Editing the truth

Measuring government bureaucratic quality using anonymous Wikipedia edits.

Patrick, J , Healy*

Department of Economics, Monash University, patrick.healy1@monash.edu

Klaus, Ackerman

Department of Econometrics and Business Statistics, Monash University, Klaus.Ackermann@monash.edu

Simon, D, Angus

Department of Economics, Monash University, simon.angus@monash.edu

Paul, A, Raschky

Department of Economics, Monash University, paul.raschky@monash.edu

Weijia, Li

Department of Economics, Monash University, Weijia.Li@monash.edu

Nathan, Lane

Department of Economics, Oxford University, nahaniel.lane@economics.ox.ac.uk

Satya, Borgohain

Faculty of Information Technology, Monash University, satya.borgohain@monash.edu

Cynthia, Huang

Department of Econometrics and Business Statistics, Monash University, Cynthia.huang@monash.edu

CCS CONCEPTS • Applied Computing • Law, social and behavioral science • Economics

Additional Keywords and Phrases: Political Economy, State Capacity, Wikipedia, Machine Learning

1 EXTENDED ABSTRACT

Information plays a crucial role in how states govern society, and the adoption of digital technology has dramatically altered what governments can learn and how they can use knowledge for political intervention. Taylor Owen discusses how there is an "arms race, between people who are empowered through free, secure communication and governments that want to monitor and limit this communication" [Owen 2015)]. Wikipedia is the largest and most heavily used reference work in history [The Economist 2021]. It plays a critical role in the struggle for free knowledge as it is decentralised and can be freely accessed and edited by any person. Research has shown that Wikipedia guides academic research [Black 2008, Thompson 2018]; informs economic decision making [Hoopes 2015]; influences political views and voting behaviour [Agarwal 2020, Yasseri 2016]; and is widely used for work and education [Singer, 2017, Lemmerich 2019].

Wikipedia is a critical piece of knowledge infrastructure in society made up of the public good creation of volunteers. It plays a crucial role in the ability of citizens to separate truth from fiction [Benkler 2002, Benkler 2020]. [Guriev 2019] write that states can utilise these new technologies to develop more sophisticated methods of controlling information and media. While [Zuboff 2019] argues that we are facing epistemic inequality in the digital world, as the internet gives governments and corporations the increased potential to control the flow of information and learning as well as the power to influence societies decision making. This raises the question of whether states are editing wikipedia, what role they take in knowledge production and whether they are editing the truth.

The emergence of the state as an actor in the digital realm through the adoption of new digital technologies can be termed the rise of the digital state'. This new dynamic raises serious questions about an enlarged state capacity to influence Wikipedia as a public good, and indeed whether increases in technological efficiency are allowing states to actively manipulate information and edit the truth with far greater reach and immediacy. Empirical literature examining state capacity has predominantly looked at traditional non-digital measures of capacity including military, tax collection, legal enforcement, regulation of markets and the collection & dissemination of information [e.g., Bäck 2008, Besley 2011, Bardhan 2016, Muralidharan 2016, Lee 2017] There is a rich empirical literature which has sought to assess at government institutional quality [La Porta 1999, Kaufmann 2008]. However there are limited multi-country empirical studies, partly due to the limited amount of relevant data available [Olsen 2006, Egeberg 2012] and many recent empirical studies have either been time consuming, [Chong 2014] or relied on data from costly surveys [Suzuki 2019]. In this paper we show the capacity of states to either positively or negatively intervene in Wikipedia as a public good and provide initial evidence that the quality of government edits serves as an effective proxy for state bureaucratic professionalism. We achieve this through: 1) creating a novel dataset of government owned IP addresses which we use to create our measure of digital state capacity to manipulate and disseminate information; and 2) Testing the correlation of our measure of digital state capacity with other determinants of state capacity.

