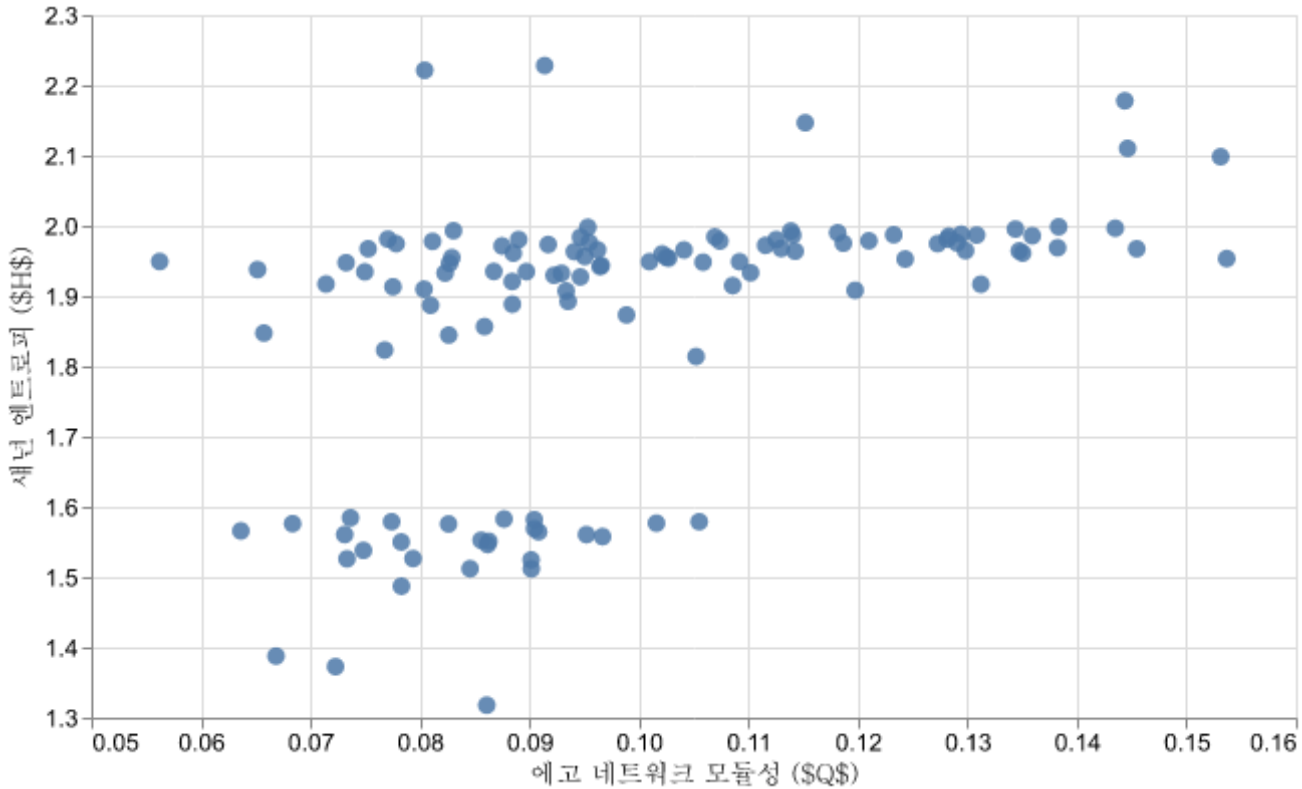

Figure 8.3 에고 네트워크 다의성 지표 평면 분포 | ‹단순 빈도 네트워크›

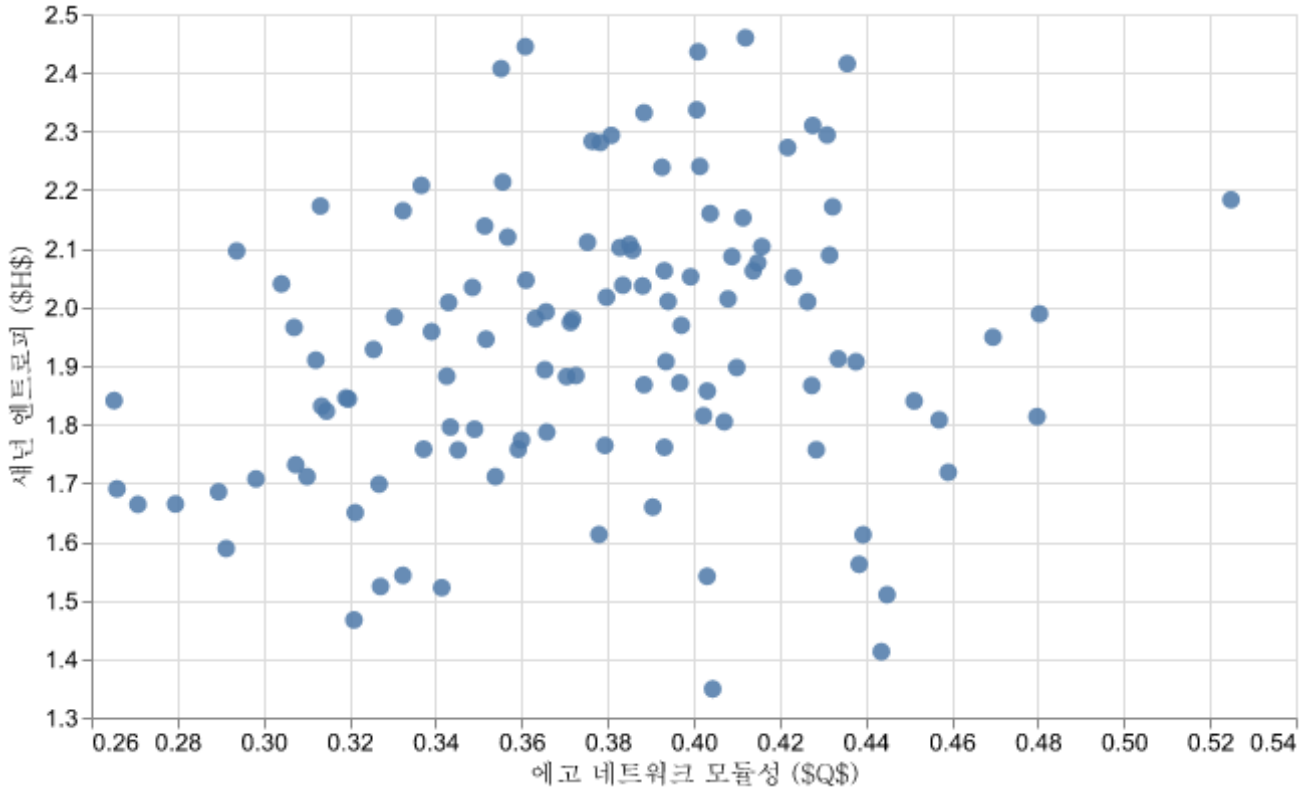

Figure 8.4 에고 네트워크 다의성 지표 평면 분포 | ‹G-score 네트워크›

두 산점도를 비교하면 네트워크 모델에 따른 구조적 차이가 나타난다. 먼저 ‹단순 빈도 네트워크›에서는 주요 본초의 이웃 노드가 서로 빽빽하게 이어져 있다. 이 때문에 에고 네트워크의 모듈성(Q, X축)이 대부분 낮은 구간(0.06 ~ 0.15)에 머물러 처방 목적에 따른 세부 문맥이 잘 구분되지 않는다.

Y축의 섀넌 엔트로피(H)를 보면 관측점들이 $H \approx 1.58$과 $H \approx 2.00$ 부근에 수평 띠 모양으로 몰려 있다. 앞서 정의한 바와 같이 섀넌 엔트로피는 각 커뮤니티의 연결 비중이 균등할 때($p_k \approx 1/K$) 이론상 최댓값인 $\log_2 K$에 도달한다. 커뮤니티가 3개이면 $\log_2 3 \approx 1.585$, 4개이면 $\log_2 4 = 2.000$이다. ‹단순 빈도 네트워크›에서는 다빈도 본초가 하위 커뮤니티 노드와 거의 고르게 연결되다 보니 세부 차이가 드러나지 않고 엔트로피 값이 커뮤니티 개수에 따라 최댓값에 수렴하는 한계를 보였다.

반면 ‹G-score 네트워크›에서는 G-score를 통해 유의성이 낮은 결합을 걷어내면서 모듈성(Q, X축)이 크게 높아졌다(0.27 ~ 0.52). 이웃 본초가 처방 목적에 맞춰 뚜렷한 커뮤니티로 나뉜다는 뜻이다. 특정 수치에 몰리던 수평 띠 현상도 사라지고 엔트로피(H, Y축) 값 역시 넓게 분포한다. 본초마다 서로 다른 배합 특성이 선명하게 드러난 셈이다.

‹G-score 네트워크›에서 모듈성과 엔트로피가 모두 높아 산점도 우상단에 위치한 본초는 배합 맥락에 따라 다양하게 쓰이는 대표적인 다의성 본초이다. 이들을 살펴보기 위해 모듈성과 엔트로피에 기준값을 적용하였다. 모듈성(Q)은 네트워크 분석에서 커뮤니티 구조가 유의미하게 형성되었다고 판단하는 일반적 기준인 $Q \geq 0.30$을 적용하였다. 섀넌 엔트로피(H)는 분할된 세부 커뮤니티 수(K)에 따라 이론적 상한($\log_2 K$)이 달라진다. 따라서 고정된 단일 상수를 적용하기보다 데이터의 실증 분포를 반영하는 중앙값($H \geq 1.945$)을 기준으로 삼았다. 중심부 공통 본초 117종 가운데 두 기준을 동시에 만족하는 본초는 다음과 같다.

다의성 핵심 본초 (총 58종):

황련(黃連), 황백(黃柏), 오두(烏頭), 영양각(羚羊角), 오약(烏藥), 대황(大黃), 육계(肉桂), 계지(桂枝), 건강(乾薑), 반하(半夏), 행인(杏仁), 백지(白芷), 부자(附子), 차전자(車前子), 감초(甘草), 인삼(人蔘), 백출(白朮), 목통(木通), 복령(茯苓), 지각(枳殼), 당귀(當歸), 맥문동(麥門冬), 택사(澤瀉), 귤홍(橘紅), 정향(丁香), 생지황(生地黃), 마황(麻黃), 천초(川椒), 천궁(川芎), 목향(木香), 세신(細辛), 감국(甘菊), 시호(柴胡), 강활(羌活), 오수유(吳茱萸), 빈랑(檳榔), 길경(桔梗), 독활(獨活), 목단피(牡丹皮), 작약(芍藥), 지모(知母), 후박(厚朴), 연교(連翹), 만형자(蔓荊子), 고본(藁本), 황금(黃芩), 고삼(苦蔘), 적석지(赤石脂), 황기(黃芪), 진교(秦艽), 현삼(玄蔘), 의이인(薏苡仁), 치자(梔子), 청피(靑皮), 석창포(石菖蒲), 천화분(天花粉), 승마(升麻), 사인(砂仁)

각 네트워크에서 에고 네트워크 모듈성(Q)을 기준으로 도출한 다의성 상위 20개 본초 목록은 다음과 같다.

