Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 11
TPCEE 2025

Research on the Global Distribution of Cybercrime Base on Machine Learning and Data Mining

Yufei Wu *, Zhirong Zhuang, Ziyou Chen

Hohai University, Nanjing, China

* Corresponding Author Email: 2308060318@hhu.edu.cn

Abstract. With the acceleration of the global informatization process, cybercrime has increasingly become a serious challenge for the world. The transnational nature of cybercrime makes the response of a single country inadequate, and there is an urgent need for global policy coordination and optimization. Through an in-depth analysis of VCDB data, this paper reveals the distribution characteristics of global cybercrime, and points out the significant differences between developed and developing countries in terms of cybersecurity incident reporting and transparency. Then, use the Grey Relational Analysis method and the five key indicators of the Global Cybersecurity Index (GCI)to evaluate the effectiveness of their national security policies. The results show that technical capability and organizational management are the core factors affecting the effectiveness of cybersecurity policies. In addition, this paper uses the entropy-weighted TOPSIS method to further analyze the relationship between demographic data and network security indicators to verify the previous conclusions. By fostering a unified approach, the international community can better address the evolving landscape of cybercrime and safeguard the digital ecosystem.

Keywords: Cybercrime; Cybersecurity Policy; Grey Relational Analysis; Entropy-Weighted TOPSIS; Global Cybersecurity Index Introduction.

1. Introduction

The emergence of the Internet has broken the advantages accumulated by traditional trade, and has prompted emerging economies and developing countries to participate in market competition and share the fruits of development [1]. However, the development of cyber technology has also made cybercrime transnational [2], which not only poses serious dangers and harms to the population and property of countries around the world, but also poses great challenges to the formulation and implementation of cyber security policies in various countries [3]. Therefore, it is worth considering how to accurately identify and analyze the current situation of global cybercrime distribution and the effectiveness of national cybersecurity policies.

Scholars have conducted in-depth research on the challenges posed by global cybercrime and how to deal with them. At present, most scholars in China have put forward opinions on cybercrime from the level of legal mechanism. Jiang Su proposed to use the United Nations as a platform to integrate the values of "diversity and inclusion" and promote the formulation of a new global convention on cybercrime [4]. Li Yan believes that emerging countries can build a new global international legal mechanism through the "two-step" path of "multilateralism", and countries can strengthen cooperation with each other to jointly combat and prevent transnational cybercrime [5]. An Keying proposed to construct a dualistic legal naturalization method of "legal governance and technological governance" from the perspective of China so that it can have jurisdiction over transnational cybercrime cases [6]. Foreign scholars have conducted research on cyber security from multiple perspectives. Qasem Abu Al-Haija et al. [7] proposed a cybercrime time series estimation model using an autoregressive model to estimate the number of global cybersecurity incidents in the coming years. Ishrat Hameed et al. [8] used Pakistan as an example to discuss and analyze the factors that cause cybercrime and measures to reduce it. Shefali Batra et al. [9] analyzed statistics on various types of cybercrime over the past few years and analyzed the global impact of cybercrime. It can be seen that cybercrime has attracted the attention of scholars from all over the world, but there is still a

lack of relevant research on accurately assessing the effectiveness of global cybersecurity policies, especially the use of mathematical models and data visualization analysis.

The global cybersecurity situation remains severe, and governments need to introduce relevant cybersecurity policies to combat cybercrime, which is critical to the protection of critical infrastructure and the development of the digital economy [10]. This paper innovatively uses Grey Relational Analysis and five key indicators of GCI to evaluate the effectiveness of cybersecurity policies. At the same time, the Entropy-Weighted TOPSIS method is used to further analyze the relationship between demographic data and network security indicators. It provides a scientific basis and suggestions for the formulation and optimization of global cybersecurity policies.

