한국언론학보
2023년 67권 6호 189~222쪽
https://doi.org/10.20879/kjjcs.2023.67.6.006
http://www.comm.or.kr

편향의 위상학

1650만 건 기사의 뉴스 정보원 연결망 분석을 통해 파악한 인용 방식의 보편적 분포로서 두터운 꼬리 분포

박대민 선문대학교 미디어커뮤니케이션학부 조교수

Topology of Media Bias*

Fat-Tailed Distribution as Universal Distribution of Quotation by Analyzing News Source Networks with 16.5 Million Articles

Daemin Park**
(Assistant Professor, School of Media and Communication, Sunmoon University)

Many studies on media bias point to a heavy reliance on official sources, such as governments and corporations, regardless of the media outlets, coverage, topics, or periods. If source bias persists despite numerous efforts to mitigate it, it may be due to its universal nature.
In this study, we shed light on the universality of source bias by structuring source citation patterns into news source networks and identifying topological features of news source networks.
We collected 16,538,668 articles from 19 national newspapers and economic journals in five categories (political, economic, social, cultural, and international) from 2000 to 2020, using BigKinds, the news big data system of Korea Press Foundation. We compared these articles by obtaining the degree exponents from 1,700 datasets, categorized by year, category, and media outlet.
Our findings showed that the degree exponents converged to 1.2. This implies that when the news source network is sufficiently mature, it exhibits a universal fat-tailed distribution across period, media, and topic, meaning that a small number of sources are cited as highly significant, while many sources are cited as relatively unimportant.
This has several methodological, theoretical, and practical implications. First, methodologically, to validly analyze news big data, including news source networks, it is necessary to ensure that the distribution of source citations, when the pattern is represented as news source networks, has a fat-tailed distribution. A fat-tailed distribution means that at least the rankings of the top sources are

* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea(이 논문은 2021년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임): NRF-2021S1A5A806921311

** dmpark@sunmoon.ac.kr

robust enough that they are not significantly affected by missing or incorrect information. To be reliable, a source's ranking should be such that it is widely co-cited in multiple articles, allowing for the emergence of one centralized hub source and one large main cluster. The collected news big data should be large enough to include most of the important sources.
Second, news big data analysis centered on top sources is better suited for media system-level analysis than cross-media comparative research. Cross-media comparisons require analyzing the content of each source in the fat tail. The problem is that there are a lot of sources in the fat tail. Analyzing these sources is ultimately only possible through qualitative research that focuses on specific individuals or groups of sources, reading and reviewing their quotes.
Third, the convergence of news source networks to a fat-tailed distribution suggests that source bias is a socio-physically universal phenomenon: it is inevitable as long as deadlines exist under objectivist journalistic practices that rely on citations to establish factuality. However, the fat-tailed distribution suggests that the harm of source bias can be reduced by actively seeking out sources in the fat tail in addition to official sources in the center.

Keywords: Source Bias, News Source Network, Fat-tailed Distribution, Scale Free Network, Degree Exponent

국문초록

많은 정보원 편향 연구는 매체나 취재 범위, 주제, 시기 등을 막론하고 정부나 기업 등 공식 정보원에 대한 의존도가 높다고 지적한다. 편향을 개선하려는 많은 노력에도 불구하고 정보원 편향이 나타난다면, 이는 정보원 편향이 보편성을 가진 현상이기 때문일지도 모른다. 본 연구에서는 정보원 인용 패턴을 뉴스 정보원 연결망으로 구조화하고, 뉴스 정보원 연결망의 위상학적 특징을 파악함으로써 정보원 편향의 보편성을 밝히고자 하였다. 이를 위해 2000년부터 2020년까지, 19개 전국지와 경제지의 정치, 경제, 사회, 문화, 국제 등 5개 지면 기사 16,538,668건을 빅카인즈에서 수집하여 연도별, 지면별, 매체별로 1,700개 데이터세트의 연결정도 지수를 구해 비교했다.
분석 결과 연도나 매체 및 지면을 막론하고 연결정도 중앙성 값이 커질수록 연결정도 지수 값은 1.2에 수렴하는 것을 확인할 수 있었다. 이는 뉴스 정보원 연결망이 충분히 성장한 경우, 뉴스 정보원 연결망이 시기와 매체, 주제를 막론하고 보편적으로 두터운 꼬리 분포를 갖는다는 것을 의미한다. 즉 극소수의 정보원이 매우 중요하게 인용되는 한편, 중요도가 매우 낮은 정보원이 다수를 차지한다. 이는 방법론적, 이론적, 실천적으로 다음과 같은 점을 시사한다.
첫째, 방법론적으로 뉴스 정보원 연결망을 비롯한 뉴스 빅데이터 분석을 타당하게 수행하려면, 정보원 인용을 뉴스 정보원 연결망으로 나타냈을 때 그 분포가 두터운 꼬리 분포를 갖는지 확인할 필요가 있다. 뉴스 정보원 연결망이 두터운 꼬리 분포라는 것은 최소한 상위권 정보원의 순위는 결측이나 오측에 크게 영향을 받지 않을 정도로 강건함을 의미한다. 정보원의 순위를 신뢰할 수 있으려면, 정보원이 여러 기사에서 다양하게 공동 인용되면서 중심 정보원과 거대 구성집단이 출현할 수 있을 정도가 되어야 한다. 수집된 뉴

스 빅데이터가 중요한 정보원은 대부분 포함돼 있을 정도로 충분히 성장한 데이터여야 한다.
둘째, 상위권 정보원을 중심으로 하는 뉴스 빅데이터 분석은 매체 간 비교 연구보다는 언론계 수준의 분석에 적합하다. 매체 간 비교 연구에는 두터운 꼬리에 해당하는 정보원에 대한 분석이 요청된다. 문제는 두터운 꼬리 부분에는 동 순위에 수많은 정보원이 있다는 점이다. 이러한 정보원에 대한 분석은 결국 특정 정보원 개인 또는 집단에 주목하여 해당 정보원의 인용문을 읽어보고 검토하는 질적 연구를 통해서만 가능하다.
셋째, 뉴스 정보원 연결망이 두터운 꼬리 분포로 수렴하는 현상은 정보원의 편향이 사회물리학적으로 보편적인 현상임을 의미한다. 즉 정보원의 편향은 인용을 통해 사실성을 확보하는 객관주의 저널리즘 관행 아래 마감 시간이 존재하는 한 불가피하다. 그러나 두터운 꼬리 분포는 중심에 해당하는 공식 정보원 외에 두터운 꼬리에 해당하는 정보원을 적극 발굴함으로써 정보원 편향의 폐해를 줄일 수 있음을 시사한다.

핵심어 : 정보원 편향, 뉴스 정보원 연결망, 두터운 꼬리 분포, 척도 없는 연결망, 연결정도 지수

1. 문제제기

언론의 편향된 뉴스 보도, 즉 미디어 편향(media bias)에 대해서는 오래전부터 연구가 진행되어 왔다(Williams, 1975). 미디어 편향 연구는 방법론적으로 수작업 또는 자동화된 방식으로 양적인 내용분석(content analysis)이나 질적인 담론분석(discourse analysis)을 수행하여 미디어의 다양한 편향을 발견한다. 분석을 통해서는 언론이 편향을 통해 의제 설정(agenda-setting), 프레이밍(framing), 프라이밍(priming) 등의 측면에서 어떻게 영향을 주는지를 살펴본다. 이러한 연구들에 따르면 언론은 정보 수집이나 기사 작성, 편집 등 기사 생산 과정 전반에서 의도적이거나 의도치 않게 편향을 드러낸다(Hamborg, Donnay, & Gipp, 2019). 언론의 편향은 수용자의 정치적 의사 결정에 악영향을 끼치는가 하면, 언론 신뢰도 하락을 초래할 수도 있기 때문에 매우 중요한 문제다(Bernhardt, Krasa, & Polborn, 2008; Tsfati & Cohen, 2005).

정보원 편향은 미디어 편향의 핵심 문제 중 하나로 다뤄졌다. 언론은 정부나 기업 등의 정보원에 편향되어 이들에게 유리한 기사를 작성할 수도 있다(Bennett, 1990; Gilens & Hertzman, 2000). 오랫동안 많은 언론인과 언론학자들을 비롯한 사회 구성원들이 이러한 정보원 편향을 지양해야 한다고 지적해왔다. 정보원 편향을 규범적 차원에서 기자나 언론사의 윤리, 정부 규제 등을 통해 개선하려는 노력도 적지 않았다. 그러나 사회 각계의 오랜 노력에도 불구하고 국적, 시기, 매체, 주제를 막론하고 정보원 편향이 계속되어 왔다(Berkowitz & Beach, 1993; Brown, Bybee, Weardem, & Straughan, 1987; Gans, 1979). 이는 정보원 편향이 보편성을 가진 현상이기 때문일지도 모른다.

본 연구는 이러한 정보원 편향의 보편성을 계량적으로 확인하고자 한다. 이를 위해 우선 분석 대상을 일반화했다. 기존 연구는 특정 매체, 특정 시기, 특정 주제의 기사만을 대상으로 분석했다. 본 연구에서는 한국언론진흥재단의 뉴스 빅데이터 분석 시스템 빅카인즈(BigKinds)[1]에서 수집한 21년치(2000-2020) 19개 매체(전국지, 경제지), 5개 지면(정치, 경제, 사회, 문화, 국제)의 1,650만여 건 기사에 등장한 정보원 약 1,100만 명[2]의 인용 패턴을 분석한다.

많은 양의 기사와 정보원 및 인용 패턴을 분석하기 위해 본 연구에서는 자동화된 내용분석

1) https://www.bigkinds.or.kr/

2) 같은 연도의 같은 매체의 같은 지면의 정보원 중복은 제거하고, 다른 연도 간, 다른 지면 간, 다른 매체 간 중복은 포함한 수치다.

방법인 뉴스 빅데이터 분석을 사용했다(박대민, 2013, 2014). 다만 기존 뉴스 빅데이터 분석이 구체적인 내용을 살펴보는데 초점을 뒀던 것과 달리, 본 연구는 구체적 내용에 영향받지 않는 정보원 인용의 위상학적 구조를 탐구하고자 했다.

정보원 편향을 수학적으로 모델링하고 검증하는 과정은 다음과 같다. 첫째, 정보원 인용 패턴을 뉴스 정보원 연결망으로 정의했다. 둘째, 정보원을 완전히 불편부당하게 인용했을 때의 가능한 뉴스 정보원 연결망과 편향되게 인용했을 때의 가능한 뉴스 정보원 연결망을 각각 3개씩 총 6개로 제시했다. 셋째, 복잡계 물리학에서 활용되는 연결망의 연결정도 지수(degree exponent)를 활용하여 각 연결망의 가능한 연결정도 지수 값을 계산했다. 넷째, 빅카인즈에서 수집한 1,650만여 개 기사의 인용문 정보를 매체별, 연도별, 지면별로 나누어 1,700개의 데이터세트로 나눈 뒤, 각 데이터세트에서 도출된 뉴스 정보원 연결망의 연결정도 지수 값을 계산하고, 이를 앞의 6개 연결망과 비교했다.

이를 통해 뉴스 정보원 연결망에 보편적 분포가 존재하는지, 그리고 그러한 분포가 정보원 인용의 편향으로 해석될 수 있는지를 살펴보았다. 결론부터 말하면 뉴스 정보원 연결망은 시기, 매체, 주제를 막론하고 충분히 성장했을 경우, 척도 없는 연결망(scale free network)의 특징을 보인다. 좀 더 정확히 말하면 뉴스 정보원 연결망은 연결정도 지수는 1.2로 수렴하며 두터운 꼬리 분포(fat-tail distribution)를 갖는다.

2. 기존 문헌검토

1) 정보원 편향

뉴스는 실재 그 자체가 아니며 언론이 구성한 산물이다(Lippmann, 1922; Shoemaker & Reese, 1996). 정보원은 언론의 현실 구성에 핵심적 역할을 한다(Gans, 1979; Schudson, 1978, Sigal, 1973). 여기서 정보원이란 기자가 기사 작성에 필요한 정보를 얻기 위해 직접 취재해 인용한 사람으로 정의할 수 있다. 정보원은 우선 실명 여부에 따라 실명 정보원과 익명 정보원으로 나눌 수 있다. 실명 정보원은 인명이 특정 여부에 따라 개인 실명 정보원과 기관명만 제시된 집단 실명 정보원으로 분류할 수 있다. 예컨대 '홍길동 문화부 장관'은 개인 실명 정보원, '문화부'는 기관 정보원, '한 관계자'는 익명 정보원으로 볼 수 있다.

언론은 사실성을 확보하는 전문적 관행을 갖고 있다(Tuchman, 1972). 특히 정보원의 인용은 직접 관찰, 수치 제시와 함께 핵심적인 사실성 기제로 꼽혀왔다(남재일, 2008; van Dijk,

1988). 즉 기사는 설명적 진술 뒤에 르포 등을 통해 직접 관찰한 사례나 수치화된 데이터, 또는 인용을 근거로 제시하는 형태로 작성된다. 이중에서도 가장 일반적으로 사용되는 것이 인용이다.

