Appendix C, Part 1: On-line Tables and Figures for El Emam et al., "A Globally Optimal k-Anonymity Method for the De-identification of Health Data"

Description	Quasi-identifiers	No. Records	No. of Nodes
Adult The adult dataset from the UC Irvine machine learning data repository. This is an extract from the US census: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/adult	• Age (3) • Profession (2) • Education (2) • Marital status (2) • Position (2) • Race (1) • Sex (1) • Country (3)	30,162	5,184
FARS Department of Transportation Fatal crash information: http://www-fars.nhtsa.dot.gov/main.cfm	• Age (4) • Race (1) • Month of Death (3) • Day of Death (2)	101,034	120
CUP Data from the Paralyzed Veterans Association on veterans with spinal cord injuries or disease: http://kdd.ics.uci.edu/databases/kddcup98/kddcup98.html	• ZIP code (5) • Age (4) • Gender (2) • Income (3)	63,441	360
Pharm Prescription records from the Children's Hospital of Eastern Ontario pharmacy for 18 months. This is for inpatients only and excludes acute cases. A de-identified version of this data is disclosed to commercial data aggregators.	• Age (4) • Postal code (FSA) (3) • Admission date (6) • Discharge date (6) • Sex (1)	7,318	2,352
ED Emergency department records from Children's Hospital of Eastern Ontario for July 2008. This data is disclosed for the purpose of disease outbreak surveillance and bio-terrorism surveillance.	• Admission date (3) • Admission time (7) • Postal Code (7) • Date of Birth (6) • Sex (1)	4,090	3,584
Niday A registry of all newborns in Ontario for 2006-2007: https://www.nidaydatabase.com/info/index.shtml	• Maternal postal code (7) • Baby DoB (4) • Mother DoB (6) • Baby sex (2) • Aboriginal status (2) • Primary language (2)	124,933	7,560

Table 1: Summary of the six datasets that we used in our comparative evaluation. For each dataset the quasi-identifiers used and the height of the generalization hierarchy are shown in the second column, the number of observations in the original dataset in column 3, and the number of nodes in the overall lattice in the last column.

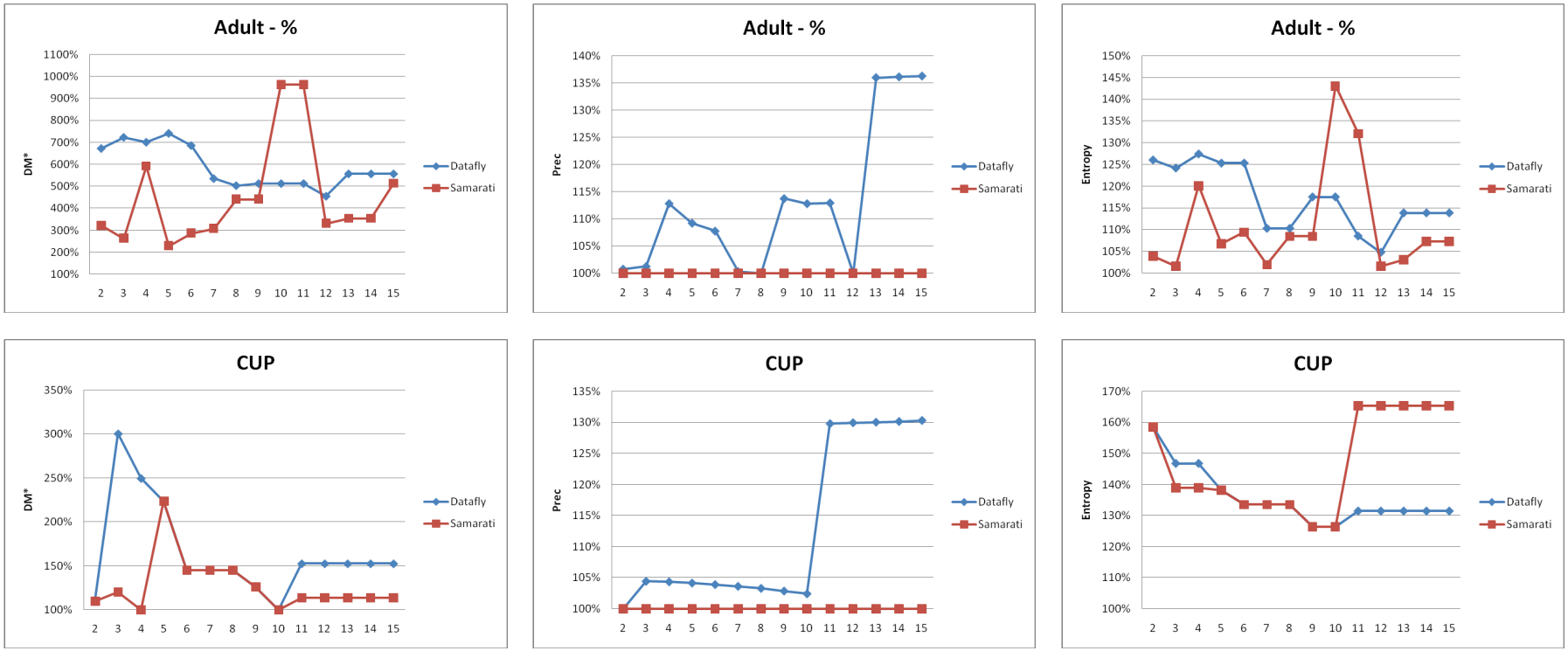


Figure 6a: Information loss results for the six datasets. The x-axis shows the value of k as it was varied from 2 to 15. The y-axis shows the information loss relative to our algorithm. The graphs show the results for a 5% suppression limit.

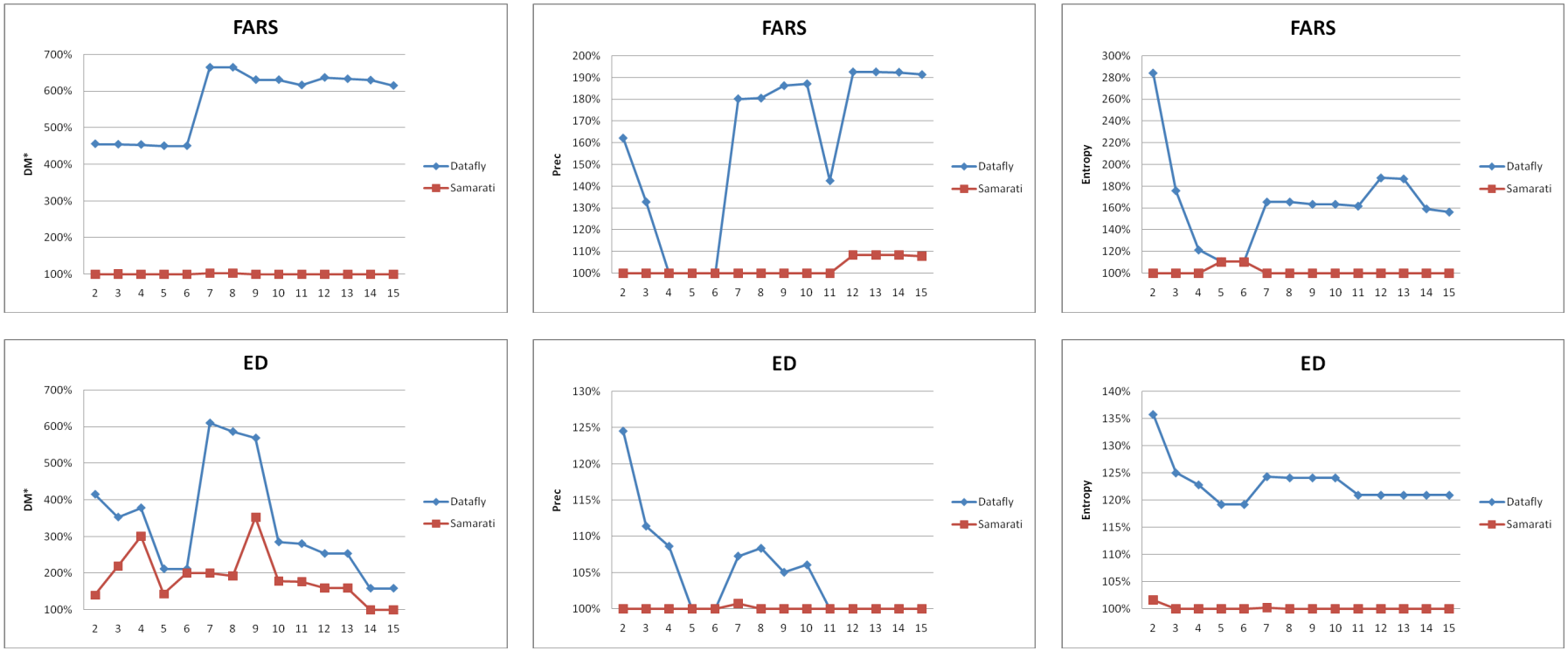


Figure 6b: Information loss results for the six datasets. The x-axis shows the value of k as it was varied from 2 to 15. The y-axis shows the information loss relative to our algorithm. The graphs show the results for a 5% suppression limit.

