

DOI: 10.13718/j.cnki.xdzk.2020.12.002

# 基于复杂网络的全球新闻流动模式分析

莫小梅<sup>1</sup>, 沈浩<sup>2</sup>, 俞定国<sup>1</sup>

1. 浙江传媒学院 媒体工程学院, 杭州 310018; 2. 中国传媒大学 新闻学院, 北京 100024

**摘要:** 因特网的出现大大拓展了全球范围内新闻报道和传播的广度和深度, 并通过在线报道的形式使世界各国前所未有地紧密联系在一起。本文根据全球整合新闻数据库(GDELDT)中的 GKG 数据集的特点, 建立国家间新闻流动网络模型; 利用复杂网络理论中平均度、图密度、平均聚类系数和平均路径长度等参数对该模型拓扑特性及边权分布特点进行分析; 并提出一种基于改进的互信息节点重要性评估方法。对 2019 年 4 月 8 日—12 日期间“一带一路”国家间新闻流动为例进行实证分析的结果表明: 国家间新闻流动网络具有小世界特性; 其网络的权重分布具有幂律分布的特性; 基于改进的互信息节点重要性评估方法可以有效挖掘出前 10 个重要节点国家。最后以节点重要性评估排名第一的中国为例进行了单节点分析。本文提出的分析方法为大数据时代的国家间新闻流动模式探索提供一个新的视角, 同时也为我国的国际传播策略和“一带一路”建议的推动提供了一些参考。

**关键词:** 全球整合新闻数据库; 复杂网络; 一带一路; 新闻流动模式; 幂律分布

**中图分类号:** TP393

**文献标志码:** A

**文章编号:** 1673-9868(2020)12-0015-10

随着通信和信息技术的飞速发展, 特别是因特网的出现, 极大地方便了信息的交流和传播, 从而大大拓展了全球范围内新闻报道和传播的广度和深度, 将全世界前所未有地紧密联系在一起, 形成了一张巨大而复杂的国际关系网络<sup>[1]</sup>。党的十八大以来, 习近平总书记反复强调: 要从战略高度进一步加快我国传播能力, 特别是国际传播能力建设。党的十九大报告明确地对中国国际传播能力建设提出了新要求, 向媒体行业布置了新任务<sup>[2]</sup>。如何从海量的全球新闻报道中全面、及时、准确地分析国际关系及其变化特征, 对推动“一带一路”建设、进一步加快中国国际传播能力建设均具有重要参考价值。

当前, 大数据领域新技术的不断涌现和迅速发展使得海量数据的获取、存储和计算成为可能, 其中全球整合新闻数据库(The Global Database of Events, Language, and Tone, GDELDT)是一个开源、大规模的全球在线新闻和事件元数据库。GDELDT 是一个涵盖全球各个国家共计 100 多种语言的新闻媒体数据库, 其中 GDELDT 2.0 实时翻译全球 65 种语言, 不仅对新闻中的人物、组织、事件等信息进行提取, 还提供了新闻的来源、情感态度、以及所涉及对象的国家编码等地理位置信息。GDELDT 数据具有很高的时效性, 每 15 分钟实时更新<sup>[3]</sup>。该数据库的这些特点, 为我们从空间和时间等角度探索全球新闻流动模式的特点及规律提供了可能。

复杂网络理论是由图论发展而来, 用来解决现实世界中网络问题而形成的一门抽象学科。复杂网络是由大量的节点和边组成的网络系统, 其特点包括自组织、自相似、吸引子、小世界、无标度中部分或全部性质<sup>[4]</sup>。任何复杂网络都可以认为是对现实世界中存在于个体之间实际关系的抽象, 例如社交网络<sup>[5]</sup>、交通网络<sup>[6]</sup>、信息网(计算机病毒<sup>[7]</sup>、页面链接<sup>[8]</sup>、微信<sup>[9]</sup>及新闻资讯<sup>[10]</sup>等)都是常见的复杂网络。利用复杂网络的理论和技术对实体之间的关系进行建模构建网络, 可以很方便地对其中蕴含的关系进行分析, 而 GDELDT 数据库包含着新闻报道的来源和被报道的对象等实体信息, 可以使用复杂网络的理论和方法很好

收稿日期: 2020-10-15

基金项目: 浙江省重点研发计划项目(2019C03138)。

作者简介: 莫小梅(1978—), 女, 硕士, 讲师, 主要从事复杂网络及应用、时空数据挖掘与大数据分析等方面的研究。

地研究不同国家实体之间的新闻流动关系。

近年来国内外开始出现一些基于 GDELT 数据集构建国家交互网络,利用复杂网络的理论与方法探索网络特征的研究,如秦昆等<sup>[11]</sup>研究不同的国家在同一个文档中出现的交互关系网络;方鹏<sup>[12]</sup>研究了南海仲裁案对跨国关系网络的影响及动态演变;也出现了从时空可视化的角度呈现一段时期内全球媒体关注的“国家共现社区”网络<sup>[13]</sup>。但目前基于 GDELT 对国家间关系网络的构建研究主要以国家的共现分析为主,即不同的国家是否在同一篇新闻报道中出现以及出现的次数等,体现在网络分析中为简单的由无向边组成的网络,未能分析国家间新闻流动模式的数量及方向性。

