数据挖掘分析报告
DATA MINING REPORT

南京中医药大学信息技术学院医学信息创新工作室
MADICAL INFORMATICS INNOVATIAN STUDIO IN INSTITUTE Of INFORMATION & TECHNOLOGY

目 录

数据挖掘研究基本信息

项目名称	乳腺癌方药配伍规律研究
挖掘目的	方药配伍规律
数据集	乳腺癌方药 177 例
研究范围	综合利用 Fp-Growth、Complex Network 等方法分析数据，探索吴门医派医家治疗便秘的用药特点。
数据数量	177 例（其中内服方 136 例，外用方 41 例）
挖掘周期	2017 年 10 月 29 日——2017 年 11 月 7 日
挖掘系统	SPSS 18.0 统计软件 XMiner V2.4 中医药数据关联分析平台 Liquorice V3.0 复杂网络分析平台 Python SciPy & MatplotPy 数据处理算法库
建模类型	Fp-Growth 模型；Complex Network 模型
算法类型	Fp-Growth; Hierarchical Network; Multiscale backbone Network
数据分析	屈丹丹、赵状、王一帆、
工程审核	包钦睿、毛美清
指导老师	杨涛
工作团队	南京中医药大学信息技术学院医学信息创新工作室

中英文对照表

英文缩写	英文全称	中文全称
Fp-Growth	Fp-Tree Growth Algorithm	频繁模式树生长算法
Complex Network	Complex Network	复杂网络
minSup	Minimal Support Rate	最小支持度
minConf	Minimal Confidence Rate	最小置信度
Multiscale networks	Multiscale networks	多尺度网络
Hierarchical backbone	Hierarchical backbone	层次骨干网
Layer Num	Layer Numer	层次数目
Degree	Degree	节点度
Already Prescriptions	Already Prescriptions	附方
Ass.	Association	关联
Clustering Analysis	Clustering Analysis	聚类分析
Confidence	Confidence	置信度
Distribution	Distribution	分布
External Association Rule	External Association Rule	外关联规则
Fre.	Frequency	频率
Internal Association Rule	Internal Association Rule	内关联规则
K-means Clustering Analysis	K-means Clustering Analysis	K-均值聚类分析
Original Words	Original Words	原生词
Principle of Treatment	Principle of Treatment	治则治法
Pulse Condition	Pulse Condition	脉象
Qua.	Quantity	频次
Se.	Sequence	序列
Site Web	Site Web	位点结构
Sta.	Standard	规范对照
Standard Words	Standard Words	规范词
Support	Support	支持度
System Clustering Analysis	System Clustering Analysis	系统聚类分析
TCM Diagnosis	Tradition Chinese Medicine Diagnosis	中医疾病诊断
TCM Syndrome Diagnosis	Tradition Chinese Medicine Syndrome Diagnosis	中医证候诊断
Western Medicine Diagnosis	Western Medicine Diagnosis	西医疾病诊断

1 研究目的

乳腺癌方药配伍规律研究。

2 研究方法

2.1医案资料来源

数据来自《外科全生集》、《饲鹤亭集方》、《医方简义》、《外科正宗》、《金鉴》、《全国中药成药处方集》、《疡医大全》、《圣济总录》、《喉科心法》、《医门补要》、《医林纂要》、《外科大成》、《外科集腋》、《医部全录》、《千金》、《药庵医学丛书·论医集》、《集验良方》、《内外科百病验方大全》、《理瀹》、《青囊秘诀》、《医门八法》、《顾氏医径》、《惠直堂方》、方出《奇方类编》、《古方汇精》、《揣摩有得集》、《扁鹊心书·神方》、《马培之医案》、《竹林女科》、《医事启源》、《疡科选粹》、《简明中医妇科学》、《疡科心得集·方汇补遗》、《仙拈集》、《玉机微义》、《外台》等。由指导老师提供。

2.2医案的预处理

对原始数据进行预处理，分为“内服方剂”136 例、“外用方剂”41 例，处理过程如下：

（1）删除药物的非法字符，如空格、制表符等，其中空格 178 个，制表符共 154 个；

（2）补全药物名称。内服方第 85 条记录有两个“公英”，删除 1 各，且将“公英”补充为“蒲公英”；

（3）删除冗余信息。内服方第 55 条记录，“炙僵蚕三”改为“炙僵蚕”；

（4）删除数字，如内服方第 48 条“海藻 1”改为“海藻”，72 条“丹参桃仁 1”拆分为“丹参”、“桃仁”，“甘草 1”改为“甘草”；78 条“天龙 3g”改为“天龙”；

（5）删除括号等字符，如 88 条、90 条删除“先煎”，93 条“天麦东（各）”改为“天门冬、麦门冬”；

（6）内服方剂 56 条，“川楝子各”修改为“川楝子”；57 条“莪术各”修改为“莪术”；85 条“陈皮各”修改为“陈皮”；86 条“牡蛎各”修改为“生牡蛎”；

（7）名称统一，“天冬”统一为“天门冬”；

（8）药物拆分，内服方第 96 条“茜草根白芥子茯苓”改为“茜草根、白芥子、茯苓”；

（9）内服方第 128 条，“当归）”改为“当归”。

详见“1 数据清洗：乳腺癌数据库数据收录.xlsx”中的“清洗说明”Sheet。

2.3数据分析过程

2.3.1 数据格式处理

对“药物组成”字段进行分析，将数据转化为分析平台要求的矩阵格式。详见“1 数据

清洗：乳腺癌数据库数据收录.xlsx”中的“内服方格式化”、“外用方格式化”的相关 sheet。

2.3.2 医案信息数据挖掘方法

本次数据研究方法采用 Python SciPy & MatplotPy 编程实现数据整理和格式化；采用 XMiner 进行 Fp-Growth 算法建模，进行方药数据关联分析；采用 Liquorice 进行 Complex Network 建模，进行网络分析。

3 研究结果

3.1描述性统计结果

3.1.1内服方药物频次分布

表 1 内服方药物频次分布（前 30 味）

药物	频次	药物	频次	药物	频次
当归	69	蒲公英	19	山慈菇	14
甘草	33	白芷	18	肉苁蓉	14
茯苓	30	莪术	18	生甘草	13
白芍	29	青皮	16	金银花	13
白术	25	乳香	16	没药	13
党参	24	山茱萸	16	生黄芪	13
黄芪	23	陈皮	16	连翘	12
川芎	22	川贝母	16	香附	12
夏枯草	21	淫羊藿	15	漏芦	12
柴胡	19	人参	14	麝香	11

注：详见“2 内服方-药物频次：乳腺癌数据库数据收录.xlsx”。

3.1.2 外用方药物频次分布

表 2 外用方药物频次分布（前 30 味）

药物	频次	药物	频次	药物	频次
冰片	6	川芎	4	麝香	4
当归	6	乌药	4	蟾酥	4
川乌	5	香附	4	轻粉	4
白蔹	5	白芷	4	半夏	3
乳香	5	白及	4	陈皮	3
没药	5	生姜	4	青皮	3
血竭	5	葱白	4	羌活	3
儿茶	5	槐枝	4	枳壳	3
香油	5	穿山甲	4	僵蚕	3
硇砂	4	樟脑	4	白芥子	3

注：详见“2 外用方-药物频次：乳腺癌数据库数据收录.xlsx”。

3.2关联分析

3.2.1 内服方频繁项集

表 3 内服方药物二项频繁集（前 20 条）

编号	项集	支持度数	支持度	编号	项集	支持度数	支持度
1	白芍, 当归	22	0.1618	11	人参, 当归	13	0.0956
2	白术, 茯苓	19	0.1397	12	肉苁蓉, 莪术	13	0.0956
3	川芎, 当归	18	0.1324	13	白术, 当归	13	0.0956
4	甘草, 当归	18	0.1324	14	淫羊藿, 山茱萸	12	0.0882
5	黄芪, 当归	17	0.125	15	山茱萸, 白术	12	0.0882
6	茯苓, 当归	16	0.1176	16	山茱萸, 莪术	12	0.0882
7	肉苁蓉, 淫羊藿	14	0.1029	17	白芷, 当归	12	0.0882
8	淫羊藿, 莪术	14	0.1029	18	肉苁蓉, 山茱萸	11	0.0809
9	党参, 茯苓	14	0.1029	19	柴胡, 当归	11	0.0809
10	没药, 乳香	13	0.0956	20	夏枯草, 当归	11	0.0809

注：其他项集（三项频繁集、四项频繁集等）详见“3 内服方-关联分析：乳腺癌数据库数据收录.xls”中的“项集”Sheet。

3.2.2 内服方关联规则

表 4 内服方药物关联分析（minSup=0.04，minConf=0.9）

编号	规则	支持度	置信度	提升度
1	熟地→当归	0.0441	1	2.0299
2	泽兰→当归	0.0441	1	2.0299
3	瓜蒌仁→漏芦	0.0515	1	11.3333
4	瓜蒌仁, 白芷→漏芦	0.0441	1	11.3333
5	鹿角片→淫羊藿	0.0588	1	9.0667
6	鹿角片→莪术	0.0588	1	7.5556
7	鹿角片→肉苁蓉	0.0588	1	9.7143
8	枸杞子→党参	0.0588	1	5.6667
9	枸杞子→白术	0.0588	1	5.44
10	枸杞子→茯苓	0.0588	1	4.6897
11	枸杞子→莪术	0.0588	1	7.5556
12	延胡索, 丹参→三棱	0.0441	1	15.1111
13	延胡索, 三棱→丹参	0.0441	1	12.3636
14	延胡索, 丹参→山慈菇	0.0441	1	9.7143