To create our dataset of government owned IP addresses [DB-IP 2019]. Large companies and government entities are listed as owners of their own IP blocks in the db-ip IP address location database. We have utilised this ownership information to classify government vs non-government IP addresses at the sub-national level. We then utilise Amazon Web Services Lambda tool to successfully extract entity types and entity descriptions from google knowledge panels for 25 percent of approximately 4 million unique owners per month the IP location database for 30 months from January 2019 to July 2021. After that we query Wikidata to build a fine-grained dataset of entity types and entity descriptions for government and other non-government classes which we use as training data for a support vector machine which achieves 89% accuracy when classifying government vs non-government entities in DB-IP ownership information matched to google knowledge panels. Next we download and parse anonymous Wikipedia edits from publicly available dumps [Wikimedia 2021]. We match our classified dataset of government IP addresses to the IP addresses of anonymous editors. Finally, we utilise the ORES API developed by [Halfaker 2020] to predict the likelihood of each government edit (1) being made in good faith with the intention of improving the quality of the article and (2) improving or damaging the quality of the article being edited. We use these prediction scores to create our measure of bureaucratic quality. We identify 46,971 Edits from 702 Government Entities in 83 Countries to 30 language versions of Wikipedia. Using a threshold of 90% in the ORES vandalism detection tool, we find the rate of vandalism detected in our government edits dataset is 9.6 percent.

Using our novel datasets of the quality of government edits to Wikipedia as a proxy for state digital capacity and bureaucratic professionalism. We follow [Chong 2014] testing for the correlation between our proxy for digital state capacity and existing determinants of state capacity. Our initial findings show that high quality government edits are positively related to indicators such as the education & cyber security skill level of public servants and the proportion of female public servants while low edits are negatively related with measures such as the proportion of female public servants and government transparency. To the best of our knowledge our study provides the most complete set of government edits to Wikipedia. The empirical analysis provides an important first step towards measuring state capacity in the digital space.

REFERENCES

Agarwal, P., Redi, M., Sastry, N., Wood, E., & Blick, A. (2020, July). Wikipedia and Westminster: Quality and dynamics of Wikipedia pages about UK politicians. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media* (pp. 161-166).

Bäck, H., & Hadenius, A. (2008). Democracy and state capacity: exploring a J-shaped relationship. *Governance*, *21*(1), 1-24.

Bardhan, P. (2016). State and development: The need for a reappraisal of the current literature. *Journal of Economic Literature*, *54*(3), 862-92.

Benkler, Y. (2002). Coase's Penguin, or, Linux and" The Nature of the Firm". *Yale law journal*, 369-446.

Benkler, Y. (2020). From Utopia to Practice and Back. *Wikipedia@ 20: Stories of an Incomplete Revolution*, 43-54.

Besley, T., & Persson, T. (2011). *Pillars of prosperity*. Princeton University Press.

Black, E. W. (2008). Wikipedia and academic peer review: Wikipedia as a recognised medium for scholarly publication?. *Online Information Review*.

Chong, A., R. LaPorta, F. Lopez-de-Silanes, and A. Shleifer. (2014). "Letter Grading Government Efficiency." Journal of European Economic Association. 12 (2): 277-299.

Dittus, M., & Graham, M. (2019). Mapping Wikipedia's geolinguistic contours. *Digital Culture & Society*, *5*(1), 147-164.

Egeberg, M. (2012). How Bureaucratic Structure Matters: An Organisational Perspective. In B. G.

Guriev, S., & Treisman, D. (2019). Informational autocrats. *Journal of Economic Perspectives*, *33*(4), 100-127.

Halfaker, A., & Geiger, R. S. (2020). Ores: Lowering barriers with participatory machine learning in wikipedia. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1-37.

Hoopes, J. L., Reck, D. H., & Slemrod, J. (2015). Taxpayer search for information: Implications for rational attention. *American Economic Journal: Economic Policy*, *7*(3), 177-208.

IP Geolocation API & Free Address Database | DB-IP. (2019). Db-Ip. https://db-ip.com/

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Vishny, R. (1999). The quality of government. *The Journal of Law, Economics, and Organization*, *15*(1), 222-279.

