Spatially-Decaying Aggregation Over a Network: Model and Algorithms

Edith Cohen
AT&T Labs–Research
180 Park Avenue
Florham Park, NJ 07932, USA
edith@research.att.com

Haim Kaplan
School of Computer Science
Tel Aviv University
Tel Aviv, Israel
haimk@cs.tau.ac.il

ABSTRACT

Data items are often associated with a location in which they are present or collected, and their relevance or influence decays with their distance. Aggregate values over such data thus depend on the observing location, where the weight given to each item depends on its distance from that location. We term such aggregation *spatially-decaying*.

Spatially-decaying aggregation has numerous applications: Individual sensor nodes collect readings of an environmental parameter such as contamination level or parking spot availability; the nodes then communicate to integrate their readings so that each location obtains contamination level or parking availability in its neighborhood. Nodes in a p2p network could use a summary of content and properties of nodes in their neighborhood in order to guide search. In graphical databases such as Web hyperlink structure, properties such as subject of pages that can reach or be reached from a page using link traversals provide information on the page.

We formalize the notion of spatially-decaying aggregation and develop efficient algorithms for fundamental aggregation functions, including sums and averages, random sampling, heavy hitters, quantiles, and L_p norms.

1. INTRODUCTION

In many applications, data items are associated with locations on some network. Data present at one location is relevant to other locations, yet, this relevance or influence decreases with the distance between the locations. Thus each location views and aggregates the data through a different distribution, where the *weight* given to each item decreases with its distance. This dependence on the distance is quantified by a *decay function*, which is a non-increasing function. Some natural decay functions are threshold functions (BALL_r), where items in the r-neighborhood have uniform weights (and other items have 0 weight), Exponential decay, where the weight decreases exponentially with the distance, and Polynomial decay, where the weight decreases polynomially with the distance. While global aggregates assigns equal weight to all items, and result in a single global value, *spatially decaying* aggregates depend on the decay function used and on the location with respect to which the aggregation is performed.

SIGMOD 2004, June 13–18, 2004, Paris, France.

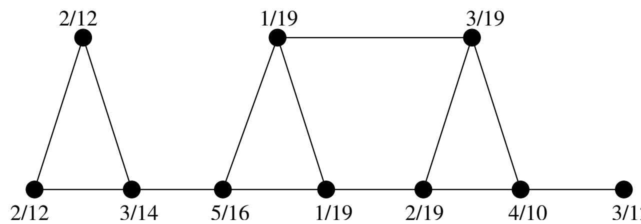


Figure 1: A network where each node is labeled by the value of the item it carries and the sum of values in its 2-neighborhood (BALL_2 decaying sum).

Aggregate functions. We consider several fundamental aggregate functions. With *spatially-decaying sum*, each node obtains an estimate on the *sum* of the weight times value product over all items. Figure 1 shows an example of a network with values at each node and the respective decaying sums under the BALL_2 decay function. A related aggregate, derivable from the sum, is the *spatially-decaying average* (a weighted average of values of items). Spatially-decaying random samples are defined with respect to the weight distribution (for a BALL_r decay function a node obtains a uniform random sample from its r-neighborhood). Using random sampling as a building block we obtain algorithms for the spatially-decaying variants of other basic aggregation problems such as approximate frequency counts [30] and quantiles [32, 31]. Over the set of *distinct elements* (set of distinct values over items with nonzero weight) we consider the counting, multiplicity (frequency) of a random distinct element, and random sampling aggregates. We also consider approximate *spatially-decaying L_p norms* for $p \in [1, 2]$ (of vectors where items constitute spatially-decaying updates to certain coordinates [14, 11]).

Our model and algorithms supports a wide range of applications that including distributed or centralized aggregation, one-time computation or periodic to support continuous queries. The desired result of the aggregation process can be a single numeric result at each node or a *summary* that enables a node to respond locally to aggregation queries for different decay functions or aggregate functions. Our algorithms can be modified to perform aggregations over the set of reverse nearest neighbors (RNN). Such aggregates are of interest for decision support and had been studied on stored data sets and on data streams [25, 26].

Applications. Our work is motivated by several application areas (both distributed and centralized):

- **Sensor networks:** Sensor networks consist of many individual sensor nodes, each collecting data on its environment, for example, temperature, density of some contaminant, identities of passing vehicles, or availability of a parking spot (see, e.g.[13, 29, 12]). Each reading by one sensor is relevant to other locations, but the relevance, for example of an available parking space, decreases with its distance from the destination [16, 17, 12]. Spatially-decaying aggregation provides each location with information on their larger neighborhood such as parking availability, average or median temperature, the number of distinct vehicles passing through, or the median speed of a vehicle. RNN aggregates can be used to decide when a certain node should start or stop providing some service to its neighboring nodes. Since power is a limiting factor in sensor nets, it is important that aggregation is performed efficiently [28, 33]. Since node distribution can be highly irregular [17, 12], it is important to handle general (or arbitrary planar) topologies.

- **P2P networks:** P2P networks consist of nodes connected via an overlay network. An important functionality they provide is search, where nodes initiate queries for data that can be present at other nodes. Some popular P2P systems (such as the Gnutella protocol [9]) deploy flood searching, where the query is propagated for a fixed number of hops, thus to a neighborhood that includes all nodes that lie within several hops away. Aggregates over the neighborhood that can be used to tune the efficiency of the search process include the number of hosts, the total number of (distinct or not) items offered, the median load or bandwidth of a node, the total number of distinct hosts issuing queries or the total number of queries seen by neighboring nodes. The search process can be improved with each node obtaining a summary of the types of content present in its neighborhood and in the neighborhood of its immediate neighbors [10].

- **Graphical databases:** Many databases have natural graph form with data present at the nodes. On hyperlink graphs of Web pages, the context or importance of a Web page is indicated by aggregate information on pages it links to. For XML documents representing bibliographic data, a context for a new article or book can be deduced by articles authored by collaborators of its author. For road maps, it can be useful to obtain for each location the number of restaurants or gas stations in its proximity. For images, spatially-decaying aggregates can assist in feature extraction.

Metrics and challenges: The algorithmic challenge in efficient spatially-decaying aggregations is that aggregate values are location-dependent. Each data items influence many locations, and to a different extent. The naive approach is to collect for each node values of all items that influence the aggregate (have a nonzero weight) and then to compute from scratch the aggregate value. This scales poorly, as it can result in a quadratic running time in a centralized computation, and in quadratic communication in a distributed setting (for flooding all values to all locations in which they are considered). It is not hard to see that for many aggregates and decay functions, such quadratic bounds are necessary for obtaining *exact* aggregate values. We thus focus on obtaining *approximate* values. Our approximations are within a small $(1+\epsilon)$ relative error with very high confidence which suffices for many applications. We develop novel summarization techniques and obtain algorithms that use per-node communication that depends *polylogarithmically* (rather than linearly) on the size of the *influencing* data. Our centralized algorithms run in nearly linear time in the size of the data base.

Related work: The spatial decay model generalizes (from an algorithmic perspective but not so much from an application perspective) the sliding-window model [11, 18] for massive data streams [3] and more generally, the computation of time-decaying aggregates on massive data streams [8]. In particular, the threshold ($\textsc{Ball}_r$) decay function generalizes sliding windows. There is a straightforward reduction from tracking a time-decaying aggregate (e.g., over a sliding window) to spatially decaying aggregation (of the same aggregate and under the same decay function) on a path network. This allows to carry over some lower bounds from the sliding window model to our model. On the other hand the spatial decay setting seems considerably more complex and existing techniques for sliding windows and time-decaying aggregates (such as Exponential Histograms [11] for sums and deterministic summaries [19, 27] for quantiles) do not seem to extend to the general spatial setting. Exponentially decaying sum, which is trivial to efficiently track as a time-decay function [22, 15] is not known to be any easier to track than other decay functions in the spatial-decay model. The spatial decay model also generalizes computation of non-decaying aggregates in a network (e.g. [33]). From a framework point of view, the spatial decay model is conceptually related to the spatial gossip model of Kempe, Kleinberg, and Demers [24, 23]. The main difference is that in spatial decay, communication is performed along the edges of an *arbitrary graph* and the cost we consider is per-node communication cost. In spatial gossip protocols, communication is performed by contacting *random* nodes with likelihood that depends on distance, and cost is measured by the time it takes information to spread. In many applications, including sensor networks and graphical databases, node and link placement are given and can be rather irregular, thus, it is important to handle general topologies. Last, basic SQL type aggregates over neighborhoods in sensor nets had been recently studied [17, 12]. Deshpande et al [12] considered a fixed partitioning into neighborhoods (for example, parking availability in Tribecca rather than parking availability within 5 minute walk of destination), which makes for a rigid handling of boundaries which can be less natural in applications where the physical world is not so partitioned. Ganeriwal et al [16] did not consider the algorithmic efficiency of performing average aggregates, but proposed Voronoi diagrams as a way of assigning a weight to each sensor reading for a per-area weighted average aggregation, an issue that is orthogonal to ours. Our algorithms use techniques we developed in [5] to efficiently sum values over neighborhoods, but several key new ideas and techniques were necessary in order to handle different aggregate functions, general decay functions, and performing the aggregations in a distributed setting.