本文基于 GDELT 中的 GKG 数据集的特点,建立国家间新闻流动网络模型;利用复杂网络理论中平均度、图密度、平均聚类系数和平均路径长度等参数对该模型拓扑特性及边权分布特点进行分析;并提出一种基于改进的互信息节点重要性评估方法。同时以 2019 年 4 月 8 日—12 日期间“一带一路”国家间的新闻流动模式为例进行实证分析。本文提出的分析方法可以为大数据时代的国家间新闻流动模式研究和探索提供一个新的视角,为“一带一路”建设的新闻宣传工作提供决策支持,进而为加快国际传播能力建设提供技术支持与分析方法。

## 1 数据、网络模型与分析方法

### 1.1 GDELT 的 GKG 数据集介绍

GDELT 全球知识图(GKG) 2.1 版本发布于 2015 年 2 月,该版本超越了对物理事件的编目,使其不仅限于事件流表面的基础信息,而是更深入地关注新闻事件的详细信息,包括每篇报道的来源、报道中的人物或组织信息,地理信息和情感态度等。GKG 数据集使用现有的一些最复杂的命名实体和地理编码算法,计算一系列编码元数据,将数据中隐藏的信息进行解码。GKG 的一系列新功能极大地增强了数据的处理能力,为研究者提供了全新的分析机会。但由于 GKG 数据集的结构相对复杂,数据量大,到目前为止相关的研究不多,仍有很大的挖掘和应用缺口。本文研究的全球新闻流动模式,可从 GDELT 的 GKG 数据集中挖掘及分析该时间段内全球的政治经济类“硬新闻”报道的数量,以及每篇报道中的报道方、被报道方等相关数据,从而进一步筛选、重构及分析相关国家间新闻流动的网络模型,进而利用复杂网络模型理论进行模式分析。本文选取 SourceCommonName 和 Locations 这 2 个字段用于构建网络。其中 SourceCommonName 是一个表示文档源的标识符。对于带有网址的 Web 网页(SourceCollectionIdentifier 字段值为 1),此字段将包含页面来源的域名。由于每个域名对应一个 Web 中的媒体机构,从而可以按照域名注册的规则查找该媒体所对应的国家编码,即可通过进一步的域名查询确定每篇 Web 新闻报道的来源国。Locations 是指在本篇新闻报道中能找到的所有位置的列表,每个位置用“;”分隔,每个位置字段中包含许多子字段,用“#”分隔,本文用到其中的第 3 个子字段 LocationCountryCode(用于表示位置的国家编码)。

### 1.2 基于 GDELT 的国家间新闻流动网络构建

在 GKG 数据集中,国家间的报道关系可定义为:在同一条记录中,由 SourceCommonName 字段的域名通过域名查询的方法所对应的国家为报道国,由 Locations 字段解析出来的每一个地名所在国家均为被报道国。根据复杂网络构建的原理,可将报道国和被报道国均定义为网络中的基本节点,从报道国到每个被报道国定义一条有向的边,一条记录有可能对应一个由报道国为起点,每个被报道国为终点且权重均为 1 的多条有向边。如果在其他记录中发现已有的报道国和被报道国的有向边,则对已有的边进行加权操作,权值为在网络中出现的有向边(即特定的新闻流动)次数。

表 1 为特定时间段内 GKG 数据集中的 4 条示例数据。首先可根据域名查询方法,将 SourceCommonName 字段中的域名转换为国家名称,即为报道国;将 Locations 字段中的地名提取出相应的国家名称,即为被报道国,有几个符合条件的被报道国即生成几对报道关系。这里被报道国的条件是指:① 不考虑与报道国相同的情况,如表 1 第一条记录中的报道国为中国,被报道国为中国和新西兰,则只生成中国对新西兰的报道关系;② 不考虑研究范围以外的被报道国。如表 1 最后一条记录中报道国为意大利,被报道国为中国和日本,当研究对象为“一带一路”范围内的国家时,则只生成意大利对中国的报道关系。同理,SourceCommonName 也只选择“一带一路”相关国家的域名,并在生成报道关系的同时累计相应报道量对

应的权重. 如表 1 第一条记录和第三条记录都出现了中国对新西兰的报道关系, 累计次数为 2, 则相应的权重也为 2. 根据上述的国家间新闻流动网络构建方法, 可将表 1 中的 GKG 数据映射为表 2 定义的新闻流动映射表, 生成的有向加权复杂网络如图 1 所示.

表 1 GKG 数据示例

SourceCommonName(文档源)	Locations(地点)
jschina.com.cn	1 # China # CH # CH # 35 # 105 # CH; 1 # New Zealand # NZ # NZ # -42 # 174 # NZ
thestar.com.my	1 # Russia # RS # RS # 60 # 100 # RS; 1 # Poland # PL # PL # 52 # 20 # PL
china.org.cn	1 # Italy # IT # IT # 42.833333 # 12.833333 # IT; 1 # New Zealand # NZ # NZ # -42 # 174 # NZ
iphoneitalia.com	1 # China # CH # CH # 35 # 105 # CH; 1 # Japan # JA # JA # 36 # 138 # JA

表 2 国家间新闻流动映射表

报道国	被报道国	权重
中国	新西兰	2
马来西亚	俄罗斯	1
马来西亚	波兰	1
中国	意大利	1
意大利	中国	1

### 1.3 国家间新闻流动网络拓扑特征定义

在全球各国, 每天都有大量的新闻报道, 既有国内的时事新闻, 也有关注他国的国际时事报道. 根据国家间新闻流动网络的特点, 定义相应的复杂网络拓扑特征如下:

**定义 1 节点数:** 节点数  $N$  指国家间新闻流动网络中国家的数量.

**定义 2 输出节点数:** 输出节点数  $N_{out}$  指国家间新闻流动网络中报道国的数量.

**定义 3 输入节点数:** 输入节点数  $N_{in}$  指国家间新闻流动网络中被报道国的数量.

**定义 4 连边数:** 连边数  $M$  指国家间新闻流动网络中从报道国到被报道国之间有向连边的数量.

**定义 5 平均度:** 平均度  $k$  代表在某时间段内一个国家平均与多少个国家有过新闻流动的关系, 计算公式为

$$k = M/N \tag{1}$$

其中,  $M$  代表连边数,  $N$  代表国家(节点)数.

