One path or many? Policy development and diffusion across Wikipedia language editions

Zhaozhi Li, Julia Wagner, Weijun Yuan, Benjamin Mako Hill, Seth Frey

Abstract

Do comparable self-governing collective action institutions converge on comparable policy systems? Do they do so via comparable developmental paths? We test both theories using data on 60 policies shared by 245 Wikipedia language editions which we use as a petri dish for the diffusion of policy in collective action institutions more broadly. We find that policies that are shared tend to be shared widely, that nearly every shared policy can be found in the English edition, and that the clearest predictor of policy adoption order is policy popularity across editions. Although we do not definitively eliminate the possibility that language editions follow multiple paths in converging on their policy systems (say, by culture or language), the evidence suggests that editions follow a single noisy developmental path, potentially suggesting strong influence across editions and a stronger role of common structural constraints than diverse cultural constraints in determining patterns of policy adoption.

Keywords: wikipedia, governance, online communities, policy development, policy diffusion

Introduction

The practice of self-governance is driven as much by environmental and resource constraints as it is by the errors and lessons of its members. This leads to striking diversity in how different communities solve the same problem, and similarly striking consonances between them. As a result, it it often difficult to predict when a set of institutions will converge on identical or divergent solutions to the same governance challenge (Ostrom, 2005). Will organizations facing a similar set of challenges build similar sets of rules? And when they converge on similar solutions, do they tend to follow the same path to get there?

In this work we focus on one platform that allows us to examine the parallel development of hundreds of similar and highly comparable collective action institutions: the several hundred different language editions of Wikipedia. Wikipedia is the free online encyclopedia, published in more than 300 languages. All versions of Wikipedia, and a number of other different non-encyclopedia knowledge bases, are hosted by a common umbrella organization, the Wikimedia Foundation. Because of its open community-driven structure, Wikipedia has already proven to be a productive platform for studying policy processes (Butler et al., 2008; Im et al., 2018; Heaberlin and DeDeo, 2016), but there is only one paper we know of that comparisons policies across language editions systematically (Hwang and Shaw, 2022). By comparing their histories of formal policy development under a common institutional umbrella (and physical server infrastructure), we are able to test several theories of institutional development and influence, particularly those from the "policy diffusion" literature (Sabatier and Weible, 2014; Blatter et al., 2022). Policy diffusion is a body of scholarship concerned with social influence processes between organizations (rather than between individuals), specifically the processes by which policy innovations diffuse among governance institutions.

With high-granularity insights over hundreds of autonomous governance institutions pursuing the same mission under the same constraints, we gain a high-level understanding of the emergence of formal governance that can inform scholarship in disciplines ranging from organizational sociology to institutional economics to comparative government and peer production (Benkler et al., 2015)

Methods

We build a dataset from WikiData. WikiData contains information on MediaWiki's cross-wiki linking feature which links pages in Wikipedia editions to pages on the same topic in other language editions. We use the fact that certain pages are categorized as policy pages in WikiData. By collecting all cross-linked papers in the project namespace and vetting at least one language version by hand to confirm that they are policies, and not project-level activity, we are able build a human-vetted, computer legible indication of cross-edition equivalence of policy structures, despite our ignorance of 326 of the 332 languages that Wikipedia is written in. With this, we extracted the order of adoption of 61 policies that we shared across the

editions. We rely on population-scale comparative analysis to permit evaluation of theories at the unit of analysis of the institution (Hill and Shaw, 2018; Frey and Sumner, 2019). We then employ several data analysis methods to characterize variation in policy adoption order across editions.

We analyze our data using a range computational methods. This includes a range of exploratory, descriptive, and inferential statistics and visualizations, a network analysis of a bipartite policy co-occurrence network, and a computational sequence clustering conducted using the ProM software which is used to conduct sequence mining.

Results

Overall, we find that most policies adopted by Wikipedia language editions are not shared, and that most editions develop their own policies. It is perhaps not surprising that very small editions tend to have very few policies, shared or otherwise. Restricting our attention to characteristics of the shared policies, the English edition has the most, having adopted nearly all of the policies that are shared across editions. Although there is evidence that at least some Wikipedias have pulled their policies directly from English or other larger editions, they frequently develop their own policies or their own versions of policies that are shared.

