

**图 1** 网络舆情热点主题识别整体架构图

### 1 特征词筛选

微博作为大众社交媒体，人们在利用微博发布消息或评论时经常存在文本无意义和内容发散的现象，增加了主题概括的难度，需要从原始词表中筛选出最重要的词汇作为特征词。与文本中的其它成份相比，名词和动词更能够解释微博中所讨论的话题，因此，将每条微博进行分词、去掉停用词操作后，只保留长度大于1的名词和动词，构成词列表*Wlist*。设微博信息总条数为*N*，统计*N*条微博的总词数记为*CN*，*N*条微博中词*w*出现的频数记为*Cw*，包含词*w*的微博条数记为*Nw*，定义词*w*的词频特征表示为*Cw*/*CN*：可以反映特征词的重要程度；定义词的文档特征表示为*Nw*/*N*，对于原创微博信息而言，同一特征词出现在越多的微博文本中，表明越多的用户使用此特征词，因而，文档特征可反映不同用户对同一特征词的偏好程度。本文针对词列表中的每个词综合考虑词频特征和文档特征，保留满足如式1所示条件的词构成特征词集合：

 (1)

其中*α*、*β*表示筛选阈值。

记者/调查/网传/红十字会/售卖/物资

红十字会/售卖/山东/寿光/蔬菜

湖北/红十字会/物资/使用/情况/引发/质疑

**红十字会**

**售卖**

**物资**

### 2 特征词共现网络构建及话题社区划分

在获得特征词集合后，基于特征词出现的位置和频数构建词共现网络图[17]。设图为一加权无向图，其中，表示节点集，由3.1节所获得特征词组成；表示边集，如果两个特征词至少同时出现在一条微博中，则在这两个特征词之间建立一条边，从边集中可以反映出特征词的共现关系；*W*表示图*G*中边集*E*所对应的权重矩阵，考虑到特征词相伴出现的次数对语义分析的影响，需要赋予边相应的权重，同时为保证权重矩阵*W*的对称性，令：

 (2)

其中，即为边的权重，、、分别表示特征词，共现的次数以及特征词和各自出现的次数。

根据语言学的特点，人们在阐述问题时，相同话题对应的词汇表述更为相近，而不同话题对应的词汇表述存在差异。这种现象反映在特征词共现网络中即表现为：相同话题特征词之间联系紧密，而与其它话题联系稀疏。相同话题的特征词联系紧密，不同话题之间特征词联系稀疏。鉴于社会网络中社区发现的思想有助于探测网络内部的隐藏规律和拓扑结构，为有效识别出特征词共现网络中存在的事件，本文采用Louvain算法[18]实现舆情事件社区发现。Louvain算法是一种基于多层次优化模块度的算法，能够刻画社区的紧密程度，得到不同规模的社区结构。

### 3 事件主题抽取

经过Louvain算法获得多个话题社区，为增加每个话题社区的概括性并解决不同社区之间交叉特征词过多的问题，需要进行话题的主题抽取。首先，针对每一个话题社区，利用Kruskal算法[19]得到每个社区的最小生成子树，然后，综合考虑特征词节点的度（Degree）、介数中心性（Betweenness）和PageRank值计算特征词的重要性，设特征词*w*对应的节点记为*v*，则特征词*w*的重要性计算如下：

 (3)

其中，*D*(*v*)表示节点*v*的度，是与*v*直接相连的节点数目，即与特征词*w*共现的特征词数目，节点*v*的度越大，说明与之联系的特征词越多，自然越重要。*B*(*v*)代表节点*v*的介数中心性，计算方法如下：

 (4)

其中，*σst*表示节点*s*和节点*t*之间存在的最短路径的数量，*σst*(*v*)表示这些最短路径中经过节点*v*的数量，反映了节点*v*作为“桥梁”的重要程度。

*P*(*v*)表示节点*v*的PageRank值，*v*的PageRank值取决于它所具有邻居节点的数量和质量：

 (5)

其中，*P*(*v*)，*P*(*j*)分别是节点*v*，*j*的PageRank值，*Mv*是节点*v*的邻居节点集合，*W*(*v*, *j*)是节点*v*，*j*之间边的权重，*D*(*j*)是节点*j*的度，*N*是节点总数，*θ*一般取1。

