Reproducibility Report for ACM SIGMOD 2021 Paper: "Looking for Trouble: Analyzing Classifier Behavior via Pattern Divergence"

YINGJUN WU, Singularity Data Inc.

The results of the paper are reproducible after direct communication with the authors.

1 INTRODUCTION

This report describes the reproducibility process for the paper "Looking for Trouble: Analyzing Classifier Behavior via Pattern Divergence" [1], by Eliana Pastor (Politecnico di Torino, Italy), Luca de Alfaro (UCSC, USA), and Elena Baralis (Politecnico di Torino, Italy). The results of the published paper were reproduced after a series of discussions and improvements with the help of the authors.

2 SUBMISSION

Detailed description:

- code
  - The code is mainly written in Python3 and thus has good readability.
  - The code consists of two parts, namely the functionality part and the display part (such as drawing charts and exporting them to PDF files).
- scripts
  - The script is written in Python3 too.
  - All experiments can be executed with one run. Users can also run every single experiment.
- data
  - There are two types of datasets, namely raw datasets, and processed datasets.

GitHub repository: https://github.com/elianap/divexplorer_SIGMOD21_experiments
Detailed readme file: https://github.com/elianap/divexplorer_SIGMOD21_experiments/blob/main/README.md
Data sources: https://github.com/elianap/divexplorer_SIGMOD21_experiments/tree/main/datasets

3 HARDWARE AND SOFTWARE ENVIRONMENTS

Table 1. Hardware environment

	Paper	Repro Device 1	Repro Device 2
CPU	Intel Core i7	AND64 Ryzen 5 4600U	Intel Core i5
cores	4	6	4
GHz	2.40	2.10	2.00
RAM	16GB	16GB	16GB
Storage	SSD	SSD	SSD

Table 2. Software environment

	Paper	Repro Device 1	Repro Device 2
OS	Ubuntu 16.04 LTS 64-bit	Windows 10 19041 64-bit	Darwin 20.4.0
Interpreter	Python 3.6.10	Python 3.6.10	Python 3.6.10
Dependencies	ipywidgets>=7.2.1	ipywidgets 7.2.1	ipywidgets 7.2.1
	matplotlib>=3.1.1	matplotlib 3.1.1	matplotlib 3.1.1
	numpy>=1.16.4	numpy 1.16.4	numpy 1.16.4
	mlxtend>=0.17.1	mlxtend 0.17.1	mlxtend 0.17.1
	pandas>=0.24.2	pandas 0.24.2	pandas 0.24.2
	plotly>=5.3.0	plotly 5.3.0	plotly 5.3.0
	python_igraph (any version)	python_igraph 0.9.8	python_igraph 0.9.8
	scikit_learn>=0.23.2	scikit_learn 0.23.2	scikit_learn 0.23.2
	kaleido>=0.2.1	kaleido 0.2.1	kaleido 0.2.1
	psutil (any version)	psutil 5.8.0	psutil 5.9.0
	requests (any version)	requests 2.26.0	requests 2.27.1

4 REPRODUCIBILITY EVALUATION

4.1 Process

- What was easily achieved
  - The environment creation and dependencies installation can be finished with a few commands. The experimental script is well written and easy to use.
  - When installing the dependencies with the command "pip install -r requirements.txt", there was an error that "psutils" was not found in pip. Then we realized that it might be a typo and its original spelling should be "psutil". We then changed it and successfully installed all the dependencies. To further verify this, we created a pull request on the GitHub-hosted repo, which was quickly accepted by the author. There are some other simple issues in the code base, such as some PDF files have duplicated file extensions. But they could be easily fixed and were either fixed by us through a patch or by the author.
- What was more complex
  - Something went wrong when we compared the outputted figures with figures in the paper. We spent a lot of time on these issues. The detailed result would be reported in the result section. But finally, these issues are all resolved.
- How you discussed with the authors
  - We mainly discussed in the issue list (https://github.com/elianap/divexplorer_SIGMOD21_experiments/issues) or in the pull request list (https://github.com/elianap/divexplorer_SIGMOD21_experiments/pulls?q=is%3Apr+is%3Aclosed) on GitHub. Particularly, we discussed a lot on this issue: https://github.com/elianap/divexplorer_SIGMOD21_experiments/issues/5.