15	延胡索, 山慈菇→丹参	0. 0441	1	12. 3636
16	延胡索, 三棱→山慈菇	0. 0441	1	9. 7143
17	延胡索, 山慈菇→三棱	0. 0441	1	15. 1111
18	延胡索, 丹参→三棱, 山慈菇	0. 0441	1	19. 4286
19	延胡索, 三棱→丹参, 山慈菇	0. 0441	1	19. 4286
20	延胡索, 丹参, 三棱→山慈菇	0. 0441	1	9. 7143

注：详见“3 内服方-关联分析：乳腺癌数据库数据收录.xls”中的“规则（minSup=0.04，minConf=0.9）”Sheet。

3.2.3 内服方依赖关系网络

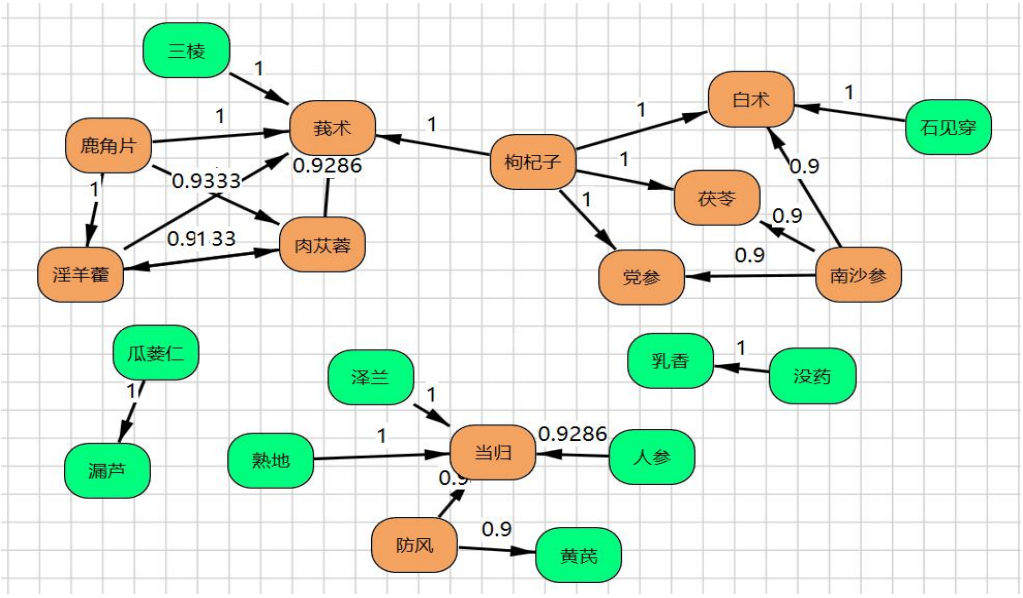


图 1 依赖关系网络（minSup=0.04，minConf=0.9）

注：关联规则、依赖关系网络详见“3 内服方-关联分析：乳腺癌数据库数据收录.xls”中的“规则（minSup=0.04，minConf=0.9）”Sheet。

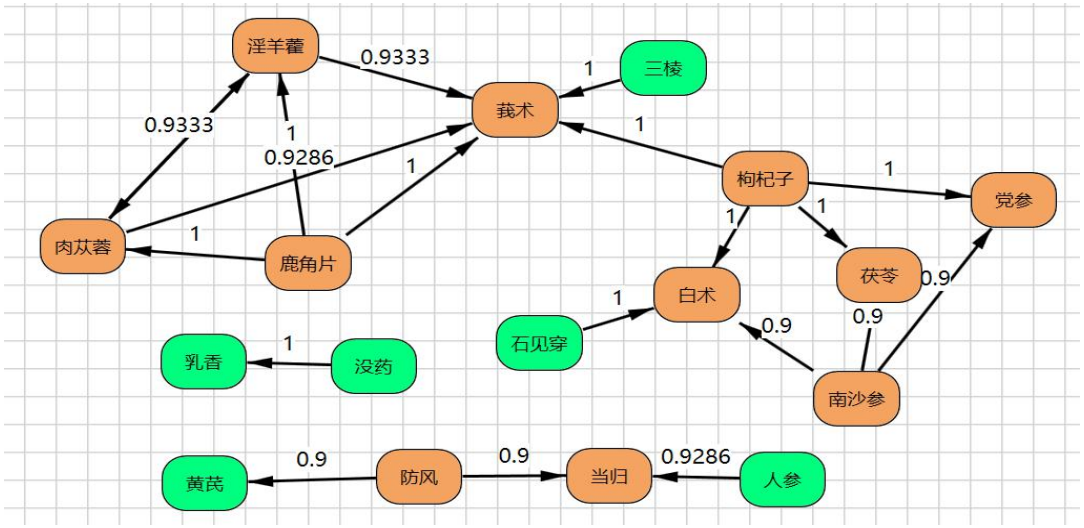


图 2 依赖关系网络（minSup=0.05，minConf=0.9）

注：关联规则、依赖关系网络详见“3 内服方-关联分析：乳腺癌数据库数据收录.xls”中的“规则（minSup=0.05，minConf=0.9）”Sheet。

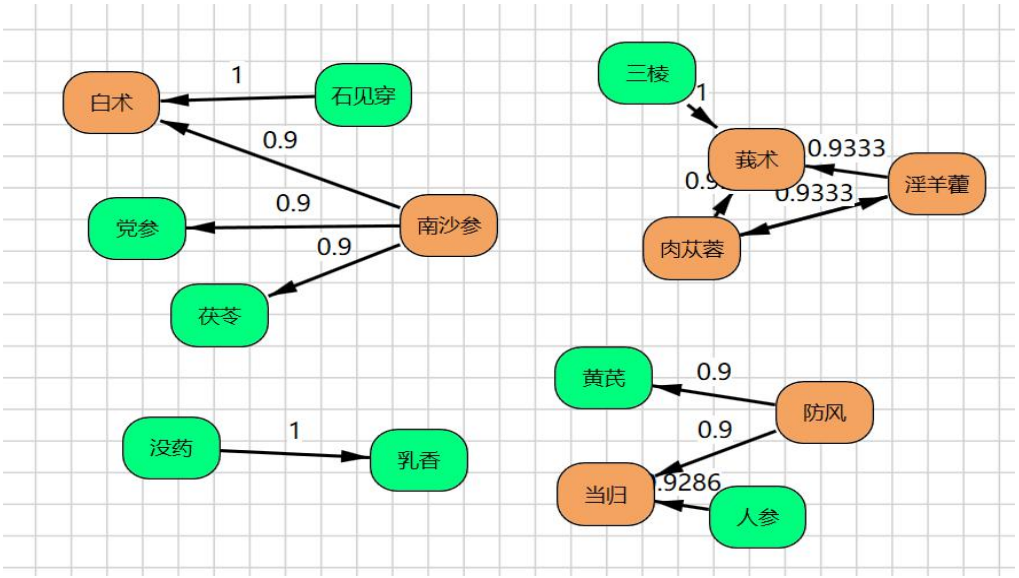


图 3 依赖关系网络（minSup=0.06，minConf=0.9）

注：关联规则、依赖关系网络详见“3 内服方-关联分析：乳腺癌数据库数据收录.xls”中的“规则（minSup=0.06，minConf=0.9）”Sheet。

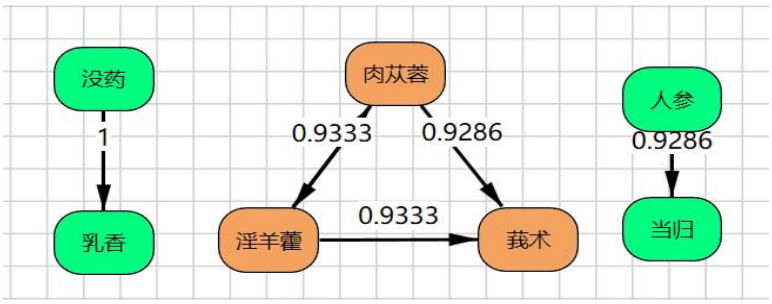


图 4 依赖关系网络（minSup=0.07，minConf=0.9）

注：关联规则、依赖关系网络详见“3 内服方-关联分析：乳腺癌数据库数据收录.xls”中的“规则（minSup=0.07，minConf=0.9）”Sheet。

3.2.4 外用方频繁项集

表 5 外用方药物二项频繁集（前 20 条）

编号	项集	支持度数	支持度	编号	项集	支持度数	支持度
1	乳香, 血竭	5	0. 122	11	葱白, 没药	4	0. 0976
2	没药, 乳香	5	0. 122	12	葱白, 乳香	4	0. 0976
3	没药, 血竭	5	0. 122	13	葱白, 生姜	4	0. 0976
4	儿茶, 没药	5	0. 122	14	葱白, 乌药	4	0. 0976
5	儿茶, 乳香	5	0. 122	15	葱白, 血竭	4	0. 0976
6	儿茶, 血竭	5	0. 122	16	白及, 白蔹	4	0. 0976

