Evolutionary algorithms for overlapping correlation clustering

[Supplementary Material]

Starkey trajectory similarity

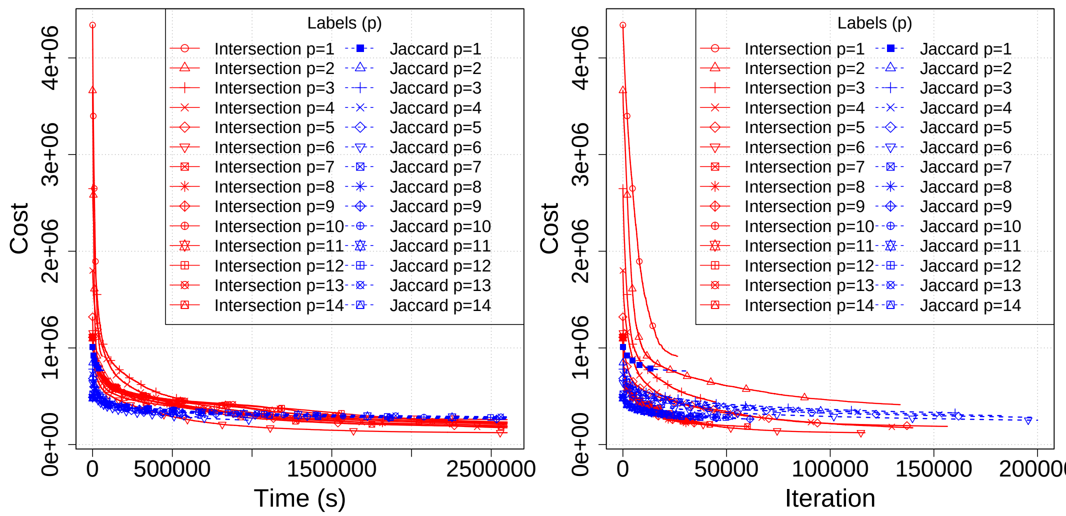
To calculate the similarity of trajectories for Starkey, we use the EDR distance [5], which is defined as following: let $P = [(x_1, y_1, t_1), \ldots, (x_n, y_n, t_n)]$ be a trajectory such that each triple (x, y, t) is a position in space and time. Denote by $r(P) = [(x_2, y_2, t_2), \ldots, (x_n, y_n, t_n)]$ the remainder of the trajectory, i.e., the original trajectory without the first point. Let P and Q be two different trajectories. For $p \in P$, $q \in Q$, we say that $m(p, q) = 1$ if $|p_x - q_x| < \varepsilon_x$ and $|p_y - q_y| < \varepsilon_y$ and $|p_t - q_t| < \varepsilon_t$, i.e., the distance in space and time is not larger than a constant factor. We take $m(p, q) = 0$ otherwise. We say the *Edit Distance in Real Sequences* is

$$EDR(P,Q) = \begin{cases} |P| & \text{if } |Q| = 0, \\ |Q| & \text{if } |P| = 0, \\ \min(EDR(r(P), r(Q)) + m(p_1, q_1), & \\ \qquad EDR(r(P), Q) + 1, & \\ \qquad EDR(P, r(Q)) + 1), & \text{otherwise.} \end{cases}$$