인용은 실명 정보원을 인용할 수도 있고 익명 정보원을 인용할 수도 있다. 일반적으로 실명 정보원 인용이 언론의 사실성 관행에 맞는 것으로 간주된다. 익명 정보원 활용은 내부 폭로자처럼 정보원을 특별히 보호할 필요가 있을 때만 타당한 것으로 여겨진다(Culbertson, 1980; Swain & Robertson, 1995). 한편 문법적으로는 발화를 그대로 전하는 직접 인용과 발화를 재정리한 간접 인용으로 나눌 수 있다. 기사에서는 직접 인용은 큰 따옴표를 사용해 표기된다. 간접 인용은 사설을 제외하면 사실상 사용되지 않는다. 언론에서는 간접 인용을 인용으로 간주하지 않으며, 인용으로 간주하는 발화는 대부분 기사 편집 과정에서 직접 인용으로 수정된다. 이는 직접 인용이 독자에게 생생함을 제공하는 것은 물론, 기자 역시 정보원 발화를 사실대로 인용해야 한다는 사실성 관행의 구속력을 부여해주기 때문이다(박대민, 2015).

정보원 편향 연구는 정보원 연구의 핵심 중 하나다. 다양한 정보원 편향 연구들은 언론이 전국지나 지방지, 신문이나 방송 및 잡지 등 매체 유형을 막론하고 정부 관료나 기업 임원 등 공식 정보원에 대한 의존도가 높다고 지적한다(Berkowitz & Beach, 1993; Brown et al., 1987; Gans, 1979). 정보원 편향은 주제 유형과도 무관하게 나타난다(Atwater & Green, 1988; Chang, 1999). 국내에서도 출입처 제도 등의 영향으로 공식적 정보원에 대한 편향이 적지 않다(장호순, 2001). 이러한 정보원 편향은 권언유착 또는 기업 광고주와의 담합 등 저널리즘 환경에 부정적인 영향을 미칠 수 있다(박동숙·조연하·홍주현, 2001; 유재천·이민웅, 1994). 공식적 정보원에 대한 편향은 지표 가설(indexing hypothesis)로 개념화됐다. 언론이 정부나 기업 중심의 공식 정보원의 의견을 지표(index)로 삼아 전체 기사를 작성한다는 것이다(Bennett, 1990). 물론 언론이 공식 정보원에 끌려다니기만 하는 것은 아니다. 언론은 편집 과정에서 정보원과 그들의 발언을 선별하고 다른 맥락에 배치한다. 기자나 언론사는 정보원 인용을 통해 자신들의 의견을 객관적이고 중립적인 것처럼 제시하기도 한다(박대민, 2015). 대안적 정보원을 발굴해 저항적 뉴스 틀을 제공하기도 한다(Entman, 2003; Harp, Loke, & Bachmann, 2010).

정보원 편향 연구는 매체 간 비교 연구에서 자주 다뤄졌다. 예컨대 북한 관련 보도에서 보수 성향 매체는 보수적 정보원에, 진보 성향 매체는 진보적 정보원에 편향돼 있다는 것이다(이종혁, 2022). 다만 언론사가 자신들의 성향에 따라 정보원을 일방적으로 편향되게 인용하는 것은 아니다. 예컨대 북한 보도에서 언론사는 보수, 또는 진보 성향에 따라 1명의 주요 정보원은 편향되게 활용하지만, 2명의 주변 정보원은 중립적으로 배분한다(한동섭·임종수, 2001). 북한 관련

사설 연구에서는 보수지와 진보지를 막론하고 정부 소속 공식 정보원을 모두 가장 많이 인용한다. 인용 방식에서 차이가 나는 정보원은 보수지의 탈북자 정보원과 진보지의 미국 정부 및 언론 정보원이었다. 이들은 상대적으로 인용 빈도가 적은 주변적 정보원이었으며 보수지는 탈북자 정보원의 대북 비판 내용을 인용했다(김경희·노기영, 2011). 코로나 19와 4대강 관련 보도에서는 언론사가 보수와 진보를 막론하고 정부 소속 공식 정보원에 크게 의존하지만, 진보 매체가 시민이나 사회, 종교 단체 정보원을 적극 활용하는 경향도 확인했다(박대민, 2015; 박주현, 2020). 정리하면 대체로 보수 매체나 진보 매체 모두 공식 정보원에 편향되어 이들을 중심 정보원으로 활용하지만, 주변 정보원 활용에서 성향에 따른 차이를 보이며, 이러한 차이는 각 매체의 성향을 반영하는 방식으로 이뤄졌다.

2) 뉴스 빅데이터 분석을 통한 정보원 편향 분석

몇몇 연구는 정보원 편향이 사회물리학적 보편성을 갖는다고 지적한다. 정보원 편향의 보편성을 파악하기 위해서는 먼저 정보원 인용을 모델링할 수 있어야 한다.

뉴스 정보원 연결망은 그러한 방법 중 하나이다. 뉴스 정보원 연결망은 정보원을 노드(node)로, 기사 공동출현 여부를 엣지(edge)로 하는 무방향(undirected) 1원 연결망(1 mode network)을 뜻한다(박대민, 2013). 다음으로 정보원 편향을 나타내는 특징을 찾아낼 수 있어야 한다. 이러한 특징은 뉴스 정보원 연결망의 위상학적 특징으로 나타날 수 있다. 즉 정보원 편향이 보편적이라면 뉴스 정보원 연결망이 특정 노드에 고도로 집중되는 멱함수 분포(power-law distribution)를 가져야 한다.

뉴스 정보원 연결망의 분포를 연결정도 지수를 통해 분석한 기존 연구에 따르면, 뉴스 정보원 연결망은 매체와 주제를 막론하고 연결정도 지수가 1.6에 수렴하는 두터운 꼬리 분포를 보인다고 밝혔다(박대민, 2014). 이는 미국 언론에 대해서도 나타난다. 미국 통신사인 UPI의 '북한(North Korea)' 및 '대통령(President)' 관련 기사의 뉴스 정보원 연결망을 분석한 결과에 따르면 연결정도 지수는 1.6에 수렴한다(Park, Kim, & On, 2016). 이는 뉴스 정보원 연결망의 분포가 한국 기사는 물론 미국 기사에도 유사하게 나타난다는 것을 의미한다. 그러나 기존 연구는 보편성을 주장하기에는 충분하지 않다. 전자의 연구는 20개 매체를 대상으로 했지만 6개월치만 분석했다. 후자의 연구는 대략 2.5년치 기사를 분석했지만 UPI 매체의 특정 주제만 다뤘다는 한계가 있다.

대규모 기사 분석, 특히 정보원 분석은 빅카인즈가 공개되면서 좀 더 수월해졌다. 예컨대 빅카인즈를 통해 8개 전국지의 정치면과 사회면의 26년치 기사 약 100만 건을 분석한 결과에 따

르면, 통념과 달리 최상위권 정보원과 주제어는 매체별로 큰 차이가 없었다(박대민, 2016). 즉 〈한겨레〉와 같은 진보지든 〈동아일보〉와 같은 보수지든 모두 같은 연도에는 정치면에서는 대통령을, 사회면에서는 교육부를 가장 중요한 정보원으로 다룬다. 다만 대통령이 바뀌거나 부처 이름이 바뀌는 등 시기가 달라지면 가장 중요한 정보원이 바뀔 수 있다(박대민, 2016). 앞서 살펴본 것처럼 매체 간 차이는 상위권 정보원이 아니라 오히려 중요도가 중간 이하인 정보원들에게서 나타난다(김경희·노기영, 2011; 박대민, 2015; 박주현, 2020; 한동섭·임종수, 2001). 언론사는 특정한 방향으로 의제 설정 또는 프레이밍(framing)을 하기 위해 발로 뛰면서 정보원을 발굴한다. 이렇게 발굴된 정보원은 공식 정보원이 아니라 현장에서 만난 주민이나 종교인 등이다. 이러한 정보원은 보통 1개의 기사에만 인용된다. 1개 기사에서 평균 3명을 인용한다고 보면, 한 정보원의 공동 인용된 정보원 수는 2명 안팎이다. 뉴스 정보원 연결망에서 같은 기사에 인용된 정보원 수는 연결정도 중앙성 값으로 구할 수 있으므로 이들의 연결정도 중앙성 값은 보통 2 이하이다. 이러한 정보원은 각각의 중요도는 낮지만, 그 수는 많다. 정리하면 연결정도 중앙성 값이 큰 소수의 공식 정보원이 아니라, 2 이하인 다수의 주변부 정보원이 매체 간 차이를 낳는다(박대민, 2015).

3. 이론적 논의

1) 두터운 꼬리 분포

복잡계 네트워크 이론에서는 자연과 사회의 다양한 현상이 좁은 세상(small world), 더 나아가 척도 없는 연결망임을 밝히고 있다. 척도 없는 연결망은 멱함수 분포를 갖는다(Barabási & Albert, 1999). 멱함수란 거듭제곱의 밑을 변수(x)로 하고, 지수를 상수(constant number)로 하는 함수이다. 멱함수 분포는 흔히 정규 분포(normal distribution)와 대비된다. 정규 분포를 따르는 집단은 평균과 표준편차로 집단을 기술할 수 있다. 즉 극단치가 없이 최대값과 최소값이 대표값인 평균에 수렴한다. 반면 멱함수를 따르는 척도 없는 연결망은 평균과 표준편차가 무의미하며, 평균과 분산으로 예측하기 어려운 극단치가 존재한다. 즉 연결이 극단적으로 집중된 소수의 결점이 존재하는 동시에 연결이 매우 적은 결점이 다수를 차지한다.

척도 없는 연결망은 연결정도(degree)가 k인 노드가 얼마나 많은지를 나타내는 연결정도 분포함수 $P_d(k)$를 통해 설명할 수 있다(강병남, 2010; Barabási & Albert, 1999). 노드 i에 연결된 엣지의 수를 연결정도(degree) k_i, k_i에 대한 분포함수(degree distribution)를 $P_d(k)$

라고 하자. $P_d(k)$는 연결정도가 k인 노드의 수 $n(k)$를 총 결점 수 N으로 나눈 값으로 정의한다. $P_d(k)$의 합계는 1이다.

$$P_d(k) = \frac{n(k)}{N}$$

척도 없는 연결망의 경우, N이 충분히 크다면 k가 작아질 때 $n(k)$가 급증하며, k가 커질 때 $n(k)$는 급감한다. 즉 k가 충분히 클 때, $P_d(k)$는 멱함수인 $k^{-\gamma}$에 근사하며 γ는 k가 매우 큰 영역에서 특정 값에 수렴한다. γ는 연결정도 지수이다. 이를 나타내는 수식은 다음과 같다.

$$P_d(k) \sim k^{-\gamma}$$

이 식을 γ 기준으로 변환하면 아래와 같다.

$$\gamma \sim -\frac{\ln(P_d(k))}{\ln(k)}$$

연결망은 γ 값에 따라 유형화할 수 있다. $\gamma > 3$의 경우 좁은 세상 연결망, $2 < \gamma < 3$인 경우 엄밀한 의미의 척도 없는 연결망이자 극도로 좁은 세상(ultra-small) 연결망이 나타난다. $1 < \gamma < 2$인 경우, 시뮬레이션해보면 전형적인 척도 없는 연결망에 비해 k가 큰 노드가 더 적고 k가 작은 노드는 더 많으며, k가 큰 구간에서 순위 경감에 따른 k 값의 차이가 큰 두꺼운 꼬리 형태가 나타난다. 좁은 세상 연결망이나 두터운 꼬리 분포는 엄밀한 의미에서 척도 없는 연결망이 아니지만 k가 매우 큰 영역에서 멱함수 형태의 연결정도 지수 분포를 따르면 척도 없는 연결망으로 본다(강병남, 2010; 박대민, 2014; Cohen & Havlin, 2003).

척도 없는 연결망을 설명하는 모형으로는 바라바시-알버트 모형(Barabasi-Albert model, 이하 BA 모형)이 대표적이다. BA 모형은 1) 단위 시간에 따라 결점과 연결이 추가되는 성장하는 연결망에서, 2) 이미 많이 연결된 결점에 더 많이 연결되는 선호적 연결(preferential attachment)이 존재하면 나타난다(Barabási & Albert, 1999).

선호적 연결은 폭발성(burst)으로도 나타난다(Barabási, 2010). 폭발성은 어떤 사건이 시간 간격이나 공간적 거리, 사용 빈도 등에서 긴 공백과 집중적인 출현이 반복되는 현상을 의미한다. 폭발성은 우선순위(priority)를 고려할 때 나타난다. 이는 특히 시간 압박(time pressure) 아래에서 선호적 연결이 출현하는 형태가 전형적이다. 예를 들면 이메일 답장, 휴대

폰 통화, 온라인 뉴스 읽기 등의 행동 패턴이 대표적이다(Barabási, 2005; Candia et al., 2008; Dezsö et al., 2006; Eckmann, Moses, & Sergi, 2004).

