Demand-aware network designs of bounded degree

Avin, Chen; Mondal, Kaushik; Schmid, Stefan

Published in:
31st International Symposium on Distributed Computing, DISC 2017

DOI (link to publication from Publisher):
10.4230/LIPIcs.DISC.2017.5

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Demand-Aware Network Designs of Bounded Degree*

Chen Avin[1], Kaushik Mondal[2], and Stefan Schmid[3]

1 **Communication Systems Engineering Department, Ben Gurion University of the Negev, Be'er Scheva, Israel**
avin@cse.bgu.ac.il

2 **Communication Systems Engineering Department, Ben Gurion University of the Negev, Be'er Scheva, Israel**
mondal@post.bgu.ac.il

3 **Department of Computer Science, Aalborg University, Denmark**
schmiste@cs.aau.dk

Abstract

Traditionally, networks such as datacenter interconnects are designed to optimize worst-case performance under *arbitrary* traffic patterns. Such network designs can however be far from optimal when considering the *actual* workloads and traffic patterns which they serve. This insight led to the development of demand-aware datacenter interconnects which can be reconfigured depending on the workload.

Motivated by these trends, this paper initiates the algorithmic study of demand-aware networks (DANs), and in particular the design of bounded-degree networks. The inputs to the network design problem are a discrete communication request distribution, $\mathcal{D}$, defined over communicating pairs from the node set V, and a bound, Δ, on the maximum degree. In turn, our objective is to design an (undirected) demand-aware network $N = (V, E)$ of bounded-degree Δ, which provides short routing paths between frequently communicating nodes distributed across N. In particular, the designed network should minimize the *expected path length* on N (with respect to $\mathcal{D}$), which is a basic measure of the efficiency of the network.

We show that this fundamental network design problem exhibits interesting connections to several classic combinatorial problems and to information theory. We derive a general lower bound based on the entropy of the communication pattern $\mathcal{D}$, and present asymptotically optimal network-aware design algorithms for important distribution families, such as sparse distributions and distributions of locally bounded doubling dimensions.

1998 ACM Subject Classification F.2 Analysis of Algorithms and Problem Complexity, C.2.1 Network Architecture and Design

Keywords and phrases Network design, reconfigurable networks, datacenter topology, peer-to-peer computing, entropy, sparse spanners

Digital Object Identifier 10.4230/LIPIcs.DISC.2017.5

1 Introduction

The problem studied in this paper is motivated by the advent of more flexible datacenter interconnects, such as ProjecToR [16, 17]. These interconnects aim to overcome a fundamental

* This work was supported by the German-Israeli Foundation for Scientific Research (GIF) Grant I-1245-407.6/2014

drawback of traditional datacenter network designs: the fact that network designers must decide *in advance* on how much capacity to provision between electrical packet switches, e.g., between Top-of-Rack (ToR) switches in datacenters. This leads to an undesirable tradeoff [25]: either capacity is over-provisioned and therefore the interconnect expensive (e.g., a fat-tree provides full-bisection bandwidth), or one may risk congestion, resulting in a poor cloud application performance. Accordingly, systems such as ProjecToR provide a reconfigurable interconnect, allowing to establish links flexibly and in a *demand-aware manner*. For example, direct links or at least short communication paths can be established between frequently communicating ToR switches. Such links can be implemented using a bounded number of lasers, mirrors, and photodetectors per node [17]. First experiments with this technology demonstrated promising results: although the interconnecting networks is of bounded degree, short routing paths can be provided between communicating nodes.

The problem of designing demand-aware networks is a fundamental one, and finds interesting applications in many distributed and networked systems. For example, while many peer-to-peer overlay networks today are designed towards optimizing the *worst-case performance* (e.g., minimal diameter and/or degree), it is an intriguing question whether the "hard instances" actually show up in real life, and whether better topologies can be designed if we are given more information about the actual communication patterns these networks serve in practice.

While the problem is natural, surprisingly little is known today about the design of demand-aware networks. At the same time, as we will show in this paper, the design of demand-aware networks is related to several classic combinatorial problems.

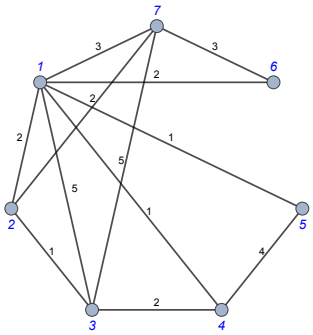
Our vision is reminiscent in spirit to the question posed by Sleator and Tarjan over 30 years ago in the context of binary search trees [10, 26]: While there is an inherent lower bound of $\Omega(\log n)$ for accessing an arbitrary element in a binary search tree, can we do better on some "easier" instances? The authors identified the *entropy* to be a natural metric to measure the performance under actual demand patterns. We will provide evidence in this paper that the entropy, in a slightly different flavor, also plays a crucial role in the context of network designs, establishing an interesting connection.

The Problem: Bounded Network Design. We consider the following network design problem, henceforth referred to as the *Bounded Network Design* problem, short *BND*. We consider a set of n nodes (e.g., top-of-rack switches, servers, peers) $V = \{1, \ldots, n\}$ interacting according to a certain *communication pattern*. The pattern is modelled by $\mathcal{D}$, a discrete distribution over *communication requests* defined over $V \times V$. We represent this distribution using a communication matrix $M_{\mathcal{D}}[\mathrm{p}(i,j)]_{n \times n}$ where the (i,j) entry indicates the communication frequency, $p(i,j)$, from the (communication) source i to the (communication) destination j. The matrix is normalized, i.e., $\sum_{ij} \mathrm{p}(i,j) = 1$. Moreover, we can interpret the distribution $\mathcal{D}$ as a weighted directed *demand graph* $G_{\mathcal{D}}$, defined over the same set of nodes V: A directed edge $(u,v) \in E(G_{\mathcal{D}})$ exists iff $\mathrm{p}(u,v) > 0$. We set the edge weight to the communication probability: $w(i,j) = \mathrm{p}(i,j)$.