7	葱白, 白蔹	4	0.0976	17	白及, 川乌	4	0.0976
8	葱白, 川乌	4	0.0976	18	白及, 葱白	4	0.0976
9	葱白, 儿茶	4	0.0976	19	白及, 儿茶	4	0.0976
10	葱白, 槐枝	4	0.0976	20	白及, 槐枝	4	0.0976

注：其他项集（三项频繁集、四项频繁集等）详见“3 外用方-关联分析：乳腺癌数据库数据收录.xls”中的“项集”Sheet。

3.2.5 外用方关联规则

表 6 外用方药物关联分析（minSup=0.05，minConf=0.8）

编号	规则	支持度	置信度	提升度
1	枯矾→白芷	0.0732	1	10.25
2	僵蚕→川乌	0.0732	1	8.2
3	僵蚕→羌活	0.0732	1	13.6667
4	羌活→僵蚕	0.0732	1	13.6667
5	僵蚕→川乌, 羌活	0.0732	1	13.6667
6	僵蚕, 川乌→羌活	0.0732	1	13.6667
7	羌活→僵蚕, 川乌	0.0732	1	13.6667
8	僵蚕, 羌活→川乌	0.0732	1	8.2
9	川乌, 羌活→僵蚕	0.0732	1	13.6667
10	僵蚕→香附	0.0732	1	10.25
11	僵蚕→川乌, 香附	0.0732	1	13.6667
12	僵蚕, 川乌→香附	0.0732	1	10.25
13	僵蚕, 香附→川乌	0.0732	1	8.2
14	川乌, 香附→僵蚕	0.0732	1	13.6667
15	僵蚕→羌活, 香附	0.0732	1	13.6667
16	羌活→僵蚕, 香附	0.0732	1	13.6667
17	僵蚕, 羌活→香附	0.0732	1	10.25
18	僵蚕, 香附→羌活	0.0732	1	13.6667
19	羌活, 香附→僵蚕	0.0732	1	13.6667
20	僵蚕→川乌, 羌活, 香附	0.0732	1	13.6667

注：详见“3 外用方-关联分析：乳腺癌数据库数据收录.xls”中的“规则（minSup=0.05，minConf=0.8）”Sheet。

3.2.6 外用方依赖关系网络

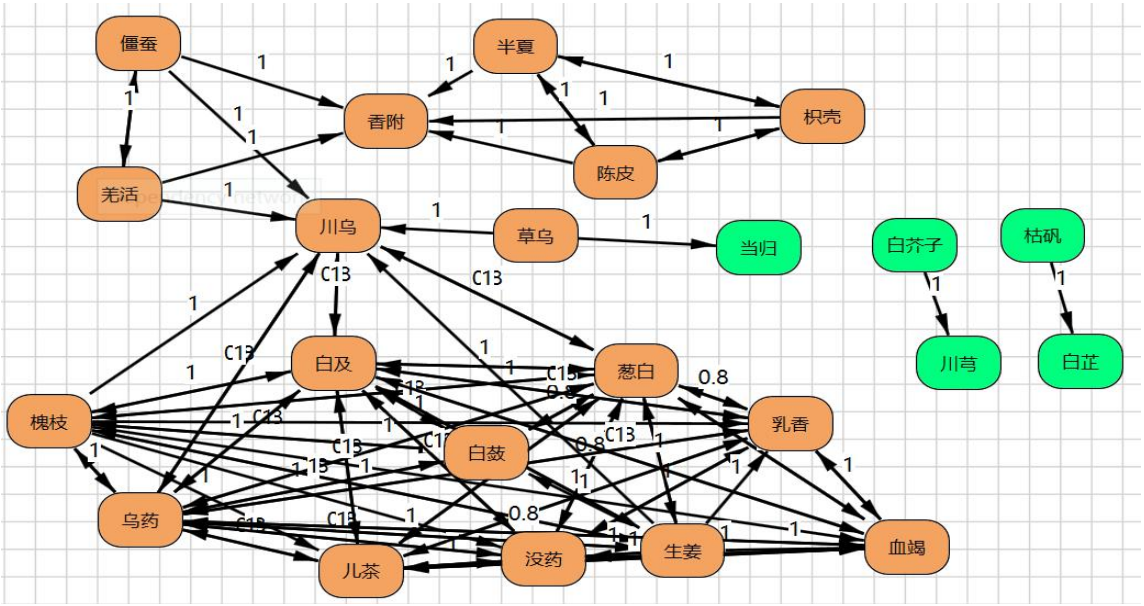


图 5 依赖关系网络（minSup=0.05，minConf=0.8）

注：关联规则、依赖关系网络详见“3 外用方-关联分析：乳腺癌数据库数据收录.xls”中的“规则（minSup=0.05，minConf=0.8）”Sheet。

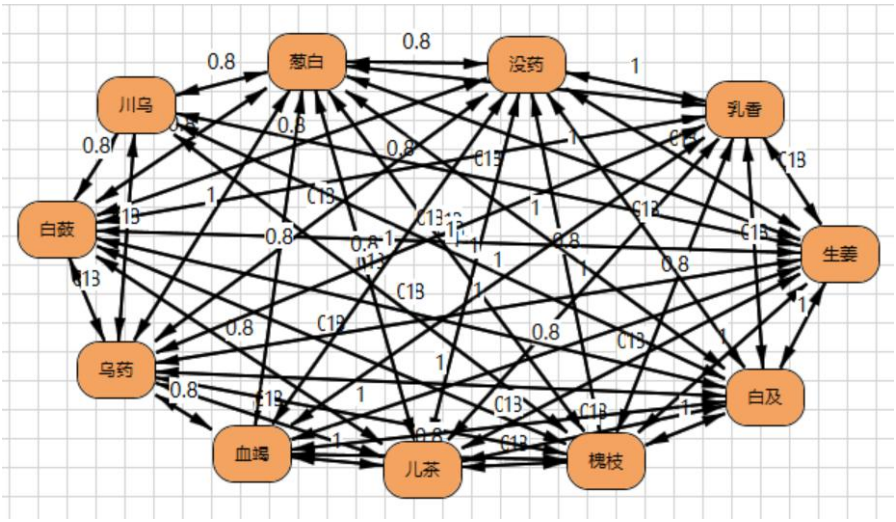


图 6 依赖关系网络（minSup=0.05，minConf=0.8）

注：关联规则、依赖关系网络详见“3 外用方-关联分析：乳腺癌数据库数据收录.xls”中的“规则（minSup=0.08，minConf=0.8）”Sheet。

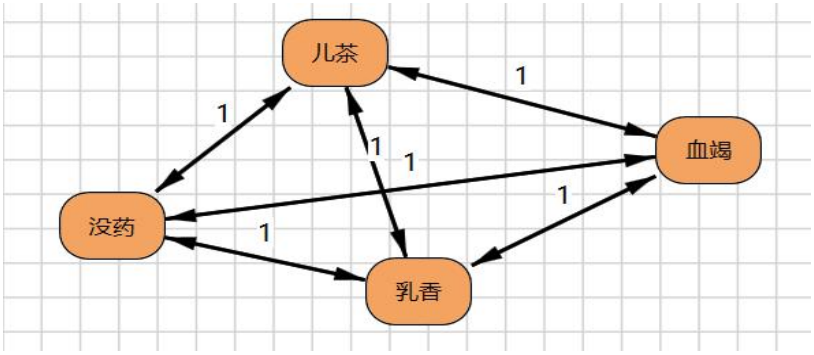


图 7 依赖关系网络（minSup=0.1，minConf=0.8）

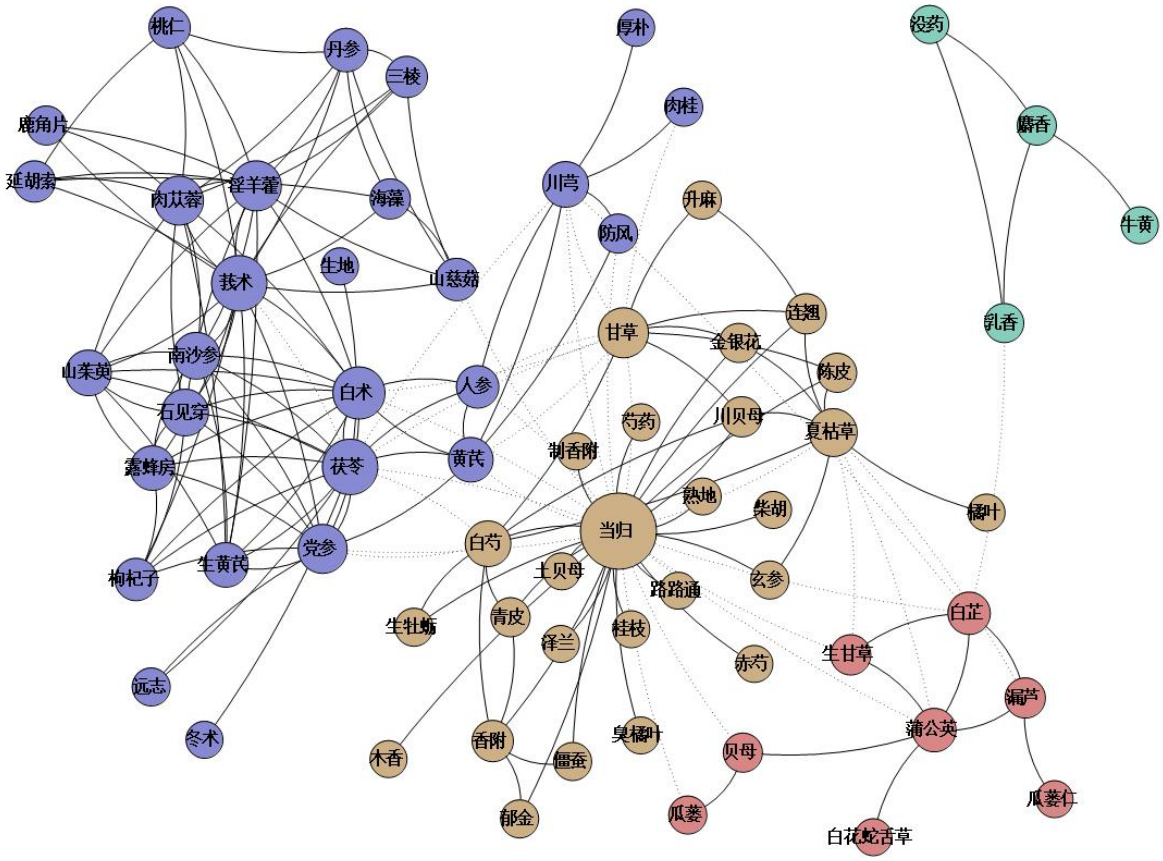
注：关联规则、依赖关系网络详见“3 外用方-关联分析：乳腺癌数据库数据收录.xls”中的“规则（minSup=0.05，minConf=0.8）”Sheet。

