SCIENCE CHINA
Information Sciences

• LETTER •

February 2020, Vol. 63 129303:1–129303:3
https://doi.org/10.1007/s11432-019-9841-4

A data analysis of political polarization using random matrix theory

Hui CHEN[1], Xiaofeng TAO[1*], Na LI[1] & Zhu HAN[2]

[1]*National Engineering Lab for Mobile Network Technologies, Beijing University of Posts and Telecommunications, Beijing* 100876*, China;*
[2]*Electrical and Computer Engineering and Computer Science, University of Houston, Houston* 77004*, TX, USA*

Received 21 January 2019/Revised 20 March 2019/Accepted 25 March 2019/Published online 12 September 2019

Citation Chen H, Tao X F, Li N, et al. A data analysis of political polarization using random matrix theory. Sci China Inf Sci, 2020, 63(2): 129303, https://doi.org/10.1007/s11432-019-9841-4

Dear editor,

Data analysis science is currently a hot topic in industry and academia due to ubiquitous and rapidly growing data [1, 2]. Data analysis technology has been applied in a variety of fields and has provided efficient and profitable data-driven decisions. For instance, since the beginning of the 20th century, large-scale opinion polls have been conducted before each U.S. presidential election to analyze or predict election results. The main method traditionally used to analyze political-related problems is probability statistics, e.g., regression analyses and causal models. To make the analysis tractable, these statistical methods usually convert high-dimensional data into univariate or multivariate sample statistics.

Generally, traditional methods, such as the law of large numbers, require that the number of samples is much larger than that of the characters, which is not always fulfilled, and may lose some of the properties of the original data when applied to high-dimensional data. For instance, in the DNA sampling problem, the number of genes is almost innumerable, and the sampling number cannot be much larger than the gene number, which results in the traditional statistical methods being unable to derive an accurate result [3, 4]. Hence, random matrix theory (RMT), which combines statistics and matrices and can reveal the structures of the data, plays an important role in this type of problem [5]. In a nutshell, this study analyzes the political polarization data analysis problem using RMT.

Data collection. To reveal U.S. biparty polarization properties, the Pew Research Center [6] conducts a large number of investigations, collecting a mass of public data and providing a foundation for further in-depth studies. Inspired by these fundamental studies, we try to study these data to derive some deep insights. We obtain four datasets of Democrats and Republicans from the Pew Research Center for the years 2004, 2011, 2014, and 2015, including N ($N = 10$) carefully designed political questions that reflect the political views of the respondents [6], where the number of respondents n is on the level of several hundreds or thousands per year.

The responses are coded as $\{-1, 0, 1\}$, representing the conservative view, neutral view and liberal view, respectively. Since the ratio $c = N/n$ is critical for RMT analysis, we choose $n = 600$ respondents equally from different parties and different years, which is the minimum data dimension of the four datasets. Although the amount of data is limited, the approximating method of RMT is highly accurate [3].

Data modeling. Next, we provide an intuitive matrix structure hypothesis of the data, which is verified by the following real data analysis. The data are organized as a two-dimensional matrix, where the rows represent different questions, the columns represent different respondents, and the

* Corresponding author (email: taoxf@bupt.edu.cn)

data represents public views of different respondents; the analysis matrix is given by

$$\boldsymbol{Y}_{N,n}^{p} = \sigma \boldsymbol{X}_{N,n}^{p} + \boldsymbol{U}_{N,n}^{p}, \tag{1}$$

where $p \in \{p_1, p_2\}$ stands for two different parties, the entries of matrix $\boldsymbol{X}_{N,n}^{p}$ are independent and identically distributed, σ is the variance, and $\boldsymbol{U}_{N,n}^{p}$ is a constant matrix. From this point forward, matrix $\boldsymbol{X}_{N,n}^{p}$ represents the random responses, and $\boldsymbol{U}_{N,n}^{p}$ represents the inherent political perspectives.

