MediaWell

News Item

# Research note: Examining potential bias in large-scale censored data | HKS Misinformation Review

By Jennifer Allen and Duncan J. Watts
July 28, 2021


> We examine potential bias in Facebook's 10-trillion cell URLs dataset, consisting of URLs shared on its platform and their engagement metrics. Despite the unprecedented size of the dataset, it was altered to protect user privacy in two ways: 1) by adding differentially private noise to engagement counts, and 2) by censoring the data with a 100-public-share threshold for a URL's inclusion. To understand how these alterations affect conclusions drawn from the data, we estimate the prevalence of fake news in the massive, censored URLs dataset and compare it to an estimate from a smaller, representative dataset. We show that censoring can substantially alter conclusions that are drawn from the Facebook dataset. Because of this 100-public-share threshold, descriptive statistics from the Facebook URLs dataset overestimate the share of fake news and news overall by as much as 4X. We conclude with more general implications for censoring data.
>
> […]



Source: Research note: Examining potential bias in large-scale censored data | HKS Misinformation Review