As shown in Figure 1, we find that the clearest predictor of policy adoption order is policy popularity across editions. In other words, policies that have been adopted by many editions are more likely to be adopted as early policies. Among shared policies, we find that Wikipedias share a common core of policies and that divergence in policy styles tends to happen among more developed encyclopedias, both in terms of having more policies and more specialized policies (see the network in Figure 2). The clustering shows very strong evidence for a single, highly interconnected cluster, and therefore no evidence for coherent "types" of governance styles across editions. To investigate variation in policy adoption paths, our sequence clustering analysis (see Figure 3) shows some variation around a relatively strong policy adoption sequence. We tentatively propose that language editions overall follow a single shared, but noisy, developmental trajectory. One explanation for this result is that shared constraints (such as resource limits or shared mission or constitutional structure) play a larger role in driving policy development than edition-specific constraints (such as language or culture).

Discussion

Our results point to a strong culture of sharing among the most common policies within Wikipedia language editions. Encouraging editions to draw from these highly shared policies, rather than developing their own, may economize on the finite time and energy of each project's volunteers. This seems to be particularly likely to be true in small projects. One caveat is that our analysis is based on the fact that two encyclopias both have, for example, a vandalism policy, not that their vandalism policies are the same. Absent a close multi-language analysis of each policy text, our conclusions only hold for the fact of a policy, and do not extend to its content.

More generally, these results are important for the study of complex collective action institutions generally. The Wikipedias provide a rare case of highly comparable policy systems developing in parallel, enabling us to test several theories of policy development, an important contribution to the policy analysis, collective action, and peer production literatures.

References

[Benkler et al.2015] Yochai Benkler, Aaron Shaw, and Benjamin Mako Hill. 2015. Peer production: A form of collective intelligence. *Handbook of collective intelligence*, 175.

[Blatter et al.2022] Joachim Blatter, Lea Portmann, and Frowin Rausis. 2022. Theorizing policy diffusion. *J European Pub Pol*, 29(6):805–825.

[Butler et al.2008] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. 2008. Don't look now, but we've created a bureaucracy. In *Proceedings of the ACM SIGCHI CHI*, pages 1101–1110.

[Frey and Sumner2019] Seth Frey and Robert W Sumner. 2019. Emergence of integrated institutions in a large population of self-governing communities. *PloS one*, 14(7):e0216335.

[Heaberlin and DeDeo2016] Bradi Heaberlin and Simon DeDeo. 2016. The evolution of wikipedia's norm network. *Future Internet*, 8(2):14.

[Hill and Shaw2018] Benjamin Mako Hill and Aaron Shaw. 2018. Studying populations of online communities. In *The Oxford Handbook of Networked Communication*. Oxford University Press.

[Hwang and Shaw2022] Sohyeon Hwang and Aaron Shaw. 2022. Rules and rule-making in the five largest wikipedias. In *Proceedings of the AAAI ICWSM*, volume 16, pages 347–357.

[Im et al.2018] Jane Im, Amy X Zhang, Christopher J Schilling, and David Karger. 2018. Deliberation and resolution on wikipedia: A case study of requests for comments. *Proceedings of the ACM CHI*, 2(CSCW):1–24.

[Ostrom2005] Elinor Ostrom. 2005. *Understanding institutional diversity*. Princeton University Press.

[Sabatier and Weible2014] Paul A Sabatier and Christopher M Weible. 2014. *Theories of the policy process*. Westview press.

