JCDL 2022 COLOGNE

Workshop-EEKE 2022

SOOCHOW UNIVERSITY

# Topic Distribution of China's Data Governance Policies: A Full-text Highlighted Clue Word Approach

**Bikun CHEN** [1], Yuxin LIU [2], Kuan BAI [2], Yi ZHOU [1]

1 School of Social Science, Soochow University (Suzhou, China)

2 School of Economics & Management, Nanjing University of Science and Technology (Nanjing, China)

# 1. Introduction

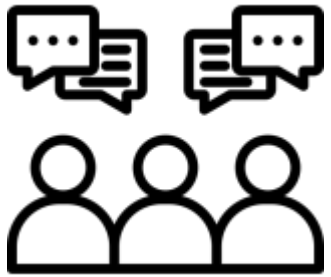
- In-text citations and entity metrics are typical examples of full-text analysis in *scientometrics* (*Ding et al. 2013*). Whether full-text analysis with academic literatures can be extended or properly applied to other full-text literature sources (eg. policy documents) is a critical topic for their robustness and flexibility.
- Quantitative analysis of policy literature (especially the scientific technology and academic policies) is a hot research topic in *scientometrics* and *public management* field, which aims to explore policy topics, intentions, evolutions or relationships among government entities.
- Topic analysis of policy documents are mainly implemented by global keyword frequency or keyword co-occurrence, which needed to be conducted in fine-grained or detailed level.

# 1. Introduction

- In order to expand application scopes of full-text approaches and enrich topic analysis of policy documents, this paper applied full-text highlighted clue word approach to analyze topic distribution and evolution of China’s data governance policies.
- It is a application study in public management domain. Domain research questions come first and are selected by domain experts.
- International consensus: data, together with labor, land, knowledge, technology and management, are regarded as important production factors.
- Data market and data trading is an effective way to utilize data.

# 1. Introduction

- Under this background, governments are required to guide and regulate data trading and data market.
- Some critical questions are needed to be addressed in data governance policies: (1) Whether the relevant subjects have data rights or legal rights (Who); (2) To what extent can the data (the relevant objects) be utilized (What and How); (3) What is the purpose of utilizing the data (Why).



# 2. Data

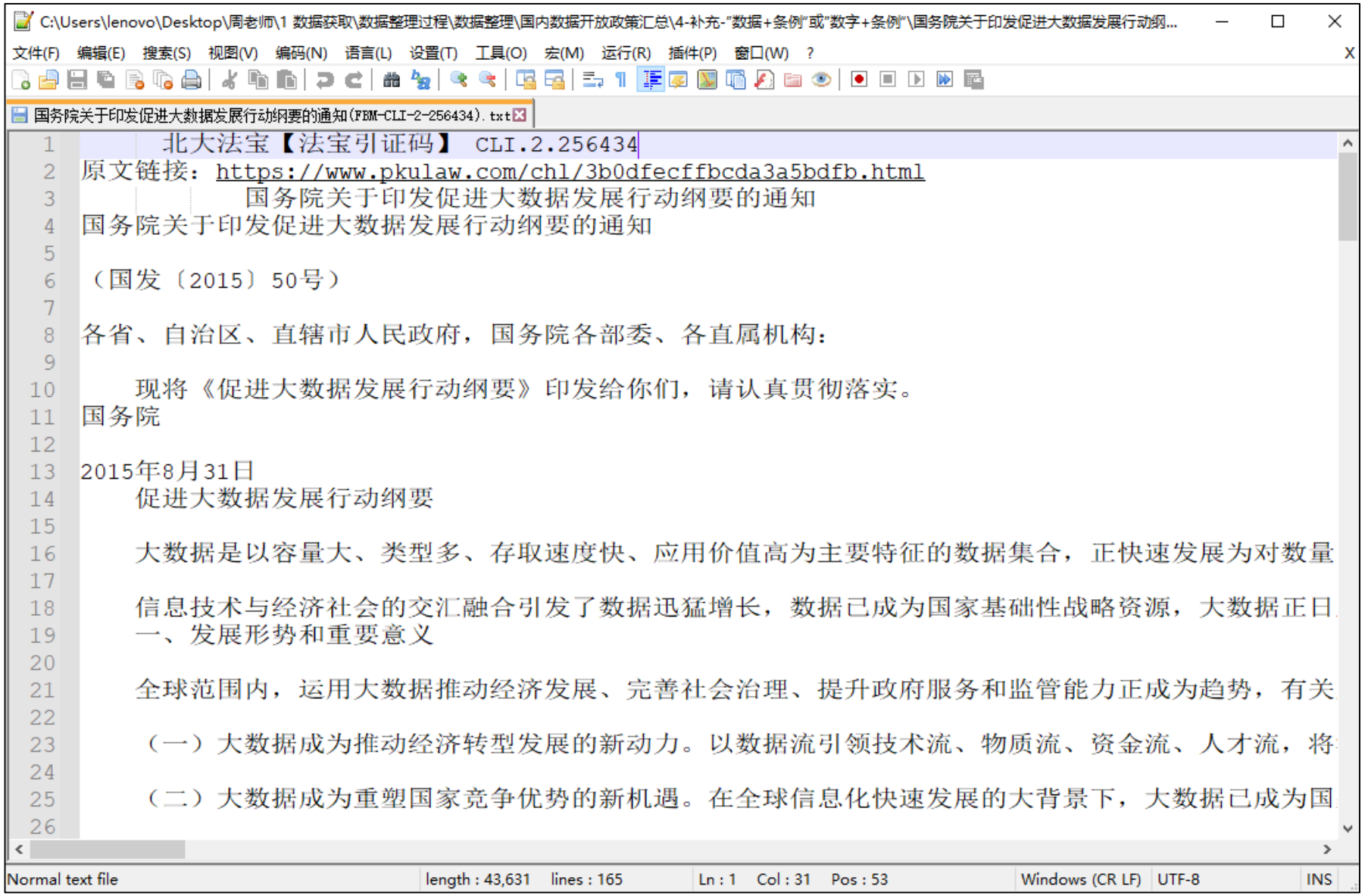
- Data governance policies cover a broad scope, ranging from government data, public data, industry data, big data and so on.
- In this study, data governance policies are acquired by searching a series of data governance related keywords (eg. government data, public data, industry data, big data, government information, public information, social information, digital regulation, data regulation and so on) in *PKULaw Database*, general search engine (eg. *Google*, *Baidu* and *Bing*) and academic literatures.
- Then, every crawled policy is manually read and evaluated by five experienced domain experts. Finally, 258 policy documents are kept as the sample.

