Automated Measurement of Policy Similarity in Legislation: Revealing Patterns of Text Reuse

Much political science research relies on identifying the substantive similarity of legislation, especially in the study of policy diffusion. Scholars mostly rely on hand-coded datasets. However, in order to study policy diffusion more broadly, for example by extending the study to dozens or more policy domains, prohibitive volumes of legislative text would have to be analyzed manually. Common text analysis algorithms such as topic models or bag-of-words based text similarity measures are too coarse to identify policy similarity. However, legislators often directly use text from bills introduced in other legislatures or bill text provided by interest groups in model legislation. We propose the use of text-sequencing algorithms to find matching text between bills. We describe a new approach to the application of sequence alignment algorithms to large amounts of legislative text, and assess the validity of text reuse as a measure of substantive bill similarity. Three key results, drawn from an analysis of 500,000 bills from US-state legislatures, demonstrate the validity of text reuse as a measure of policy similarity. First, we show that bills introduced by ideologically similar sponsors are more likely to exhibit a high degree of text reuse. Second, we show that the main topical themes underlying strings of reused text map closely onto major areas of public policy in the US States. Third, we show that rates of text reuse across state borders correlates strongly with the state policy diffusion ties data recently introduced by Desmarais, Harden and Boehmke (APSR, 2015).