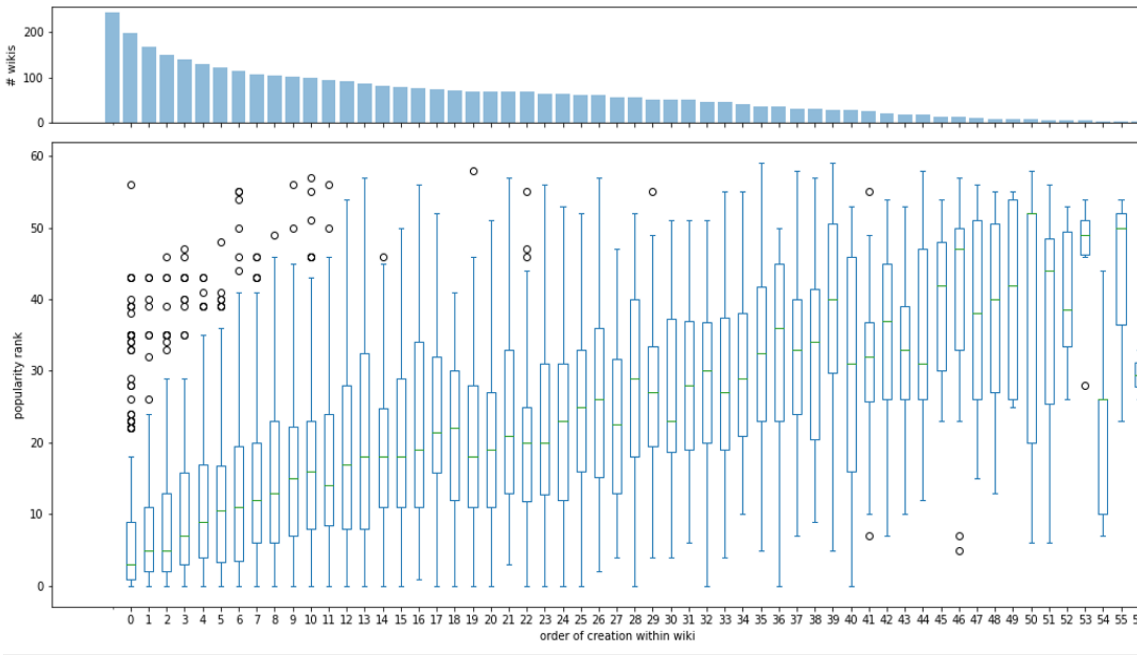

Figure 1: The policies first adopted by a Wikipedia also tend to be the ones that have been adopted by the greatest number of Wikipedias. This is consistent with one policy path. The top bar graph further illustrates this relationship by showing the extent of adoption of each policy across editions.

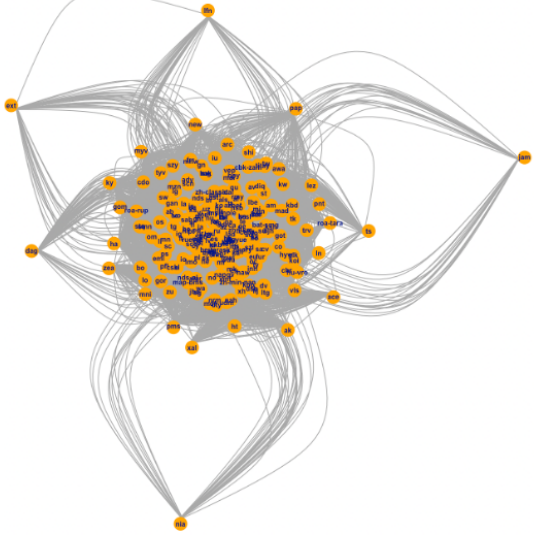

Figure 2: A bipartite network of language editions linked by shared policy shows a dense central core, indicating high overlap in policy coverage across editions.

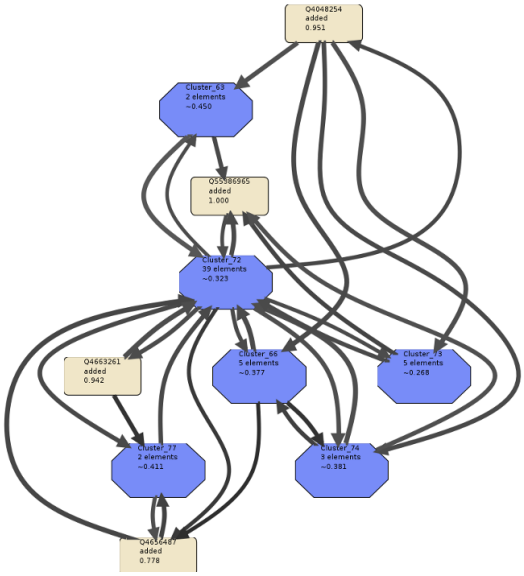

Figure 3: Sequence clustering analysis groups any sequence of things based on shared subsequences. We use it to cluster policy adoption sequences, for evidence of distinct policy adoption paths. The resulting evidence is most consistent with a single (noisy) typical path.