4.2 Results

The verification-needed outputs of the code:

- 6 tables: describe the statistics of the datasets
- 12 figures: generated by the algorithm

We will continue with more details.

4.2.1 Tables. After a discussion and adjustment related to the display format (see https://github.com/elianap/divexplorer_SIGMOD21_experiments/pull/3), all the outputted tables are fully matched with tables in the paper. Since the tables contain some irregular words and thus are not functional in Microsoft Excel, we open and display them in the raw CSV form. Following is the comparison of tables in the paper with tables in our local output.

Itemset	
age=25-45, #prior>3, race=African-Am, sex=Male	FPR=0.308
age>45, race=Caucasian	FNR=0.929
race=African-Am, sex=Male	FPR=0.150
race=African-Am, sex=Male, #prior>3	FPR=0.267
race=African-Am, sex=Male, #prior=0	FPR=0.097

Table 1: Example of patterns in the *COMPAS* dataset, along with their FPR or FNR. The overall FPR and FNR are 0.088 and 0.698.

```
itemsets,metric
"frozenset({'sex=Male', 'age_cat=25 - 45', 'priors_count=>3', 'race=African-American'})",d_fpr=0.308
"frozenset({'age_cat=Greater than 45', 'race=Caucasian'})",d_fnr=0.929
"frozenset({'sex=Male', 'race=African-American'})",d_fpr=0.150
"frozenset({'sex=Male', 'priors_count=>3', 'race=African-American'})",d_fpr=0.267
"frozenset({'sex=Male', 'priors_count=0', 'race=African-American'})",d_fpr=0.097
```

Fig. 1. Reproducing Table 1

Itemset	Sup	Δ_{FPR}	t
age=25-45, #prior>3, race=Afr-Am, sex=Male	0.13	0.22	7.1
age=25-45, #prior>3, race=Afr-Am	0.15	0.211	7.4
age=25-45, charge=F, #prior>3, race=Afr-Am	0.11	0.202	6.2
	Sup	Δ_{FNR}	t
age=25-45, stay<week, #prior=0	0.15	0.236	12.1
charge=M, stay<week, #prior=[1,3]	0.10	0.233	12.2
age>45, race=Cauc	0.10	0.231	10.3
	Sup	Δ_{ER}	t
age<25, stay<week, race=Afr-Am	0.10	0.098	4.7
age<25, stay<week, sex=Male	0.13	0.095	5.2
age<25, race=Afr-Am, sex=Male	0.11	0.090	4.5
	Sup	Δ_{ACC}	t
stay<week, #prior=0, race=Cauc	0.12	0.141	8.4
charge=M, stay<week, #prior=0	0.15	0.133	8.6
charge=M, #prior=0	0.16	0.129	8.5

Table 2: Top-3 divergent patterns with respect to FPR, FNR, error rate (ER) and accuracy (ACC) for the *COMPAS* dataset. The support threshold is $s = 0.1$.

```
itemsets,sup,Δ_fpr,t_fp
"age=25-45, #prior>3, race=Afr-Am, sex=Male",0.13,0.22,7.1
"age=25-45, #prior>3, race=Afr-Am",0.15,0.211,7.4
"age=25-45, charge=F, #prior>3, race=Afr-Am",0.11,0.202,6.2
itemsets,sup,Δ_fnr,t_fn
"age=25-45, stay<week, #prior=0",0.15,0.236,12.1
"charge=M, stay<week, #prior=[1,3]",0.1,0.233,12.2
"age>45, race=Cauc",0.1,0.231,10.3
itemsets,sup,Δ_error,t_fp_fn
"age<25, stay<week, race=Afr-Am",0.1,0.098,4.7
"age<25, stay<week, sex=Male",0.13,0.095,5.2
"age<25, race=Afr-Am, sex=Male",0.11,0.09,4.5
itemsets,sup,Δ_accuracy,t_tp_tn
"stay<week, #prior=0, race=Cauc",0.12,0.141,8.4
"charge=M, stay<week, #prior=0",0.15,0.133,8.6
"charge=M, #prior=0",0.16,0.129,8.5
```