Our Contributions: We formalize and motivate the notion of *spatially-decaying* aggregation and develop algorithmic techniques to efficiently approximate fundamental aggregates. The aggregates we consider include basic SQL aggregates and had been previously studied over data sets, data streams, or time-decaying aggregation over data streams, but our spatial model is more complex and general and necessitated developing and applying a new set of techniques. Our algorithms are nearly optimal (are within polylogarithmic factors from the lower bounds) and vastly improve over naive approaches.

Outline We organize our presentation as follows. In section 2 we introduce the spatial decay model in its general setting, define the main problems we study and some natural decay functions. Section 3 is concerned with computing decaying sums for any decay function. Section 4 focuses on computing MV/D lists, a basic technique used for computing decaying sums and other aggregate functions. In Section 5 we develop algorithms for random sampling, Section 6 is concerned with approximate L_p norms, and Section 7 with counting distinct values. Section 8 deals with the important case of the Euclidean metric. In Section 9 we present lower bounds based on the relations between spatial decay and sliding windows and in Section 10 we explore a unified model which captures both spatial and time decay. Finally, Section 11 contains an experimental study which compares and evaluates different techniques.

2. PRELIMINARIES

We model the network as a graph $G = (V, E)$, where $V = \{v_1, \ldots, v_n\}$ is the set of nodes, and there is an edge between two nodes if and only if the two nodes can communicate. We denote the number of edges by m. Edges can have nonnegative lengths associated with them, which are interpreted as distances. We consider both directed and undirected graphs (symmetric or asymmetric distances). We denote by $\textsc{dist}(v_i, v_j)$ the distance between two nodes v_i and v_j with respect to the shortest-path metric on the edge lengths. Nodes in the network have a set of data items $\{(f_i, \ell_i)|i \in I\}$, where for each item $i \in I$, $f_i \geq 0$ is the value of the item (which can generally be a vector or non-numeric entities) and $\ell_i \in V$ is the location of the item.

A *decay function* is a non-increasing function $g(x) \geq 0$ defined for $x \geq 0$. The decay function determines the *weight* of a remote item as a function of its distance. The weight of an item $i \in I$ as viewed by a node $u \in V$ is $w_{u,g}(i) = g(\textsc{dist}(u, \ell_i))$. An *aggregate function* is a function $h()$ defined on a multiset of a value-weight pairs. Given a network G with data items $\{(f_i, \ell_i)|i \in I\}$, a decay function $g()$, and a node u, the spatially-decaying $h()$-aggregate at the node u is the value of h on the multiset $\bigcup_{i \in I}\{(f_i, w_{u,g}(i))\}$. Our definitions and results can be generalized to the case where items have arbitrary "initial" weights, w_i^0, and the weight of the item at distance ℓ_i is $w_{u,g}(i) = w_i^0 g((\textsc{dist}(u, \ell_i)))$. For simplicity of presentation we subsequently only treat the case where "initial" item weights are uniform.

We denote the value of an aggregate H according to decay function $g()$ at location u by $H_g(u)$. Our goal is to obtain (estimates) of the decaying aggregate value $H_g(u)$ at all locations $u \in V$. We seek algorithms that minimize the amount of communication between nodes, as well as the computation and storage required at each node. The algorithms we develop here are general in the sense that rather producing values for a particular aggregate and a specific decay function they produce a *summary* at each node u from which an approximate value of the aggregate $H_g(u)$ can be computed for **any** decay function $g()$.

2.1 Aggregate functions

Sum and count. The *spatially-decaying sum* is such that the function h sums up the value components $f_i \geq 0$ of the pairs in the multiset, each attenuated by the corresponding weight, that is, the spatially-decaying sum at node u is

$$S_g(u) = \sum_{i \in I} f_i w_{u,g}(i) \ .$$

In the special case where the values f_i are binary, we refer to this aggregate as the *spatially-decaying count*. Our algorithms produce a summary at each node, that allows it to obtain with very high confidence, for *any* decay function $g()$, $(1 \pm \epsilon)$-approximate estimates, $S'_g(u)$, where $\epsilon > 0$ is fixed. That is, $S'_g(u)$ are such that $\frac{|S'_g(u) - S_g(u)|}{S_g(u)} \leq \epsilon$ for every $u \in V$.

Average. A related aggregate is the *spatially-decaying average*, defined as

$$A_g(u) = \frac{\sum_{i \in I} f_i w_{u,g}(i)}{\sum_{i \in I} w_{u,g}(i)} \ .$$

The numerator is the spatially decaying sum and the denominator is the spatially-decaying count with respect to the data items $\{(1, \ell_i)|i \in I\}$. Thus, an approximate decaying average can be obtained from corresponding approximate decaying sum and approximate decaying count.

Random sampling:. For a node u and a decay function $g()$, $R_g(u)$ is a random variable that returns an item $i \in I$ with probability proportional to its weight, that is, the probability that i is drawn is equal to

$$\frac{w_{u,g}(i)}{\sum_{i \in I} w_{u,g}(i)} \ .$$

Each application of our algorithm results in a summary at each node u from which we obtain an item drawn according to $R_g(u)$ for any decay function $g()$. (We allow dependencies between items drawn for different decay functions and different nodes).

Quantiles:. For some $0 \leq p \leq 1$ and any decay function $g()$, u can obtain an item with value that is with high probability a $[p \pm \epsilon]$-quantile of the distribution of values over weighted items. Using an existing folklore technique, an approximate quantile with constant confidence arbitrarily close to 1 can be obtained using a constant number of independent random samples.

Frequency counts (heavy hitters):. The *spatially-decaying frequency counts (heavy hitters)* is to find all elements such that the total relative weight of their occurrences is at least $p+\epsilon$. If an element is reported then its total weight is at least $p-\epsilon$. Approximate heavy hitters can be computed from independent random samples [30].

L_p norms ($p \in [1,2]$):. Each item $i \in I$ is an update of a coordinate of a d-dimensional vector and is specified by a triplet (c_i, a_i, ℓ_i). The value $c_i \in [d] = \{0,\ldots,d-1\}$ specifies the coordinate which this item updates. The value $a_i \in \{0,\ldots,M\}$ is the amount by which we increment the target coordinate, and $\ell_i \in V$ is the item's location. The d-dimensional vector $\mathbf{V}(g,u)$ associated with location $u \in V$ is then defined by the coordinate values

$$\mathbf{V}(g,u)_j = \sum_{i|c_i=j} w_{u,g}(i)a_i \ .$$

For a fixed $p \in [1,2]$, we show how to obtain, using $o(d)$ communication per node, at each location u, a summary of size $o(d)$, such that for any decay function $g()$, the node can obtain an approximate value of $||\mathbf{V}(g,u)||_p$ (the L_p norm of $\mathbf{V}(g,u)$). One ingredient in our solution is a technique by Indyk [21] for computing sketches of vectors which preserve approximate L_p differences (for $p \in [1,2]$). The problem Indyk considered (for non-decaying data streams) was introduced in [14]. For the problem of computing approximate L_p norms in the sliding window model, Datar et al [11] developed an algorithm which combines Indyk's vector sketches with their Exponential Histogram (EH) technique. Their solution, however, does not seem to generalize to the spatial setting.

Distinct values:. For any node u and decay function $g()$, $D_g(u) = |\{f_i | w_{u,g}(i) > 0\}|$ is the number of distinct values over items that have a positive weight (The problem is therefore equivalent to its restriction to BALL_r decay functions, that is, for any given $r \geq 0$, obtain an approximate number of distinct values present in the r-neighborhood.) We show how each node u can obtain a summary such that for any $g()$ it can obtain an approximate value of $D_g(u)$. We are also able to obtain a *random* distinct element, that is, a distinct element drawn uniformly at random, and (approximate) multiplicity of a random distinct element.

