Incremental Record Linkage

Anja Gruenheid
Xin Luna Dong
Divesh Srivastava

Introduction

- What is record linkage?

  The task of linking records that refer to the same real-world entity.

- Why do we need incremental record linkage?

  Batch computing record linkage is costly. If the underlying data set is modified only slightly, it is more efficient to use an incremental approach.

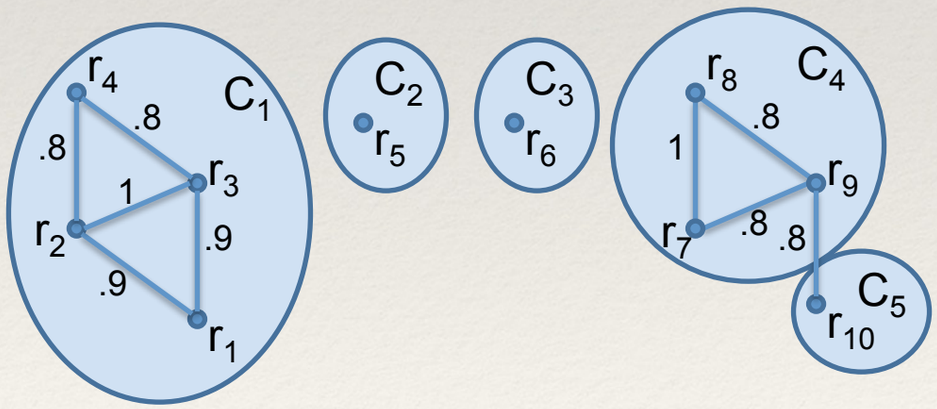
Example: IRL

	BizID	ID	name	street address	city	phone
$\mathbf{D_0}$	B_1	r_1	Starbucks	123 MISSION ST STE ST1	SAN FRANCISCO	4155431510
	B_1	r_2	Starbucks	123 MISSION ST	SAN FRANCISCO	4155431510
	B_1	r_3	Starbucks	123 Mission St	San Francisco	4155431510
	B_2	r_4	Starbucks Coffee	340 MISSION ST	SAN FRANCISCO	4155431510
	B_3	r_5	Starbucks Coffee	333 MARKET ST	SAN FRANCISCO	4155434786
	B_3	r_6	Starbucks	MARKET ST	San Francisco	
	B_4	r_7	Starbucks Coffee	52 California St	San Francisco	4153988630
	B_4	r_8	Starbucks Coffee	52 CALIFORNIA ST	SAN FRANCISCO	4153988630
	B_5	r_9	Starbucks Coffee	295 California St	San Francisco	4159862349
	B_5	r_{10}	Starbucks	295 California St	San Francisco	

Example: IRL

	BizID	ID	name	street address	city	phone
$\mathbf{D_0}$	B_1	r_1	Starbucks	123 MISSION ST STE ST1	SAN FRANCISCO	4155431510
	B_1	r_2	Starbucks	123 MISSION ST	SAN FRANCISCO	4155431510
	B_1	r_3	Starbucks	123 Mission St	San Francisco	4155431510
	B_2	r_4	Starbucks Coffee	340 MISSION ST	SAN FRANCISCO	4155431510
	B_3	r_5	Starbucks Coffee	333 MARKET ST	SAN FRANCISCO	4155434786
	B_3	r_6	Starbucks	MARKET ST	San Francisco	
	B_4	r_7	Starbucks Coffee	52 California St	San Francisco	4153988630
	B_4	r_8	Starbucks Coffee	52 CALIFORNIA ST	SAN FRANCISCO	4153988630
	B_5	r_9	Starbucks Coffee	295 California St	San Francisco	4159862349
	B_5	r_{10}	Starbucks	295 California St	San Francisco	

apply batch record linkage



Example: IRL

	BizID	ID	name	street address	city	phone
$\mathbf{D_0}$	B_1	r_1	Starbucks	123 MISSION ST STE ST1	SAN FRANCISCO	4155431510
	B_1	r_2	Starbucks	123 MISSION ST	SAN FRANCISCO	4155431510
	B_1	r_3	Starbucks	123 Mission St	San Francisco	4155431510
	B_2	r_4	Starbucks Coffee	340 MISSION ST	SAN FRANCISCO	4155431510
	B_3	r_5	Starbucks Coffee	333 MARKET ST	SAN FRANCISCO	4155434786
	B_3	r_6	Starbucks	MARKET ST	San Francisco	
	B_4	r_7	Starbucks Coffee	52 California St	San Francisco	4153988630
	B_4	r_8	Starbucks Coffee	52 CALIFORNIA ST	SAN FRANCISCO	4153988630
	B_5	r_9	Starbucks Coffee	295 California St	San Francisco	4159862349
	B_5	r_{10}	Starbucks	295 California St	San Francisco	

	BizID	ID	name	street address	city	phone
$\mathbf{\Delta D}_1$	B_6	r_{11}	Starbucks Coffee	201 Spear Street	San Francisco	4159745077
$\mathbf{\Delta D}_2$	B_3	r_{12}	Starbucks Coffee	MARKET ST	San Francisco	4155434786
	B_3	r_{13}	Starbucks	333 MARKET ST	San Francisco	4155434786
$\mathbf{\Delta D}_3$	B_1	r_{14}	Starbucks	123 MISSION ST STE ST1	SAN FRANCISCO	4155431510
	B_1	r_{15}	Starbucks	123 Mission St Ste St1	San Francisco	4155431510
$\mathbf{\Delta D}_4$	B_5	r_{16}	Starbucks	295 CALIFORNIA ST	SAN FRANCISCO	4159862349
	B_4	r_{17}	Starbucks	52 California Street	SF	4153988630

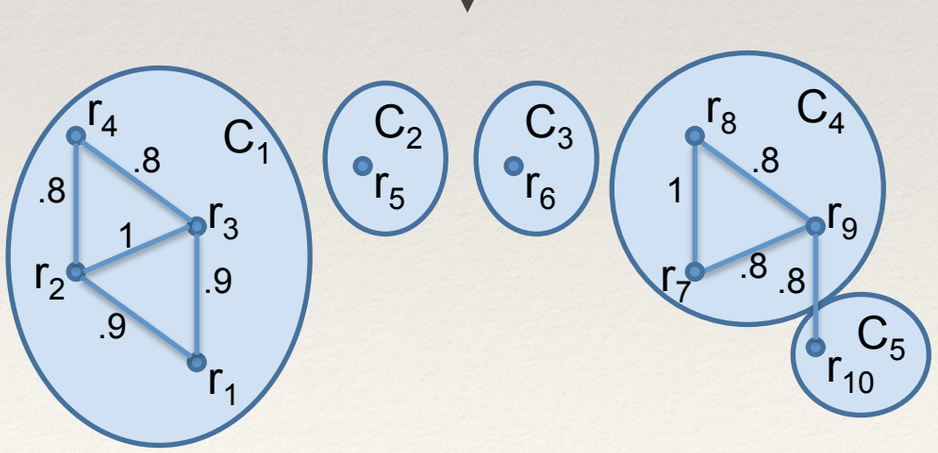
Example: IRL

	BizID	ID	name	street address	city	phone
$\mathbf{D_0}$	B_1	r_1	Starbucks	123 MISSION ST STE ST1	SAN FRANCISCO	4155431510
	B_1	r_2	Starbucks	123 MISSION ST	SAN FRANCISCO	4155431510
	B_1	r_3	Starbucks	123 Mission St	San Francisco	4155431510
	B_2	r_4	Starbucks Coffee	340 MISSION ST	SAN FRANCISCO	4155431510
	B_3	r_5	Starbucks Coffee	333 MARKET ST	SAN FRANCISCO	4155434786
	B_3	r_6	Starbucks	MARKET ST	San Francisco	
	B_4	r_7	Starbucks Coffee	52 California St	San Francisco	4153988630
	B_4	r_8	Starbucks Coffee	52 CALIFORNIA ST	SAN FRANCISCO	4153988630
	B_5	r_9	Starbucks Coffee	295 California St	San Francisco	4159862349
	B_5	r_{10}	Starbucks	295 California St	San Francisco	