Fig. 2. Reproducing Table 2

I	corr. item	Δ(I)	Δ(I ∪α)	c_f	t
FPR					
race=Afr-Am, sex=Male	#prior=0	0.062	0.009	0.053	2.8
race=Afr-Am	#prior=0	0.051	-0.001	0.051	3.4
stay<week, #prior=0	race=Afr-Am	-0.044	-0.003	0.041	3.1
FNR					
charge=F, race=Afr-Am, sex=Male	#prior=[1,3]	-0.123	-0.011	0.112	3.8
charge=F, race=Afr-Am	#prior=[1,3]	-0.113	0.004	0.109	4.3
race=Afr-Am, sex=Male	charge=M	-0.090	-0.001	0.089	3.3

Table 3: Top corrective items for FPR and FNR of *COMPAS* dataset.

```
I,corr. item,$Δ_{FPR}(I)$,$Δ_{FPR}(I \cup α)$,c_f,t_corr
"race=Afr-Am, sex=Male",#prior=0,0.062,0.009,0.053,2.8
race=Afr-Am,#prior=0,0.051,-0.001,0.051,3.4
"stay<week, #prior=0",race=Afr-Am,-0.044,-0.003,0.041,3.1
I,corr. item,$Δ_{FNR}(I)$,$Δ_{FNR}(I \cup α)$,c_f,t_corr
"charge=F, race=Afr-Am, sex=Male","#prior=[1,3]",-0.123,-0.011,0.112,3.8
"charge=F, race=Afr-Am","#prior=[1,3]",-0.113,0.004,0.109,4.3
"race=Afr-Am, sex=Male",charge=M,-0.09,-0.001,0.089,3.3
```

Fig. 3. Reproducing Table 3

dataset	$\|D\|$	$\|A\|$	$\|A\|_{cont}$	$\|A\|_{cat}$
adult	45,222	11	4	7
bank	11,162	15	6	9
COMPAS	6,172	6	2	4
german	1,000	21	7	14
heart	296	13	5	8
artificial	50,000	10	0	10

Table 4: Dataset characteristics. A_{cont} is the set of continuous attributes, A_{cat} of categorical ones.

```
dataset,|D|,|A|,|A|_cont,|A|_cat
adult,45222,11,4,7
bank,11162,15,6,9
compas,6172,6,2,4
german,1000,21,7,14
heart,296,13,5,8
artificial_10,50000,10,0,10
```

Fig. 4. Reproducing Table 4

Itemset	Sup	Δ_{FPR}	t
gain=0, status=Married, occup=Prof, race=White	0.05	0.469	25.8
gain=0, loss=0, status=Married, occup=Prof	0.05	0.462	26.6
loss=0, status=Married, occup=Prof, race=White	0.06	0.458	25.3
Itemset	**Sup**	**Δ_{FNR}**	**t**
age≤28, gain=0, hoursXW≤40, status=Unmarried	0.17	0.61	21.8
gain=0, loss=0, edu=HS, hoursXW≤40, status=Unmarried	0.14	0.61	28.2
gain=0, loss=0, status=Unmarried, relation=Own-child	0.12	0.61	18.9

Table 5: Top-3 divergent itemsets for FPR and FNR. *adult* dataset, $s = 0.05$.

```
itemsets,sup,Δ_fpr,t_fp
"gain=0, status=Married, occup=Prof, race=White",0.05,0.469,25.8
"gain=0, loss=0, status=Married, occup=Prof",0.05,0.462,26.6
"loss=0, status=Married, occup=Prof, race=White",0.06,0.458,25.3
itemsets,sup,Δ_fnr,t_fn
"age≤28, gain=0, hoursXW≤40, status=Unmarried",0.17,0.61,21.8
"gain=0, loss=0, edu=HS, hoursXW≤40, status=Unmarried",0.14,0.61,28.2
"gain=0, loss=0, status=Unmarried, relation=Own-child",0.12,0.61,18.9
```

Fig. 5. Reproducing Table 5

Itemset	Sup	Δ_{FPR}	t
status=Married, occup=Prof	0.07	0.434	26.1
occup=Prof, relation=Husband	0.06	0.423	23.4
edu=Bachelors, status=Married	0.09	0.413	29

Table 6: Top-3 divergent itemsets for FPR with redundancy pruning. *adult* dataset, $\epsilon = 0.05$, $s = 0.05$.

```
itemsets,sup,Δ_fpr,t_fp
"status=Married, occup=Prof",0.07,0.434,26.1
"occup=Prof, relation=Husband",0.06,0.423,23.4
"edu=Bachelors, status=Married",0.09,0.413,29.0
```