Another family of important aggregates is the (approximate) spatially-decaying variance and moments. In [7] we reduced this problem to computing (approximate) spatially-decaying sums and medians, which are studied here.

2.2 Decay functions

Our results hold for general decay functions. Families of decay functions that are of particular interest are: *Ball decay functions (BALL_r)* are parameterized by the radius r, and have $g(x)=1$ for $x \leq r$ and $g(x)=0$ otherwise. That is, all data values within a distance of r have equal importance and all other data values have 0 weight. *Exponential decay (EXPD_λ)* is such that for a parameter $\lambda > 0$, $g(x) = \exp(-\lambda x)$. With EXPD, the relative significance of each value decreases exponentially with its distance. Exponential decay is commonly used in practice for time-decay. One reason for its common use is that it can be maintained easily as a time-decay function, using a single register (It is not clear, however, if it is simpler to compute than other decay functions in the spatial setting on general networks). *Polynomial decay* (POLYD_α) is such that for a parameter α, $g(x) = 1/x^\alpha$. Polynomial decay is often a natural choice when a smooth decay is desired and when Exponential decay is too drastic. In many natural graphs (like d-dimensional grids), the neighborhood size increases polynomially with the distance. For such graphs, EXPD would suppress longer horizons. Many natural effects (for example, electro-magnetic radiation) have polynomial decrease with distance.

2.3 Model assumptions

Spatially-decaying aggregation has applications in a centralized setting, where performance is measured by running time and storage and distributed settings where we consider communication and per-node storage. In a distributed setting we also consider a routing mechanism that allows nodes to broadcast along a shortest path tree. In some settings the complete topology or at least some routing table can be obtained at each node (see [33]). In other more dynamic settings (such as P2P networks) each node is only aware of its neighbors. The distance metric can be edge-length or Euclidean depending on the application. Edge-length metrics may capture number of network hops (with uniform edge lengths) or propagation time (where the length of each edge is the sum of the latency and the processing time at the end nodes).

3. DECAYING SUM

We will make use of data structures that we term *Neighborhood summaries* (NH-summaries). These data structures are maintained in each node of the network and support *neighborhood-sum queries*: For any given radius $r \geq 0$, the NH-summary provides node u with $(1+\epsilon)$ approximation of $S_{\text{BALL}_r}(u)$: the total sum of values of items that are within distance r. Also of interest are *d-limited NH-summaries*, which support neighborhood-sum queries over r-neighborhoods for $r \leq d$.

NH-summaries generalize *Window-summaries* on data streams which provide, for each time window t, the sum of values of items observed in the last t time units. (Approximate) window summaries can be obtained using *Exponential Histograms* (EH) introduced by Datar et al [11]. They use $O(\epsilon^{-1}\log^2 W)$ memory bits for obtaining $(1+\epsilon)$-approximate estimates for windows up to size W.

We next argue that NH-summaries are general enough to support the computation of spatially-decaying sums under any decay function. The following lemma generalizes a lemma in [8] that shows that in the context of time decay, Window-summaries can be used to compute time-decaying sums under general decay functions.

LEMMA 3.1. *NH-summaries allow us to compute spatially-decaying sums under any decay function. If the summaries are approximate then so are the sums obtained.*

PROOF. Consider some node u, and assume the items are numbered in increasing order of their distance to u. For convenience, assume that distances are integers. Let $\text{DIST}(u,\ell_i) = d_i$. Let $\Delta_j = g(j) - g(j+1)$. Since $g()$ is non-increasing, $\Delta_j \geq 0$ for all j. (w.l.o.g. we assume that $g(x) \to 0$ as x increases. Thus, $g(x) = \sum_{k \geq x} \Delta_k$. Otherwise, the problem is composed of a non-decaying problem and a decaying sum problem where the decay function goes to 0.) The decaying sum for u can be written as:

$$\begin{aligned} S_g(u) &= \sum_{i \in I} f_i g(d_i) = \sum_i f_i \sum_{k \geq d_i} \Delta_k = \sum_i \sum_{k \geq d_i} f_i \Delta_k = \\ &= \sum_k \Delta_k \sum_{i | d_i \leq k} f_i = \sum_k \Delta_k S_{\text{BALL}_k}(u) \ . \end{aligned}$$

Thus, $S_g(u)$ is expressed as a linear combination of $S_{\text{BALL}_r}(u)$ decaying sums. □

Lemma 3.1 can be extended to using d-limited NH-summaries to express spatially-decaying sums for all *d-limited decay functions* (decay functions with $g(d) = 0$ and thus $g(x) = 0$ for all $x \geq d$).

When computing the sum, the number of summands can be reduced: When the approximate summary provide the same estimate $\hat{S}_{\text{BALL}_r}(u)$ of $S_{\text{BALL}_r}(u)$ for a range of values $a \leq r \leq b$, we can "replace" the sub-sum $= \sum_{k=a}^{b} \Delta_k \hat{S}_{\text{BALL}_k}(u)$ with the single term $(g(a) - g(b+1))\hat{S}_{\text{BALL}_a}(u)$.

The data structure that we use for our NH-summaries are *Min-value/Distance lists* (MV/D lists). MV/D lists were introduced in [5] in the context of a fast estimation algorithm for neighborhood sizes of nodes in a graph and the following was established:

LEMMA 3.2. *MV/D lists with ranks drawn independently at random can be used as NH-summaries to answer NH queries. When using k (independent) lists, the probability of relative error larger than ϵ is* $\exp(-\Omega(\epsilon^2 k))$. *(Thus for a fixed ϵ, the probability that relative error exceeds ϵ can be made an arbitrarily low constant using a constant number of MV/D lists and polynomially-low using a logarithmic number of lists.)*

Combined with Lemma 3.1, we thus far reduced the problem of obtaining approximate decaying sums to computing MV/D lists.

4. MV/D LISTS

We use MV/D list computation in our algorithms for many aggregates. MV/D lists are defined with respect to numerical values r_i $(i \in I)$, called *ranks*, that we associate with each data item. We use ranks that are drawn independently at random. In principle, many distributions can be used to draw the ranks, but a convenient distribution is the Exponential distribution with parameter f_i. (Items with $f_i \equiv 0$ get ranks $r_i = +\infty$.) For simplicity of the presentation we assume that there is at most one item in each node and we focus on the decaying count problem, where the values of items are binary. We outline how to extend our results to allow for multiple items in each node and arbitrary nonnegative values. The term n (that corresponds to the number of nodes) in the upper bounds should be replaced by the ratio of the sum of values of all items to the smallest value of an item, subsequently we refer to this quantity as n^*. Note that if the number of items is bounded by a polynomial in n and each value is integral and bounded by a polynomial in n, then n^* is polynomial in n and $\log n^* = O(\log n)$. Moreover, if we restrict our attention to d-limited decay functions then n, and n^* in the derived bounds can be replaced by the respective quantities, n_d or n_d^* (number and ratio of sum to smallest values in a d-neighborhood). d-limited decay functions do not include smooth functions like polynomial or exponential decay, for such functions, all items in the system can influence any location. But when we can restrict ourselves to such functions, the saving can be considerable: for bounded degree networks the size of an r-neighborhood is bounded by a function of r and the communication is thus polynomial in r (polylogarithmic in the size of the r-neighborhood and independent of the total number of nodes).

The *MV/D list* of a node u is a summary that allows us to retrieve, for any distance d, the minimum rank associated with an item that lies within distance d from u. An MV/D list has the form

$$((0 = d_0, R_0), (d_1, R_1), (d_2, R_2), (d_3, R_3), \ldots)$$

where $d_j < d_{j+1}$ are increasing and $R_j > R_{j+1}$ are decreasing. We refer to each distance-rank pair as an *element* of the list. The MV/D list has the property that for all j, the minimum rank found within distance $d \in [d_j, d_{j+1})$ from u is R_j. The following was established in [5]:

LEMMA 4.1. *When ranks are drawn independently at random, the expected* size *of the respective MV/D lists (in terms of number of elements in the list) is logarithmic in the number of nodes n. (When item values are arbitrary and there are multiple items per node, the size is logarithmic in n^*.)*

4.1 Computing MV/D lists

We start by reviewing a main-memory algorithm to compute MV/D lists [5] that was used for estimating neighborhood sizes. We then address interesting and subtle issues regarding the implementation of the technique of [5] in a distributed setting. For example we relax the rigid order in which nodes announce their ranks in [5] and show that it costs only a logarithmic factor in time (or communication in our distributed setting). For each node, we refer to the minimum rank of an item present at a node u, as the rank $r(u)$ of the node. Nodes that have no item associated with them get a rank of infinity. All nodes of the graph (with finite ranks) are sorted according to their ranks. In increasing order of ranks, we then perform from each node a *truncated BFS* pass (for uniform edge lengths) or a *truncated Dijkstra shortest-paths* pass (for non-uniform edge lengths). We refer to a truncated pass starting at a node u as an *announcement* made by u.