Table 8.6 에고 네트워크 기반 다의성 상위 본초 (상위 20개) | ‹단순 빈도 네트워크›

본초	커뮤니티 수	모듈성(Q)	엔트로피(H)
감초(甘草)	4	0.1537	1.9534
당귀(當歸)	5	0.1531	2.0984
복령(茯苓)	4	0.1455	1.9673
천궁(川芎)	5	0.1446	2.1102
작약(芍藥)	5	0.1444	2.178
인삼(人蔘)	4	0.1435	1.9966
육계(肉桂)	4	0.1383	1.9986
방풍(防風)	4	0.1382	1.9688
목향(木香)	4	0.1359	1.9859
황련(黃連)	4	0.135	1.9612
대황(大黃)	4	0.1347	1.9644
백지(白芷)	4	0.1344	1.9955
생지황(生地黃)	4	0.1312	1.9169
부자(附子)	4	0.1308	1.9869
황금(黃芩)	4	0.1298	1.9647
백출(白朮)	4	0.1294	1.9881
세신(細辛)	4	0.1291	1.9758
반하(半夏)	4	0.1283	1.9853
건강(乾薑)	4	0.1282	1.9809

본초	커뮤니티 수	모듈성(Q)	엔트로피(H)
황기(黃芪)	4	0.1273	1.9747

Table 8.7 에고 네트워크 기반 다의성 상위 본초 (상위 20개) | ‹G-score 네트워크›

본초	커뮤니티 수	모듈성(Q)	엔트로피(H)
황련(黃連)	7	0.5249	2.1833
황백(黃柏)	6	0.4804	1.9887
몰약(沒藥)	6	0.4798	1.8134
오두(烏頭)	4	0.4696	1.9489
서우각(犀牛角)	4	0.4592	1.7184
하수오(何首烏)	5	0.4571	1.8078
침향(沉香)	5	0.4513	1.84
모과(木瓜)	3	0.445	1.5096
천마(天麻)	4	0.4436	1.4124
전갈(全蠍)	4	0.4394	1.6119
소회향(小茴香)	3	0.4385	1.5614
연호색(延胡索)	5	0.4377	1.907
영양각(羚羊角)	6	0.4357	2.4155
선태(蟬蛻)	4	0.4336	1.912
오약(烏藥)	5	0.4323	2.1712
대황(大黃)	6	0.4316	2.0887
육계(肉桂)	5	0.431	2.2937
백강잠(白僵蠶)	5	0.4285	1.7568
계지(桂枝)	5	0.4277	2.31
창출(蒼朮)	6	0.4275	1.8664

다의성 지표가 두드러진 대표적 본초인 황련(黃連)의 배합 문맥을 구체적으로 살펴보자. 황련(黃連)의 에고 네트워크에서 커뮤니티를 탐지한 결과, ‹단순 빈도 네트워크›에서는 4개 커뮤니티($Q = 0.1350$, $H = 1.9612$)로 나뉘었고, ‹G-score 네트워크›에서는 7개 커뮤니티($Q = 0.5249$, $H = 2.1833$)로 구분되었다. ‹단순 빈도 네트워크›에 비해 ‹G-score 네트워크›에서 다의성 지표가 한층 높게 나타났다. 이는 배합 문맥의 경계가 훨씬 선명하게 구분된다는 뜻이다. 세부 커뮤니티 구성은 다음과 같다.

‹단순 빈도 네트워크›로 분석한 황련(黃連)의 문맥

문맥 A: 103종

감초(甘草), 복령(茯苓), 작약(芍藥), 황금(黃芩), 방풍(防風), 대황(大黃), 생지황(生地黃), 맥문동(麥門冬), 시호(柴胡), 길경(桔梗), 행인(杏仁), 승마(升麻), 강활(羌活), 치자(梔子), 마황(麻黃), 석고(石膏), 형개(荊芥), 생강(生薑), 목통(木通), 지모(知母), 상백피(桑白皮), 서우각(犀牛角), 갈근(葛根), 현삼(玄蔘), 전호(前胡), 지골피(地骨皮), 영양각(羚羊角), 박하(薄荷), 연교(連翹), 감국(甘菊), 천화분(天花粉), 패모(貝母), 진교(秦艽), 차전자(車前子), 방기(防己), 자완(紫菀), 백질려(白蒺藜), 고삼(苦蔘), 우방자(牛蒡子), 의이인(薏苡仁), 정력자(葶藶子), 용담초(龍膽草), 만형자(蔓荊子), 두시(豆豉), 금은화(金銀花), 관동화(款冬花), 선복화(旋覆花), 사간(射干), 밀(蜜), 구맥(瞿麥), 화마인(火麻仁), 죽여(竹茹), 위령선(威靈仙), 자소자(紫蘇子), 백합(百合), 과루(瓜蔞), 용치(龍齒), 죽엽(竹葉), 사삼(沙蔘), 위유(葳蕤), 인진(茵陳), 대청(大靑), 구등(鉤藤), 목적(木賊), 백선피(白鮮皮), 결명자(決明子), 청상자(靑葙子), 나미(糯米), 유핵(蕤核), 과루인(瓜蔞仁), 괴각(槐角), 백모근(白茅根), 석결명(石決明), 노근(蘆根), 담죽엽(淡竹葉), 누로(漏蘆), 통초(通草), 괴화(槐花), 갱미(粳米), 충울자(茺蔚子), 호마(胡麻), 자초(紫草), 등심초(燈心草), 토과(土瓜), 진피(秦皮), 조각자(皂角刺), 천초근(茜草根), 지부자(地膚子), 차전초(車前草), 천패모(川貝母), 하엽(荷葉), 유백피(楡白皮), 곡정초(谷精草), 해백(薤白), 하고초(夏枯草), 산두근(山豆根), 밀몽화(密蒙花), 토복령(土茯苓), 백두옹(白頭翁), 이(梨), 은시호(銀柴胡), 황약자(黃藥子), 사함(蛇含)

문맥 B: 92종

당귀(當歸), 세신(細辛), 사향(麝香), 주사(硃砂), 백지(白芷), 망초(芒硝), 황백(黃柏), 유향(乳香), 백반(白礬), 웅황(雄黃), 오두(烏頭), 천남성(天南星), 빙편(冰片), 파두(巴豆), 전갈(全蠍), 천마(天麻), 몰약(沒藥), 경분(輕粉), 활석(滑石), 백강잠(白僵蠶), 우황(牛黃), 조각(皂角), 백부자(白附子), 청대(靑黛), 유황(硫黃), 선태(蟬蛻), 연단(鉛丹), 초오(草烏), 노사(磠砂), 오령지(五靈脂), 호박(琥珀), 호황련(胡黃連), 붕사(硼砂), 한수석(寒水石), 백렴(白斂), 울금(鬱金), 혈갈(血竭), 적소두(赤小豆), 노회(蘆薈), 섬서(蟾蜍), 천축황(天竺黃),

지룡(地龍), 자연동(自然銅), 호분(胡粉), 오초사(烏梢蛇), 백급(白笈), 청염(靑鹽), 목별자(木鱉子), 송향(松香), 금박(金箔), 오배자(五倍子), 철(鐵), 진주(珍珠), 밀타승(密陀僧), 사태(蛇蜕), 망초(莽草), 혈여탄(血餘炭), 오공(蜈蚣), 웅담(熊膽), 대자석(代赭石), 담남성(膽南星), 지마유(芝麻油), 노봉방(露蜂房), 여로(藜蘆), 녹두(綠豆), 비마자(蓖麻子), 아다(兒茶), 백초상(百草霜), 은박(銀箔), 연상(鉛霜), 저지(豬脂), 과체(瓜蒂), 야명사(夜明砂), 노감석(爐甘石), 백정향(白丁香), 유지(柳枝), 자황(雌黃), 인중백(人中白), 녹반(綠礬), 해표소(海螵蛸), 괴지(槐枝), 비석(砒石), 자위피(刺蝟皮), 현정석(玄精石), 와우(蝸牛), 대풍자(大風子), 괴백피(槐白皮), 석연(石燕), 부용엽(芙蓉葉), 여(茹), 저담(豬膽), 염사담(蚺蛇膽)

문맥 C: 68종

백출(白朮), 진피(陳皮), 목향(木香), 반하(半夏), 건강(乾薑), 지각(枳殼), 계지(桂枝), 빈랑(檳榔), 후박(厚朴), 청피(靑皮), 창출(蒼朮), 가자(訶子), 정향(丁香), 향부(香附), 오수유(吳茱萸), 도인(桃仁), 사인(砂仁), 침향(沉香), 지실(枳實), 신곡(神麴), 별갑(鱉甲), 육두구(肉豆蔻), 삼릉(三稜), 봉아출(蓬莪朮), 자소엽(紫蘇葉), 곽향(藿香), 견우자(牽牛子), 오매(烏梅), 소회향(小茴香), 맥아(麥芽), 고량강(高良薑), 연호색(延胡索), 홍화(紅花), 대복피(大腹皮), 오약(烏藥), 초두구(草豆蔻), 모과(木瓜), 저령(豬苓), 울리인(鬱李仁), 산사(山楂), 귤홍(橘紅), 백두구(白豆蔻), 건칠(乾漆), 익지인(益智仁), 천련자(川楝子), 무이(蕪荑), 감수(甘遂), 사군자(使君子), 흑두(黑豆), 내복자(萊菔子), 비파엽(枇杷葉), 아위(阿魏), 앵속각(罌粟殼), 백개자(白芥子), 강황(薑黃), 백편두(白扁豆), 소목(蘇木), 청목향(靑木香), 포강(炮薑), 곤포(昆布), 뇌환(雷丸), 향유(香薷), 관중(貫衆), 석류피(石榴皮), 학슬(鶴蝨), 춘피(椿皮), 몰식자(沒食子), 고련(苦楝)

문맥 D: 59종

인삼(人蔘), 천궁(川芎), 육계(肉桂), 부자(附子), 황기(黃芪), 숙지황(熟地黃), 오미자(五

味子), 우슬(牛膝), 독활(獨活), 목단피(牡丹皮), 택사(澤瀉), 천초(川椒), 원지(遠志), 아교(阿膠), 산약(山藥), 용골(龍骨), 산수유(山茱萸), 육종용(肉蓯蓉), 석창포(石菖蒲), 두중(杜仲), 애엽(艾葉), 대조(大棗), 석곡(石斛), 모려(牡蠣), 토사자(菟絲子), 속단(續斷), 적석지(赤石脂), 천동(天冬), 산조인(酸棗仁), 파극천(巴戟天), 백자인(柏子仁), 지유(地楡), 포황(蒲黃), 단삼(丹蔘), 녹용(鹿茸), 구기자(枸杞子), 비해(萆薢), 고본(藁本), 천웅(天雄), 오가피(五加皮), 사상자(蛇床子), 하수오(何首烏), 연자(蓮子), 종유석(鍾乳石), 측백엽(側柏葉), 상표소(桑螵蛸), 택란(澤蘭), 백미(白薇), 우여량(禹餘糧), 복룡간(伏龍肝), 석위(石韋), 오적골(烏賊骨), 조핵(棗核), 구판(龜板), 백석지(白石脂), 저실자(楮實子), 계내금(鷄內金), 당삼(黨蔘), 상이(桑耳)