# 2. Data

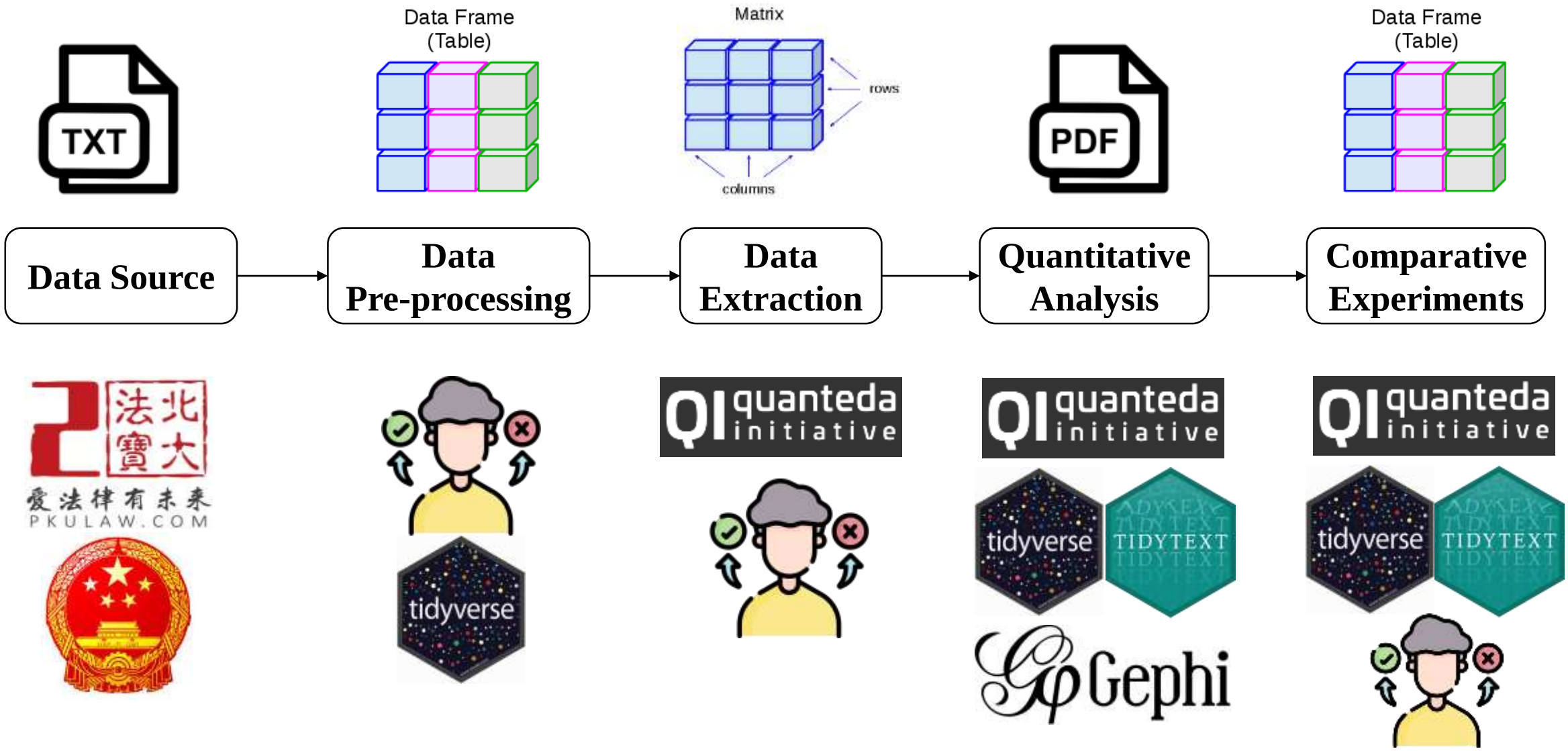
**Figure 1.** A sample policy document in *PKULaw Database*.

# 2. Data

**Figure 2.** A sample policy document downloaded from *PKULaw Database*.



# 3. Method


TXT
Data Frame
(Table)
Matrix
rows
columns
PDF
Data Frame
(Table)
Data Source
Data
Pre-processing
Data
Extraction
Quantitative
Analysis
Comparative
Experiments
北大法寶
愛法律有未來
PKULAW.COM
tidyverse
quanteda initiative
quanteda initiative
quanteda initiative
tidyverse
TIDYTEXT
tidyverse
TIDYTEXT
Gephi

# 3. Method

## 3.1 Classification of Policy Documents.

- Every policy is manually read and evaluated by five experienced domain experts in sample data collection. It is found that policies titled “government information publicity” are firstly issued, then policies titled “government data publicity”, “public information or data publicity”, “big data development” and “certain industry data development” (eg. scientific data and transportation data) are released by governments at all levels in China.
- Therefore, based on the policy titles and trajectory of policy issuance, data governance policies in this study are classified into four categories: government data, public data, industry data and big data.

# 3. Method

## 3.1 Classification of Policy Documents.

| Type | # of policy documents |
|---|---|
| government data policy | 129 |
| public data policy | 31 |
| industry data policy | 79 |
| big data policy | 19 |

# 3. Method

## 3.2 Extraction of Policy Elements.

- This study is mainly based on two policy elements: policy-issuing time and location of policy-issuing agency. Policy-issuing time is the specific time when a policy was released to the public; location of policy-issuing agency refers to China' provinces where the government department that formulated and released the policy located.
- Policy-issuing time and location are the metadata listed in the policy document and can be directly extracted.