3.3复杂网络

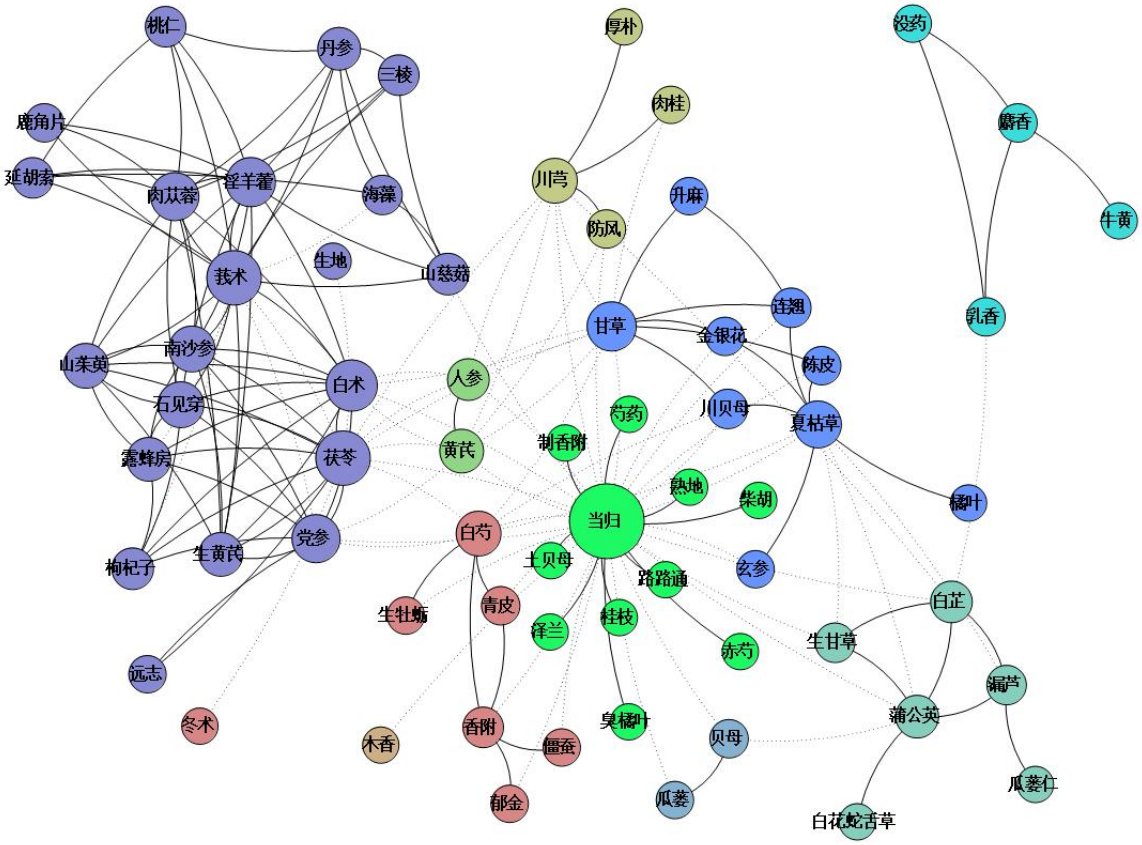
3.3.1 内服方单层药物网络

设定 Algorithm：Multiscale backbone, Wconfidence=0.95

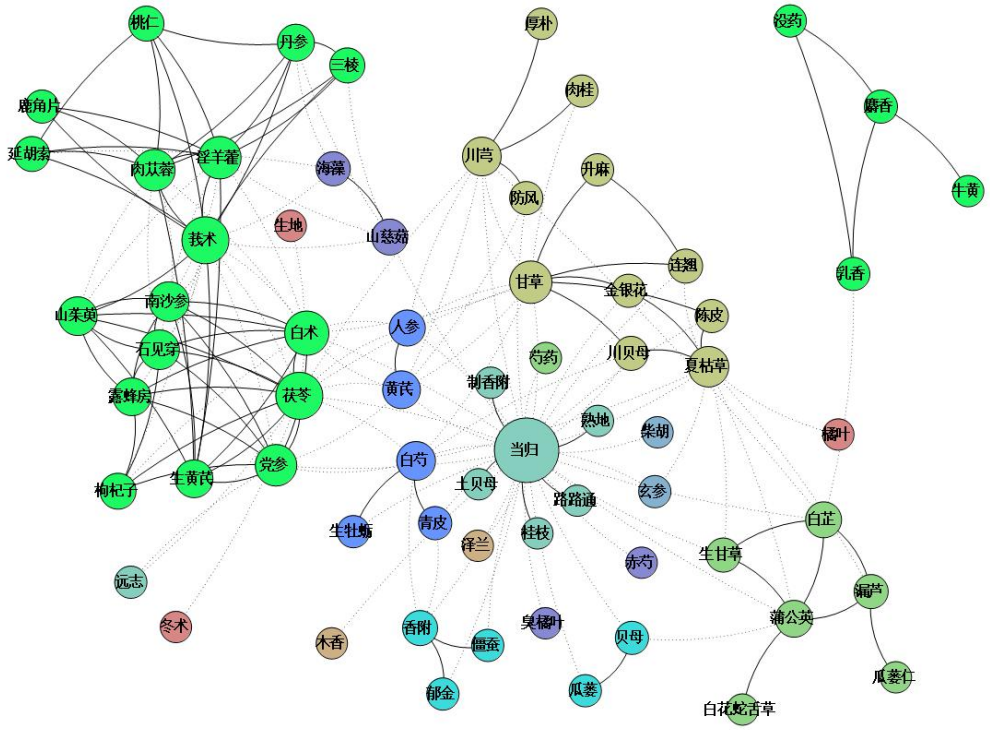
（1）聚类系数设为 30，网络拓扑：



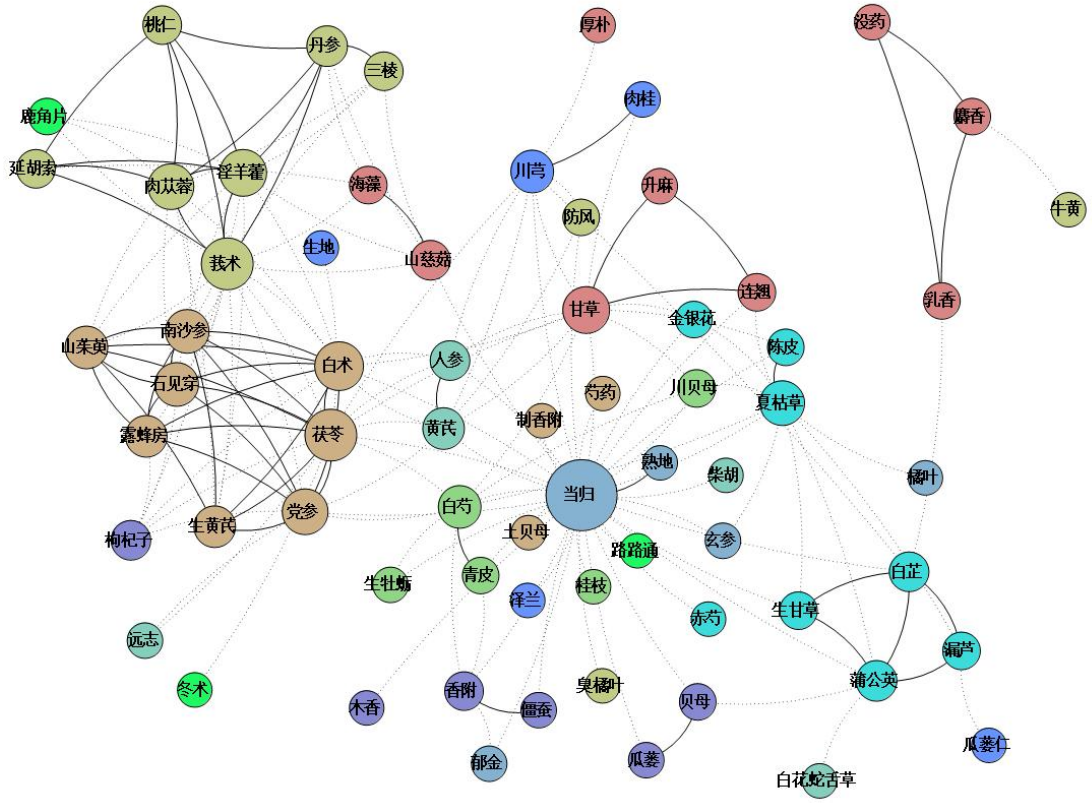
（2）聚类系数设为 60，网络拓扑：



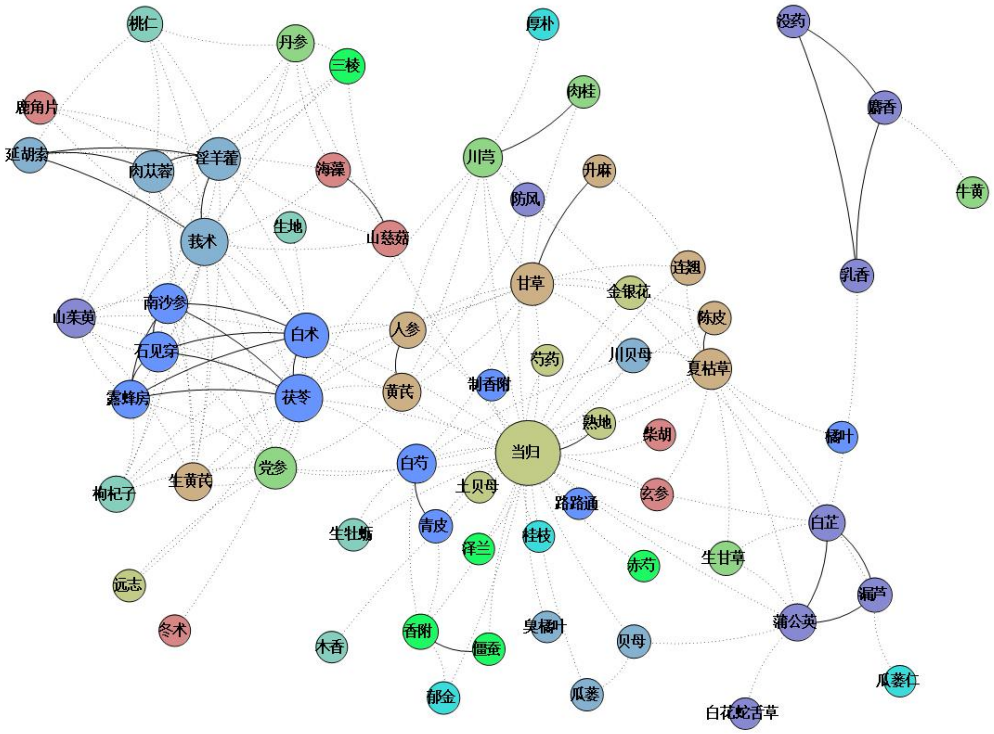
（3）聚类系数设为 90，网络拓扑：



（4）聚类系数设为 120，网络拓扑：