综上所述，特征词*w*的所对应的节点*v*的重要性计算公式如下：

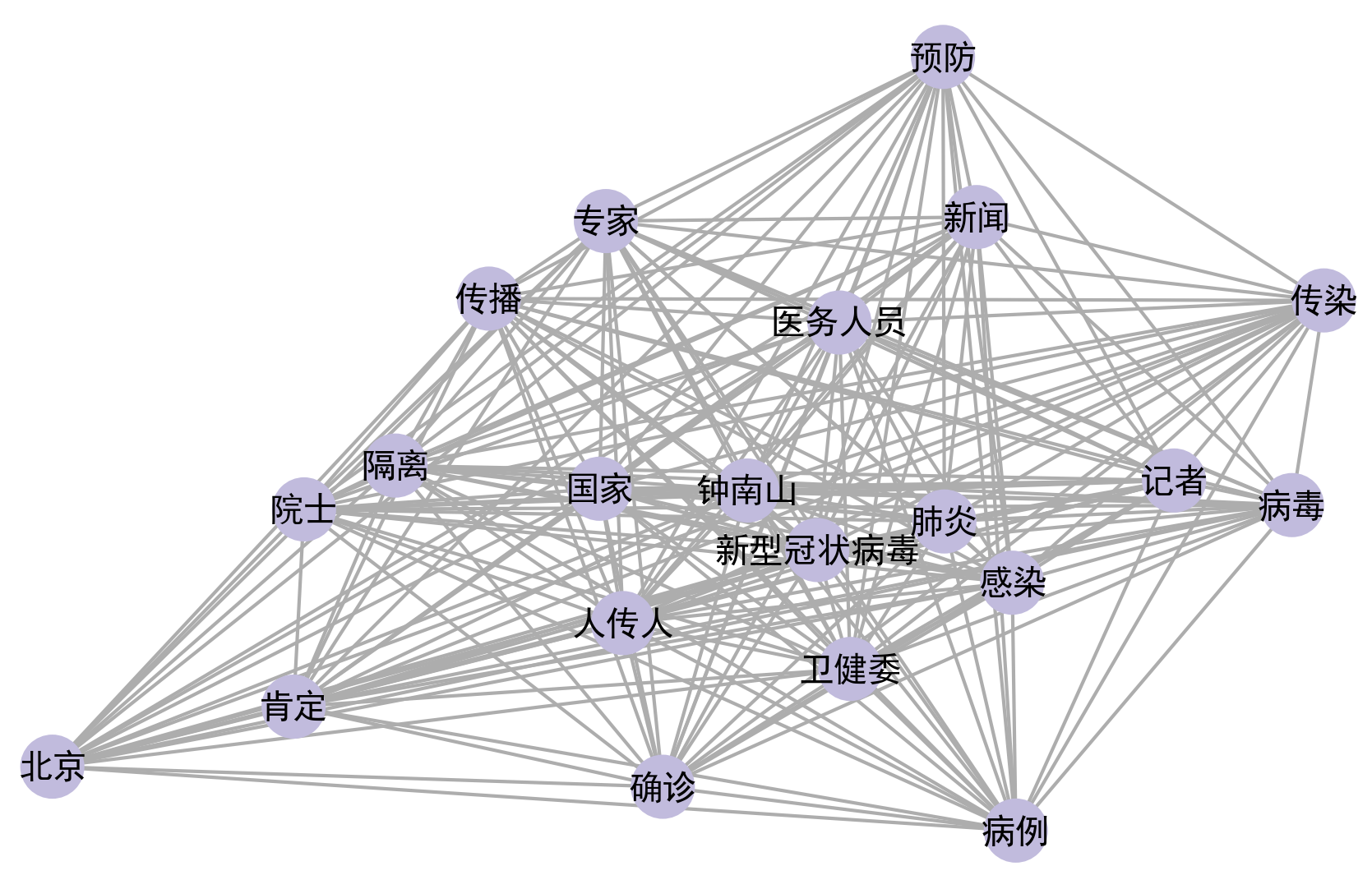