척도 없는 연결망에 대해서는 여러 논쟁이 있다. 첫째, 엄밀한 의미의 척도 없는 연결망은 적다는 주장이다(Broido & Clauset, 2019). 그러나 척도 없는 연결망의 주창자인 바라바시를 필두로 많은 연구자들이 느슨한 의미에서 다양한 사례를 척도 없는 연결망으로 인정한다(Holme, 2019). 둘째, 척도 없는 연결망로 보이는 것이 실은 푸아송 분포와 같은 다른 분포로 설명된다는 반론이다(Stouffer, Malmgren, & Amaral, 2005). 그러나 학계에서는 대체로 바라바시의 입장을 따른다(Barabási, 2005; Barabási, Goh, & Vazquez, 2005; Vazquez et al., 2006). 셋째, 두터운 꼬리는 엄밀한 의미에서는 척도 없는 연결망은 아니다. 다만 많은 경우 두터운 꼬리는 넓은 의미의 척도 없는 연결망의 하나로 간주된다(강병남, 2010). 특히 이 연구에서는 두터운 꼬리의 연결정도 지수 공식의 계수를 조정해 척도 없는 연결망의 연결정도 지수 공식과 일치시켰다. 이렇게 하면 연결정도 지수 값의 범위만 고려해 두터운 꼬리 분포 여부를 판정할 수 있다. 두터운 꼬리 분포는 척도 없는 연결망 중에서도 좀 더 허브에 집중되는 동시에 꼬리가 두터운 형태로 간주할 수 있다.

2) 뉴스 정보원 연결망의 위상학적 모형

(1) 연결망의 위상학

위상학(topology)은 간단히 말해 어떤 공간이나 도형 등의 위상을 연구하는 학문을 뜻한다. 위상학적 특징은 늘이기, 비틀기, 구부리기 등과 같은 연속적 변형(continuous deformations)에도 불변하는 특징을 갖는다. 즉 위상학은 연속적 변형으로 척도가 명확하지 않아 보이는 상황에서도 보편성을 탐구하고자 하는 시도로 이해할 수 있다. 그래프 위상학(graph topology)에 따르면 그래프 또는 연결망도 위상학적으로 표현될 수 있다. 연결망은 꼭지점, 또는 점에 해당한 버텍스(vertex) 또는 노드와 변, 또는 선에 해당하는 엣지로 표현할 수 있다. 어떤 두 연결망의 노드 수가 같고, 노드별 엣지 수도 같으면 두 그래프는 동형사상(isomorphism)이다. 예를 들어 아래의 두 연결망은 서로 다른 노드로 구성된 다른 연결망이지만, 동형사상이다. 노드 1과 노드 4를 각각 노드 a와 d의 위치로 이동시키면 동일한 모양이 될 수 있기 때문이다(Wilson, 1979).

뉴스 기사와 같은 문서의 위상학적 생산 구조를 파악하려는 다양한 시도가 있었다. 동시 인용 분석(co-citation analysis), 동시 단어 분석(co-word analysis), 공동연구 연결망 분석(collaboration network, 공저자 분석(coauthorship analysis) 등 지식 연결망(knowledge

network) 분석이 대표적이다(Callon, Courtial, Turner, & Bauin, 1983; Moody, 2004; Newman, 2001; Small, 1973). 사회과학 분야나 커뮤니케이션학 분야의 공저자 연결망 분석 내용을 소개하면, 연구자들은 공저자 연결망의 가능한 위상학적 특성으로 척도 없는 연결망, 파편화된 연결망, 구조적 응집성 연결망(structural cohesion network), 작은 세계 연결망 등을 검토한다(Moody, 2004; Song, Eberl, & Eisele, 2020). 연구 결과를 검토해보면 공저자 연결망은 고도로 집중되어 있거나 파편화돼 있기보다는 어느 정도 구조적 응집성을 갖다. 즉 공저자 연결망은 다음 절에서 검토할 별형 연결망(star network)이나 성긴 연결망(sparse network)이 아닌, 넓은 의미의 척도 없는 연결망 중에서도 어느 정도 응집성을 갖춘 두터운 꼬리 분포를 따른다.

(2) 뉴스 정보원 연결망을 통한 정보원 편향과 불편부당성의 위상학적 정의

본 연구에서는 불편부당한 정보원 인용과 편향된 정보원 인용의 위상학적 특징을 분석하기 위해 뉴스 정보원을 분석할 것이다. 〈Figure 2〉는 뉴스 정보원 연결망의 개념도이다. 화살표 왼쪽은 기사 1(A1)에 정보원 1, 2, 3, 기사 2(A2)에 정보원 1, 4, 5가 인용된 것을 시각화한 것이다. 이를 함께 시각화하면, 화살표 오른쪽과 같다. 두 기사에 모두 인용된 정보원 1을 기준으로 2개의 연결망이 하나로 묶인다.

뉴스 정보원 연결망의 연결정도 중앙성 값은 같은 기사에 공동 인용된 정보원 수를 의미한다. 〈Figure 2〉에서 정보원 1의 연결정도 중앙성 값은 4이다. 기사는 기자가 개설한 하나의 작은 토론방과 같다. 정보원은 기자가 초청한 논객이다. 따라서 5명의 정보원 중 가장 중요한 정보원은 여러 주제에 다양한 논객과 함께 대결하는 정보원 1이다.

모든 노드가 연결된 완전 연결망(complete network)은 의미연결망에서 하나의 문서(document)를 뜻한다. 사회연결망의 파당(clique)에 해당하며, 뉴스 정보원 연결망에서는 하나의 기사를 뜻한다. 예컨대 〈Figure 2〉에서 화살표 오른쪽의 연결망에는 정보원 1, 2, 3으로 구성된 완전연결망과 정보원 1, 4, 5로 구성된 완전연결망 2개가 있으며 각각은 기사 1과 기사 2에 해당한다. 반면 같은 기사에 인용되지 않은 정보원 2와 정보원 4, 정보원 3과 정보원 4 등은 서로 연결돼 있지 않다.

뉴스 정보원 연결망은 앞서 소개한 공저자 연결망과는 근본적인 차이가 있다. 공저자 연결망은 근본적으로 논문 공저라는 행위에 기반한 사회연결망이다. 노드는 행위자이며 엣지는 저술이라는 상호작용으로 구성된다. 반면 뉴스 정보원 연결망은 의미연결망이다. 개별 기사에 인용되는 정보원 수는 기사 분량 등의 제한을 받기는 하지만, 전체 기사에서 한 정보원과 공동 인용 가

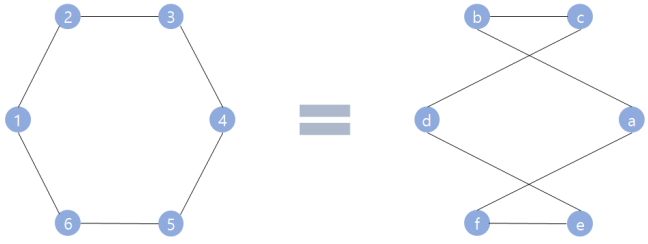


Figure 1. Isomorphism of two networks

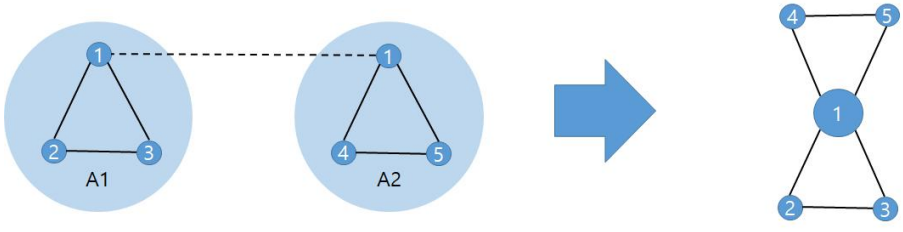


Figure 2. Visualization of news source network

능한 정보원 수는 훨씬 더 많다. 같은 기간에 생산되는 문서 수도 논문보다 기사가 훨씬 많다. 즉 뉴스 정보원 연결망에서 한 정보원의 연결정도 중앙성 값은 공저자 연결망에서 한 저자의 연결정도 중앙성 값에 비해 훨씬 크다.

이어 정보원 편향을 뉴스 정보원 연결망으로 표현해보자. 불편부당한 정보원 인용은 정보원들을 모두 동일한 중요도를 갖고 인용하는 것을 의미한다고 하자. 불편부당한 인용의 위상학적 형태는 이론적으로 다음과 같이 도출된다. 첫째, 완전 연결망이다. 모든 정보원은 같은 수의 정보원과 공동 인용된다. 그러나 이는 1개의 기사에 모든 정보원을 인용한다는 것을 의미한다. 둘째, 동일한 크기의 완전연결망으로 구성된 성긴 연결망이다. 그러나 이는 논의들이 유기적으로 연결되어 심층적으로 다뤄지지 않고 산발적으로 표피적인 형태로 다뤄지는 형태일 가능성이 크다(Park et al., 2016). 셋째, 베테 격자(Bethe lattice) 형태다. 베테 격자는 모든 노드가 같은 수의 이웃(가지)을 갖고 있으며, 순환 고리가 없는 트리 구조의 그래프를 뜻한다(Katsura & Takizawa,1974). 베테 격자는 경계가 없이 무한하게 가지치기를 한다. 베테 격자 형태의 뉴스 정보원 연결망은 모든 정보원이 같은 수의 기사에 같은 수의 정보원과 함께 인용되는 상태를 나타낼 수 있다. 이웃 수가 동일하므로 정보원의 연결정도 중앙성 값은 모두 같다. 불편부당함의 이상적인 상태로 볼 수 있다. 그러나 베테 격자는 노드, 즉 정보원과 엣지, 즉 공동 인용이 무한 확장돼야 한다. 즉 현실적으로 불가능하다. 넷째, 고리형(ring) 연결망도 있을 수 있다. 그

러나 고리형 연결망은 기사당 정보원 수가 2명이며, 하나의 정보원은 2개 기사에만 인용되어야만 하는 형태로 본 연구의 검토에서는 제외한다.

다음으로 편향된 인용의 위상학적 형태를 검토해보자. 불편부당한 형태가 아닌 모든 연결망은 정도의 차이는 있지만 편향된 인용을 표현한다. 먼저 정규 격자(regular lattice) 형태의 뉴스 정보원 연결망을 생각해보자. 언뜻 생각하면 연결망 이론에서 논의되는 정규 격자 역시 불편부당한 인용처럼 여겨질 수 있다. 그러나 〈Table 1〉의 네 번째 그래프에서 보듯이 정규 격자 형태의 뉴스 정보원 연결망은 어떤 형태든 주변이 덜 중요한 편향된 인용 패턴을 보여준다. 노드의 숫자는 연결정도 중앙성 값이다. 정규 격자를 아무리 확대해도 모든 정보원의 연결정도 중앙성 값을 일치시킬 수 없다. 또한 기사 등장 빈도 역시 동일할 수 없다. 중심에 위치한 정보원은 항상 4개의 기사에 4명의 정보원과 함께 인용된 반면, 모서리에 위치한 정보원은 항상 2개 기사에 2명의 정보원과 함께 인용된다.

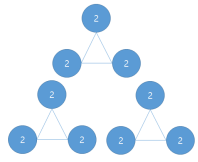
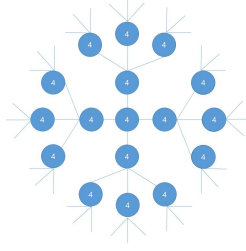
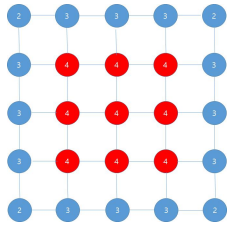
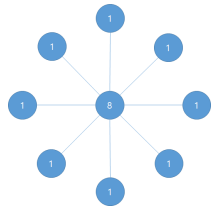
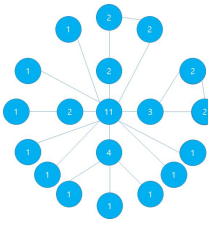
극단적인 경우, 한 사람만 집중 인용하는 별형 연결망을 생각할 수 있다(Shin, 2015). 이는 독재자를 항상 인용하는 언론에서 나타날 수 있다. 앞서 말한 것처럼 정규 격자 형태도 있다. 그러나 일반적으로는 베테 격자와 별형 연결망의 사이가 일반적이다. 척도 없는 연결망은 그러한 수준의 편향을 위상학적으로 표현하는 가장 유력한 형태로 보인다.

각 연결망의 특징을 보다 엄밀히 정의하기 위해 각 연결망의 특징에 따라 연결정도 지수 값을 시뮬레이션해보자. 여기서는 고리형 연결망은 제외한다. 우선 전체 노드 수를 N이라고 하고 연결정도 중앙성 값이 가장 큰 중심 노드인 허브(hub)의 연결정도 중앙성 값을 k_{max}라고 하자.