To obtain a good perspective of the experimental result, the raw data matrix needs to be preprocessed with the following operation: $\hat{X} = \frac{1}{L}\sum_{i=1}^{L} X_i$, which means that the raw data matrix is randomly divided by column into L same dimension matrices X_i, added together and averaged by L, denoted as L trials. Note that the sampling is conducted in a typical time of a year, and the respondents are randomly chosen from the public, which causes the samples to be independent from one another, and the form of the raw data eigenvalue distribution remains unchanged. This operation fulfills the theory of RMT and can provide us a more explicit result.

RMT spectral analysis. The numerical distribution of the empirical spectral distribution (e.s.d.) is depicted in Figure 1(a), where the histograms are the estimated spectral distribution, and the solid lines are the e.s.d. (the theory of e.s.d. can be found in [3]). This figure illustrates that the e.s.d. fits the real raw data very well, where the small gaps are caused by differences between our model and reality. Generally, if the raw data are impacted by different factors, the spectral distribution of their covariance can be separated into different segments, which is termed a spiked model in RMT. In our research scenario, the largest eigenvalue interval represents the spectral distribution of the constant matrix in (1), which in turn represents the degree of polarization for our raw data. In addition to this intuitive analysis, we adopt a metric denoted average entropy to evaluate the polarization of the raw data.

Polarization evaluation. Based on the data structure drawn in previous sections, we proposed a reasonable data evaluation metric. Based on (1), the difference model of opinions is widely used in social influence analysis and is given by

$$\boldsymbol{Y}_{N,n}^{p_1,p_2} = \boldsymbol{Y}_{N,n}^{p_1} - \boldsymbol{Y}_{N,n}^{p_2} = \sigma^{p_1,p_2}\boldsymbol{X}_{N,n}^{p_1,p_2} + \boldsymbol{U}_{N,n}^{p_1,p_2}, \tag{2}$$

where $(\sigma^{p_1,p_2})^2 = (\sigma^{p_1})^2 + (\sigma^{p_2})^2$, and σ^{p_1} and σ^{p_2} are the variances of the two data sources. Matrix $\boldsymbol{U}_{N,n}^{p_1,p_2}$ is a constant matrix that represents the polarization of the two parties, while $\boldsymbol{X}_{N,n}^{p_1,p_2}$ has unit variance and zero mean entries.

To evaluate the difference between these two datasets, we propose an average entropy model as follows:

$$I^{p_1,p_2} = \frac{1}{N}\log_2 \det\left(\boldsymbol{I}_N + \frac{1}{n}\frac{\boldsymbol{U}_{N,n}^{p_1,p_2}\boldsymbol{U}_{N,n}^{p_1,p_2\mathrm{T}}}{(\sigma^{p_1,p_2})^2}\right), \tag{3}$$

where T is the matrix transpose. The model is a measure of the uncertainty in the field of information theory with units of bits. The polarization measure is a mapping $M : \mathbb{S} \to \mathbb{R}^+$, where $\mathbb{S}$ is the multidimensional support. The ratio in the brackets represents the ratio between the intrinsic gap and the random noise of the two parties. This entropy model can reflect how much information we can derive from the polarization of two different parties, where polarization is primarily reflected by the gap between two constant masses, and the other part can be viewed as noise.

To derive the average entropy of the model, we should first calculate the unknown parameters in (3) from the raw data. We adopt two techniques to estimate the parameters, namely, the large-dimensional approach (LDA) and RMT.

LDA estimation. The LDA assumes that the respondents are numerous to achieve great diversity, i.e., $n \to \infty$ and $c = \frac{N}{n} \to 0$. Therefore, the largest eigenvalue of matrix $\boldsymbol{Z} \triangleq \frac{1}{n}\boldsymbol{Y}_{N,n}^{p_1,p_2}\boldsymbol{Y}_{N,n}^{p_1,p_2\mathrm{T}}$ is almost surely composed of $\sigma^2 + \eta$, where $\eta = \mathbb{E}\big(\frac{1}{n}\boldsymbol{U}_{N,n}^{p_1,p_2}\boldsymbol{U}_{N,n}^{p_1,p_2\mathrm{T}}\big)$, which represents the empirical effect, and σ is the covariance. Parameter η is used to derive the useful factor of the data, while the other eigenvalues of this matrix almost surely converge to value σ^2.

