Set of Proofs Showing that RPS is Side Effect Free

Natasha de Kriek and Henry Zwart

September 2020

RPS the algorithm to which these proofs correspond, was first outlined by [1].

To begin, we lay down some notation. Observe that given a universal set of items I, and a closed itemset model M, the support of an itemset $X \subseteq I$ in the model described by M can be calculated as the maximum support across all closed itemsets in M which contain X. We denote this value $\sigma_M(X)$.

We introduce the following definition to formalise the concept of RPS sanitising a particular closed itemset:

Definition 1. Sanitisation Tree
Given a closed itemset i, define the Sanitisation Tree, T_i, to be a rooted tree where the root has value i. Let x be a node in T_i. If x contains no sensitive itemset, then it is a leaf. Otherwise, if x contains some sensitive itemset s, then the children of x in T_i are the itemsets: $x \setminus \{s_j\}$ for each $s_j \in s$.

Observe that this definition encapsulates the concept of RPS' recursive sanitisation process for a single closed itemset. As such, every node in a tree T_i has support $\sigma_M(i)$. Furthermore, the set of leaves in a given T_i is the set of closed itemsets introduced during the sanitisation of the closed itemset i.

Theorem 1. Let I be a universal set of items, and D a transactional database. Let M be a closed itemset model for D with minimum support σ_{model}, and M' the resulting closed itemset model after RPS has sanitised M for some set of sensitive itemsets. Take non-empty $X \subseteq I$ to be any itemset which contains no sensitive itemset, where $\sigma(X) \geq \sigma_{model}$. Let $Y \in M$ be a closed itemset such that $X \subseteq Y$ and $\sigma(X) = \sigma_M(Y)$. Then there exists some closed itemset $Z \in M'$ such that $X \subseteq Z$ and $\sigma_{M'}(Z) = \sigma_M(Y)$.

Proof:
Let I be a universal set of items and M a closed itemset for a transactional database D over I. Take some non-empty itemset $X \subseteq I$ which occurs in D with support $\sigma(X) \geq \sigma_{model}$. Find any $Y \in M$ such that $X \subseteq Y$ and $\sigma(X) = \sigma_M(Y)$. Such a Y exists as M contains all closed itemsets for D with minimum support σ_{model}. We may assume that Y is frequent and contains some sensitive itemset, as otherwise it remains unmodified by RPS, otherwise set $Z = Y$.

Let T_Y be the sanitisation tree rooted at the closed itemset Y. As noted earlier, the leaves of T_Y are exactly the set of closed itemsets which are introduced to the sanitised model once Y has been sanitised.

We proceed via induction to show that for each level of T_Y, either the level contains some node which contains X, or a leaf on a previous level contains X. We will then show that T_Y is finite, implying that X is contained inside some closed itemset in the sanitised model M'.

In the base case, T_Y has a single node Y which contains X by definition. Suppose now that the inductive hypothesis holds for level t in the sanitisation tree, and consider level $t+1$. By our inductive hypothesis, at least one of the following is true:

1. $X \subseteq Z$, where Z is a leaf node on a previous level.

2. There is a node W on level t such that $X \subseteq W$.

In the first case, Z is a leaf on a level prior to t, so is also a leaf on a level prior to level $t+1$, as required. In the second case, we may assume that W contains a sensitive itemset, otherwise it is a leaf on level t, satisfying the induction. Suppose then that there is some sensitive itemset $s \subseteq W$. By definition $s \not\subseteq X$, so there is some $s_i \in s$ such that $s_i \notin X$. The children of W are defined as $\{W \setminus \{s_j\} | s_j \in s\}$ so some child $Z = W \setminus \{s_i\}$ of W will contain X, completing the induction.

We now observe that the size of the sets in T_Y decreases strictly monotonically with the levels of the tree, and that no empty-set contains a sensitive itemset. It follows that T_Y is finite and so the santization of Y will terminate. By this result, we find that one of the leaves of T_Y contains X, and thus that X appears in some $Z \in M'$.

We now show that $\sigma_{M'}(Z) = \sigma_M(Y)$. Recall that M' is constructed by adding all closed itemsets from M which do not need to be sanitised, and inserting any new closed itemsets to M' the first time they are generated during sanitisation.

Let Z, Y be as above, and take V such that $Z \subseteq V$ and $\sigma_M(Y) > \sigma_M(V)$. We proceed to prove that RPS will sanitize Y first, and hence that Z will occur in M' with the highest support across all of it's supersets in M. This is equivalent to proving that $|Y| < |V|$.