The truncated BFS pass proceeds very much like a complete BFS pass initiated at u. When the pass determines for a node u' that its distance from the announcing node u is $\text{DIST}(u, u')$, we proceed as follows: if $r(u)$ is smaller than the minimum rank of a node seen so far of distance at most $\text{DIST}(u, u')$, the MV/D list of u' is updated to include the entry $(\text{DIST}(u, u'), r(u))$ and the pass proceeds through the outgoing edges of u'. Otherwise, the pass is halted from u' on (but may proceed from other nodes of the same distance, if the MV/D list of those nodes was augmented).

An announcement uses at most $O(m)$ time, but if truncated could use much fewer edges. It is shown in [5] that when nodes make announcements in increasing order of their ranks, the (centralized version of the) computation takes $O(m \log n)$ expected time. This bound stems from the following considerations. First, the number of announcements propagated by a certain node is exactly the size of its MV/D list (to see this, observe that since smaller ranks are announced before higher ranks, when an announcement reached a node and provoked an update of its list, this update can not be overridden by future announcements.) Second, the work performed per announcement is proportional to the number of edges incident to the nodes reached by the announcement. Third, when the ranks order constitutes a random permutation (in particular, are not correlated with distances), the expected size of the MV/D list is logarithmic (see Lemma 4.1).

We will next consider the computation of the MV/D lists in a distributed network, where nodes communicate with each other along the edges of the network. Our main building block remains the "announcement process" where a node announces its rank to other nodes, but two important issues arise. First we need to consider the communication involved in making a single announcement. Second we need to consider the order in which announcements are made, since in a distributed setting it is undesirable to sort and synchronize so nodes are invoked in the order of their ranks.

We first consider the distribution of a single announcement in the network. When the topology of the network is known at all nodes, each node can compute shortest path trees and use them to make sure each announcement follows the shortest path tree of its origin. This way an announcement would require number of messages that is linear in the number of tree edges incident to visited nodes. When the complete topology is not available to nodes but distances correspond to propagation delay then each announcement can be performed via flooding from the source, using number of messages proportional to the number of edges incident to the nodes it reaches. Otherwise, if topology is not known the nodes can perform a distributed BFS or shortest path computations [2]. BFS can be performed in $O(D^{1+\epsilon})$ time and $O(m^{1+\epsilon})$ messages, where D is the diameter of the network and a time unit is the time required for a message to pass through an edge. In our setting, D can be replaced by the diameter of the truncated network and m can be replaced by the number of edges that are incident to visited nodes. Shortest paths can be computed in $O(n^{1+\epsilon}\log W)$ time and $O(m^{1+\epsilon}\log W)$ messages, (W is the maximum edge length assuming integral lengths) where n, and m can be replaced in our setting by the respective quantities for the truncated search.

4.2 Order of announcements

Announcing in increasing rank order makes it possible to charge each message to an item in the MV/D list of the propagating node. We can carry out the announcements in a different order while maintaining at each node a suitable MV/D list for the announcements seen thus far. However in this case deletions can occur from an MV/D list. Deletions occur when a node v appears on the MV/D list of u, but another node v' that is at least as close to u as v and has smaller rank than v makes an announcement later than v. When this happens u has to remove v from its MV/D list and insert v' instead. Since u propagated the rank of v to its neighbors when it received it we can no longer charge each message to an item on the MV/D list of u (that ended up without v). So this charging scheme breaks down if we use a different order. Note that the content of the MV/D list at any point depends only on the set of nodes that made announcements so far (not on their order), and the expected size of the list is logarithmic in the number of announced notes. Hence, the expected amount of storage at each node is logarithmic regardless of the order in which announcements are made.

The following lemma proves that if nodes make their announcements in a *random order* (independent of the ranks and inter-node distances), we get only a logarithmic increase in the number of messages. Random order is relatively easy to achieve in a distributed system. The proof bounds the expected number of insertions to the MV/D list of any node.

LEMMA 4.2. *When nodes announce their ranks in a random order, the expected number of announcements that a node sends is $O(\log^2 n)$. (Note that it is $O(\log^2 n^*)$ in the general setting and $O(\log^2 n^*_d)$ for d-limited MV/D lists).*

PROOF. Nodes make their announcements at some random order that is unrelated to their ranks. We refer to the nodes that did not make an announcement yet as *active* nodes.

Lets look at the process from the viewpoint of some node u. The node at each point maintains a current MV/D list that is correct with respect to all nodes that made their announcement so far.

Consider another node v. We compute the likelihood that v will be on the MV/D list of u when v makes an announcement. Node v can be on the MV/D list of u if and only if no node with smaller rank than v is both closer to u than v and makes an announcement before the node v. Suppose the node v has the kth-smallest rank. The likelihood that i of the k nodes ($i = 0, \ldots, k-1$) with smaller rank are closer than v to u is equal to $1/k$. The likelihood that none of these values make its announcement before v does is $1/(i+1)$. It follows that the likelihood that the kth-smallest-rank node appears on the MV/D list of u is equal to

$$\frac{1}{k}\sum_{j=1}^{k} 1/j = \frac{H_k}{k} \ ,$$

where H_k is the k'th Harmonic number.

Since a node is equally likely to have its rank in any position, the expected number of nodes that at some point appear on the MV/D list of a node u is equal to

$$\sum_{k=1}^{n} \frac{H_k}{k} = \Theta(\log^2 n) \ .$$

To see the last claim observe that

$$\sum_{i=1}^{k} (\log i)/i \leq \sum_{\ell=1}^{\log k} \ell \sum_{j=2^{\ell-1}}^{2^\ell} 1/j \approx \sum_{\ell=1}^{\log k} \ell \approx \log^2 k/2 \ .$$

(Since the inner sum is always a constant – about 1.) □

We next show that an *arbitrary* order of announcements can require much more communication. (Since we can always assume that the announcement order is independent of rank, by arbitrary order we refer to dependence on the location.) Consider a network arranged as a path $v_0, v_1, \ldots, v_{n-1}$. Suppose that the nodes make their announcements in order $v_0, v_1, \ldots$. When the node v_i makes its announcement, it is the closest node that made an announcement thus far to nodes $v_{i+1}, \ldots, v_{n-1}$. Thus (regardless of what the rank of v_i is), the announcement must reach all nodes v_j with $j > i$. Therefore, this order causes $\Omega(n)$ messages to go through most nodes. Observe that even though communication is large, the expected size of the MV/D list at any given point

Announcement order	storage	communication
Ordered	$O(\log n)$	$O(\log n)$
Random	$O(\log n)$	$O(\log^2 n)$
Arbitrary	$O(\log n)$	$O(n)$ or $O(D \log n)$

Table 1: The storage and communication bounds for an MV/D list computation. The communication is the number of announcements reaching a node (Each announcement message consists of distance and rank and is typically of logarithmic size.) The storage is in terms of number of elements (where each element is a distance-rank pair, and possibly item identifier). When nodes contain multiple items with arbitrary values, the n in these bounds should be replaced by n^*. If we restrict our applications to d-limited decay functions, then n and n^* can be replaced by n_d and n_d^*. We do not count here messages that may be required to compute shortest path trees.

in time is logarithmic (the analysis holds since ranks are independent of location).

When announcements are made in arbitrary order but there are only $D \ll n$ distinct distances, we can derive a tighter bound. Clearly there cannot be more than one element of each particular distance at any particular time on a single MV/D list. Since the ranks are independent of the order of announcements, it is easy to see that the expected total number of elements of each particular distance that appear on the list at any point during the computation is $O(\log n)$. The number of different distances is small if the diameter is small or if distances are discretized into a small number of buckets. Table 1 summarizes the storage (in terms of number of elements on the MV/D list) and communication (in terms of number of announcements reaching a node) using different orders of announcements. Experimental results on comparing the different orders of announcements are given in Section 11.