2. Methodology

2.1. Description of the Dataset

The experimental data of this paper comes from the VCDB dataset and official authoritative reports, which are available in https://www.itu.int/epublications/publication/global-cybersecurity-index-2024 and http://verisframework.org/vcdb.html.The Global Cybercrime Report released by cybersecurity companies ranks the Internet security of 94 countries and regions around the world in terms of national security index and global security index. The sample size of the GCI report provided by ITU includes 194 countries, and countries are scored according to five indicators: legal, technical, organizational, capacity building and cooperation.

2.2. Data Preprocessing

The Global Cybercrime Report selects the top 10 and bottom 10 countries, as well as the 94 countries disclosed in the report, based on the national security index and risk index. The top 10 and bottom 10 countries are evaluated by using five indicator data provided by GCI, combined with Grey Relational analysis and Entropy-Weighted TOPSIS method.

2.3. Grey Relational Analysis

Grey Relational Analysis is a method based on the grey system theory, which measures the degree of correlation between factors according to the degree of similarity or difference in the development trend between factors. It has the characteristics of low data requirements, simple calculation, wide applicability and dynamic analysis capabilities. Therefore, this paper uses Grey Relational Analysis to analyze the effectiveness of cybersecurity policies.

Table 1 is a table of the selected data. "x_1, x_2, x_3, x_4, x_5" represent Legal Measures, Technical Measures, Organizational Measures, Capacity Development and Cooperation Measures. "ω" represents Global Cybersecurity index, （GCI, $\omega = x_1 + x_2 + x_3 + x_4 + x_5$）. "$y$"represents Cyber-Safety Score.

Table 1. Some country-specific data

Country	x_1	x_2	x_3	x_4	x_5	ω	y
Denmark	19.3	18.94	18.98	19.48	15.89	92.6	8.91
Netherland	20	19.84	18.98	18.82	19.41	97.05	8
Germany	20	19.54	18.98	19.48	19.41	97.41	8.76
Singapore	20	19.54	18.98	20	20	98.52	7.96
United States	20	20	20	20	20	100	8.73
United Kingdom	20	19.54	20	20	20	99.54	8.44
Australia	20	19.08	18.98	20	19.41	97.47	8.16
Canada	20	18.27	20	20	19.41	97.67	8.35

Japan	20	19.08	18.74	20	20	97.82	8.09
Isreal	19.68	16.99	15.02	19.24	20	90.93	7.75
Georgia	17.75	17.13	14.67	15.89	15.63	34.11	5.98
Bosnia and Herzegovina	10.41	6.56	1.02	3.12	8.33	29.44	3.46
Tunisia	20	19.54	12.21	16.96	12.52	86.23	4.72
Panama	10.41	10.94	2.37	6.12	4.26	35.23	3.66
Dominica	0.85	0	3.35	0	0	4.2	4.98
Hungary	18.16	16.82	18.29	18.6	19.41	91.28	6.27
Saudi Arabia	20	19.54	20	20	20	99.54	5.54
Honduras	2.2	0	0	0	0	2.2	3.13
Russia	20	19.08	18.98	20	20	98.06	6.39
Myanmar	9.39	3.64	4.71	8.92	9.75	36.41	2.22

First, make a reference data column $\boldsymbol{x_0}$.

$$x_0 = x_0(1), x_0(2), \ldots \ldots x_0(n) \tag{1}$$

The comparison series in association analysis is often noted as $\boldsymbol{x_i}$.

$$x_i = x_i(1), x_i(2), \ldots, x_i(n), i = 1,2, \ldots, m \tag{2}$$

$$X = \begin{vmatrix} x_1(1) & x_2(2) & \cdots & x_1(n) \\ x_2(1) & x_2(2) & \cdots & x_2(n) \\ \vdots & \vdots & \cdots & \vdots \\ x_m(1) & x_m(2) & \cdots & x_m(n) \end{vmatrix} \tag{3}$$

Before the evaluation of multi-objective decision-making, the calculation correlation series should be standardized and converted into dimensionless data.

$$x_j = \frac{(x_j - minx_j)}{(maxx_j - minx_j)} \tag{4}$$

For a reference data column x_0, comparison sequence x_i, the following relationship is used to express the degree of relevance between each object to be compared and the reference object.

$$\xi_i(k) = \frac{min_i min_k |x_0(k) - x_i(k)| + \zeta max_i max_k |x_0(k) - x_i(k)|}{|x_0(k) - x_i(k)| + \zeta max_i max_k |x_0(k) - x_i|} \tag{5}$$

where,$\xi_i(k)$ is called the correlation coefficient between x_i and x_0 regarding the k index. ζ is the resolution coefficient, $\zeta \in [0,1]$, usually take $\zeta \leq 0.5$.