**定义 6 图密度:** 图密度  $D$  是指国家间新闻流动网络中国家间新闻报道连接的紧密程度. 计算公式为

$$D = M/N(N - 1) \tag{2}$$

**定义 7 聚类系数:** 聚类系数  $C_i$  是指所有与节点  $i$  相连的节点之间实际相连的边数占这些点可能的最大连边数目的比例, 聚类系数反映网络节点的聚类情况, 计算公式为

$$C_i = \frac{2E_i}{k_i(k_i - 1)} \tag{3}$$

其中,  $k_i$  为与节点  $i$  连接的其他节点的个数;  $E_i$  为这  $k_i$  个节点间相互连接的边数(两点之间的双向边和单向边都按 1 条计数).

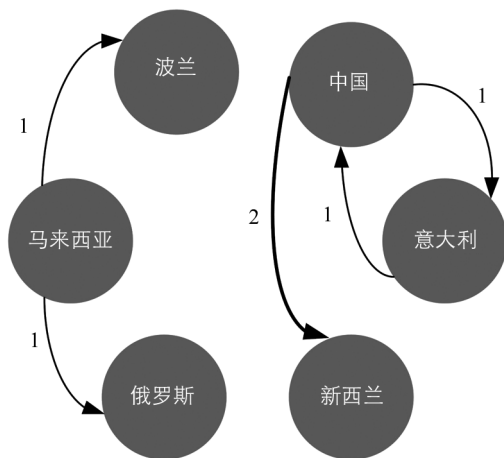


图 1 国家间新闻流动网络示例

**定义 8** 平均聚类系数: 平均聚类系数  $C$  定义为网络中所有节点聚类系数的平均值, 计算公式为

$$C = \frac{1}{N} \sum_{i=1}^N C_i \quad (4)$$

其中,  $C$  取 1 说明网络中的所有节点都相连.

**定义 9** 平均路径长度: 相互连接的 2 个节点  $i$  和  $j$  之间边数最少的路径所包含的边数即为这 2 个节点间的距离  $d_{ij}$ . 取尽网络中任意两节点的组合, 网络的平均路径长度  $L$  定义为所有组合之间距离的平均值.

$$L = \frac{1}{N(N-1)} \sum_{i,j} d_{ij} \quad (5)$$

#### 1.4 基于改进的互信息节点重要性评估

在“一带一路”国家间新闻流动网络中, 如何评价各国在网络中的作用和地位? 如何综合衡量各国在网络中自身的报道量、被报道热度, 以及由新闻流动关系连接的国家数量和相关联国家的特点与连接能力? 这些都涉及复杂网络中的节点重要性问题. 目前, 复杂网络的节点重要性评估方面较成熟的方法主要集中在无向无权网络<sup>[14]</sup>, 少量适用于无向加权网络<sup>[15-16]</sup>. 近年来, 逐渐出现了关于有向加权网络的节点重要性的研究, 如王雨等<sup>[17]</sup>提出的基于多重影响力矩阵的评估方法, 刘臣等<sup>[18]</sup>提出的基于出度、入度和权重的节点重要性指标: 交叉度(cw)指标, 及基于该指标的 cw 壳分解方法. 前者构建了 3 个影响力矩阵, 考虑的角度较多, 方法有一定的复杂性, 且不适用于入度为 0 的节点; 后者算法较为简单, 但更适用于对网络中信息的传播和扩散效果的研究. 马润年等<sup>[19]</sup>提出的基于互信息的通信网络节点重要性度量方法由于考虑到了节点自身的强度、相邻节点的强度、节点的连边权重和权重比例系数等, 为我们提供了一个较好的参考方案. 此外, SHANG K K 等<sup>[20]</sup>通过对路透社舆论关键词即重要节点的时变性分析, 洞悉了美国 911 恐怖事件上舆论的发酵过程, 基于链路预测技术预测了舆论关键词的走向. 综合以上研究成果, 同时基于 GDELT 数据库中 GKG 数据集构建的国家间新闻流动网络的自身特点, 我们对参数中的出度部分改用节点对应新闻报道国的实际报道总次数来代替, 并采用这种改进的互信息节点重要性度量方法对具体的新闻流动网络进行了节点重要性评估.

##### 1.4.1 基于互信息的有向加权网络节点重要性度量方法

首先将有向加权网络抽象为图  $G = \{V, E, W\}$ , 其中  $V = \{v_1, v_2, \dots, v_N\}$  为节点集合,  $E = \{e_1, e_2, \dots, e_M\}$  为有向边集合,  $W = \{w_{e_1}, w_{e_2}, \dots, w_{e_M}\}$  为有向边的权重集合.  $(i, j) \in E$  表示节点  $i$  到节点  $j$  的一条有向边,  $w(i, j) \in W$  表示有向边  $(i, j)$  的权值, 类似的,  $w(j, i)$  表示有向边  $(j, i)$  的权值, 其中  $w(i, j)$  并不一定等于  $w(j, i)$ . 节点强度可分为入强度和出强度,  $S_{in}(i)$  表示节点  $i$  的入强度:  $S_{in}(i) = \sum_{k \in V_{in}(i)} w(k, i)$ , 其中  $V_{in}(i)$  表示指向节点  $i$  的所有节点集合;  $S_{out}(i)$  表示节点  $i$  的出强度:  $S_{out}(i) = \sum_{k \in V_{out}(i)} w(i, k)$ , 其中  $V_{out}(i)$  表示节点  $i$  所指向的所有节点集合, 入强度和出强度之和即为节点  $i$  的总强度:  $S(i) = S_{in}(i) + S_{out}(i)$ .