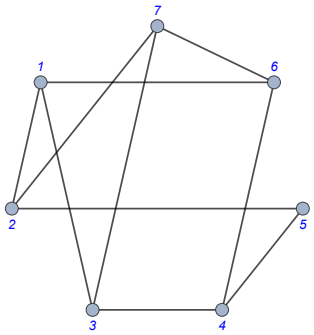
In turn, our objective is to design an unweighted, undirected *Demand-Aware Network* (`DAN`) defined over the set of nodes V and the distribution $\mathcal{D}$, henceforth denoted as $N(\mathcal{D})$ or just N when $\mathcal{D}$ is clear from the context. The objective is that $N(\mathcal{D})$ optimally serves the communication requests from $\mathcal{D}$ under the constraint that N must be chosen from a certain family of *desired topologies* $\mathcal{N}$. In particular, we are interested in *sparse* networks (i.e., having a *linear number* of edges) with *bounded* degree Δ (i.e., nodes have a small number of lasers [17]), and we denote the family of Δ-bounded degree graphs by $\mathcal{N}_\Delta$.

	1	2	3	4	5	6	7
1	0	$\frac{2}{65}$	$\frac{1}{13}$	$\frac{1}{65}$	$\frac{1}{65}$	$\frac{2}{65}$	$\frac{3}{65}$
2	$\frac{2}{65}$	0	$\frac{1}{65}$	0	0	0	$\frac{2}{65}$
3	$\frac{1}{13}$	$\frac{1}{65}$	0	$\frac{2}{65}$	0	0	$\frac{1}{13}$
4	$\frac{1}{65}$	0	$\frac{2}{65}$	0	$\frac{4}{65}$	0	0
5	$\frac{1}{65}$	0	$\frac{3}{65}$	$\frac{4}{65}$	0	0	0
6	$\frac{2}{65}$	0	0	0	0	0	$\frac{3}{65}$
7	$\frac{3}{65}$	$\frac{2}{65}$	$\frac{1}{13}$	0	0	$\frac{3}{65}$	0

(a)



(b)



(c)

Figure 1 Example of the *bounded network design* problem. (a) A given demand distribution $\mathcal{D}$ (which in this case is *symmetric*). (b) The demand graph $G_{\mathcal{D}}$ (with non-normalized weights). Nodes 1, 3, and 7 have a degree more than 3. (c) An optimal solution DAN N with $\Delta = 3$. In this case, the solution is not a subgraph but contains auxiliary edges (e.g., $\{2, 5\}$), and $\mathrm{EPL}(\mathcal{D}, N) = 1.19$ while $H(X \mid Y) = 1.08$ (the Shannon entropy to the base 3 is $H(X) = 1.68$).

Note that the designed network can be seen as "hosting" the served communication pattern, i.e., the demand graph is embedded on the designed network. Accordingly, we will sometimes refer to the demand graph as the *guest network* and to the designed network as the *host network*.

Our objective is to minimize the *expected path length* [1, 2, 24] of the designed host network $N \in \mathcal{N}$: For $u, v \in V(N)$, let $\mathrm{d}_N(u, v)$ denote the shortest path between u and v in N. Given a distribution $\mathcal{D}$ over $V \times V$ and a graph N over V, the *Expected Path Length* (EPL) of route requests is defined as:

$$\mathrm{EPL}(\mathcal{D}, N) = \mathbb{E}_{\mathcal{D}}[\mathrm{d}_N(\cdot, \cdot)] = \sum_{(u,v)\in\mathcal{D}} \mathrm{p}(u, v) \cdot \mathrm{d}_N(u, v)$$

Since routing across the host network usually occurs along shortest paths, the expected path length captures the average hop-count of a route (e.g., latency incurred or energy consumed along the way).

Succinctly, the Bounded Network Design (BND) problem is to minimize the expected path length and is defined as follows:

▶ **Definition 1** (Bounded Network Design)**.** Given a communication distribution, $\mathcal{D}$ and a maximum degree Δ, find a host graph $N \in \mathcal{N}_\Delta$ that minimizes the expected path length:

$$\mathrm{BND}(\mathcal{D}, \Delta) = \min_{N\in\mathcal{N}_\Delta} \mathrm{EPL}(\mathcal{D}, N)$$

See Figure 1 for an example of these definitions.

Our Contributions. This paper initiates the study of a fundamental problem: the design of demand-aware communication networks. While our work is motivated by recent trends in datacenter network designs, our model is natural and finds applications in many distributed and networked systems (e.g., peer-to-peer overlays). The main contribution of this paper is to establish an interesting connection of the network design problem to the conditional entropy of the communication matrix. In particular, we present a lower bound on the expected path length of a network with maximum degree Δ which is proportional to the conditional entropy

of $\mathcal{D}$, $H_\Delta(X \mid Y) + H_\Delta(Y \mid X)$ where Δ is the base of the logarithm used for calculating the entropy. While this lower bound can be as high as $\log n$, for many distributions it can be much lower (even constant). Our main results are presented in Theorem 7 which proves a matching upper bound for the case when $\mathcal{D}$ is a sparse distribution. It is important to note the real datacenters traffic shows evidence that the demand distributions are indeed sparse [23, 17]. Additionally Theorem 12 proves a matching upper bound for the case when $\mathcal{D}$ is a regular and uniform (but maybe dense) distribution of a locally bounded doubling dimension. Also in these two cases the conditional entropy could range from a constant and up to $\log n$. At the heart of our technical contribution is a novel technique to transform a low-distortion network of maximum degree Δ to a low-degree network whose maximum degree equals the average degree of the original network, while maintaining an expected path length in the order of the conditional entropy. Moreover, we show an interesting reduction of uniform and regular distributions to graph spanners in Theorem 8.

Paper Organization. The remainder of this paper is organized as follows. We put our work into perspective with respect to related work in Section 2 and introduce some preliminaries in Section 3. We derive lower bounds in Section 4 and present algorithms to design networks for sparse distributions resp. regular and uniform distributions in Section 5 resp. Section 6. We conclude our work and outline directions for future research in Section 7. Due to space constraints, some details are omitted in this paper, and we refer the reader to our accompanying technical report [3].