As $\sigma(Z) = \sigma(Y)$ and $Z \subseteq Y$, Y and Z always occur together in transactions. Similarly as $Z \subseteq V$, all transactions which contain V also contain Z, and so contain Y. As V is a closed itemset, it has no direct superset with the same support. It follows that V contains Y, as otherwise V would have the same support as the superset $V \cup Y$. Finally, as $\sigma(V) < \sigma(Y)$, we find that Y is a direct subset of V. We conclude that Y is shorter than V, and so will be sanitised first by RPS.

Finally, observe that the closed itemsets in M' are a subset of the itemsets which occur in the various sanitisation trees for M, and the support of any itemset in a sanitisation tree is the same as that of the root. It follows that there is no closed itemset in M' which contains X and has a larger support than Y. Therefore $\sigma_{M'})(X) = \sigma_M(X) = \sigma(X)$.

Tying the various results together, we have proven that there exists some $Z \in M'$ such that $X \subseteq Z$ and $\sigma(X) = \sigma_M(Y) = \sigma_{M'}(Z)$. We then proved that $\sigma(X) = \sigma_{M'}(Z)$, and finally that $\sigma_{M'}(X) = \sigma(X)$.

□

Corollary 1. Let I be the universal set of items and D a transactional database over I. Let M be the closed itemset model of D with minimum support σ_{model}. Take M' to be the sanitized model returned by RPS given the input the model M and a set of sensitive frequent itemsets S. For all non-sensitive itemsets $X, Y \subseteq I$ with $X \cup Y$ non-sensitive, $\sigma(X), \sigma(Y) \geq \sigma_{model}$, and $X \cap Y = \emptyset$ we have that $\gamma_M(X \Rightarrow Y) = \gamma_{M'}(X \Rightarrow Y) = \gamma(X \Rightarrow Y)$

Proof:
Let I, D, M, and M' be as stated, and let $X, Y \subseteq I$ be itemsets with $\sigma(X), \sigma(Y) \geq \sigma_{model}$, which contain no sensitive itemsets. By Theorem 1, we have that

$$\sigma(X) = \sigma_M(X) = \sigma_{M'}(X)$$

and

$$\sigma(Y) = \sigma_M(Y) = \sigma_{M'}(Y)$$

Then it follows that:

$$\gamma_{M'}(X \Rightarrow Y) = \frac{\sigma_{M'}(X \cup Y)}{\sigma_{M'}(X)} = \frac{\sigma_M(X \cup Y)}{\sigma_M(X)} = \gamma_M(X \Rightarrow Y) = \gamma(X \Rightarrow Y)$$

And so the confidence of the non-sensitive rule $(X \Rightarrow Y)$ is unchanged.

□

Corollary 2. RPS produces no Artifactual Patterns (False Positives), and no Misses Cost (False Negatives).

Proof:
Let I be a universal set of items, and D a transactional database over I. Let M be a closed itemset model for D with minimum support σ_{model}, and let M' be the resulting closed itemset model as output by RPS. Let $X, Y \subset I$ be itemsets such that $X \cap Y \neq \emptyset$ and $\sigma(X), \sigma(Y) \geq \sigma_{model}$, and the rule $(X \Rightarrow Y)$ is non-sensitive.

By Theorem 1 and Corollary 1, we find that $\sigma(X \Rightarrow Y) = \sigma_{M'}(X \Rightarrow Y)$, and $\gamma(X \Rightarrow Y) = \gamma_{M'}(X \Rightarrow Y)$. So we find that $(X \Rightarrow Y)$ is frequent in M if, and only if, it is frequent in M'. Thus there no Artifactual Patterns nor Misses Cost are produced.

$\square$

Theorem 2. RPS is side-effect free.

Proof:
By Corollary 2 RPS produces no Artifactual Patterns nor Misses Cost, so we proceed to prove that it successfully hides all sensitive itemsets.
Let I be a universal set of items, and D a transactional database over I. Let M be a closed itemset model for D, and M' the resulting closed itemset model produced by running RPS on M with a hiding threshold σ_{min}.
Note that M is finite, and so there are a finite number of sensitive itemsets in M. For each sensitive itemset in M, X, with $\sigma(X) \geq \sigma_{min}$, let T_X be the associated sanitisation tree. As proved in Theorem 1, each sanitisation tree is finite, so RPS is guaranteed to terminate. At the end of the sanitisation process for a sensitive closed itemset $X \in M$, all internal nodes in T_X have been removed from M', and all leaves which were not already in M' have been added to M'. These leaves are non-sensitive by definition.
Thus we find that M' contains no sensitive closed itemsets with support at least σ_{min}, and so it has no hiding failures. We conclude that RPS is side-effect free.

$\square$

References

[1] S. H and M. H. S, “Hiding sensitive itemsets without side effects,” *Applied Intelligence*, vol. 49, no. 4, pp. 1213–1227, Apr. 2019. [Online]. Available: https://doi.org/10.1007/s10489-018-1329-5