对于有向加权网络, 节点之间的连边具有方向和权重两重特性, 有向加权边表示该边的 2 个端节点以一定数量进行单向报道. 此时, 同一节点的每条边的概率分布将有所不同, 节点的边的概率将与该节点的出入强度、该边的方向和权值有关, 因为一个节点有可能同时既输出信息又接收信息, 这里定义节点的出边概率和入边概率, 如有从节点  $i$  指向节点  $j$  的有向边  $(i, j)$ , 则定义节点  $i$  的出边  $(i, j)$  的概率为<sup>[19]</sup>

$$p_{i \rightarrow j} = \frac{w(i, j)}{S_{out}(i)} \quad (6)$$

**定义 10**<sup>[19]</sup> 设节点  $i$  到节点  $j$  的互信息为  $I(i, j)$ , 定义如下

$$I(i, j) = \begin{cases} \ln \frac{1}{p_{i \rightarrow j}} - \ln \frac{1}{p_{j \leftarrow i}} \\ 0 \end{cases}$$

$$\begin{cases} \ln \frac{S_{out}(i)}{w(i, j)} - \ln \frac{S_{in}(j)}{w(i, j)} \\ 0 \end{cases} \\
 \begin{cases} \ln \frac{S_{out}(i)}{S_{in}(j)} & i \text{ 到 } j \text{ 有连边} \\ 0 & \text{其他} \end{cases}
 \end{cases} \quad (7)$$

**定义 11**<sup>[19]</sup> 节点  $i$  的信息量是节点  $i$  到其指向的所有节点的互信息之和减去所有指向  $i$  的节点到节点  $i$  的互信息之和, 即

$$I(i) = \sum_{j \in V_{out}(i)} I(i, j) - \sum_{k \in V_{in}(i)} I(k, i) \quad (8)$$

其中,  $V_{out}(i)$  表示节点  $i$  所指向的所有节点集合,  $V_{in}(i)$  表示指向节点  $i$  的所有节点集合. 根据式(7)和(8)可证明整个网络中的信息量之和  $\sum_{i=1}^N I(i)$  为 0. 计算出所有节点的信息量之后按照从大到小进行排序, 信息量越大的节点重要性越强.

#### 1.4.2 基于改进的互信息节点重要性评估

对基于 GDELT 中 GKG 数据集构建的国家间新闻流动网络而言, 由于它统计了一篇新闻报道中的多个位置, 其中每个符合条件位置的国家都与同一个报道国产生连边, 从而导致对一个报道国的报道次数的重复计算, 因此本文改用统计报道国实际报道的总次数(即在符合条件的记录中, SourceCommonName 字段被解析为该国的新闻报道总数)代替节点的出度作为计算互信息和信息量的依据.

## 2 实验与分析

由于本文关注的是一段时期内“一带一路”国家间的新闻流动模式, 而与中国签订“一带一路”合作文件的国家数量可能会随着时间增加, 我们以 2019 年 4 月 8 日-12 日这一周(5 个工作日)为研究的时间段, 以截至 2019 年 4 月 12 日前与中国签订“一带一路”合作文件的 126 个国家和中国(共 127 个国家)作为国家间新闻流动网络中的基本节点, 利用复杂网络的理论和方法进行新闻流动的网络模式分析: 从拓扑特征统计及边权分布特点分析网络的整体特性; 采用改进的互信息算法进行节点重要性评估; 并对以中国为例的单节点进行分析.

### 2.1 “一带一路”国家间新闻流动网络拓扑特征分析

本文首先以一天、一周 2 种时间尺度对“一带一路”国家间新闻流动网络的基本特性进行分析. 从表 3 可以看到, 2 种时间尺度中, 输入节点、输出节点及连边数都比较大, 尤其是网络输入节点数达到了当时研究范围内“一带一路”的国家总数 127, 说明在这段时间内“一带一路”各国都有被其他国家报道, 而一天及一周的输出节点数都比较大, 说明大部分国家的媒体都有对其他国家的报道. 平均度  $k$  代表在某时间段内一个国家平均与多少个国家有过新闻流动的关系, 可见 4 月 8 日这一天每个国家平均与 30 多个国家有新闻流动的关系, 一周内虽然新闻流动的平均国家数增加到 50 多个, 但还不能达到成倍的增加, 且数量与国家总数量 127 个相比还是比较低的报道比例, 说明在短时期内国家间新闻流动的平均数量不多, 即有一个相对小范围互相关注的“朋友圈”. 这也可以从图密度  $D$  看得出来, 图密度代表国家间连接的紧密程度, 在 4 月 8 日这一天里图密度不到 0.3, 在一周的尺度里也仅能达到 0.454. 此外, 在 2 种时间尺度的网络中, 平均聚类系数相对较大, 平均路径长度都较小, 反映了国家间新闻流动网络具有小世界特性.