2 Putting Things Into Perspective and Related Work

There are at least three interesting perspectives on our problem. The first one arises when trying to gain some intuition about the problem complexity. If $\Delta = n$, the problem is simple: the demand (or guest) graph $G_\mathcal{D}$ itself can be used as the host graph or `DAN` $N \in \mathcal{N}_\Delta$, providing an ideal expected path length 1. If a sparse host graph is desired, a star topology could be used as a `DAN` to provide an expected path length of at most 2. At the other end of the spectrum, if $\Delta = 2$ (and the host network is required to be connected) the `DAN` N must be a line or a ring graph. However, the problem of how to arrange nodes on the linear chain or the ring such that the expected path length is minimized, is already NP-hard: the problem is essentially a Minimum Linear Arrangment (MinLA) problem [7, 11, 15]. One perspective to see our contribution is that in this paper, we are interested in what happens between these extremes, for other values of Δ, in particular for a constant Δ which guarantees that our host network will be sparse, i.e., has a linear number of edges. In contrast to the general arrangement problem which asks for an embedding of the guest graph on a *specific* and *given* host graph, in our network design problem we are free to *choose the best* host graph from a given family of graphs (i.e., bounded degree graphs). One might wonder: does this flexibility make the problem easier? Existing works on low maximum resp. low average degree networks, e.g., in the context of publish/subscribe overlays [8, 20, 21], do not provide formal performance guarantees.

Sparse and distance-preserving spanners open a second perspective on our work: intuitively, a good host graph N for $G_\mathcal{D}$ "looks similar" to $G_\mathcal{D}$. But in contrast to classic spanner problems in the literature which are primarily concerned with minimizing the worst-case *distortion* (resp. the average distortion) among *all* node pairs [4, 22], we are only interested in the *local distortion*. Namely, we aim to find a good "spanner" which preserves *locality of neighborhoods*, i.e., 1-hop neighborhoods in the demand graph. Second, unlike classic spanner

problems but similar to geometric (metric) spanners, the designed network N does not have to be a subgraph and may include edges which do not exist in the demand network $G_{\mathcal{D}}$, i.e., 0-entries in the corresponding communication matrix $M_{\mathcal{D}}$. We refer the corresponding edges as *auxiliary edges* (a.k.a. shortcut edges [19]). It is easy to see that auxiliary edges can indeed be required to compute optimal network designs, and yield strictly lower communication costs than subgraph spanners (e.g. Figure 1). Third, in contrast to the frequently studied sparse graph spanner problem variants, we require that nodes in the designed network are of degree at most Δ. Finally, we are not aware of any work studying the relationship between spanners and entropy. This makes our model fundamentally different from existing models studied in the literature.

The fact that our matrix represents a distribution provides some interesting structure. In particular, it leads us to a third connection, namely to information and coding theory. It is known that the expected path length in binary search trees [26] as well as in network designs providing local routing [2, 24] is upper bounded by the entropy $H(X)$ (over the (empirical) distribution of accessed elements X in the data structure). The conditional entropy of the distribution, $H(X|Y) + H(Y|X)$, is a lower bound on the expected path length of local routing tree designs [24] where X, Y are the random variables distributed according to the marginal distribution of the sources and destinations in $\mathcal{D}$. This bound is tight for the limited case where D is a product distribution (i.e., $p(i,j) = p(i)p(j)$). Additionally the optimal binary search tree can be computed efficiently for every $\mathcal{D}$ using dynamic programming [24]. In the current work we extend this line of research by studying more general distributions and a larger family of host networks.

3 Preliminaries

We start with some notation about $\mathcal{D}$. Let $\mathcal{D}[i,j]$ or $p(i,j)$ denote the probability that source i routes to destination j. Let $p(i)$ denote the probability that i is a source, i.e., $p(i) = \sum_j p(i,j)$. Similarly let $q(j)$ denote the probability that j is a destination. Let X, Y be random variables describing the marginal distribution of the sources and destinations in $\mathcal{D}$, respectively. Let $\overrightarrow{\mathcal{D}}[i]$ denote the *normalized* i'th row of $\mathcal{D}$, that is, the probability distribution of destinations given that the source is i. Similarly let $\overleftarrow{\mathcal{D}}[j]$ denote the normalized j'th column of $\mathcal{D}$, that is the probability distribution of sources given that the destination is j. Let Y_i and X_j be random variables that are distributed according to $\overrightarrow{\mathcal{D}}[i]$ and $\overleftarrow{\mathcal{D}}[j]$, respectively. We say that $\mathcal{D}$ is *regular* if $G_{\mathcal{D}}$ is a regular graph (both in terms of in and out degrees). We say that $\mathcal{D}$ is *uniform* if for every $\mathcal{D}[i,j] > 0$, $\mathcal{D}[i,j] = \frac{1}{m}$ and m is the number of edges in $G_{\mathcal{D}}$. We say that $\mathcal{D}$ is *symmetric* if $\mathcal{D}[i,j] = \mathcal{D}[j,i]$.

We will show that a natural measure to assess the quality of a designed network relates to the *entropy* of the communication pattern. For a discrete random variable X with possible values $\{x_1, \ldots, x_n\}$, the entropy $H(X)$ of X is defined as

$$H(X) = \sum_{i=1}^{n} p(x_i) \log_2 \frac{1}{p(x_i)} \tag{1}$$

where $p(x_i)$ is the probability that X takes the value x_i. Note that, $0 \cdot \log_2 \frac{1}{0}$ is considered as 0. If $\bar{p}$ is a discrete distribution vector (i..e, $p_i \geq 0$ and $\sum_i p_i = 1$) then we may write $H(\bar{p})$ or $H(p_1, p_2, \ldots p_n)$ to denote the entropy of a random variable that is distributed according to $\bar{p}$. If $\bar{p}$ is the uniform distribution with support s (s being the number of places in the distribution with $p_i > 0$, i.e., $p_i = 1/s$) then $H(\bar{p}) = \log s$.

Using the decomposition (a.k.a. grouping) properties of entropy the following are well-known [9]:

$$H(p_1, p_2, p_3 \dots p_m) \geq H(p_1 + p_2, p_3 \dots p_m) \tag{2}$$

$$H(p_1, p_2, p_3 \dots p_m) \geq (1 - p_1) H(\frac{p_2}{1-p_1}, \frac{p_3}{1-p_1} \dots \frac{p_m}{1-p_1}) \tag{3}$$

For a joint distribution over X, Y, the *joint entropy* is defined as

$$H(X,Y) = \sum_{i,j} p(x_i, y_j) \log_2 \frac{1}{p(x_i, y_j)} \tag{4}$$

Also recall the definition of the *conditional entropy* $H(X|Y)$:

$$\begin{aligned} H(X|Y) &= \sum_{i,j} p(x_i, y_j) \log_2 \frac{1}{p(x_i \mid y_j)} = \sum_j p(y_j) \sum_i p(x_i \mid y_j) \log_2 \frac{1}{p(x_i \mid y_j)} \\ &= \sum_{j=1}^{n} p(y_j) H(X|Y = y_j) \end{aligned} \tag{5}$$

For X and Y defines as above from $\mathcal{D}$ we also have

$$H(X|Y) = \sum_{j=1}^{n} p(y_j) H(X|Y = y_j) = \sum_{j=1}^{n} q(j) H(\overleftarrow{\mathcal{D}}[j]) = \sum_{j=1}^{n} q(j) H(X_j) \tag{6}$$

$H(Y|X)$ is defined similarly and we note that it may be the case that $H(X|Y) \neq H(Y|X)$. We may simply write H for the entropy if the purpose is given by the context. By default, we will denote by H the entropy computed using the binary logarithm; if a different logarithmic basis Δ is used to compute the entropy, we will explicitly write H_Δ.