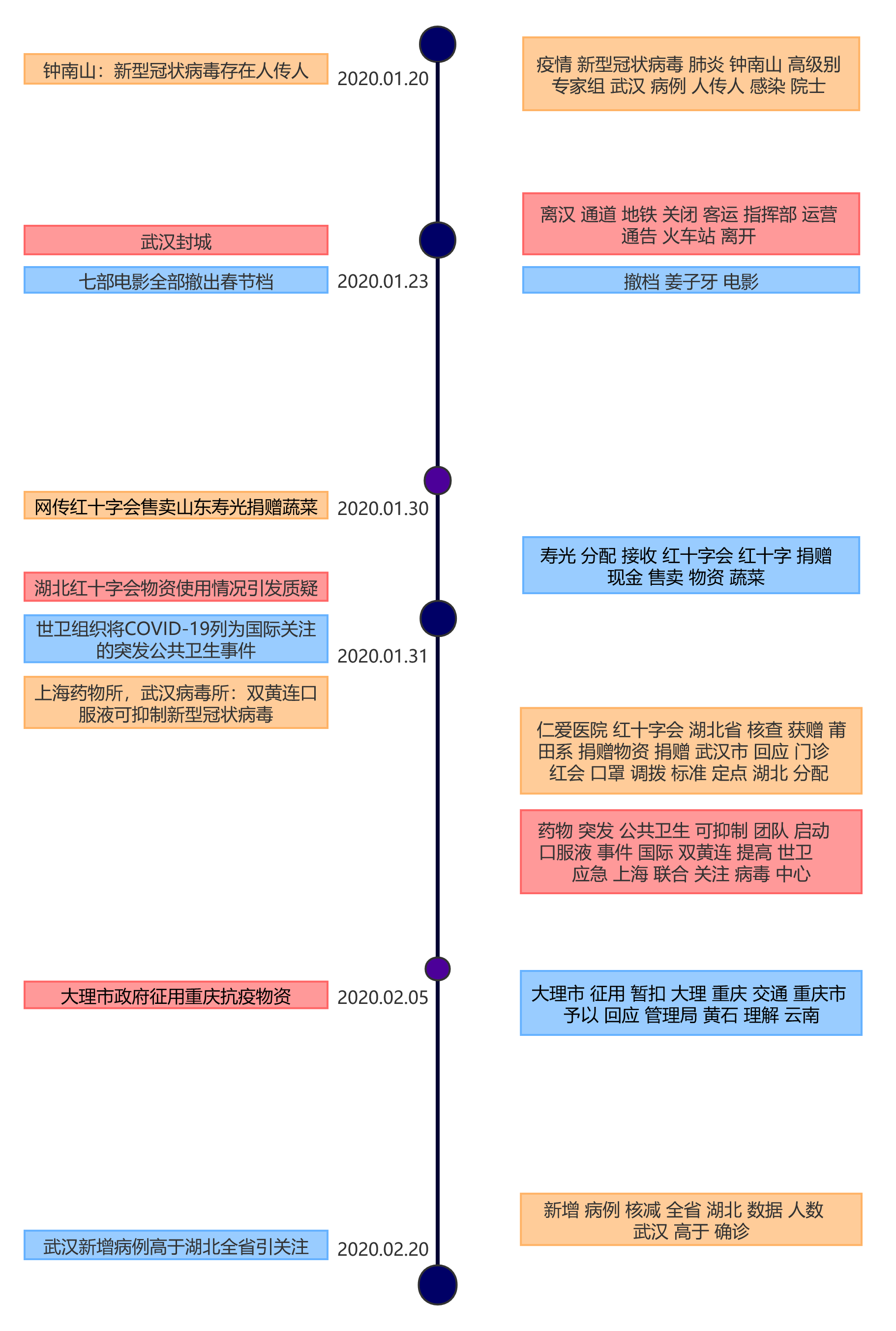
 (6)