（5）聚类系数设为 150，网络拓扑：

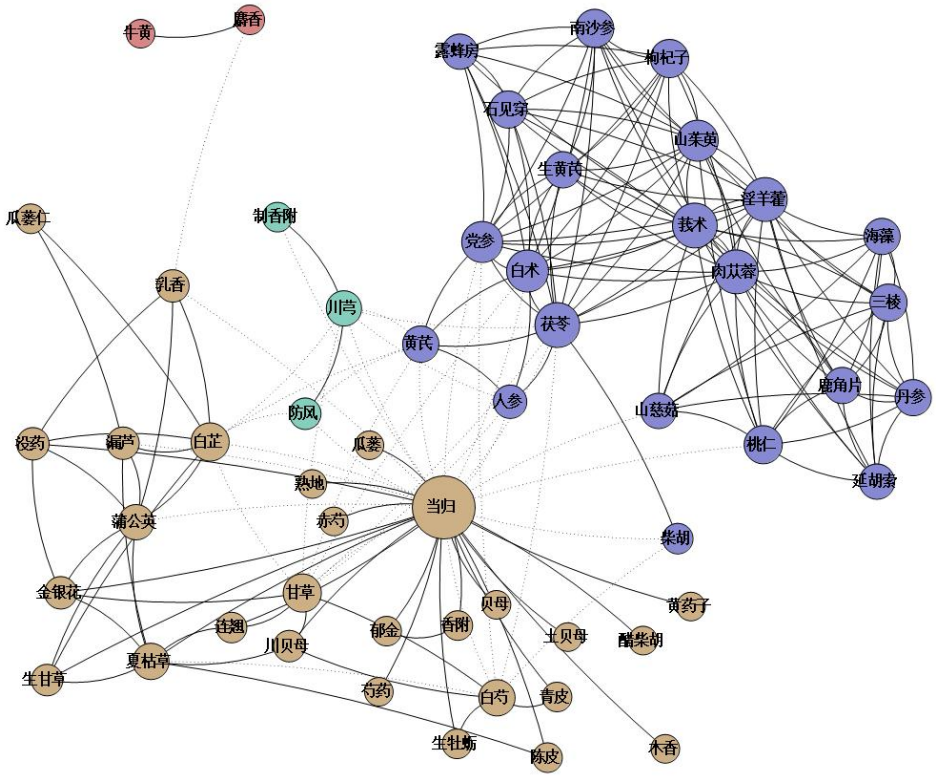


注：节点度、紧密度、拓扑结构详见“4 内服方-网络分析：乳腺癌数据库数据收录.xls”。

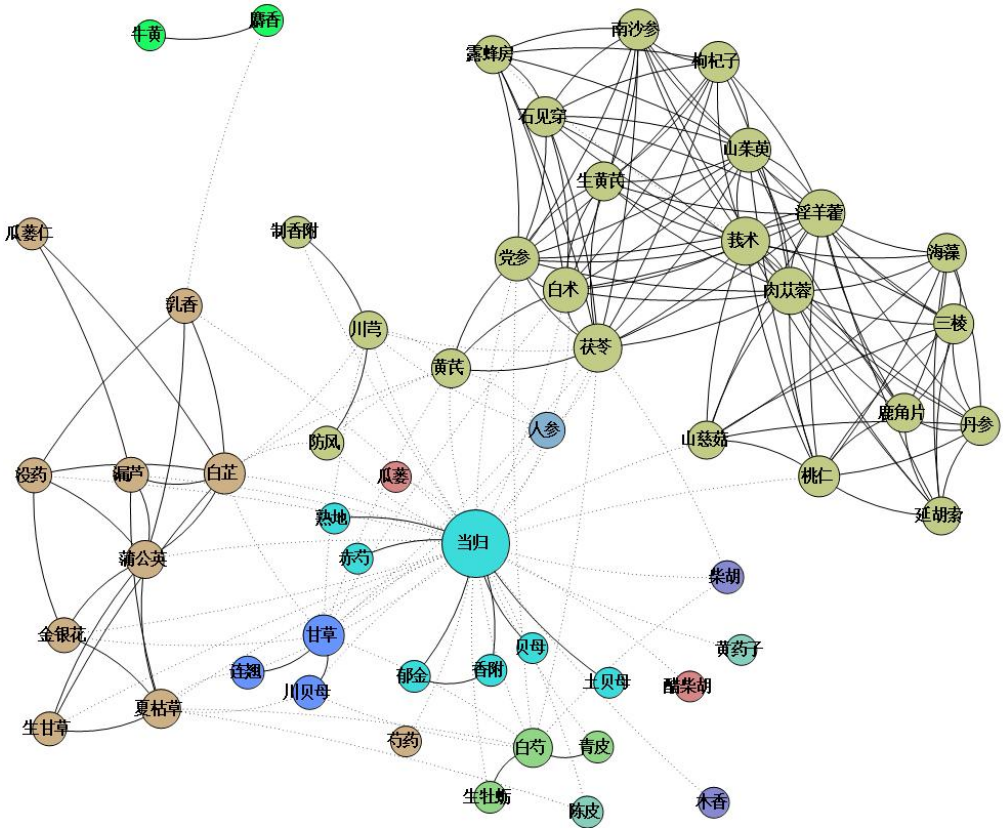
3.3.2 内服方两层药物网络

设定 Algorithm:Hierarchical Newwork，Layer Num=2，Degree Coefficient=1.3

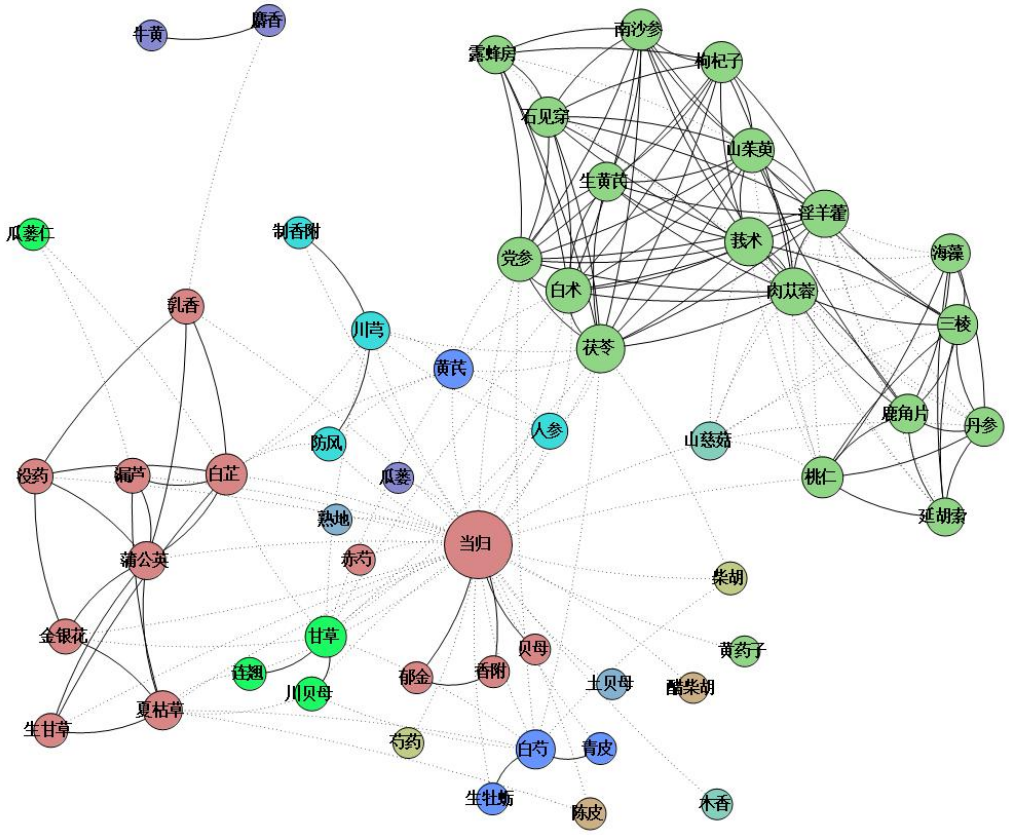
（1）聚类系数设为 30，网络拓扑：



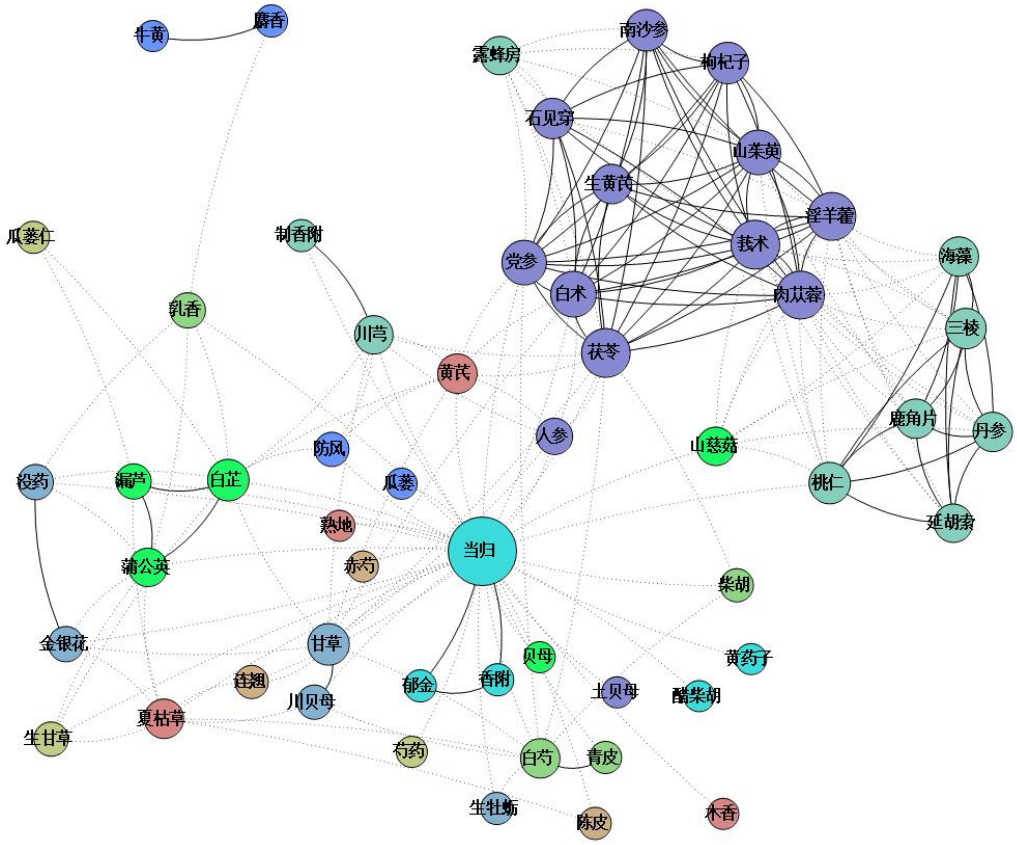