# 3. Method

## 3.3 Extraction Full-Text Highlighted Clue Words (FHCW).

- Inspired by entity metrics (*Ding et al. 2013*), full-text highlighted clue words are proposed to solve the research questions and defined as follows: they are a series of notional words in full text of literatures with certain logic (eg. “subjective-objective”, “theory-method-application”, “structural-dynamical”) and significance (eg. representation of sentiments, motivation, behavior or scenario), primarily selected by experienced domain experts.

# 3. Method

## 3.3 Extraction Full-Text Highlighted Clue Words (FHCW).

- Traditional term frequency, n-gram and co-word analysis usually focus on author-assigned, database-assigned or bibliography-extracted keywords with high frequency, TF-IDF method focus on unique term in each document, while FHCW approach emphasize on notional words in full text with any frequency, no matter how common or unique.
- Besides, FHCW approach is similar to expert content analysis because notional words are selected and arranged logically by experienced domain experts, but FHCW are automatically extracted by software and expert content analysis are usually conducted by manually reading and coding.

# 3. Method

## 3.3 Extraction Full-Text Highlighted Clue Words (FHCW).

- In this study, FHCW are selected by five experienced domain experts in terms of “policy subjects, policy objects and application scenarios” logic and orderly arranged on the basis of rights of policy subjects (mainly from low intensity to high intensity), openness degree of policy objects (from low degree to high degree) and specific application scenarios (from specific to general).

| **Type** | **Highlighted clue words** |
|---|---|
| rights of policy subjects | reserve the right; confirm the right; authorization; rights; legal rights |
| openness degree of policy objects | share; openness; development; utilization |
| application scenarios | data security; data assets; digital economy; digital government; digital society |

# 3. Method

## 3.3 Extraction Full-Text Highlighted Clue Words (FHCW).

- Operatively, FHCW in each policy document are extracted by *quanteda* package in R language (*Benoit et al. 2018*).
- In order to evaluate the advantage of FHCW approach, comparative experiments between FHCW approach and traditional global keywords analysis (unigram, bigram and TF-IDF methods) are also conducted.
- For TF-IDF method, top10 keywords in each policy document are extracted and then aggregated globally.

# Comparative Experiments (Global Level)

| Unigram | Bigram | TF-IDF | FCHW |
|---|---|---|---|
| share | healthcare | government section | reserve the right |
| information | information resources | government information resources | confirm the right |
| data | share of government information resources | sharing platform | authorization |
| department | medical big data | openness | rights |
| management | catalogue of government information resources | government data | legal rights |
| government information resources | scientific data | big data | share |
| operation | legal person | public data | openness |
| agency | perform duty | medical care | development |
| big data | public credit | health | utilization |
| construction | administrative region | administrative agency | data security |
| resources | agency of public administration & service | service agency | data assets |
| service | geographic space | open platform | digital economy |
| catalogue | national secret | leading group | digital government |
| health | development and reform | data center | digital society |

# 3. Method

## 3.4 Quantitative Analysis of FHCW.

- FHCW in every policy document are counted and aggregated into each policy type, year and province in China. Then, in order to reduce the influence caused by the unbalanced number in each policy type, year and province, mean value of every FHCW are calculated by dividing the total number of policy documents in each group. Mean value are also normalized between each group (horizontal level) and in each group (vertical level).
- Co-occurrence network of FHCW is constructed based on the co-occurrence relationships in each policy document by *Gephi* software.

