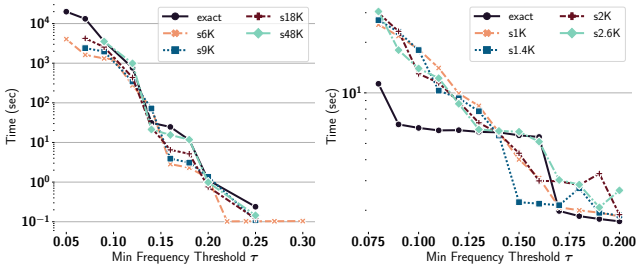

Figure 6: Running time of MaNIACS and the exact algorithm, varying sample size and min frequency threshold τ, on Mico (left) and Citeseer (right).

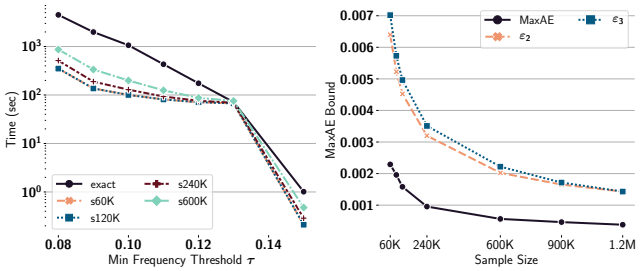

Figure 7: YouTube: running time of MaNIACS and the exact algorithm, varying sample size, and min frequency threshold τ (left); and Max Absolute Error (MaxAE), ε_2, ε_3, ε_4, and ε_5, for various sample size and min frequency threshold $\tau = 0.16$ (right).

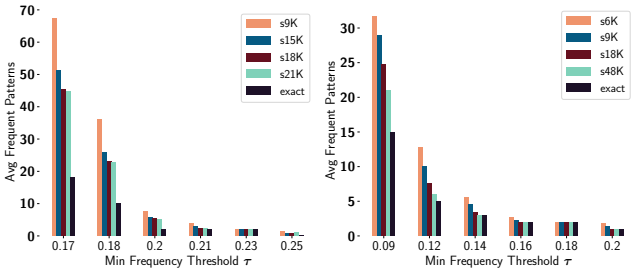

Figure 8: Average number of patterns found by MaNIACS, together with the exact number of frequent patterns, varying minimum frequency threshold τ, in Phy-Cit (left) and Mico (right).

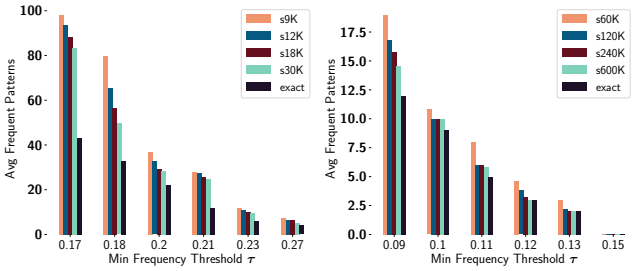

Figure 9: Average number of patterns found by MaNIACS, together with the exact number of frequent patterns, varying minimum frequency threshold τ, in Patents (left) and YouTube (right).

holds that S is, simultaneously, an η_i-sample for $(V, \mathcal{R}_i)$ for every $1 \le i \le k$. Assume for the rest of the proof that that is the case.

We show inductively that, at the end of every iteration of the "main" loop of MaNIACS (lines 4–14), it holds that

(1) $\mathcal{Q}$ contains a triplet $(P, \mathrm{f}_S(P), \varepsilon_i)$ for each $P \in \mathcal{F}_i$, and the triplet is such that

$$|\mathrm{f}_V(P) - \mathrm{f}_S(P)| \le \varepsilon_i \ ;$$

(2) $\mathcal{F}_i \subseteq \mathcal{H}_i$, for $i \le k$.

At the beginning of the first iteration, i.e., for $i = 1$, it obviously holds $\mathcal{F}_1 \subseteq \mathcal{H}_1$ from the definition of $\mathcal{H}_1$ (line 3). Thus, at the first iteration of the do-while loop on lines 6–11, the value ε_1 computed on line 8 using Thm. 3.1 is not smaller than η_1, because b_1^* is an upper bound to the eVC-dimension of $(V, \mathcal{R}_1)$ on S, thanks to Lemmas 4.2 to 4.4, and the value η on the l.h.s. of (6) is monotonically increasing with the value d used on the r.h.s. of the same equation. It then follows, from this fact and from Corol. 4.1, that no pattern $P \in \mathrm{FP}_\tau(V)$ may have $\mathrm{f}_S(P) < \tau - \varepsilon_1$, therefore the refinement of $\mathcal{H}_1$ on line 10 is such that it still holds $\mathcal{F}_1 \subseteq \mathcal{H}_1$ at the end of the first iteration of the do-while loop. Following the same reasoning one can show that this condition and the fact that $\varepsilon_1 \ge \eta_1$ throughout every iteration of the do-while loop.

The set $\mathcal{Q}$, updated on line 12, therefore contains, among others, a triplet for every pattern $P \in \mathrm{FP}_\tau(V)$, and the properties from the thesis hold because of this fact and the fact that $\varepsilon_1 \ge \eta_1$, thus completing the base case for point (1) in the list above. Point (2), i.e., that $\mathcal{F}_2 \subseteq \mathcal{H}_2$, then follows from the anti-monotone property of the MNI-frequency (Fact 2).

Assume now that points (1) and (2) hold at every iteration of the while loop from $i = 1, \ldots, i^* < k$. The proof that they hold at the end of iteration $i^* + 1$ follows the same reasoning as above. □

A.3 Additional experiments

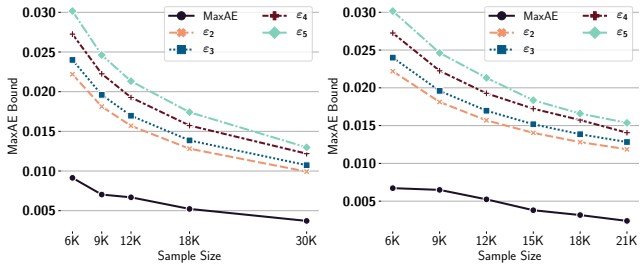

Figure 5: Max Absolute Error (MaxAE), ε_2, ε_3, ε_4, and ε_5, for various sample size, min frequency threshold $\tau = 0.16$, in Patents (left) and in Phy-Cit (right).