（2）聚类系数设为 60，网络拓扑：



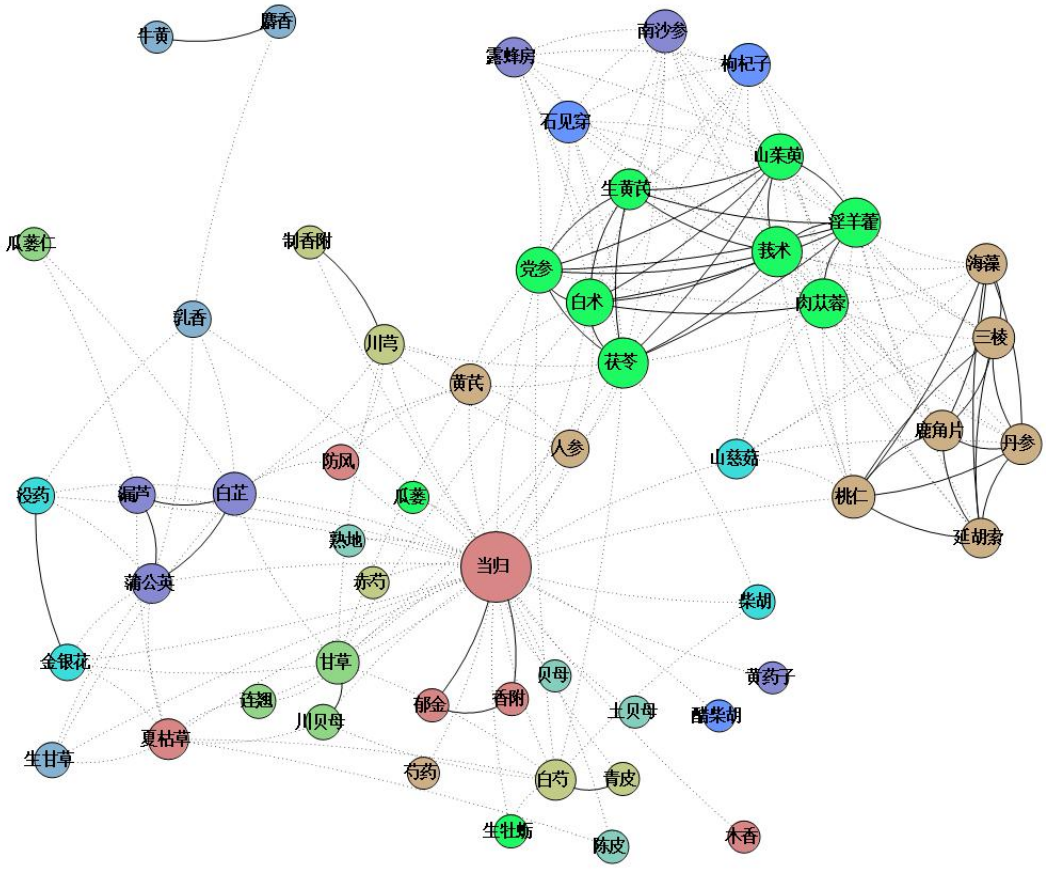
（3）聚类系数设为 90，网络拓扑：



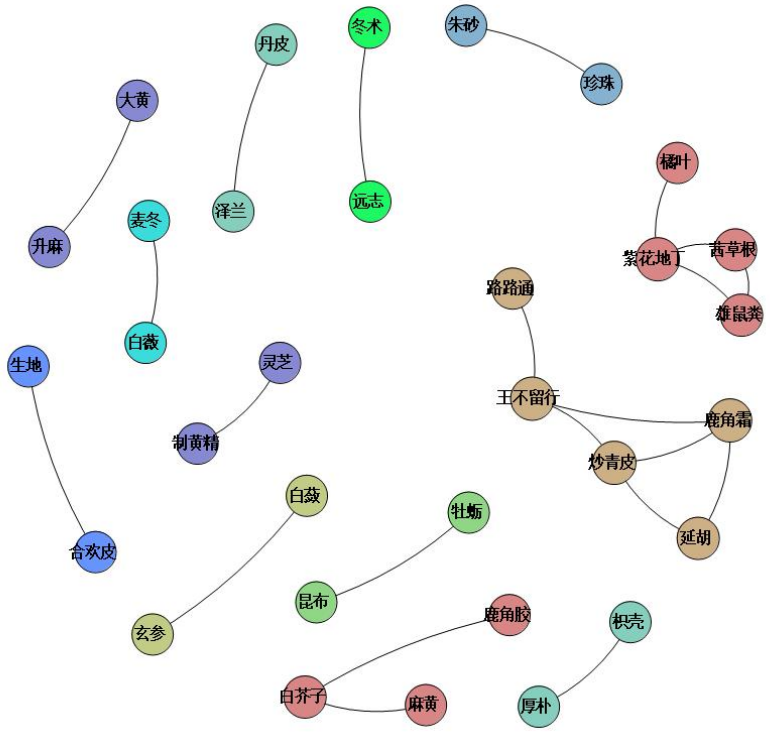
（4）聚类系数设为 120，网络拓扑：



（5）聚类系数设为 150，网络拓扑：



（6）二层网络拓扑（中药加减）

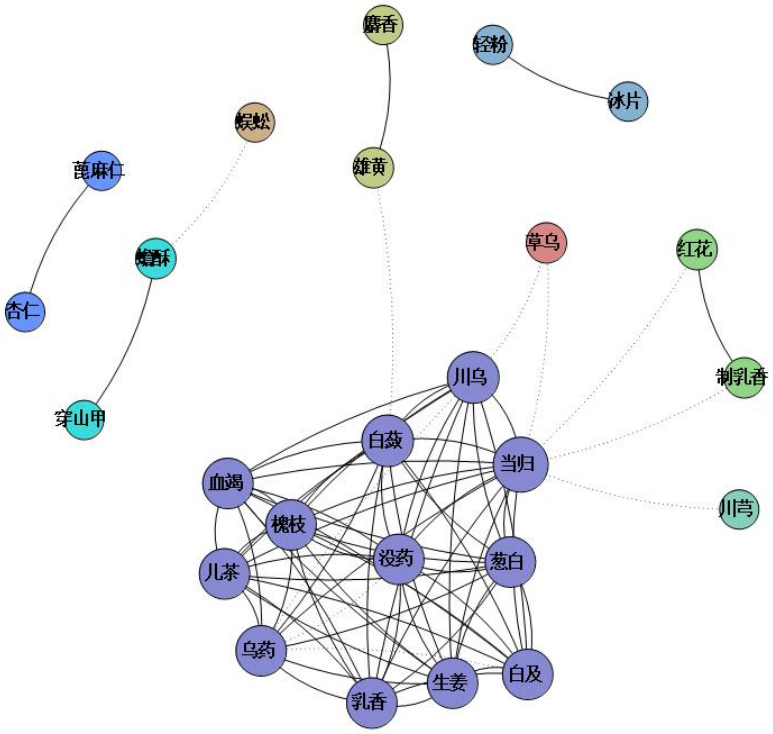


注：节点度、紧密度、拓扑结构详见“4 内服方-网络分析：乳腺癌数据库数据收录.xls”。

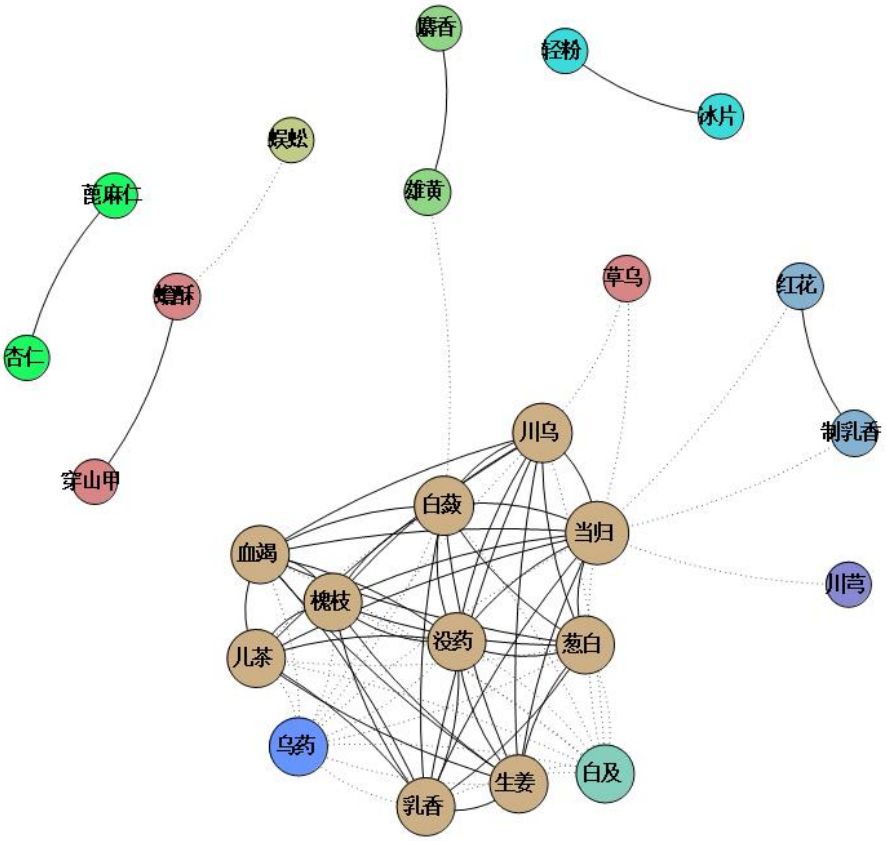