먼저 완전 연결망을 살펴보자. 모든 노드가 자신을 제외한 다른 모든 노드와 서로 연결돼 있으므로 k_{max}는 자기 자신을 뺀 나머지 노드의 수인 N-1과 같다. 모든 노드의 연결정도 중앙성 값 k는 k_{max}이다. 모든 노드가 허브이므로 허브의 개수 n(k_{max})는 N과 같다. N과 n(k)가 같으므로 앞의 공식에 따라 허브에서 $P_d(k_{max})$는 1이며 γ는 0이 된다. 예를 들면 전체 노드 수가 10,000개라고 할 때, 모든 노드의 연결정도 중앙성 값은 9,999가 된다. k_{max} 역시 9,999이다. n(k_{max})는 전체 노드 수와 같은 10,000이다. 연결정도 분포함수 $P_d(k_{max})$의 값은 1, 허브의 연결정도 지수 값 γ는 0이 된다.

다음은 동일한 수의 노드를 갖는 여러 개의 완전 연결망들로 구성된 성긴 연결망을 살펴보자. 앞의 단일 완전 연결망과 마찬가지로 모든 노드의 연결정도 중앙성 값은 동일하게 k_{max}이다. 역시 모든 노드가 허브이므로 n(k_{max})는 N이며 $P_d(k_{max})$는 1, γ는 0이다. 예를 들어 전체 노드 수가 10,000개, 각각의 작은 완전 연결망은 4개의 노드로 구성된다고 가정하면, 모든 노드의 연결정도 중앙성 값은 4이다. 즉 k_{max}는 4가 된다. n(k_{max})는 N과 같은 10,000개, $P_d(k_{max})$의 값

Table 1. Theoretical Topology of News Source Networks

Category		Visualization	N	k_{max}	n(k_{max})	$P_d(k_{max})$	γ
Unbiased	Complete network		N	N-1	N	1	0
	Sparse network with complete networks		N	k_{max}	N	1	0
	Bethe lattice		∞	k	∞	0	No solution
Biased	Regular lattices		N	4	$(\sqrt{N}-2)^2$	$1-\frac{4}{\sqrt{N}}+\frac{4}{N}$	$0<\gamma\leqq 1$
	Star network		N	N-1	1	$\frac{1}{N}$	$1<\gamma<1.585$
	Scale free network		N	k_{max}	1	$\frac{1}{N}$	$1<\gamma$

은 앞의 완전연결망과 동일하게 1이며, γ 역시 동일하게 0이 된다.

베테 격자의 경우, k_{max}을 무엇으로 가정하든 N과 $n(k_{max})$가 무한대(∞)이다. 따라서 $P_d(k_{max})$은 0이 되며, γ는 해가 없다.

정규 격자에서 가능한 k값은 N이 3*3, 즉 9 이상일 때 그 크기가 얼마나 크든, 노드 수가 몇 개이든 상관없이 2, 3, 4이다. 따라서 k_{max}는 항상 4이다. 정규 격자의 크기를 N이라고 하면, 정규 격자는 둘레의 가로와 세로에 동일하게 N의 제곱근에 해당하는 수의 노드가 배열된다. k_{max}를 갖는 허브는 둘레에서 노드 하나씩을 뺀 것(〈Table 1〉의 정규 격자 그래프에서 붉은색 노드)에 해당한다. 따라서 $n(k_{max})$는 N의 제곱근에서 2를 뺀 수의 자승이다. N이 9일 때, k_{max}는 4, $n(k_{max})$는 1, $Pd(k_{max})$는 0.11111(이하 소수점 여섯번째 자리에서 반올림), γ는 1.58496이다. 가로와 세로 둘레의 노드 수가 각각 4로 1개 늘어나면 γ는 1이 된다. 정규 격자에서 γ는 1.58496과 1 사이에 값이 없다. N이 커지면 γ는 작아진다. N이 무한대가 되면, $P_d(k_{max})$는 1, k_{max}는 4, γ는 0이 된다. 즉 γ는 1.58496(이 때 N=9) 또는 $0 < \gamma \leq 1$이다.

다음으로 허브가 1개이고 허브가 모든 다른 노드와 연결돼 있으며 다른 노드는 서로 연결돼 있지 않는 극단적인 별형 연결망을 가정하자. 이 별형 연결망은 전체 노드 수가 N일 때 k_{max}는 N-1이며, $n(k_{max})$는 1이다. 참고로 허브를 제외한 다른 노드의 연결정도 중앙성 값 ki는 모두 1이다. 허브와 노드 간 k 값의 차이가 존재하는 N의 최솟값은 3이다. 이때 k_{max}는 2, $n(k_{max})$는 1, $P_d(k_{max})$는 1.58496이다. N이 커질수록 γ는 작아지며, N이 무한대에 가까워지면 $P_d(k_{max})$는 0에, γ는 1에 가까워진다. 즉 별형 연결망에서 $1 < \gamma < 1.59$이며 γ의 값은 N이 커짐에 따라 빠르게 감소한다. 예컨대 N=10,000일 때 γ는 1.00001이다.

척도 없는 연결망은 k_{max}가 정규 격자의 4보다는 크고 별형 연결망의 N-1보다는 작다. 또한 별형 연결망처럼 N이 충분히 클 때 $n(k_{max})$는 1이다. 이러한 조건에서 N=10,000일 때 시뮬레이션을 해보자. 앞서 3-1)에서 γ의 구간을 1, 2, 3으로 나누었다. γ가 3 이상이라면 k_{max}는 21 이하이다. γ가 2 이하라면 k_{max}는 100 이하이다. k_{max}가 9,998이면 극단적인 별형 연결망에 가장 가까운 형태로, 별형 연결망과 마찬가지로 γ는 1에 근접한다. k_{max}가 2100, 즉 허브가 10,000명 중 2,100명의 정보원과 공동 인용됐다면, γ는 1.20401이다. 참고로 N=1,000일 때 허브가 같은 비율(21%)의 정보원과 공동 인용됐다면, γ는 1.29187이 된다. N=100일 때 γ는 1.51261이다. 즉 전체 정보원 수 N에서 허브와 공동 인용된 정보원 수 k_{max}의 비중이 같다면, N이 작아질수록 γ가 커진다. 반대로 N 대비 k_{max}의 비율이 동일하다면, N이 커질수록 γ는 작아진다. 또한 γ가 일정 값으로 수렴해 유지된다면, N이 커질수록 k_{max}는 감소한다는 의미이며, 이는 허브의 집중이 줄어드는 것을 의미한다.

4. 연구문제, 연구대상, 연구방법

1) 연구문제

정리하면, 언론사의 정보원 인용은 뉴스 정보원 연결망으로 구조화할 수 있다. 또한 정보원 인용의 편향은 뉴스 정보원 연결망의 위상학으로 모델링할 수 있다. 뉴스 정보원 연결망의 위상학적 특성은 이론적으로 가장 불편부당한 형태인 베테 격자와 극단적으로 편향된 별형 연결망 사이의 척도 없는 연결망 형태로 나타날 것으로 추정된다. 척도 없는 연결망은 충분히 규모가 큰 연결망에서 연결정도 지수가 특정 값으로 수렴하는지 여부로 판단할 수 있다. 다양한 매체와 주제, 시기에서 연결정도 지수가 수렴한다면 정보원 편향은 일정한 형태로 보편적이라고 할 수 있다. 매체별, 주제별, 시기별 연결정도 지수가 다르지만 각 매체나 주제, 시기에는 일관된다면 매체간 비교 연구나 주제에 따른 비교 연구, 또는 시계열 분석의 토대가 될 수 있다.

이에 따른 연구문제는 다음과 같다. 첫째, 뉴스 정보원 연결망의 분포는 어떻게 되는가? 이는 뉴스 정보원 연결망의 연결정도 지수를 계산함으로써 파악할 수 있다. 본 연구에서는 연도별, 매체별, 지면별로 연결정도가 가장 큰 노드에서 연결정도 지수를 계산한 뒤 비교한다.

둘째, 뉴스 정보원 연결망의 위상학적 특성은 무엇인가? 연결정도 지수가 특정 값에 수렴하는지, 수렴하는 값의 범위가 어떻게 되는지를 통해 파악한다. 뉴스 정보원 연결망은 베테 격자나 별형 연결망의 중간 상태인 척도 없는 연결망의 형태를 띨 것으로 추정되는데, 이를 검증한다. BA 모형의 경우, 연결망의 시간에 따라 노드와 엣지가 늘어나는 연결망의 성장(network growth)과 선호적 연결을 확인하는 것으로 살펴볼 수 있다.

셋째, 뉴스 정보원 연결망의 위상학적 특징을 저널리즘 관행과 연결해 해석하고 편향을 해소하는 현실적인 대안을 제안한다. 뉴스 정보원 연결망의 위상학적 특성이 척도 없는 연결망으로 고정돼 있다면, 이는 편향이 특정한 저널리즘 관행으로부터 필연적으로 도출되는 것을 의미한다. 따라서 편향의 개선은 편향을 유지한 채로 개선하는 방식으로 모색돼야 한다.

2) 연구대상

분석 대상은 2000년부터 2020년까지 전국지와 경제지 19개 매체의 주요 5개 지면 기사 전체에 직접 인용된 정보원이다. 인용문 데이터 수집 데이터베이스는 한국언론진흥재단 뉴스 빅데이터 분석 시스템 빅카인즈이다. 직접 인용문은 큰따옴표 내지 그것을 나타내는 마크업(markup) 표식을 기준으로 추출하는 것으로 알려져 있다. 정보원은 직접 인용된 개인 실명 정보원, 집단 정보원, 익명 정보원을 모두 추출한다. 데이터는 재단을 통해 빅카인즈의 개발사로부터 직접 받았

다. 수집 기간은 2021년 1월 16일부터 2월 26일까지다. 구체적인 분석 매체, 분석 지면, 분석 기간은 다음과 같다.

분석 매체: 전국지 11개(경향신문, 국민일보, 내일신문, 동아일보, 문화일보, 서울신문, 세계일보, 조선일보, 중앙일보, 한겨레, 한국일보), 경제지 8개(매일경제, 머니투데이, 서울경제, 아시아경제, 아주경제, 파이낸셜뉴스, 한국경제, 헤럴드경제) 등 총 19개 매체

분석 지면: 정치, 경제, 사회, 문화, 국제 등 5개 지면

분석 기간: 2000년 1월 1일부터 2020년 12월 31일까지 연도별로 21년치

21개 연도, 19개 매체, 5개 지면을 단순 곱하면 1995개의 데이터 세트가 산출된다. 실제 분석 데이터 세트는 총 1,700개이다. 이는 아시아경제처럼 2000년 이후 창간한 경우, 조선일보처럼 저작권 협의에 따라 최근 연도의 기사만 제공된 경우, 일부 기간에 아카이브에서 누락된 경우 등의 사유로 빅카인즈에서 기사 제공을 못하고 있기 때문이다.

분석 기사 수는 16,538,668개이다. 인용문이 없는 기사는 분석 대상에서 제외했다. 인용문 없는 기사는 주로 홍보성 기사나 단신 기사, 인사, 부고, 동정, 사진 등이 해당된다. 연도별, 매체별, 지면별 평균 분석 기사 수는 9,729개다. 인용문은 총 66,867,349개로 분석한 객체 수와 같다. 수집 당시에는 같은 정보원이 하나의 기사에 여러 번 인용될 경우, 각각의 인용문을 나눠서 제공하면서 인용문 수가 늘어났다[3].

정보원 수는 10,779,412명[4]이었다. 연도별, 매체별로는 평균 정치 5,775명, 경제 9,564명, 사회 6,271명, 문화 5,025명, 국제 5,160명을 분석했다. 공백과 'unknownactor'만 일괄 삭제하는 것 외에 정보원 데이터를 추가로 정제하지는 않았다. 정제를 많이 하지 않은 이유는 우선 데이터를 수작업으로 정제하기에는 데이터가 너무 많기 때문이다. 또한 빅카인즈의 개체명 인식(NER, named entity recognition) 성능은 우수한 편이다. 빅카인즈 매뉴얼에 따르면 빅카인즈의 NER는 SSVM(Structured Support Vector Machine)과 BERT(Bidirectional Encoder Representations from Transformers)를 결합해 F1 점수 0.915의 준수한 성능을

3) 2023년 8월 현재 빅카인즈에서는 같은 기사 내 같은 정보원의 인용문은 하나로 묶어서 제공한다.

4) 특정 매체의 특정 연도, 특정 지면 내의 정보원 중복은 제거됐다. 예컨대 조선일보 2020년 정치면 기사 내에서 '문재인 대통령'이 정보원으로 100번 등장했다고 하더라도 1명으로 산정된다. 매체 간, 지면간, 연도간 중복은 포함된다. 즉 조선일보와 한겨레에 2019년과 2020년 정치면과 경제면에 '문재인 대통령'이 정보원으로 등장했다면, 2개 매체*2개 연도*2개 지면에 걸쳐 16명으로 산정된다.