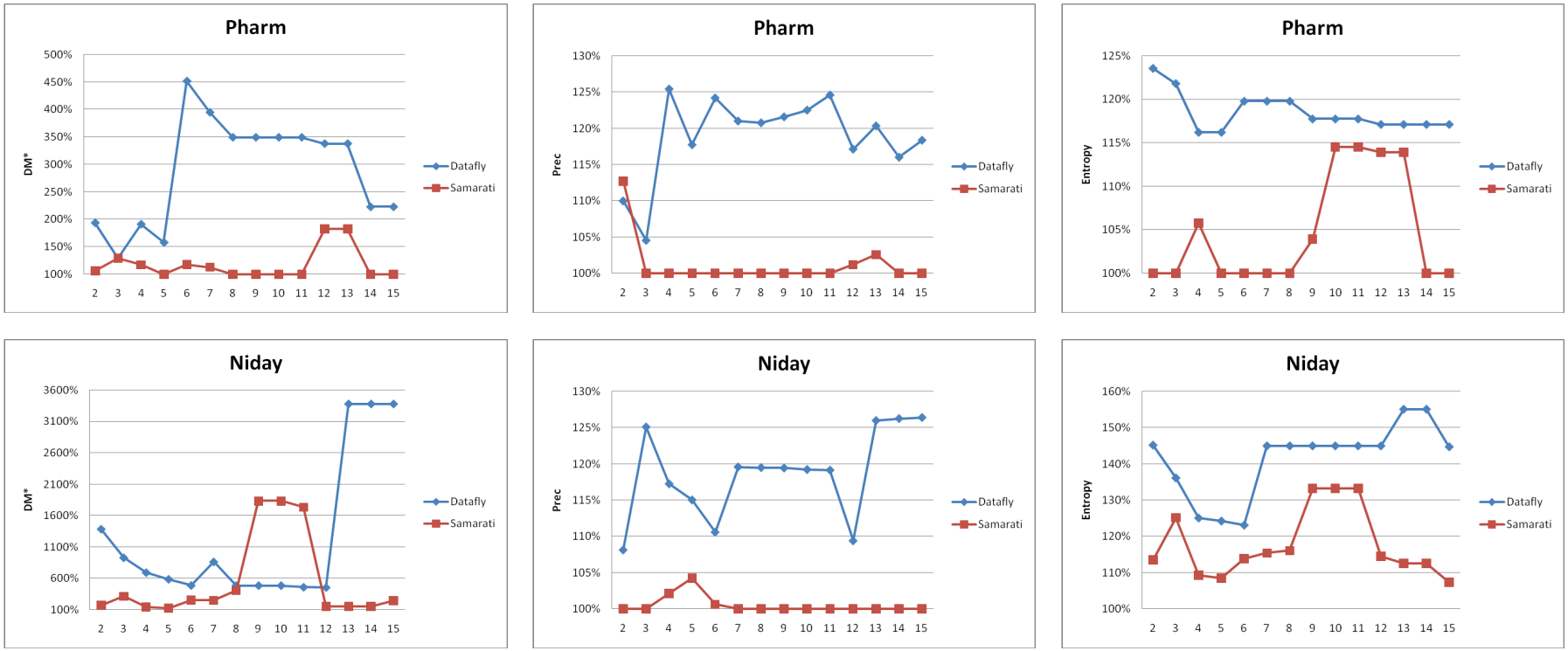


Figure 6c: Information loss results for the six datasets. The x-axis shows the value of k as it was varied from 2 to 15. The y-axis shows the information loss relative to our algorithm. The graphs show the results for a 5% suppression limit.

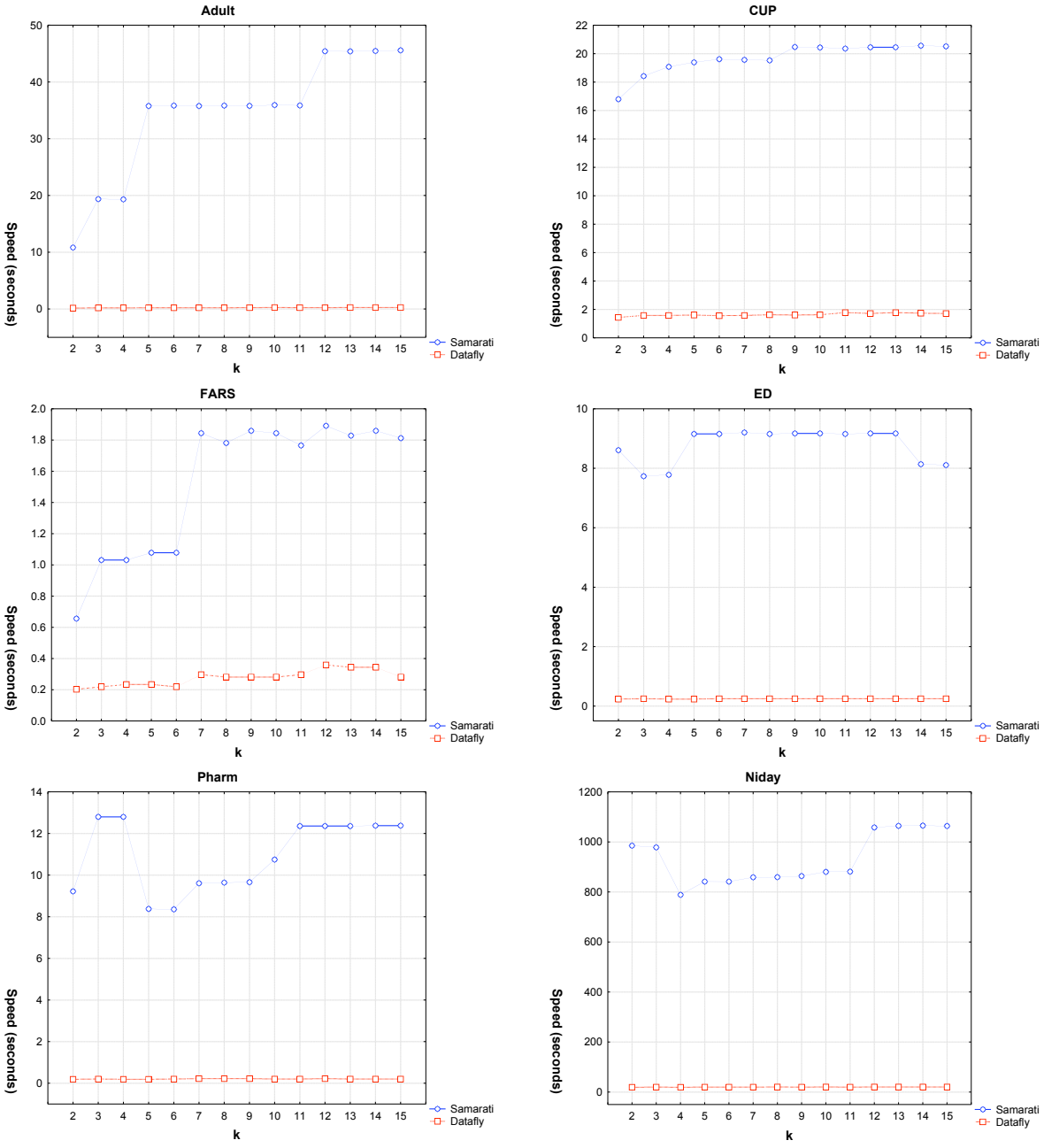


Figure 7: The speed to obtain a solution for Datafly and Samarati on the six datasets across values of k 2 to 15 and a maximum suppression value of 5%. The timing is in seconds and includes all pre- and post- processing that may be needed to run each algorithm. These results were obtained on a dual core 3.2 GHz Pentium D processor with 3GB of RAM.

