Application: Digitization Tutorial

Guangrui(Kayla) LI
Department of Information Systems, Business Statistics and Operations Management
School of Business and Management, Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
Email: glian@connect.ust.hk
Phone: +852 5222-7619

Dear Dr. Goldfarb and Dr. Greenstein,

I am writing to apply for the tutorial on the Economics of Digitization organized by NBER. I am currently a 4th-year-PhD-candidate in Information Systems at the Hong Kong University of Science and Technology. I respectfully submit this letter of application, for I believe my training in online platform and big data analytics make me well qualified to the tutorial meeting, and the tutorial would help me a lot to further develop as a junior researcher.

My education experience includes 2-year (2014-2016) study as a Mphil Student, and 4-year (2016-current) study as a PhD candidate majored in Information Systems at the Hong Kong University of Science and Technology. I've taken courses related to the economics of digitization during my study, include: doctoral seminar on Digital Innovation, Platforms and the Sharing Economy, doctoral seminar on Healthcare IS and Big Data in IS Research, doctoral seminar on Cybersecurity and Social Network Analysis, IS Research for Societal Good, Econometrics, Microeconomic Theory, etc.

My research is motivated by the prevalence of social media platforms and online communities in our daily life. Given the wide reach of these platforms, even minor change on the platform design/policy would lead to big societal impacts. In addition, enormous number of platform users generate huge amount of data, which brings statistical challenges when using such "big data". My research tries to answer the above questions: (1) the societal impacts of online platform governance strategy and (2) the statistical challenge brought by big data. In the following section, I describe these two streams, research objectives within each stream, projects and future work related to each stream.

The overarching goal of the first stream is to understand the impact of online platform governance strategy on various societal issues. Such an understanding would help platforms and governments avoid undesirable societal outcomes. I have *two* broad research objectives in this research stream.

The first objective is to understand the societal impacts of the behavior of artificial intelligence systems on social media platforms. Machines powered by artificial intelligence increasingly mediate our social, cultural, economic and political interactions. Understanding the behavior of artificial intelligence systems is essential to our ability to control their actions, reap their benefits and minimize their harms. I focus on one specific artificial intelligence system – the ranking algorithm and its impacts in a specific context – social media platforms. My first project examines whether the implementation of ranking algorithm on social media platform creates filter bubble – less exposure to diverse-content and attitude-challenging viewpoints. The existence of filter bubble

could have a negative effect on the public sphere and the democratic opinion forming process and contributes to social sorting, political polarization and discriminatory practices. We relied on a natural experiment on Sina Weibo, where ranking algorithm was implemented at June 2012, and conducted Difference-in-differences (DID) analysis to identify the existence of filter bubble. We found significant evidence for the creation of filter bubble: the number of unique topics appear in users' posts decreased greatly (34.8%) even after controlling for the number of posts, indicating users have less access to diverse content; the sentiment of users' posts became less negatively, indicating users are less likely to encounter attitude-challenging viewpoints. Furthermore, we found the existence of filter bubble would discourage users' contribution to the platform by 35.1%, resulting in adverse effects on the development of platforms. The second project examines the impact of ranking algorithm on attention inequality on social media platforms. Attention is the currency that facilitates transactions between producers (authors) and consumers (readers) on the Internet. In this sense, the distribution of attention corresponds to the distribution of income, raising concerns about attention inequality that can have negative consequences similar to the ones that are associated with income inequality. Similarly, we employed DID design and found the ranking algorithm (1) suppresses fan attraction for unpopular users but helps to attract more interactions received by their posts at the user level, (2) helps to attract more discussion for unpopular topics at the topic level.

The second objective is to understand the impact of real-name identification policy on users' piracy behavior. Piracy is socially undesirable due to its possible negative impact on legitimate sales, more than $50 billion is lost every year due to online piracy[1]. It's important to find effective ways to combat piracy. I investigate whether the real-name identification policy could deter online users from piracy behavior, as the policy requires all users to verify their mobile phone number to enjoy the services provided by different websites, which increases the certainty of punishment for piracy. By using a natural experiment on Baidu Tieba and considering Pinggu BBS as the control group, I conducted a Difference-in-differences (DID) analysis to assess the impact of real-name identification policy. Interestingly, the findings suggest the policy only decreases piracy-link sharing on online forums within a very short time window – around 10 days, while significantly increasing the sharing in the long run.

The second research stream is to address the statistical challenges caused by the prevalence of extremely large sample in academic researches. This project tries to deal with the 'deflated *p* value' problem, and the role of prediction versus explanation, in big data research. The 'deflated *p* value' problem describes the phenomenon that the *p* value decreases as sample size increases, and most variables will be significant in the big data context, raising concerns of the practical meaning of *p* value. Although researchers have started reporting the effect size (θ) –the sensitivity of a dependent variable to changes in an independent variable – along with p values, this univariate effect size method cannot be generalised to a multiple-variable scenario. E.g. socioeconomic status is determined by the three variables of income, education and occupation, and categorical factors are usually operationalised as a set of dummy variables. In this context, the effect size at the factor level would be of more interest, but cannot be directly measured by the univariate effect size method by summation (i.e., $\theta_1 + \theta_2 + \theta_3$). This project proposes multivariate effect size measures the joint effect of all variables as which the factor is operationalized, thus overcoming the

[1] Gaille, Brandon. "35 Startling Internet Piracy Statistics." *BRANDONGAILLE Small Business & Marketing Advice*, 23 May.2017, https://brandongaille.com/34-startling-internet-piracy-statistics/

disadvantage of the traditional effect size method (θ), i.e., that it can only be applied at the variable level. The TE-based statistic asymptotically follows the chi-square distribution, but because factors can be operationalized as different numbers of variables, a direct comparison is possible only if the factors have the same number of degrees of freedom. This project proposes a quantile-matching based transformation method to address this problem. This method has three advantages: (1) Feasibility – it has a one-to-one corresponding relationship with *p* values under perfect matching, and thus inherits the merits of *p* values in comparisons, but can be applied to scenarios where the *p* value (quantile) method fails due to the use of big data. (2) Flexibility – the quantile-matching transformation can be performed on a full distribution or a specific interval, providing tremendous flexibility in real applications such as the big data context (where most *p* values approach 0, such that matching on the right tail of the distribution would be more helpful). (3) Fast calculation – it is very easy to implement and calculate. Furthermore, an absolute multivariate effect size statistic is developed based on the transformed multivariate effect size, to facilitate the drawing of conclusions without comparison. The above methods are tested on a very large dataset (n = 6,230,253) and it could effectively differentiate factors with various effect sizes in the big data context. The result obtained from this method are compared with that of ROC analysis and consistent results are found: explanatorily influential factors are usually also predictively influential in a large sample scenario.

I believe my expertise is well suitable for the tutorial on the Economics of Digitization. And I look forward to hearing back from you!

Yours sincerely,
Guangrui(Kayla) Li