5. RANDOM SAMPLING

We start with the simpler problem of computing, for each node, a summary that would allow it to obtain, for each r, an item selected uniformly at random from its r-neighborhood. (Dependencies between selection for different r and different locations are allowed). This can be performed using a single MV/D lists computation: Each item computes a random rank for itself, and an MV/D list is computed for each node. For each $r \geq 0$, all items in the r-neighborhood are equally likely to have the minimum rank. Thus, the minimum ranked item constitutes a uniform random sample from the r-neighborhood. More generally, a random sample of size k can be obtained either by performing the above k times (using k different MV/D lists), or more efficiently, by computing *k-min MV/D lists* which provide for each r the k-smallest ranked items in the r-neighborhood. It is not hard to modify our 1-min MV/D list algorithms to compute a k-min MV/D list while incurring (a necessary) k-fold increase in storage and communication – an item is placed on the list if and only if there are at most $k-1$ other items which are at least as close and have smaller ranks.

We next address the problem of obtaining samples with respect to an *arbitrary* decay function $g()$, where the probability of an item i to be drawn at location u is proportional to its weight $w_{u,g}(i)$.

Let i_r be a uniform random sample from the r-neighborhood of u. As argued above, (dependent) samples i_r for all $r \geq 0$ can be obtained using a single MV/D lists computation. Let N_r be the number of items in the r-neighborhood of u, and $N \equiv N_{r_{\max}}$ be the number of items in the system. For notational convenience we assume integral distances and define $g(r) \equiv 0$ for $r > r_{\max}$. We now consider a node u and a decay function $g()$.

LEMMA 5.1. *The following process draws the item $i \in I$ with probability $w_{u,g}(i)/\sum_{i' \in I} w_{u,g}(i')$:*

- *Draw $0 \leq r \leq r_{\max}$ randomly according to $(p_0, \ldots, p_{r_{\max}})$, where*
$$p_r = \frac{(g(r) - g(r+1))N_r}{\sum_{j \geq 0}(g(j) - g(j+1))N_j}$$
(Thus, r is selected with probability proportional to $(g(r) - g(r+1))N_r$.)
- *Return the sample i_r.*

PROOF. Note that $\sum_{j \geq 0}(g(j) - g(j+1))N_j = \sum_{i \in I} w_{u,g}(i)$. This is a 2-stage sampling process, where in the first stage an "event" $j \in 0, \ldots, r_{\max}$ is drawn according to the probability distribution $(p_0, \ldots, p_{r_{\max}})$. In the second stage, an item is drawn according to a distribution $q_{j\bullet}$, where q_{ji} is the probability that the item $i \in I$ is selected. This 2-stage sampling process is equivalent to drawing according to the probability distribution where item i is drawn with probability equal to $\sum_{j=0}^{r_{\max}} p_j q_{ij}$. In our setting, each event j corresponds to a distinct distance value r, and the distribution $q_{j\bullet}$ constitutes a uniform distribution over the items present in the j-neighborhood (and nullifies on more distant items). Substituting $q_{ji} = 1/N_j$ for i such that $j \geq \text{DIST}(u, \ell_i)$ and $q_{ji} = 0$ otherwise, we obtain that item i is drawn with probability

$$\begin{aligned}
\sum_{j=0}^{r_{\max}} p_j q_{ij} &= \sum_{j=\text{DIST}(u,\ell_i)}^{r_{\max}} p_j q_{ij} \\
&= \frac{\sum_{j=\text{DIST}(u,\ell_i)}^{r_{\max}} (g(j) - g(j+1))}{\sum_{i' \in I} w_{u,g}(i')} \\
&= \frac{g(\text{DIST}(u, \ell_i))}{\sum_{i' \in I} w_{u,g}(i')} = \frac{w_{u,g}(i)}{\sum_{i' \in I} w_{u,g}(i')}
\end{aligned}$$

□

The nodes use estimates $\hat{N}_r$ instead of the (unavailable) exact values N_r. We can obtain estimates such that the relative error is small with very high probability, that is, $|\hat{N}_r - N_r|/N_r \leq \epsilon$ for $r \geq 0$. When a node uses estimates we have $q_{ji} = 1/N_j$ but $\hat{p}_j = \frac{(g(j)-g(j+1))\hat{N}_j}{\sum_{j' \geq 0}(g(j')-g(j'+1))\hat{N}_j}$. Both the numerator and denominator of $\hat{p}_j$ are within relative error ϵ from the respective quantities for p_j, thus for $\epsilon < 1/3$ we have $|p_j - \hat{p}_j|/p_j \leq 3\epsilon$. Therefore,

$$|\sum_{j=0}^{r_{\max}} p_j q_{ij} - \sum_{j=0}^{r_{\max}} \hat{p}_j q_{ij}| / \sum_{j=0}^{r_{\max}} p_j q_{ij} \leq 3\epsilon \ .$$

Thus, the probability that an item i is drawn is in the range $(1 \pm 3\epsilon)\frac{w_{u,g}(i)}{\sum_{i' \in I} w_{u,g}(i')}$. This approximation is sufficient for most applications of random sampling, including approximate median.

6. APPROXIMATE L_P NORMS (FOR $P \in [1, 2]$)

The data stream version of the problem [11] is as follows: There is an underlying d dimensional vector. Each item consists of a pair (i, a), where $i \in [d]$ is a coordinate and $a \in \{0, \ldots, M\}$ is an increment to the ith dimension of the underlying vector. Every window W represent a vector, defined by applying all the items in the window to the $\mathbf{0}$ vector. The $L_p(W)$ norm of that vector is given by

$$L_p(W) = (\sum_{i \in [d]} |s_i|^p)^{1/p} ,$$

where s_i is the sum of the increment values of all items of the form (i, a) obtained in the last W time units. The goal is to efficiently maintain an approximate value of $L_p(W)$ using storage that is significantly smaller than the dimension d. Datar et al [11] showed that any constant relative error and confidence can be obtained using $O(\log N(\log N + \log M) + \log M \log d)$ storage (where N is the maximum window size we would like to support).

The definition of the problem in the spatial setting was given in Section 2.1. We extend the results to this setting. Observe that for $p = 1$, all coordinates can be aggregated together and the problem reduces to the decaying sum problem. For $p > 1$, clearly, an approximate value of each coordinate of the vector $\mathbf{V}(g, u)$ can be obtained by computing a spatially decaying sum over the update values associated with that coordinate. An approximate value of the L_p norm can be obtained from these estimates. This requires, however, applying the decaying sum computation d times, which imposes a factor d increase in communication. The goal again is to significantly reduce the dependence on d.

We briefly describe Indyk's sketching technique and then explain how to apply it in our setting. The simplified sketching algorithm works as follows. For some fixed value L, which depends on the desired accuracy and confidence, the algorithm uses dL independent random variables X_i^k (where $i \in [d]$ and $k \in \{1, \ldots, L\}$) which are drawn from some distribution that depends on p (the parameter of the norm). Each node has to have access to these numbers. Indyk argued [21] that it can generate them via some shared random seed or on the fly from seeds. The implementation affects storage and computation at each node.

For each $k \in \{1, \ldots, L\}$, each node v then computes a value

$$f(u)_k = \sum_{i | \ell_i = v} a_i X_{c_i}^k .$$

We then perform L decaying sum computations using the values $f(u)_k$ $(k = 1, \ldots, L)$. The estimated norm at each node can then be computed according to Indyk's method (the computation depends on the norm parameter p).

7. DISTINCT ELEMENTS

The problem of distinct elements is to (approximately) count the number of distinct values occurring in neighborhoods of each node. This can be performed using an NH-summary computation based on a variant on MV/D lists. The difference is that each node draws a pseudo-random rank that depends on its value (e.g., by applying a random hash function to its value and mapping the result to be a pseudo random sample from an Exponential distribution). For simplicity of presentation we assume availability of hash functions with perfect random properties, but this assumption can be relaxed (see techniques in [1]). The key point here is that all items with identical values obtain the same rank, but other crucial properties, namely independence of the rank choice for different values and their distribution are preserved. As with random ranks, we can use several MV/D list computations to obtain an NH-summary that can estimate the number of distinct values in each neighborhood.