Fig. 6. Reproducing Table 6

4.2.2 Figures. However, the issues about the figures are more difficult to resolve. Here are our discussions: https://github.com/elianap/divexplorer_SIGMOD21_experiments/issues/5. In conclusion, there are two types of problems:

- The figure formats between the camera-ready paper and our outputs are mismatched. Actually, we could easily tell that the results of some figures (figure_3, figure_7, figure_8) were reproduced even without these adjustments. Figure_11 is a more complicated graph and even a little change (such as label words abbreviations and device differences) would make the graph looks considerably different. But the topology between the figure in the paper and in our local output is the same.
- Some libraries behave slightly differently on different OS and thus make the output graph slightly different from the figures in the paper. After several rounds of discussions and retries, we came to realize that the figure_9 could not be fully reproduced in our Windows 10 device with the library "orca". There are some known issues related to "ocra", a figure generator library. After replacing "ocra" with the library "kaleido", the figure_9 is reproduced in our Windows.

After a long journey, we reproduced the result both in our Windows and macOS machines. The following comparison contains the figures in the paper and our reproduced results. For each item, we first display the figure in the paper, then display the figures we reproduced in Windows and macOS individually.

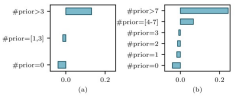

Figure 1: Individual item divergence for false-positive rate of *prior* attribute value of the *COMPAS* dataset where the attribute is discretized in 3 (a) and 6 (b) intervals (s=0.05.)

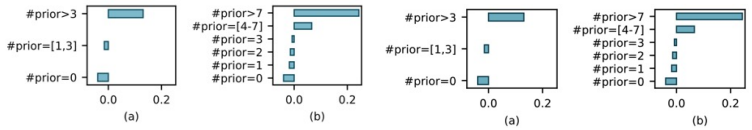

Fig. 7. Reproducing Figure 1

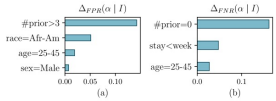

Figure 2: Contributions of individual items to the divergence of the *COMPAS* frequent patterns having greatest false-positive and false-negative divergence.

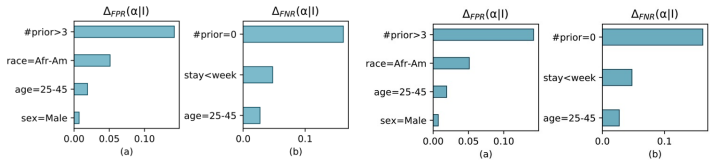

Fig. 8. Reproducing Figure 2

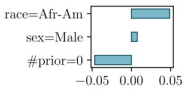

Figure 3: An itemset where an item has a negative divergence contribution.

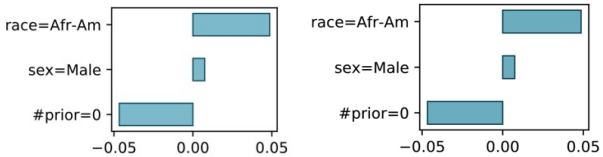

Fig. 9. Reproducing Figure 3

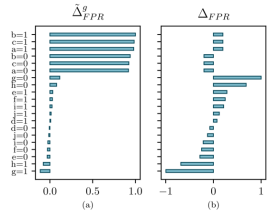

Figure 4: Relative magnitudes of $\tilde{\Delta}^g(\cdot, s)$ and individual item divergence, for false-positive rate in the artificial dataset. The attributes a, b, c give raise to divergence when appearing together: global divergence captures this.

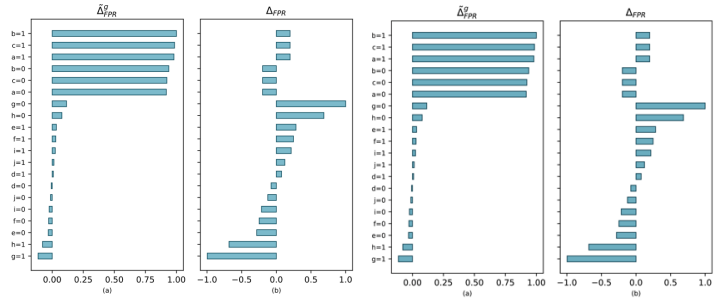

Fig. 10. Reproducing Figure 4

5 SUMMARY

After a series of discussions and improvements, finally, all the tables and figures displayed in the paper are reproduced successfully.