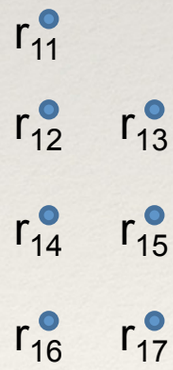
apply incremental record linkage

+

	BizID	ID	name	street address	city	phone
$\Delta\mathbf{D}_1$	B_6	r_{11}	Starbucks Coffee	201 Spear Street	San Francisco	4159745077
$\Delta\mathbf{D}_2$	B_3	r_{12}	Starbucks Coffee	MARKET ST	San Francisco	4155434786
	B_3	r_{13}	Starbucks	333 MARKET ST	San Francisco	4155434786
$\Delta\mathbf{D}_3$	B_1	r_{14}	Starbucks	123 MISSION ST STE ST1	SAN FRANCISCO	4155431510
	B_1	r_{15}	Starbucks	123 Mission St Ste St1	San Francisco	4155431510
$\Delta\mathbf{D}_4$	B_5	r_{16}	Starbucks	295 CALIFORNIA ST	SAN FRANCISCO	4159862349
	B_4	r_{17}	Starbucks	52 California Street	SF	4153988630



+



= ?

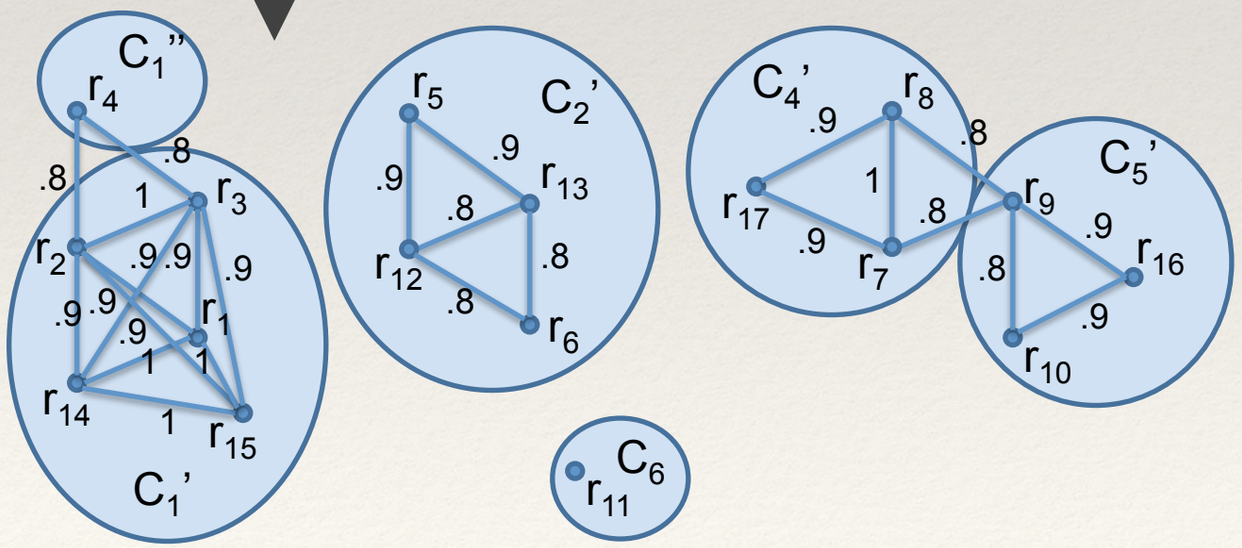
Example: IRL

	BizID	ID	name	street address	city	phone
$\mathbf{D_0}$	B_1	r_1	Starbucks	123 MISSION ST STE ST1	SAN FRANCISCO	4155431510
	B_1	r_2	Starbucks	123 MISSION ST	SAN FRANCISCO	4155431510
	B_1	r_3	Starbucks	123 Mission St	San Francisco	4155431510
	B_2	r_4	Starbucks Coffee	340 MISSION ST	SAN FRANCISCO	4155431510
	B_3	r_5	Starbucks Coffee	333 MARKET ST	SAN FRANCISCO	4155434786
	B_3	r_6	Starbucks	MARKET ST	San Francisco	
	B_4	r_7	Starbucks Coffee	52 California St	San Francisco	4153988630
	B_4	r_8	Starbucks Coffee	52 CALIFORNIA ST	SAN FRANCISCO	4153988630
	B_5	r_9	Starbucks Coffee	295 California St	San Francisco	4159862349
	B_5	r_{10}	Starbucks	295 California St	San Francisco	

apply incremental record linkage

+

	BizID	ID	name	street address	city	phone
$\Delta\mathbf{D}_1$	B_6	r_{11}	Starbucks Coffee	201 Spear Street	San Francisco	4159745077
$\Delta\mathbf{D}_2$	B_3	r_{12}	Starbucks Coffee	MARKET ST	San Francisco	4155434786
	B_3	r_{13}	Starbucks	333 MARKET ST	San Francisco	4155434786
$\Delta\mathbf{D}_3$	B_1	r_{14}	Starbucks	123 MISSION ST STE ST1	SAN FRANCISCO	4155431510
	B_1	r_{15}	Starbucks	123 Mission St Ste St1	San Francisco	4155431510
$\Delta\mathbf{D}_4$	B_5	r_{16}	Starbucks	295 CALIFORNIA ST	SAN FRANCISCO	4159862349
	B_4	r_{17}	Starbucks	52 California Street	SF	4153988630



Optimal Approaches

- Connected Component Approach

  Update the connected component (set of clusters) that is or was connected to the modified record.