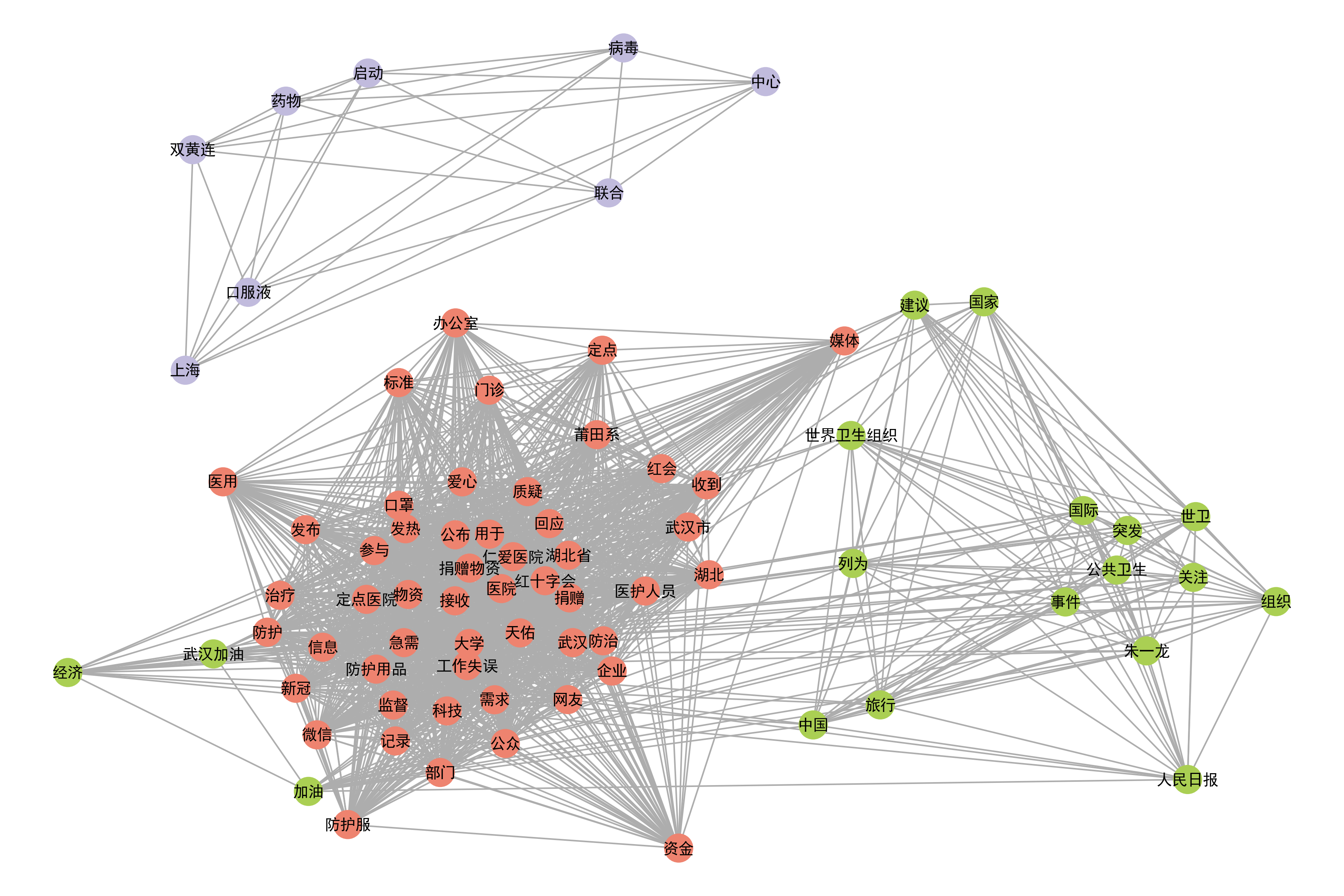
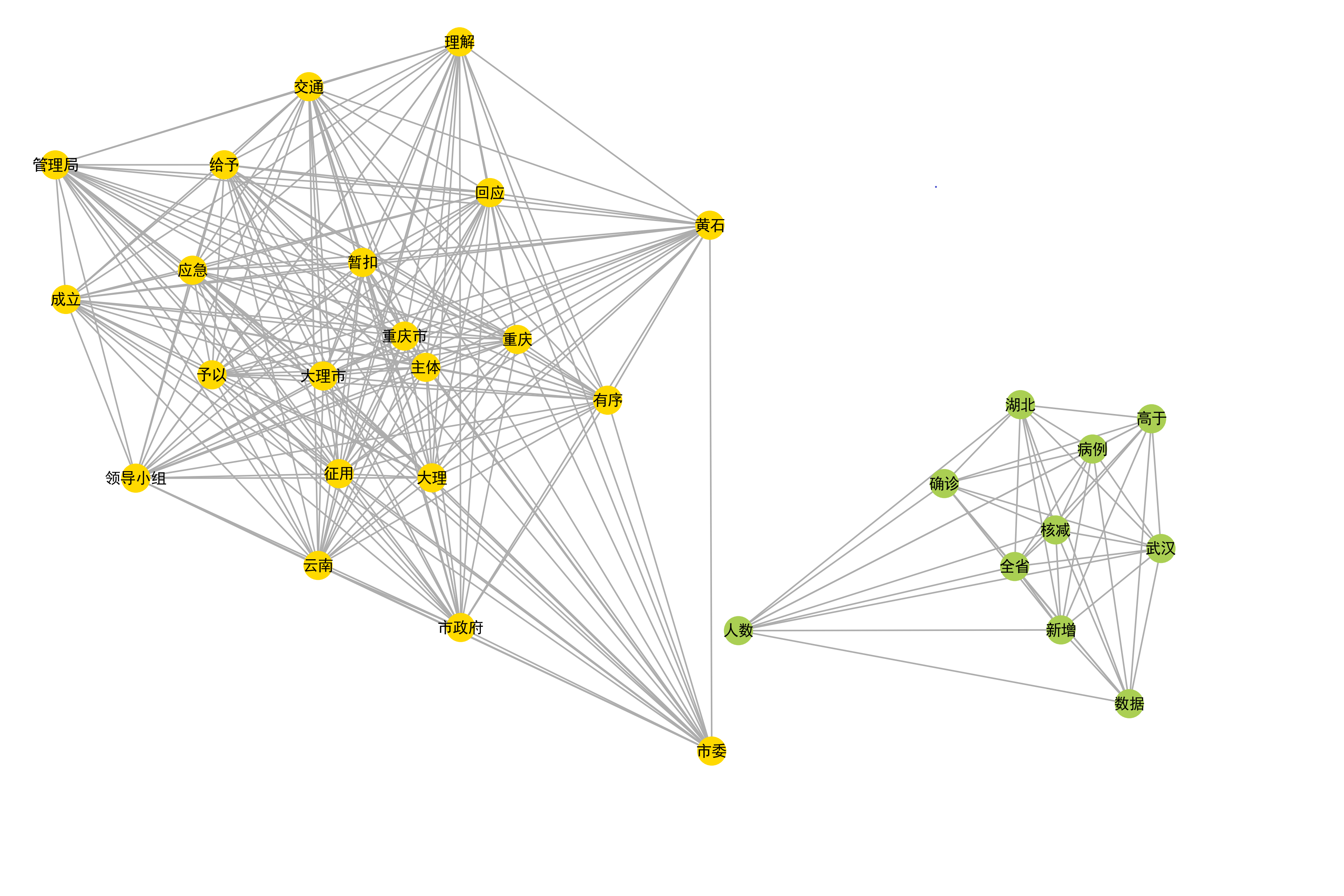