2.4. Entropy-Weighted TOPSIS method

The Entropy-weighted TOPSIS method determines the weights through the Entropy weight method, which makes the allocation of weights more objective and reasonable, while the TOPSIS method provides an effective ranking mechanism, which can be used together to improve the effectiveness and accuracy of multi-attribute decision analysis.

Therefore, use the Entropy-weighted TOPSIS method to analyze the correlation between national demographics and the distribution of cybercrime.

In order to analyze the correlation between national demographics and the distribution of cybercrime, this paper selects **five** statistics as representative data of national demographics and organize them into the table 2.

Table 2. Representative Data—National Demographic Data

Representative Data	National Demographic Data
Fragile States Index	Degree of Political Stability
State Terrorism Index	Degree of National Security
Internet Usage	Internet Penetration
GDP	Level of Economic Development
Proportion of Government Spending on Education	Level of Education

Subsequently, this paper finds data on the demographics of the countries on the websites of the Peace Foundation, the Institute for Economics and Peace, and the World Bank, and compile them into the table 3.

Table 3 is a table of the selected data. "$\rho_1, \rho_2, \rho_3, \rho_4, \rho_5$"relatively represent Fragile States Index, State Terrorism Index, Internet Usage, GDP and Proportion of Government Spending on Education. "y"represents Evaluation Index.

Table 3. Representative Data

Country	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	y
Denmark	15.9	0	0.950	2142470.91	0.118	8.91
Netherland	20.6	0.577	0.970	1154361.31	0.116	8
Germany	24	2.782	0.920	4525703.9	0.092	8.76
Singapore	25.4	0	0.940	501427.5	0.101	7.96
United States	44.5	4.141	0.970	27720709	0.127	8.73
United Kingdom	40.8	2.737	0.027	3070214.1	0.950	8.44
Australia	19.6	1.475	0.950	407091.92	0.139	8.16
Canada	18.6	1.753	0.750	1728057.32	0.111	8.35
Japan	30.2	1.189	0.850	42044944.8	0.075	8.09
Isreal	51.5	0	0.920	513611.1	0.175	7.75
Georgia	69	0	0.820	30777.83	0.122	5.98
Bosnia and Herzegovina	71	0	0.830	27514.78	0.104	3.46
Tunisia	67.2	2.914	0.740	48529.6	0.181	4.72
Panama	47.7	0	0.740	83318.18	0.119	3.66
Cambodia	78.6	0	0.570	42335.65	0.157	2.63
Hungary	46.2	0	0.910	212388.91	0.104	6.27
Saudi Arabia	63.2	1.366	1.000	1067582.93	0.193	5.54

Honduras	78.1	0	0.600	34000.51	0.154	3.13
Russia	81.6	3.016	0.830	2021421.48	0.189	6.39
Myanmar	100	7.532	0.075	66757.62	0.440	2.22

Firstly, use the improved entropy method to assign weights to the evaluation indexes:

positive indicators: $a_{ij} = \frac{\rho_{ij} - min(\rho_j)}{max(\rho_j) - min(\rho_j)}$

negative indicators: $a_{ij} = \frac{\rho_{ij} - max(\rho_j)}{max(\rho_j) - min(\rho_j)}$

In this formula,a_{ij} represents the normalized value of the data metric, ρ_{ij} represents the original value of the indicator **i** for country **j**.

