18 DIMSUM

We'd like to compute entries for $A^T A$ for which $\cos(r_i, r_j) \geq s$ for some threshold $s$. Columns of $A$ are vectors, and vectors can have similarities. We need the following notion of similarity of two vectors [Cosine Similarity] The *cosine similarity* between two columns $c_i$ and $c_j$ is defined as

$$\cos(c_i, c_j) = \frac{c_i^T c_j}{c_i^T c_j}.$$

**Algorithm 1** DIMSUMMapper ($r_i$)

```
for all pairs ($a_{ij}, a_{ik}$) in $r_i$ do With probability $\min\{1, \frac{1}{c_j^T c_k}\}$, emit (($j, k) \rightarrow a_{ij} a_{ik}$)
```

**Algorithm 2** DIMSUMReducer (($i, j), v_1, \ldots, v_R$)

```
if $\frac{1}{c_i^T c_j} > 1$ then return $b_{ij} \rightarrow \frac{1}{c_j^T c_j} \sum_{i=1}^{R} v_i$ return $b_{ij} \rightarrow \frac{1}{\sum_{i=1}^{R} v_i}
```

The *Dimension Independent Matrix Square using MapReduce* (DIMSUM) algorithm is described in [1] and [2]. The DIMSUM algorithm outputs the cosine similarities (in fact probabilistic estimates of the cosine similarities). Also note that you need to compute the norms of columns beforehand (which requires all-to-all communication).

**References**