Lemmerich, F., Sáez-Trumper, D., West, R., & Zia, L. (2019, January). Why the world reads Wikipedia: Beyond English speakers. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (pp. 618-626).

Kiesel, J., Potthast, M., Hagen, M., & Stein, B. (2017, May). Spatio-temporal analysis of reverted wikipedia edits. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 11, No. 1).

Kaufmann, D., Kraay, A., & Mastruzzi, M. (2009). Governance matters VIII: aggregate and individual governance indicators, 1996-2008. *World bank policy research working paper*, (4978).

Lee, M.M., and Zhang, N. (2017). "Legibility and the Informational Foundations of State Capacity." The Journal of Politics, 79(1), 118–132.

Muralidharan, K., Niehaus, P., & Sukhtankar, S. (2016). Building state capacity: Evidence from biometric smartcards in India. *American Economic Review*, *106*(10), 2895-2929.

Owen, T. (2015). *Disruptive power: The crisis of the state in the digital age*. Oxford Studies in Digital Poli.

Peters & J. Pierre (Eds.), The SAGE handbook of public administration (pp. 157–168). London: SAGE Publications Ltd

Penney, J. W. (2016). Chilling effects: Online surveillance and Wikipedia use. *Berkeley Tech. LJ*, *31*, 117.

Poerner, N., Waltinger, U., & Schütze, H. (2019). E-BERT: Efficient-yet-effective entity embeddings for BERT. *arXiv preprint arXiv:1911.03681*.

Samoilenko, A., Karimi, F., Edler, D., Kunegis, J., & Strohmaier, M. (2016). Linguistic neighbourhoods: explaining cultural borders on Wikipedia through multilingual co-editing activity. *EPJ data science*, *5*, 1-20.

Singer, P., Lemmerich, F., West, R., Zia, L., Wulczyn, E., Strohmaier, M., & Leskovec, J. (2017, April). Why we read Wikipedia. In *Proceedings of the 26th international conference on world wide web* (pp. 1591-1600).

Suzuki, K., & Demircioglu, M. A. (2019). The association between administrative characteristics and national level innovative activity: Findings from a cross-national study. *Public Performance & Management Review*, *42*(4), 755-782.

Thompson, N., & Hanley, D. (2018). Science is shaped by wikipedia: Evidence from a randomized control trial.

The Economist. (2021, January 14). *Wikipedia is 20, and its reputation has never been higher*. Retrieved January 15, 2022, from https://www.economist.com/international/2021/01/09/wikipedia-is-20-and-its-reputation-has-never-been-higher

Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., & Tang, J. (2021). KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, *9*, 176-194.

West, A. G., Kannan, S., & Lee, I. (2010, April). Detecting wikipedia vandalism via spatio-temporal analysis of revision metadata?. In *Proceedings of the Third European Workshop on System Security* (pp. 22-28).

Wikimedia. (2021). *Wikimedia Downloads*. Wikimedia Dumps. Retrieved July 2021, from https://dumps.wikimedia.org/

Yang, D., Halfaker, A., Kraut, R., & Hovy, E. (2016, March). Who did what: Editor role identification in Wikipedia. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 10, No. 1).

Yasseri, T., & Bright, J. (2016). Wikipedia traffic data and electoral prediction: towards theoretically informed models. *EPJ Data Science*, *5*(1), 1-15.

Zhang, X. M., & Zhu, F. (2011). Group size and incentives to contribute: A natural experiment at Chinese Wikipedia. *American Economic Review*, *101*(4), 1601-15.

Zuboff, S. (2019). The age of surveillance capitalism: the fight for a human future at the new frontier of power.

Zuboff, S. (2020, January 25). Opinion | You Are Now Remotely Controlled. *The New York Times*. https://www.nytimes.com/2020/01/24/opinion/sunday/surveillance-capitalism.html