In practice, if σ^2 is unknown, we can use the $N-1$ smaller eigenvalues to estimate the covariance. Hence, the estimated result η can be derived by $\hat{\eta}_1 = \lambda_N - \hat{\sigma^2}$, where $\hat{\sigma^2} = \frac{1}{N-1}\sum_{i=1}^{N-1}\lambda_i$, and $\lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_N$ are the increasingly ordered eigenvalues of the real data matrix $\boldsymbol{Z}$.

Moreover, when the ratio $c = \frac{N}{n} \to 0$, the two eigenvalue sets in Figure 1(a) can be derived by [7] $[\lambda_1^-, \lambda_1^+] = [\sigma^2 - \mathcal{O}(1/n), \sigma^2 + \mathcal{O}(1/n)]$, $[\lambda_2^-, \lambda_2^+] = [(\lambda_N + \sigma^2) - \mathcal{O}(1/\sqrt{n}), (\lambda_N + \sigma^2) + \mathcal{O}(1/\sqrt{n})]$, where λ_N is the largest eigenvalue of matrix $\boldsymbol{Z}$. As $n \to \infty$, while $\mathcal{O}(1/n)$ and $\mathcal{O}(1/\sqrt{n}) \to 0$, we can derive the same result $\hat{\eta}_2 = \lambda_N - \hat{\sigma^2}$.

RMT estimation. By using RMT, we can keep the sampling matrix unchanged while quantifying the mutual differences between the respondents of two parties. When the condition $c = \frac{N}{n} \to 0$ is not satisfied, the constant value η can be estimated using RMT. First, we need to calculate the moment of the elements in (2). The kth moment of a matrix $\boldsymbol{A}$ is defined as $t_{\boldsymbol{A}}^k = \mathrm{E}\left[\mathrm{Tr}(\boldsymbol{A}^k)\right] = \int \lambda^k \mathrm{d}F_{\boldsymbol{A}}(\lambda)$, where $\mathrm{Tr}(\cdot)$ denotes matrix trace, and $\mathrm{E}(\cdot)$ denotes

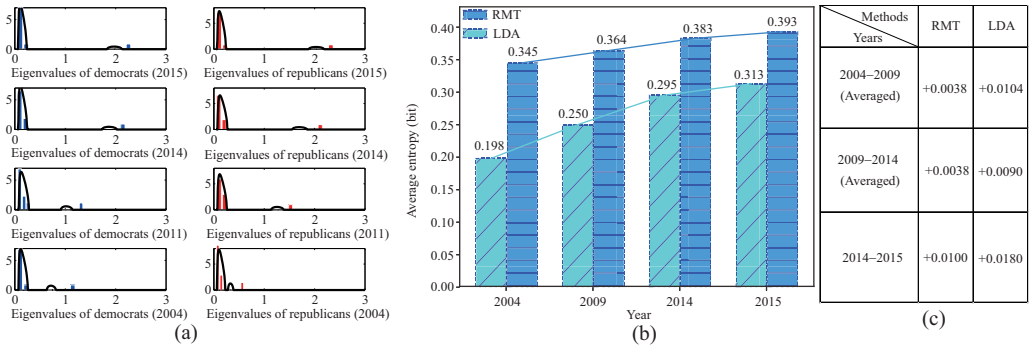

Figure 1 (Color online) Results. (a) Eigenvalue distributions of Democrats and Republicans with the number of questions $N = 10$ and the number of respondents $n = 100$; (b) entropy estimations using RMT and LD; (c) list of the entropy increments.

expectation.