3.3.3 外用方单层药物网络

设定 Algorithm：Multiscale backbone, Wconfidence=0.85

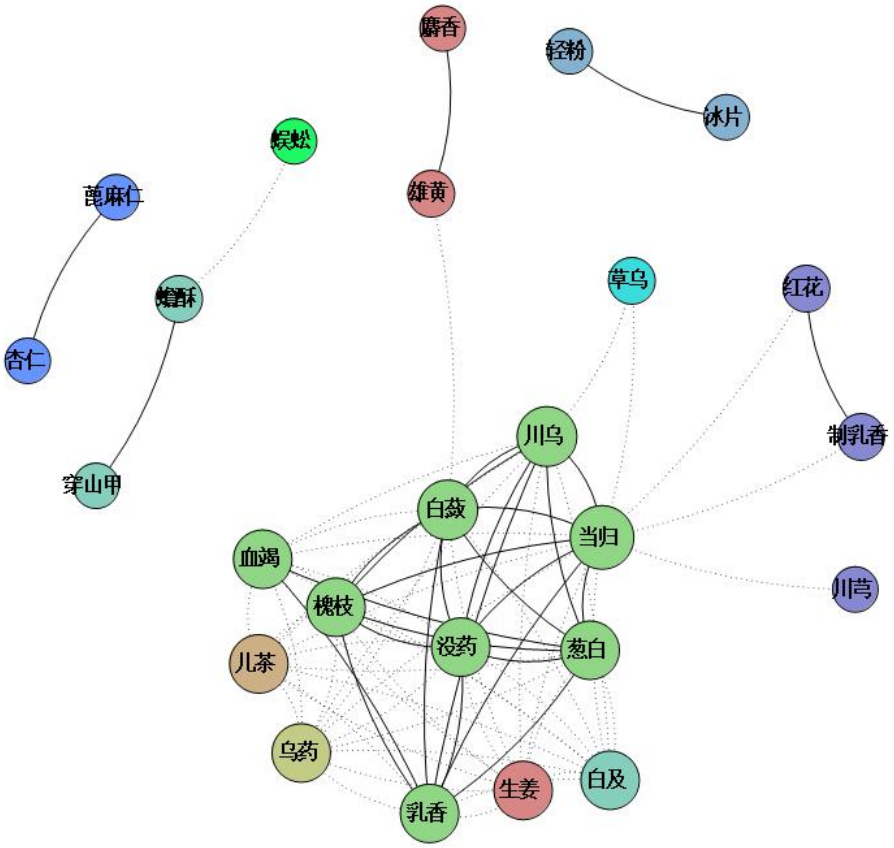
（1）聚类系数设为 10，网络拓扑：



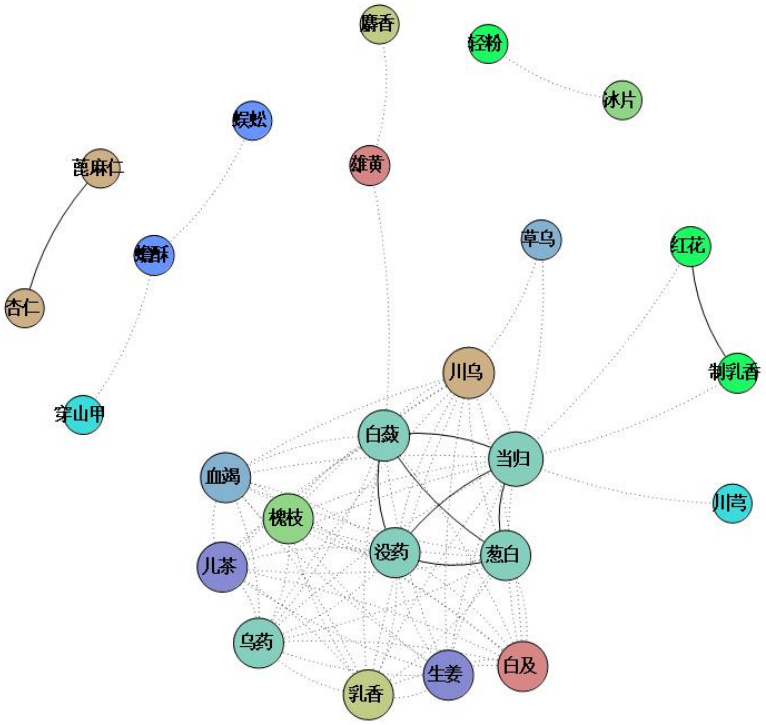
（2）聚类系数设为 30，网络拓扑：



（3）聚类系数设为 50，网络拓扑：



（4）聚类系数设为 70，网络拓扑：

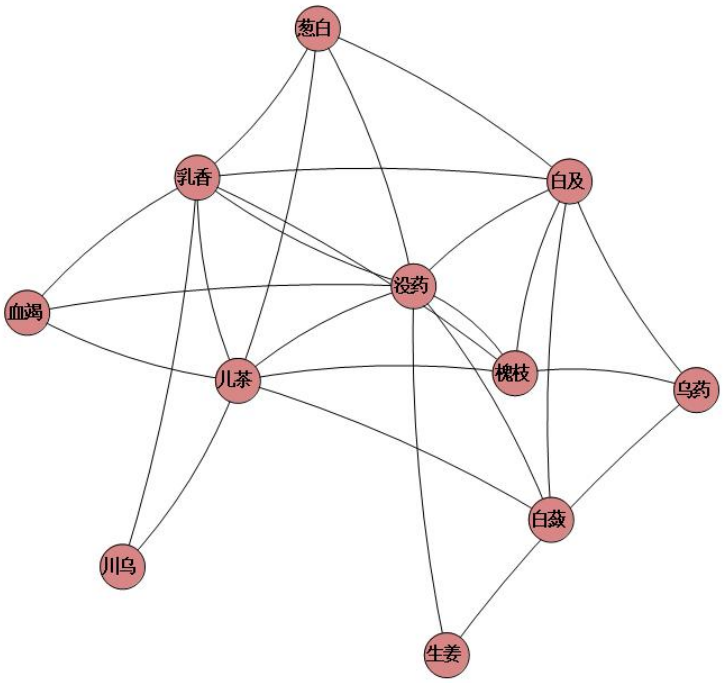


注：节点度、紧密度、拓扑结构详见“4 外用方-网络分析：乳腺癌数据库数据收录.xls”。