보여준다. 또한 오측 정보원은 중요도가 낮은 경우가 많아 다른 지면이나 매체, 연도에 나타나는 경우가 비교적 적다. 연결정도 중앙성 계산 시 개별 오측 정보원은 아무리 많이 등장하더라도 중복 제거된 채로 합산되기 때문에 중요 정보원의 연결정도 중앙성 값에 크게 영향을 주지 않는다. 설사 중요도가 높게 나오는 소수의 오측 정보원이 있더라도 그 영향력은 여러 정보원에 분산된다. 오히려 연구자가 부분적으로 임의 정제할 경우 전체 데이터의 분포를 왜곡할 우려도 있다.

3) 연구방법

본 연구는 자연어처리와 의미연결망 분석이 결합된 뉴스 빅데이터 분석을 실시했다. 자연어처리 데이터는 빅카인즈를 통해 얻었다. 빅카인즈는 기사에서 정보원과 인용문을 추출해 매핑한 인용문 데이터를 제공한다. 정보원은 이름, 소속, 직함에 대해 각각 NER을 진행해 결합한 데이터를 제공한다. 인용문 데이터에서는 'INFOSRC' 속성으로 제시된다. 기사 식별자(identifier)는 매체 식별자와 날짜, 일련번호가 결합된 'ART_ID'로 제공한다.

의미연결망 분석은 뉴스 정보원 연결망 분석을 실시했다. 뉴스 정보원 연결망의 노드 정보는 'INFOSRC'에, 엣지 정보는 'ART_ID'에 있다. 이를 활용하면 링크드 리스트(linked list) 형태의 연결망 데이터를 구할 수 있다.

연결정도 지수는 중복이 제거된 정보원의 수, 연결정도 중앙성 값이 가장 큰 정보원의 연결정도 중앙성 값을 알아야 한다. 이를 위해 본 연구에서는 연도별, 매체별, 지면별로 모든 정보원에 대한 연결정도 중앙성 값을 구한 뒤, 연결정도 중앙성 값에 따라 정보원을 정렬하고, 연결정도 중앙성 값이 가장 큰 정보원의 개수와 해당 정보원의 연결정도 지수 값을 산출했다.

본 연구에서는 JAVA 기반으로 자체 개발한 SNAnalyzer(https://bit.ly/47esex6)를 사용했다(Park, Lee, & Jeong, 2022). 본 연구에서는 1,700개의 데이터셋을 수집해 데이터를 추출하고, 자료구조를 변환하고, 중복을 제거하고, 데이터를 병합 또는 분할하고, 데이터를 부분 정제하고, 연결정도 중앙성, 연결정도 지수, 기술통계 값을 파일별로 구했다. 전체 노드 수는 정보원 수와 같은 1,078만개, 전체 엣지 정보는 인용문 수와 같은 6,687만개, 사회연결망의 파당에 해당하는 문서 단위의 완전연결망은 기사 수와 같은 1,654만개나 된다. 따라서 분석 값을 단계별로 구하기 위해 UCINET이나 NETMINER와 같은 연결망 분석 프로그램을 사용하는 것은 불가능에 가까우며, Python이나 R을 활용하는 것 역시 적지 않은 컴퓨팅 자원이 필요하다.

SNAnalyzer는 파일 병합 및 분할 모듈인 projector, 정제 모듈인 Cleanser, 파일명 표준화 모듈인 FileNameStandardizer, 연결정도 중앙성 계산 모듈인 DegreeAnalyzer, 기술통계 및 순위화 모듈인 StatisticsAnalyzer, 연결정도 지수 계산 모듈인

DegreeExponentAnalyzer 등으로 구성된다. SNAnlyzer는 1개의 엑셀(Excel) 파일이 가질 수 있는 최대 행의 수인 2의 20승, 즉 1,048,576개에서 컬럼명으로 사용되는 1행을 제외한 1,048,575개 행에 기록된 객체를 한 번에 처리한다. 또한 입력 폴더에 저장된 엑셀을 순차적으로 읽어 들여서 분석한 뒤 출력 폴더에 저장하는 폴더 대 폴더(folder to folder) 기능을 지원함으로써 수천 개의 데이터세트를 모듈별로 한 번에 분석할 수 있다. 즉 본 연구와 같이 수천 개의 엑셀 파일로 저장된 데이터세트의 수천만 개의 노드와 엣지 정보를 개인용 컴퓨터에서도 처리할 수 있다.

SNAnalyzer에서 지원하지 않는 방식의 기술통계 산출은 Python을, 링크드 리스트의 연결망 데이터를 '정보원×정보원'의 행렬(matrix)로 변환하는 것은 Netminer를 사용했다. 연결망 시각화는 UCINET의 NetDraw를 활용했다. 연결정도 지수 등의 시각화는 엑셀의 차트(Chart) 기능을 활용했다.

5. 결과 및 해석

1) 뉴스 정보원 연결망의 성장

뉴스 정보원 연결망은 시간의 누적에 따라 성장하는 연결망이다. 〈Figure 3〉는 한겨레 2020년 정치면의 뉴스 정보원 연결망 시각화 결과이다. 각각 1일치, 1주치, 1개월치, 6개월치, 1년치의 기사에서 뉴스 정보원 연결망을 도출했다.

우선 1일치의 경우 다양한 크기의 소규모 완전 연결망들이 대다수를 차지하는 성긴 연결망에서 시작한다. 뉴스 정보원 연결망의 정의에 따라 하나의 기사는 하나의 완전 연결망을 이루기 때문이다.

1주치 기사의 뉴스 정보원 연결망의 경우 뉴스 정보원 연결망 간에 결합이 부분적으로 이뤄진다. 전체적으로는 완전 연결망의 비중이 감소한 성긴 연결망 구조로 볼 수 있다.

1달치 기사의 뉴스 정보원 연결망은 중앙에 주 구성 집단(main component)이, 주변에 띠 형태로 점들과 작은 크기의 분리된 연결망들이 모인 형태로 그려져 있다. 주변에 존재하는 정보원은 한 기사에 단독 인용된 고립자(isolated), 2명이 인용된 펜던트(pendant), 또는 3명이 인용된 트라이앵글(triangle) 형태의 분리된 연결망을 구성한다.

1달치 이후 연결망은 중심의 주 구성집단과 주변의 띠 형태가 위상학적으로 유지되면서 노드와 엣지만 증가하는 모습을 보여준다. 주 구성집단의 크기가 커지며 주요 정보원의 연결정도

중앙성 값은 빠르게 증가한다. 연결정도 중앙성 기준 가장 중요한 정보원을 살펴보면 1일차에는 '김(정은) 위원장'(연결정도 중앙성 값 9), 1주차에는 '청와대 관계자'(15)였다. 그러나 1개월차부터는 '문(재인) 대통령'이 1위 자리를 뺏기지 않는다. '문 대통령'의 연결 1개월차 연결정도 중앙성 값은 54, 6개월차는 201, 1년차는 402로 급증한다. '문 대통령'과 마찬가지로, 일정한 기간이 지나 한 번 상위권에 자리 잡은 정보원의 연결정도 중앙성은 대부분의 하위권의 정보원에 비해 더 빠르게 늘어나는 마태 효과(Matthew effect)가 나타난다. 상위권 정보원들 간의 연결도 우선적으로 추가되는 경향이 있다.

정리하면 뉴스 정보원 연결망은 초기에는 성긴 연결망에서 점차 거대한 주 구성 집단이 등장하는 형태로 성장한다. 〈Figure 3〉에서 보듯이, 뉴스 정보원 연결망은 직관적으로 봐도 별형 연결망이나 정규 격자가 아님을 알 수 있다. 또한 주요 정보원에 대한 편향은 강화되지만, 구성 집단도 커지면서 응집성도 강화되고 있음을 확인할 수 있다.

2) 뉴스 정보원 연결망의 두터운 꼬리 분포

〈Figure 4〉는 연결정도 중앙성의 값(k)에 따른 연결정도 지수의 값을 시각화한 것이다. 첫번째 그래프는 1,700개 데이터세트 전체를 분석한 결과를 시각화한 것이다. 추세선 식은 $y=3.3013x^{-0.132}$이며, 결정계수(R^2)는 0.8081이다. 연결정도 지수 값은 연결정도 중앙성 값이 10 이하로 작은 경우 0.7에서 3.3까지 사실상 발산하지만, 연결정도 중앙성이 큰 경우 1.2에 수렴한다. 연결정도 중앙성 값이 100 인근에서는 대부분 연결정도 지수가 1.8±0.2이었으며, 연결정도 중앙성 값이 200을 초과하는 796건에서 연결정도 지수 값은 모두 1.2에서 1.8 사이에 분포했다. 나머지 5개 그래프는 전체 데이터를 5개 지면별로 나누어 놓은 것으로, 지면별로도 연결정도 중앙성 값에 따라 연결정도 지수 값이 유사한 패턴으로 변화함을 확인할 수 있다.

연결정도 중앙성이 2,299로 가장 컸던 '서울경제-경제면-2009년' 데이터세트의 경우 연결정도 지수 값은 1.217이었다. 연결정도 지수 값이 0.801로 가장 작은 '아주경제-정치면-2007년' 데이터세트는 연결정도 중앙성 값이 7로 사실상 이상값이다. 실제로 아주경제는 2007년 창간해 빅카인즈에는 2007년 11월부터 기사가 수집돼 있다. 때문에 2007년 정치면의 정보원 수도 37명에 불과하다. 연결정도 지수 값이 두번째로 작은 데이터세트는 '아시아경제-문화-2014년' 데이터세트였다. 연결정도 지수 값은 1.201이었으며 중심 노드의 연결정도 중앙성 값은 2,040로 전체 데이터세트 중에 두 번째로 컸다. 다만 연결정도 중앙성 값이 커지면 무조건 연결정도 지수 값이 1.2에 가까워지는 것은 아니다. 예를 들어 '서울경제-사회-2009년' 데이터셋의 연결정도 중앙성 값은 599인데 연결정도 지수 값은 1.264이다. 그러나 '머니투데이-경제-2011' 데이터셋의

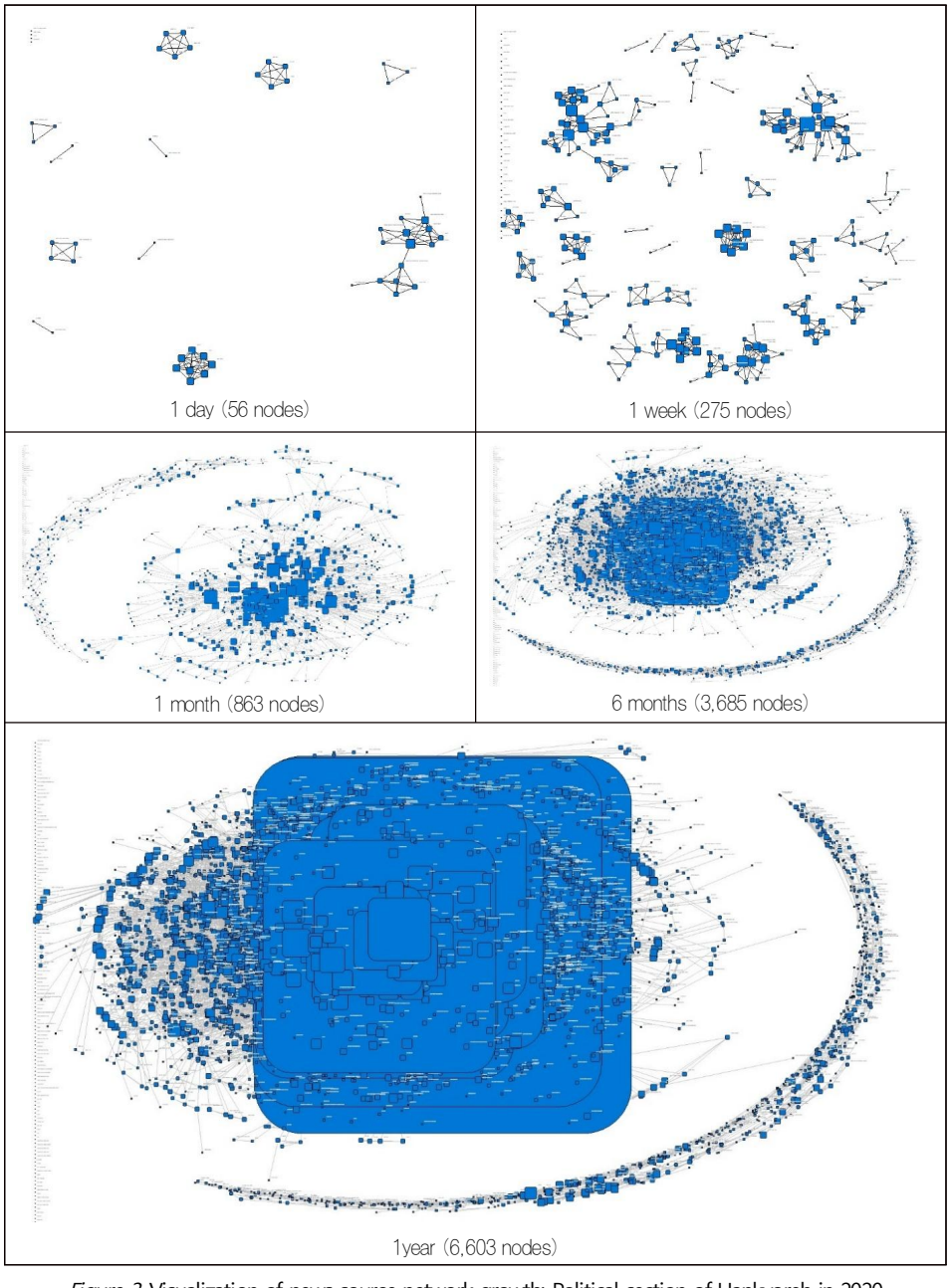
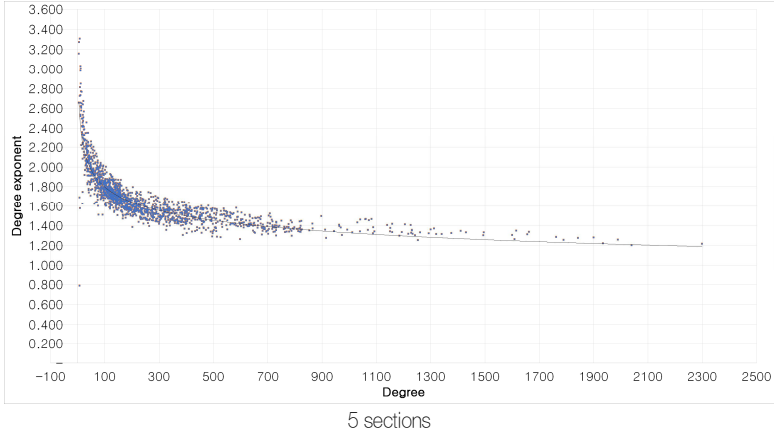
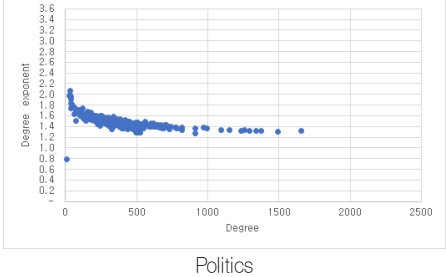