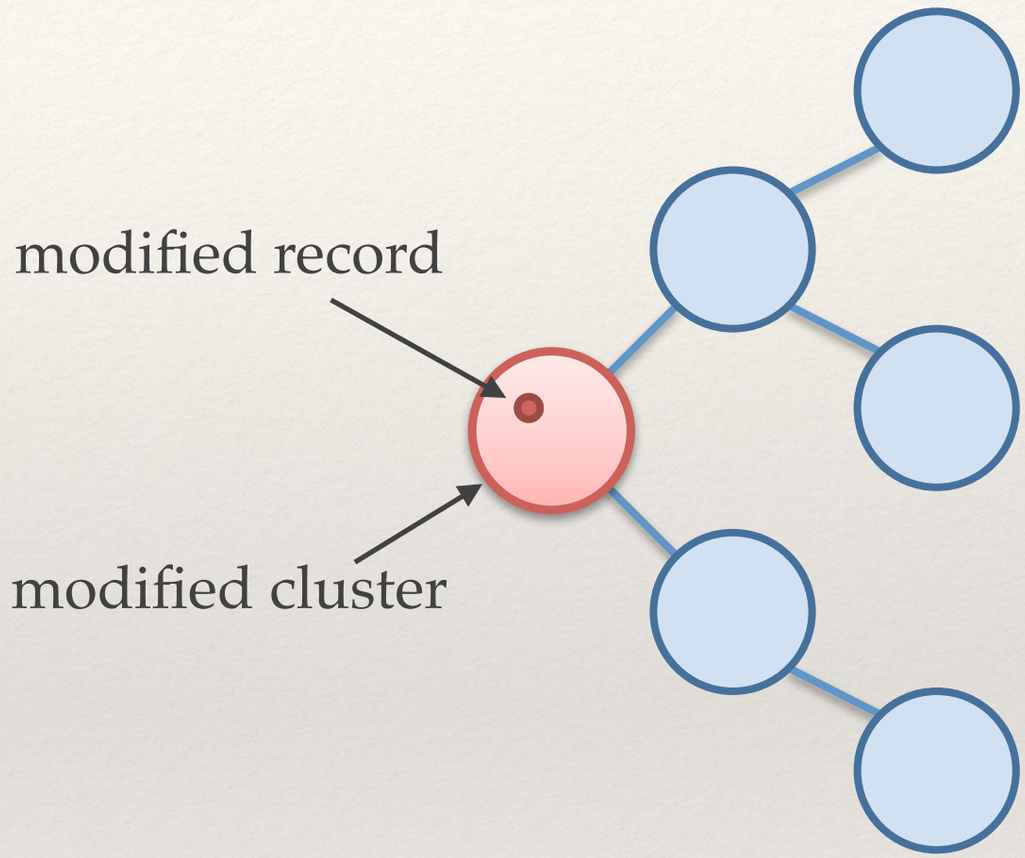
- Iterative Approach

  Iteratively propagate the update through clusters in the connected component.

Example: Iterative Approach

1) A modified record is associated with a modified cluster.

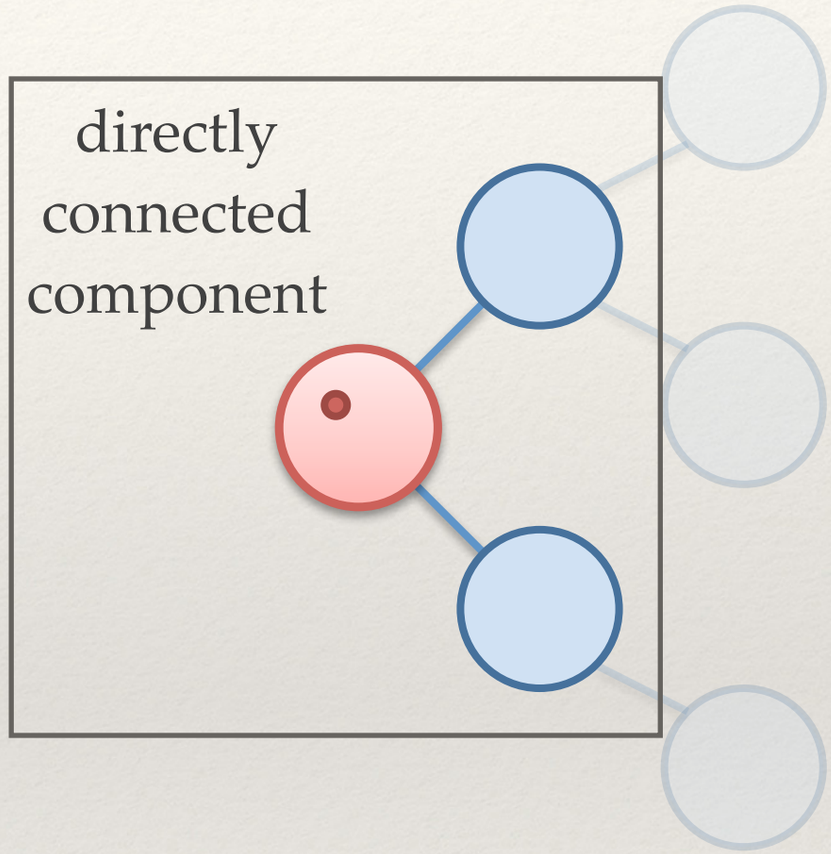
Modified clusters can be singleton clusters if record cannot be associated with an existing cluster.



Example: Iterative Approach

2) The directly connected component is evaluated with a batch algorithm.

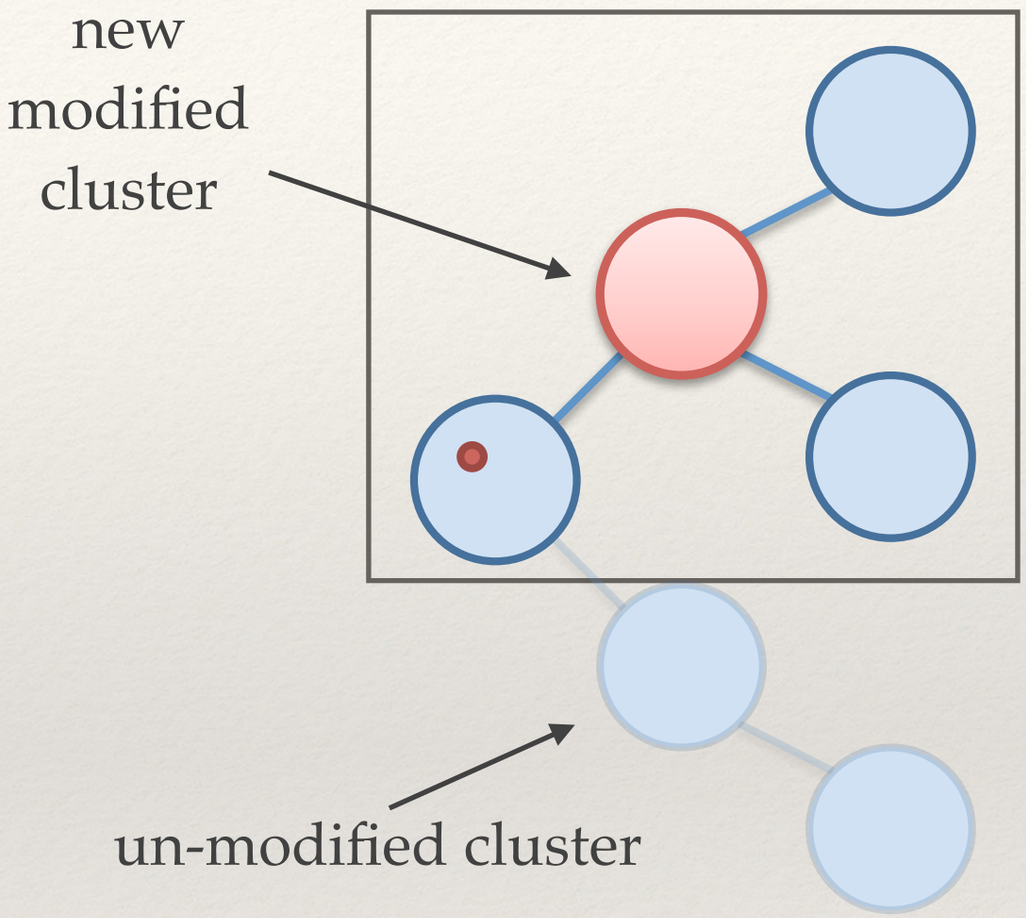
A directly connected component are those clusters directly connected to the modified cluster.