Considering the situation that the deterministic part in the raw data is of rank 1, which can be derived from Figure 1(a), the RMT estimation procedure is given by the following: the first moment of $\boldsymbol{Z}$ is $h = \mathrm{Tr}(\boldsymbol{Z}) = \sum_{i=1}^{N} \lambda_i$; the estimated first moment h_f of matrix $\boldsymbol{H} \triangleq \frac{1}{n}\boldsymbol{U}_{N,n}^{p_1,p_2}\boldsymbol{U}_{N,n}^{p_1,p_2\mathrm{T}}$ is then $h_f = h - \sigma^2$; and, finally, the estimated eigenvalue of $\boldsymbol{H}$ is $\lambda_N = \mathrm{Tr}(\boldsymbol{H}) = h_f$. Therefore, the estimated value is $\hat{\eta}_3 = h - \sigma^2 = \sum_{i=1}^{N} \lambda_i - \sigma^2$.

In the general case, the value $\hat{\eta}_3$ derived by the RMT is larger than the value $\hat{\eta}_1$ and $\hat{\eta}_2$ derived by the LDA, which can be seen directly from the results of these two approaches. In the special case, when the ratio $c = \frac{N}{n} \to 0$ is satisfied, the LDA method converges to the RMT method.

Results. Figure 1(b) illustrates the estimated entropy in (3) using the LDA and RMT. The histogram with line segment is derived by RMT, while the histogram with slash is derived by the LDA. The entropy increased over the years, which means that the average difference between the two parties has expanded. The values of these two histograms slightly differ, but the increasing entropy trend remains the same because of the sampling limitation of the raw data.

Figure 1(c) lists the entropy increments derived by the two estimation methods. The increments of every year are positive, which means that polarization is increasing. The evaluated polarization value is used to describe the degree of the ideological difference. If the evaluated value increases, the voice of opposition between these two parties increases. By contrast, as shown by the RMT method, the increments in the second and third rows are almost the same, while that in the fourth row is almost doubled, which means that the polarization rate remained constant over the past few years but doubled in recent years.

Conclusion. This study adopts RMT to analyze political polarization data. Based on the real data experimental result, a spiked model is adopted to analyze the problem. Then, an average entropy metric is proposed to calculate the differences between two parties. Finally, the LDA and RMT are used to evaluate the entropy. The proposed method not only provides a clear perspective into the data structure but also can be used to solve the problem when the law of large numbers is not satisfied. Additionally, we provide a time-efficient and cost-effective mathematical tool for analyzing social networks, which opens a new door for future comprehensive mathematical research.

Acknowledgements This work was supported in part by National Science Fund for Distinguished Young Scholars (Grant No. 61325006), in part by National Nature Science Foundation of China (Grant No. 61631005), in part by Beijing Municipal Science and Technology Project (Grant No. Z181100003218005), and in part by 111 Project of China (Grant No. B16006).

References

1 Bello-Orgaz G, Jung J J, Camacho D. Social big data: recent achievements and new challenges. Inf Fusion, 2016, 28: 45–59
2 Cui Q, Gong Z, Ni W, et al. Stochastic online learning for mobile edge computing: learning from changes. IEEE Commun Mag, 2019, 57: 63–69
3 Couillet R, Debbah M. Random Matrix Methods for Wireless Communications. Cambridge: Cambridge University Press, 2011
4 Bai Z, Silverstein J W. Spectral Analysis of Large Dimensional Random Matrices. New York: Springer, 2010
5 Yang Y, Shen F, Huang Z, et al. Discrete nonnegative spectral clustering. IEEE Trans Knowl Data Eng, 2017, 29: 1834–1845
6 Michael D, Carroll, D, Jocelyn K, et al. Political polarization in the American public. Pew Research Center, 2014. http://www.people-press.org/2014/06/12/political-polarization-in-the-american-public/
7 Vallet P, Loubaton P, Mestre X. Improved subspace estimation for multivariate observations of high dimension: the deterministic signals case. IEEE Trans Inform Theor, 2012, 58: 1043–1068