Tecnica CAHD

Correlation Aware Anonymization of High-dimensional data

- Davide Caputo
- Fabio Parodi

Introduzione

- La tecnica proposta fornisce un metodo di anonimizzazione per dati transazionali. I classici metodi che si basano sulla generalizzazione o permutazione (k-Anonymity, l-diversity…) non sono adatti per questo tipo di dati.
- La tecnica CAHD cerca di combinare i vantaggi dei metodi sopracitati anche per dati sparsi e ad alta dimensionalità garantendo un certo grado di privacy e allo stesso tempo mantenendo comunque un buon valore dell'utilità dei dati stessi.

Notazioni

- Insiemi di transazioni da anonimizzare $T = \{t_1, \dots, t_n\}$; $n = |T|$.
- Insieme di item $I = \{i_1, \dots, i_d\}$; $d = |I|$.
- QID, quasi-identifier, se associati a conoscenza esterna può portare all'identificazione dell'individuo.
- S, sensitive item, che rappresentano una minaccia alla privacy qualora riescano ad essere associati ad una certa transazione.

	Wine	Strawberries	Meat	Cream	Pregnancy Test	Viagra
Bob	X		X			X
David	X		X			
Claire		X		X	X	
Andrea		X	X			
Ellen	X		X	X		

(a) Original Data

	Wine	Meat	Cream	Strawberries	Pregnancy Test	Viagra
Bob	X	X				X
David	X	X				
Ellen	X	X	X			
Andrea		X		X		
Claire			X	X	X	

(b) Re-organized Data

	Wine	Meat	Cream	Strawberries	Sensitive Items
Bob	X	X			Viagra: 1
David	X	X			
Ellen	X	X	X		
Andrea		X		X	Pregnancy Test: 1
Claire			X	X	

(c) Published Groups

Fig. 1. Purchase Transaction Log Example

- Gli step della tecnica in esame sono:

1. Ottenere una nuova rappresentazione che sfrutta la sparsità dei dati, effettuando permutazioni di righe e colonne, in modo da ottenere transazioni vicine il più simili possibili (Fig. 1b).
2. Unire le transazioni vicine in gruppi, separando i QID dagli SD in tabelle riassuntive distinte (Fig. 1c), garantendo al tempo stesso il raggiungimento di un certo grado di privacy e il mantenimento dell'utilità dei dati .

Formalmente gli obiettivi posti sono due:

1. Requisito di privacy.

   Nel gruppo G la privacy risulta $p^G = \min_{t=1\ldots m} |G|/f_i^G$

   f_i^G, numero di occorrenze dell'item sensibile *i* nel gruppo G.

   Nell'intero partizionamento P di T $p^P = \min_{G \in P} p^G$

2. Requisito di utilità.

   Nel nostro caso si vuole minimizzare l'errore di ricostruzione. Si fornisce quindi una metrica per valutare il totale delle informazioni perse dopo aver eseguito l'anonimizzazione dei dati

$$KL\_Divergence(Act, Est) = \sum_{\forall cell\ C} Act_c^s \log \frac{Act_c^s}{Est_c^s}$$

1. Rappresentazione Band Matrix

Al fine di semplificare le operazioni di raggruppamento si vuole «avvicinare» transazioni tra loro definite come simili (maggior numero di QID in comune).

La tabella dei dati viene convertita in una matrice A, eseguendo permutazioni di righe e colonne in modo tale che le righe consecutive condividano un grande numero comune di items.

La matrice ottenuta viene detta *Band Matrix*.

E' un problema NP-C.

Band Matrix

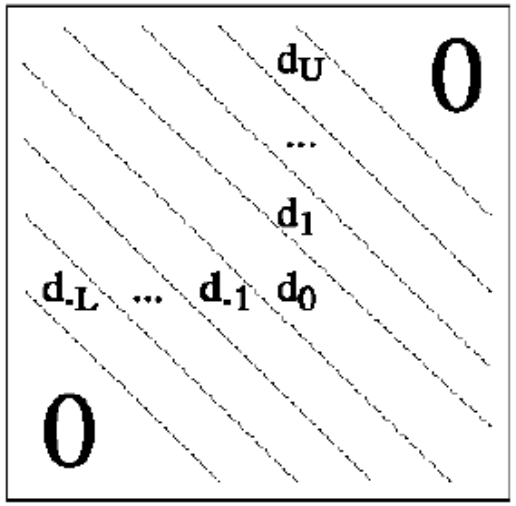
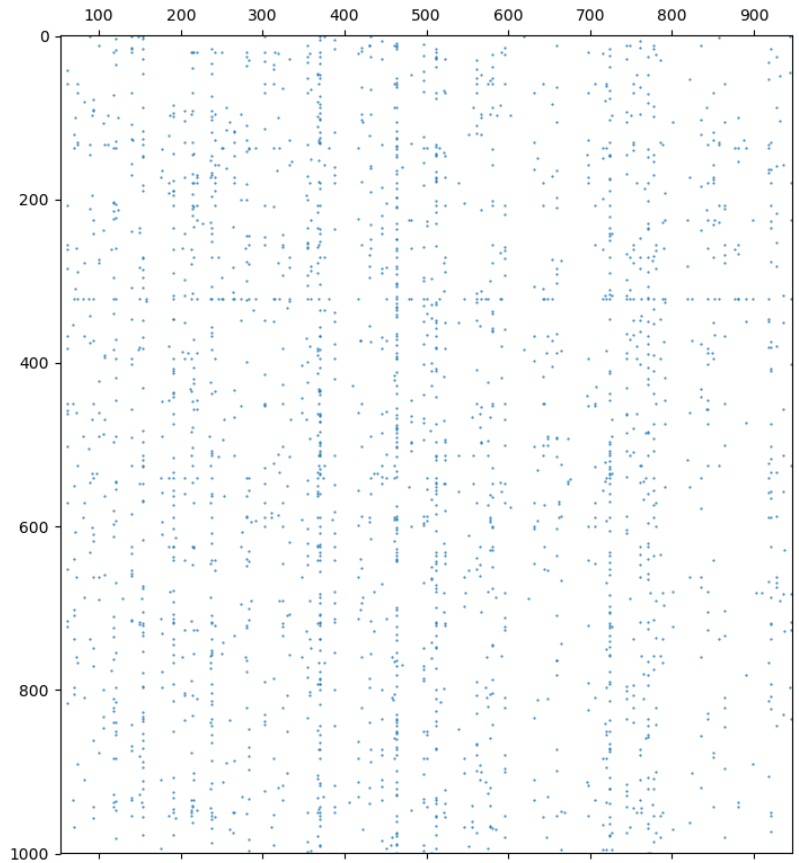