Example: Iterative Approach

3) Iteratively proceed along modified clusters only.

The modified clusters are iteratively explored to avoid unnecessary clustering for non-modified clusters.



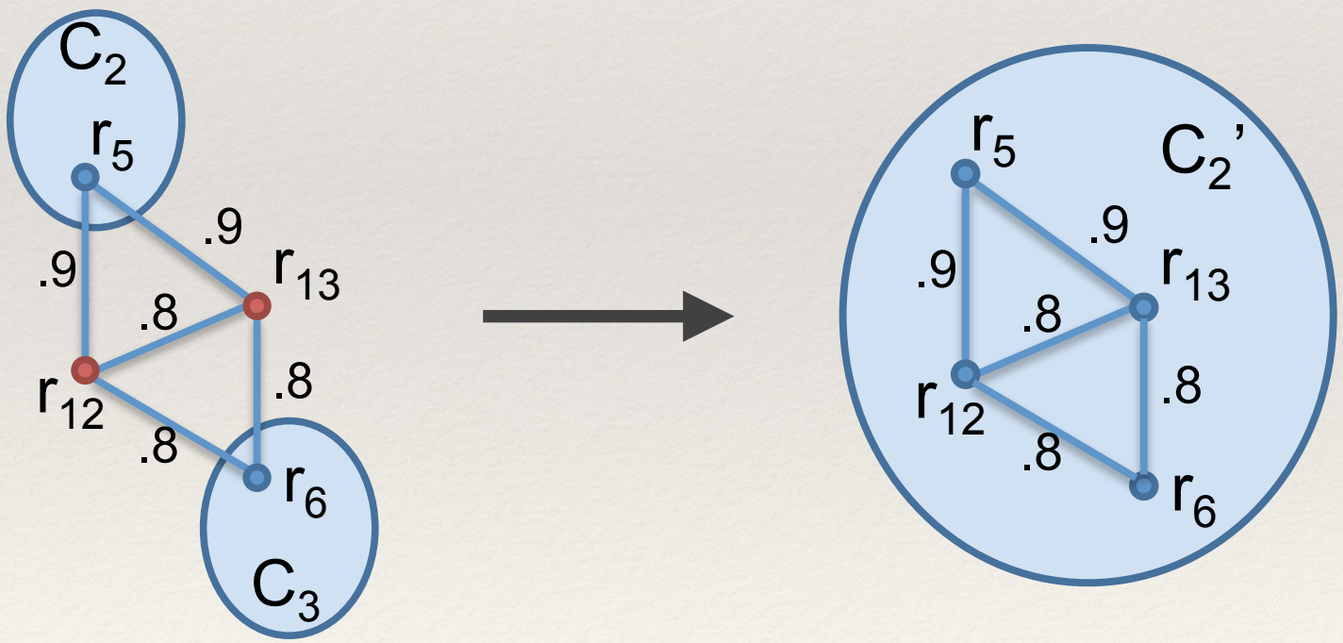
Approximation Approach

- The greedy variation of the iterative approach…
  - uses the iterative mechanism of propagating modifications through modified clusters only.
  - uses a locally optimal decision function to create, merge, split, or move records across clusters.

Greedy Operations

- Merge

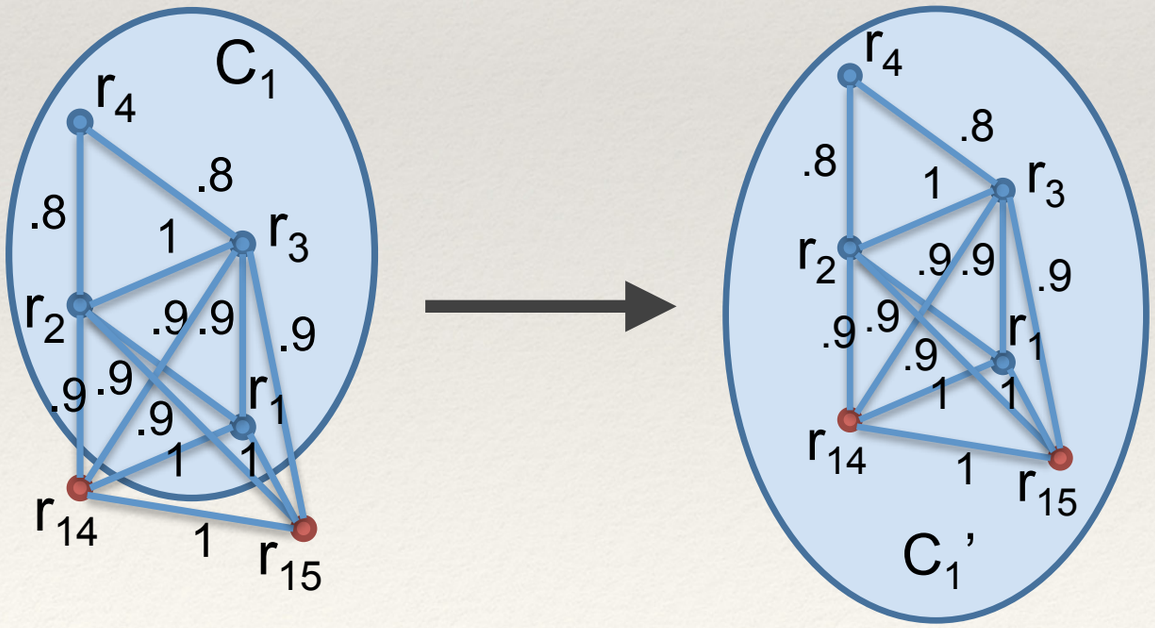
  If the benefits of merging the records into one cluster outweigh the penalties, then merge them.