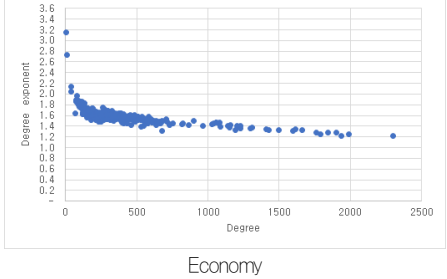
Figure 3. Visualization of news source network growth: Political section of Hankyoreh in 2020

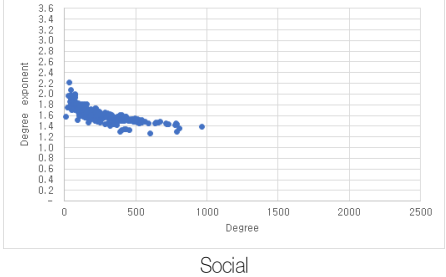
연결정도 중앙성 값은 1083이지만 연결정도 지수 값은 1.467로 다소 크다.

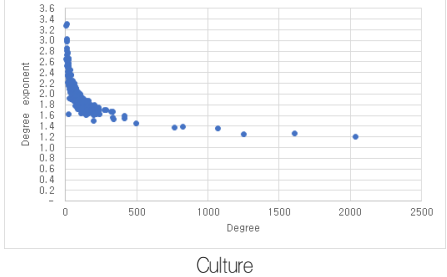
연결정도 중앙성 값이 충분히 크다면 지면별, 매체 간 연결정도 지수 값의 차이도 크지 않











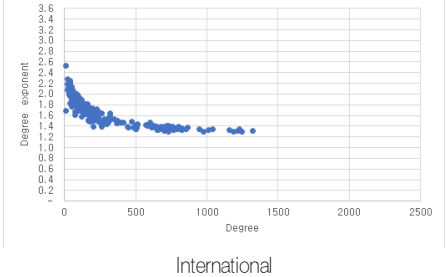


Figure 4. Degree exponents of 5 sections of 19 newspapers by years

다. 지면별 차이를 보면 연결정도 중앙성 값(x축)의 최대값은 경제면이 2299로 가장 컸고 사회면이 964로 다른 지면에 비해 상대적으로 가장 작았다. 그러나 지면별 연결정도 지수 값의 최소값은 경제 1.217, 사회 1.264로 큰 차이가 나지 않았다. 문화면의 연결정도 지수 값 최소는 1.201, 국제는 1.299였다. 즉 연결정도 중앙성 값이 일정 크기를 넘어서면 지면에 상관없이 연결정도 지수 값은 1.2에 수렴하는 양상을 보여준다. 매체별로 연결정도 중앙성이 충분히 큰 값에서 연결정도 지수가 최소인 값을 비교하면 '내일신문-국제-2018년' 세트가 1.386으로 가장 크고 '아시아경제-문화-2014년' 세트가 1.201로 가장 작았다. 전자는 연결정도 중앙성 값이 202였고 후자는 2040로 차이가 컸다.

허브 노드 이외의 연결정도 지수 값도 검토해보자. 〈Figure 5〉는 '조선일보-정치면-2020년', '서울경제-경제면-2009년', '한겨레-사회면-2019년', '매일경제-문화면-2015', '경향신문-국제면-2020'에서 연결정도 중앙성 값에 따른 연결정도 지수 값을 나타낸 그래프이다. 연결정도 지수 값은 지면별, 매체별, 연도별 차이와 무관하게 연결정도 중앙성 값이 작은 꼬리에서는 발산, 즉 다양한 값을 갖지만, 연결정도 중앙성 값이 큰 허브로 가까워지면서 특정 값 1.2로 수렴한다. 〈Figure 5〉에 포함시키지 않았지만 다른 매체, 연도, 지면 역시 결과는 대동소이하다.

앞서 시뮬레이션 값과 비교해보면, 뉴스 정보원 연결망은 연결정도 지수가 0이거나 해가

Table 2. Comparison of Degree Exponents by Media Outlet

Media	Section	Year	Degree	Number of hub node	Degree exponent
Kyunghyang	International	2020	764	1	1.335
Kookmin Ilbo	International	2019	822	1	1.336
Naeil News	International	2018	202	1	1.386
Donga Ilbo	International	2018	702	1	1.317
Maeil Business Newspaper	Culture	2015	1609	1	1.262
Money Today	International	2019	1243	1	1.303
Munwha Il-bo	International	2018	499	1	1.342
Seoul Economic Daily	Economy	2009	2299	1	1.217
The Seoul Shinmun Daily	International	2018	705	1	1.366
The Segye Times	Culture	2012	1253	1	1.254
The Asia Business Daily	Culture	2014	2040	1	1.201
Aju Business Daily	Politics	2020	690	1	1.376
Chosun Ilbo	Politics	2020	1376	1	1.324
JoongAng Ilbo	Politics	2019	1494	1	1.301
Financial News	Culture	2014	764	1	1.372
Hankyoreh	International	2018	653	1	1.326
The Korea Economic Daily	Politics	2019	818	1	1.339
The Hankook Ilbo	International	2018	779	1	1.339
Korea Herald Business	International	2017	813	1	1.344

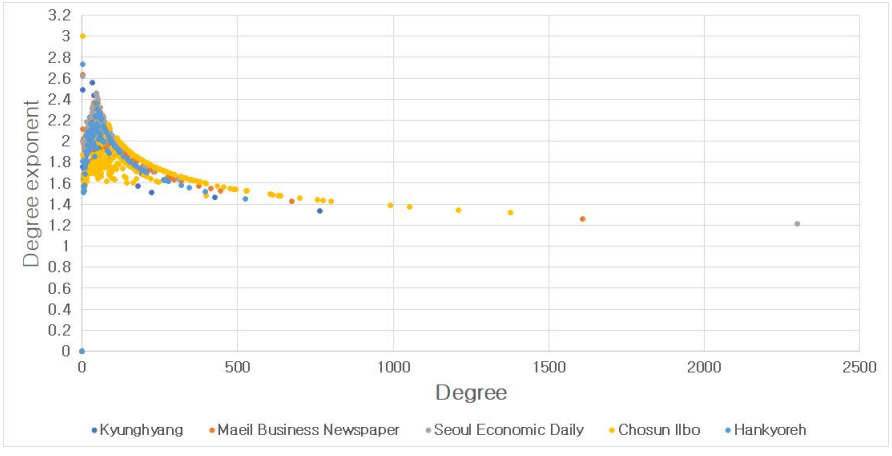


Figure 5. Degree exponents of each node in 5 different sections by degrees

없는 불편부당한 연결망과는 위상학적으로 차이가 난다. 편향된 연결망과 비교하면, 우선 연결정도 중앙성 최댓값이 4보다 훨씬 크고 연결정도 지수 값이 1보다 크므로 정규 격자가 아니다.

별형 연결망과 비교하면, 뉴스 정보원 연결망은 연결정도 지수 값이 1과 1.685 사이인 1.2로 별형 연결망의 연결정도 지수 가능값의 범위에는 있다. 그러나 뉴스 정보원 연결망은 노드 수가 충분히 크고 같을 때 연결정도 지수 값이 훨씬 크다. 즉 노드 수가 충분히 클 때 뉴스 정보원 연결망에서 허브의 연결정도 중앙성 값이 별형 연결망에 비해 훨씬 작다는 것을 의미한다.

앞서 3-2)-(2)에서 살펴본 것처럼 노드 수가 10,000일 때 별형 연결망에서 연결정도 지수 값은 1.00001이다. 뉴스 정보원 연결망은 같은 노드 수에서 연결정도 지수 값은 대략 1.2와 1.3 사이이다. 이는 허브의 연결정도 중앙성 값이 대략 1200(이때 γ는 1.29905)에서 2100(이때 γ는 1.20401)임을 의미한다. 별형 연결망 허브의 연결정도 중앙성 값인 9,999에 비해 훨씬 작다. 즉 허브의 집중도가 낮으며 연결정도 중앙성 값이 작은 부분에서 꼬리의 수가 많고 두텁다는 것을 의미한다. 아래 〈Figure 6〉은 〈Figure 5〉에서 보여준 5개 데이터세트를 합하여 연결정도 중앙성별로 노드 숫자를 나타낸 그래프이다. 연결정도 중앙성 값이 3이하인 노드가 전체 노드의 58.8%를 차지하며, 전체 노드의 90%가 연결정도 중앙성 값이 12 이하이다. 연결정도 중앙성 값 최대치는 '서울경제-경제-2009년' 데이터세트의 2,299로 12,295명 중 중심 정보원에 2,299명이 공동 인용(중복은 제거)됐음을 의미한다. 이 데이터세트에서 연결정도 중앙성 값이 3 이하인 노드 수(정보원 수)는 7,378명으로 60%가 넘는다. 즉 뉴스 정보원 연결망은 허브에 집중되어 있기는 하지만, 별형 연결망에 비해서는 훨씬 분산된 두터운 꼬리 분포의 연결망이다.

정리하면 매체나 지면의 종류와 무관하게 어떤 뉴스 정보원 연결망이 일정 규모를 넘어선

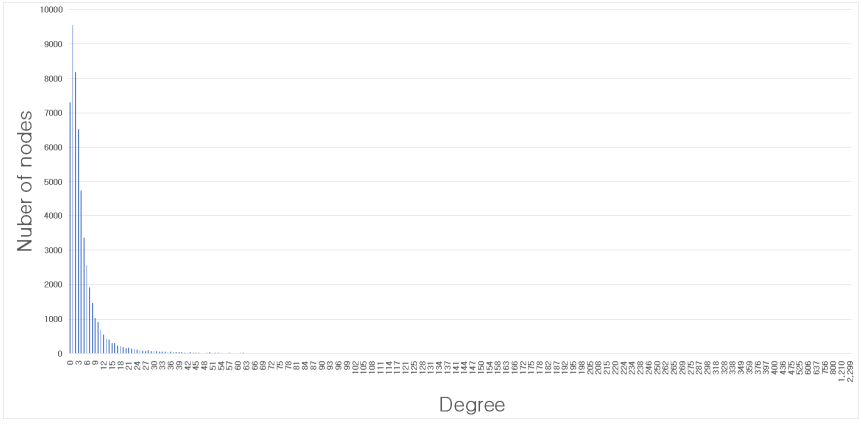


Figure 6. Number of nodes by degrees

경우, 연결정도 중앙성 값이 최대값인 허브에 해당하는 정보원을 기준으로 연결정도 지수를 구했을 때, 전체 정보원 수가 많아지고 연결정도 중앙성 값이 커질수록 연결정도 지수 값은 대체로 1.2로 수렴하는 경향을 보인다. 이는 뉴스 정보원 연결망이 두터운 꼬리 분포를 따른다는 것을 시사한다.

