We thank the reviewers for providing thorough and helpful reviews!

High-level overview of components, novelty, and comparison with [9]. One concern that was raised by multiple reviewers is that the submission does not explain how the components described in Sections 4 and 5 fit in the narrative, and which parts are the most interesting/novel technical contributions. We provide here an overview and will update the camera-ready version accordingly.

In Section 3, we represent $f(\nu_x) = \mathcal{L}^c[a](\nu_x)_0^\gamma + \mathcal{L}^c[a](\nu_x)_\gamma^\infty$, and for each summand we maintain a separate sample of size k (which will later be merged). This representation was used in [9] for the simpler task of estimating the full f-statistics of the data but in this work we need to produce samples according to these contributions.

For $\mathcal{L}^c[a](\nu_x)_0^\gamma$, we maintain a standard PPSWOR sketch. However, to facilitate sampling according to $\mathcal{L}^c[a](\nu_x)_\gamma^\infty$, we needed to develop novel techniques. The first challenge was to get a handle on the contributions $\mathcal{L}^c[a](\nu_x)_\gamma^\infty$, and for that purpose we designed the SumMax sketch structure (Section 5). The input elements are mapped to output elements with multiple subkeys, so that the sum over subkeys of the maximum element is a random variable with expectation $\mathcal{L}^c[a](\nu_x)_\gamma^\infty$. The SumMax sampling structure provides a PPSWOR sample according to this sum of maxima.

Another challenge is that for each x, $\mathsf{SumMax}(x)$ is not the exact value $\mathcal{L}^c[a](\nu_x)_\gamma^\infty$, but a random variable with this value in expectation. For that, we introduce the analysis of PPSWOR with stochastic inputs (Section 4). In that analysis, we establish the conditions that are needed in order for the sample according to the random values to be close to a sample according to the expected values. We then design a sketch structure that meets these conditions.

The most interesting technical contributions of the paper, that are also of independent interest, are the SumMax sampling structure and especially the analysis of PPSWOR with stochastic inputs, which shows that under relatively mild conditions, we can get a sample with good estimators when the input to PPSWOR is randomized.

To clarify following the questions posed by Reviewer 4: Each of the two samples we maintain (the PPSWOR and SumMax samples) have a fixed size and store at most k keys at any time. The γ threshold is chosen to guarantee that we get the desired approximation ratio. The only structure that can use more space is the Sideline structure. As part of the analysis, we bound the size of the Sideline and show that in expectation, it is $O(k)$ and also provide worst case bounds on its maximum size during the run of the algorithm. The output elements that are processed by the SumMax sketch have a value that depends on γ (which changes as we process the data), and the purpose of the Sideline structure is to store elements until γ decreases enough that their value is fixed (and then they are removed from the Sideline and processed by the SumMax sketch).

Reviewers 1 (comment 3) and 2 asked what are the differences between this submission and the results in [9]. For any soft concave sublinear function f, the sketch in [9] can estimate the statistics $\sum_{x\in\mathcal{X}} f(\nu_x)$ over the entire dataset. Our sketch outputs a sample that can be used to estimate statistics of the form $\sum_{x\in H} L_x f(\nu_x)$ for any $H \subseteq \mathcal{X}$ and $L_x \geq 0$.

It is highly non-trivial to develop the additional components needed for a sampling structure. In particular, the sketch in [9] gives us a way to map elements into "output" elements with the desired expected value. However, using the framework of [9] to produce a sample required the introduction of stochastic PPSWOR sampling, the SumMax sketch, and a modification of the way γ is chosen.

Tightness of the upper bounds. Reviewer 1 comment 2: The upper bounds for PPSWOR are essentially tight for datasets that are not very skewed (and that is why it makes sense to compare to them when designing an algorithm for the worst case). When a few elements dominate and their weight is a large fraction of the total weight, the variance can be lower. In particular, if we look at the variance of the benchmark PPSWOR in the experimental results, we see that the NRMSE is at the bound $1/\sqrt{k}$ in almost all experiments.

Reviewer 3: We do not have a lower bound to match our result. Reducing the constant 4 in the theoretical analysis is an interesting direction for further research. Going deeper into the analysis of stochastic PPSWOR, we can hope to improve 4 to 2, but at the expense of using more space. However, from a practical perspective, the error that we get in the experiments is very close to the optimal bound of PPSWOR (without the factor 4). The additional factor of $(1 + 1/(e-1))^2$ when considering non-soft concave sublinear functions is needed, as our approach relies on approximating this family using soft concave sublinear functions.

Answers to additional questions by Reviewer 1: 5. When we say in line 100 that all previous methods (including PPSWOR) require aggregation, we refer to the problem of sampling with respect to a concave sublinear function to the frequencies. In line 114, we say that if we sample according to the frequencies themselves (without applying any function), PPSWOR can be used on unaggregated data.
8. In order to compute an estimate of the target statistics, we use a conditional variant of the Horvitz-Thompson estimator ($\nu_x / \Pr[x \text{ in sample}]$). The term "conditional inclusion probabilities" in line 219 refers to the probability in the denominator, which we explain how to compute in the supplement/full version of the paper.