‹G-score 네트워크›로 분석한 황련(黃連)의 문맥

문맥 a: 21종

형개(荊芥), 박하(薄荷), 연교(連翹), 감국(甘菊), 차전자(車前子), 백질려(白蒺藜), 고삼(苦蔘), 우방자(牛蒡子), 선태(蟬蛻), 만형자(蔓荊子), 목적(木賊), 결명자(決明子), 청상자(靑葙子), 유핵(莥核), 괴각(槐角), 석결명(石決明), 진피(秦皮), 지부자(地膚子), 곡정초(谷精草), 밀몽화(密蒙花), 만청(蔓菁)

문맥 b: 20종

황금(黃芩), 대황(大黃), 시호(柴胡), 승마(升麻), 치자(梔子), 망초(芒硝), 황백(黃柏), 지모(知母), 서우각(犀牛角), 현삼(玄蔘), 지골피(地骨皮), 천화분(天花粉), 용담초(龍膽草), 울금(鬱金), 과루(瓜蔞), 위유(葳蕤), 대청(大靑), 담죽엽(淡竹葉), 토과(土瓜), 계내금(鷄內金)

문맥 c: 16종

목향(木香), 지각(枳殼), 후박(厚朴), 창출(蒼朮), 오수유(吳茱萸), 지실(枳實), 신곡(神麴), 육두구(肉豆蔻), 석창포(石菖蒲), 맥아(麥芽), 산사(山楂), 귤홍(橘紅), 백편두(白扁豆), 과루인(瓜蔞仁), 담남성(膽南星), 향유(香薷)

문맥 d: 16종

당귀(當歸), 작약(芍藥), 생지황(生地黃), 아교(阿膠), 용골(龍骨), 애엽(艾葉), 오매(烏梅), 적석지(赤石脂), 지유(地楡), 괴화(槐花), 백석지(白石脂), 석류피(石榴皮), 천초근(茜草根), 춘피(椿皮), 백두옹(白頭翁), 백백피(柏白皮)

문맥 e: 15종

백반(白礬), 빙편(冰片), 경분(輕粉), 연단(鉛丹), 붕사(硼砂), 백렴(白蘞), 호분(胡粉), 백급(白笈), 오배자(五倍子), 노감석(爐甘石), 백정향(白丁香), 해표소(海螵蛸), 여(茹), 도피(桃皮), 백사밀(白沙蜜)

문맥 f: 13종

청대(靑黛), 호황련(胡黃連), 노회(蘆薈), 무이(蕪荑), 섬서(蟾蜍), 사군자(使君子), 웅담(熊膽), 야명사(夜明砂), 몰식자(沒食子), 고련(苦楝), 은시호(銀柴胡), 저담(豬膽), 염사담(蚺蛇膽)

문맥 g: 5종

목통(木通), 택사(澤瀉), 활석(滑石), 저령(豬苓), 등심초(燈心草)

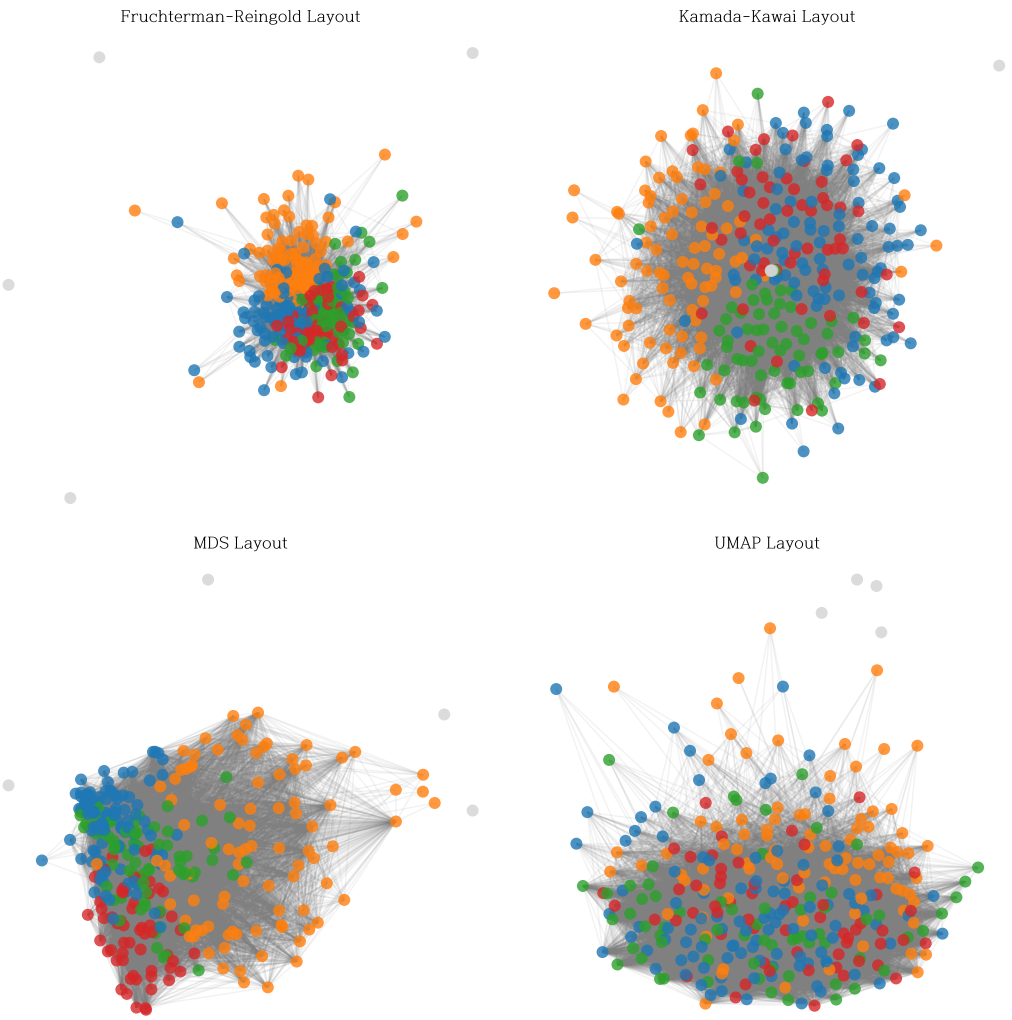

Figure 8.5 에고 네트워크와 맥락 분포 | ‹단순 빈도 네트워크›

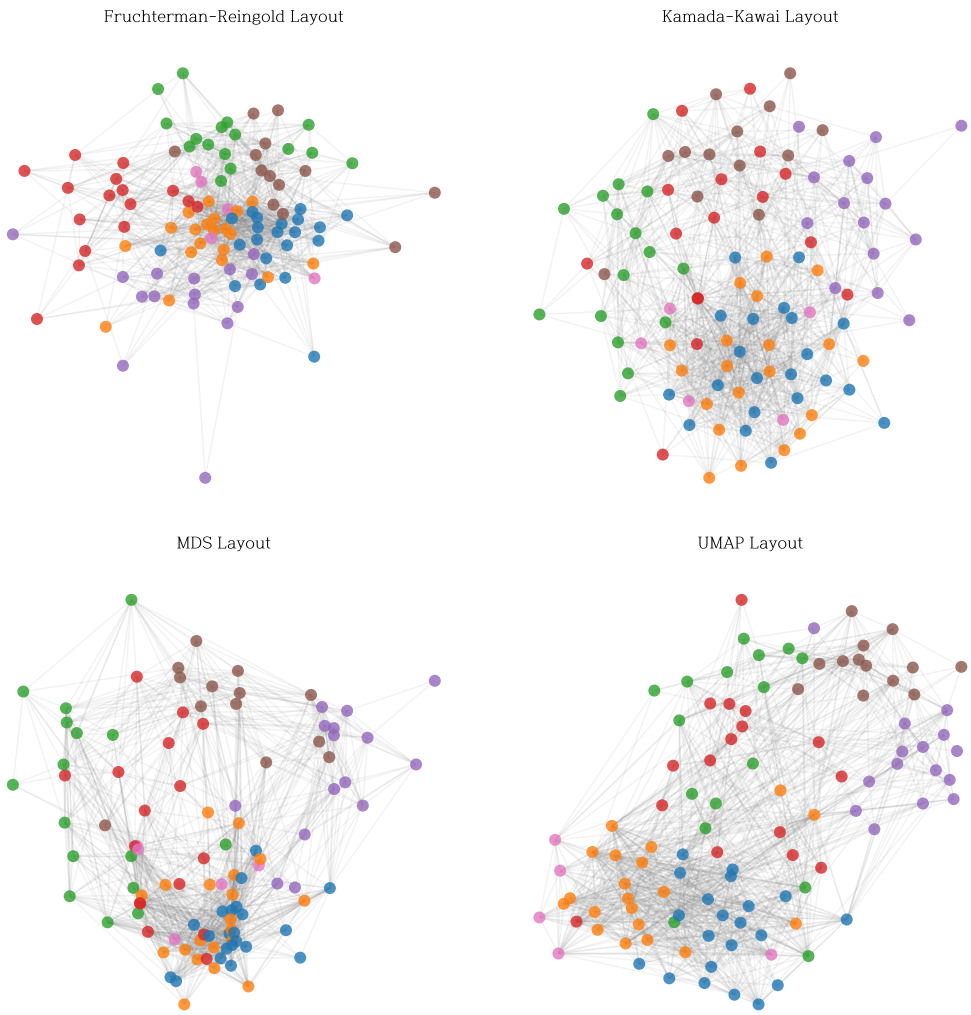

Figure 8.6 에고 네트워크와 맥락 분포 | ‹G-score 네트워크›

커뮤니티 탐지와 에고 네트워크 분석은 처방 배합의 구성 원리를 객관적으로 보여준다. 본초는 무작위로 배합되지 않고 특정 병증과 치료 목적에 따라 커뮤니티를 이룬다. 네트워크 전체에 나타난 이러한 구조는 한의학 처방이 목표 효능을 중심으로 체계적인 기능 모듈을 이루고 있음을 뒷받침한다.

아울러 주요 본초는 단일한 효능에 머물지 않는다. 에고 네트워크 모듈성과 섀넌 엔트로피에서 드러나듯 여러 커뮤니티를 넘나들며 함께 쓰이는 본초 조합에 따라 다양한 치료 맥락을 만들어낸다. 같은 본초라도 결합하는 이웃에 따라 발휘하는 주요 효능이 달라지는 것이다. 네트워크 분석은 이처럼 처방의 기능적 구조와 본초의 다의성을 데이터를 바탕으로 살펴볼 수 있게 한다.