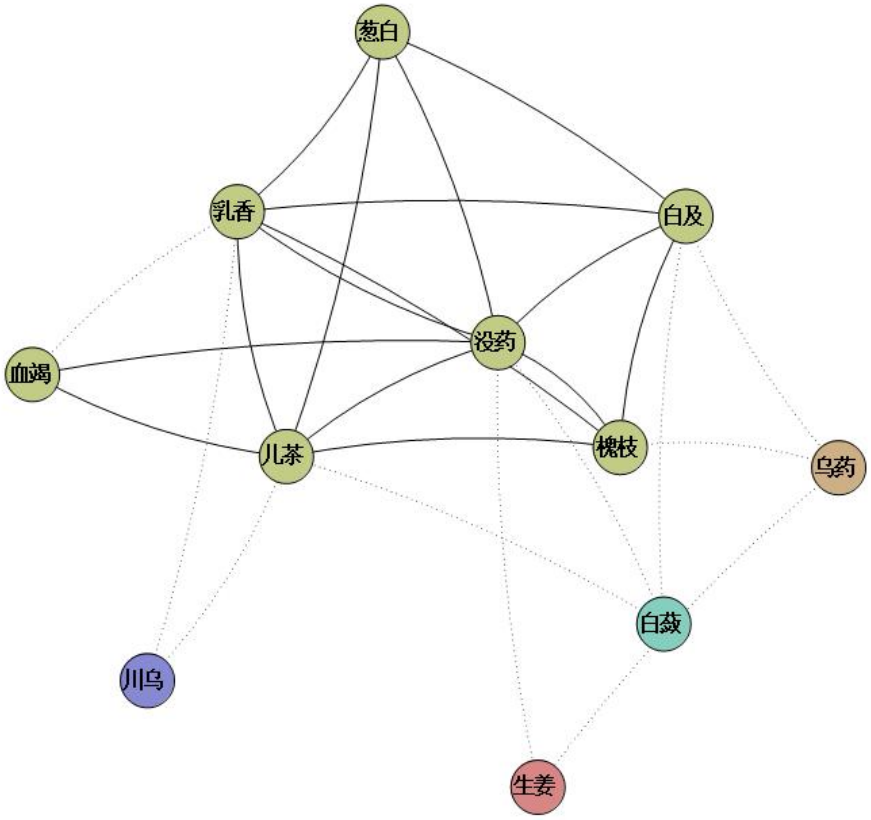
3.3.4 外用方两层药物网络

设定 Algorithm:Hierarchical Newwork，Layer Num=2，Degree Coefficient=1.3

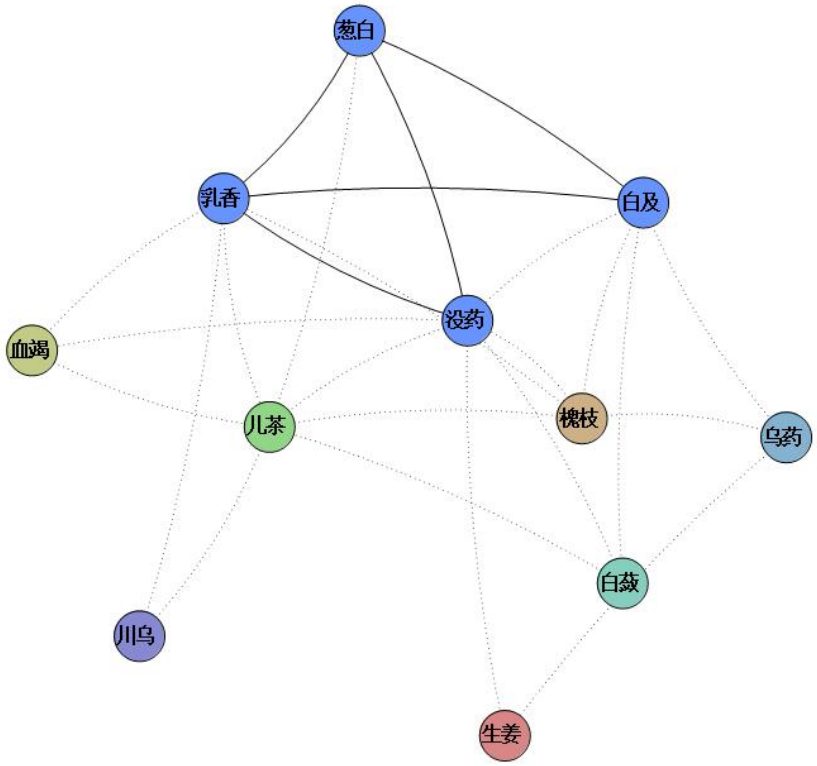
（1）聚类系数设为 0，网络拓扑：



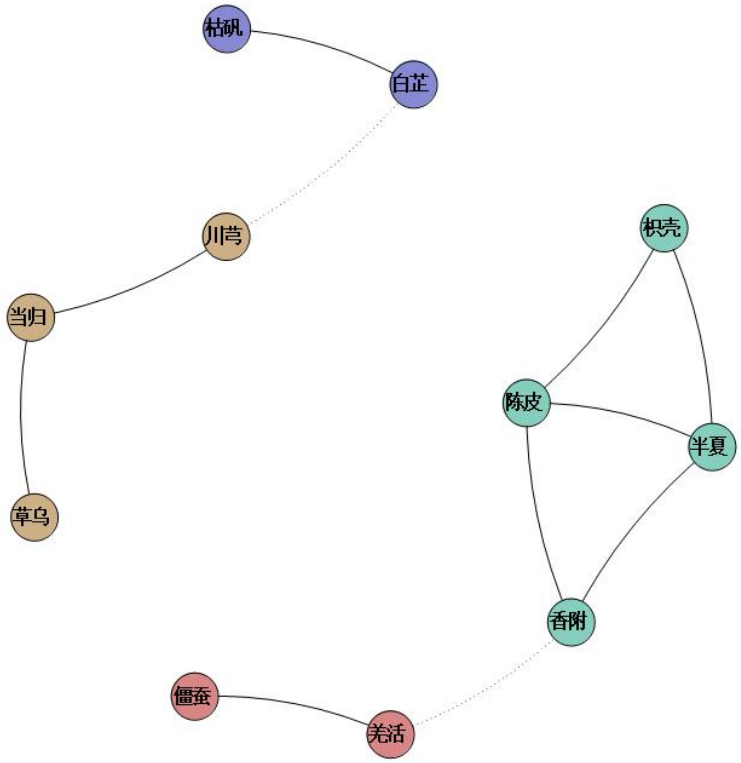
（2）聚类系数设为 10，网络拓扑：



（3）聚类系数设为 20，网络拓扑：



（6）二层网络拓扑（中药加减）



注：节点度、紧密度、拓扑结构详见“4 外用方-网络分析：乳腺癌数据库数据收录.xls”。