Splink: Free software for probabilistic record linkage at scale.

Robin Linacre
Ministry of Justice

Sam Lindsay
Ministry of Justice

Theodore Manassis
Ministry of Justice

Zoe Slade
Ministry of Justice

Tom Hepworth
Ministry of Justice

Abstract

Funded by ADR UK, a new data linking team at the Ministry of Justice set out to link administrative datasets across the justice space, for internal use and sharing with external researchers. To achieve this aim we sought a linkage implementation that was probabilistic, flexible, scalable and ideally open source.

Taking into account the tools available at the MoJ, existing open-source software (and paid alternatives) failed to meet our desired criteria. It was decided to develop a software package that builds on FastLink's implementation in R of an Expectation-Maximisation algorithm to estimate a Fellegi-Sunter linkage model, adding a range of technical improvements, increased functionality and customisation options. Distributed computing offered by Spark could facilitate comparable linkage jobs that run on much

larger datasets and much faster. Working with government data, accountability and transparency are vital, so the data and models are made accessible by a range of intuitive visualizations.

The Splink python package has been downloaded 2 million times. This initially used Spark to deliver its superior performance, but Splink v3 caters for various SQL backends and more potential users. Splink and supplementary python libraries are publicly visible on GitHub and provide assistance in all aspects of data linkage:

- splink_data_standardisation - functions to perform general data standardisation

- splink_cluster_studio - creates an interactive HTML dashboard to analyse record clusters (building on splink_graph - a library for generating graph metrics)

- splink_synthetic_data - generating realistic synthetic data for testing linkage algorithms

- splink_demos - interactive demo/tutorial notebooks for a range of features of the Splink libraries

These tools are in continuous development and have already been used to deliver deduplicated and linked data products for the entire criminal justice system.

Through technical innovation and user-focused development, Splink has improved access to cutting-edge data linkage, and created groundbreaking research opportunities at MoJ and beyond. The team is grateful to ONS and other collaborators for testing and adopting these tools, and will continue to explore ways to improve performance and user experience.

How to Cite

Linacre, R., Lindsay, S., Manassis, T., Slade, Z. and Hepworth, T. (2022) “Splink: Free software for probabilistic record linkage at scale”., International Journal of Population Data Science, 7(3). doi: 10.23889/ijpds.v7i3.1794.

Copyright

(https://ijpds.org/issue/view/25)

Download/View PDF (https://ijpds.org/article/view/1794/3457)

Download XML (https://ijpds.org/article/view/1794/3458)

Published: Aug 25, 2022

IJPDS

DOI: https://doi.org/10.23889/ijpds.v7i3.1794 (https://doi.org/10.23889/ijpds.v7i3.1794)

Keywords:

ISSN: 2399-4908

open-source, probabilistic data linking, administrative data

Publisher

Statistics from Altmetric (https://altmetric.com)

Strategic Partner

INTERNATIONAL Population Data Linkage NETWORK

Members Of