We recall the definition of a graph *spanner*. Given a graph $G = (V, E)$, a subgraph $G' = (V, E')$ is a t-spanner of G if for every $u, v \in V$, $t \cdot d_G(u, v) \geq d_{G'}(u, v)$ and t is the *distortion* of the spanner. We say that $G' = (V, E')$ is a *graph metric* t-spanner if it is not a subgraph of G, i.e., it may have additional edges that are not in G.

4 A Lower Bound

We now establish an interesting connection to information theory and show that the conditional entropy serves as a natural metric for bounded network designs. In particular, we prove that the expected path length $\mathrm{BND}(\mathcal{D}, \Delta)$ in any demand-aware bounded network design, is at least in the order of the conditional entropy. Formally:

▶ **Theorem 2.** *Consider the joint frequency distributions $\mathcal{D}$. Let X, Y be the random variables distributed according to the marginal distribution of the sources and destinations in $\mathcal{D}$, respectively. Then*

$$\mathrm{BND}(\mathcal{D}, \Delta) \geq \Omega(\max(H_\Delta(Y|X), H_\Delta(X|Y))$$

Before delving into the proof, let $\mathrm{EPL}(\bar{p}, T)$ denote the expected path length in a tree T from the root to its nodes where the expectation it taking over a distribution $\bar{p}$. That is $\mathrm{EPL}(\bar{p}, T) = \sum_i p_i d_T(root, i)$. We recall the following well-known theorem:

▶ **Theorem 3** ([18], restated.). *Let $H(\bar{p})$ be the entropy of the frequency distribution $\bar{p} = (p_1, p_2, \dots, p_n)$. Let T be an optimal binary search tree built for the above frequency distribution. Then* $\mathrm{EPL}(\bar{p}, T) \geq \frac{1}{\log 3} H(\bar{p})$.

Namely, the entropy $H(\bar{p})$, is a lower bound on the expected path length from the root to the nodes in the tree. Note that, the proof of Theorem 3 in [18] holds for any optimal binary tree T, not necessarily a search tree. For higher degree graphs, we can extend the result:

▶ **Lemma 4.** *Let $H_\Delta(\bar{p})$ be the entropy (calculated using the logarithm of base Δ) of frequency distribution $\bar{p} = (p_1, p_2, \ldots, p_n)$. Let T be an optimal Δ-ary tree built for the above frequency distribution. Then,* $\mathrm{EPL}(\bar{p}, T) \geq \frac{1}{\log(\Delta+1)} H_\Delta(\bar{p})$.

The proof almost directly follows from the proof of Theorem 3 in [18], by extending properties of binary trees to Δ-ary trees, see [3] for details. We now prove the lower bound.

Proof of Theorem 2. The proof idea is to view any network as the union of n optimal trees, one for each individual node. While the resulting network may violate the degree requirement, it constitutes a valid lower bound. So we start by finding an optimal structure for each source node i, according to all its communication destinations $\overrightarrow{\mathcal{D}}[i]$: We construct n Δ-ary trees, and let T^i_Δ be the optimal tree for source node i built using $\overrightarrow{\mathcal{D}}[i]$. From Lemma 4, we have:

$$\mathrm{EPL}(\overrightarrow{\mathcal{D}}[i], T^i_\Delta) = \sum_{j=1}^{n} p(j|i) d_{T^i_\Delta}(i,j) = \Omega(H_\Delta(Y \mid X = i))$$

where $\mathrm{EPL}(\overrightarrow{\mathcal{D}}[i], T^i_\Delta)$ denotes the expected path length of T^i_Δ given $\overrightarrow{\mathcal{D}}[i]$ and $d_{T^i_\Delta}$ denotes the shortest path in T^i_Δ. Now consider any bounded degree network N_Δ and compare it to the forest T made up of n trees $T^1_\Delta, T^2_\Delta, \ldots, T^n_\Delta$. Then,

$$\begin{aligned}\mathrm{EPL}(\mathcal{D}, N_\Delta) &= \sum_{i=1}^{n} p(i) \cdot \mathrm{EPL}(\overrightarrow{\mathcal{D}}[i], N_\Delta) \geq \sum_{i=1}^{n} p(i) \cdot \mathrm{EPL}(\overrightarrow{\mathcal{D}}[i], T^i_\Delta) \\ &\geq \sum_{i=1}^{n} p(i) \cdot H_\Delta(Y \mid X = i) = \Omega(H_\Delta(Y|X))\end{aligned}$$

Similarly we can consider a set of trees optimized toward the incoming communication of node j, $\overleftarrow{\mathcal{D}}[j]$, and the marginal destination probability. We show:

$$\mathrm{EPL}(\mathcal{D}, N_\Delta) \geq \Omega(H_\Delta(X \mid Y))$$

Hence the theorem follows. ◀

5 Network Design for Sparse Distributions

We now present families of distributions which enable DANs that match the lower bound. Our approach will be based on replacing neighborhoods with near optimal binary (or Δ-ary) trees. Following the lower bound of Lemma 4, it is easy to show a matching upper bound (for a constant Δ).

▶ **Lemma 5.** *Let $\bar{p}$ be a probability distribution on a set of node destinations (sources) V, and let u be a single source (destination) node. Then one can design a tree T with u as a root node with degree one, connected to a Δ-ary tree over V such that the expected path length from u to all destinations (or from all sources to u) is:*

$$\mathrm{EPL}(\bar{p}, T) = \sum_i p_i \cdot d_T(u,i) \leq O(H_\Delta(\bar{p})) \tag{7}$$

Proof. The proof follows by designing a Huffman Δ-ary code over $\bar{p}$ (with expected code length less than $H_\Delta(\bar{p}) + 1$ [9]) and using it to build a rooted Δ-ary tree. While the nodes in the Huffman code are tree leaves, we can move them up to become internal nodes, which only improves the expected path length. ◀

5.1 Tree Distributions

A most fundamental class of distributions for which we can construct optimal network designs is based on trees.