제 9 장

서브네트워크

지금까지 처방 데이터를 바탕으로 구축한 본초 네트워크를 거시적으로 검토하였다. 전체 네트워크를 조망하는 거시적 분석은 데이터의 전반적인 규칙성과 질서를 파악하는 데 효과적이다. 그러나 우리에게 익숙한 대상은 방대한 전체 네트워크가 아니라 각각의 개별 처방이다. 이 장에서는 전체 네트워크의 일부인 서브네트워크를 소개하고 한의학 처방이 전체 본초 네트워크 속에서 어떠한 국소 위상을 지니는지 살펴본다.

9.1 서브네트워크와 에고 네트워크

방대한 복잡계를 전반적으로 살펴보기 위해 전체 네트워크를 분석했다면, 이를 더 작은 단위로 나누어 집중적으로 분석할 수도 있다. 전체 네트워크에서 일정 부분을 추출한 부분집합을 **서브네트워크(Subnetwork)**라고 부른다. 전체 그래프($G = (V, E)$)에서 임의의 노드 부분집합($V' \subseteq V$)과 임의의 엣지 부분집합($E' \subseteq E$)을 취하여 구성한 모든 그래프를 가리킨다.

하지만 서브네트워크라고 하면 보통 '유도 서브네트워크'를 가리킨다. **유도 서브네트워크(Induced Subnetwork, $G[V']$)**는 노드 사이의 상호작용을 온전히 보존하는 서브네트워크다. 노드

부분집합($V' \subseteq V$)을 결정했을 때 원래 네트워크(G)에서 해당 노드 부분집합에 속하는 모든 엣지를 빠짐없이 유지하여 구성한다. 유도 서브네트워크는 전체 네트워크의 연결성을 온전히 보존하므로 선택된 집단 내부의 국소 밀도(Local Density)와 결속 강도를 왜곡하지 않는다. 유도 서브네트워크는 아래와 같이 수학적으로 정의한다.

$$G[V'] = (V', E'), \quad E' = \{(u, v) \in E \mid u, v \in V'\}$$

($G[V']$: 노드 부분집합 V'에 의해 유도된 서브네트워크, V': 선택된 대상 본초 노드 집합, E': 양 끝 노드가 모두 V'에 속하는 원래 네트워크의 모든 엣지 집합)

서브네트워크 가운데 특수한 형태로 '에고 네트워크'가 있다. **에고 네트워크(Ego Network, G_v)**는 단일 중심 노드(Ego, v)와 중심 노드에 직접 1차로 연결된 이웃 노드 집합(Alters, $N(v)$)만을 취하여 만든다.

$$G_v = G[\{v\} \cup N(v)]$$

i 홉(Hop)과 에고 네트워크

네트워크 분석에서 **홉(Hop)**은 노드와 노드 사이를 이동할 때 거치는 엣지의 단계를 뜻하는 거리 척도이다. 1-홉(1-hop)은 중심 노드와 다른 노드 사이에 중간 매개 노드 없이 단 하나의 엣지로 직접 이어지는 거리를 의미한다.
홉 관점에서 **에고 네트워크**는 중심 노드로부터 특정 홉 거리 이내에 도달할 수 있는 노드와 노드 간 유도 엣지로 구성된 국소 네트워크를 뜻한다. 앞서 본문에서 수식으로 정의한 에고 네트워크(G_v)는 중심 노드로부터 1-홉 거리 안에 있는 이웃(1차 이웃)만을 포함한 '1-홉 에고 네트워크(1-hop Ego Network)'에 해당한다. 일반적으로 에고 네트워크는 이 '1-홉 에고 네트워크'를 의미한다.
그러나 에고 네트워크의 범위를 2-홉(이웃 본초의 이웃) 또는 그 이상의 다중 홉(k-홉)으로 확장할 수도 있다. 다만 홉 수가 늘어날수록 중심 노드와 이웃 노드의 연결 특성이 희석되고 전체 네트워크의 거시 구조로 점차 회귀한다. 따라서 개별 노드를 중심으로 국소 특성을 살필 때는 통상 1-홉 에고

> 네트워크를 기본 분석 단위로 사용한다.

한의학 처방 데이터를 바탕으로 만든 전체 본초 네트워크는 본초 배합 관계를 전역(Global) 관점에서 보여준다. 이때 개별 처방은 전체 본초 네트워크의 서브네트워크에 해당한다. 따라서 특정 처방 F를 구성하는 본초 집합을 $V_F = \{h_1, h_2, \dots, h_k\}$라 할 때 처방 F는 전체 본초 네트워크(G)에서 유도된 서브네트워크 $G[V_F]$로 표현할 수 있다.

특정 본초 h를 중심으로 국소 배합 관계를 살피려면 해당 본초의 에고 네트워크를 구축할 수 있다. 본초 h에 직접 1-홉으로 연결된 이웃 본초 집합을 $N(h)$라 할 때 본초 h의 에고 네트워크 G_h는 $G_h = G[\{h\} \cup N(h)]$로 나타낼 수 있다. 이는 본초 h와 함께 배합되는 모든 동시 출현 본초뿐만 아니라 이웃 본초 사이의 연결 관계까지 온전히 보존한다. 따라서 단일 본초의 관점에서 주변 배합 구조의 결속력과 이웃 본초와의 상호작용을 파악하는 데 효과적이다.

9.2 서브네트워크로서의 처방

전체 본초 네트워크에서 서브네트워크로서 처방 하나를 추출하여 분석해 보자. 앞선 분석과 마찬가지로 ‹단순 빈도 네트워크›와 ‹G-score 네트워크› 두 가지 네트워크를 전체 본초 네트워크로 사용하였다. 분석 대상 처방으로는 쌍화탕(雙和湯)을 선택하였다. 이 처방은 사물탕(四物湯)과 황기건중탕(黃芪建中湯)을 합방으로 허로(虛勞)로 심신(心腎)이 모두 허할 때 기(氣)와 혈(血)을 두루 보하고 음양(陰陽)을 조화롭게 하는 처방이다.

먼저 처방 데이터와 전역 네트워크를 바탕으로 대상 처방을 구성하는 각 본초의 기본 통계값을 살펴본다. 대상 처방인 쌍화탕은 총 9종의 본초(감초, 당귀, 작약, 천궁, 황기, 계지, 숙지황, 생강, 대조)로 구성되어 있다. 전체 네트워크에서 구성 본초는 모두 핵심 본초에 해당했다. 제5장에서

수행한 k-core 분해 결과로 보면 ‹단순 빈도 네트워크›에서 모든 본초가 최대 k-core($k = 91$)에 속하였다. ‹G-score 네트워크›에서는 대다수 본초가 최대 k-core($k = 37$)에 속하였고, 나머지도 높은 코어니스 계층($k = 32$)에 속하였다. 이는 대상 처방을 구성하는 본초가 주변부 본초가 아니라 한의학 처방을 지탱하는 핵심 본초라는 사실을 정량적으로 보여준다.

Table 9.1 대상 처방 구성 본초 기본 통계 | ‹단순 빈도 네트워크›

본초명	출현 빈도	빈도 순위	연결 차수	가중 차수	코어니스 / 최대 k-core
감초(甘草)	16,581	1위	464	2892.3	91 / 91
당귀(當歸)	9,922	3위	440	2653.7	91 / 91
작약(芍藥)	7,869	5위	394	2278.1	91 / 91
천궁(川芎)	6,269	7위	378	2188.6	91 / 91
황기(黃芪)	3,729	20위	305	1662.5	91 / 91
계지(桂枝)	3,245	22위	277	1511.2	91 / 91
숙지황(熟地黃)	2,908	26위	249	1365.5	91 / 91
생강(生薑)	1,783	43위	193	963.7	91 / 91
대조(大棗)	841	108위	140	635.2	91 / 91

Table 9.2 대상 처방 구성 본초 기본 통계 | ‹G-score 네트워크›

본초명	출현 빈도	빈도 순위	연결 차수	가중 차수	코어니스 / 최대 k-core
감초(甘草)	16,581	1위	123	18307.8	37 / 37
당귀(當歸)	9,922	3위	152	24917.8	37 / 37
작약(芍藥)	7,869	5위	124	18646.3	37 / 37
천궁(川芎)	6,269	7위	136	23777.3	37 / 37
황기(黃芪)	3,729	20위	117	15111.8	37 / 37
계지(桂枝)	3,245	22위	106	7664.3	37 / 37
숙지황(熟地黃)	2,908	26위	111	20582.6	37 / 37
생강(生薑)	1,783	43위	58	3905.8	32 / 37
대조(大棗)	841	108위	56	2766.7	32 / 37

대상 처방의 구성 본초를 바탕으로 두 네트워크에서 서브네트워크($G[V_{\text{쌍화탕}}]$)를 추출하고 위상 지표를 측정하였다.

Table 9.3 대상 처방 서브네트워크 위상 구조 통계

위상 지표	‹단순 빈도 네트워크›	‹G-score 네트워크›
고유 본초 수 (노드 수, $\|V'\|$)	9 종	9 종
이론적 최대 연결 수 (Potential Connections)	36 개	36 개
실제 관찰 연결 수 (Actual Connections, $\|E'\|$)	36 개	29 개
네트워크 밀도 (Density)	1.0000	0.8056
평균 연결 차수 (Average Degree)	8.00	6.44
평균 군집 계수 (Average Clustering Coefficient)	1.0000	0.8693
평균 최단 경로 길이 (Average Path Length)	1.00	1.19

위상 통계에서 드러나듯 ‹단순 빈도 네트워크›의 서브네트워크는 모든 노드 쌍 사이에 빠짐없이 연결선이 존재하는 **완전 그래프(Complete Graph)**를 형성하여 이론적 최대 밀도인 1.0에 도달하였다. 즉 9종의 구성 본초가 둘씩 짝지어질 수 있는 모든 연결(36개)이 처방 내부에서 전부 이어진 구조를 띤다. 한편 ‹G-score 네트워크› 서브네트워크에서는 엣지 수가 29개로 줄어들고 밀도는 0.806로 감소하였다. ‹G-score 네트워크›에서 연결되지 않고 탈락한 7개의 연결선은 감초-숙지황, 당귀-생강, 대조-숙지황, 대조-천궁, 생강-숙지황, 생강-천궁, 생강-황기 조합이다. 이는 생강·대조 등 조화 약재와 일부 구성 본초의 관계가 통계적 유의성 기준을 통과하지 못해 걸러진 결과이다.