3) 두터운 꼬리 분포의 해석

뉴스 정보원 연결망이 척도 없는 연결망, 특히 두터운 꼬리 분포를 따른다는 것의 의미는 무엇일까? 우선 일반적인 척도 없는 연결망과 마찬가지로 공동 인용된 정보원이 압도적으로 많은 극소수의 정보원이 존재한다는 것을 의미한다. 데이터가 충분히 많은 경우 중요도 1위 정보원은 거의 대부분 1명이다. 또한 최상위권 정보원들 간에는 순위에 따라 연결정도 중앙성 값의 격차가 크다. 예컨대 '조선일보-정치-2020년' 데이터셋에서 1위 정보원인 '문 대통령'의 연결정도 중앙성 값은 1,376이지만 2위인 '민주당'은 1,210으로 그 차이가 166이나 된다. 1위와 2위만이 아니라 3위 '문재인 대통령'(1054), 4위 '청와대'(990), 5위 '정부'(800), 6위 '진중권 전 동양대 교수'(774) 등 20위권 이내의 상위권 정보원 간 중요도 차이도 크다. 이는 기사를 추가로 분석해서 인용 정보를 더한다고 해도 상위권에서는 순위가 잘 바뀌지 않고 안정적임을 의미한다. 반대로 데이터가 적어서 연결망이 충분히 성장하지 않았으며 그 결과 연결정도 지수도 발산하고 두터운 꼬리 분포를 확인할 수 없다면, 분석 결과의 순위를 신뢰할 수 없으며 그 순위는 언제든 바뀔 수 있다는 것을 시사하는 것이기도 하다. 즉 기간이 짧거나 잘못된 표본 추출로 인해 중요한 정보원이 데이터 수집에서 누락됐을 가능성이 크다.

다음으로 두터운 꼬리 분포를 따르기 때문에 연결정도 중앙성 기준 최하위권 정보원, 즉 공

동 인용된 정보원이 없거나 2명 이하인 정보원의 절대적 숫자는 매우 많다. 즉 개별 정보원의 중요도는 낮지만, 수적으로는 압도적인 우위를 보인다. 예를 들어 앞의 '조선일보-정치-2020년' 데이터셋에서 연결정도 중앙성 값이 200이상인 정보원은 72명에 불과하다. 그러나 연결정도 중앙성 값이 2이하인 최하위권 정보원의 수는 4,128명으로 전체 정보원수 14,275명의 29%나 된다. 중앙성 값이 20미만인 정보원 수는 12,711명으로 전체의 89%를 차지한다.

중심부에 위치한 최상위권 정보원은 지표 이론에서 말하는 공식 정보원이 다수를 차지하고 있다. 기존 연구를 살펴보면 이러한 정보원은 정치 분야의 대통령과 중앙부처를 비롯해, 경제 분야에서 대기업, 금융기관, 애널리스트, 컨설턴트, 문화 분야에서 대학과 교수 등이 주를 이룬다. 한편 주변부 정보원은 현장에서 만난 주민이나 시민운동가, 종교인, 특수한 분야의 전문가 등이다(박대민, 2013, 2015).

매체와 지면, 연도와 무관하게 기사 수가 일정 수준 이상 많아지기만 하면 두터운 꼬리가 나타난다는 사실은 정보원 편향이 보편성을 갖고 있음을 시사한다. 즉 당파성을 완화하는 등 언론 윤리를 강화하는 방식으로는 정보원 편향을 피할 수 없다. 그러나 두터운 꼬리 분포는 또다른 현상을 시사한다. 언론은 소수의 공식 정보원이나 전문가들에게 편향되어 있지만, 다른 한편으로는 중요도는 낮지만 절대 다수를 차지하는 정보원을 발굴한다.

언론사 간의 차이를 만들어내는 것은 소수의 상위권 정보원이 아니라 두터운 꼬리의 정보원일 수 있다. 예컨대 대통령의 발언은 모든 언론이 인용하지만, 시민의 목소리는 기자가 발로 뛰면서 발굴해야 한다. 즉 언론의 정보원 인용은 중심에서 불평등하지만, 두터운 꼬리를 통해 담론의 민주화를 달성하고 있다. 또한 언론사 간 차이가 두터운 꼬리에서 두드러진다는 사실은 상위권 정보원의 차이만 분석한 기존의 매체 간 비교연구를 재검토할 필요성을 제시한다. 매체 간 비교 연구에서는 중위권 이하 정보원에 주목해야 한다. 상위권 정보원 분석은 분석 수준이 개별 매체인 경우보다 거시적인 언론계 수준의 분석일 경우에 적합하다.

한편 뉴스 정보원 연결망은 BA 모형의 가정에 부합하는 것처럼 보인다. 즉 시간에 따라 척도 없는 연결망으로 성장하는 연결망이며, 빈익빈 부익부를 초래하는 선호적 연결이 존재한다. 정보원 인용에서 선호적 연결이 나타나는 이유는 인용을 핵심으로 하는 객관주의 저널리즘의 사실성 관행과 마감시간이라는 시간 압박 때문이다.

기사의 사실성을 담보하기 위해 기자는 정보원을 인용한다. 특히 중요한 기사일수록 3명 이상의 다수 정보원을 인용하려 한다. 현장에서 정보원을 인용하는 경우는 물론, 기자 회견이나 출입처, 전화 등을 이용한다고 해도, 기자가 한 명의 정보원을 인용하기 위해 적지 않은 시간을 투입해야 한다. 게다가 정쟁은 물론 투자 등 조금이라도 논란이 예상되는 기사의 경우 정보원을

인용할 때는 반드시 이해가 대립되는 양쪽 관계자들을 찾아서 인용한다. 3명 이상의 정보원을 인용하면서 1개 이상의 기사를 보통 하루 단위로 설정되는 마감 시간 내에 작성하기는 쉽지 않다. 이러한 시간 압박 속에서 기자는 인용된 3명 중 최소 1명은 정부 등 공식 정보원이나 출입처나 컨설턴트, 애널리스트 등 평소 잘 알고 선호해온 정보원에 의존하게 된다. 게다가 이러한 정보원은 기사를 통해 공개되어 다른 기자들에게도 공유된다.

6. 요약 및 제언

본 연구에서는 빅카인즈에서 수집한 1,650만여 건의 기사를 분석하여 뉴스 정보원 연결망의 위상학적 특징을 파악했다. 19개 전국지와 경제지, 2000년부터 2020년까지 21년치, 정치, 경제, 사회, 문화, 국제 등 5개 지면을 대상으로 연도별, 매체별, 지면별로 1700개의 뉴스 정보원 연결망 데이터세트에서 구한 연결정도 지수를 비교한 결과, 연도나 매체 및 지면을 막론하고 연결정도 중앙성 값이 커질수록 연결정도 지수 값은 1.2에 수렴하는 것을 확인할 수 있었다. 이는 기사가 일정한 수를 넘어서 논의가 충분히 진행된 상태일 경우, 뉴스 정보원 연결망이 시기와 매체, 주제를 막론하고 보편적으로 두터운 꼬리 분포를 갖는다는 것을 의미한다. 이는 방법론적, 이론적, 실천적으로 다음과 같은 점을 시사한다.

첫째, 방법론적으로 뉴스 정보원 연결망을 비롯한 뉴스 빅데이터 분석, 더 나아가 내용분석 등을 통한 정보원 분석을 타당하게 수행하려면, 정보원 인용을 뉴스 정보원 연결망으로 나타냈을 때 그 분포가 두터운 꼬리 분포를 갖는지 확인할 필요가 있다. 즉 우선 기사가 많아야 하고, 기사 속 정보원 수도 충분해야 한다. 또한 정보원의 순위를 신뢰할 수 있으려면, 정보원이 여러 기사에서 다양하게 공동 인용되면서 중심 정보원과 거대 구성집단이 출현할 수 있을 정도가 되어야 한다. 수집된 뉴스 빅데이터가 중요한 정보원은 대부분 포함돼 있을 정도로 충분히 성장한 데이터여야 하기 때문이다. 뉴스 정보원 연결망이 두터운 꼬리 분포라는 것은 최소한 상위권 정보원의 순위는 결측이나 오측에 크게 영향을 받지 않을 정도로 강건함을 의미한다. 데이터가 추가되더라도 상위 정보원과 하위 정보원의 격차는 더 커질 것이다.

둘째, 상위권 정보원을 중심으로 하는 뉴스 빅데이터 분석은 매체 간 비교 연구보다는 언론계 수준의 분석에 적합하다. 매체 간 비교 연구에는 두터운 꼬리에 해당하는 정보원에 대한 분석이 요청된다. 문제는 두터운 꼬리 부분에는 동 순위에 수많은 정보원이 있다는 점이다. 이러한 정보원에 대한 분석은 결국 특정 정보원 개인 또는 집단에 주목하여 해당 정보원의 인용문을 읽

어보고 검토하는 질적 연구를 통해서만 가능하다.

셋째, 뉴스 정보원 연결망이 두터운 꼬리 분포로 수렴하는 현상은 정보원의 편향이 사회물리학적으로 보편적인 현상임을 의미한다. 즉 정보원의 편향은 인용을 통해 사실성을 확보하는 객관주의 저널리즘 관행 아래 마감 시간이 존재하는 한 불가피하다. 그러나 정보원의 편향을 방치하자는 것은 아니다. 뉴스 정보원 연결망의 두터운 꼬리 분포는 기자의 전문직 윤리나 독자의 리터러시보다 더 적확한 실천적 전략을 제공한다. 바로 두터운 꼬리를 강화하는 것이다. 정보원의 편향은 불가피하더라도 소수의 정보원끼리 단 한 번만 인용되는 정보원이 많으면 많을수록 언론계의 다양성은 확대되며 담론은 민주화될 것이다.

이는 여러 수준에서 이뤄져야 한다. 우선 기자가 보도 자료나 기자회견, 유선상의 취재에 대한 의존을 줄이고 다양한 정보원을 발굴하고자 노력해야 한다. 그러나 이는 기자의 취재 부담을 늘리는 것을 의미한다. 사실 기자 개인이 정보원, 그것도 거의 인용되지 않은 비공식 정보원을 늘리는 데에는 한계가 있다. 따라서 궁극적으로 편향의 해소는 언론사 단위, 더 나아가 언론계 단위에서 이뤄져야 한다. 즉 한 언론사의 기자가 다른 경쟁 언론사의 기자가 취재하지 않은 정보원을 새롭게 발굴하고자 노력하고, 언론사는 이를 장려한다면 언론계 수준에서 두터운 꼬리가 강화되는 효과가 발생할 수 있다. 뉴스 사용자는 특히 다양한 정보원, 특히 비공식적 정보원의 목소리에 주목하는 방식으로 정보원 편향의 부작용을 극복할 수 있다. 디지털 환경 수준의 개선도 필요하다. 포털이나 소셜 미디어의 추천 알고리듬 역시 의외성(serendipity) 알고리듬 등을 활용해 다양한 정보원의 목소리를 사용자에게 전달할 수 있도록 개선되어야 한다(Reviglio, 2019). 즉 언론사나 기자 단위가 아닌, 정보원 단위에서 다양성이 제공돼야 한다.

본 연구의 결론을 일반화하기 위해서는 추가적인 연구가 필요하다. 우선 본 연구에서 적지 않은 매체의 기사를 분석했지만, 추가적인 매체 분석이 필요할 수 있다. 본 연구가 분석한 전국지와 경제지 외에 빅카인즈의 수록 매체 중 지역지나 전문지, 방송 등 성격이 다른 매체에서도 동일한 분포가 확인된다면 두터운 꼬리 분포의 보편성을 좀 더 강하게 주장할 수 있을 것이다. 또한 비록 결측이나 오측이 많은 것으로 보이지만, 분석 기간도 2000년 이전으로 확대해볼 수도 있다. 이러한 분석에서는 결측과 오측이 분포를 왜곡하는 패턴을 파악하는 한편, 두터운 꼬리 분포 여부로 데이터의 결측과 오측을 탐지할 수 있음을 확인해볼 수도 있다. 더 나아가 미국이나 일본 등 객관주의 저널리즘 관행을 따르는 국가의 매체와 프랑스나 중국, 북한 등 저널리즘 관행이 차이가 나는 국가의 매체로 확대해 분포를 살펴볼 필요도 있다. 예컨대 권위주의 국가의 매체에서는 뉴스 정보원 연결망이 별형 연결망에 가까운 형태로 나타날 것인가? 아니면 알려진 바와 달리 그러한 국가의 매체 역시 동일한 분포를 가지고 있으며, 따라서 객관주의 저널리즘 관행을

따르는 국가의 매체와 비슷한 수준의 담론 민주화를 달성하고 있는가?