# 4. Results

**Figure 3.** Temporal (a) and Spatial (b) Distribution of Different Data Governance Policies.

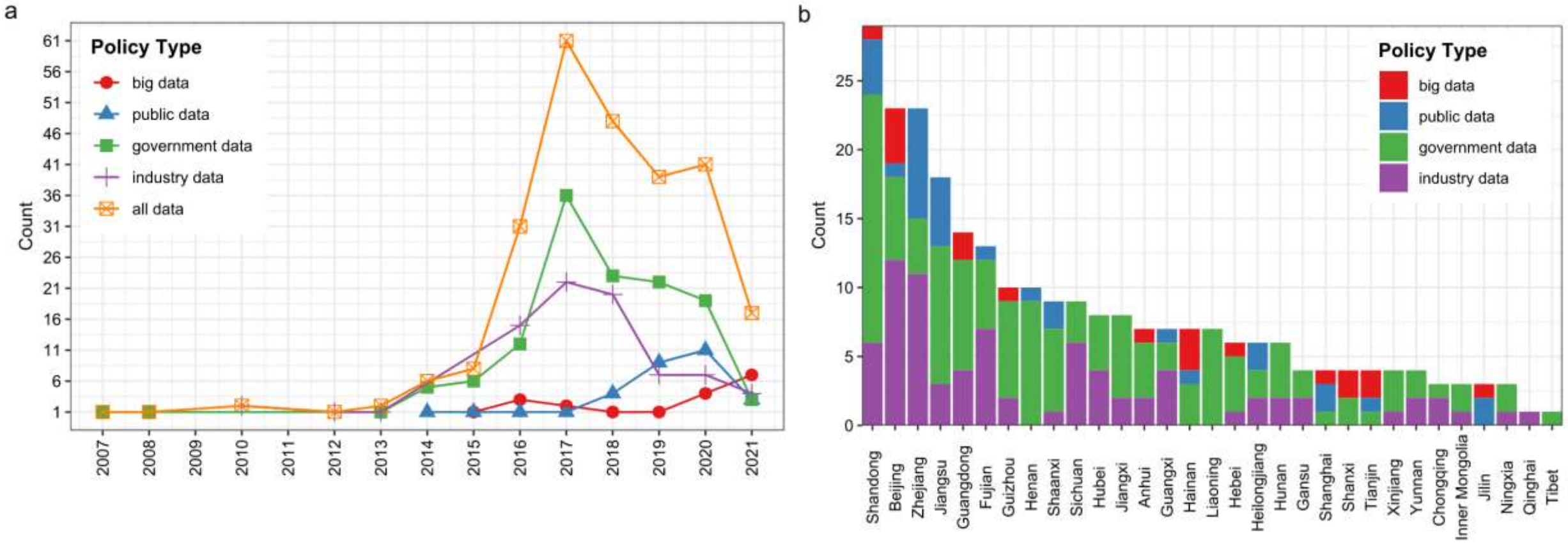
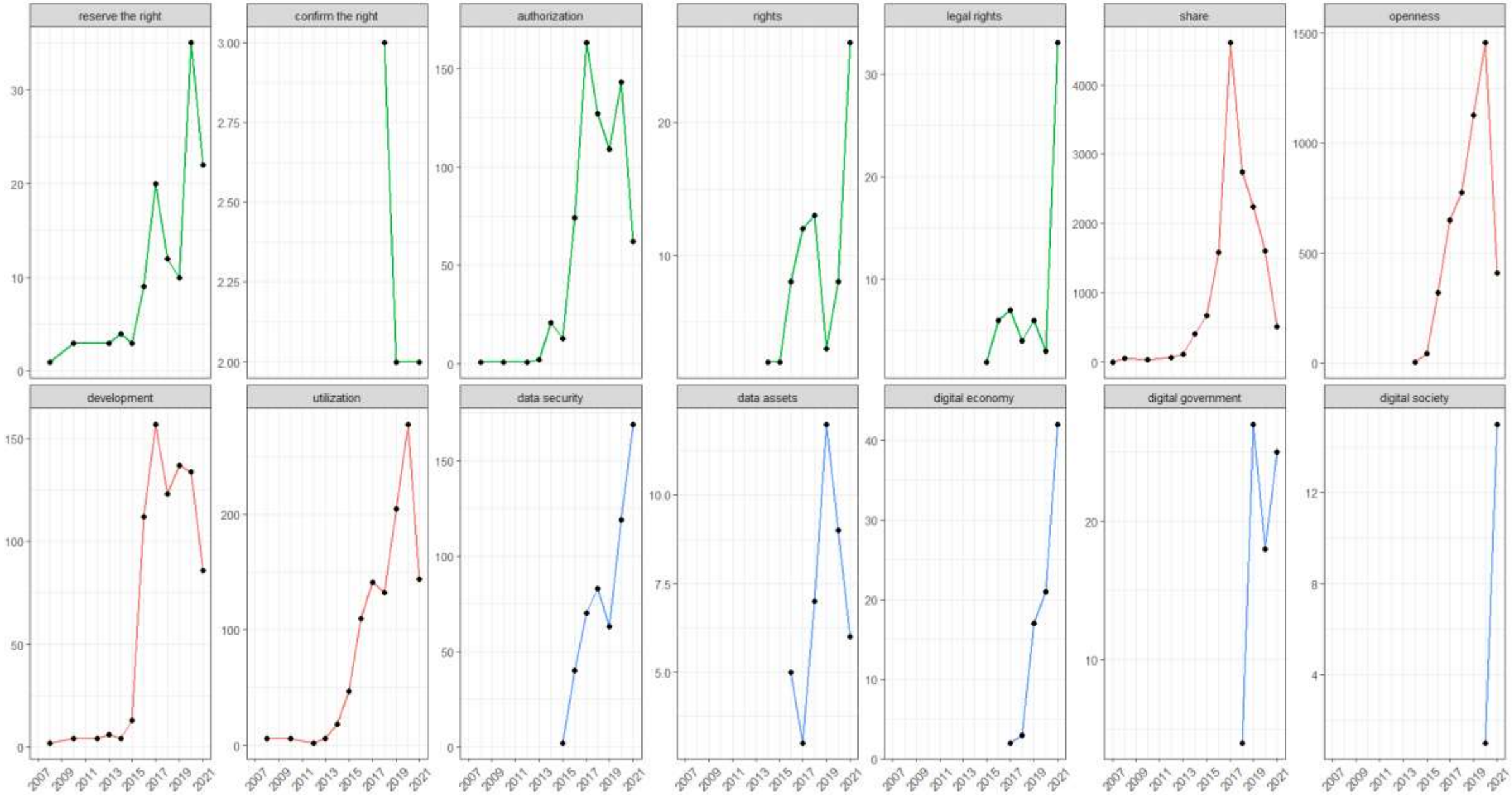


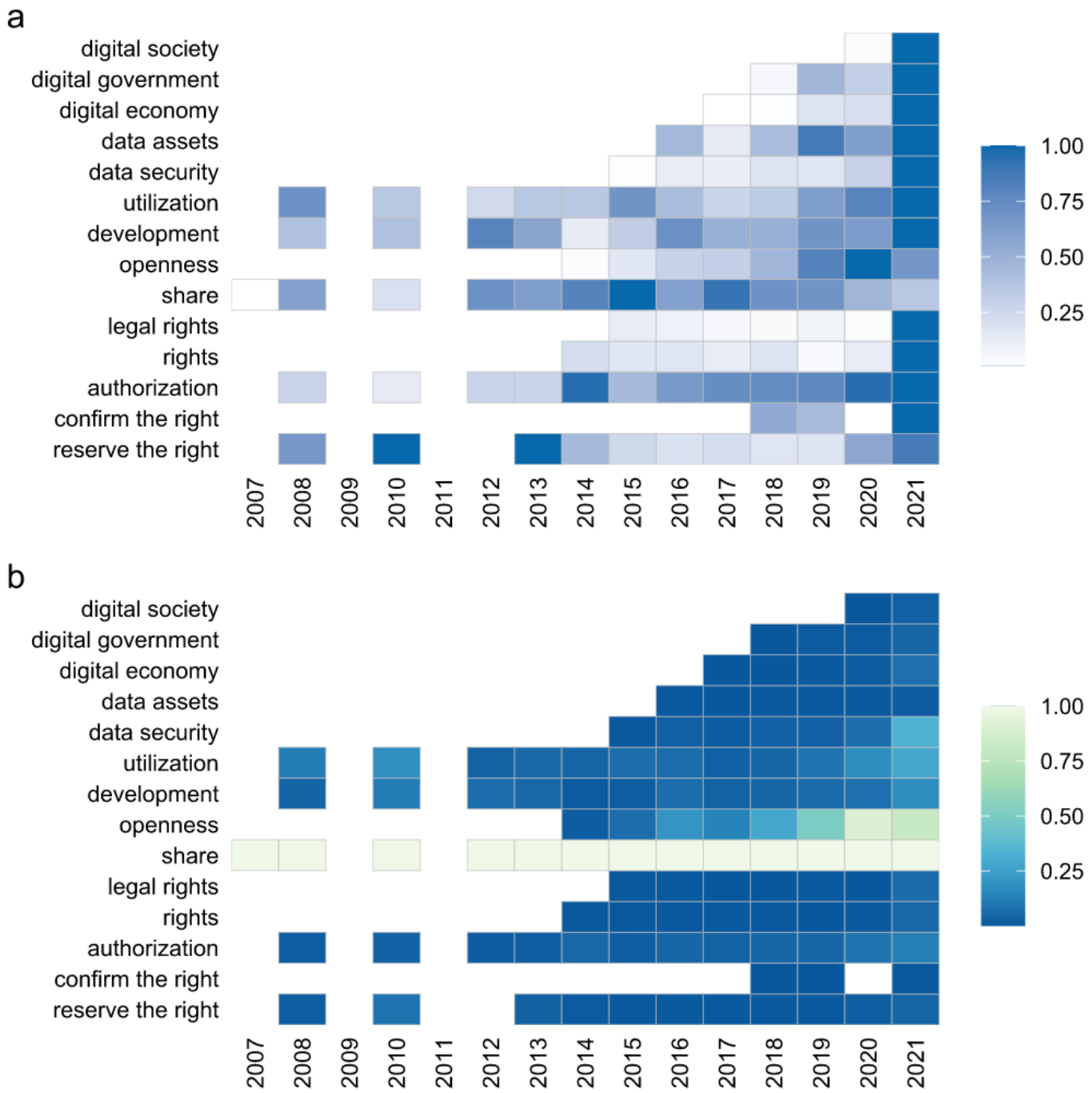
**Figure 4.** Temporal Distribution of FHCW (original value).



