This document provides the reasoning behind the following statement used in papers of Vladimir Mic et al.:

**Lemma 1.** *The error in symmetry of function $p(x, b)$ is negligible when using a sketching techniques that produce low correlated bits.*

also, we provide its empirical verification.

Let us recall, that probability function $p(x, b)$ is derived by composition of $\lambda$ instances of function $p_i(x, 1)$. If these instances of $p_i(x, 1)$ are pairwise independent, the symmetry of function $p(x, b)$ is given by the symmetry of the binomial distribution. Let us consider a set of $n = |X|$ sketches with balanced bits and two arbitrarily selected bits $i$ and $j$. We denote $H_i$ the list of all $n^2$ Hamming distances measured using just one bit $i$. Please notice, that distribution of values in list $H_i$ determines probability function $p_i(x, 1)$ for bit $i$. Therefore, a level of independence of instances $p_i(x, 1)$ and $p_j(x, 1)$ for bits $i$ and $j$ is given by correlation of lists $H_i$ and $H_j$. Mic et al. prove by Theorem 8 in [1], that if the bits $i$ and $j$ are balanced, the correlation $Corr(i, j)$ determines the correlation of lists $H_i$, $H_j$:

$$Corr(H_i, H_j) = Corr(i, j)^2 \tag{1}$$

both measured by Pearson correlation coefficient. Therefore, small correlations of bits $i$ and $j$ imply even smaller correlations of lists $H_i$, $H_j$, due to the second power in Equation 1, which ensures high level of independence of instances $p_i(x, 1)$ for bits $i$ and $j$. As a result, symmetry of function $p(x, b)$ can be damaged only negligibly for low correlated bits.

An empirical verification of this reasoning is shown in Table 1. All the correlations are measured on the DeCAF dataset, and the average correlation $Corr(H_i, H_j)$ is tiny in all cases. Therefore, we can expect the function $p(x, b)$ to be nearly perfectly symmetric around the main peak.

Despite of high independence of particular instances $p_i(x, 1)$, it is necessary to take pairwise correlations into account. If not, the final binomial analogue does not match the reality at all. An example is provided in Figure 1. In this experiment, we show (1) measured values of probability $p(x, b)$, depicted in black colour and denoted $p_{measured}(x, b)$, (2) the proposed binomial analogue, denoted $p(x, b)$, and (3) wrong binomial analogue assuming independent instance of probability

| Sketch length $\lambda$ | Average $Corr(i, j)$ | Average $Corr(H_i, H_j)$ |
|---:|---:|---:|
| 32 | 0.085 | 0.007 |
| 64 | 0.101 | 0.010 |
| 128 | 0.111 | 0.012 |
| 205 | 0.117 | 0.014 |
| 256 | 0.121 | 0.015 |
| 4,096 | 0.150 | 0.023 |

Table 1: Measured average correlations

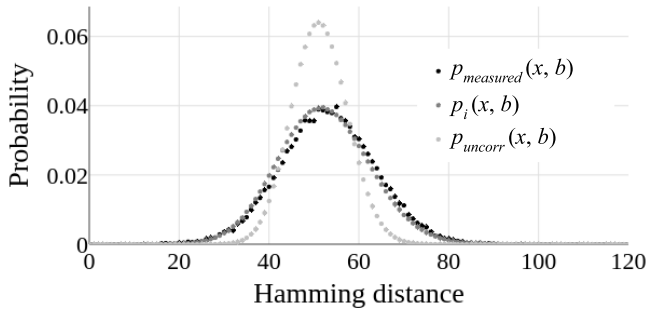


Fig. 1: Correct and wrong binomial analogue $p(x, b)$ for fixed $x$ implying $p(x, 1) = 0.25$

functions $p_i(x, 1)$, denoted $p_{uncorr}(x, b)$. Clearly, the last mentioned binomial analogue does not match measured values well.

## References

1. Mic, V., Novak, D., Zezula, P.: Designing sketches for similarity filtering. In: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW). pp. 655–662 (Dec 2016)