Fig. 3. Band Matrix Representation

La miglior euristica attualmente presente è la *Reverse Cuthill-McKee (RCM) Algorithm.*

Nelle figure sottostanti è possibile osservare la rappresentazione a bande dei dataset analizzati



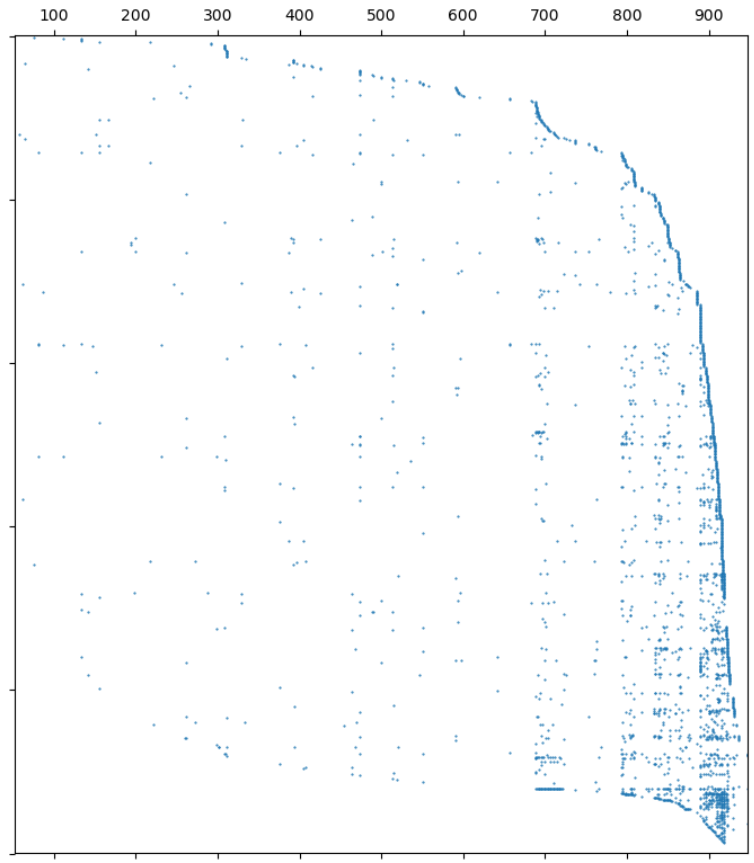


Fig. 1bn Matrice a Bande del dataset BMS1

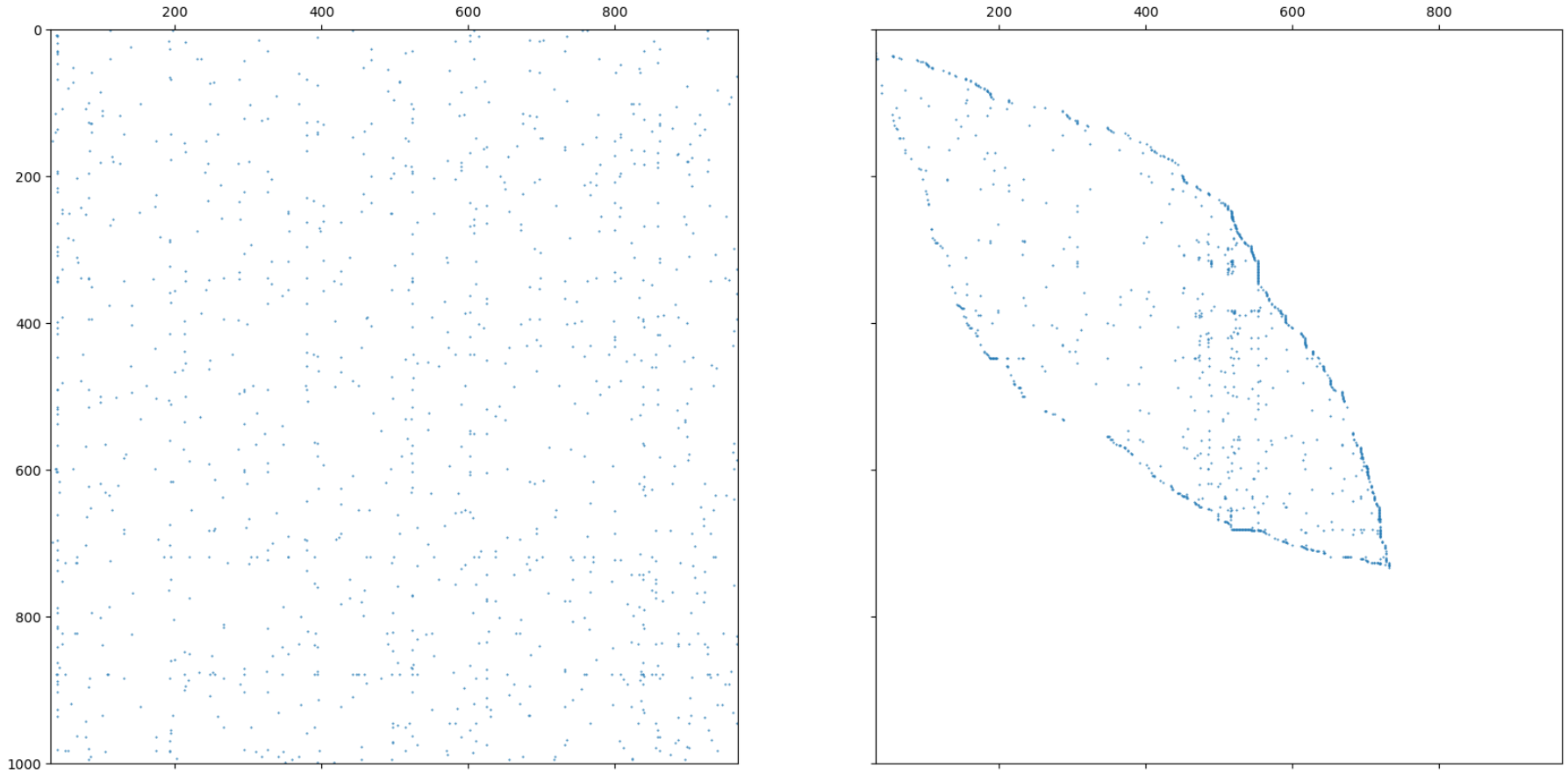


Fig. 2bn Matrice a Bande del dataset BMS2

2. Formazione Gruppi Anonimi

Dopo aver trasformato i dati in accordo con l'euristica RCM, lo step successivo è quello di creare gruppi di transazioni. Per soddisfare il requisito della privacy desiderato, ogni transazione sensibile ha bisogno di essere raggruppata con quelle non sensibili o con altre transazioni che non contengono lo stesso item sensibile.