The similarity between two trajectories u and v is given by

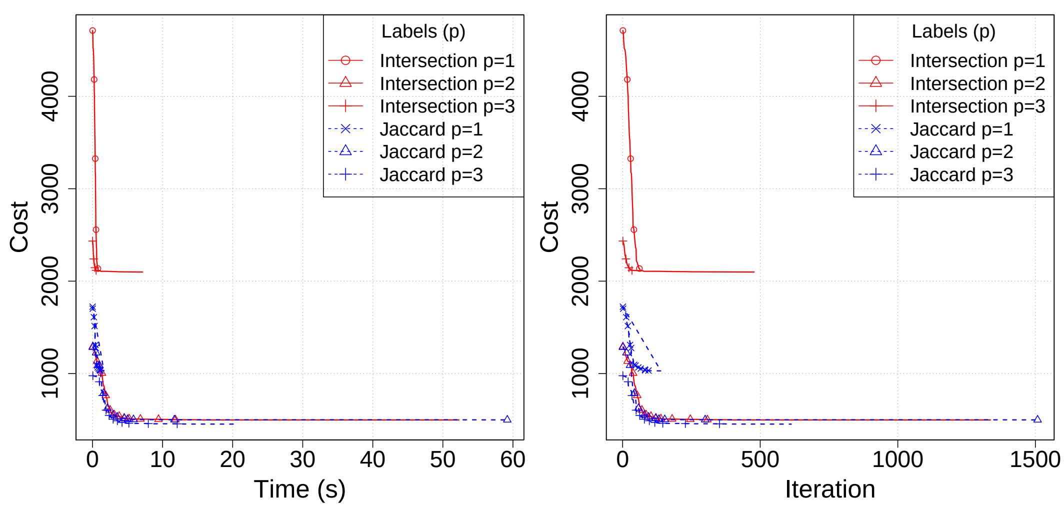
$$s(u, v) = 1 - EDR(u, v). \tag{1}$$

Additional plots of Section 4.4



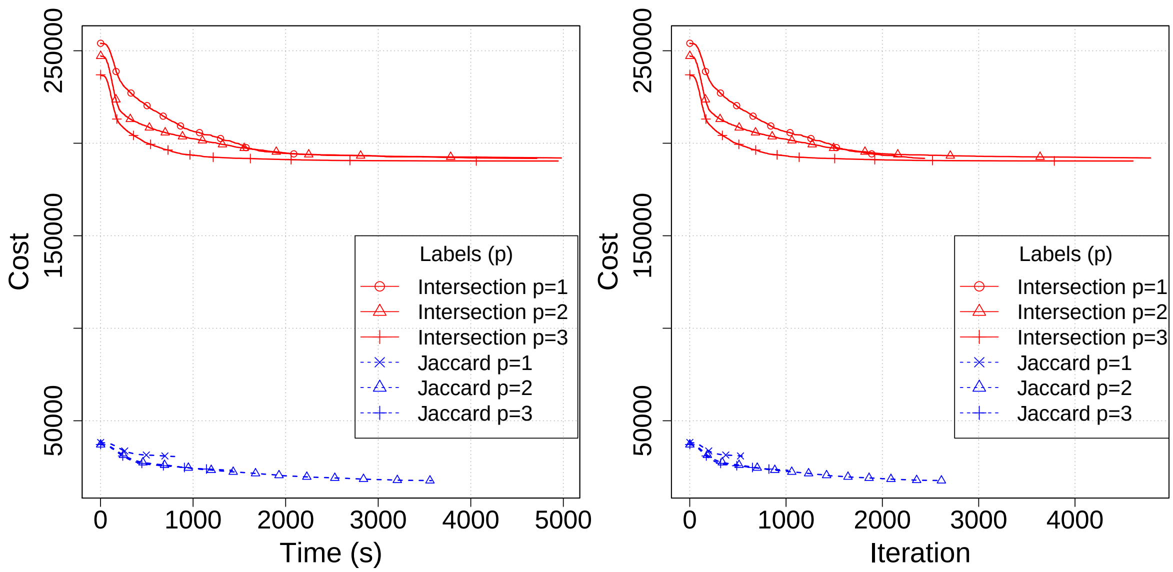
(a) (b)

Figure 1: Evolution of the cost for the Yeast dataset.



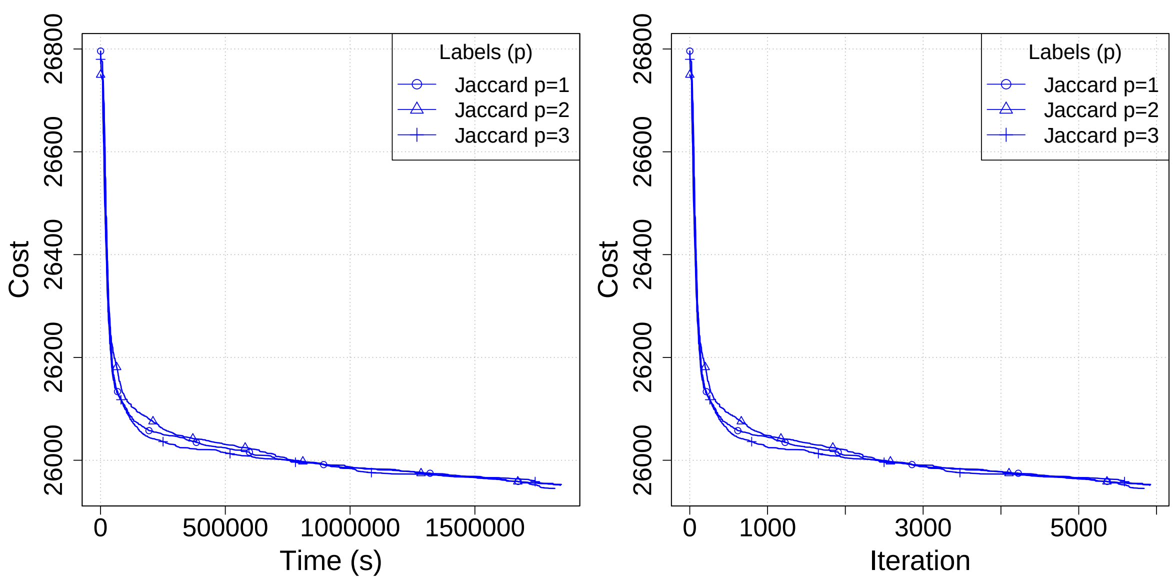
(a) (b)