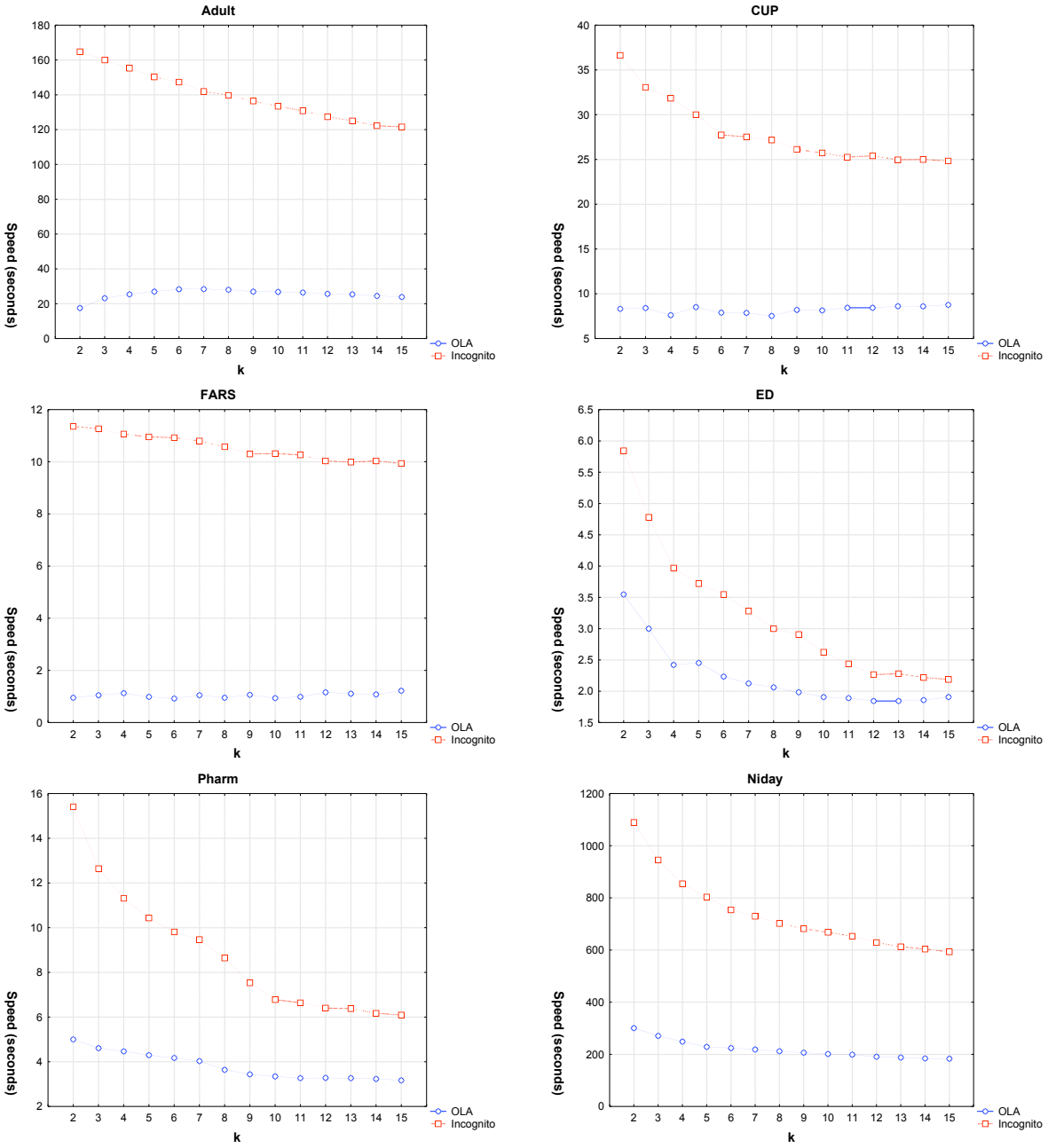
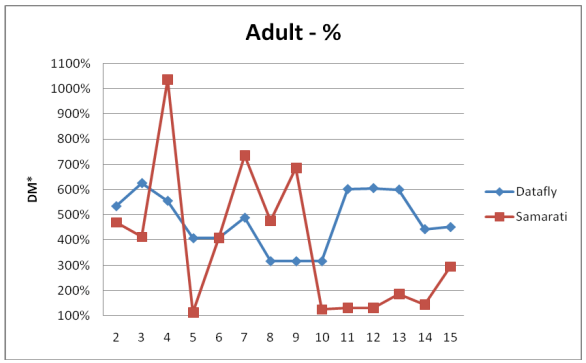


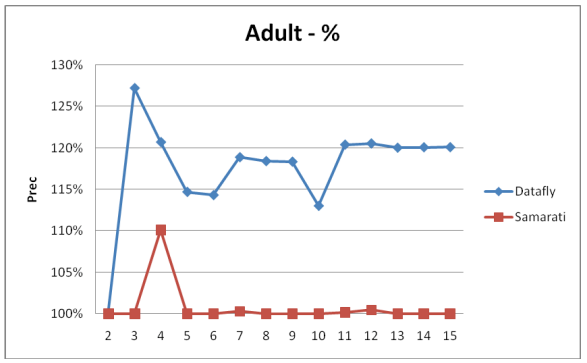
Figure 9: The speed to obtain a globally optimal solution using OLA and Incognito on the six datasets across values of k 2 to 15 and a maximum suppression value of 5%. The timing is in seconds and includes all pre- and post- processing that may be needed to run each algorithm. These results were obtained on a dual core 3.2 GHz Pentium D processor with 3GB of RAM.

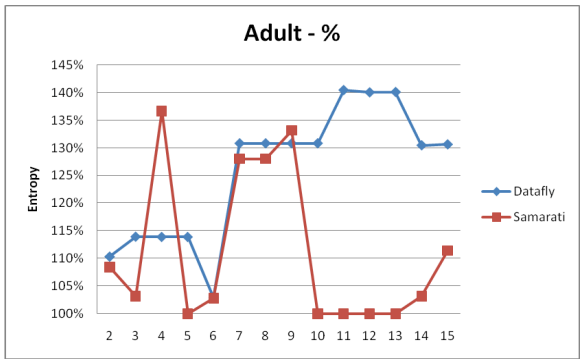
Appendix C, Part 2: Results for 1% and 10% Suppression Limits

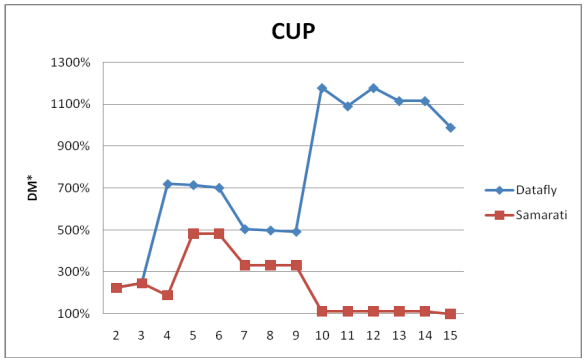
In this section we present all of the results graphs for the 1% and 10% suppression limits. The way these graphs should be interpreted is described in the main body of the paper. These results are consistent with the 5% suppression limit results.

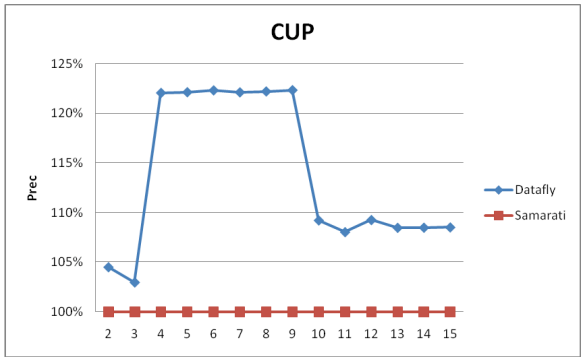
Information Loss Results for the 1% Suppression Limit