9.3 전체 네트워크와 서브네트워크의 비교

전체 네트워크에서 도출한 전역 특성은 특정 본초가 한의학 처방 전반에서 보이는 보편적인 모습으로 '기댓값'의 성격을 띤다. 반면 개별 처방 서브네트워크에서 측정한 국소 특성은 해당 처방에서 나타나는 고유한 모습으로 '관측값'의 성격을 지닌다. 따라서 전역 특성과 국소 특성을 교차 비교하면 보편적 기댓값 대비 처방 내부 관측값이 어떠한 편차를 보이는지 추적할 수 있다. 특정 본초의 국소 위상이 전역 위상보다 높게 나타난다면 대상 처방 내부에서 해당 본초를 중심

으로 한 국소적 결속 구조가 더 조밀하게 형성되었음을 의미한다. 반대로 전역 위상에 비해 국소 위상이 낮게 나타난다면 전역적으로 광범위한 처방군에 걸쳐 다빈도로 쓰이는 본초라도 해당 처방 내부에서는 특정 핵심 약대 조합에 비해 상대적 결속 가중치가 낮게 분배되었음을 보여준다.

전역 네트워크와 서브네트워크에서 구성 본초의 중심성과 커뮤니티를 비교해 보자.

9.3.1 중심성 비교

먼저 구성 본초의 전역 중심성과 국소 중심성을 비교한다. 전역 중심성이 해당 본초가 처방 전체에서 얼마나 광범위하게 영향을 미치는지 보여준다면 국소 중심성은 각 본초가 해당 처방을 지탱하고 유지하는 데 얼마나 기여하는지를 나타낸다. 중심성 비교에는 페이지랭크(PageRank) 중심성을 활용하였다.[1]

서로 다른 두 네트워크 사이에서 페이지랭크 중심성을 비교할 때는 척도의 특성에 주의해야 한다. 페이지랭크는 네트워크 내부의 모든 노드 확률값을 합산하면 항상 1이 되는 확률분포 특성을 지닌다. 따라서 네트워크를 이루는 노드 수에 따라 수치 규모가 달라진다. 즉 수천 종의 본초로 구성된 전역 네트워크의 기본 기댓값($1/|V|$)과 몇 종의 본초로 한정된 서브네트워크의 기댓값($1/|V'|$)은 절대적인 수치 크기가 매우 다르다. 그러므로 두 네트워크에서 중심성 수치를 직접 일대일로 비교하기보다는 각 네트워크 내부에서 본초가 차지하는 상대적인 순위 변동을 중심으로 해석하는 편이 타당하다.

[1] 중심성에 대해서는 제6장을 참고하자.

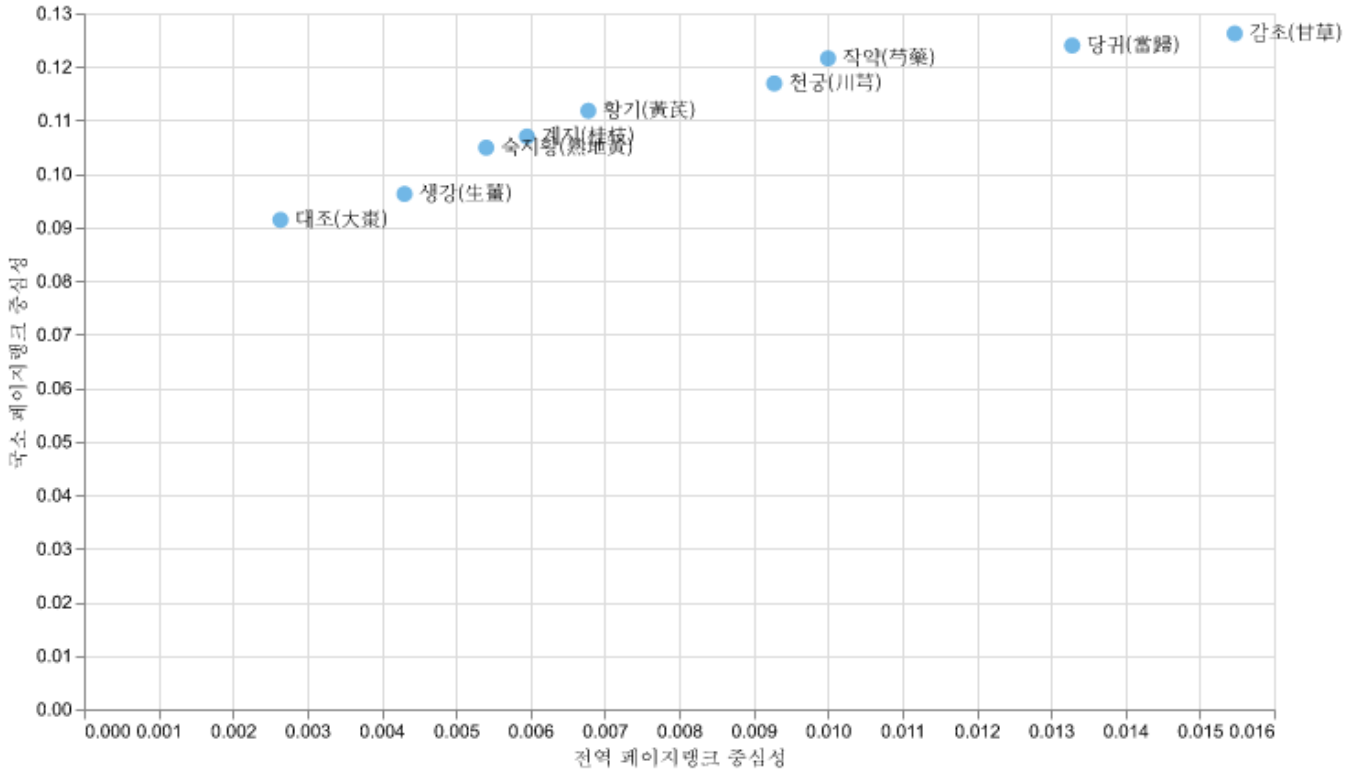

Figure 9.1 전역 vs 국소 페이지랭크 중심성 비교 | ‹단순 빈도 네트워크›

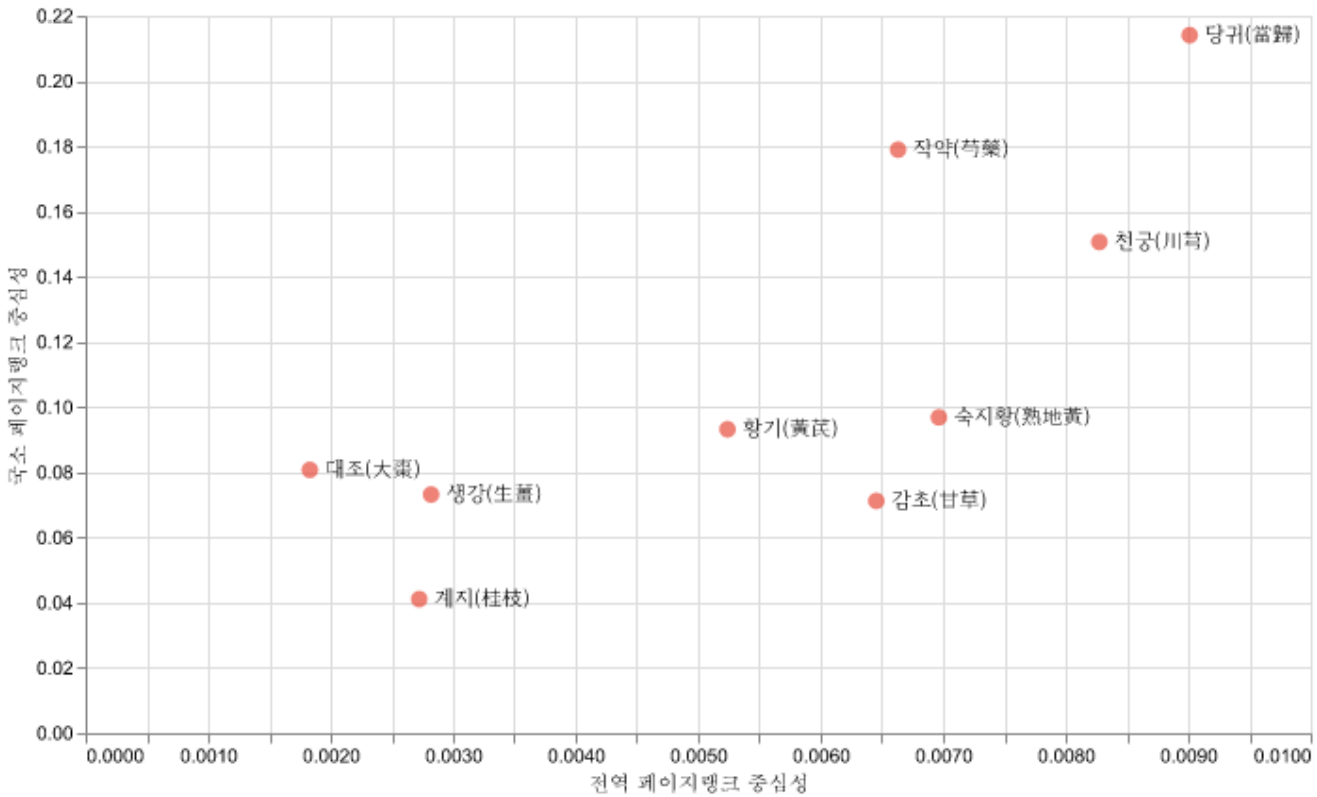

Figure 9.2 전역 vs 국소 페이지랭크 중심성 비교 | ‹G-score 네트워크›

산점도를 통해 전역 중심성과 국소 중심성 양상을 비교해 보았다. ‹단순 빈도 네트워크›에서는 서브네트워크의 모든 노드가 연결되어 완전 그래프를 이룬다. 무방향 네트워크에서 페이지랭크는 랜덤 워크의 정상분포가 노드의 연결 강도(가중 차수)에 비례하는 수학적 성질을 지닌다. 이에 따라 국소 페이지랭크는 균등 기준선인 약 0.111 주변(0.091 ~ 0.126)에 밀집하면서도 처방 내부 동시 출현 가중치 합이 큰 본초를 중심으로 완만한 선형성을 유지하였다.

반면 ‹G-score 네트워크›에서는 중심성 재배치가 일어났다. 구체적으로 대조(大棗)·작약(芍藥)·황기(黃芪)의 순위가 상승한 반면 감초(甘草)·천궁(川芎)·숙지황(熟地黃)·계지(桂枝)의 순위는 하락하였다. 상세 변동을 보면 대조(大棗)는 전역 중심성 9위에서 국소 중심성 6위로, 작약(芍藥)은 전역 중심성 4위에서 국소 중심성 2위로, 황기(黃芪)는 전역 중심성 6위에서 국소 중심성 5위로 순위가 상승하였다. 그러나 감초(甘草)는 전역 중심성 5위에서 국소 중심성 8위로, 천궁(川芎)은 전역 중심성 2위에서 국소 중심성 3위로, 숙지황(熟地黃)은 전역 중심성 3위에서 국소 중심성 4위로, 계지(桂枝)는 전역 중심성 8위에서 국소 중심성 9위로 각각 순위가 하락하였다.

