**Abstract**

We study distributed protocols for finding all pairs of similar vectors in a large dataset. Our results pertain to a variety of discrete metrics, and we give concrete instantiations for Hamming distance. In particular, we give improved upper bounds on the overhead required for similarity defined by Hamming distance $r > 1$ and prove a lower bound showing qualitative optimality of the overhead required for similarity over any Hamming distance $r$. Our main conceptual contribution is a connection between similarity search algorithms and certain graph-theoretic quantities. For our upper bounds, we exhibit a general method for designing one-round protocols using edge-isoperimetric shapes in similarity graphs. For our lower bounds, we define a new combinatorial optimization problem, which can be stated in purely graph-theoretic terms yet also captures the core of the analysis in previous theoretical work on distributed similarity joins. As one of our main technical results, we prove new bounds on distance correlations in subsets of the Hamming cube.