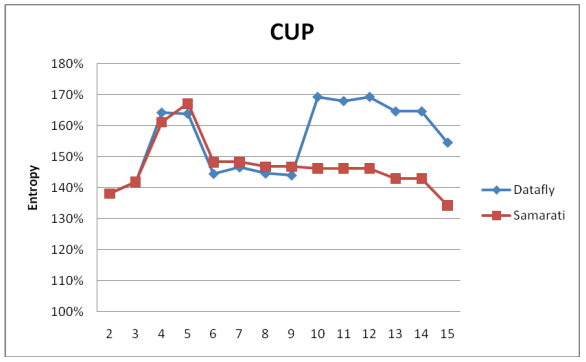


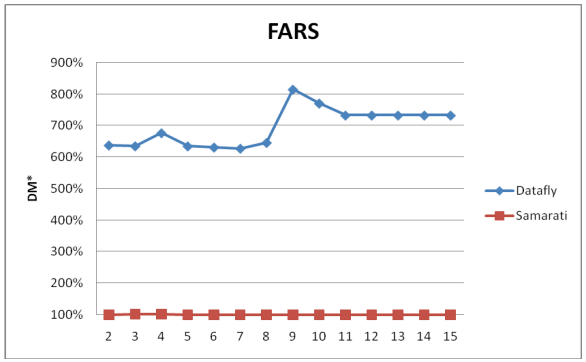


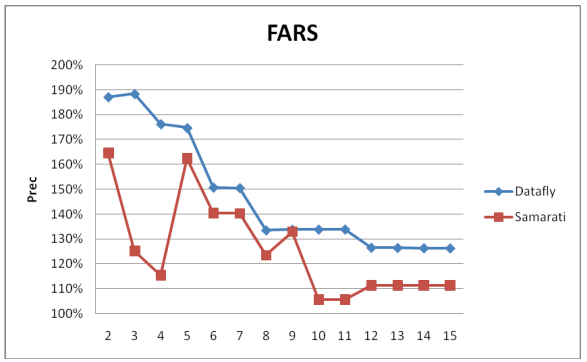


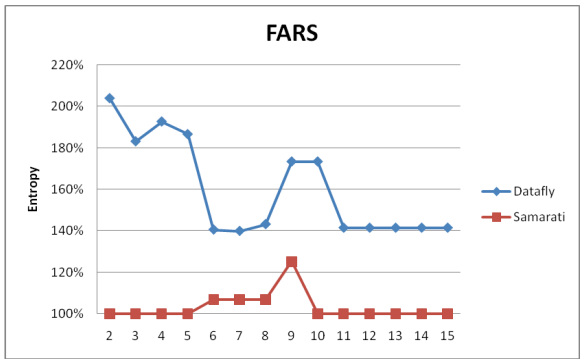


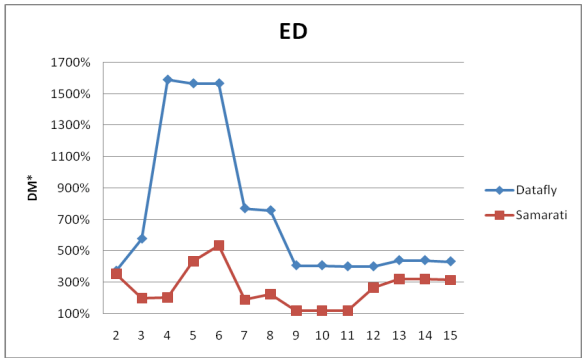


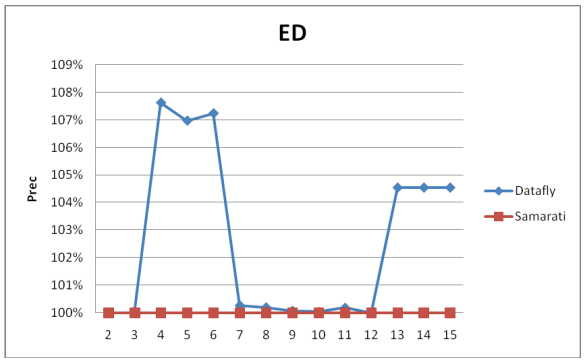


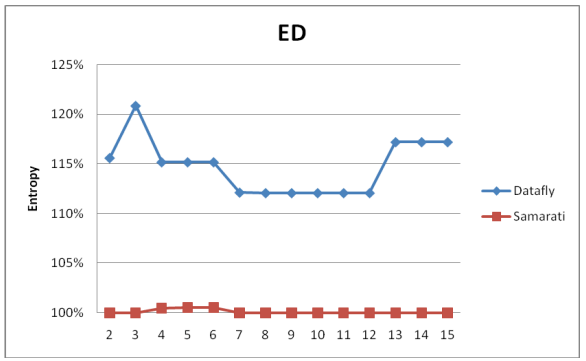

FARS
DM*
900%
800%
700%
600%
500%
400%
300%
200%
100%
2 3 4 5 6 7 8 9 10 11 12 13 14 15
Datafly
Samarati


FARS
Prec
200%
190%
180%
170%
160%
150%
140%
130%
120%
110%
100%
2 3 4 5 6 7 8 9 10 11 12 13 14 15
Datafly
Samarati


FARS
Entropy
220%
200%
180%
160%
140%
120%
100%
2 3 4 5 6 7 8 9 10 11 12 13 14 15
Datafly
Samarati


ED
DM*
1700%
1500%
1300%
1100%
900%
700%
500%
300%
100%
2 3 4 5 6 7 8 9 10 11 12 13 14 15
Datafly
Samarati


ED
Prec
109%
108%
107%
106%
105%
104%
103%
102%
101%
100%
2 3 4 5 6 7 8 9 10 11 12 13 14 15
Datafly
Samarati


ED
Entropy
125%
120%
115%
110%
105%
100%
2 3 4 5 6 7 8 9 10 11 12 13 14 15
Datafly
Samarati

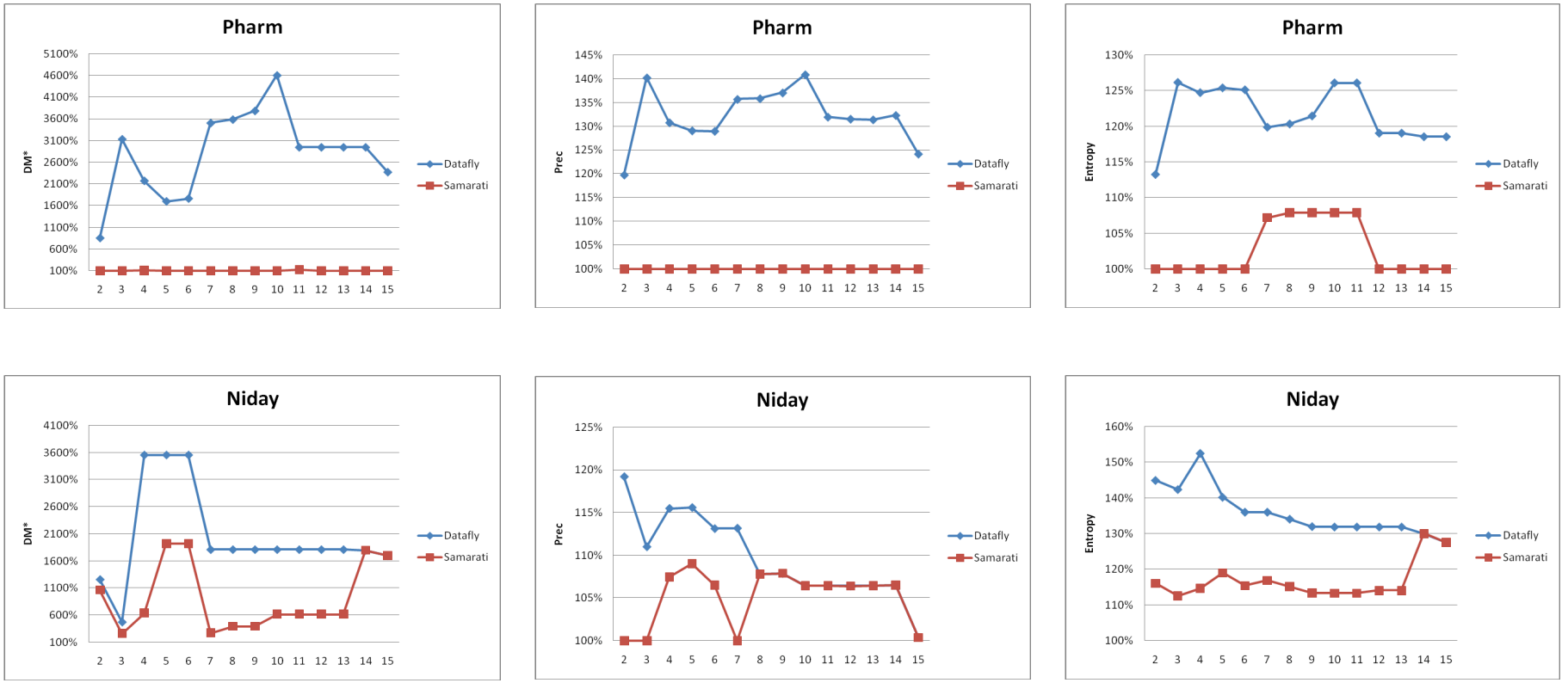
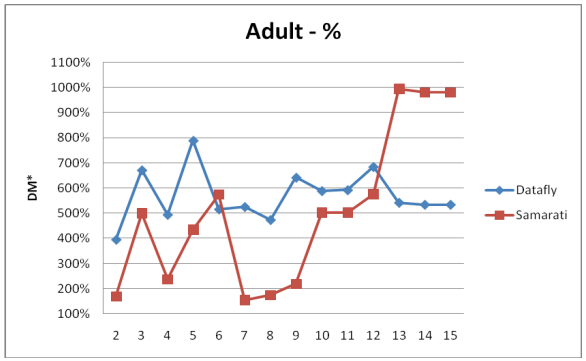
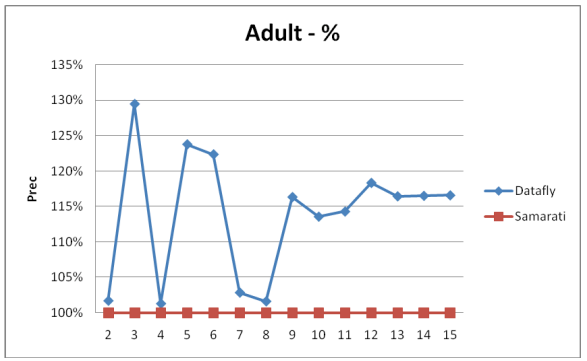
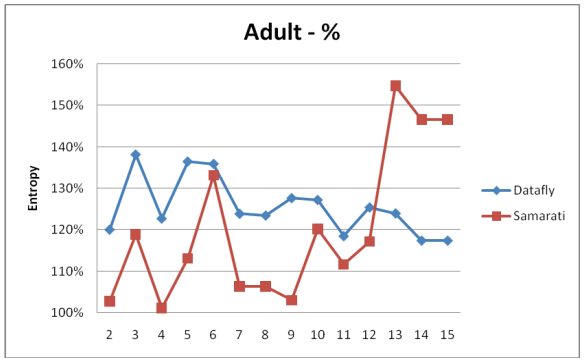


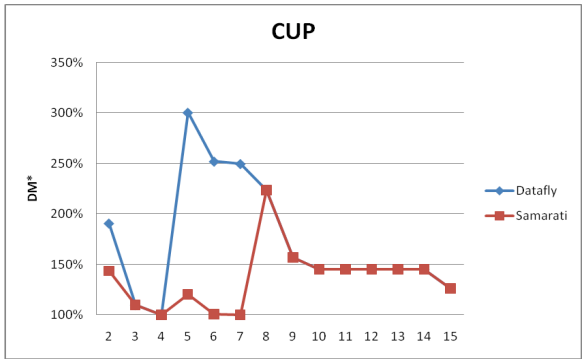
Figure 1: Results summary for the six data sets. The x-axis shows the value of k as it was varied from 2 to 15. The graphs show the information loss results for a 1% suppression limit for the three information loss metrics: DM^*, *Prec*, and non-uniform entropy.