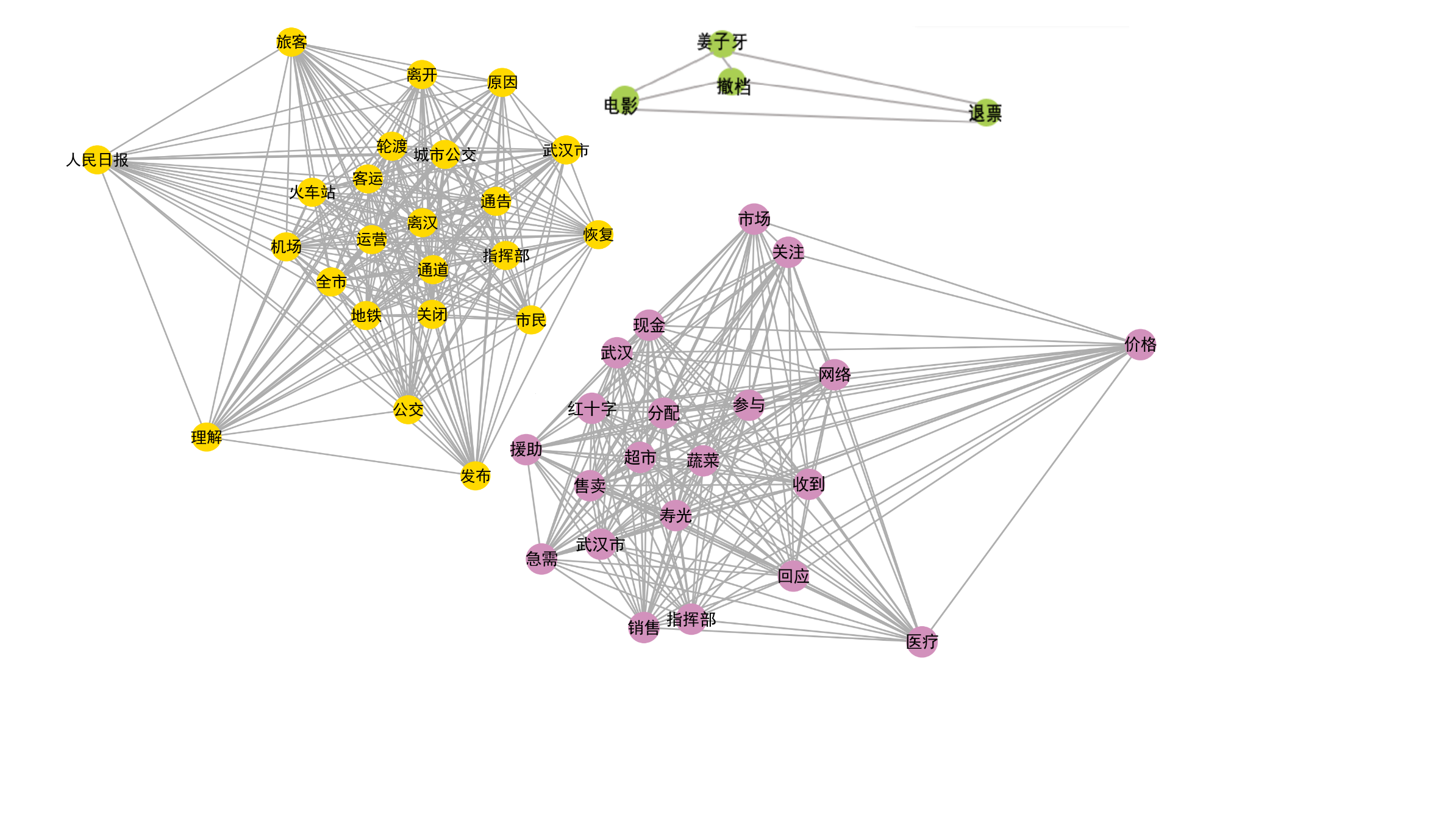
### 3.4 事件主题演化