表 3 “一带一路”国家间新闻流动网络拓扑特征统计结果

时间尺度	$N_{out}$	$N_{in}$	$M$	$k$	$D$	$C$	$L$
1 天	116	127	4 502	35.449	0.281	0.573	1.758
1 周	120	127	7 258	57.15	0.454	0.654	1.533

注: 1 天指 2019 年 4 月 8 日; 1 周指 2019 年 4 月 8 日-12 日;  $N_{out}$  代表输出节点数;  $N_{in}$  代表输入节点数;  $M$  代表连边数;  $k$  代表平均度;  $D$  代表图密度;  $C$  代表平均聚类系数;  $L$  代表平均路径长度.

## 2.2 “一带一路”国家间新闻流动网络边权分布特点

“一带一路”国家间新闻流动模式网络作为一种加权有向的网络,其中一个重要的特点是在一段时间内不同新闻流动关系所对应连边的权重会有很大的差别,如 2019 年 4 月 8 日-12 日这 5 天内乌克兰对俄罗斯的报道量高达 7 762 条,俄罗斯对乌克兰的报道达 7 406 条,两者加起来占总报道量的近 4%。此外,大量的报道关系中边权只有 1,2,3,在 7 258 条有向边代表的国家间报道关系中,权重为 1 的边数有 1 105 条,占总边数的 15.2%,权重为 2 的边数占 11.4%,权重为 3 的占 7.5%,权重为 7 以内的边数达到 51.2%。图 2 给出“一带一路”国家间新闻流动网络权重分布图,其中横坐标为已有的边权值,纵坐标为该权值的边数占总边数的百分比。

图 3 为权重分布的双对数图,从双对数图看出网络的权重分布具有幂律分布的特性。可见在“一带一路”各国中,少数国家间的报道量大,大多数国家间的报道量少。如权重最大的前 20 对国家间流动关系在 7 258 个关系中,以不到 0.3%的关系比例,其报道总量(边权的和)却占了总报道量的 16%。这意味着“一带一路”国际间新闻流动模式网络是一个中心性较强的复杂网络,大部分国家之间并不存在明显的新闻流动,关注少量的国家对就能够大体反映国家间新闻流动模式网络的主要面貌。权重排名靠前的连边对应的节点国家就报道能力和被关注的热度这 2 个角度来看都是新闻流动模式网络中的重要节点。

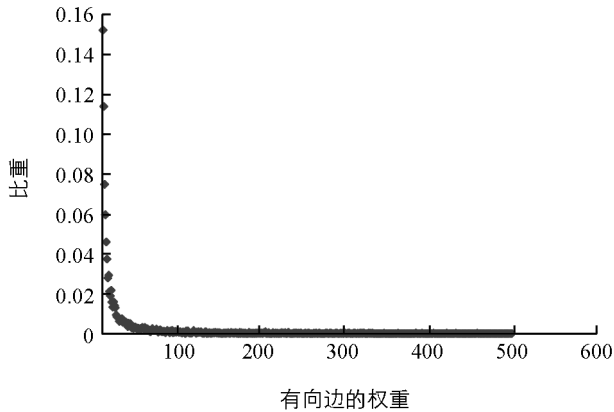


图 2 “一带一路”国家间新闻流动网络权重分布

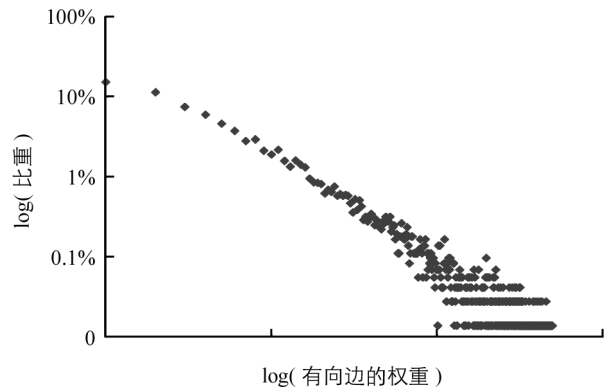


图 3 “一带一路”国家间新闻流动网络权重分布双对数图

图 4 显示了 2019 年 4 月 8 日-12 日这一周内权重排名前 20 的新闻流动关系,其中连边的粗细与权重成正比,连边旁用数字标出实际的权重值。可以看出,乌克兰和俄罗斯的相互新闻流动量是最大的,且都在 7 400 条以上,明显超过了其他各国间的新闻流动。而现实中,乌克兰和俄罗斯两国有着长期特殊的历史渊源,从而形成了双方媒体大量的相互关注和报道新闻流。在其余的新闻流动关系中,中国和俄罗斯出现的次数最多,中国在报道国和被报道国中分别出现了 6 次和 4 次,俄罗斯则分别出现了 3 次和 4 次。这些边权分布的特点也为我们评估国家间新闻流动网络中的重要节点提供了一定的启示。

## 2.3 基于改进的互信息的国家间新闻流动网络节点重要性评估

图 5 显示了使用本文改进的互信息算法计算得到的全部节点的重要性名次与信息量的关系。其中横坐标为节点的重要性名次,纵坐标为相关排名对应的信息量。从图 5 可以看出,重要性排名前 10 的节点信息量区分度最大:不仅信息量明显大于 10 名以后的节点,且相互之间的数值变化也较大;而 10 名以后的节点信息量及变化率都明显变小。由此得到的结论是在“一带一路”国家间新闻流动网络中通过改进的互信息算法可以有效挖掘出网络中重要性最突出的前 10 个节点国家。

表 4 和表 5 显示了分别使用文献 19 的互信息算法及本文提出的改进的互信息算法评估的“一带一路”国家间新闻流动网络中排名前 10 的节点及其信息量,其中加粗的国名为使用 2 种方法评估得出的不同结果。