Figure 2: Evolution of the cost for the Starkey project dataset.



(a) (b)

Figure 3: Evolution of the cost for the protein alignment dataset 1.



(a) (b)

Figure 4: Evolution of the cost for the newsgroup messages.

Additional tables of Section 4.6

Table 1: Difference in median location for cost distributions for Starkey dataset using Wilcoxon-Mann-Whitney U test with 95% of confidence. The bottom-left block shows p-values that are greater than 0.05. A negative value means that the median of the "line" algorithm is smaller/better than the "column" algorithm. A dash (—) indicates that the results for that pair of algorithms are identical.

$\mathcal{H}$	p	Algorithm	`Bonchi`	`OLS-Comp`	`OLS-Ext`	`BLS-Comp`	`BLS-Ext`
Jaccard	1	`Bonchi`	**0.61**	-0.27	-0.27	0.60	0.61
		`OLS-Comp`		**0.89**	0.00	0.87	0.88
		`OLS-Ext`		0.16	**0.89**	0.87	0.88
		`BLS-Comp`				**0.01**	0.01
		`BLS-Ext`					**0.003**
	2	`Bonchi`	**0.98**	0.80	0.73	0.96	0.97
		`OLS-Comp`		**0.17**	-0.07	0.15	0.16
		`OLS-Ext`			**0.24**	0.22	0.23
		`BLS-Comp`				**0.01**	0.008
		`BLS-Ext`					**0.009**
	3	`Bonchi`	**0.98**	0.91	0.94	0.96	0.96
		`OLS-Comp`		**0.06**	0.02	0.04	0.04
		`OLS-Ext`			**0.03**	0.01	0.01
		`BLS-Comp`				**0.01**	-0.001
		`BLS-Ext`				0.67	**0.01**
Set-intersection	1	`Bonchi`	**0.85**	0.85	0.85	—	—
		`OLS-Comp`		**0.00**	0.00	-0.85	-0.85
		`OLS-Ext`			**0.00**	-0.85	-0.85
		`BLS-Comp`				**0.85**	—
		`BLS-Ext`					**0.85**
	2	`Bonchi`	**0.40**	0.40	-0.04	—	—
		`OLS-Comp`		**0.003**	-0.44	-0.40	-0.40
		`OLS-Ext`			**0.44**	0.04	0.04
		`BLS-Comp`				**0.40**	—
		`BLS-Ext`					**0.04**
	3	`Bonchi`	**0.62**	0.59	-0.07	—	—
		`OLS-Comp`		**0.03**	-0.65	-0.59	-0.59
		`OLS-Ext`	0.06		**0.67**	0.07	0.07
		`BLS-Comp`			0.06	**0.62**	—
		`BLS-Ext`					**0.62**

Table 2: Difference in median location for cost distributions for SCOP datasets using Wilcoxon-Mann-Whitney U test with 95% of confidence. The bottom-left block shows p-values that are greater than 0.05. A negative value means that the median of the "line" algorithm is smaller/better than the "column" algorithm. A dash (—) indicates that the results for that pair of algorithms are identical.