▶ **Theorem 6.** *Let $\mathcal{D}$ be a communication request distribution such that $G_\mathcal{D}$ is a tree (i.e., ignoring the edge direction, $G_\mathcal{D}$ forms a tree). Let X, Y be the random variables of the sources and destinations in $\mathcal{D}$, respectively. Then, it is possible to generate a* DAN *N with maximum degree 8, such that*

$$\mathrm{EPL}(\mathcal{D}, N) \leq O(H(Y \mid X) + H(X \mid Y))$$

This is asymptotically optimal.

Proof. We generate N as follows. Consider an arbitrary node as the root of the tree $G_\mathcal{D}$, call this tree $T_\mathcal{D}$, and consider the parent-child relationship implied by the root. Let $\pi(i)$ denote the parent of node i. Let $\overrightarrow{c}_i$ denote the communication distribution *from* v_i to its children (not including its single parent) and $\overrightarrow{\mathcal{D}}[i]$ denote the communication distribution *from* i to its neighbors (children and parent). Let $p_i^\pi = \overrightarrow{\mathcal{D}}[i][\pi(i)]$ denote the corresponding entry in $\overrightarrow{\mathcal{D}}[i]$ for the parent of i. From entropy Eq. (3), we have that $(1 - p_i^\pi)H(\overrightarrow{c}_i) \leq H(\overrightarrow{\mathcal{D}}[i])$. Similarly we define $\overleftarrow{c}_i$ and $\overleftarrow{\mathcal{D}}[i]$ as the communication distribution *to* v_i, from its children and neighbors respectively.

The construction has two phases. In the first phase we replace outgoing edges. For each node i replace the edges between i and its *children* with a binary tree according to $\overrightarrow{c}_i$ and the method of [18] (or Lemma 5 for a general Δ) for creating a near optimal binary tree. Let $\overrightarrow{B}_i$ denote this tree and recall that $\mathrm{EPL}(\overrightarrow{c}_i, \overrightarrow{B}_i) \leq O(H(\overrightarrow{c}_i))$. Note that every node i may appear in at most two trees $\overrightarrow{B}_i$ and $\overrightarrow{B}_{\pi(i)}$; in $\overrightarrow{B}_i$ its degree is one and in $\overrightarrow{B}_{\pi(i)}$ its degree is at most 3, so the outgoing degree of each node is at most 4 after this phase.

In the second phase we take care of the remaining incoming edges from children to parents. For each node i replace the edges from its *children* to it with a binary tree according to $\overleftarrow{c}_i$ and the method of [18] for creating a near optimal binary tree. Let $\overleftarrow{B}_i$ denote this tree and recall that $\mathrm{EPL}(\overleftarrow{c}_i, \overleftarrow{B}_i) \leq O(H(\overleftarrow{c}_i))$. Note that every node i may appear in at most two trees $\overleftarrow{B}_i$ and $\overleftarrow{B}_{\pi(i)}$; in $\overleftarrow{B}_i$ i's degree is one and in $\overleftarrow{B}_{\pi(i)}$ i's degree is at most 3. Thus, the incoming degree of each node is bounded by 4 after this phase.

Now we bound $\mathrm{EPL}(\mathcal{D}, N)$ by bounding the expected path lengths in the corresponding binary trees of each node, first considering all edges from parent to children and then all edges from children to parent. Let $p(i)$ and $q(i)$ denote the probabilities that node i will be a source and a destination of a communication request, respectively. Then:

$$\begin{aligned}\mathrm{EPL}(\mathcal{D},N) &\leq \sum_{(u,v)\in\mathcal{D}} p(u,v)d_N(u,v)\\ &= \sum_{(\pi(i),i)\in T_{\mathcal{D}}} p(\pi(i),i)d_N(\pi(i),i) + \sum_{(i,\pi(i))\in T_{\mathcal{D}}} p(i,\pi(i))d_N(i,\pi(i))\\ &= \sum_{i=1}^{n} p(i)\mathrm{EPL}(\overrightarrow{c}_i,\overrightarrow{B}_i) + \sum_{i=1}^{n} q(i)\mathrm{EPL}(\overleftarrow{c}_i,\overleftarrow{B}_i)\\ &\leq \sum_{i=1}^{n} p(i)H(\overrightarrow{\mathcal{D}}[i]) + \sum_{i=1}^{n} q(i)H(\overleftarrow{\mathcal{D}}[i]) = H(Y\mid X)+H(X\mid Y)\end{aligned}$$

This matches our lower bound in Theorem 2. ◀

5.2 General Sparse Distributions

Asymptotically optimal demand-aware networks can even be designed for general sparse distributions.

▶ **Theorem 7.** *Let $\mathcal{D}$ be a communication request distribution where Δ_{avg} is the average degree in $G_{\mathcal{D}}$ (so the number of edges $m = \frac{\Delta_{\mathrm{avg}}\cdot n}{2}$). Let X, Y be the random variables of the sources and destinations in $\mathcal{D}$, respectively. Then, it is possible to generate a* DAN *N with maximum degree $12\Delta_{\mathrm{avg}}$, such that*

$$\mathrm{EPL}(\mathcal{D},N) \leq O(H(Y\mid X)+H(X\mid Y)) \tag{8}$$

This is asymptotically optimal when Δ_{avg} is a constant.

Proof. Recall that $G_{\mathcal{D}}$ (for short G) is a directed graph and define in-degree and out-degree in the canonical way. Let the (undirected) degree of a node, be the sum of its in-degree and out-degree and denote the average degree as Δ_{avg}. Denote the $n/2$ nodes with the lowest degree in G as *low degree* nodes and the rest as *high degree* nodes. Note that each low degree node has a degree at most $2\Delta_{\mathrm{avg}}$ and both its in-degree and out-degree must be low. A node with out-degree (in-degree) larger than $2\Delta_{\mathrm{avg}}$ is called a *high out-degree* (*high in-degree*) node (some nodes are neither low or high degree nodes).