Greedy Operations

- Split

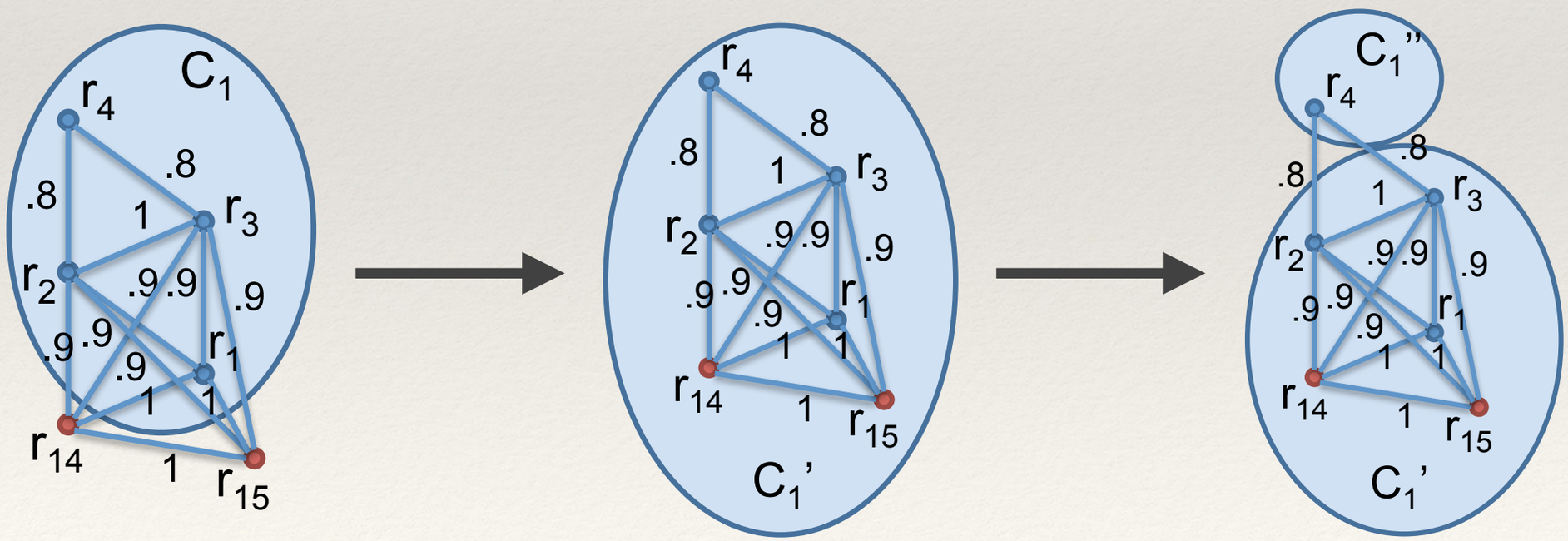
  If the benefits of separating the records in one cluster into two clusters outweigh the penalties, then split them.



Greedy Operations

- Split

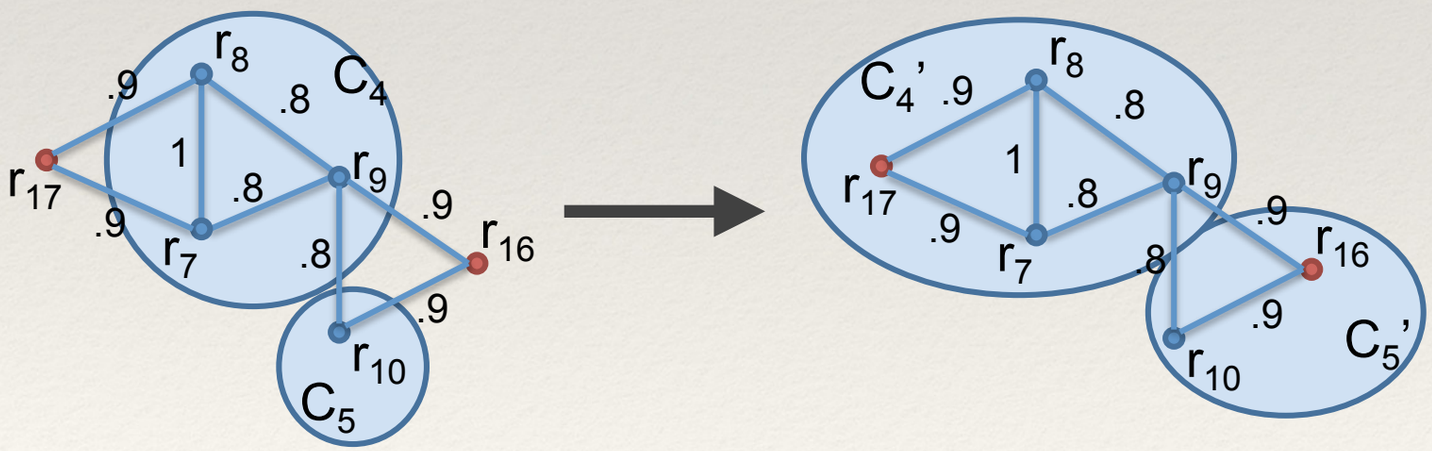
  If the benefits of separating the records in one cluster into two clusters outweigh the penalties, then split them.



Greedy Operations

- Move

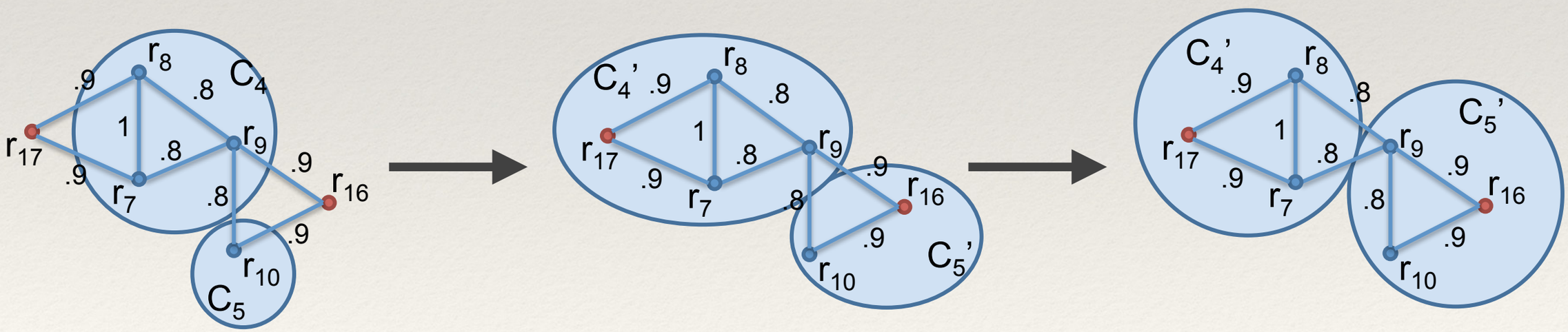
  If removing a record from one cluster and adding it to another decreases the overall penalty, then move the record.



Greedy Operations

- Move

  If removing a record from one cluster and adding it to another decreases the overall penalty, then move the record.