$\mathcal{H}$	p	Algorithm	`Bonchi`	`OLS-Comp`	`OLS-Ext`	`BLS-Comp`	`BLS-Ext`
Jaccard	1	`Bonchi`	**0.96**	-0.03	-0.03	0.70	0.95
		`OLS-Comp`		**1.00**	—	0.75	0.99
		`OLS-Ext`			**1.00**	0.75	0.99
		`BLS-Comp`				**0.24**	0.23
		`BLS-Ext`					**0.00**
	2	`Bonchi`	**0.99**	0.97	0.98	0.19	0.92
		`OLS-Comp`		**0.01**	0.001	-0.77	-0.02
		`OLS-Ext`		0.62	**0.00**	-0.74	-0.00
		`BLS-Comp`				**0.79**	0.73
		`BLS-Ext`		0.28	0.32		**0.04**
	3	`Bonchi`	**0.98**	0.85	0.97	0.00002	0.40
		`OLS-Comp`		**0.07**	0.07	-0.85	-0.43
		`OLS-Ext`			**0.0003**	-0.95	-0.46
		`BLS-Comp`	0.92			**0.95**	0.42
		`BLS-Ext`					**0.48**
Set-intersection	1	`Bonchi`	**1.00**	-0.00001	-0.00002	-0.00002	-0.00002
		`OLS-Comp`	0.52	**1.00**	0.00003	0.00003	0.00003
		`OLS-Ext`		0.65	**1.00**	—	—
		`BLS-Comp`		0.65		**1.00**	—
		`BLS-Ext`		0.65			**1.00**
	2	`Bonchi`	**1.00**	-0.00002	-0.00002	-0.00002	-0.00002
		`OLS-Comp`		**1.00**	-0.00003	—	—
		`OLS-Ext`	0.34	0.48	**1.00**	0.0000004	0.0000004
		`BLS-Comp`			0.48	**1.00**	—
		`BLS-Ext`			0.48		**1.00**
	3	`Bonchi`	**1.00**	-0.00002	-0.00001	-0.00002	-0.00002
		`OLS-Comp`		**1.00**	-0.00003	—	—
		`OLS-Ext`	0.52	0.65	**1.00**	0.00003	0.00003
		`BLS-Comp`			0.65	**1.00**	—
		`BLS-Ext`			0.65		**1.00**

Table 3: Difference in median location for cost distributions for newsgroup messages using Wilcoxon-Mann-Whitney U test with 95% of confidence. The bottom-left block shows p-values that are greater than 0.05. A negative value means that the median of the "line" algorithm is smaller/better than the "column" algorithm. A dash (—) indicates that the results for that pair of algorithms are identical.

$\mathcal{H}$	p	Algorithm	Bonchi	OLS-Comp	OLS-Ext	BLS-Comp	BLS-Ext
Jaccard	1	Bonchi	**0.001**	-0.52	-0.92	-0.46	-0.51
		OLS-Comp		**0.52**	-0.40	0.05	0.006
		OLS-Ext			**0.92**	0.45	0.44
		BLS-Comp				**0.46**	-0.04
		BLS-Ext		0.70			**0.51**
	2	Bonchi	**0.99**	0.99	0.90	0.002	0.75
		OLS-Comp		**0.00**	-0.09	-0.99	-0.23
		OLS-Ext			**0.09**	-0.89	-0.14
		BLS-Comp	0.70			**0.99**	0.75
		BLS-Ext					**0.23**
	3	Bonchi	**0.98**	0.98	0.94	-0.01	0.77
		OLS-Comp		**0.0002**	-0.04	-0.99	-0.21
		OLS-Ext			**0.04**	-0.95	-0.16
		BLS-Comp	0.10			**0.99**	0.78
		BLS-Ext					**0.21**
Set-intersection	1	Bonchi	**0.000007**	-0.99	-0.99	-0.001	-0.001
		OLS-Comp		**0.99**	-0.005	0.992	0.99
		OLS-Ext			**0.99**	0.998	0.99
		BLS-Comp	0.10			**0.001**	-0.0001
		BLS-Ext	0.10			0.70	**0.001**
	2	Bonchi	**0.00001**	-0.98	-0.99	-0.001	-0.001
		OLS-Comp		**0.98**	-0.01	0.98	0.98
		OLS-Ext			**0.99**	0.99	0.99
		BLS-Comp	0.10			**0.001**	-0.0003
		BLS-Ext	0.10			0.10	**0.001**
	3	Bonchi	**0.00003**	-0.99	-0.99	-0.001	-0.001
		OLS-Comp		**0.99**	-0.007	0.99	0.99
		OLS-Ext			**0.99**	0.99	0.99
		BLS-Comp	0.01			**0.001**	-0.000
		BLS-Ext	0.01			0.10	**0.001**