전역 네트워크의 중심성을 기댓값, 서브네트워크의 중심성을 관측값이라고 할 때 이러한 순위 변동을 바탕으로 처방 내부에서 본초 간 결속 강도가 어떻게 재편되는지 추적할 수 있다. 실제로 쌍화탕에서 국소 중심성 순위가 상승한 본초는 처방의 주된 효능을 이끈다. 반대로 순위가 하락한 본초는 그 고유 효능이 처방의 목표 효능과 다소 차이 나거나 이웃 본초를 돕는 역할을 맡는다.

9.3.2 커뮤니티 비교

전체 네트워크 분석에서 확인한 커뮤니티가 개별 처방 내부에서도 일관되게 유지되는지, 아니면 처방의 고유한 배합 목적에 맞추어 새롭게 재편되는지 살펴본다. 이를 위해 제8장에서 적용한 가중 라이덴(Leiden) 커뮤니티 탐지 알고리즘을 쌍화탕 서브네트워크에 동일하게 적용하였다.

Table 9.4 전역 vs 국소 커뮤니티 교차 분류 | ‹단순 빈도 네트워크›

국소 커뮤니티	전역 커뮤니티 B	전역 커뮤니티 C	전역 커뮤니티 D	전역 커뮤니티 E
국소 커뮤니티 1	작약(芍藥), 감초(甘草), 생강(生薑), 대조(大棗)	숙지황(熟地黃), 황기(黃芪), 당귀(當歸)	계지(桂枝)	천궁(川芎)

Table 9.5 전역 vs 국소 커뮤니티 교차 분류 | ‹G-score 네트워크›

국소 커뮤니티	전역 커뮤니티 b	전역 커뮤니티 c	전역 커뮤니티 d	전역 커뮤니티 e	전역 커뮤니티 f
국소 커뮤니티 1	감초(甘草)	숙지황(熟地黃), 황기(黃芪)	천궁(川芎)	계지(桂枝)	작약(芍藥), 당귀(當歸)
국소 커뮤니티 2	생강(生薑), 대조(大棗)	-	-	-	-

교차 분류 결과에서 두 네트워크의 커뮤니티 차이를 살펴볼 수 있다. ‹단순 빈도 네트워크›의 전역 네트워크 관점에서 9종의 본초는 4개의 서로 다른 전역 커뮤니티에 나누어져 있었다.[2] 하지만 ‹단순 빈도 네트워크› 서브네트워크는 커뮤니티 분화가 일어나지 않고 본초 전체가 하나의 국소 커뮤니티로 묶였다. 모든 노드가 연결된 완전 그래프(밀도 1.0)에서는 모든 연결선이 단일 커뮤니티 내부에 포함된다. 따라서 네트워크를 분할하더라도 구조적 이득이 생기지 않아 모듈성(Q)이 0.000으로 측정되었다.

한편 ‹G-score 네트워크›의 전역 네트워크 관점에서 9종의 본초는 5개의 서로 다른 전역 커뮤니티에 나누어져 있었다. ‹G-score 네트워크› 서브네트워크에서는 2개의 국소 커뮤니티로 구분되었다. 모듈성(Q)은 0.110으로 나타났다.

전역 커뮤니티와 국소 커뮤니티는 처방의 배합 구조를 다각도로 해석하는 데 상호 보완적인 통찰을 제공한다. 전역 커뮤니티는 전체 처방이라는 거시적 지형에서 각 본초의 효능 배경을 보여준다. 구성 본초가 한의학 처방 전체에서 어떻게 활용되었는지 조망하게 돕는다. 반면 국소 커뮤니티는 해당 처방 내부에서 본초들이 실제로 맺고 있는 직접적인 역할을 미시적으로 보여준다. 두 관점을 교차하여 살펴보면 서로 다른 전역 기능을 담당하던 본초들이 단일 처방 안에서 어떻게 협력하는지 더욱 입체적으로 파악할 수 있다.

9.4 처방 시각화 모형

시각화는 대상을 직관적으로 파악하는 데 큰 도움을 준다. 서브네트워크를 바탕으로 단순한 본초 나열을 넘어 본초 사이의 관계망으로 대상 처방을 표현할 수 있다.

[2] 전역 커뮤니티의 탐지 알고리즘 및 커뮤니티 구조에 대한 상세한 내용은 제8장의 '커뮤니티'를 참고한다.

9.4.1 엣지 가중치 기반 덴드로그램

본초와 본초 사이의 관계는 엣지 가중치로 표현하였다. 이를 이용하여 계층적 군집화를 수행하면 본초 간 결속 강도를 기준으로 구성 본초를 이진 트리 형태로 나타낼 수 있다(Johnson 1967). **계층적 군집화(Hierarchical Clustering)**는 개별 본초를 독립된 단일 군집에서 출발하여 거리가 가장 가까운 군집 쌍을 단계적으로 병합해 나감으로써 전체 개체를 이진 트리 구조로 묶어내는 상향식 군집화(Agglomerative Clustering) 알고리즘이다. **덴드로그램(Dendrogram)**은 이러한 계층적 병합 과정을 계층 트리 구조로 시각화한 도표이다. 가지가 갈라져 합쳐지는 분기 높이(또는 길이)로 군집 간 결속 거리와 비유사도 크기를 직관적으로 나타낸다. 이 방법으로 대상 처방을 시각화한다.

계층적 군집화는 거리 데이터를 사용하므로 엣지 가중치를 거리로 선형 반전시켰다.[3] 아울러 상향식 병합 과정에서 군집 간 거리를 정의하고 병합 순서를 결정하는 기준을 연결법(Linkage Method)이라고 한다. 본 분석에서는 연결법으로 와드 연결법(Ward's Linkage)을 적용하였다.[4]

덴드로그램은 네트워크 특성에 따라 처방 내부의 결속 구조를 잘 드러낸다. ‹단순 빈도 네트워크› 덴드로그램에서는 전역 빈도가 매우 높은 당귀(當歸)·작약(芍藥)·감초(甘草) 등이 중심을 구성하고 나머지 약재가 순차적으로 결합하였다.

‹G-score 네트워크› 덴드로그램에서는 한의학의 기능적 약대 구조가 더 잘 드러났다. 당귀(當歸)와 작약(芍藥)이 가장 먼저 최단 거리로 결합하고 천궁(川芎)이 합류하였다. 여기에 황기(黃芪)와

[3] 본초와 본초 사이의 결속 거리는 제6장의 '유사도(Similarity)와 거리(Distance)'에서 정립한 선형 반전 공식을 적용하였다. 연결된 본초 쌍에는 가중치 선형 반전 거리를 부여하고 서브네트워크 내에 엣지가 없는 비연결 쌍에는 페널티 거리 계수($\beta = 1.5$)를 적용하여 가장 나중에 결합하도록 구성하였다.

[4] 와드 연결법(Ward's Linkage)은 군집을 병합할 때 군집 내부의 분산(오차 제곱합) 증가량이 최소가 되는 쌍을 순차적으로 결합해 나가는 방식이다. 다른 연결 방식에 비해 크기가 균형 잡히고 결속 밀도가 높은 응집성 군집을 도출하는 장점을 지닌다.

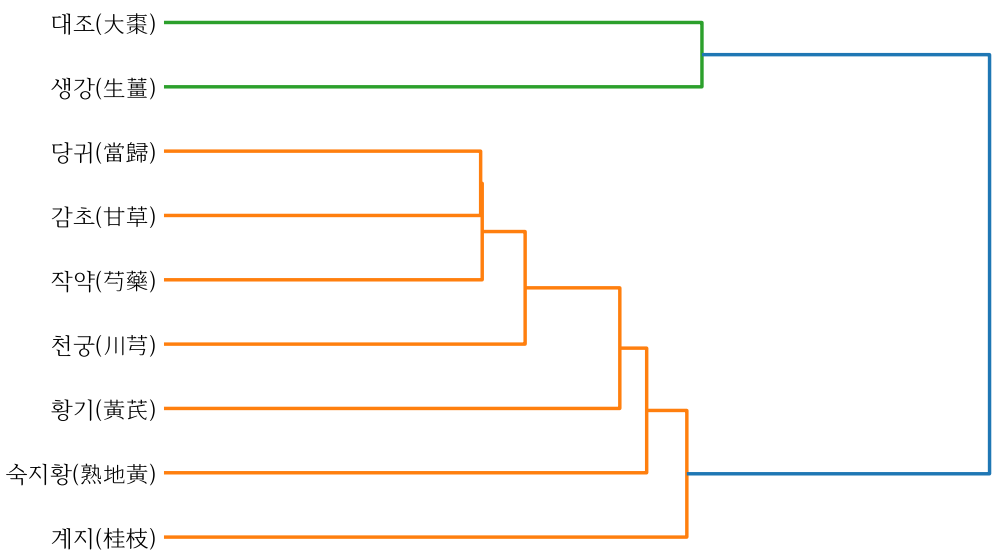

Figure 9.3 대상 처방 엣지 가중치 기반 계층적 군집화 덴드로그램 | ‹단순 빈도 네트워크›

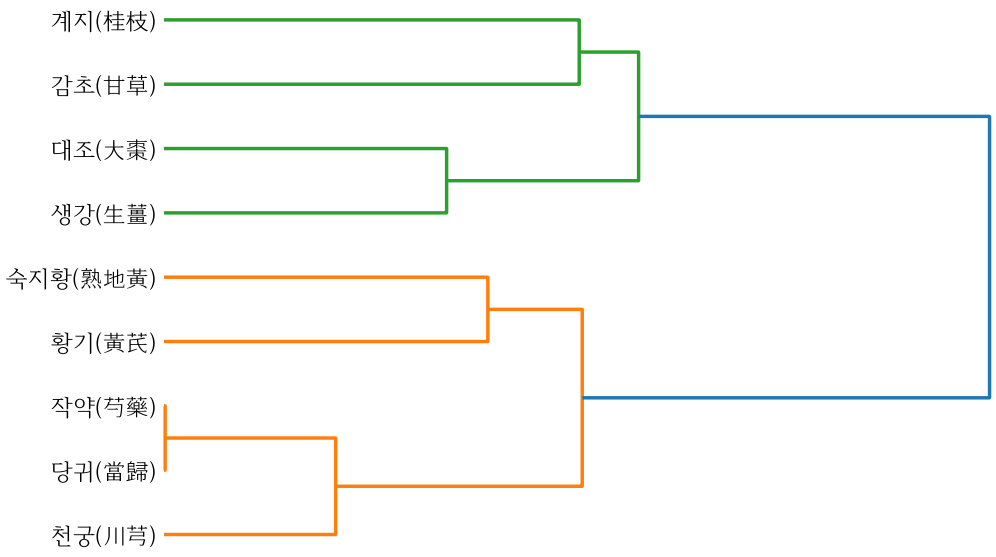

Figure 9.4 대상 처방 엣지 가중치 기반 계층적 군집화 덴드로그램 | ‹G-score 네트워크›

숙지황(熟地黃)이 결합하여 처방의 얼개를 이루었다. 이와 별도로 계지(桂枝)·감초(甘草)와 생강(生薑)·대조(大棗)가 도 다른 군집을 이루고 있었다. 이를 통해 사물탕(四物湯)과 계지탕(桂枝湯)의 약대를 구성을 확인할 수 있었다.