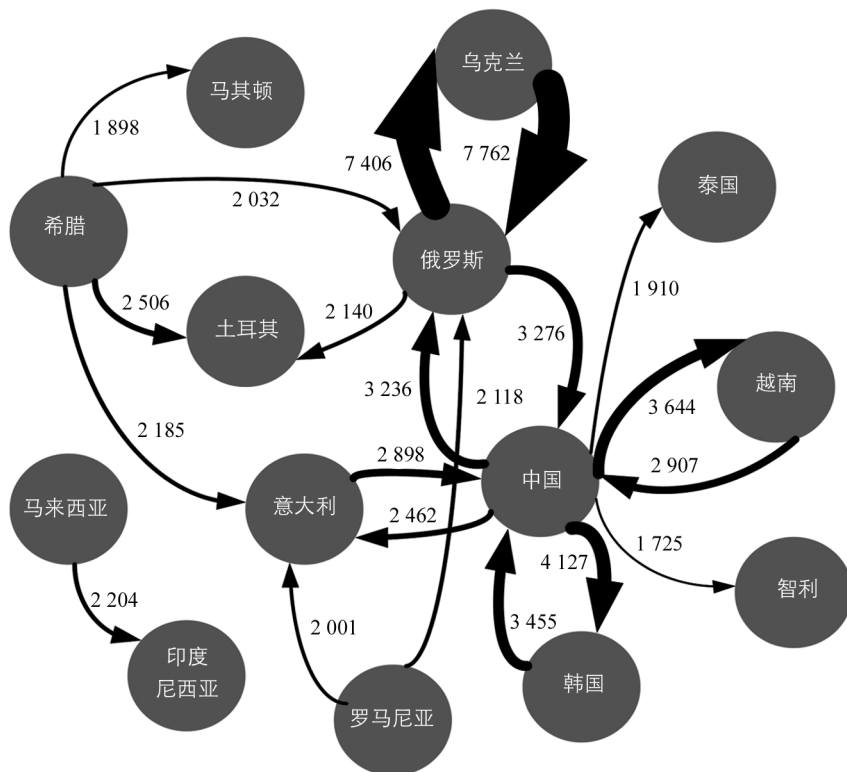


图 4 “一带一路”国家间新闻流动关系前 20

表 4 2 种互信息方法对应节点重要性排名前 1~5 的国家及信息量

	No. 1	No. 2	No. 3	No. 4	No. 5
文献 19 方法	中国	俄罗斯	意大利	土耳其	乌克兰
信息量	850.11	790.67	624.11	477.32	452.05
本文改进方法	中国	俄罗斯	意大利	土耳其	乌克兰
信息量	855.99	786.64	631.24	463.12	477.85

表 5 2 种互信息方法对应节点重要性排名前 6~10 的国家及信息量

	No. 6	No. 7	No. 8	No. 9	No. 10
文献 19 方法	希腊	埃及	韩国	罗马尼亚	波兰
信息量	416.20	325.06	319.24	311.39	280.76
本文改进方法	希腊	韩国	埃及	罗马尼亚	波兰
信息量	378.21	333.00	311.96	295.04	265.07

由表 4 和表 5 可见: 在节点重要性排名前 10 的国家中, 2 种方法的排名结果基本是一致的, 只有埃及和韩国在使用文献 19 算法的结果中排名分别为 7, 8, 使用本文改进的互信息算法则交换了排名顺序. 实际上, 韩国不论是在入度、出度、还是报道的篇数上都大于埃及, 从这些主要的指标来看重要性应在埃及之前, 而韩国的排名在埃及之前也更符合现实中人们对两国国际影响力的评估. 可见使用本文改进的互信息算法可以校正使用传统出度计算互信息和信息量产生的个别排名误差, 对节点重要性的评估结果也更符合实际的排名.

## 2.4 单节点分析

中国作为“一带一路”国家间新闻流动网络中互信息量最大, 节点重要性评估中排名第一的国家, 尤其值得关注的问题包括: 本国媒体对“一带一路”其他国家的报道情况, 以及其他国家媒体对我国的报道情

况。为此，我们对所建立的复杂网络模型中以“中国”这一单节点为起点和终点的连边及相关节点进行了统计和分析，并给出了相应的解答。

图 6 显示了在 2019 年 4 月 8 日-12 日期间中国对“一带一路”各国报道量的排名及分布，其中横坐标表示以中国报道的国家按报道量进行排名的名次，纵坐标表示相关名次的报道量占比。从图 6 可以看出，中国对其他国家的报道大量集中在排名靠前的少数国家，其中对前 10 国的报道合计占比达到了 51%，而对前 20 国的报道量合计占比达到了 70%。

