

# String distances for near-duplicate detection

Near-duplicate detection is important when dealing with large, noisy databases in data mining tasks. We present the results of applying the **Rank** distance and the **Smith-Waterman** distance, along with the classic **Levenshtein** distance, together with a disjoint set data structure powered by the **Union-Find** method.

Iulia Dănilă  
Liviu P. Dinu  
Vlad Niculae  
Octavia-Maria Șulea

danailaiulia@yahoo.com  
ldinu@fmi.unibuc.ro  
vlad@vene.ro  
mary.octavia@gmail.com

Faculty of Mathematics and Computer Science  
University of Bucharest

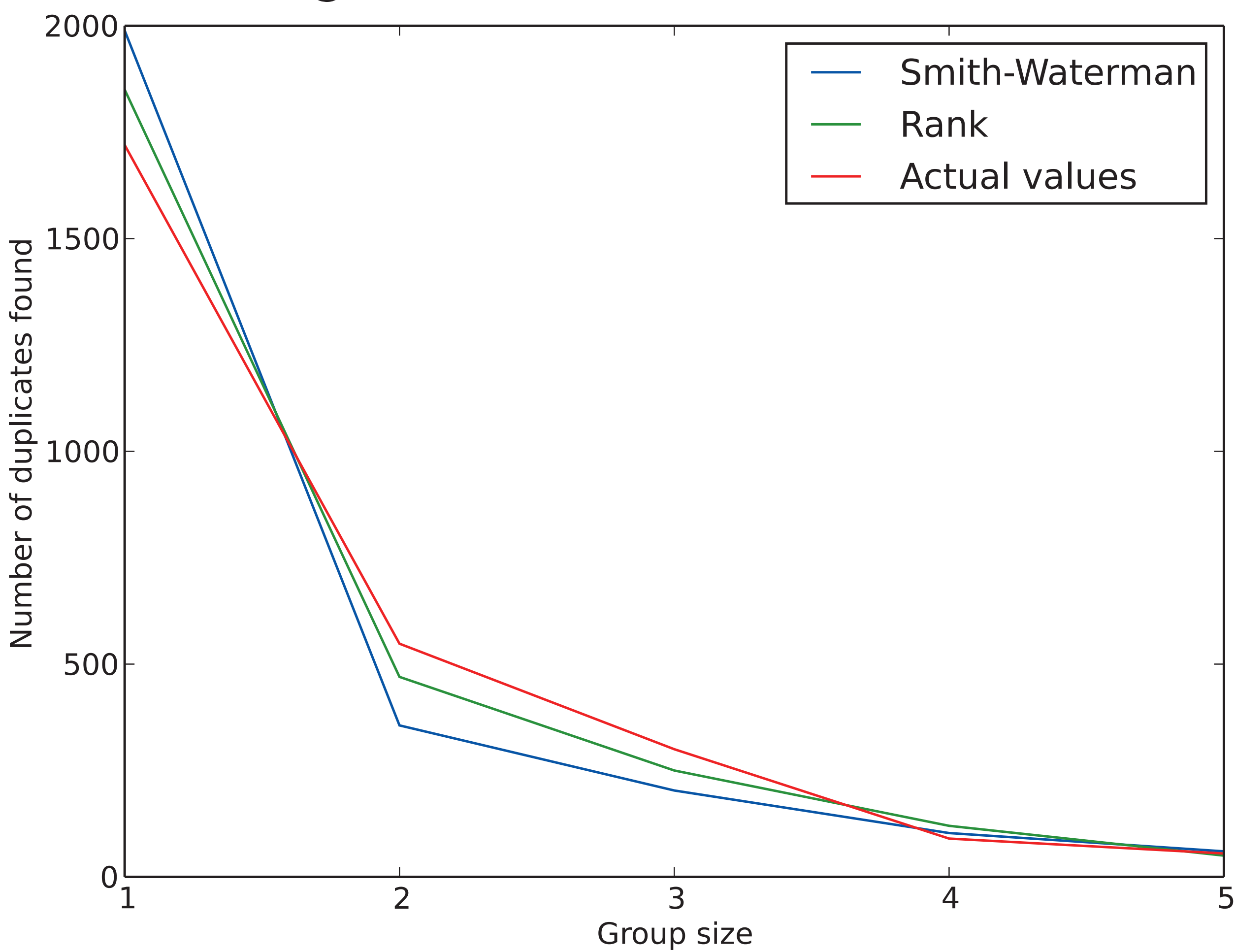
Metric	Perfect matching	Near matching
Rank distance	$d = 0$	$d \leq \frac{n_a n_b}{2}$
Smith-Waterman	$d = 2 \max(n_a, n_b)$	$d \geq \min(n_a, n_b)$
Levenshtein	$d = 0$	$d \leq \frac{\max(n_a, n_b)}{2}$
Similar-text	$p = 100\%$	$p \geq 50\%$

Thresholds and critical values for the distances used. Rank distance is an ordinal metric for comparing two possibly disjoint rankings. Smith-Waterman is a dynamic programming algorithm from molecular biology. Similar-text is the algorithm implemented by the PHP function with the same name.