To obtain a *random distinct value* we can simply take the value that had the minimum hashed rank. To obtain *multiplicity of a random distinct value*, we do as follows. We first discuss small multiplicities. Multiplicities up to M can be obtained exactly with M factor on communication: when computing the MV/D lists, we continue to propagate all items with rank equal to the current minimum rank value. To do so, we may need to record up to M identifiers at each node, and pay an additional factor of M on communication. For high multiplicities, we can reduce the overhead to $O(\log M)$ factor and obtain approximate multiplicities as follows. Consider each item as having a (hashed by value) rank and an additional random rank. For each minimum 1st rank item we consider the second rank and propagate only if smaller. These second ranks allow us to estimate the multiplicity of that value (we need several such 2nd ranks for better accuracy.). Since we don't know the multiplicity in advance, we can apply the two approached simultaneously with a cutoff of M.

8. EUCLIDEAN MV/D LISTS

We thus far focused on edge-length metrics on the underlying network. We now consider the L_2 metric on the Euclidean plane, both when the data is centrally processed and in a distributed setting. MV/D list computation is a basic ingredient in our aggregation algorithms, and simply substituting it with a *Euclidean MV/D list computation* yields algorithms for the Euclidean version of these aggregates.

In the Euclidean MV/D list problem, item locations are points in the Euclidean plane. Items have ranks and each point in the plane has an MV/D list defined according to the Euclidean metric. The goal is for each node to obtain the MV/D list for its location or more generally, to construct a data structure that would allow us to efficiently obtain the MV/D list of any query point.

8.1 Centralized

The centralized version of the problem amounts to computing and performing point location in a particular collection of Voronoi regions (taken from different diagrams): (proof is immediate from definitions)

LEMMA 8.1. *Let items be indexed in increasing rank order $(p_1, \ldots, p_N)$. Let c_i be the Voronoi region of the location of item p_i in the Voronoi diagram defined by the locations of $\{p_1, \ldots, p_i\}$. The MV/D list of a query point consists of all i such that the point resides in the region c_i.*

Moreover, it is not hard to see that the expected size of the MV/D list associated with a point is logarithmic (The arguments used in [5] for edge metrics carry over).

Our centralized Euclidean MV/D list computation can be stated in terms of constructing and querying *incremental Voronoi diagram* under a random order of point insertions. The goal is to find the closest neighbor for a query point not only in the final diagram, but with respect to all prefixes of the insertion order. A paper by Guibas et al [20] constructs an appropriate data structure in $O(n \log n)$ expected time. The expected size of the structure is linear $O(n)$. The expected query time is doubly logarithmic $O(\log^2 n)$ (but is conjectured in [20] to be logarithmic.)

8.2 Distributed

In the distributed setting, data items reside at nodes connected by an underlying network. Generally, however, the Euclidean metric may not be representable as an edge metric on the network (unless the network is a full mesh or very degenerate, eg, when a path network lies on a line). We construct an example that shows that sometimes a Euclidean MV/D list computation necessitates quadratic communication: Consider a U-shaped path network, where the nodes coordinates in the Euclidean plane are $(0, i)$ and $(100, i)$ (for $i = 0, \ldots, N$); the network edges are $((0, 0), (100, 0))$, $((0, i), (0, i+1))$, and $((100, i), (100, i+1))$ (for $0 \leq i < N$). Items are present at nodes $(0, i)$ (for $N \geq i \geq 0$). The closest item to a node $(100, i)$ is the one residing on node $(0, i)$, thus, this particular item is present on the MV/D list of $(100, i)$. It turns out that for any $\ell \geq 0$, all announcements from nodes $(0, i)$ for $i \geq \ell$ must flow through the nodes $(0, \ell)$ and $(100, \ell)$. We thus obtain that during an MV/D list computation, an average $\Omega(N)$ announcement messages must traverse a node.

This bad example exploits the mismatch between the network topology and the Euclidean metric: even though the size of the MV/D list of each node is logarithmic, nodes must pass on information through the network that is not relevant to their own MV/D list. On some networks, in particular on grids, it is possible to efficiently compute Euclidean MV/D lists distributively, with nearly linear total communication. We associate the grid square (*cell*) defined by the points $(i \pm 1/2, j \pm 1/2)$ with each node (i, j). Each node builds a data structure that allows it to produce the MV/D list of any query point in its cell.

Consider the Voronoi regions $\{c_1, \ldots, c_N\}$ as in Lemma 8.1. In order to determine the Euclidean MV/D list for all query points that lie within its grid cell, it is sufficient (and necessary) that a node obtains the list of items i for which the region c_i intersects its cell. We next observe that for each region, the set of grid nodes that their cell intersects the region is connected. This holds because each region is a convex polygon and the grid cell of a node neighbors only the grid cells of the neighbors of the node. We accordingly define an announcement process that outputs the list of all regions that intersect the grid cell of a node: When a node makes an announcement, a receiving node checks if it "updates" the current list of any point in its cell, if it does, it propagates the message to all its neighbors. The analysis of the number of messages is deferred to the full version.

9. LOWER BOUNDS

Consider a problem instance of computing some time-decaying aggregate over N time units where item of value f_i is observed at time $t_i = i$. Also consider a directed path network of size N, where edges have unit lengths and item of value f_i is present at node $\ell_i = i$. Suppose that the same aggregate function and the same decay function apply in both setting (decay over j elapsed time units is equal to decay over distance j). The aggregate value at node i in the spatial-decay setting is equal to the value at time i in the time-decay setting. There is direct correspondence between algorithms for the time-decaying variant to algorithms for the spatial-decay variant on this path network: the size of a summary that is propagated from time i to time $i+1$ in order to make sure we can know the value of the aggregate in later times corresponds to the communication along the edge $(i, i+1)$ (and the storage at node i). Thus, known lower bounds on the space required to maintain time-decaying aggregates together with the reduction described above imply lower bounds on the maximum number of bits sent along an edge when performing the respective spatially-decaying aggregation on path networks.

We list some examples of such lower bounds. Datar et. al. [11] showed that $\Omega(\frac{1}{\epsilon} \log^2 N)$ bits are needed to maintain ($1 \pm \epsilon$ approximate) time-decaying sums over a binary stream where $g()$ is a sliding window of size N. This implies that when computing spatially-decaying sums when $g() \equiv \text{BALL}_N$ at least $\Omega(\frac{1}{\epsilon} \log^2 N)$ bits must traverse some edge. Similarly, a lower bound of $\Omega(\log R)$ bits on per-edge communication for computing (approximate) polynomial and exponential spatially-decaying sums over R items (of value at most polynomial in R) follows from the corresponding lower bounds for polynomial or exponential time-decaying sums that were provided in [8]. Linear ($\Omega(N)$) lower bound on the per-edge communication needed to obtain the minimum value of an item in the N-neighborhood of each node follows from the linear lower bounds on space in the sliding-window model [11].

10. DECAY ACROSS SPACE AND TIME

We extend our model to aggregation over items that arrive in different locations *and at different times.* We use f_i, t_i and $\ell_i \in V$ to denote the value, time, and location of the ith item. There are two different decay functions: g_D captures decay as a function of distance and g_T captures decay as a function of elapsed time. The weight of an item i with $t_i \leq T$, at time T as viewed from location u is $g_D(\text{DIST}(u, \ell_i)) g_T(T - t_i)$. The decaying sum at location u at time T is

$$S_{g_D, g_T}(u) = \sum_{i \in I | t_i \leq T} f_i g_D(\text{DIST}(u, \ell_i)) g_T(T - t_i) \ .$$

We measure performance by both communication *per time unit* and the amount of storage per node, so that at any time and at any location we can obtain approximate value of the respective aggregate.

We allow for a tolerance of Δ time units in the interpretation of time stamps. This is necessary, since otherwise every single item that arrives at a distinct time must be broadcasted to the whole system, since it constitutes the most recent arrival. With Δ tolerance, we can use plain spatial aggregations for all items arriving within Δ time units (as they are treated as having the same time stamp). For efficiency, we would further like the order of announcements within each Δ period to be unrelated to location, which can be achieved by adding random delays to item arrival times.