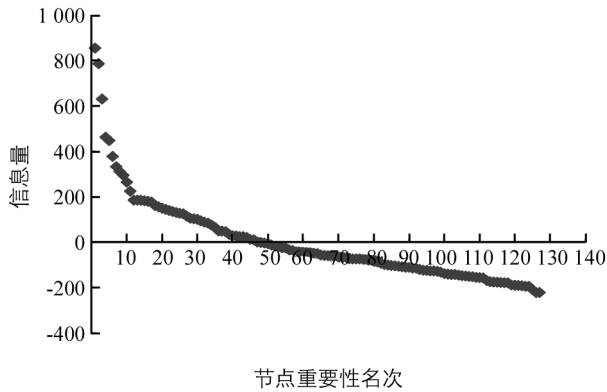


图 5 节点重要性名次及信息量

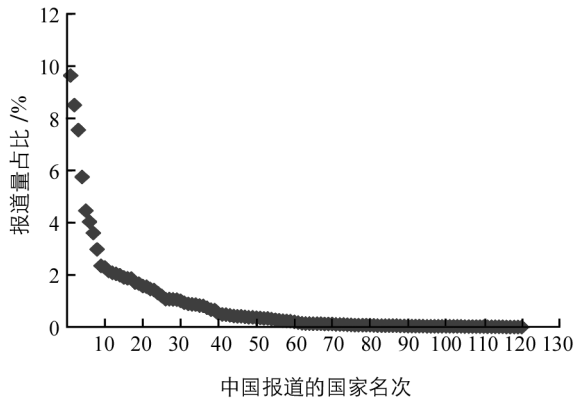


图 6 中国报道的国家名次及报道量占比

图 7 显示了在中国报道的各国新闻中报道量排名前 20 的国家，其中纵坐标表示中国对该国的报道数量在中国对各国报道总量中的比重。可以看出，在 2019 年 4 月 8 日-12 日这一周内，中国报道量较多的国家为韩国、越南、俄罗斯、意大利等，占比都在 5% 以上。

图 8 显示了对中国的新闻报道量排名前 20 的国家，其中纵坐标表示对中国的报道在本国总报道量中的比重。可见对中国的报道量最多的 20 国中有 13 国对中国的报道占比都在 9% 以上，其中，韩国、越南、新加坡对中国的报道占比分别达到了 34%，41%，23%。可见中国在“一带一路”各国对外的报道中占到了较高的比例，也说明我国在“一带一路”各国媒体中得到了较多的关注。

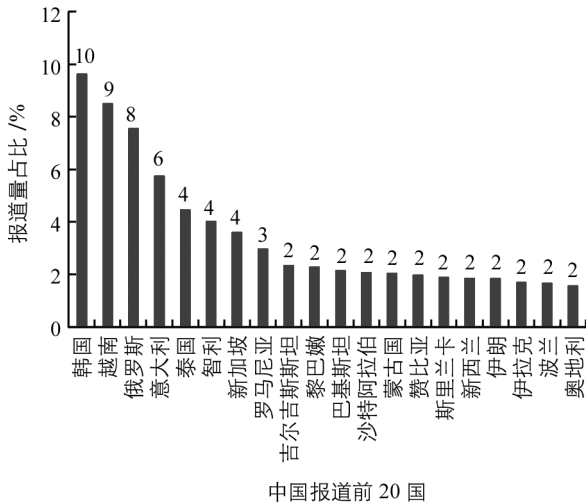


图 7 中国报道占比前 20 国(2019 年 4 月 8-12 日)

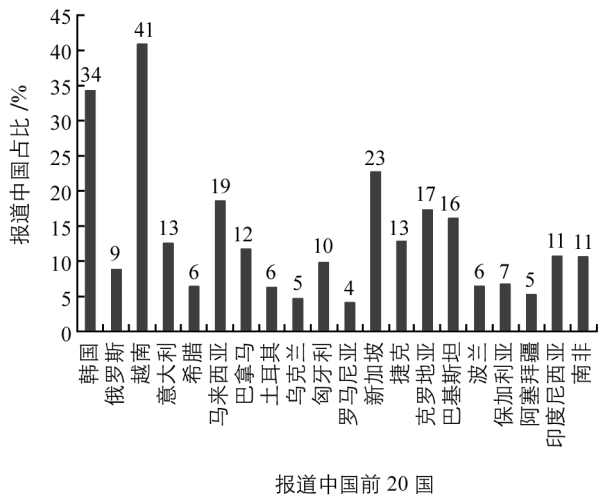


图 8 各国报道中国的数量占比(2019 年 4 月 8-12 日)

### 3 结论与展望

本文基于 GDELT 中 GKG 数据集的特点，提出了一种利用复杂网络的理论和方法构建国家间新闻流动网络的方法，并以特定时间段内“一带一路”国家间的新闻流动为例对新闻的流动模式进行实证分析：建



立了“一带一路”国家间新闻流动的网络模型,分析了网络的拓扑特征及边权分布特点;采用改进的互信息节点重要性评估算法进行重要节点挖掘,并以节点重要性评估排名第一的中国为例进行了单节点分析,主要结论如下:

1) 基于 GDELT 中的 GKG 全球新闻数据集构建国家间新闻流动网络,通过对“一带一路”国家间新闻流动网络的拓扑特征进行统计和分析,发现在短时间内国家间新闻流动的平均数量不多,即有一个相对小范围互相关注的“朋友圈”。另一方面,网络平均聚类系数相对较大,平均路径长度都较小,反映了国家间新闻流动网络具有小世界特性。

2) 对“一带一路”国家间新闻流动网络的边权分布特点进行统计和分析,发现网络的权重分布具有幂律分布的特性,即少数国家间的新闻流动量大,大多数国家间的新闻流动量小。

3) 根据 GKG 数据集的特点,采用报道国实际报道的总次数代替出度作为改进的互信息方法对“一带一路”国家间新闻流动网络进行节点重要性评估,可以有效挖掘出网络中信息量最大,区分度最高的前 10 个重要节点国家,且可以校正使用传统出度计算互信息和信息量产生的个别排名误差。

4) 对 2019 年 4 月 8 日-12 日期间“一带一路”国家间新闻流动网络,以中国为例进行单节点分析,发现中国对其他国家的报道大量集中在排名靠前的少数国家,且中国在“一带一路”各国对外的报道中占到了较高的比例。