The construction of N will be done in two phases. In the first phase, we consider only (directed) edges (u,v) between a high out-degree u and a high in-degree node v. We subdivide each such edge with two edges that connect u to v via a helping low degree node ℓ, i.e., removing the directed edge (u,v) and adding the edges (u,ℓ) and (v,ℓ). Note that there are at most m such edges, so we can distribute the help between low degree nodes in such a way that each low degree node helps at most Δ_{avg} such edges. Call the resulting graph G'.

Accordingly, we also create a new matrix $\mathcal{D}'$ which, initially, is identical to $\mathcal{D}$. Then for each (u,v) and ℓ as above we set $\mathcal{D}'(u,v)=0$, $\mathcal{D}'(u,\ell)=\mathcal{D}(u,l)+\mathcal{D}(u,v)$ and $\mathcal{D}'(\ell,v)=\mathcal{D}(l,v)+\mathcal{D}(u,v)$. Note that $\mathcal{D}'$ is not a distribution matrix anymore, as the sum of all the entries is more than one, but it has the following property: For each high degree node i, we have $H(\overrightarrow{\mathcal{D}'}[i]) \leq H(\overrightarrow{\mathcal{D}}[i]$ and $H(\overleftarrow{\mathcal{D}'}[i]) \leq H(\overleftarrow{\mathcal{D}}[i])$ (see Eq. (2)).

In the second phase, we construct N from G'. Consider each node i with high out-degree and create a nearly optimal binary tree $\overrightarrow{B}^i$ according to $\overrightarrow{\mathcal{D}'}[i]$ using the method of [18]. Add an edge from i to the root of $\overrightarrow{B}^i$ and delete all the out-edges from i from G'. Similarly consider each node j with high in-degree and create a nearly optimal binary tree $\overleftarrow{B}_j$ according to $\mathcal{D}'[i]$ using the method of [18]. Add an edge from j to the root of $\overleftarrow{B}_j$ and delete all the in-edges of j from G'. This completes the construction of N.

We first bound the maximum degree in N. First consider a low degree node ℓ, helping an edge (u, v), i.e., u is high out-degree and v is high-indegree. So ℓ is part of both u's and v's binary tree, hence ℓ's degree increases by at most 6 (two times degree 3 for being an internal node). Note that ℓ needs to help at most Δ_{avg} edges itself. For each of these Δ_{avg} edges, ℓ's degree will be at most 6, resulting in a degree of $6\Delta_{\text{avg}}$. Since ℓ's degree was at most $2\Delta_{\text{avg}}$, in the worst case, ℓ was associate with $2\Delta_{\text{avg}}$ high in-degree or out-degree nodes, i.e., ℓ will be present in all these $2\Delta_{\text{avg}}$ trees, which results in another $6\Delta_{\text{avg}}$ degrees for ℓ. In total, ℓ's degree is $12\Delta_{\text{avg}}$. If a node h has both high out-degree and high in-degree, then its degree will be two: h is connected to the root of the tree created of its out-edges and to the root of the tree created of its in-edges. If a node u is only a high out-degree node, its degree in N is bounded by $6\Delta_{\text{avg}} + 1$ (and similarly for a node u which is only a high in-degree node). If a node is neither high nor low degree, then its degree in N is bounded by $12\Delta_{\text{avg}}$ (originally it was up to $4\Delta_{\text{avg}}$ in G'). We now bound $\text{EPL}(\mathcal{D}, N)$. Recall that from Lemma 5 and Eq. (2), we have,

$$\text{EPL}(\overrightarrow{\mathcal{D}'}[i], \overrightarrow{B}_i) \leq O(H(Y \mid X = i))$$

and

$$\text{EPL}(\overleftarrow{\mathcal{D}'}[j], \overleftarrow{B}_j) \leq O(H(X \mid Y = j))$$

For each request (i, j) in $\mathcal{D}$ there are two possibilities for the route on N: either the edge $(i, j) \in N$ is a direct route, or the route goes via $\overrightarrow{B}_i$ or $\overleftarrow{B}_j$ or both. Let $\mathcal{O}$ and $\mathcal{I}$ be the set of high out-degree and in-degree nodes, respectively. Then:

$$\begin{aligned}
\text{EPL}(\mathcal{D}, N) &= \sum_{(u,v)\in\mathcal{D}} p(u,v) d_N(u,v) \\
&\leq \sum_{(i,j)\notin\mathcal{O}\cup\mathcal{I}} p(u,v) + \sum_{i\in\mathcal{O}} p(i)\text{EPL}(\overrightarrow{\mathcal{D}}[i]), \overrightarrow{B}_i) + \sum_{j\in\mathcal{I}} q(j)\text{EPL}(\overleftarrow{\mathcal{D}}[j])\overleftarrow{B}_j) \\
&= \sum_{i\notin\mathcal{O}} p(i) + \sum_{j\notin\mathcal{I}} q(j) + \sum_{i\in\mathcal{O}} p(i)\text{EPL}(\overrightarrow{\mathcal{D}}[i]), \overrightarrow{B}_i) + \sum_{j\in\mathcal{I}} q(j)\text{EPL}(\overleftarrow{\mathcal{D}}[j])\overleftarrow{B}_j) \\
&\leq O(H(X \mid Y) + H(Y \mid X))
\end{aligned}$$

This matches our lower bound in Theorem 2. ◀

6 Regular and Uniform Distributions

Another large family of distributions for which demand-aware networks can be designed are regular and uniform distributions $\mathcal{D}$. While it is easy to see that both conditions can be relaxed such that the supported distributions can be "*nearly* regular" and "*nearly* uniform", for ease of presentation, we keep the conditions strict in what follows.

We first establish an interesting connection to spanners. As we will see, this connection will provide a simple and powerful technique to design a wide range of demand-aware networks meeting the conditional entropy lower bound.

▶ **Theorem 8.** *Let $\mathcal{D}$ be an arbitrary (possibly dense) regular and uniform request distribution. It holds that if we can find a constant and sparse (i.e., constant distortion, linear sized) spanner for $G_{\mathcal{D}}$, we can design a constant degree* DAN *N providing an expected path length of*

$$\text{EPL}(\mathcal{D}, N) \leq O(H(Y \mid X) + H(X \mid Y)) \tag{9}$$

This is asymptotically optimal.