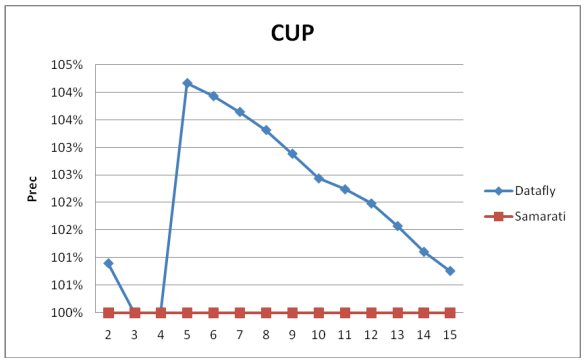
Information Loss Results for 10% Suppression Limit

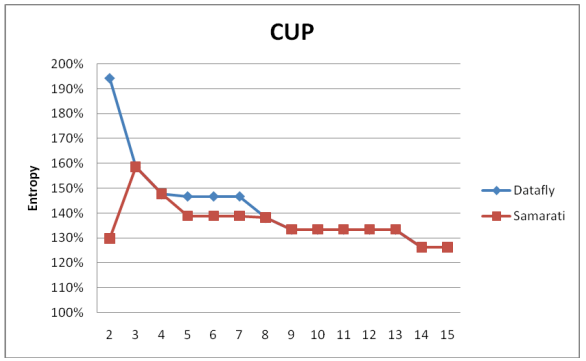


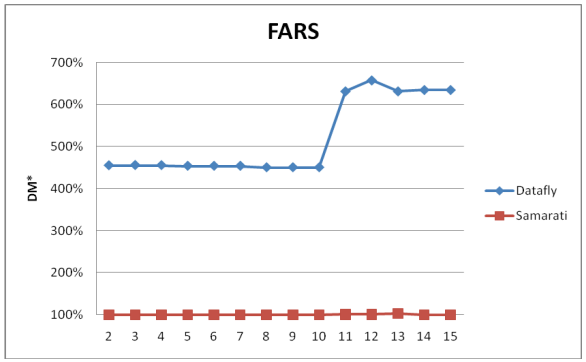


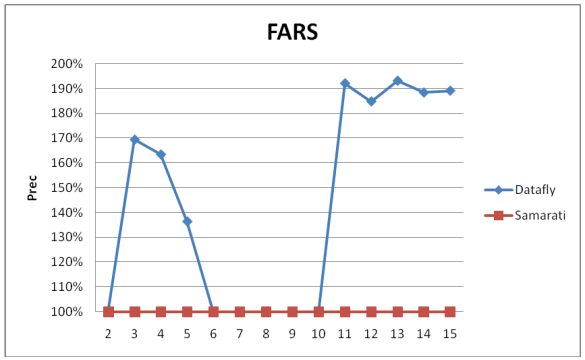


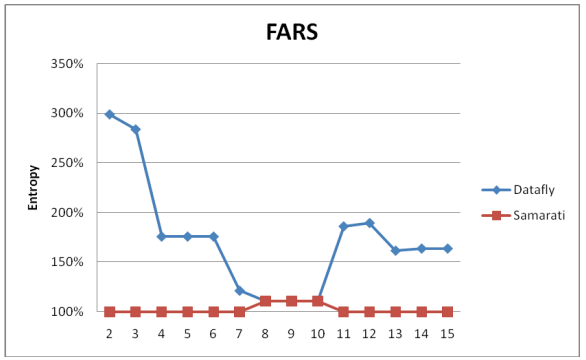


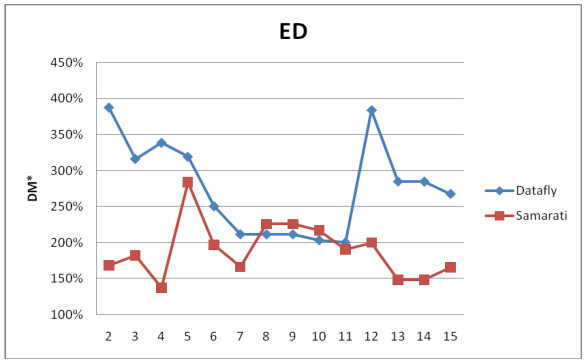


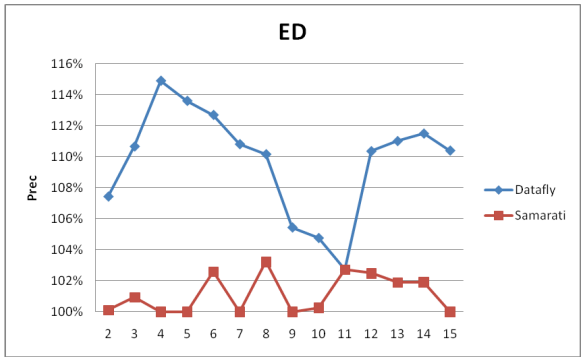


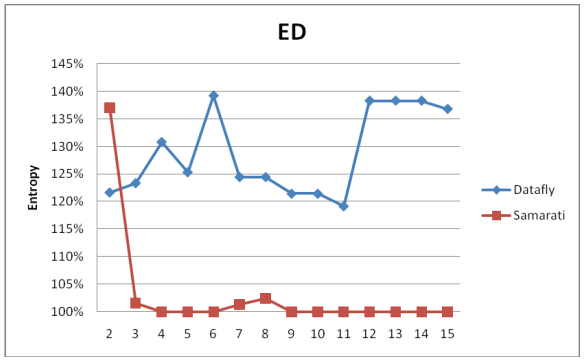

FARS
DM*
700%
600%
500%
400%
300%
200%
100%
2 3 4 5 6 7 8 9 10 11 12 13 14 15
Datafly
Samarati


FARS
Prec
200%
190%
180%
170%
160%
150%
140%
130%
120%
110%
100%
2 3 4 5 6 7 8 9 10 11 12 13 14 15
Datafly
Samarati


FARS
Entropy
350%
300%
250%
200%
150%
100%
2 3 4 5 6 7 8 9 10 11 12 13 14 15
Datafly
Samarati


ED
DM*
450%
400%
350%
300%
250%
200%
150%
100%
2 3 4 5 6 7 8 9 10 11 12 13 14 15
Datafly
Samarati


ED
Prec
116%
114%
112%
110%
108%
106%
104%
102%
100%
2 3 4 5 6 7 8 9 10 11 12 13 14 15
Datafly
Samarati


ED
Entropy
145%
140%
135%
130%
125%
120%
115%
110%
105%
100%
2 3 4 5 6 7 8 9 10 11 12 13 14 15
Datafly
Samarati