Secondly, calculate the entropy value:

$$e(a_j) = -\sum_{i=1}^{m}(a_{ij}\, ln\, a_{ij}) \tag{6}$$

$$e_j = \frac{e(a_j)}{ln\, m} \tag{7}$$

$$d_j = 1 - e_j \tag{8}$$

In this formula, m represents subjects of the study, n indicates the evaluation index, j=1, 2, ……, m; i=1, 2, ……, n.

Thirdly, solve the entropy weight:

$$w_j = \frac{d_j}{\sum_{i=1}^{n} d_i} \tag{9}$$

In this formula, j=1, 2, ……, m; i=1, 2, ……, n.

After calculating the entropy weight, this paper uses the TOPSIS method to calculate the Euclidean distance and proximity between the index and the positive and negative ideal solutions.

Firstly, construct a weighted decision matrix:

$$v = (v_{ij})_{m\times n}\begin{bmatrix} w_1 y_{11} \cdots & w_n y_{1n} \\ w_1 y_{m1} \cdots & w_n y_{mn} \end{bmatrix} \tag{10}$$

In this formula,w_j is the weight, y_{ij} is standardized data.

Secondly, calculate the positive and negative ideal solutions:

$$v^{+} = \{max(v_{ij}) \,|\, i = 1,2,\cdots,m\} \tag{11}$$

$$v^{-} = \{min(v_{ij}) \,|\, i = 1,2,\cdots,m\} \tag{12}$$

Thirdly, calculate the distance:

$$d_i^+ = \sqrt{\sum_{j=1}^{n}\left(v_{ij} - v_j^+\right)^2} \tag{13}$$

$$d_i^- = \sqrt{\sum_{j=1}^{n}\left(v_{ij} - v_j^-\right)^2} \tag{14}$$

In this formula, i=1, 2, ……, m.

Finally, calculate the proximity:

$$F_i = \frac{d_i^-}{d_i^- + d_i^+} \tag{15}$$

In this formula, i=1, 2, ……, m.

Sort the relative closeness from large to small, and finally choose the scheme with the larger value.

3. Result and Discussion

3.1. Global Distribution of Cybercrime

This paper has created a map of the distribution of cybercrime, a map of the national cybersecurity index, and a map of the risk index, and marked the countries with a high incidence of cybercrime.

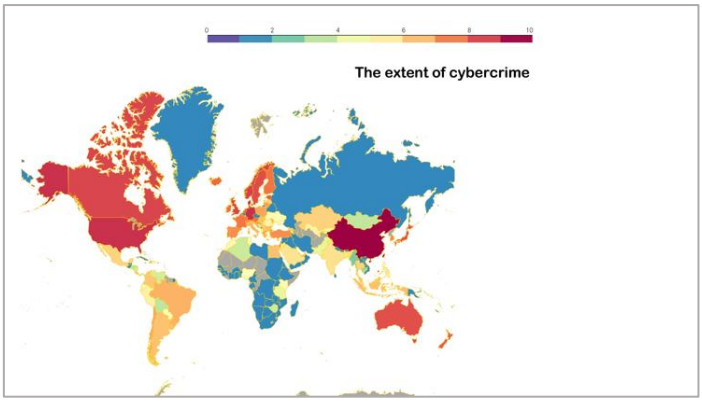

Figure 1. The Distribution of Cybercrime

As can be seen from the figure 1, there are significant differences in the number of cybercrimes in different countries, with the highest incidence of cybercrime in the United Kingdom, the United States, Spain, India, Brazil. It reflects that these countries usually have a higher level of economic development, Internet penetration, or poor cybersecurity management mechanisms and cybersecurity awareness.