# 4. Results

**Figure 4.** Temporal Distribution of FHCW (a. normalized between any year; b. normalized in any year).

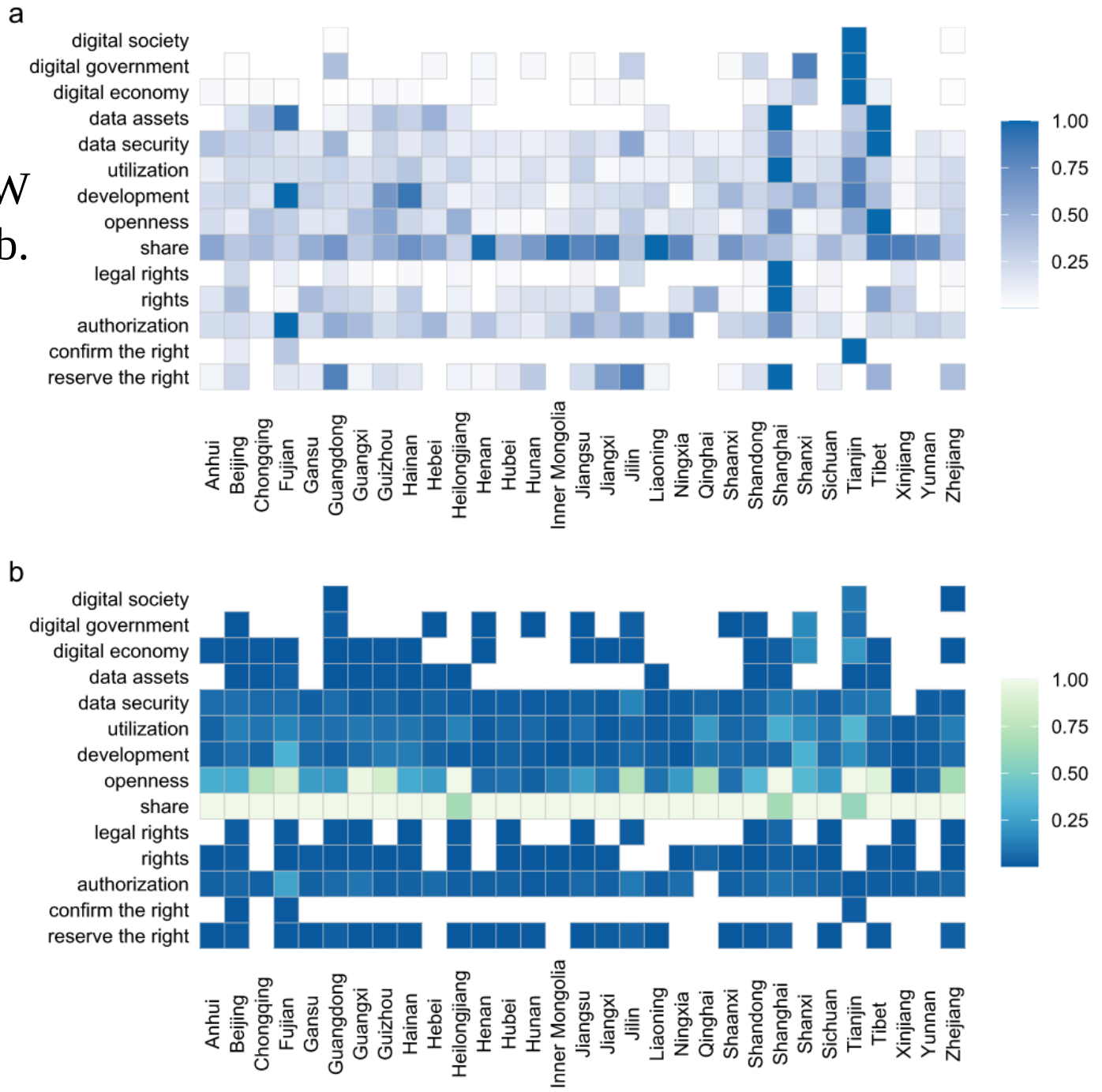
- From Figure 4a, “reserve the right” shows “increase-decrease-increase” trajectory, “authorization”, “confirm the right”, “rights” and “legal rights” show overall growing trend.
- From Figure 4b, “share” is mostly mentioned in almost every year, then “openness”, “utilization” and “data security”; “share” is solely accounting for a large proportion in early years but proportions of “openness”, “utilization” and “data security” are gradually increasing in recent years.