Si è utilizzato CAHD (Correlation Aware Anonymization of High-dimensional Data), una euristica greedy che si concentra sul mantenimento della correlazione dei dati, raggruppando insieme transazioni che sono vicine nella rappresentazione band matrix.

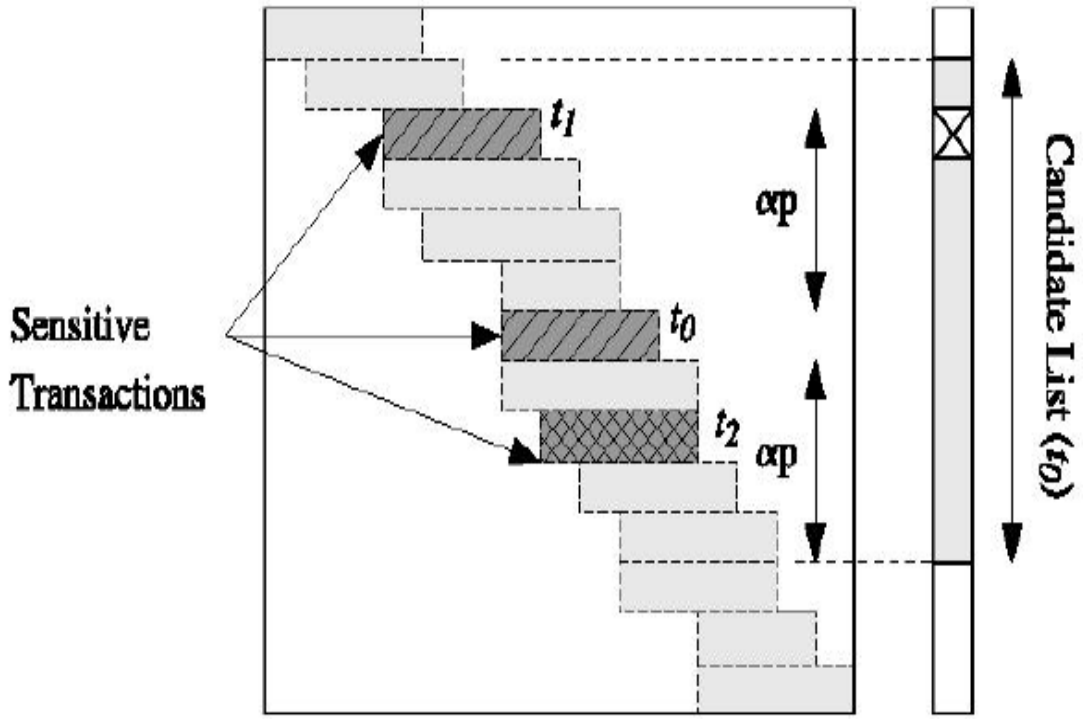


Fig. 7. Group Formation Heuristic

CAHD scansiona l'insieme delle transazioni T nell'ordine delle righe, trova la prima transazione sensibile nella sequenza e cerca di formare un gruppo anonimizzato per essa.

Due transazioni sono in conflitto se presentano almeno un dato sensibile in comune.

Assumiamo di voler anonimizzare t_0 con un grado di privacy p. In questo caso, abbiamo bisogno di raggrupparla con almeno p -1 differenti transizioni.

Viene adottata l'euristica «one-occurence-per-group».

Funzionamento CAHD

Data una transazione sensibile t_0, viene formata una lista di candidate (CL) con αp transazioni che precedono e seguono t_0, che non sono in conflitto con quest'ultima e con nessun'altra all'interno della lista. Più α è grande maggiori sono le probabilità di includere in CL transazioni simili (anche con valori piccoli si possono ottenere buoni risultati grazie all'organizzazione della band matrix).

Il gruppo anonimo viene quindi formato scegliendo le p-1 transazioni che hanno il maggior numero di QID in comune con t_0 tra le $2\alpha p$ in CL, in altre parole le più simili a t_0.

Più le transizioni hanno in comune gli stessi QID più è piccolo l'errore di ricostruzione.