本文应用复杂网络的理论与方法,提出了一种基于全球整合新闻数据库的国家间新闻流动复杂网络建模的方法,并分析了“一带一路”国家间新闻流动的关系模式,从而为我国的国际传播策略和“一带一路”建议的推动提供了一些参考。进一步的研究包括:① 研究国家间新闻流动网络的社区挖掘方法;② 与 GDELT 中的事件数据库联合分析报道中的事件特性;③ 研究网络化数据挖掘的时空演变与预测分析方法。

## 参考文献:

- [1] 王 骏. 基于文本挖掘的国际关系网络研究 [D]. 北京:北京邮电大学,2013.
- [2] 李青昱. 新时代中国国际传播能力建设思考 [J]. 中国广播电视学刊,2018(10): 107-110.
- [3] The GDELT project. GDELT 2.0: Our Global World in Realtime [EB/OL]. (2015-02-19) [2020-09-27]. <https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/>.
- [4] 孙玺菁,司守奎. 复杂网络算法与应用 [M]. 北京:国防工业出版社,2015.
- [5] 艾 均,矫雅楠,苏 湛. 基于观点邻域状态变化的复杂网络传播与干扰模型 [J]. 软件导刊,2020,19(8): 186-191.
- [6] 许 葭,宋守信,翟怀远,等. 基于复杂网络的北京地铁网络化发展分析 [J]. 都市快轨交通,2020,33(5): 88-93.
- [7] 李婵婵. 几类复杂社团网络上的病毒传播建模及免疫策略研究 [D]. 南京:南京邮电大学,2018.
- [8] 韩忠明,李胜男,郑晨焯,等. 基于动态网络表示的链接预测 [J]. 物理学报,2020,69(16): 332-345.
- [9] 朱宏森,张生太,闫 辛. 微信群中隐性知识传播模型研究 [J]. 科研管理,2019,40(2): 106-115.
- [10] 王思佳. 新闻 APP 的信息传播溯源技术的研究与实现 [D]. 北京:北京邮电大学,2019.
- [11] 秦 昆,罗 萍,姚博睿. GDELT 数据网络化挖掘与国际关系分析 [J]. 地球信息科学学报,2019,21(1): 14-24.
- [12] 方 鹏. 基于 Django 的海量媒体数据分析平台的设计与实现 [D]. 南宁:广西大学,2017.
- [13] The GDELT project. Mapping Media Geographic Networks: The News Co-occurrence Globe [EB/OL]. (2015-06-01) [2020-09-27]. <https://blog.gdeltproject.org/mapping-media-geographic-networks-the-news-co-occurrence-globe/>.
- [14] MARTÍN GONZÁLEZ A M, DALSGAARD B, OLESEN J M. Centrality Measures and the Importance of Generalist Species in Pollination Networks [J]. Ecological Complexity, 2010, 7(1): 36-43.
- [15] MELLO D A A, WALDMAN H, QUITÉRIO G S. Interval Availability Estimation for Protected Connections in Optical Networks [J]. Computer Networks, 2011, 55(1): 193-204.
- [16] BRANDES U. On Variants of Shortest-path Betweenness Centrality and Their Generic Computation [J]. Social Net-

works, 2008, 30(2): 136-145.

- [17] 王 雨, 郭进利. 基于多重影响力矩阵的有向加权网络节点重要性评估方法 [J]. 物理学报, 2017, 66(5): 19-30.
- [18] 刘 臣, 李丹丹, 韩 林, 等. 加权有向网络中心节点识别的分解算法研究 [J]. 计算机应用研究, 2019, 36(1): 37-40.
- [19] 马润年, 王 班, 王 刚, 等. 基于互信息的通信网络节点重要性度量方法 [J]. 电子学报, 2017, 45(3): 747-752.
- [20] SHANG K K, YAN W S, SMALL M. Evolving Networks-Using Past Structure to Predict the Future [J]. Physica A: Statistical Mechanics and its Applications, 2016, 455: 120-135.

## Analysis of Global News Flow Patterns Based on Complex Networks

MO Xiao-mei<sup>1</sup>, SHEN Hao<sup>2</sup>, YU Ding-guo<sup>1</sup>

1. College of Media Engineering, Communication University of Zhejiang, Hangzhou 310018, China;

2. School of Journalism, Communication University of China, Beijing 100024, China

**Abstract:** The advent of the Internet has greatly expanded the breadth and depth of news coverage and dissemination on a global scale, and online reporting has brought the world closer together than ever before. Based on the characteristics of the GKG dataset in the Global Database of Events, Language and Tone (GDELT), this paper established a news flow network model among countries. The topological characteristics and edge-weight distribution characteristics of the model are analyzed using the parameters of the complex network theory such as average degree, graph density, average clustering coefficient and average path length; and a node importance assessment method based on improved mutual information is proposed. The results of an empirical analysis of the example of news flows between the B&R countries during April 8-12, 2019 show that the network of news flows between countries has small-world characteristics; and the weight distribution of its network is characterized by a power-law distribution. Based on the improved mutual information node importance assessment method, the top 10 countries of node importance are effectively uncovered. Finally, China, which ranks the first in node importance evaluation, is taken as an example for single node analysis. The analysis method proposed in this paper provides a new perspective for the exploration of news flow pattern among countries in the era of big data, and also provides some references for China's international communication strategy and the promotion of B&R suggestions.

**Key words:** The Global Database of Events, Language, and Tone(GDELT); complex network; Belt and Road (B&R); news flow model; power-law distribution

责任编辑 崔玉洁