# 4. Results

**Figure 5.** Spatial Distribution of FHCW (a. normalized between any province; b. normalized in any province).

- From Figure 5a, On the whole, provinces in east China (eg. Tianjin, Shanghai, Beijing, Guangdong, Zhejiang and Jiangsu) mention the majority of FHCW, introduce the emerging FHCW (eg. confirm the right, digital government and digital society) and shift their focus from “share” to “development” and “utilization”.
- From Figure 5b, it is shown that “share” is mostly mentioned in most provinces.



# 4. Results

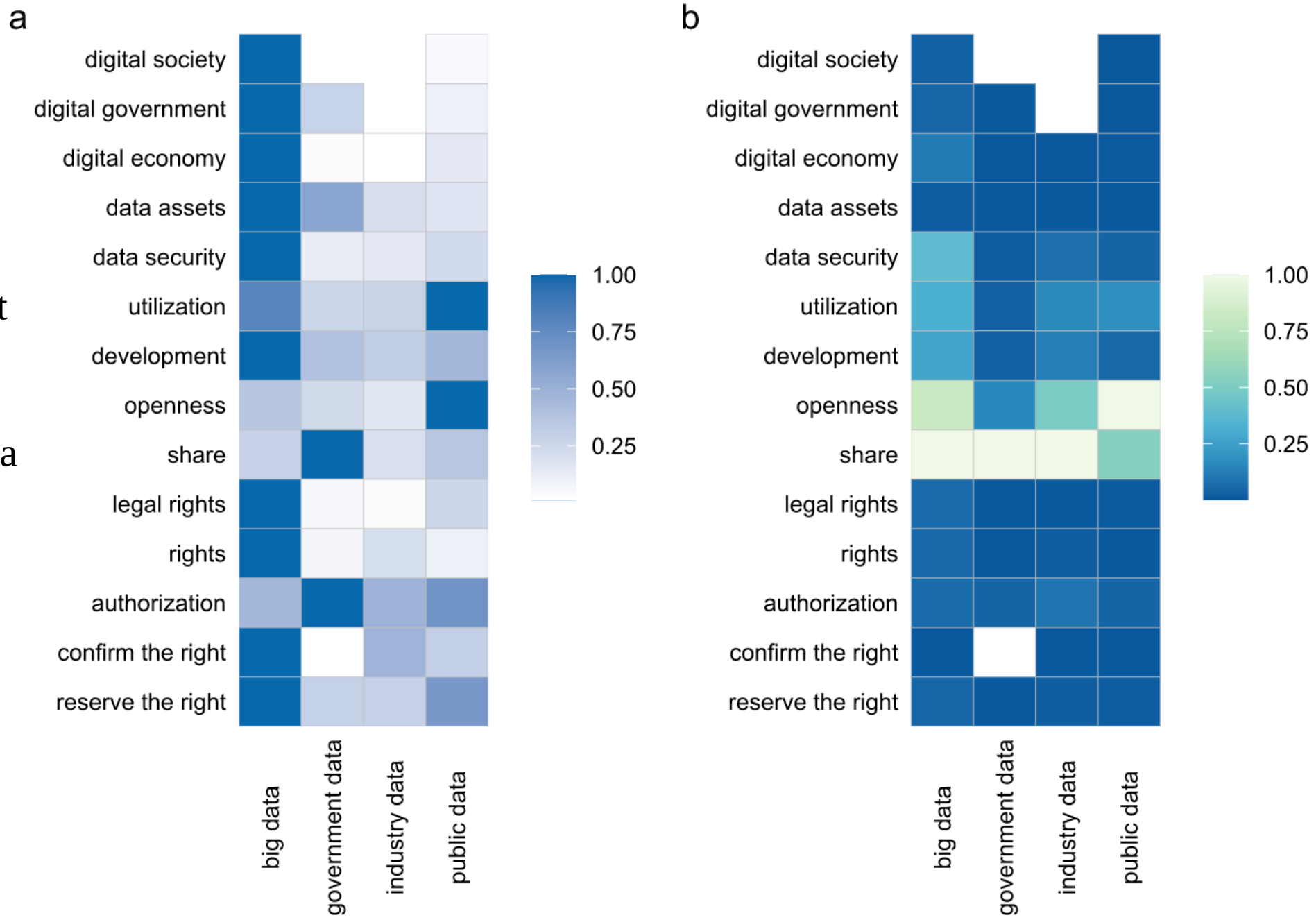



**Figure 6.** FHCW Distribution of Different Data Governance Policies (a. normalized between any kind of data governance policies; b. normalized in any kind of data governance policies).

# 4. Results

**Figure 7.** Global Co-occurrence Network of FHCW (green: rights of policy subjects; pink: openness degree of policy objects; blue: application scenarios).