In other words, for regular and uniform distributions, the network design problem boils down to finding a constant[1] sparse spanner: as we will see, we can automatically transform this spanner into an efficient network (which may contain auxiliary edges). The remainder of this section is devoted to the proof of the theorem.

Assume that $\mathcal{D}$ is r-regular and uniform. Recall that in this case $H(Y \mid X) = H(X \mid Y) = \log r$, so $\text{BND}(\mathcal{D}, \Delta) \geq \Omega(H(Y \mid X))$ where Δ is a constant. We now describe how to transform a constant sparse (but potentially irregular) spanner for $G_{\mathcal{D}}$ into a constant-degree host network N with $\text{EPL}(\mathcal{D}, N) \leq O(\log r)$. This will be done using a similar degree reduction technique as discussed earlier (in the proof of Theorem 7).

▶ **Lemma 9.** *Let G be a graph of maximum degree $\Delta_{\max}$ and an average degree Δ_{avg}. Then, we can construct a graph G' with maximum degree $8\Delta_{\text{avg}}$ which is a graph metric $\log \Delta_{\max}$-spanner of G, i.e., $d_{G'}(u, v) \leq 2 \log \Delta_{\max} \cdot d_G(u, v)$.*

Proof. Let us call the $n/2$ nodes with the lowest degree in G the *low degree* nodes and the remaining nodes *high degree* nodes. By the pigeon hole principle, each low degree node has a degree at most $2\Delta_{\text{avg}}$. The construction of G' proceeds in two phases. In the first phase we take every edge between high degree nodes u, v and subdivide it with two edges that connect u to v via a helping low degree node ℓ, i.e., removing the edge (u, v) and adding the edges (u, ℓ) and (v, ℓ). Note that there are at most m edges connecting high degree nodes so we can distribute the help between low degree nodes such that each low degree node helps to at most Δ_{avg} such edges.

In the second phase we consider each high degree node u and replace the set of edges between u and its neighbors, $\Gamma(u)$, with a balanced binary tree that connects u as the root and $\Gamma(u)$ as remaining nodes of the tree. Denote as B_u this tree and note that the height of B_u is at most $\log(|\Gamma(u)| + 1)$. We leave edges between low degree nodes as in G.

Let us analyze the degrees in G'. Since every high degree node u in G' only connects to low degree nodes, it is only a member of B_u and its degree reduces to 2 in G'. Now consider a low degree node ℓ: for each edge (u, v) it helps, ℓ participates in B_u and B_v. Hence, its degree increases by at most 6 in G' compared to G. Overall, for helping high degree nodes, the degree of ℓ can increase by $6\Delta_{\text{avg}}$. Together with its original neighbors from G, the degree of ℓ in G' can be at most $8\Delta_{\text{avg}}$.

Next consider the distortion of G'. The distortion between neighboring low degree nodes is one. The distortion between neighboring high degree nodes is at most twice $\log \Delta_{\max}$ and the distortion between a neighboring high and low degree is at most $\log \Delta_{\max}$.

So, $d_{G'}(u, v) \leq 2 \log \Delta_{\max} \cdot d_G(u, v)$ for all u, v in G'. ◀

We will apply Lemma 9 to prove Theorem 8.

Proof of Theorem 8. Let S be a constant and sparse spanner of $G_{\mathcal{D}}$ (G could be either a subgraph or a metric spanner of max degree asymptotically not larger than $G_{\mathcal{D}}$) of degree at most r. Lemma 9 then tells us how to transform S to a DAN N of degree Δ_{avg}. Since S is a constant spanner there is a constant c such that,

$$\text{EPL}(\mathcal{D}, S) = \sum_{(u,v)\in\mathcal{D}} \text{p}(u, v) \cdot d_S(v, v) = c \tag{10}$$

[1] To be precise, a spanner with constant *average* distortion will be sufficient (see [3] for details). However, for simplicity, we leave it as a constant spanner.

Since S has maximum degree r, according to Lemma 9, it has a graph metric spanner N such that, the distance of any source-destination pair of $G(\mathcal{D})$ in N is at most $2\log r$ times their distance in S. Hence:

$$\begin{aligned}\mathrm{EPL}(\mathcal{D}, N) &= \sum_{(u,v)\in\mathcal{D}} \mathrm{p}(u,v)\cdot d_N(u,v) \leq \sum_{(u,v)\in\mathcal{D}} \mathrm{p}(u,v)\cdot d_S(u,v)\cdot 2\log r \\ &\leq \log r \cdot\ \mathrm{EPL}(\mathcal{D}, S) = O(\log r) = O(H(Y \mid X))\end{aligned}$$

The last equality holds since $\mathcal{D}$ is r-regular and uniform. The bound is asymptotically optimal when Δ is a constant: it matches our lower bound in Theorem 2. ◀

Theorem 8 allows us to simplify the design of asymptotically optimal networks for uniform and regular distributions $\mathcal{D}$ where $G_{\mathcal{D}}$ has a constant sparse spanner. In particular, the approach can be used to design optimal networks for the following large families of distributions which are known to have a constant and sparse graph spanners.

▶ **Corollary 10.** *Let $\mathcal{D}$ describe a uniform and regular communication request distribution. Then, it is possible to generate a constant degree* DAN *N such that*

$$\mathrm{EPL}(\mathcal{D}, N) \leq O(H(Y \mid X) + H(X \mid Y)) \tag{11}$$

in the following scenarios:

- *If, for a constant $c \geq 1$, $G_{\mathcal{D}}$ has a minimum degree $\Delta \geq n^{\frac{1}{c}}$.*[2]
- *If $G_{\mathcal{D}}$ forms a hypercube with $n\log n$ edges.*
- *If $G_{\mathcal{D}}$ forms a (possibly dense) chordal graph.*

See [3] for the proof.

We round off our study of uniform and regular distributions by considering one more interesting family of (possibly very dense) distributions: distributions $\mathcal{D}$ which describe a bounded and *local* doubling dimension, note that this family is more general than the standard bounded doubling dimension graphs which are sparse.