9.4.2 서브네트워크 기반 레이아웃

다음으로 네트워크 레이아웃 알고리즘을 적용하여 서브네트워크 관점에서 대상 처방을 2차원 공간에 시각화한다.[5] 여기서는 여러 레이아웃 중 다차원 척도법(MDS) 레이아웃을 사용하였다. 구성 본초 수가 적은 개별 처방 서브네트워크에서는 본초 간 가중 최단 경로 거리 행렬을 대수적으로 직접 분해하여 2차원 평면에 투영하는 다차원 척도법(MDS)이 적합하기 때문이다.[6]

시각화에서는 노드의 색상에 전체 네트워크의 커뮤니티 소속을 반영하고 엣지의 두께에는 가중치 정규화 값을 반영하여 결속 강도를 시각적으로 식별할 수 있도록 구성하였다.

두 네트워크의 시각화 결과를 비교하면 공간 배치에 차이가 있지만 커뮤니티 구성과 처방의 의미를 모두 관찰할 수 있었다. ‹단순 빈도 네트워크› 서브네트워크 레이아웃에서 중앙을 기준으로 사물탕(四物湯)과 계지탕(桂枝湯) 조합을 엿볼 수 있었다. ‹G-score 네트워크› 서브네트워크 레이아웃에서도 같은 조합이 나타났다. 다만 전자에서 외곽에 위치했던 계지(桂枝)가 후자에서 내부로 이동하고 대신 황기(黃芪)가 외곽으로 이동하는 변화를 보였다.[7]

---

[5] 네트워크 시각화를 위한 레이아웃에 대해서는 제7장을 참고하자.

[6] 힘 기반(Spring, Kamada-Kawai) 알고리즘은 노드의 초기 무작위 좌표와 물리적 시뮬레이션 반복 계산에 의존하므로 실행 시점이나 파라미터에 따라 노드의 회전 방향이나 국소 배치가 달라질 수 있다. 반면 MDS는 확률적 무작위성이나 시뮬레이션 반복에 따른 왜곡 없이 본초 간의 거리 구조를 수학적으로 보존하는 결정론적(Deterministic) 알고리즘이다. 따라서 처방 내부의 결속 강도와 본초 간 상대적 거리를 객관적이고 일관되게 관찰할 수 있다.

[7] 네트워크를 2차원 평면에 투영하는 과정에서 기하학적 왜곡이 뒤따른다. 따라서 노드의 평면 좌표나 중심·외곽 배치는 네트워크 구조를 직관적으로 살피는 보조 수단으로 활용해야 한다. 특정 노드의 공간적 위치를 바탕으로 처방 내부의 위계를 추측하기보다는 중심성 순위나 엣지 가중치 등 정량적 지표와 함께 해석해야 한다.

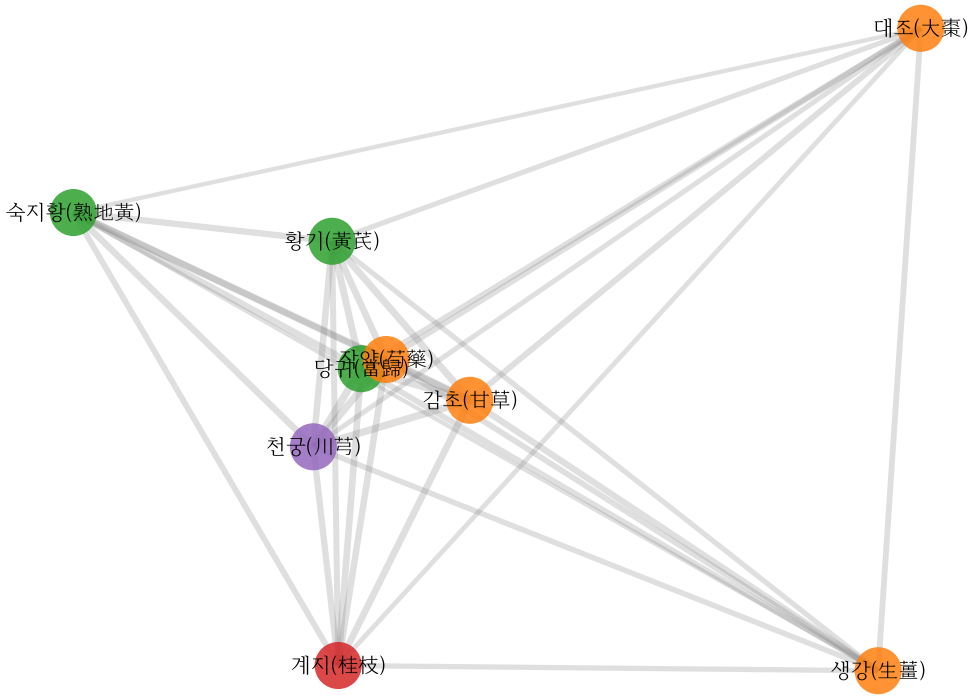

Figure 9.5 대상 처방 서브네트워크 다차원 척도법(MDS) 레이아웃 | ‹단순 빈도 네트워크›

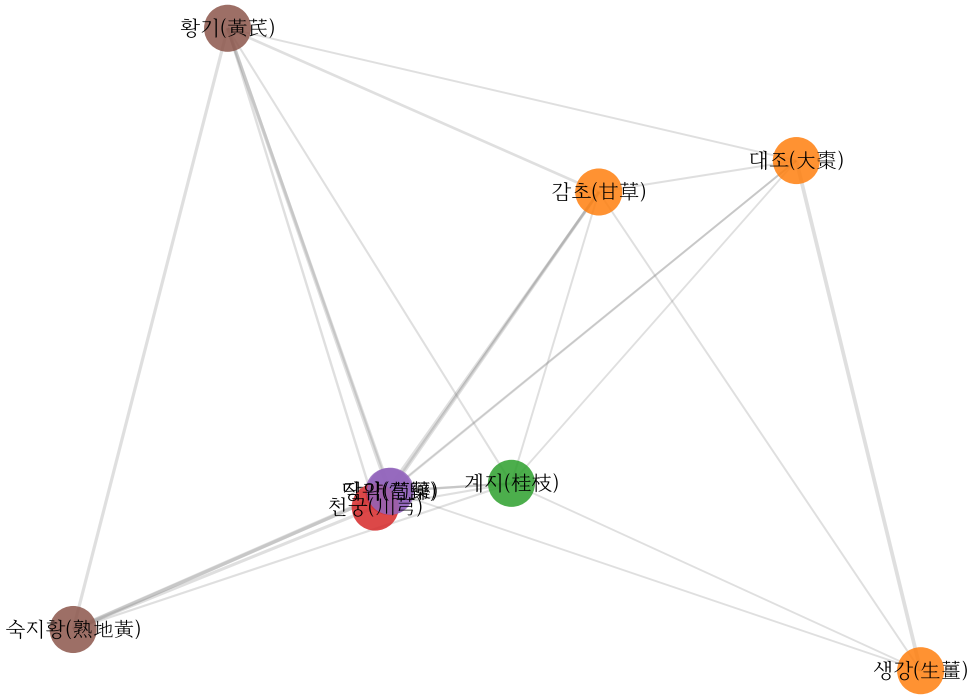

Figure 9.6 대상 처방 서브네트워크 다차원 척도법(MDS) 레이아웃 | ‹G-score 네트워크›

이 장에서는 전체 네트워크의 부분집합인 서브네트워크를 소개하고 우리에게 익숙한 처방이 본초 네트워크의 서브네트워크라는 관점에서 대상 처방을 간단히 분석하였다. 서브네트워크의 특성을 전체 네트워크와 비교하여 특정 처방에서 각 본초의 영향력과 주변 본초와의 관계가 어떻게 변화하는지 살펴보았다. 또 한의학 처방의 본초 구성을 직관적으로 살펴보기 위해 서브네트워크를 시각적으로 표현하는 방법도 다루었다.

맺음말

이 책에서는 네트워크 분석 방법을 한의학 처방이라는 전통 의학 지식에 적용하여 체계적으로 탐색하였다. 먼저 방대한 고전 처방 속 본초 배합 관계를 노드와 엣지로 모델링하고 출현 빈도의 헤비테일(Heavy-tailed) 분포 양상을 확인하였다. 아울러 처방 데이터를 네트워크로 모델링하고자 대표적인 엣지 가중치 산출 방식을 소개하고 빈도 편향을 보완하는 방안도 함께 검토하였다.

한의학 처방 데이터를 바탕으로 본초 네트워크를 구축하여 본초 배합 구조를 다각도로 분석하였다. 무작위 모델과 비교하여 본초 네트워크가 단순한 우연의 산물이 아니며, 좁은 세상 구조와 삼각 모티프 그리고 중심부-주변부 계층 질서를 지니고 있음을 검증하였다. 이어 중심성 지표로 개별 본초의 영향력을 정량화하였다. 아울러 강건성 시뮬레이션으로 핵심 중심 거점(Hub) 본초가 네트워크 전체의 안정성에 미치는 영향을 살펴보았다.

나아가 다차원 공간 배치를 활용해 복잡한 본초 관계를 2차원 평면에 직관적으로 시각화하고 모듈성 최적화 기법으로 기능적 배합 단위인 커뮤니티를 포착하였다. 또한 에고 네트워크와 정보 엔트로피를 응용하여 처방 맥락에 따라 배합 양상이 분화하는 본초의 다의적 특성을 분석하였다. 끝으로 개별 처방을 전체 네트워크의 유도 서브네트워크로 간주하고 전역 지형과 국소 처방 사이의 위상 변화를 비교하였다. 이를 바탕으로 처방의 고유한 특성을 파악하고 처방 구조를 시각적으로 표현하였다.

이러한 분석은 본초 간 관계 구조가 처방 전체의 성격을 규정하며 본초의 의미는 주변 본초와의 관계 속에서 비로소 완성된다는 네트워크 분석의 철학을 전제로 한다. 이는 전통 한의학의 핵심 배합 이론과도 긴밀하게 맞닿아 있다. 통계적으로 유의미한 결속과 삼각 모티프는 두세 본초가 짝을 이루어 나타내는 약대(藥對) 배합을 뒷받침하는 객관적 단서가 된다. 또한 중심부-주변부 계층과 중심성 편차는 거시적 관점에서 본초의 구조적 위상과 역할 분화를 조망할 수 있는 정량적 참고 지표를 제공한다.

물론 이 책에서 다룬 분석은 고전 문헌에 기록된 본초의 출현 여부를 바탕으로 한 정적(Static) 무방향 관계망이라는 방법론적 한계가 있다. 본초의 용량 차이나 병증에 따른 가감 변화, 단일 처방 내 주종 관계가 형성하는 비대칭적이고 입체적인 구조를 온전히 담아내지는 못했다.