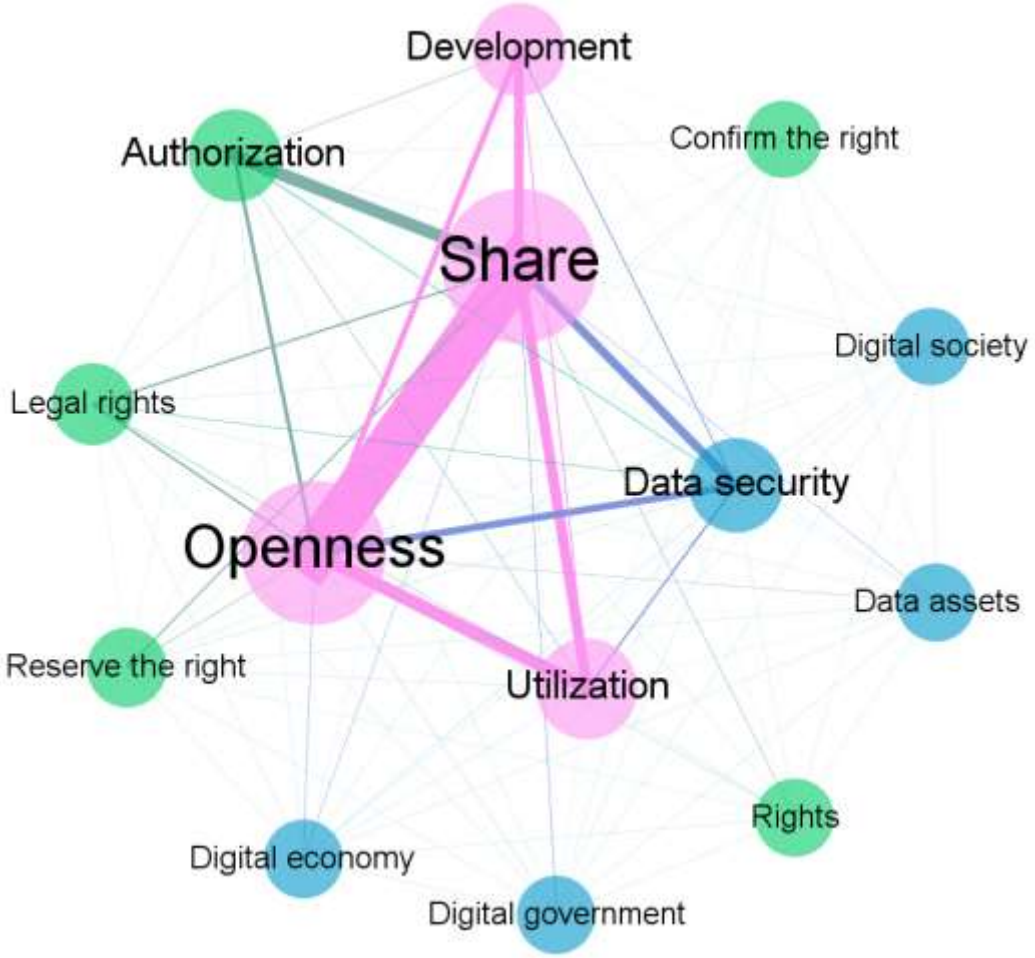
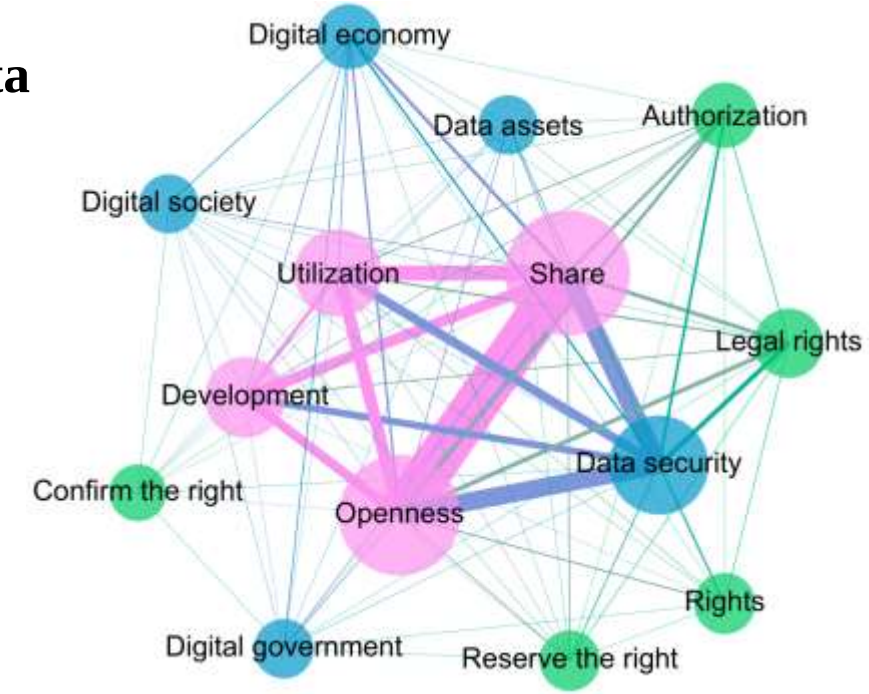


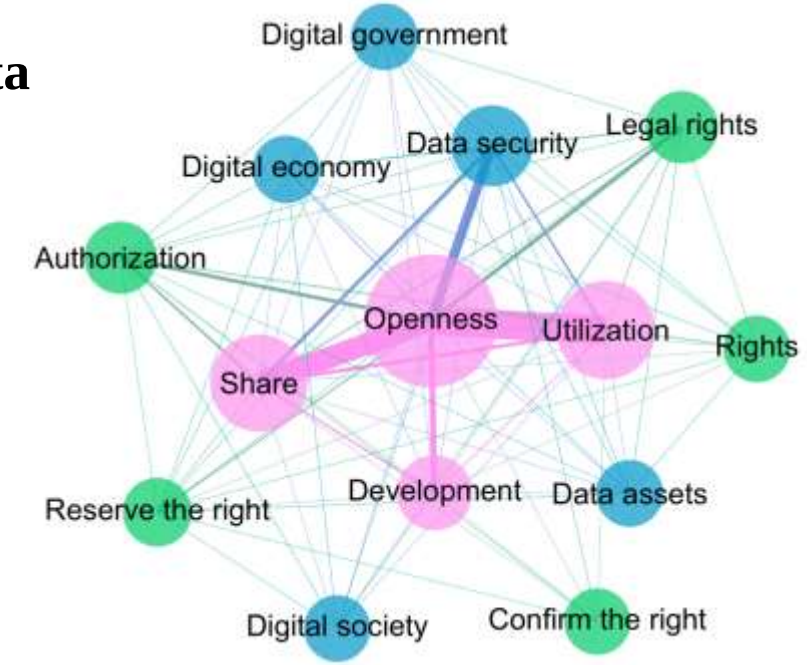

**Table 1.** Top 10 Normalized Co-occurrence Value of Node Pairs.

| Node pairs | Normalized co-occurrence value |
|---|---|
| share - openness | 0.333 |
| utilization - openness | 0.095 |
| share - authorization | 0.09 |
| share - utilization | 0.074 |
| share - development | 0.065 |
| share - data security | 0.062 |
| data security - openness | 0.051 |
| development - openness | 0.042 |
| authorization - openness | 0.027 |
| openness - legal rights | 0.015 |