Experiments

- 3 (real-world and synthetic) datasets

  Business dataset - contains records from businesses registered in the SFO area

  Cora dataset - widely used publications dataset

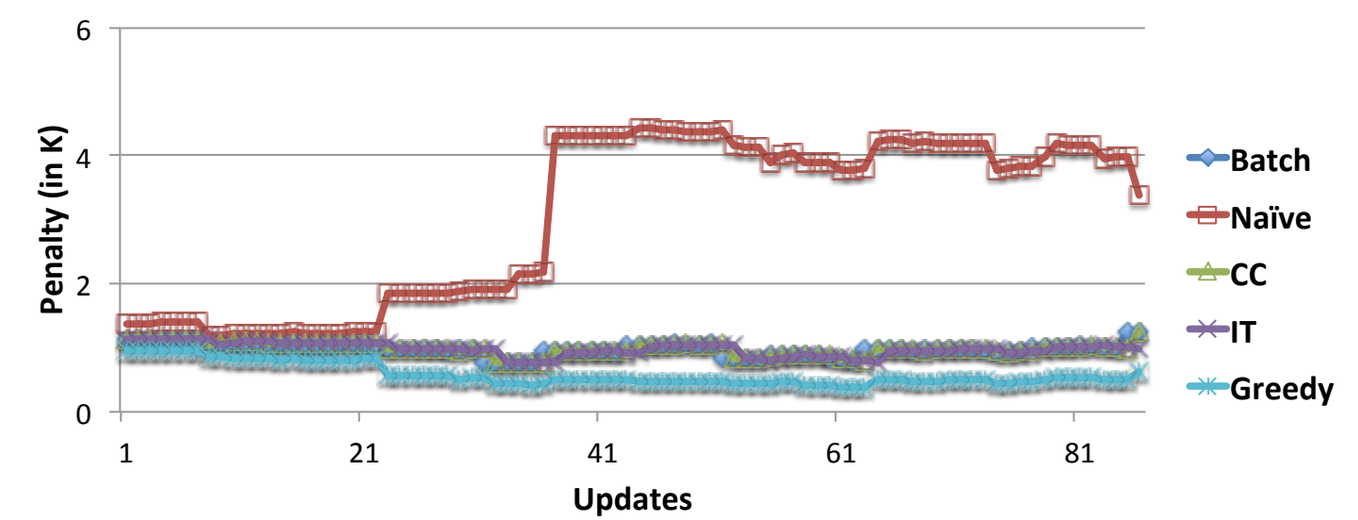
  Febrl dataset - dataset generator

- 2 batch algorithms and 4 incremental approaches

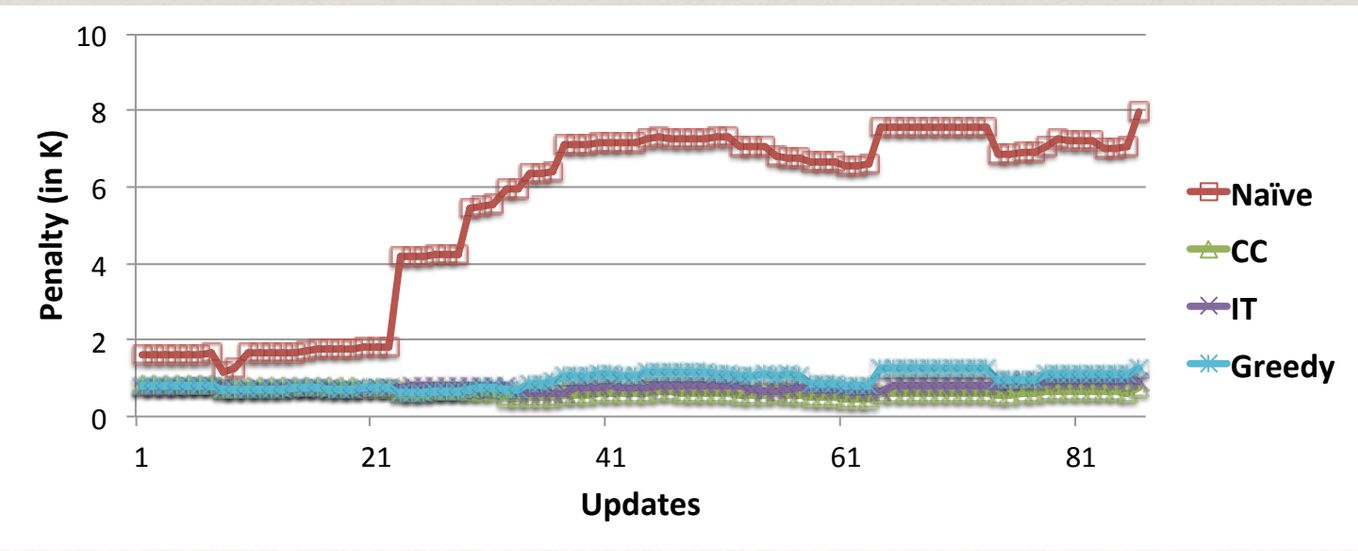
1) Cautious correlation clustering	1) Naive
2) DB-Index	2) Connected component (CC)
	3) Iterative (IT)
	4) Greedy

Experiments: Penalty

Penalty for Business dataset with Correlation Clustering:

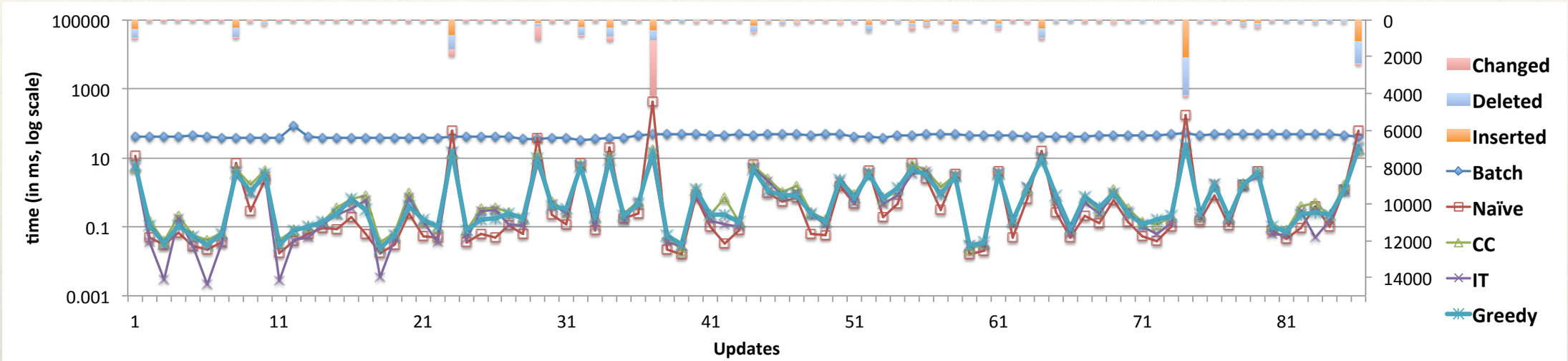


Penalty for Business dataset with DB-Index:

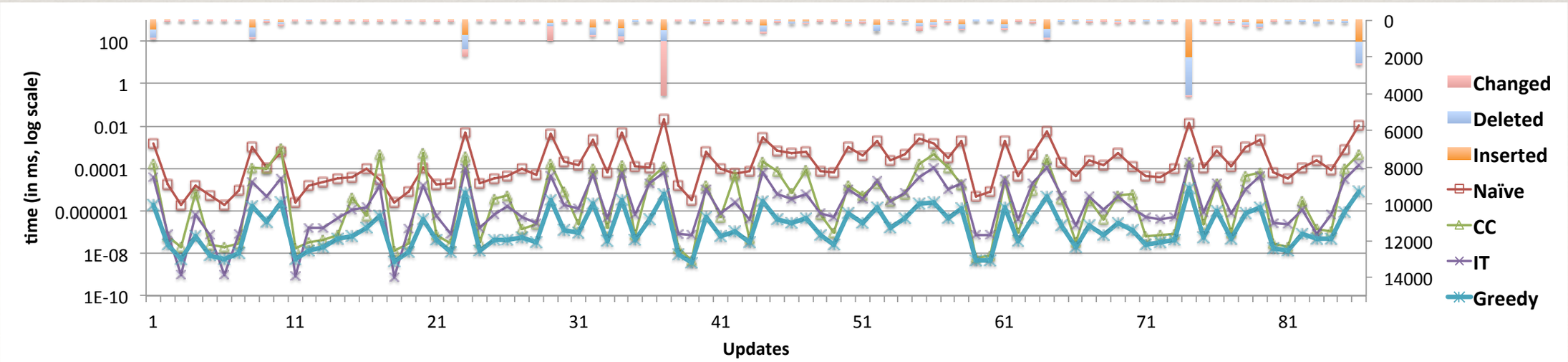


Experiments: Execution Time

Execution time for Business dataset with Correlation Clustering:



Execution time for Business dataset with DB-Index:



Conclusion

- Incremental record linkage is an essential mechanism to improve the overall performance of linkage algorithms.
- The performance and quality trade-offs for incremental record linkage are dependent on the applied objective function.
- Greedy approximations provide a good alternative to optimal incremental record linkage algorithms.

Conclusion

- Incremental record linkage is an essential mechanism to improve the overall performance of linkage algorithms.
- The performance and quality trade-offs for incremental record linkage are dependent on the applied objective function.
- Greedy approximations provide a good alternative to optimal incremental record linkage algorithms.

Thank you!

anja.gruenheid@inf.ethz.ch

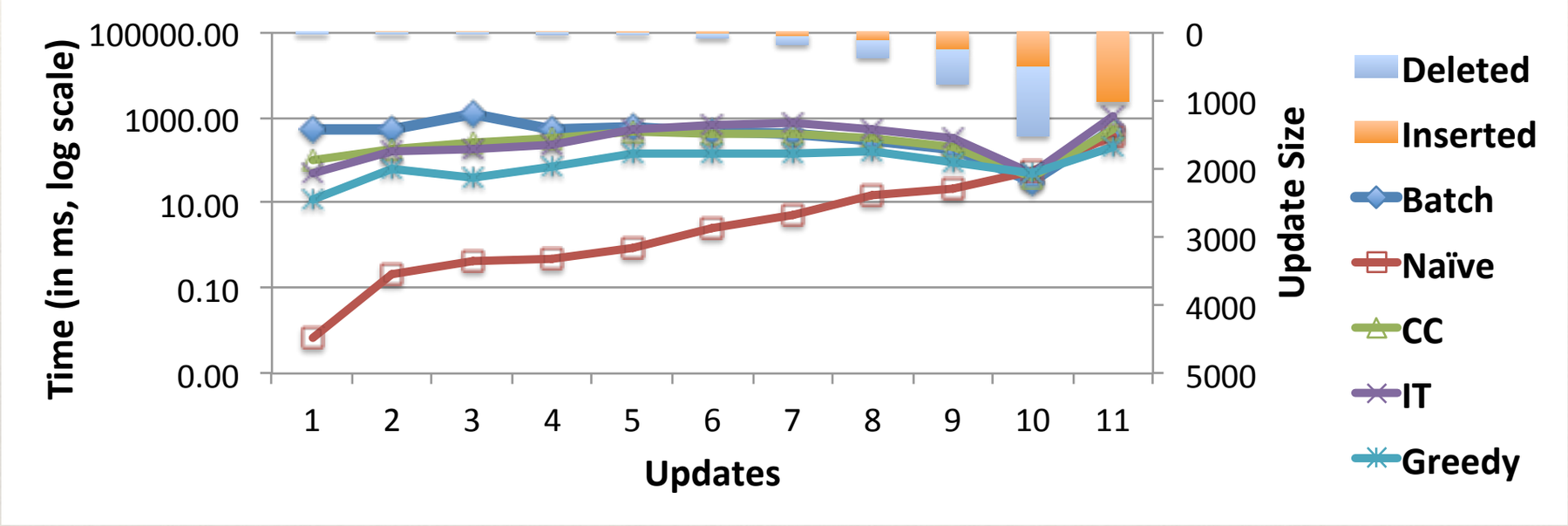
Experiments: Business

Measurements for Business dataset with Correlation Clustering and DB-Index:

Method			Time (s)	Impro.	Penalty
Corr Clust.		Batch	3.7	-	988
	Cont	Naive	.86	76.7%	3037
		CC	.18	78.7%	988
		IT	0.16	81.4%	981
		Greedy	**0.14**	**84.1%**	**592**
	Reset	Naive	0.79	79.7%	1072
		CC	0.20	74.2%	987
		IT	**0.17**	**77.7%**	987
		Greedy	0.20	74.3%	**922**
DB-Index	Cont	Naive	997	99.9%	5426
		CC	57.1	94.3%	**651**
		IT	14.4	98.6%	783
		Greedy	**.79**	**99.9%**	941

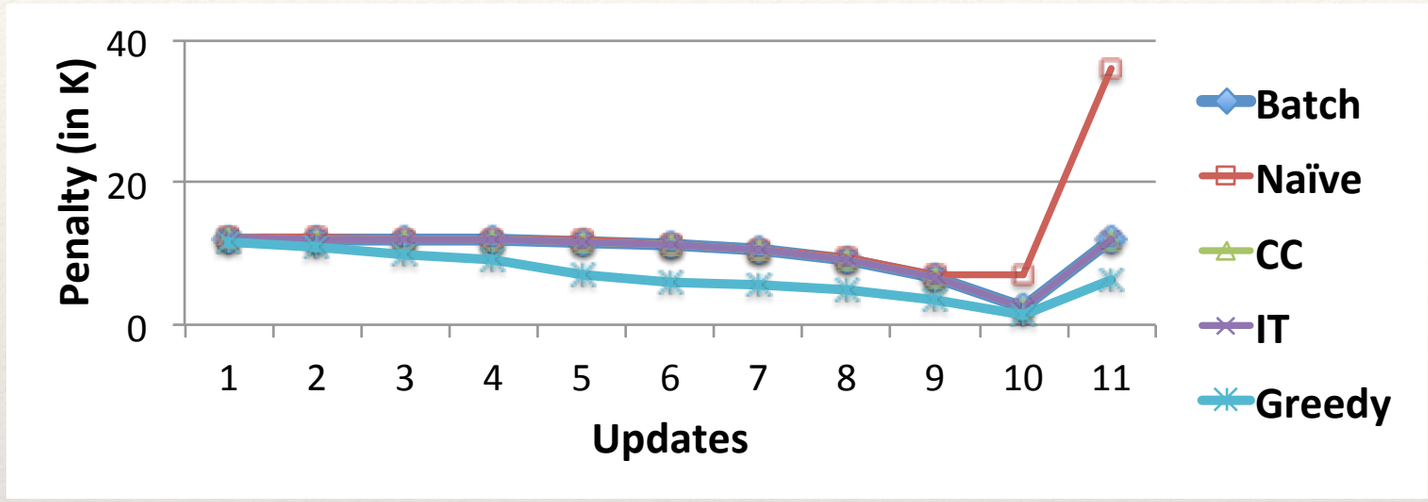
Experiments: Execution Time

Execution time for Cora dataset with Correlation Clustering:

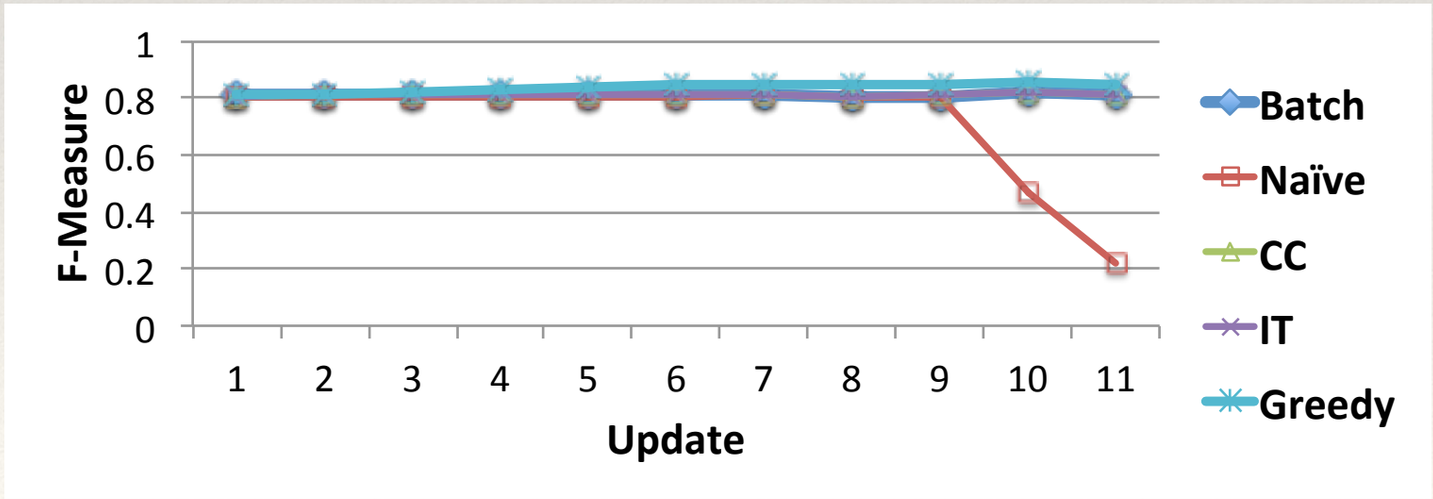


Experiments: Quality

Penalty for Cora dataset with Correlation Clustering:

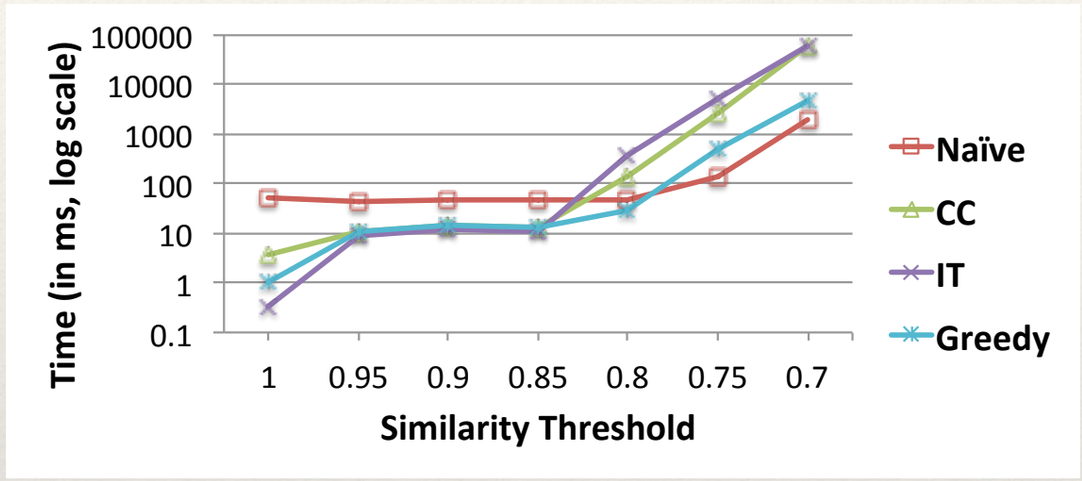


F-Measure for Cora dataset with DB-Index:

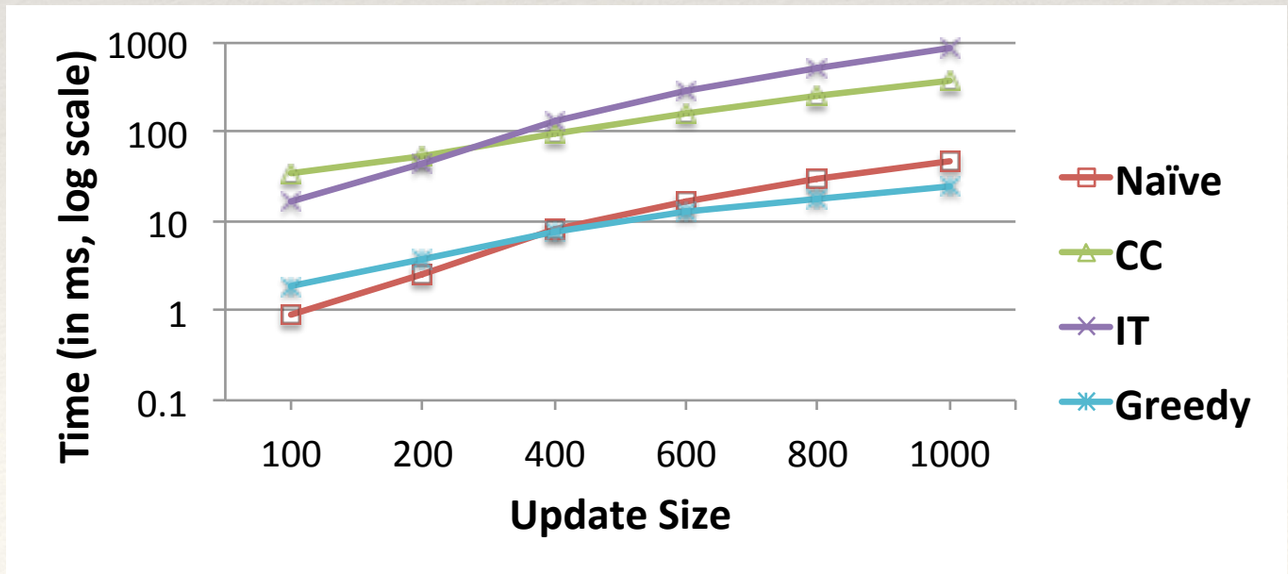


Experiments: Execution Time

Execution time for Febrl dataset with Correlation Clustering and varying similarity thresholds



Execution time for Febrl dataset with Correlation Clustering and varying update sizes



Experiments: Quality

F-Measure for Febrl dataset with Correlation Clustering and varying similarity thresholds