또한 본 연구에서는 정보원만 분석했지만, 주제어 등 다양한 언어학적 요소로 분석을 확대해볼 필요도 있다. 예컨대 빅카인즈에서 제공하는 인용문별 주제어를 대상으로 뉴스 주제어 연결망을 구성한 뒤 그 분포를 확인해볼 수도 있다. 만일 주제어에서도 보편적 분포가 나타난다면, 정보원뿐만이 아니라 개별 주제어 측면에서도 매체 간 동조화는 중심을 포함한 최상위권에서, 양극화 내지 파편화는 두터운 꼬리에서 확인할 수 있을 것이다. 이는 기사에서 정보원과 주제의 동조화와 파편화가 동시에 일어난다는 것을 시사한다.

본 연구에서는 개인 실명 정보원, 집단 정보원, 익명 정보원을 모두 함께 분석했다. 익명 정보원과 집단 정보원은 별개의 정보원인데도 동일하게 표기될 수 있다. 때문에 이들의 중요성이 다소 과장됐을 수도 있다. 후속 연구에서는 주요 정보원에 속하는 익명 정보원을 제거하고 분석한다면 좀 더 정확하게 연결정도 지수 값을 계산하여 두터운 꼬리 분포 양상을 좀 더 구체적으로 파악할 수 있을 것이다.

또한 본 연구는 매체, 지면, 연도를 불문한 편향의 보편성 자체는 발견했지만, 구체적으로 누구에게 편향됐는지를 분석하지는 않았다. 기존 연구를 비추어 미루어 보았을 때, 중요도 상위권에서는 동일한 공식적 정보원에 대한 편향이 매체를 불문하고 나타날 것으로 추정된다(박대민, 2015, 2016; Bennett, 1990). 즉 매체별로 약간의 순위 차이는 있을 수 있지만 대통령이나 장관, 재계 회장 등 주요 공식적 정보원은 보수와 진보와 같은 매체 특성을 막론하고 중요한 상위권 정보원으로 다뤄질 것으로 예상된다. 추가로 공식 정보원은 아니지만 애널리스트나 컨설턴트, 교수 등 전문가 집단의 중요도도 확인할 수도 있다(박대민, 2013). 매체 간 차이는 두터운 꼬리를 차지하는, 르포 과정에서 만난 한 번만 등장하는 정보원을 통해 부각되는 것으로 예상된다. 후속 연구에서 연도별, 매체별, 지면별 정보원 전수를 순위에 따라 분석해본다면 매체 간 비교 연구 관점에서 정보원의 편향 문제를 보편적이면서도 구체적으로 분석할 수 있을 것으로 기대한다.

References

Atwater, T., & Green, N. F. (1988). News sources in network coverage of international terrorism. *Journalism Quarterly*, *65*(4), 967-971.

Barabási, A. L. (2005). *The origin of bursts and heavy tails in human dynamics. Nature, 435*(7039), 207-211.

Barabási, A. L. (2010). *Bursts: the hidden patterns behind everything we do, from your e-mail to bloody crusades*. Penguin.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509-512.

Barabási, A. L., Goh, K. I., & Vzquez, A. (2005). Reply to comment on "the origin of bursts and heavy tails in human dynamics". arXiv: physics/0511186.

Bennett, W. L. (1990). Toward a theory of press-state relations in the United States. *Journal of Communication, 40*(2), 103-125.

Berkowitz, D., & Beach, D. (1993). News sources and news context: The effect of routine news, conflict, and proximity. *Journalism Quarterly*, *70,* 4-12.

Bernhardt, D., Krasa, S., & Polborn, M. (2008). Political polarization and the electoral effects of media bias. *Journal of Public Economics, 92*(5-6), 1092-1104.

Broido, A. D., & Clauset, A. (2019). Scale-free networks are rare. *Nature Communications*, *10*(1), 1017.

Brown, J. D., Bybee, C. R., Weardem, S. T., & Straughan, D. M. (1987). Invisible power: Newspaper news sources and the limits of diversity. *Journalism Quarterly, 64*(1), 45-54.

Callon, M., Courtial, J. P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information, 22*(2), 191-235.

Candia, J., Gonzlez, M. C., Wang, P., Schoenharl, T., Madey, G., & Barabási, A. L. (2008). Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, *41*(22), 224015.

Chang, H. (2001). Analysis of the sources and paths of Korean newspapers. *Proceedings of the Korean Journalism Association Conference.* 179-193. [장호순 (2001). <한국 신문의 취재원과 취재경로 분석>. 한국언론학회 학술대회 발표논문집. 179-193.]

Chang, K. (1999). Auto trade policy and the press: Auto elite as a source of the media agenda. *Journalism and Mass Communication Quarterly, 76,* 312-324.

Cohen, R., & Havlin, S. (2003). Scale-free networks are ultrasmall. *Physical Review Letters*, *90*(5), 058701.

Culbertson, H. M. (1980). Veiled attribution: An element of style? *Journalism Quarterly, 55*(3), 456-465.

Dezsö, Z., Almaas, E., Lukács, A., Rácz, B., Szakadát, I., & Barabási, A. L. (2006). Dynamics of information access on the web. *Physical Review E*, *73*(6), 066132.

Eckmann, J. P., Moses, E., & Sergi, D. (2004). Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences of the United States of America, 101*(40), 14333-14337.

Entman, R. M. (2003). Cascading activation: Contesting the White House's frame after 9/11. *Political Communication, 20*(4), 415-432.

Gans, H. (1979). *Deciding what's news*. Austin: University of Texas Press.

Gilens, M., & Hertzman, C. (2000). Corporate ownership and news bias: Newspaper coverage of the 1996 Telecommunications Act. *Journal of Politics, 62*(2), 369-386.

Hamborg, F., Donnay, K., & Gipp, B. (2019). Automated identification of media bias in news articles: An interdisciplinary literature review. *International Journal on Digital Libraries, 20*(4), 391-415.

Han, D., & Lim, J. (2001). An examination of media's utilization of news sources and hegemonic struggle. *Proceedings of the Fall Conference of the Korean Journalism Association*, 27-56. [한동섭·임종수 (2001). 미디어의 뉴스원 활용과 헤게모니 투쟁에 대한 고찰. <한국언론학회 가을철 정기학술대회 발표논문집>, 27-56. 서울: 서강대학교.]

Harp, D., Loke, J., & Bachmann, I. (2010). Voices of dissent in the Iraq war: Moving from deviance to legitimacy?. *Journalism & Mass Communication Quarterly, 87*(3-4), 467-483.

Holme, P. (2019). Rare and everywhere: Perspectives on scale-free networks. *Nature Communications, 10*(1), 1016.

Kang, B. N. (2010). *Complex network science*. Seoul: Jipmundang. [강병남 (2010). <복잡계 네트워크 과학>. 서울: 집문당.]

Katsura, S., & Takizawa, M. (1974). Bethe lattice and the Bethe approximation. *Progress of Theoretical Physics*, *51*(1), 82-98.

Kim, K., & Noh, G. (2011). A comparative study of news reporting about North Korea on newspapers in South Korea. *Korean Journal of Journalism & Communication Studies, 55*(1), 361-387. [김경희·노기영 (2011). 한국 신문사의 이념과 북한 보도방식에 대한 연구. <한국언론학보>, 55권 1호, 361-387.]

Lee, J. (2022). Differences in the use of North Korean experts by conservative and progressive media : Focusing on KPF-BERT-based deep-learning analysis of expert quotes. *Korean Journal of Journalism & Communication Studies, 66*(6), 154-194. [이종혁 (2022). 보수 언론과 진보 언론의 북한 전문가 활용 방식의 차이 탐색: 인용문에 대한 KPF-BERT 기반 딥러닝 분석을 중심으로. <한국언론학보>, 66권 6호, 154-194.]

Lippmann, W. (1922). *Public opinion.* NY: Macmillan.

Malmgren, R. D., Stouffer, D. B., Motter, A. E., & Amaral, L. A. (2008). A Poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences, 105*(47), 18153-18158.

Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review, 69,* 213-238.

Nam, J. (2008). The cultural particularity of objectivism in Korea : The structural feature of routine reporting activities of police reporters. *Journal of Communication Science, 8*(3), 233-270. [남재일 (2008). 한국 객관주의 관행의 문화적 특수성 : 경찰기자 취재관행의 구조적 성격. <언론과학연구>, 8권 3호, 233-270.]

Newman, M. E. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences, 98*(2), 404-409.

Park, D. (2013). News source network analysis as big data analytics of news articles. *Korean Journal of Journalism & Communication Studies, 57*(6), 234-262. [박대민 (2013). 뉴스 기사의 빅데이터 분석 방법으로서 뉴스 정보원 연결망 분석. <한국언론학보>, 57권 6호, 234-262.]

Park, D. (2014). Biases by bursts of quoting sources in journalism. *Communication Theories*, *10*(1), 295-324. [박대민 (2014). 뉴스 정보원 인용에서의 폭발성과 언론의 편향성. <커뮤니케이션 이론>, 10권 1호, 295-324.]

Park, D. (2015). A study of double validity claims in quotations: News source network analysis of news on the Four Major Rivers Project in the Dong-A Ilbo and the Hankyoreh. *Korean Journal of Journalism & Communication Studies*, *59*(5), 121-151. [박대민 (2015). 사실기사의 직접인용에 대한 이중의 타당성 문제의 검토. <한국언론학보>, 59권 5호, 121-151.]

Park, D. (2016). Automated time series content analysis with news big data analytics: Analyzing sources and quotes in one million news articles for 26 years. *Korean Journal of Journalism & Communication Studies, 60*(5), 353-407. [박대민 (2016). 장기 시계열 내용 분석을 위한 뉴스 빅데이터 분석의 활

용 가능성: 100만 건 기사의 정보원과 주제로 본 신문 26년. <한국언론학보>, 60권 5호, 353-407.]

Park, D., Cho, Y., & Hong, J. (2001). A qualitative study of news source-reporter relations - On the problems of beat repeorting system. *Korean Journal of Journalism & Communication Studies, 45*(special issue), 367-396. [박동숙·조연하·홍주현 (2001). 공적 업무 수행을 위한 사적 친분 고리: 출입처에서의 정보원과 기자의 상호작용에 대한 질적 탐구. <한국언론학보>, 45권 특별호, 367-397.]

Park, D., Kim, G., & On, B. (2016). Understanding the network fundamentals of the news sources associated with a specific topic. *Information Sciences, 327*, 32-52.

Park, D., Lee, H., & Jeong, S. (2022). Production and correction of misinformation about fine dust in the Korean news media: A big data analysis of news from 2009 to 2019. *American Behavioral Scientist.*

Park, J. (2020). A comparative study on the 'Corona19' news frame based on ideological orientation of media. *Korean Journal of Journalism & Communication Studies, 64*(4), 40-85. [박주현 (2020). 언론의 이념성향에 따른 '코로나19' 보도 프레임 비교 연구. <한국언론학보>, 64권 4호, 40-85.]

Reviglio, U. (2019). Serendipity as an emerging design principle of the infosphere: Challenges and opportunities. *Ethics and Information Technology*, *21*(2), 151-166.

Schudson, M. (1978). *Discovering the news: A social history of American newspapers.* NY: Basic Books.

Shin, B. (2015). *A practical introduction to computer networking and cybersecurity*. Montezuma Publishing.

Shoemaker, P. J., & Reese, S. D. (1996). *Mediating the message: Theories of influences on massmedia content.*. NY: Longman.

Sigal, L. V. (1973). *Reporters and officials: The organization and politics of newsmaking.* Lexington, MA: DC Heath.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science, 24*(4), 265-269.

Song, H., Eberl, J. M., & Eisele, O. (2020). Less fragmented than we thought? Toward clarification of a subdisciplinary linkage in communication science, 2010-2019. *Journal of Communication, 70*(3), 310-334.

Stouffer, D. B., Malmgren, R. D., & Amaral, L. A. (2005). Comment on Barabási. *Nature, 435,* 207.

Swain, B. M., & Robertson, M. J. (1995). The Washington Post and the Woodward problem. *Newspaper Research Journal, 16*(1), 2-21.

Tsfati, Y., & Cohen, J. (2005). The influence of presumed media influence on democratic legitimacy: The

case of Gaza settlers. *Communication Research, 32*(6), 794-821.

Tuchman, G. (1972). Objectivity as strategic ritual: An examination of newsmen's notions of objectivity. *American Journal of Sociology, 77*(4), 660-679.

van Dijk, T. A. (1988). *News as discourse.* NJ: Lawrence Erlbaum.

Vázquez, A., Oliveira, J. G., Dezsö, Z., Goh, K. I., Kondor, I., & Barabási, A. L. (2006). Modeling bursts and heavy tails in human dynamics. *Physical Review E, 73*(3), 036127.

Wilson, R. J. (1979). *Introduction to graph theory.* Pearson Education India.

Williams, A. (1975). Unbiased study of television news bias. *Journal of Communication, 25*(4), 190-199.

Yoo, J., & Lee, M. (1994). *Government and the press*. Seoul: Nanam. [유재천·이민웅 (1994). <정부와 언론>. 서울: 나남.]

최초 투고일 2023년 08월 13일
게재 확정일 2023년 11월 15일
논문 수정일 2023년 11월 16일