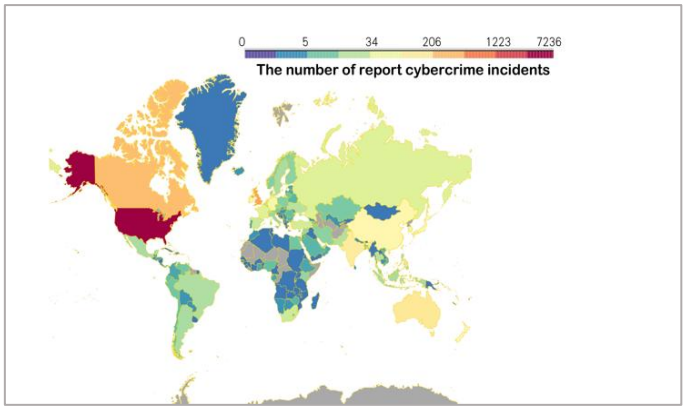

Figure 2. Distribution of Cybercrime Reporting Rates

As you can see from the figure 2, cybercrime has been successfully carried out in countries such as Myanmar and Cambodia, and has been stopped in countries such as Denmark and Germany. Some European and American countries, such as the United Kingdom, the United States, Spain, usually have high rates of reporting and prosecuting cybercrimes. It is not difficult to find that some developed countries are willing to open and transparent cybercrime incidents in order to enhance the public's awareness of cybersecurity and continuously improve the cybersecurity management mechanism. However, in some developing countries, cybercrime incidents may go viral and under-reported.

3.2. Evaluation of National Cybersecurity Policies

This paper calculates the grey correlation coefficient to determine the degree of relevance of each comparator to the reference object. The results are shown in Figure 3, Figure 4, and Table 4.

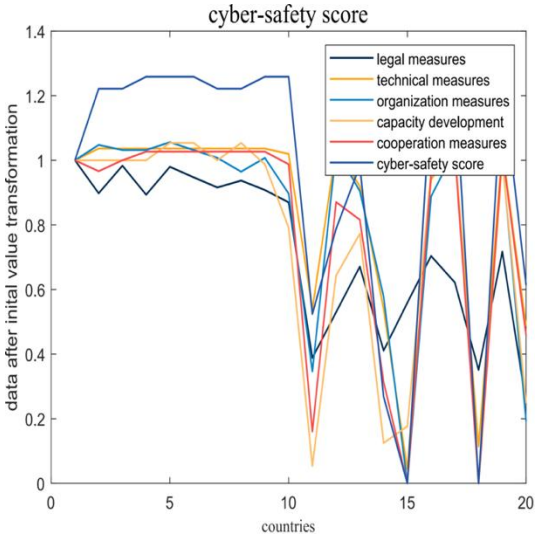

Figure 3. National Cybersecurity Score Graph

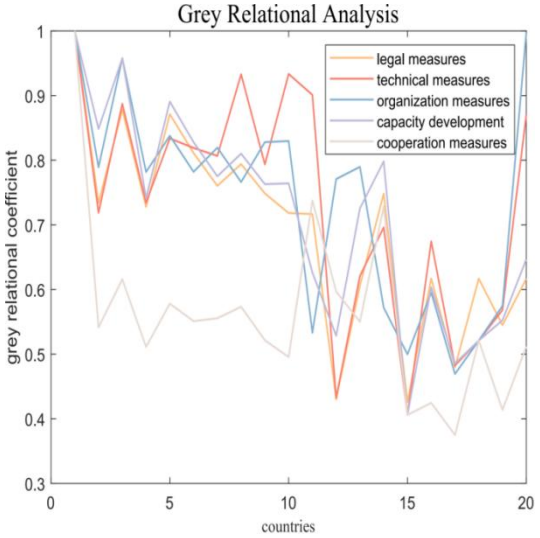

Figure 4. Grey Relational Analysis Graph

Table 4. Correlation coefficient

x_1	x_2	x_3	x_4	x_5
0.6923	0.7316	0.7358	0.7136	0.5604

As a result, here comes to the following conclusions.

Legal, technology, organization and capacity development have a great impact on the effectiveness of national cybersecurity policies. However, cooperation has little impact.

Because the data of technical measures and organizational measures have small differences, this paper thinks technology and organization have the biggest impact.

3.3. Correlation Analysis of Cybercrime Distribution

This paper uses the TOPSIS method to derive the overall level ranking of the countries in the table 1 and compared it with the Cyber-Safety ranking, as shown in the figure.