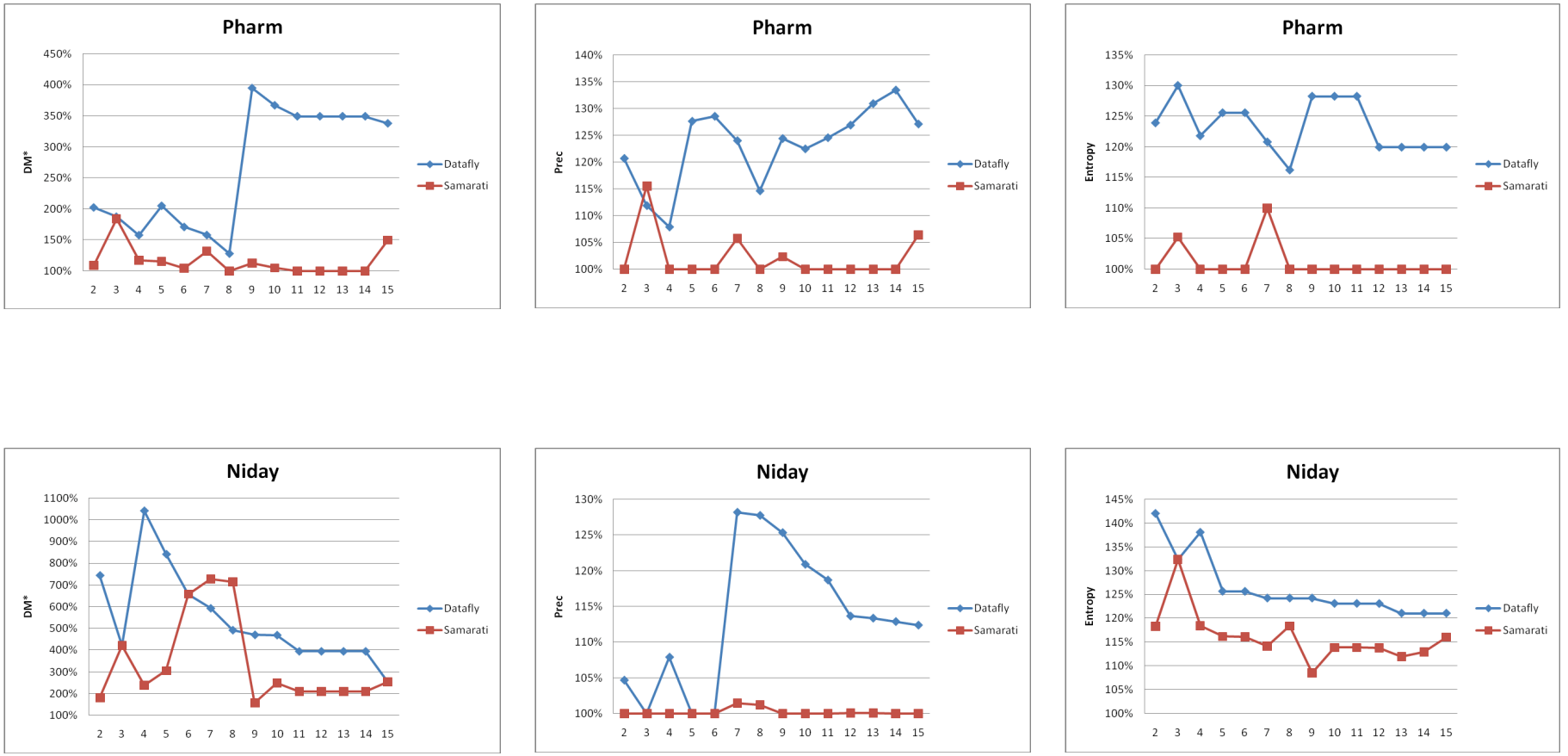


Figure 2: Results summary for the six data sets. The x-axis shows the value of k as it was varied from 2 to 15. The graphs show the information loss results for a 10% suppression limit for the three information loss metrics: DM^*, *Prec*, and non-uniform entropy.

Comparison of Datafly vs. Samarati Speed at 1% Suppression Limit

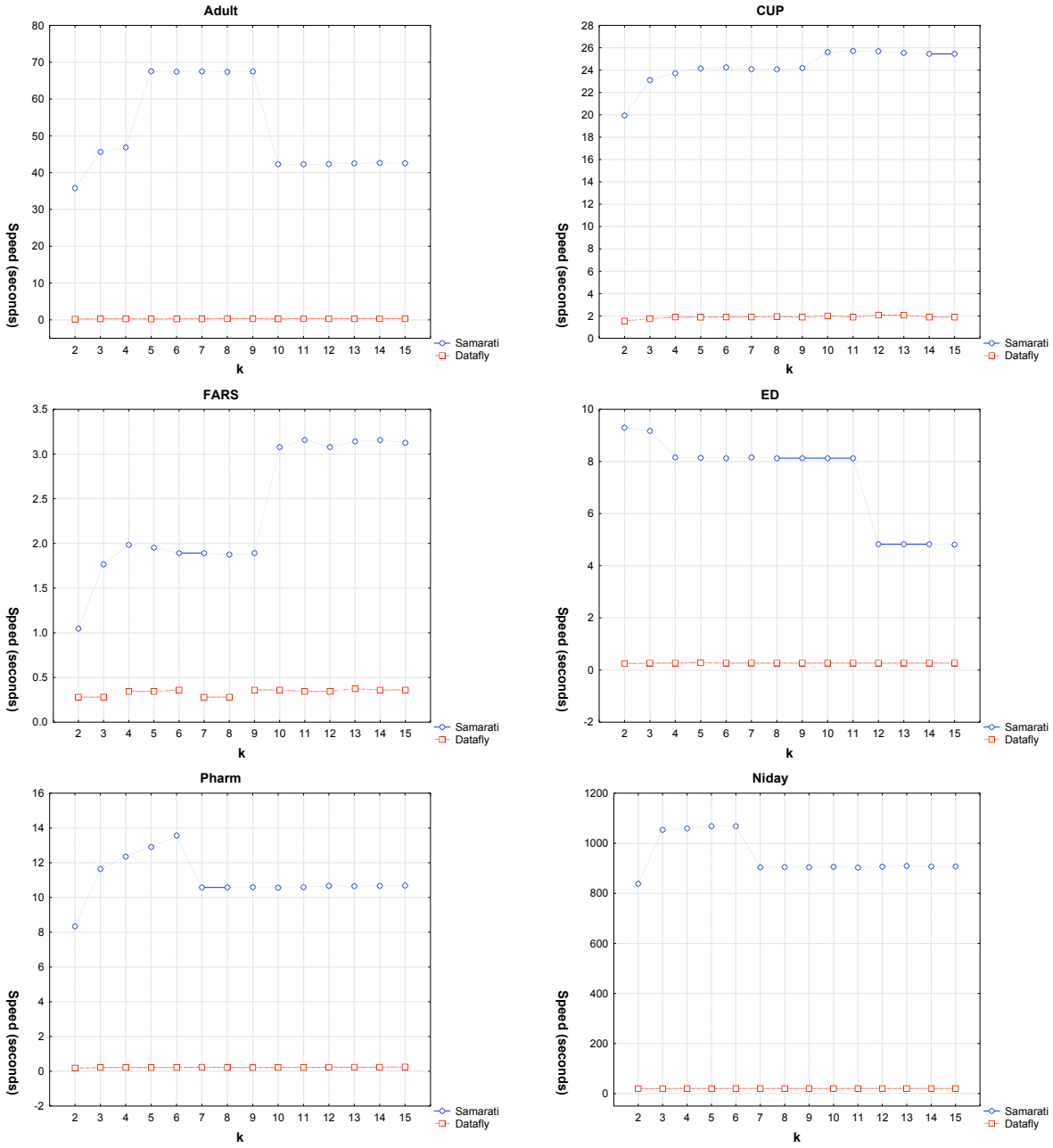


Figure 3: The speed to obtain a solution for Datafly and Samarati on the six data sets across values of k 2 to 15 and a $_{maximum}$ suppression value of 1%. The timing is in seconds and includes all pre- and post- processing that may be needed to run each algorithm. These results were obtained on a dual core 3.2 GHz Pentium D processor with 3GB of RAM.