REFERENCES

[1] Eliana Pastor, Luca de Alfaro, and Elena Baralis. 2021. Looking for Trouble: Analyzing Classifier Behavior via Pattern Divergence. In *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021.* ACM, 1400–1412. https://doi.org/10.1145/3448016.3457284

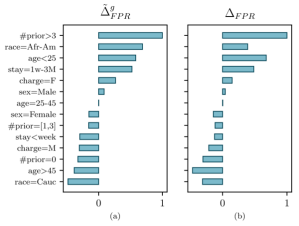

Figure 5: Relative magnitudes of global Shapley value and individual item divergence, for false-positive rate in the *COMPAS* dataset with $s = 0.1$

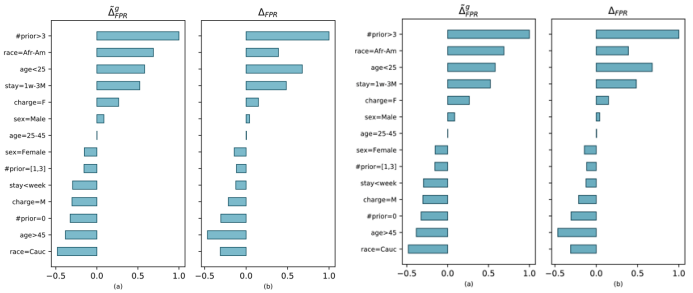

Fig. 11. Reproducing Figure 5

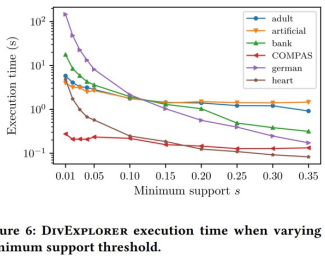

Figure 6: DivExplorer execution time when varying the minimum support threshold.

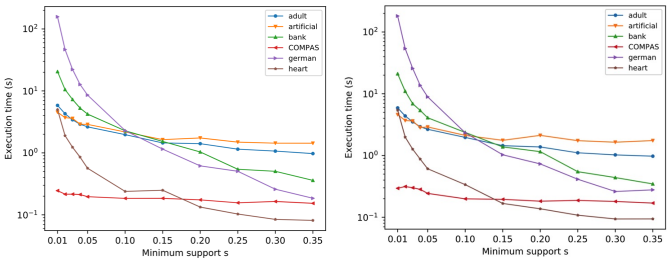

Fig. 12. Reproducing Figure 6

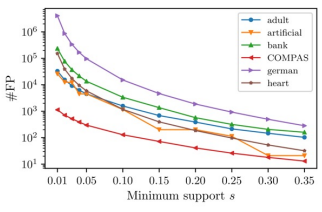

Figure 7: Number of frequent itemsets when varying the minimum support threshold.

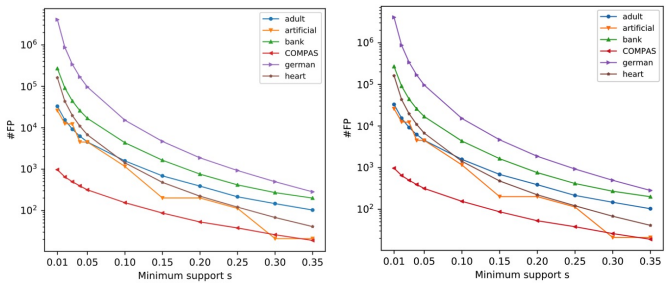

Fig. 13. Reproducing Figure 7

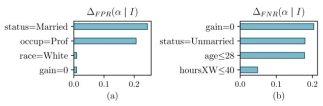

Figure 8: Contributions of individual items to the divergence of the adult frequent patterns having greatest FPR (Line 1 of Table 5) and FNR (Line 4 of Table 5) divergence.