We next consider storage. Storing a spatial summary for each Δ period can be very consuming. To support queries of decaying sum for all (time and distance) decay functions, we should be able to obtain, for each t' and d', approximate sum of values of all items that occurred within the last t' Δ periods and within distance d'. Unfortunately, the size of such a "2-d NH-summary" can be linear: Consider a path network with nodes $v_0, v_1, \ldots, v_{n-1}$ and items (of value either 0 or 1) $(v_0, t_0), (v_1, t_1), \ldots$, where $t_i < t_{i+1}$ (items that are further from v_0 occur later in time). Consider the information that needs to be stored by v_0 at some time T (that is later than all item arrival times). Each item has a unique time-distance neighborhood in which it lies: the item (v_i, t_i) is the only item within the $(T - t_i)$-recent i-neighborhood of v_0. Thus, in order to be able to obtain estimates of $\text{BALL}_{r,t}$ for all possible values of r and t, we may need storage that is linear in (the smaller of) the number of distinct distances from our location and distinct time units.

Storage requirements can be reduced with some restrictions: Observe that for each set of items occurring at the same distance (possibly at different times), we can maintain a single window summary. Thus, if distances are discretized into D meaningful levels, it suffices to maintain D window summaries. A second observation is that if we restrict ourselves to a fixed spatial decay function at each node then it suffices that each node maintains a single window summary. The information incorporated in the summary in each time unit t is a single numeric value: an approximate spatially decaying sum

$$S_{g_D,t}(u) = \sum_{i|t_i=t} f_i g_D(\text{DIST}(u, \ell_i)) \ ,$$

obtained using the NH-summary (MV/D lists) compiled for the current Δ period.

11. EXPERIMENTAL EVALUATION

For the simulation part of our evaluation we used terrestrial air temperature and precipitation data. The data was provided for grid points in half-degree spacings on the terrestrial grid. For each grid point it included the mean annual temperature and precipitation (over the time period 1950 to 1999) [4]. We treated the grid as a network. Overall, our grid included 85,794 points.

We counted the number of messages per node (with each message containing a rank and the location of the announcing node) for computing a single (∞-limited) MV/D list and the size of this list, for the three orders of announcements considered in Section 4: random order of announcements, arbitrary order (where the announcements follow a lexicographic order on the longitude, latitude coordinates), and optimal order (increasing rank). We looked at the size of the lists and number of messages for precipitation, temperature, and size values. Recall that although the communication depends on the announcement order, the size and content of the MV/D lists depend only on the ranks.

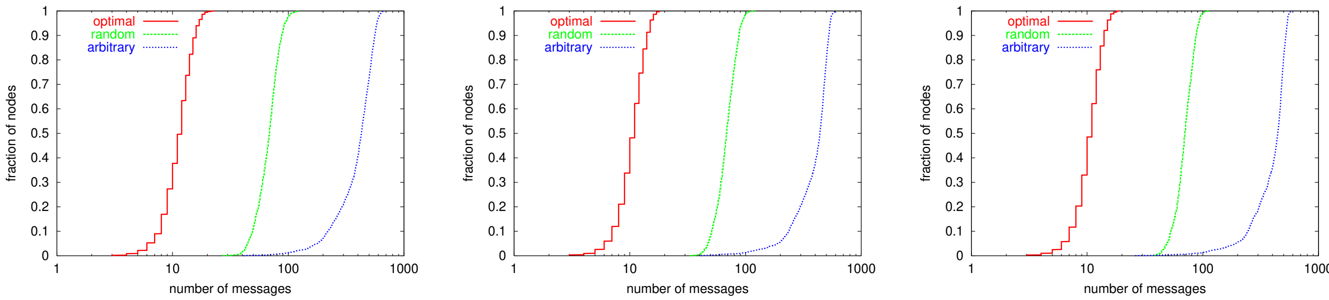
Table 2 lists the average number of messages per node under each order. Figure 2 shows the cumulative distributions. On this 86K node network, random order required about 6-7 times more messages than optimal and arbitrary order about 40 times more messages than optimal. This suggests that a combined random-minimum approach that biases earlier announcements to lower ranks can improve on the performance of random order. Theory provides precise expressions for the distribution of the size of the MV/D lists and for the accuracy of estimates using a certain number of MV/D lists when items' values are binary (this is the case when counting or computing sizes [5, 6]). For arbitrary values we provided in Section 4 a worst-case bound in terms of n^*, which depends on the ratio of the weights of the largest to smallest values. Interestingly, even though there is large variance in precipitation values, the size of the corresponding MV/D lists was only slightly larger than with binary values (11.6 versus 10.7).

	optimal	random	arbitrary
precipitation	11.6	69	409
temperature	10.7	69	404
size	10.7	71	408

Table 2: Average number of messages to obtain an MV/D list under optimal, arbitrary, and random order of announcements. For precipitation levels, temperature, and number of sensors.

We next compare our methods that produces estimates based on MV/D lists to the naive method, which produces exact values. The naive algorithm sends every data point to all nodes that it influences. Thus, the communication and storage at each node are proportional to the number of data points that influence it. The influence sets depends on the decay function. For general decay functions it involves 85793 data points from all peer nodes. For d-limited decay functions, the influence set contains only items in the d-neighborhood. Our method produces estimates with quality that increases with the number k of MV/D lists. The expected size of each list is logarithmic in the influence set. The NH-summary computed from the k lists has size which is at most the sum of sizes of the k lists. It can provide estimated aggregate values for *any* decay function. The communication depends on the order of announcements, and as shown in Section 3, is logarithmic for optimal order and log-squared for random order. Thus, asymptotically there is an exponential gap in communication and storage between the naive exact method and our MV/D list based approximation.

We drew ranks from the Exponential distribution with parameter λ, where λ was the value (for sizes values were binary; for temperatures we added a value of 60 to obtain nonnegative values). The NH-summary contained a list of distance intervals, and for each distance, we used the unbiased estimate $(k-1)/\sum_{i=1}^{k} r_i$, where r_i was the minimum rank according to MV/D list i for the corresponding distance. Figure 3 shows the distribution of the ratio of the estimate value to the exact aggregate values. The approximate aggregate values were produced using an NH-summary that constituted of 10, 20, 50, or 100 MV/D lists. The figure shows, in agreement with analysis, that the error decreases with the number of lists, and that its distribution is the same for different BALL_r functions (it does not depend on neighborhood size). The aggregate for the inverse-squared decay function $1/(1 + \text{DIST})^2$ was calculated by applying Lemma 3.1 to the NH-summary. This estimate is more accurate than over neighborhoods since it combines estimates over different neighborhoods (estimation errors for different neighborhood are only partially dependent). Figure 4 shows the number of messages per node for d-limited aggregation as a function of d. For the naive method and for the approximate method using $k = 1, 20, 100$ lists. The communication required with the naive exact computation grows much faster than with the approximate methods. Thus, exact computation is favorable for smaller d or when the number of MV/D lists needed to reach the precision that an



precipitation | temperature | size

Figure 2: Cumulative number of messages per node needed for computing an MV/D list for precipitation, size, and temperature.

application requires makes the communication cost comparable to that of the exact computation. The average number of influencing nodes (and per-node average communication cost for the naive exact computation) is 56.1 for 2.5 degree neighborhood, 192.3 for 5 degree neighborhood, and 662.9 for 10 degree neighborhood. For general decay functions (such as inverse-square of the distance), the influence set of each point contains all other (85793) nodes. The figure shows that when using 20 or fewer MV/D lists, d which is at least 10 degrees, and random order of announcements, the approximate method is preferable to the naive method; with optimal order of announcements, the estimation method is preferable even with d starting at 5-degrees. For smooth decay functions, the estimation method even with hundreds of MV/D lists is considerably more efficient than the naive method.

12. CONCLUSION

We introduced a model for spatially decaying aggregation, which is motivated by emerging applications including P2P and sensor networks where data is associated with its location and its relevance decays with distance. We developed basic techniques and efficient algorithms for some fundamental aggregate functions.

13. REFERENCES

[1] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *J. Comput. System Sci.*, 58:137–147, 1999.

[2] B. Awerbuch. Distributed shortest paths algorithms. In *Proc. 21st Annual ACM Symposium on Theory of Computing*, pages 490–500. ACM, 1989.

[3] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom. Models and issues in data stream systems. In *Proc. of the 2002 ACM Symp. on Principles of Database Systems (PODS 2002)*. ACM, 2002.

[4] Terrestrial air temperature and precipitation: monthly and annual climatologies (version 3.02), 2001. `http://climate.udel.edu/~climate`.

[5] E. Cohen. Size-estimation framework with applications to transitive closure and reachability. *J. Comput. System Sci.*, 55:441–453, 1997.