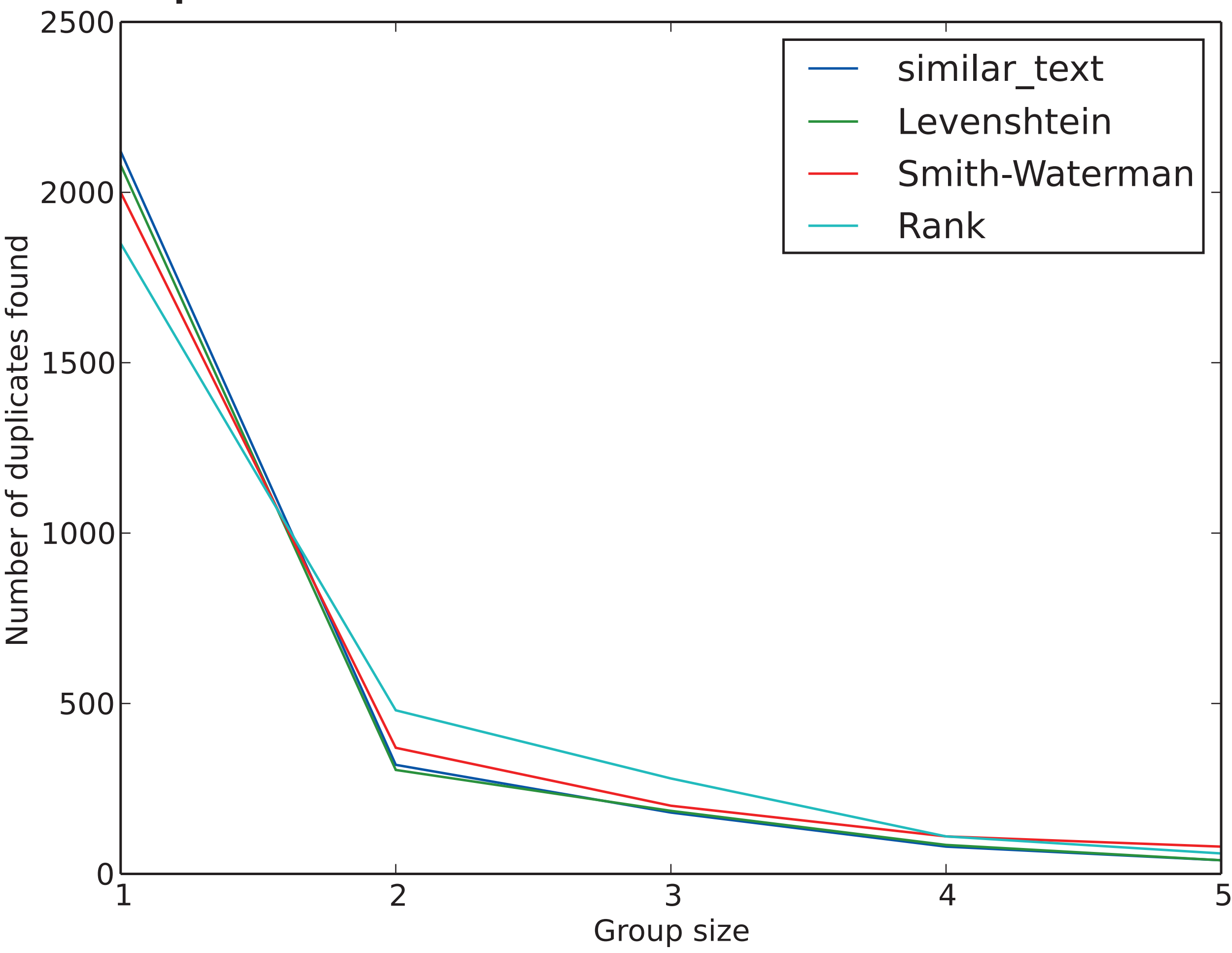
Artificial data was generated starting with entries from a product database mined from the Internet, to which randomly distorted duplicates were added. This was done in order to have a ground truth for evaluation. The distribution of duplicates is shown below.

Group size	Groups number	Input number	Percent
1	1720	1720	62.77%
2	274	548	20%
3	96	288	10.51%
4	26	104	3.79%
5	16	80	2.91%
Total	2132	2740	100%

First two distances compared to the ground truth on artificial data



Comparison of all methods on artificial data



The second database is a real-world, undistorted bibliographic collection in BibTeX format, from which we extracted only the title and the author names, in order to lighten the workload given our assumption that the most errors occur in these fields. The source of the data is "A Collection of Computer Science Bibliographies". For example here are two duplicate entries:

```
{author: "Andrew G. Barto and S. J. Bradtke and  
Satinder P.Singh", title: "Learning to Act Using  
Real Time Dynamic Programming"}  
{author: "Andrew Barto, J. S. Bradtke and S. P.  
Singh", title: "Learning to Act Using Realtime  
Dynamic Programming"}
```

The figures are distribution plots, the y-value at the position  $x = k$  showing the number of documents that can be captured in groups of  $k$ .

In the case of the artificially generated noisy database, we have access to the ground truth. The results found by Rank distance are closer to the real distribution of duplicates than the ones found by the Smith-Waterman distance.

For the bibliographic entry database no algorithm found more than 3 duplicate entries for the same information. Under human evaluation, the identified duplicates look correct, confirming the precision of the methods. Rank distance seems to have a slower decay rate than the other methods, which can be interpreted as higher recall in the distribution's tail, assuming good precision.

Comparison of all methods on real-world BibTeX author-title pairs