Comparison of Datafly vs. Samarati Speed at 10% Suppression Limit

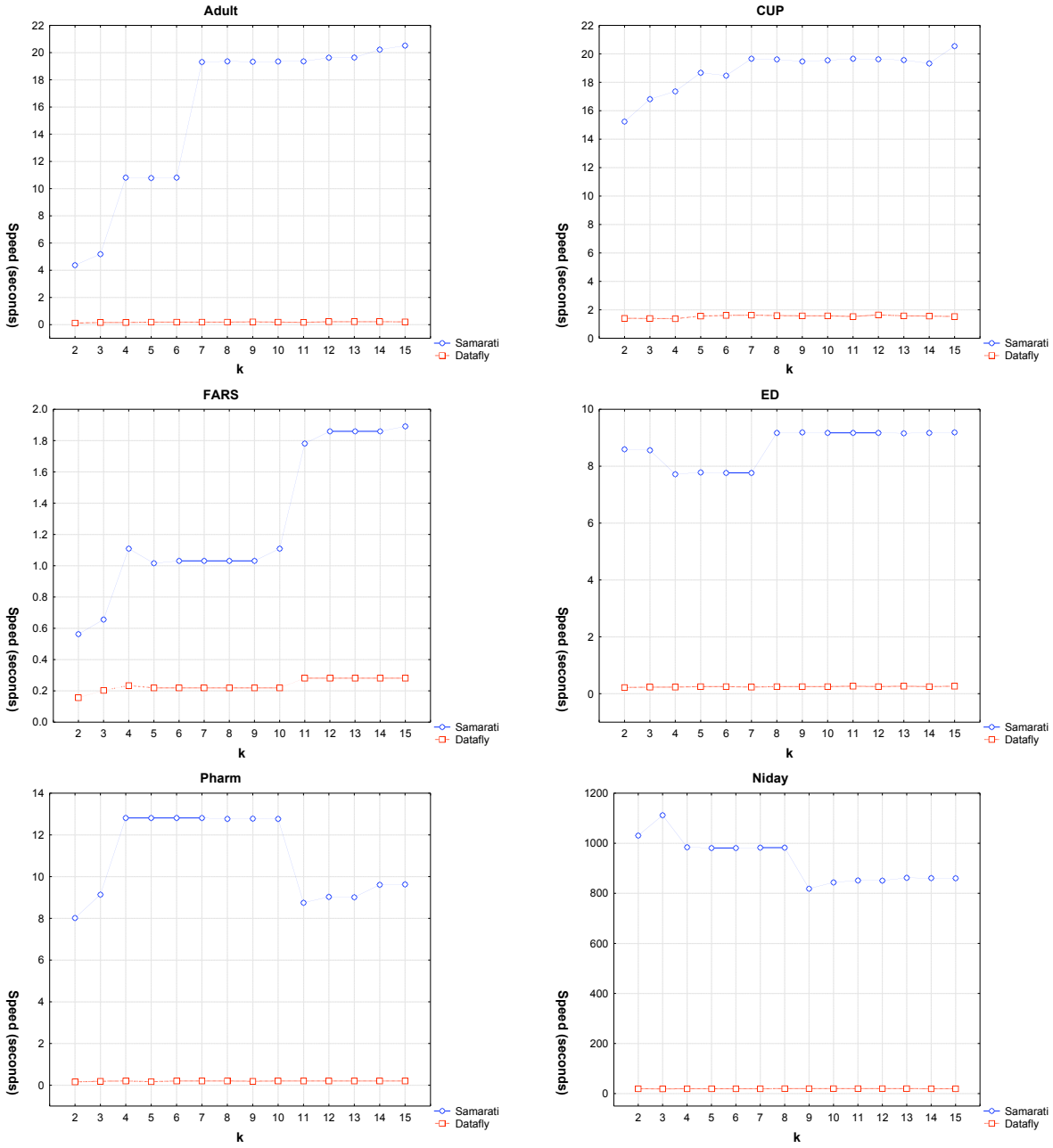


Figure 4: The speed to obtain a solution for Datafly and Samarati on the six data sets across values of k 2 to 15 and a maximum suppression value of 10%. The timing is in seconds and includes all pre- and post- processing that may be needed to run each algorithm. These results were obtained on a dual core 3.2 GHz Pentium D processor with 3GB of RAM.

Comparison of Computations at 1% Suppression Limit

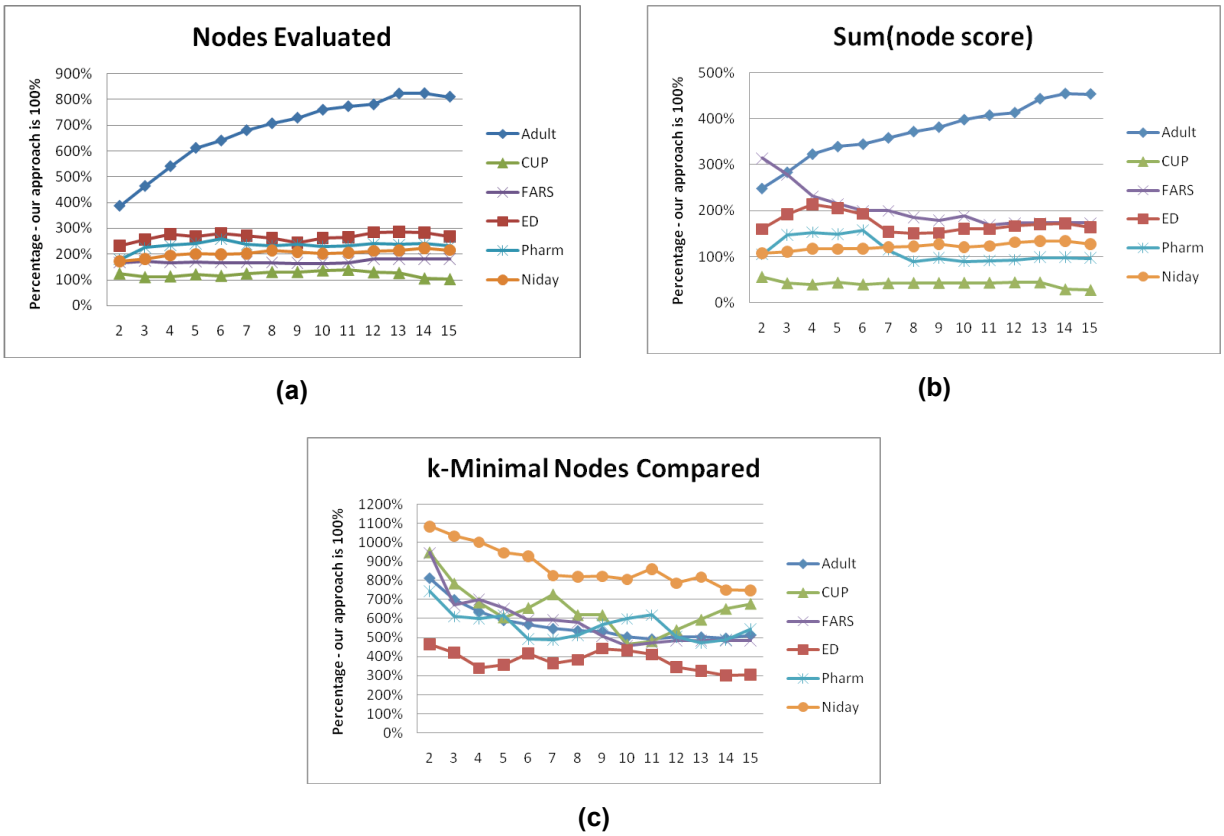


Figure 5: The performance metrics comparing our algorithm to Incognito. The results are for the 1% suppression limit. Our algorithm is the 100% value on the y-axis, and if Incognito performs more computations then its value is above 100%, and if it performs less computation then its value is below 100%. The panels show: (a) the total number of nodes for which we need to compute if they are k-anonymous, (b) the node complexity score given by equation (3), and (c) the number of nodes for which information loss needs to be computed.

Comparison of Computations at 10% Suppression Limit

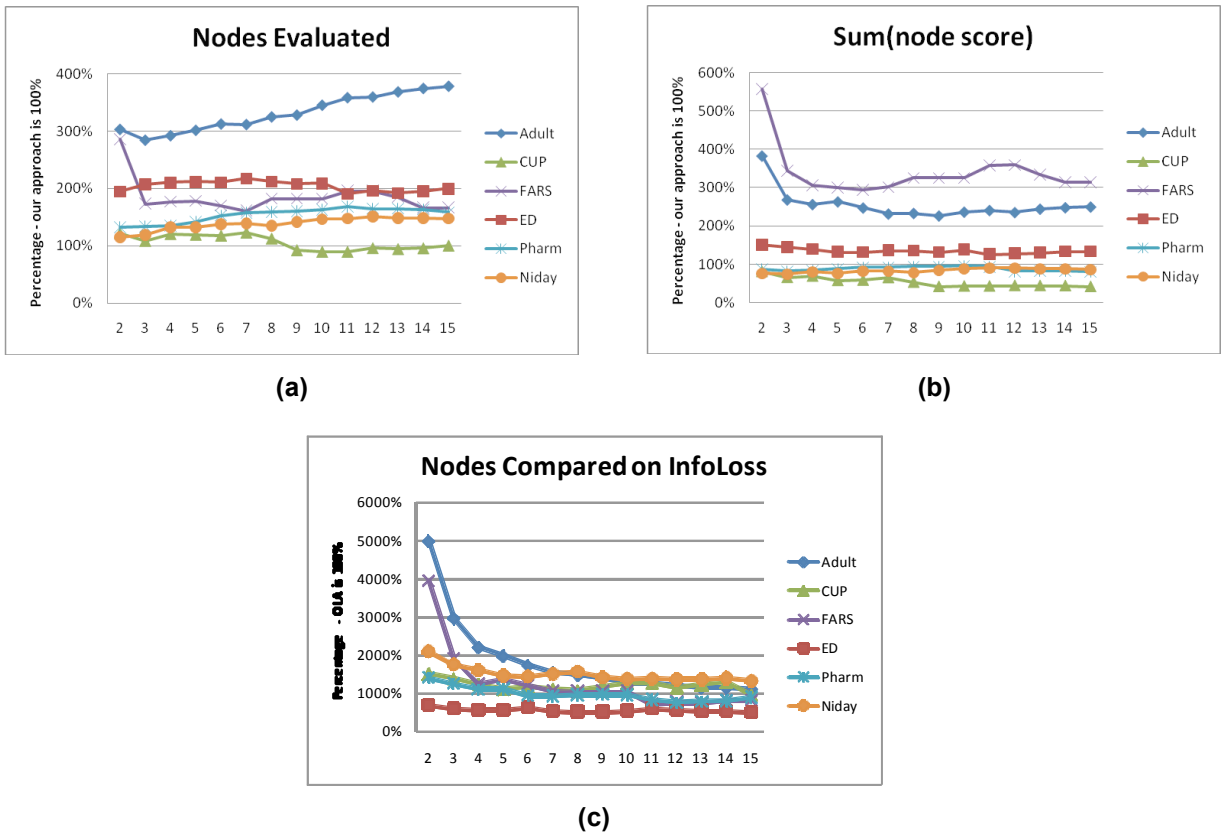


Figure 6: The performance metrics comparing our algorithm to Incognito. The results are for the 10% suppression limit. Our algorithm is the 100% value on the y-axis, and if Incognito performs more computations then its value is above 100%, and if it performs less computation then its value is below 100%. The panels show: (a) the total number of nodes for which we need to compute if they are k-anonymous, (b) the node complexity score given by equation (3), and (c) the number of nodes for which information loss needs to be computed.