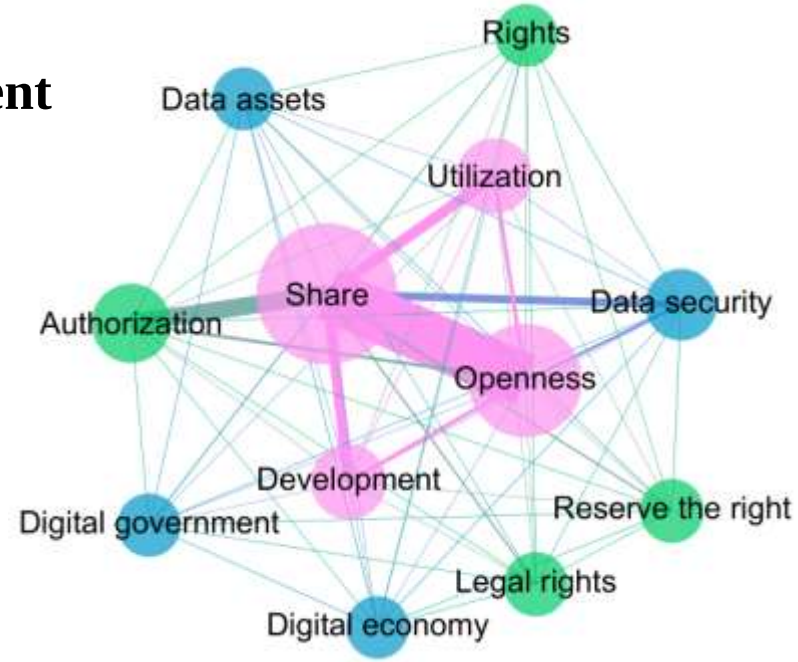
## Big Data




## Public Data




## Government Data




## Industry Data

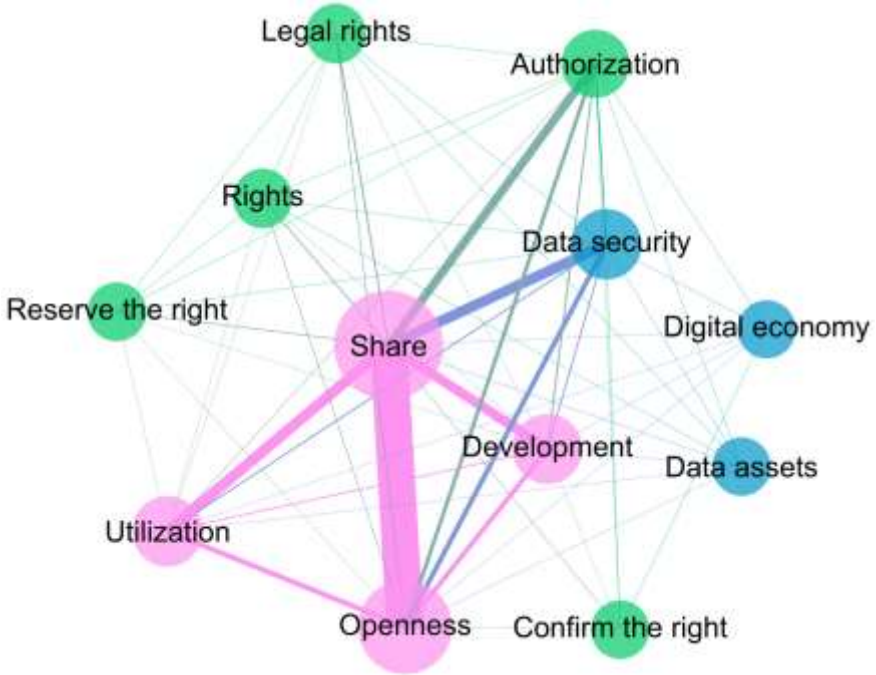


**Table 2.** Top 11 Normalized Co-occurrence Value of Node Pairs in Different Policies.

| Big Data | Public Data | Government Data | Industry Data |
|---|---|---|---|
| share-openness (0.205) | utilization-openness (0.288) | share-openness (0.411) | share-openness (0.324) |
| share-data security (0.085) | share-openness (0.219) | share-authorization (0.137) | share-utilization (0.096) |
| openness-data security (0.085) | openness-data security (0.099) | share-utilization (0.087) | share-data security (0.093) |
| share-utilization (0.07) | openness-development (0.062) | share-development (0.082) | share-development (0.087) |
| utilization-openness (0.057) | openness-legal rights (0.041) | share-data security (0.062) | share-authorization (0.075) |
| share-utilization (0.052) | openness-authorization (0.038) | development-openness (0.034) | utilization-openness (0.046) |
| utilization-openness (0.045) | share-utilization (0.037) | utilization-openness (0.032) | openness-data security (0.043) |
| utilization-data security (0.044) | share-data security (0.036) | authorization-openness (0.024) | openness-development (0.043) |
| development-data security (0.033) | share-authorization (0.022) | openness-data security (0.024) | openness-authorization (0.033) |
| data security-legal rights (0.021) | utilization-data security (0.021) | share-reserve the right (0.015) | development-authorization (0.014) |
| openness-legal rights (0.02) | share-development (0.02) | share-legal rights (0.014) | share-legal rights (0.014) |

# 5. Discussion and Conclusion

- Objects of data governance policies has been expanded from government data to industry data, public data and big data.
- Concerning rights of policy subjects, policy orientation has shifted from data access to protection of data rights.
- Concerning openness degree of policy objects, policy orientation has shifted from data share and openness to data development and utilization.
- Concerning application scenarios, policy orientation has shifted from data-oriented governance to society-oriented governance, paying more and more attention to digital economy, digital government and digital society.

# 5. Discussion and Conclusion

- This study is still in progress. Limited by required pages, only temporal and spatial elements are extracted to explore FHCW and only traditional global keywords analysis (unigram, bigram, TF-IDF and co-occurrence methods) is conducted.
- The selected FHCW are enough to address the research questions in global level, more sub-class FHCW are needed to be incorporated.
- In further study, more policy elements (eg. policy instruments and policymakers) and more advanced method (eg. word embedding and dynamic network analysis) will be introduced and compared.

# Questions?

**E-mail**: *bkchen@suda.edu.cn*