그러나 복잡계 네트워크 과학을 바탕으로 수천 년간 축적된 한의학 고전 문헌 속에서 자생적 배합 질서와 다의적 본초 구조를 데이터로 실증한 점은 전통 지식을 현대 시스템 과학의 언어로 재해석하는 중요한 출발점이 된다. 이러한 시도는 향후 처방 지식 그래프(Knowledge Graph) 구축, 인공지능 기반 처방 설계, 다성분·다표적 치료 메커니즘 규명으로 나아가는 든든한 디딤돌이 될 것이다.

참고문헌

Albert, Réka, Hawoong Jeong, and Albert-László Barabási. 2000. “Error and Attack Tolerance of Complex Networks.” *Nature* 406 (6794): 378–82. https://doi.org/10.1038/35019019.

Autor, David H. 2014. *Polanyi’s Paradox and the Shape of Employment Growth*. Working Paper No. 20485. National Bureau of Economic Research. https://doi.org/10.3386/w20485.

Barabási, Albert-László, and Réka Albert. 1999. “Emergence of Scaling in Random Networks.” *Science* 286 (5439): 509–12. https://doi.org/10.1126/science.286.5439.509.

Barabási, Albert-László, Natali Gulbahce, and Joseph Loscalzo. 2011. “Network Medicine: A Network-Based Approach to Human Disease.” *Nature Reviews Genetics* 12 (1): 56–68. https://doi.org/10.1038/nrg2918.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. "Fast Unfolding of Communities in Large Networks." *Journal of Statistical Mechanics: Theory and Experiment* 2008 (10): P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.

Bonacich, Phillip. 1987. "Power and Centrality: A Family of Measures." *American Journal of Sociology* 92 (5): 1170–82. https://doi.org/10.1086/228631.

Brown, Peter F, John Cocke, Stephen A Della Pietra, et al. 1990. "A Statistical Approach to Machine Translation." *Computational Linguistics* 16 (2): 79–85. https://aclanthology.org/J90-2002/.

Church, Kenneth Ward, and Patrick Hanks. 1990. "Word Association Norms, Mutual Information, and Lexicography." *Computational Linguistics* 16 (1): 22–29. https://aclanthology.org/J90-1003/.

Clauset, Aaron, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. "Power-Law Distributions in Empirical Data." *SIAM Review* 51 (4): 661–703. https://doi.org/10.1137/070710111.

Di Marco, Antonio, and Roberto Navigli. 2013. "Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction." *Computational Lin-*

guistics 39 (3): 709–54. https://doi.org/10.1162/coli_a_00148.

Dunning, Ted. 1993. “Accurate Methods for the Statistics of Surprise and Coincidence.” *Computational Linguistics* 19 (1): 61–74. https://aclanthology.org/J93-1003/.

Erdős, Paul, and Alfréd Rényi. 1959. “On Random Graphs I.” *Publicationes Mathematicae Debrecen* 6: 290–97. https://doi.org/10.5486/PMD.1959.6.3-4.12.

Evert, Stefan. 2005. “The Statistics of Word Co-Occurrences: Word Pairs and Collocations.” PhD thesis, University of Stuttgart. https://doi.org/10.18419/opus-2556.

Evert, Stefan. 2008. “Corpora and Collocations.” In *Corpus Linguistics: An International Handbook*, vol. 2. Walter de Gruyter. https://doi.org/10.1515/9783110213881.2.1212.

Freeman, Linton C. 1978. “Centrality in Social Networks Conceptual Clarification.” *Social Networks* 1 (3): 215–39. https://doi.org/10.1016/0378-8733(78)90021-7.

Fruchterman, Thomas MJ, and Edward M Reingold. 1991. “Graph Drawing by Force-Directed Placement.” *Software: Practice and Experience* 21 (11): 1129–64. https:

//doi.org/10.1002/spe.4380211102.

Girvan, Michelle, and Mark EJ Newman. 2002. “Community Structure in Social and Biological Networks.” *Proceedings of the National Academy of Sciences* 99 (12): 7821–26. https://doi.org/10.1073/pnas.122653799.

Grover, Aditya, and Jure Leskovec. 2016. “Node2vec: Scalable Feature Learning for Networks.” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 855–64. https://doi.org/10.1145/2939672.2939754.

Halevy, Alon, Peter Norvig, and Fernando Pereira. 2009. “The Unreasonable Effectiveness of Data.” *IEEE Intelligent Systems* 24 (2): 8–12. https://doi.org/10.1109/MIS.2009.36.

Hopkins, Andrew L. 2008. “Network Pharmacology: The Next Paradigm in Drug Discovery.” *Nature Chemical Biology* 4 (11): 682–90. https://doi.org/10.1038/nchembio.118.

Humphries, Mark D, and Kevin Gurney. 2008. “Network ’Small-World-Ness’: A Quantitative Method for Determining Canonical Network Equivalence.” *PLoS One* 3 (4): e2051. https://doi.org/10.1371/journal.pone.0002051.

Jaccard, Paul. 1901. "'Etude Comparative de La Distribution Florale Dans Une Portion Des Alpes Et Des Jura." *Bulletin de La Soci'et'e Vaudoise Des Sciences Naturelles* 37: 547–79. https://doi.org/10.5169/seals-266450.

Johnson, Stephen C. 1967. "Hierarchical Clustering Schemes." *Psychometrika* 32 (3): 241–54. https://doi.org/10.1007/BF02289588.

Kamada, Tomihisa, and Satoru Kawai. 1989. "An Algorithm for Drawing General Undirected Graphs." *Information Processing Letters* 31 (1): 7–15. https://doi.org/10.1016/0020-0190(89)90102-6.

Kruskal, Joseph B. 1964. "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis." *Psychometrika* 29 (1): 1–27. https://doi.org/10.1007/BF02289565.

Maslov, Sergei, and Kim Sneppen. 2002. "Specificity and Stability in Topology of Protein Networks." *Science* 296 (5569): 910–13. https://doi.org/10.1126/science.1065103.

McInnes, Leland, John Healy, and James Melville. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." *arXiv Preprint arXiv:1802.03426*, ahead of print. https://doi.org/10.48550/arXiv.1802.03426.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." *arXiv Preprint arXiv:1301.3781*, ahead of print. https://doi.org/10.48550/arXiv.1301.3781.

Milo, Ron, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. 2002. "Network Motifs: Simple Building Blocks of Complex Networks." *Science* 298 (5594): 824-27. https://doi.org/10.1126/science.298.5594.824.

Molloy, Michael, and Bruce Reed. 1995. "A Critical Point for Random Graphs with a Given Degree Sequence." *Random Structures & Algorithms* 6 (2-3): 161-80. https://doi.org/10.1002/rsa.3240060204.

Newman, Mark. 2018. *Networks*. 2nd ed. Oxford University Press. https://doi.org/10.1093/oso/9780198805090.001.0001.

Orsini, Chiara, Marija Mitrović Dankulov, Pol Colomer-de-Simón, et al. 2015. "Quantifying Randomness in Real Networks." *Nature Communications* 6: 8627. https://doi.org/10.1038/ncomms9627.

Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank Citation Ranking: Bringing Order to the Web*. Nos. 1999-66. Stanford InfoLab. http://infolab.stanford.edu/~backrub/google.html.

Raghavan, Usha Nandini, R'eka Albert, and Soundar Kumara. 2007. "Near Linear Time Algorithm to Detect Community Structures in Large-Scale Networks." *Physical Review E* 76 (3): 036106. https://doi.org/10.1103/PhysRevE.76.036106.

Rosvall, Martin, and Carl T Bergstrom. 2008. "Maps of Random Walks on Complex Networks Reveal Community Structure." *Proceedings of the National Academy of Sciences* 105 (4): 1118–23. https://doi.org/10.1073/pnas.0706851105.

Seidman, Stephen B. 1983. "Network Structure and Minimum Degree." *Social Networks* 5 (3): 269–87. https://doi.org/10.1016/0378-8733(83)90028-X.

Shannon, Claude E. 1948. "A Mathematical Theory of Communication." *The Bell System Technical Journal* 27 (3): 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Sutton, Richard. 2019. "The Bitter Lesson." *Incomplete Ideas (Blog)* 13 (1): 12. http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

Torgerson, Warren S. 1952. "Multidimensional Scaling: I. Theory and Method." *Psychometrika* 17 (4): 401–19. https://doi.org/10.1007/BF02288916.

Traag, Vincent A, Ludo Waltman, and Nees Jan Van Eck. 2019. "From Louvain to

Leiden: Guaranteeing Well-Connected Communities." *Scientific Reports* 9 (1): 5233. https://doi.org/10.1038/s41598-019-41695-z.

Watts, Duncan J, and Steven H Strogatz. 1998. "Collective Dynamics of 'Small-World' Networks." *Nature* 393 (6684): 440–42. https://doi.org/10.1038/30918.

Zhang, Yifang, Xiaohua Li, Yuhong Shi, et al. 2023. "ETCM V2.0: An Update with Comprehensive Resource and Rich Annotations for Traditional Chinese Medicine." *Acta Pharmaceutica Sinica B* 13 (6): 2559–71. https://doi.org/10.1016/j.apsb.2023.03.012.

Zipf, George Kingsley. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press. https://archive.org/details/in.ernet.dli.2015.90211.

오준호. 2016. "고의서에 나타난 경혈과 병증의 연관성 측정 및 시각화: 침구자생경 분석 예를 중심으로." *Korean Journal of Acupuncture* 33 (1): 18–32. https://doi.org/10.14406/acu.2016.003.

오준호. 2019. "한의학 고문헌 데이터 분석을 위한 단어 임베딩 기법 비교: 자연어처리 방법을 적용하여." *대한한의학원전학회지* 32 (1): 61–74. https://doi.org/10.14369/jkmc.2019.32.1.061.

오준호. 2022. “기계 학습을 이용한 한의학 용어 유의어 사전 구축 방안.” *대한한의학원전학회지* 35 (1): 93-102. https://doi.org/10.14369/jkmc.2022.35.1.093.

한국한의학연구원. 2024. *한의학고전DB*. https://mediclassics.kr/.

출판사항

서명: 네트워크 분석으로 보는 한의학 처방과 본초 : 이론편

저자: 오준호

발행인: 고성규

발행처: 한국한의학연구원

발행일: 2026-10-31

ISBN: 978-89-5970-637-2 (PDF) / 978-89-5970-636-5 (세트)

주소: 대전시 유성구 유성대로 1672

이 책의 저작권은 저작권자인 한국한의학연구원에 있습니다. 저작권법에 의해 보호를 받는 저작물로서 이에 대한 무단 편집, 복제, 전재 및 배포를 금합니다.

COPYRIGHT© 2026 BY KOREA INSTITUTE OF ORIENTAL MEDICINE. ALL RIGHTS RESERVED.

감사의 글

이 책에는 함초롬바탕, 나눔고딕, 나눔고딕코딩 글꼴이 사용되었습니다.

비매품/무료

95510

9 788959 706372

ISBN 978-89-5970-637-2 (PDF)
ISBN 978-89-5970-636-5 (세트)