Comparison of OLA vs. Incognito Speed at 1% Suppression Limit

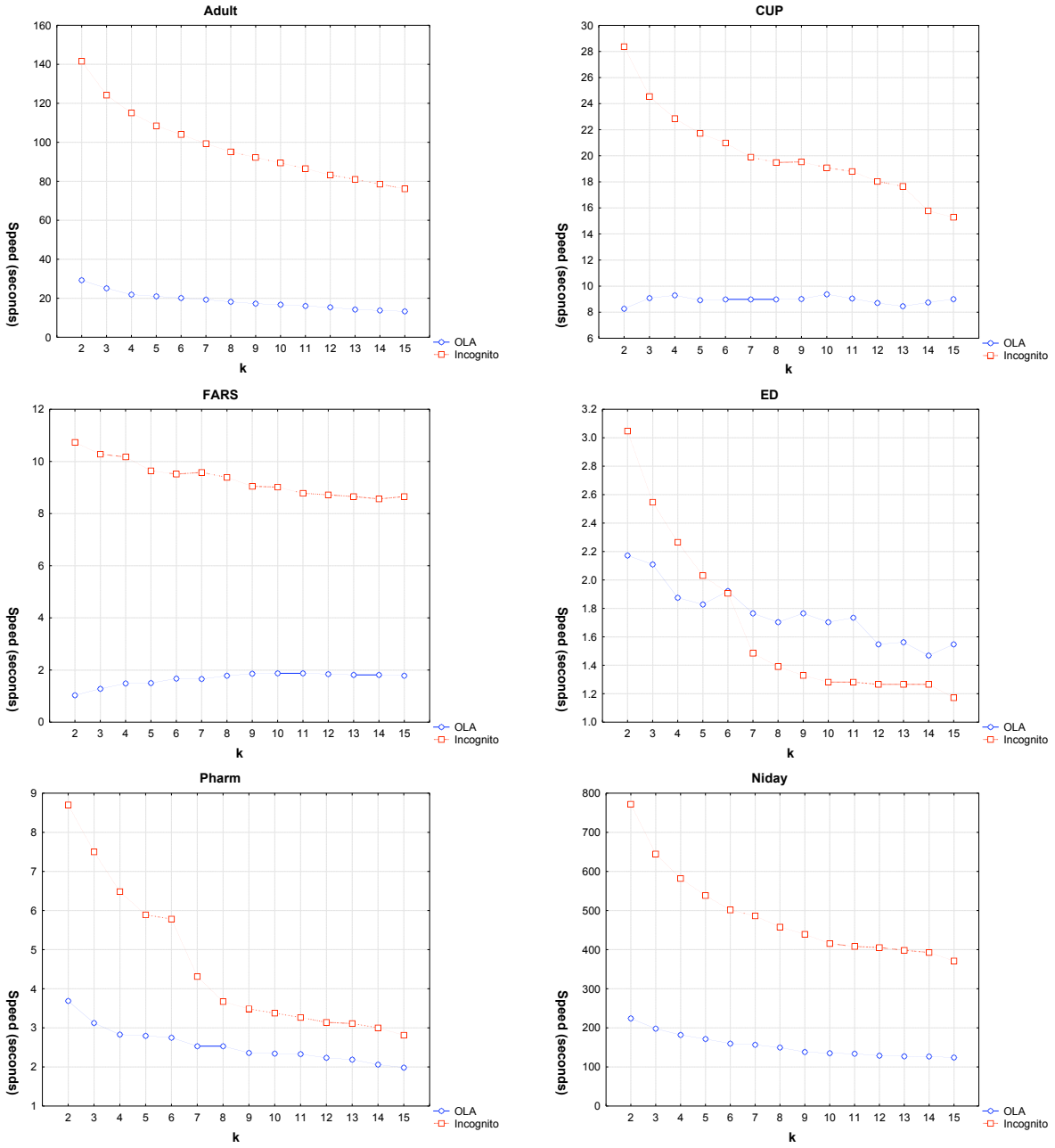


Figure 7: The speed to obtain a globally optimal solution using OLA and Incognito on the six data sets across values of k 2 to 15 and a maximum suppression value of 1%. The timing is in seconds and includes all pre- and post-processing that may be needed to run each algorithm. These results were obtained on a dual core 3.2 GHz Pentium D processor with 3GB of RAM.

Comparison of OLA vs. Incognito Speed at 10% Suppression Limit

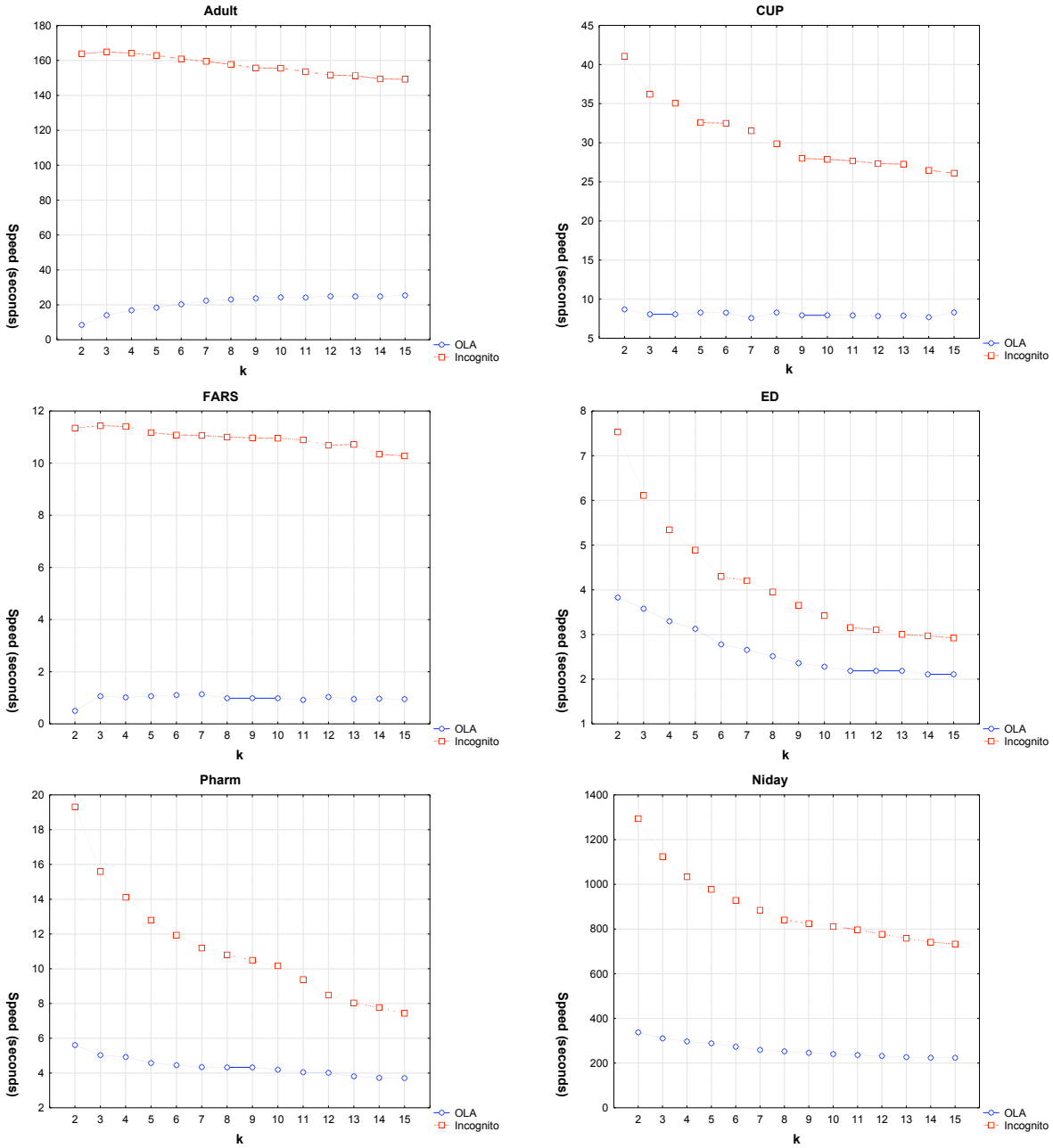


Figure 8: The speed to obtain a globally optimal solution using OLA and Incognito on the six data sets across values of k 2 to 15 and a maximum suppression value of 10%. The timing is in seconds and includes all pre- and post-processing that may be needed to run each algorithm. These results were obtained on a dual core 3.2 GHz Pentium D processor with 3GB of RAM.