CAHD Group Formation Heuristic

Input: transaction set T, privacy degree p

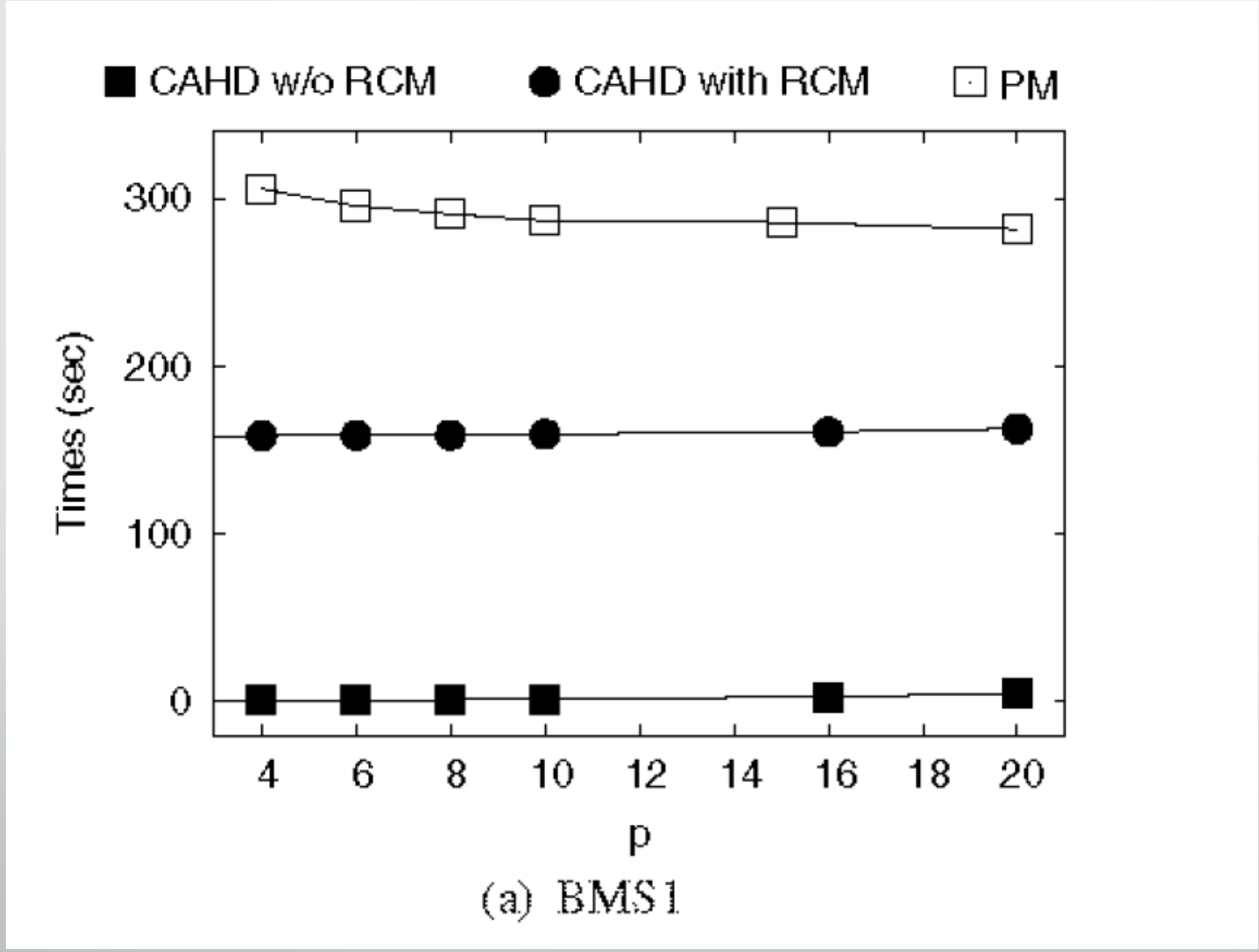
1. initialize histogram H for each sensitive item $s \in S$
2. $remaining = |T|$
3. **while** $(\exists t \in T | t \ is \ sensitive)$ **do**
4. t = next sensitive transaction in T
5. $CL(t)$ = non-conflicting αp pred. and αp succ. of t
6. $G = \{t\} \cup p-1$ trans. in $CL(t)$ with closest QID to t
7. update H for each sensitive item in G
8. **if** $(\nexists s | H[s] \cdot p > remaining)$
9. $remaining = remaining - |G|$
10. **else**
11. roll back G and continue
12. **end while**
13. output remaining transactions as a single group