为了更好的把握各个舆情发展阶段的舆论主题，本文对舆情数据进行时间切片，针对每一阶段的舆情社会网络，划分社区并提取主题，和下一阶段的主题词进行节点重合度比较，如果重合度高于设定的阈值，就认为存在主题演化关系。节点重合度的计算公式如下：

 (7)

其中，是社区的加权节点重合度，表示两个社区重合的节点各自的度之和，表示两个社区所有节点度之和。



****

**图4 微博事件追踪及社区发现可视化**

**表1事件主题抽取效果表**

|  |  |  |
| --- | --- | --- |
| 日期 | 舆论热点 | 主题抽取结果 |
| 2020.01.20 | 钟南山：新型冠状病毒存在人传人 | 疫情 新型冠状病毒 肺炎 钟南山 高级别 专家组 武汉 病例 人传人 感染 院士 防控 专家 确诊 成立 病毒 卫健委 预防 |
| 2020.01.23 | 武汉宣布封城 | 离汉 通道 地铁 关闭 客运 指挥部 运营 通告 火车站 离开 公交 市民 旅客 车站 理解 武汉市 人民日报 机场 全市 原因 感谢 |
| 春节档7部新片全部改档，另行上映。 | 撤档 姜子牙 电影 |
| 2020.01.30 | 网传红十字会售卖山东寿光350吨蔬菜 | 寿光 分配 接收 红十字会 红十字 捐赠 现金 售卖 物资 蔬菜 武汉市 网络 急需 指挥部 超市 医疗 公告 武汉 收到 援助 回应 参与 销售 关注 |
| 2020.01.31 | 湖北省红十字会物资使用情况引发质疑 | 仁爱医院 红十字会 湖北省 核查 获赠 莆田系 捐赠物资 捐赠 武汉市 回应 门诊 红会 口罩 调拨 标准 定点 |
| 上海药物所，武汉病毒所：双黄连口服液可抑制新型冠状病毒  世卫组织将新型冠状病毒列为国际关注的突发公共卫生事件 | 药物 口服液 病毒 双黄连 联合 上海 中心 启动 中心  国际 突发 世卫 中国 事件 加油 武汉加油 旅行 朱一龙 关注 人民日报 世界卫生组织 公共卫生 组织 列为 |
| 2020.02.05 | 大理市政府征用重庆抗疫物资 | 大理市 征用 暂扣 大理 重庆 交通 重庆市 予以 回应 管理局 黄石 理解 云南 |
| 2020.02.20 | 武汉新增病例高于湖北全省引关注 | 新增 病例 核减 全省 湖北 数据 人数 武汉 高于 确诊 |