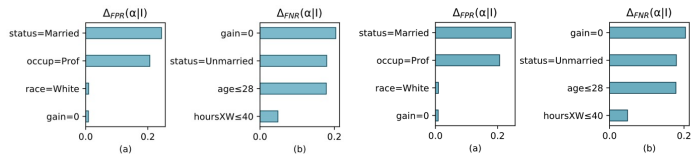

Fig. 14. Reproducing Figure 8

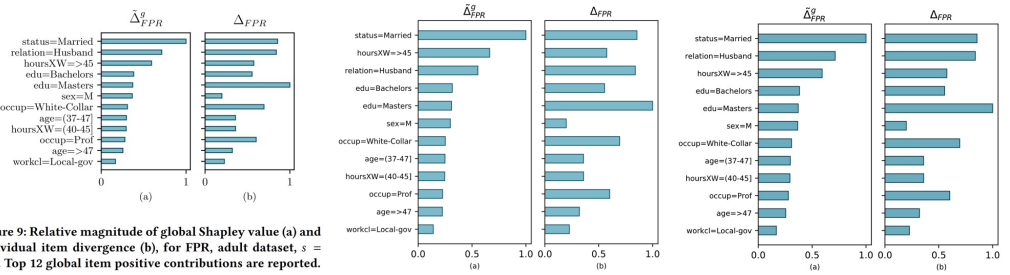

Figure 9: Relative magnitude of global Shapley value (a) and individual item divergence (b), for FPR, adult dataset, $s = 0.05$. Top 12 global item positive contributions are reported.

Fig. 15. Reproducing Figure 9

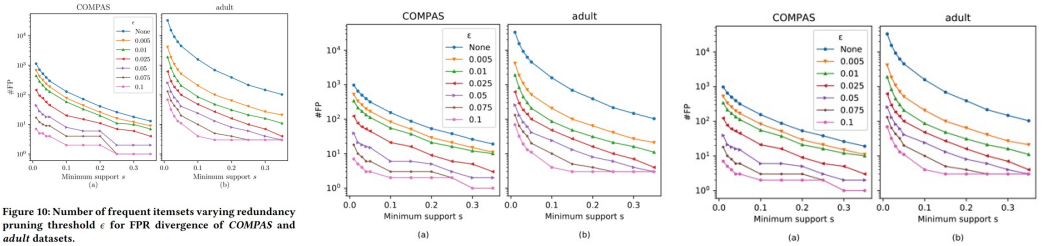

Figure 10: Number of frequent itemsets varying redundancy pruning threshold ϵ for FPR divergence of *COMPAS* and *adult* datasets.

Fig. 16. Reproducing Figure 10

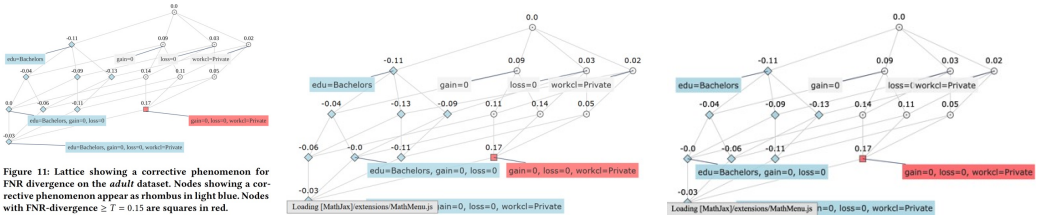

Figure 11: Lattice showing a corrective phenomenon for FNR divergence on the *adult* dataset. Nodes showing a corrective phenomenon appear as rhombus in light blue. Nodes with FNR-divergence $\geq T = 0.15$ are squares in red.

Fig. 17. Reproducing Figure 11

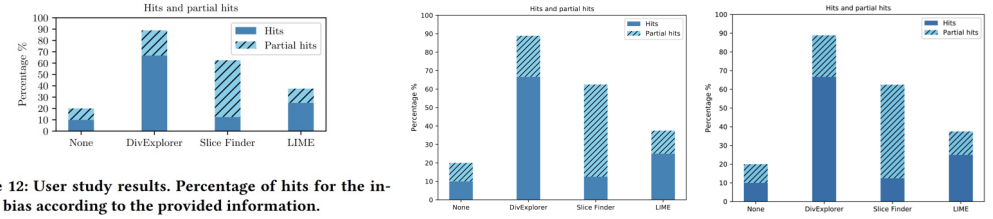

Figure 12: User study results. Percentage of hits for the injected bias according to the provided information.

Fig. 18. Reproducing Figure 12