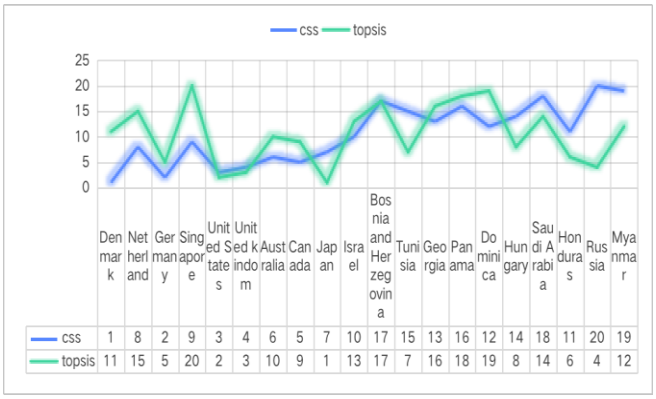

Figure 5. National Rating Rank of CSS and TOPSIS SCORE

From figure 5, it can be found that the trends of the three curves have a very high similarity, and the correlation coefficient of the two curves is high, indicating that the statistics of each country do have a strong correlation with the level of cybersecurity of the country.

Subsequently, study the relationship between various statistics and Cyber-Safety Scores, and calculate the between the rankings and Cyber-Safety Scores, as shown in the table 5.

Table 5. Pearson correlation coefficient

ρ_1	ρ_2	ρ_3	ρ_4	ρ_5
0.166749	-0.842348	0.700003	0.828571	-0.529323

There is a strong positive correlation between GDP, and in combination with the conclusion of the second question, countries with high GDP mean more money for infrastructure construction, such as building data centers. At the same time, in terms of R&D investment, high GDP also means that more funds are invested in ICT R&D, so as to promote the update and iteration of cybersecurity measures, optimize the cybersecurity landscape, and provide important technical support for cybersecurity. In addition, high-GDP countries often show significant advantages in terms of investing in education resources to cultivate cybersecurity talents. Objectively, countries with high GDP usually have a relatively level of high income, so the information market is larger, and the introduction of corresponding policies on the basis of such rigid needs to ensure the information experience of users will naturally become the focus of government attention. The above shows that GDP does support the level of cyber security at the technical level, and the above conclusion is indeed the highest among the five influencing factors, so it is consistent with the theory.

The next is internet penetration, which has a slightly lower correlation than GDP. The higher the Internet penetration rate, the higher the demand for Internet security and the government also attaches more importance to it, and more devices are connected to the network, which provides the basis for the wide application of network security technologies, such as firewall intrusion detection technology, which is conducive to improving the overall network security protection capability, so as to gain an advantage in the GCI technical maintenance score. In addition, in countries with high Internet penetration, the government needs to establish a more complete management system, and even

management agencies, such as CISA in the United States and BFI in Germany, have actually improved the level of national cybersecurity from the organizational dimension.

This is followed by the State Terrorism Index and the Integrity State Vulnerability Index, which represent some overlapping aspects. This paper analyses them together.

First of all, the State Terrorism Index represents the level of terrorism in the country, and the higher the level of terrorism, the more often information equipment and the like will be destroyed. As a result, investors are more cautious about investing in ICT technology, which will directly affect the country's cybersecurity technology update, and the inability to sell ICT products will directly affect the country's Internet penetration rate.

The social unrest and security threats caused by the high terrorism index, such as the large-scale telecom fraud in Myanmar, directly increase cybersecurity risks. In addition, it also increases the difficulty of government policy implementation, and the low rate of government policy implementation often has little effect on reducing the cybercrime rate and improving cybersecurity. The degree of policy implementation is directly related to the scores of the above five elements, which have been verified above to have a greater correlation with network security, so the terrorism index shows a huge negative correlation with the network security ranking, while the complete and vulnerable country index is similar, poor political stability, and political turmoil will also cause the increase of network security risks.