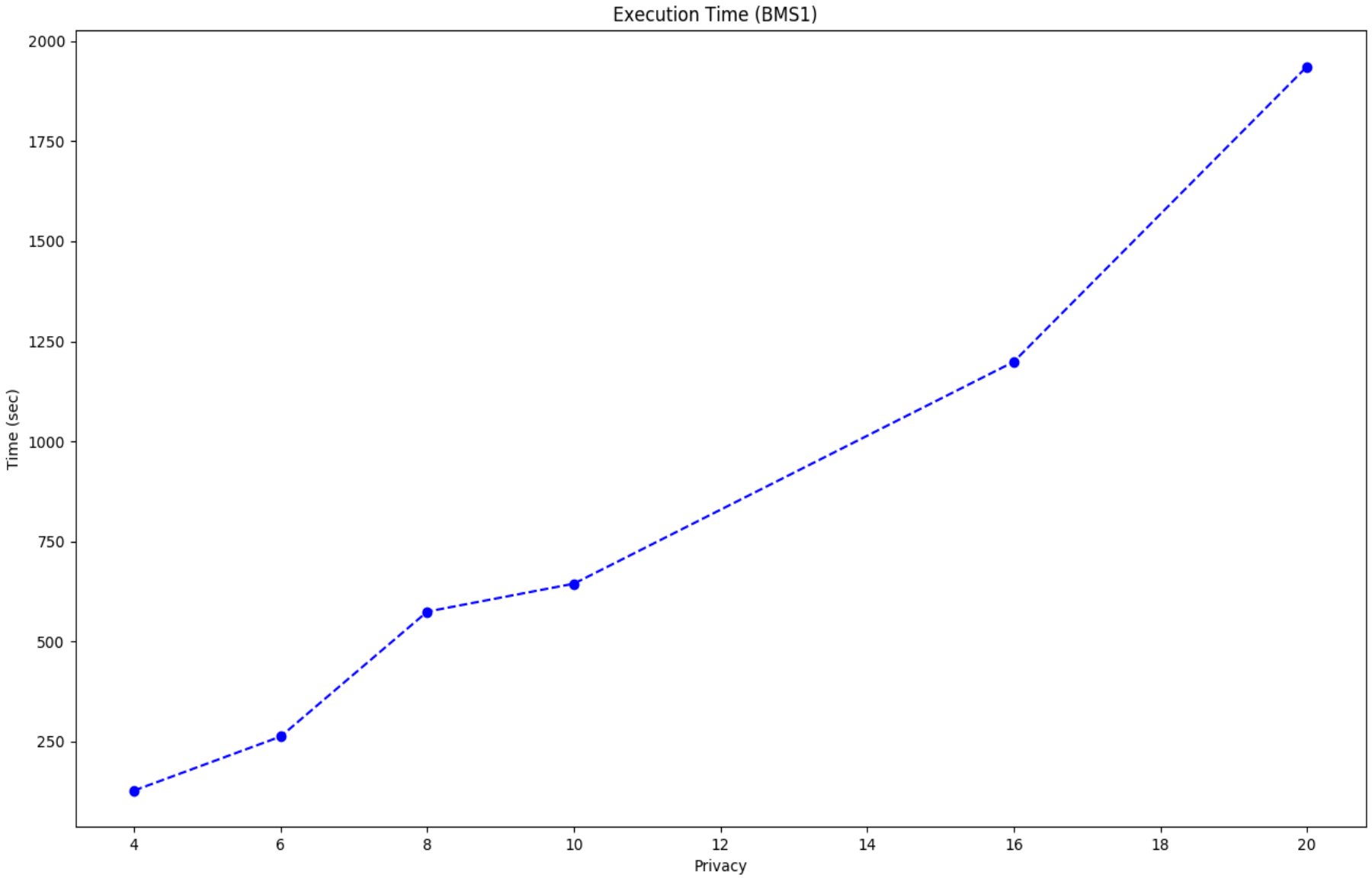
Fig. 8. CAHD Pseudocode

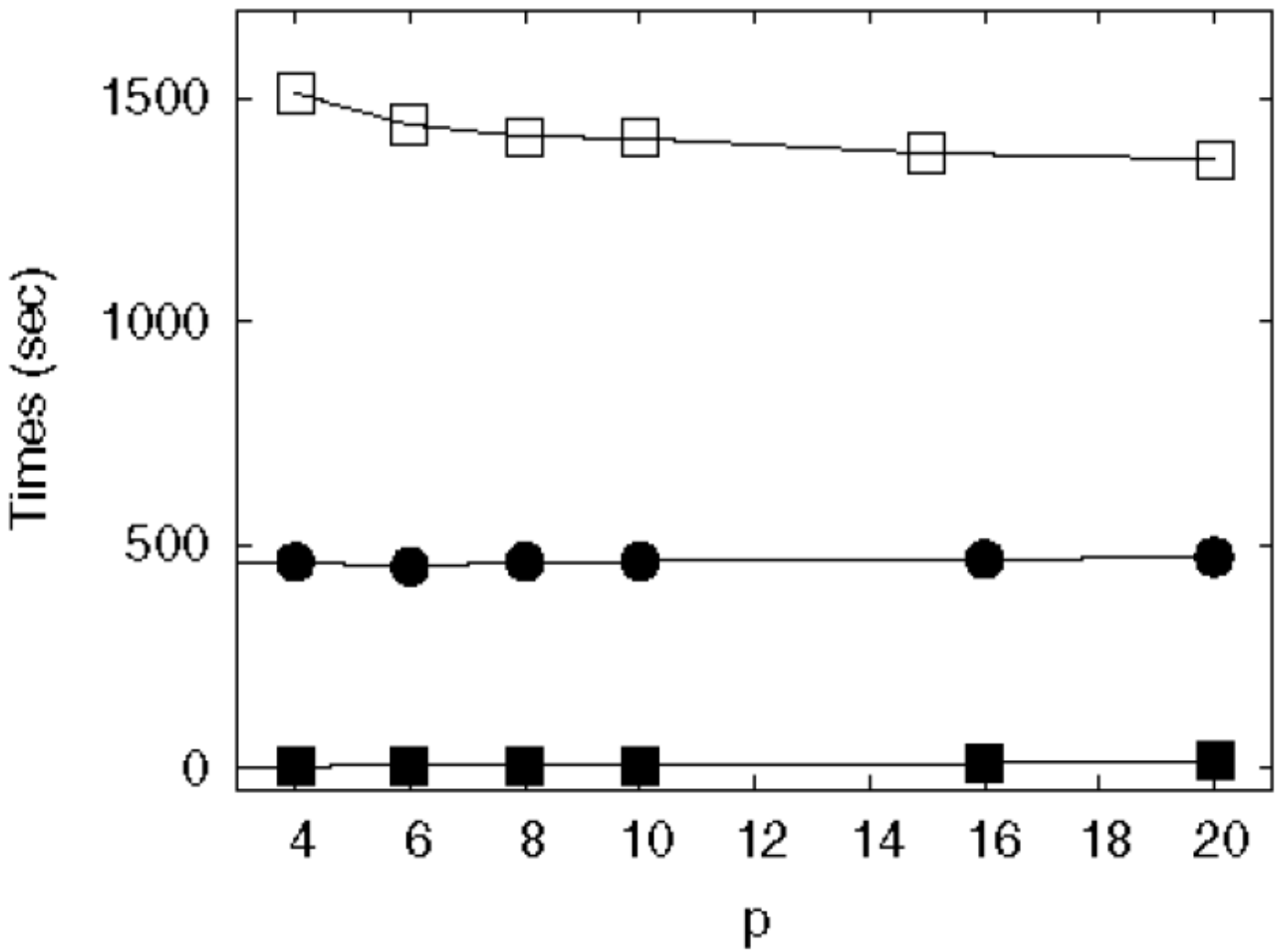
Utilizzando un algoritmo greedy si ha la necessità di garantire che venga trovata una soluzione.

Ogni volta che si forma un gruppo, quindi, non devono rimanere insiemi di transazioni che non possano essere anonimizzate. Per ovviare a questo problema viene mantenuto un istogramma che viene aggiornato ogni volta che un gruppo viene creato.