[6] E. Cohen. Structure prediction and computation of sparse matrix products. *J. Combinatorial Optimization*, 2:307–332, 1999.

[7] E. Cohen and H. Kaplan. Efficient estimation algorithms for neighborhood variance and other moments. In *Proc. 15th ACM-SIAM Symposium on Discrete Algorithms*. ACM-SIAM, 2004.

[8] E. Cohen and M. Strauss. Maintaining time-decaying stream aggregates. In *Proc. of the 2003 ACM Symp. on Principles of Database Systems (PODS 2003)*. ACM, 2003.

[9] Open Source Community. Gnutella. In *http://gnutella.wego.com/*, 2001.

[10] A. Crespo and H. Garcia-Molina. Routing indices for peer-to-peer systems. In *ICDE*, 2002.

[11] M. Datar, A. Gionis, P. Indyk, and R. Motwani. Maintaining stream statistics over sliding windows. *SIAM J. Comput.*, 31(6):1794–1813, 2002.

[12] A. Deshpande, P. Gibbons, S. Nath, and S. Seshan. Cache-and-query for wide area sensor databases. In *Proceedings of the ACM SIGMOD*, 2003.

[13] D. Estrin, L. Girod, G. Pottie, and M. Srivastava. Instrumenting the world with wireless sensor networks. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001.

[14] J. Feigenbaum, S. Kannan, M. Strauss, and M. Viswanathan. An approximate L1-difference algorithm for massive data streams. In *Proc. 39th IEEE Annual Symposium on Foundations of Computer Science*, pages 501–511. IEEE, 1999.

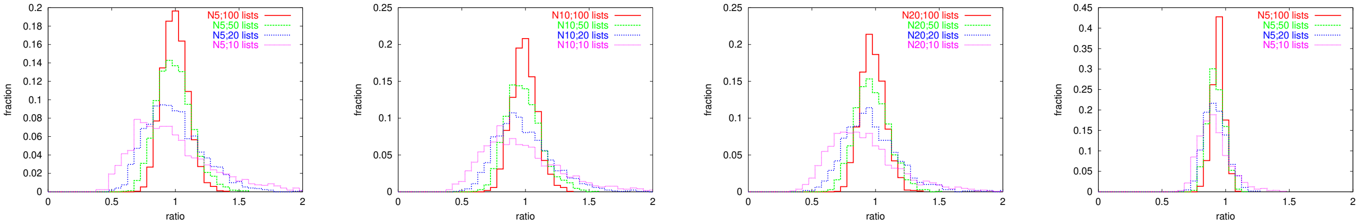
[15] S. Floyd and V. Jacobson. Random early detection gateways for congestion avoidance. *IEEE/ACM Transaction on Networking*, 1(4), 1993.

[16] S. Ganeriwal, Han C. C., and M. B. Srivastava. Spatial average of a continous physical process in sensor networks (poster). In *Proceedings of the 1st ACM conference on embedded networked sensor systems (Sensys)*, 2003.

[17] D. Ganesan, S. Ratnasamy, H. Wang, and D. Estrin. Coping with irregular spatio-temporal sampling in sensor netowrks. In *Proceedings of the ACM SIGCOMM HOTNETS-II workshop*, 2003.

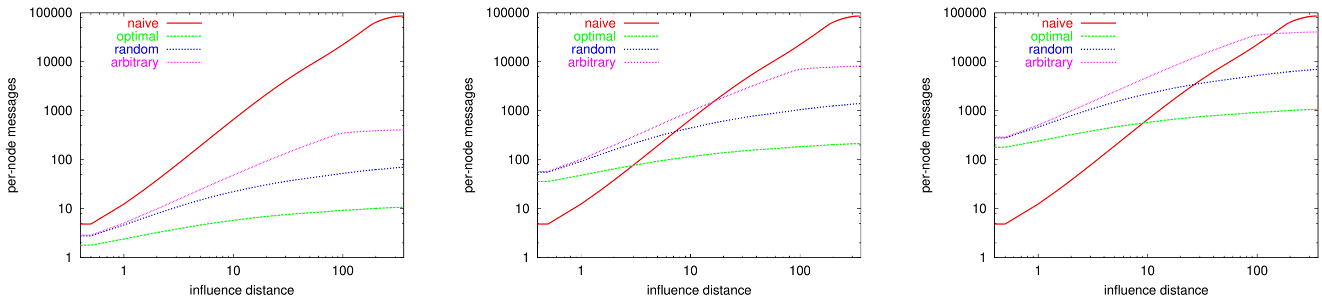
[18] P. B. Gibbons and S. Tirthapura. Distributed streams algorithms for sliding windows. In *Proc. of the 14th Annual ACM Symposium on Parallel Algorithms and Architectures*, pages 63–72. ACM, 2002.

[19] M. Greenwald and S. Khanna. Space-efficient online computation of quantile summaries. In *Proceedings of*



2.5 degrees | 5 degrees | 10 degrees | $1/(1 + \text{DIST})^2$

Figure 3: Histogram of the distribution of the ratio of the estimate we obtain to the exact quantity for precipitation levels for 2.5, 5, and 10 degree neighborhoods and for $1/(1 + \text{DIST})^2$ decay.



1 list | 20 lists | 100 lists

Figure 4: Number of messages as a function of the distance of influence under each method: the exact naive method and NH-summaries generated using 1, 20, or 100 MV/D lists generated with optimal, random, or arbitrary order of announcements.

the ACM SIGMOD, 2001.

[20] L. J. Guibas, D. E. Knuth, and M. Sharir. Randomized incremental construction of Delaunay and Voronoi diagrams. *Algorithmica*, 7:381–413, 1992.

[21] P. Indyk. Stable distributions, pseudorandom generators, embeddings and data stream computation. In *Proc. 41st IEEE Annual Symposium on Foundations of Computer Science*, pages 189–197. IEEE, 2001.

[22] V. Jacobson. Congestion avoidance and control. In *Proceedings of the ACM SIGCOMM'88 Conference*, August 1988.

[23] J. Kleinberg and D. Kempe. Protocols and impossibility results for gossip-based communication mechanisms. In *Proc. 43rd IEEE Annual Symposium on Foundations of Computer Science*, pages 471–480. IEEE, 2002.

[24] J. Kleinberg, D. Kempe, and A. J. Demers. Spatial gossip and resource location protocols. In *Proc. 33rd Annual ACM Symposium on Theory of Computing*, pages 163–172. ACM, 2001.

[25] F. Korn and S. Muthukrishnana. Influence sets based on reverse nearesr neighbors. In *Proc. of the 1000 ACM SIGMOD conference on Management of Data*, pages 201–212. ACM, 2000.

[26] F. Korn, S. Muthukrishnana, and D. Srivastava. Reverse nearest neighbor aggregates over data streams. In *Proc. of the 28th international conference on Very Large Databases (VLDB 2002)*. ACM, 2002.

[27] X. Lin, H. Lu, J. Xu, and J. X. Yu. Continously maintaining quantile sumaries of the most recent n elements over a data stream. In *Proceedings of the 20th International conference on data engineering (ICDE)*. IEEE, 2004.

[28] S. Madden, M. Franklin, J. Hellerstein, and W. Hong. Tag: a tiny aggregation service for ad-hoc sensor networks. In *Proceedings of the 5th USENIX Symposium on Operating Systems Design and Implementation (ODSI)*. USENIX Association, 2002.

[29] A. Mainwaring, J. Polastre, R Szewczyk, D. Culler, and J. Anderson. Wireless sensor networks for habitat monitoring. In *Proceedings of the 1st ACM International workshop on wireless sensor networks and applications*, 2002.

[30] G. S. Manku and R. Motwani. Approximate frequency counts over data streams. In *Proc. of the 28th international conference on Very Large Databases (VLDB 2002)*, pages 346–357. ACM, 2002.

[31] G. S. Manku, S. Rajagopalan, and B. G. Lindsay. Approximate medians and other quantiles in one pass and with limited memory. In *Proc. of the 1998 ACM SIGMOD conference on Management of Data*, pages 426–435. ACM, 1998.

[32] M. R. Paterson. Progress in selection. Dept. of Computer Science, University of Warwick, Coventry, UK, 1997.

[33] J. Zhao, R. Govindan, and D. Estrin. Computing aggregates for monitoring wireless sensor networks. In *Proc. of First IEEE international Workshop on Sensor Network Protocols and Applications*, 2003.