Moreover, the proportion of education expenditure can reflect the importance of education to a certain extent, but its negative correlation with GDP, that is, the higher the proportion of education expenditure in countries may be due to insufficient government funds, and education is an important expenditure of national governments, so the proportion is high. Therefore, it has a negative impact on the cultivation of national cybersecurity talents and the subsequent development of cybersecurity technology. It is at a significant disadvantage in terms of technology and national capacity building, and the education level field it represents has a negative relationship with itself. Countries with a high level of education have a clear advantage in terms of technology, organizational ability, etc., which does not contradict the theory presented in this work.

In summary, the statistical analysis of the relationship between the demographic data of many countries and the five factors discussed above proves the correctness of the conclusions of this work, and also confirms the view of this work in terms of the importance of the influencing factors.

4. Conclusion

This paper innovates an in-depth analysis of the global distribution of cybercrime and the evaluation of the effectiveness of national security policies through Grey Relational Analysis and Entropy-Weighted TOPSIS method. The study finds that there are significant differences in cybersecurity incident reporting and transparency between developed and developing countries, reflecting the uneven distribution of cybercrime globally. The effectiveness of national security policies is closely related to technical and management factors, which indicates that improving technical capabilities and management levels is the key to improving the effectiveness of cybersecurity policies. Based on the research results, this paper puts forward targeted policy optimization suggestions, including strengthening international cooperation, enhancing public awareness of cybersecurity, and improving laws and regulations. Although this study reveals the distribution characteristics and policy effectiveness of global cybercrime to a certain extent, there are still some problems such as data acquisition limitations and method applicability. Future research can further expand data sources, optimize analysis methods, and delve into cybersecurity policy practices in specific countries or regions.

References

[1] ROSENZWEIG P. The international governance framework for cybersecurity[J]. Canada-United States Law Journal,2012(2):405-432.

[2] Jie CHEN, Lei ZENG. Practical challenges and countermeasures of global governance of cybercrime[J]. Journal of Southwest University (Social Sciences),2021,47(04):48-58+228. DOI:10.13718/j.cnki.xdsk.2021.04.005.

[3] Wenlong Zhang. Challenges and Responses: Crime Governance in the Context of Globalization[J]. Academic exchanges,2016,(09):96-102.

[4] Su Jiang. New mechanism of international law to combat cybercrime[J]. Jurisprudence,2022, (11):45-59.

[5] Yan Li. The dilemma and path design of the construction of an international legal mechanism for cybercrime[J]. Journal of Yunnan University for Nationalities (Philosophy and Social Science),2019,36(06): 135-144.DOI: 10.13727/j.cnki.53-1191/c.2019.06.019.

[6] Keying An. China's participation in the international governance of transnational cybercrime[J]. Journal of Yunnan University for Nationalities(Philosophy and Social Science),2019,36(03):155-160. DOI:10.13727/j.cnki.53-1191/c.2019.03.025.

[7] Q. A. Al-Haija and L. Tawalbeh. Autoregressive Modeling and Prediction of Annual Worldwide Cybercrimes for Cloud Environments[C]//*2019 10th International Conference on Information and Communication Systems (ICICS)*, Irbid, Jordan, 2019, pp. 47-51.DOI: 10.1109/IACS.2019.8809125.

[8] I. Hameed and S. A. A. Naqvi. An Analysis of the factors affecting Cybercrime against individuals in Pakistan[C]//2021 15th International Conference on Open Source Systems and Technologies (ICOSST), Lahore, Pakistan, 2021, pp. 1-6.DOI: 10.1109/ICOSST53930.2021.9683986.

[9] S. Batra, M. Gupta, J. Singh, D. Srivastava and I. Aggarwal. An Empirical Study of Cybercrime and Its Preventions[C]//2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), Waknaghat, India, 2020, pp. 42-46.DOI: 10.1109/PDGC50313.2020.9315785.

[10] Hengyang Li. Adjustment and future challenges of the Trump administration's cybersecurity policy[J]. American Studies,2019,33(05):41-59+5-6.