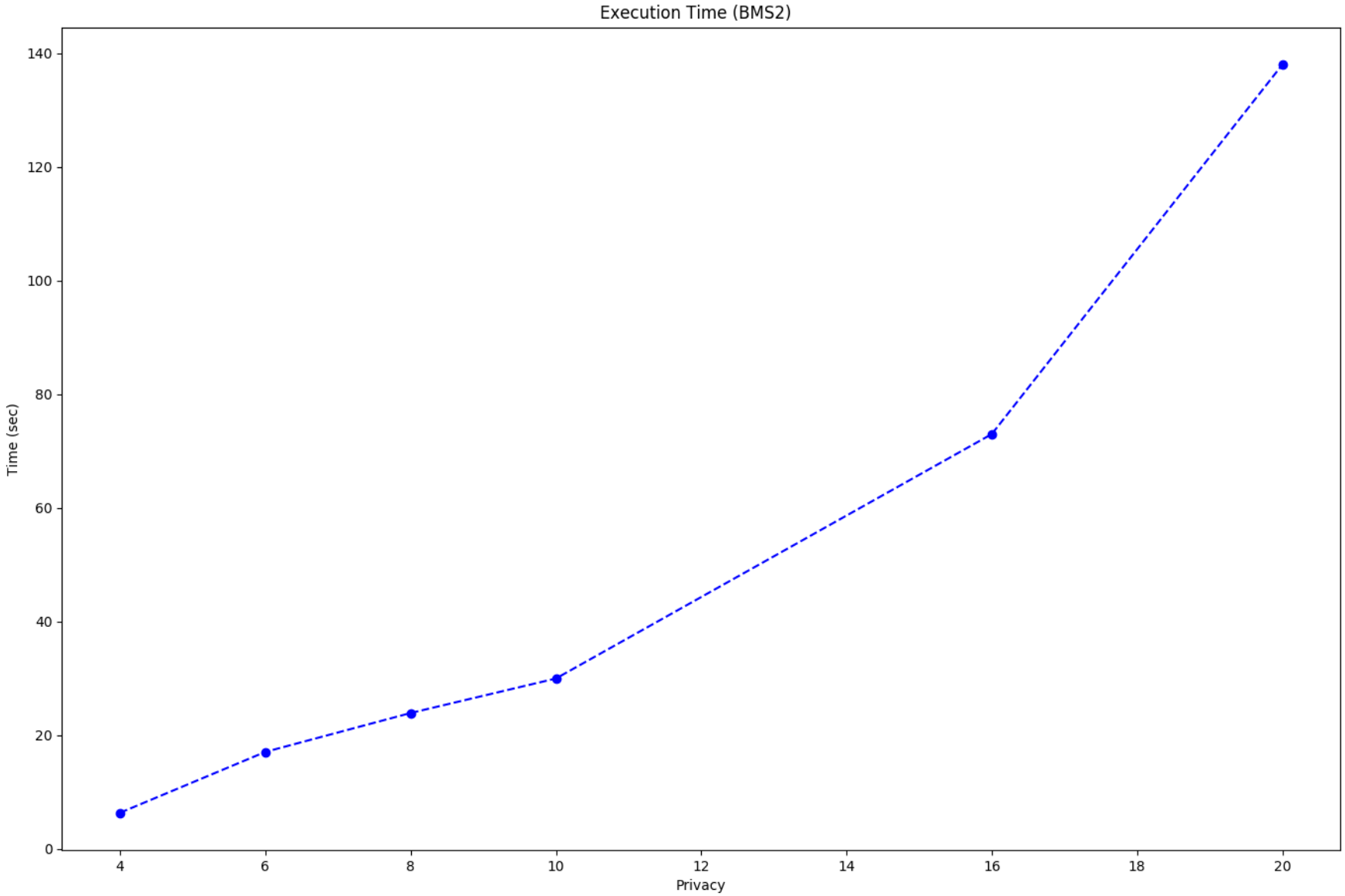
Ogni gruppo deve quindi essere convalidato.

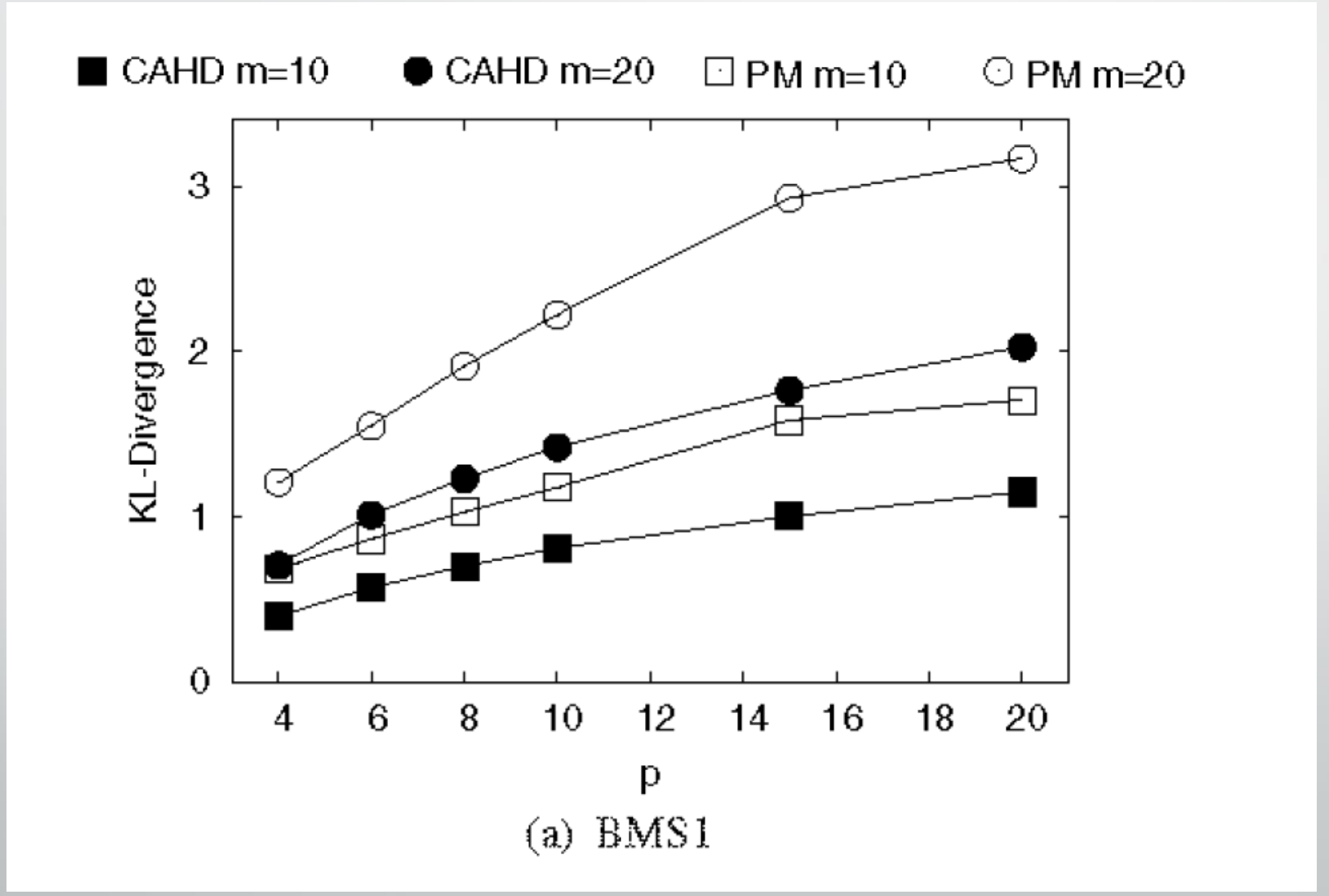

CAHD w/o RCM
CAHD with RCM
PM
Times (sec)
300
200
100
0
4
6
8
10
12
14
16
18
20
p
(a) BMS1


Execution Time (BMS1)
Time (sec)
2000
1750
1500
1250
1000
750
500
250
4
6
8
10
12
14
16
18
20
Privacy

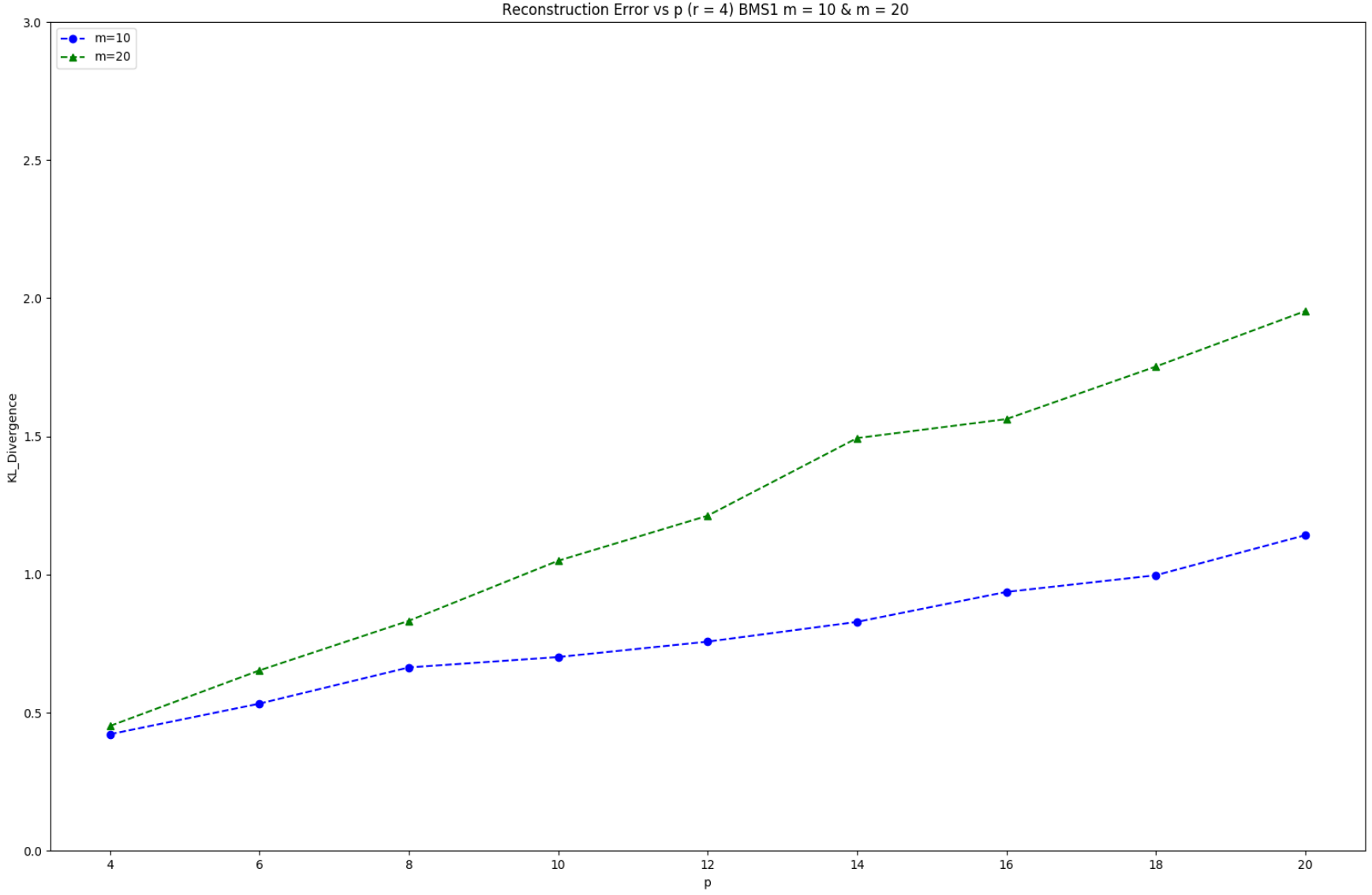


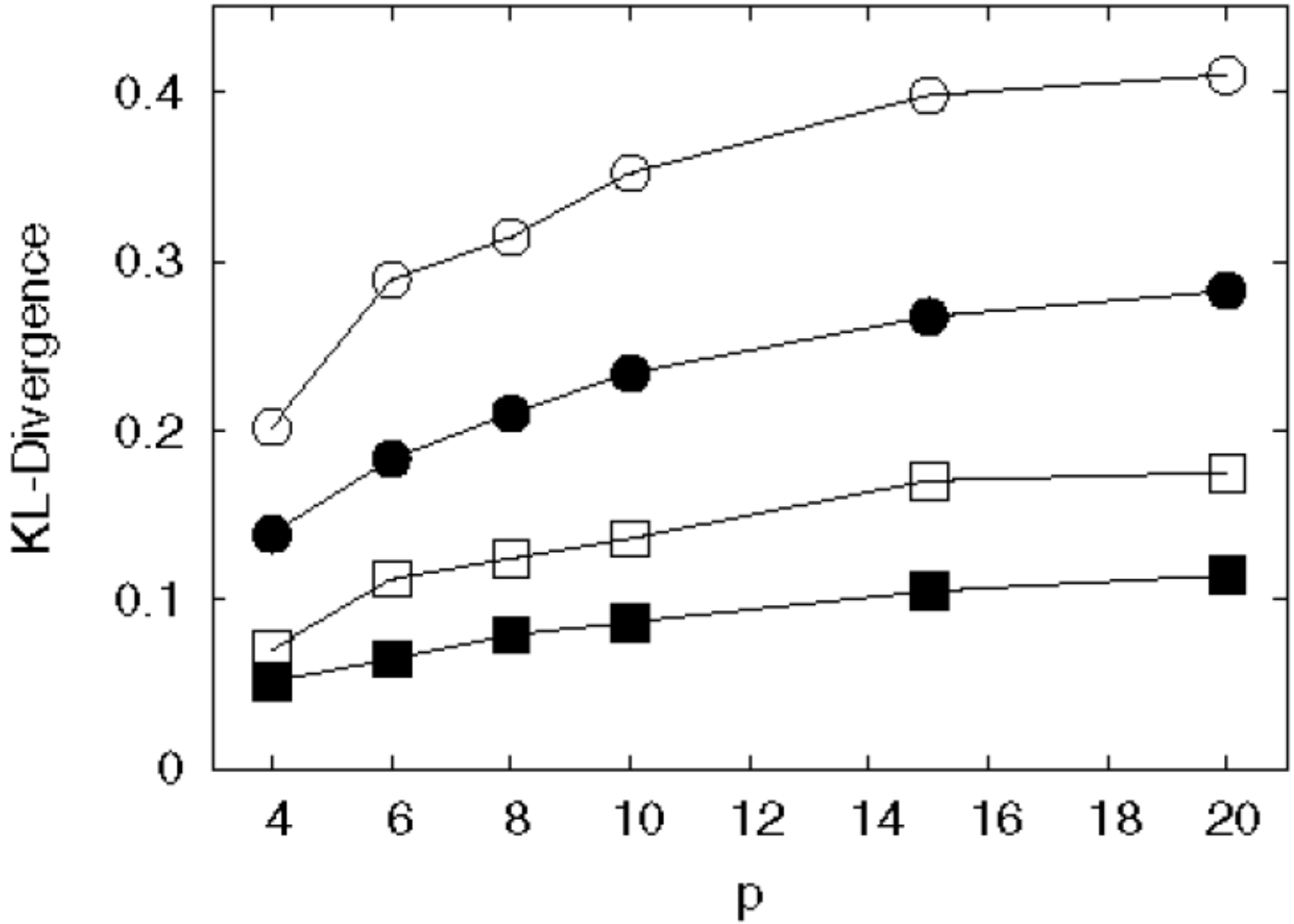
(b) BMS2


Execution Time (BMS2)
Time (sec)
Privacy
0
20
40
60
80
100
120
140
4
6
8
10
12
14
16
18
20

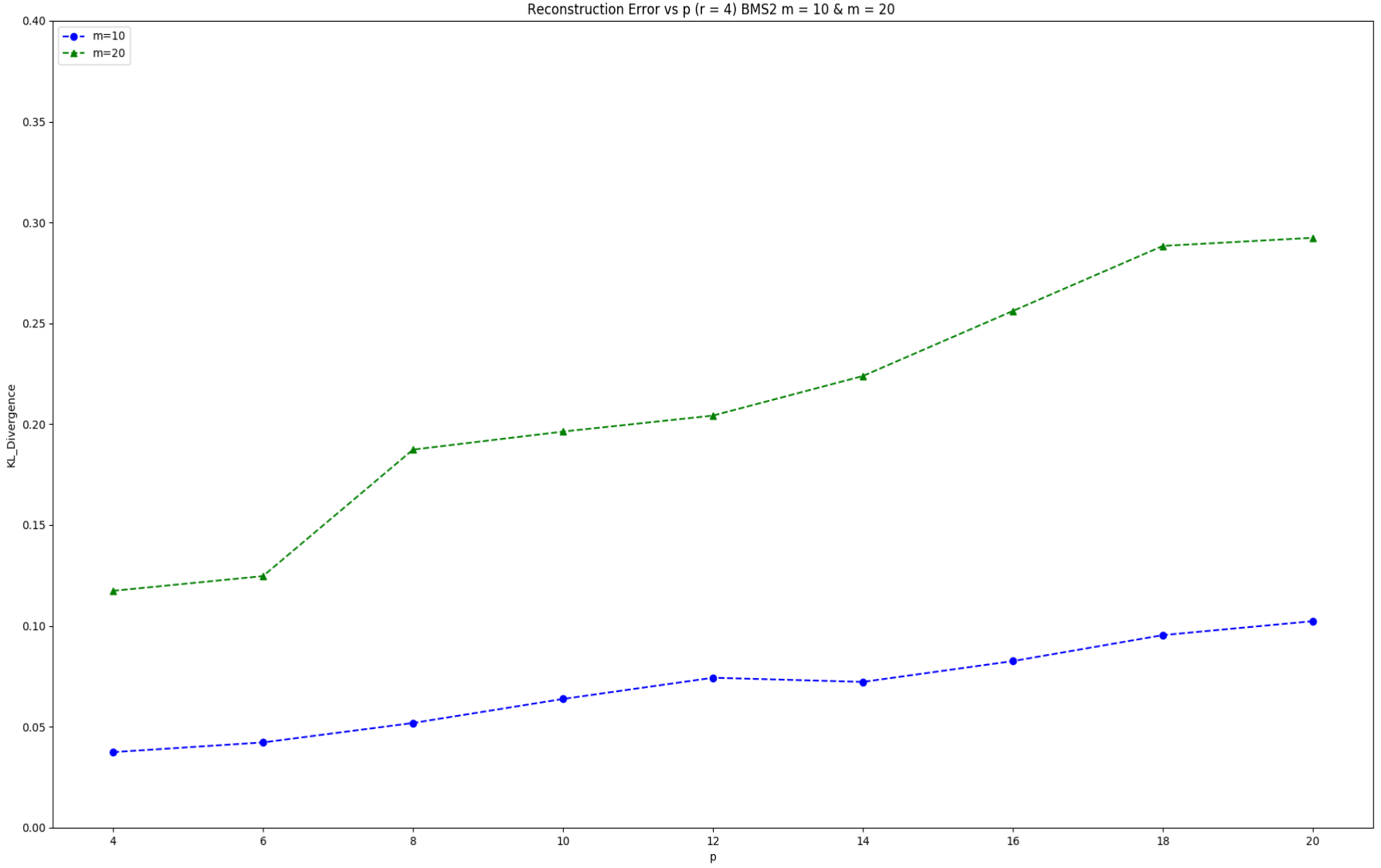

CAHD m=10
CAHD m=20
PM m=10
PM m=20
KL-Divergence
0
1
2
3
4
6
8
10
12
14
16
18
20
p

(a) BMS1


Reconstruction Error vs p (r = 4) BMS1 m = 10 & m = 20
m=10
m=20
KL_Divergence
p
0.0
0.5
1.0
1.5
2.0
2.5
3.0
4
6
8
10
12
14
16
18
20



(b) BMS2


Reconstruction Error vs p (r = 4) BMS2 m = 10 & m = 20
m=10
m=20
KL_Divergence
p
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
4
6
8
10
12
14
16
18
20