First, recall that a metric space (V, d) has a constant doubling dimension if and only if there exists a constant λ such that every ball of radius r in V can be covered by λ balls of half the radius $r/2$, for all $r \geq 1$. In general, the smallest λ which satisfies this property for a metric space is called *doubling constant* and $\log_2 \lambda$ is called the *doubling dimension* [6, 12, 13, 14]. A metric space is called *bounded* (a.k.a. constant or low) doubling dimension if λ is a constant. There has been a wide range of work on spanners for bounded doubling dimension metrics [5, 6, 13, 14]. However, if the metric is imposed by a graph metric (via shortest paths) then a bounded doubling dimension implies that the graph is nearly regular, of bounded (constant) degree and sparse. Theorem 7 already solved the case of sparse $G_{\mathcal{D}}$, even for non-uniform and irregular distributions.

In the following, however, we are interested in a more general notion of doubling dimension, which allows a higher density, unbounded degree: we call it *locally-bounded doubling dimension*:

▶ **Definition 11** (Locally-Bounded Doubling Dimension (LDD))**.** $G_{\mathcal{D}}$ implied by the distribution $\mathcal{D}$ has a *locally-bounded doubling dimension* if and only if there exists a constant λ such that

[2] In this case the constant in the O notation depends linearly on c.

the 2-hop neighbors of any node u are covered by at most λ 1-hop neighbors. Formally, for each $u \in V$, there exists a set of nodes $y_1, y_2, ...y_\lambda$, such that:

$$B(u,2) \subseteq \bigcup_{i=1}^{\lambda} B(y_i,1)$$

where $B(u,r)$ are the set of nodes that are at distance at most r-hops from u in $G_\mathcal{D}$.

Clearly, every bounded doubling dimension graph is also of locally-bounded doubling dimension, but the converse is not true. In particular, the latter enables graphs which could be dense, with unbound degree, and possibly with irregularity of degree.

In the remainder of this section, we will prove the following theorem.

▶ **Theorem 12.** *Let $\mathcal{D}$ describe a uniform and regular communication request distribution of locally-bounded doubling dimension. Then it is possible to design a constant degree* DAN N *such that*

$$\mathrm{EPL}(\mathcal{D}, N) \leq O(H(Y \mid X) + H(X \mid Y)) \tag{12}$$

This is asymptotically optimal.

Proof. Again, our proof strategy is to employ Theorem 8. Accordingly, we show that a constant sparse spanner exists for locally-bounded doubling dimension networks. In particular, we will design this spanner based on an ϵ-net construction. We first recall the definition of *ϵ-nets* [6].

▶ **Definition 13** (ϵ-net)**.** A subset V' of V is an *ϵ-net* for a graph $G = (V, E)$ if it satisfies the following two conditions:

1. for every u, $v \in V'$, $d_G(u,v) > \epsilon$
2. for each $w \in V$, there exists at least one $u \in V'$ such that, $d_G(u,w) \leq \epsilon$

Let $G_\mathcal{D} = (V, E)$ be a locally-bounded doubling dimension network. We now first construct a spanner S' of $G_\mathcal{D}$ which is a subgraph of $G_\mathcal{D}$, using the following ($\epsilon = 2$)-net: we sort all nodes according to decreasing (remaining) degrees, and iteratively select the high-degree nodes into the 2-net one-by-one and remove their 2-neighborhoods. Clearly, after this process, each node is either part of the 2-net or has a 2-net node at distance at most 2-hops, and we have computed a legal 2-net.

To form the spanner S, we next arbitrarily match each node u not in the 2-net to one of its nearest 2-net nodes v, and select the edges along a shortest path from u to v into the spanner S. This results in a forest of connected components (2-layered stars). We call these connected components *clusters* and the corresponding nodes in the 2-net *cluster heads*. We denote the cluster associated to the net node u by $Cl(u)$.

We next connect the connected clusters to each other, in a sparse manner. Towards this end, we connect each pair of clusters, with an arbitrary single edge, if they contain at least one pair of communicating nodes in $G_\mathcal{D}$. We can claim the following.

▶ **Lemma 14.** *S is a constant and sparse spanner of $G_\mathcal{D}$ (with distortion 9) .*

Proof. Let (u,v) be an edge in $G_\mathcal{D}$ and $u \in Cl(u)$, $v \in Cl(v)$. By construction, there are nodes $x \in Cl(u)$ and $y \in Cl(v)$ that are connected by an edge in S, and hence there is a path $u, C(u), x, y, C(v), v$ in S. Therefore, $d_S(u,v) \leq d_S(u, Cl(u)) + d_S(Cl(u), x) + d_S(x,y) + d_S(y, Cl(v)) + d_S(Cl(v), v) \leq 9$.

The spanner is also sparse: in a nutshell, due to the 2-net properties, we know that the distance between communicating cluster heads is at most 5: since in a locally doubling dimension graph the number of cluster heads at distance 5 is bounded, only a small number of neighboring clusters will communicate. More formally, after associating each node to some unique cluster, we have a linear number of edges in the spanner. Next we bound the number of outgoing edges from each cluster. Let (u, v) be such an edge where $u \in Cl(u)$, $v \in Cl(v)$. Let the cluster heads of $Cl(u)$ and $Cl(v)$ be i and j, respectively. By construction i and j are at most at distance 5 in $G_{\mathcal{D}}$, i.e., $d_{G_{\mathcal{D}}}(i, j) \leq 5$. So, if we can bound the number of 2-net nodes which lie within 5 hops from some net node i, it will give us a bound on the number of edges which we add between $Cl(u)$ and other clusters. According to Definition 11, all the two hop neighbors of i can be covered within one hop neighbors of λ nodes, where λ is the corresponding doubling constant. If we consider two hop neighbors of all these λ many nodes, they cover all the 3 hop neighbors of i. To cover the 2 hop neighbors of each of these nodes, we again require one hop neighbors of λ nodes. So, to cover all 3 hop neighbors of i, we require at most λ^2 one hop neighbors. Inductively, by repeating this argument, we require one hop neighbors of at most λ^4 nodes to cover all the 5 hop neighbors of i. Since we constructed a 2-net, each of these λ^4 balls with radius one contains at most one 2-net node. Hence there are at most λ^4 2-net nodes which are at a distance 5 hops or less from i. Consequently, there are at most λ^4 inter-cluster edges associated to some cluster $Cl(u)$, or cluster head i. Since there can not be more than linear number of clusters, hence the number of edges in S' is also linear. ◀

Using Lemma 14 we can directly use Theorem 8 and conclude the proof of Theorem 12. ◀

In fact, it turns out that if we consider a *metric* spanner, and by using auxiliary edges, we can improve the above distortion and constract better constant sparse spanner S'. The idea is to add inter-cluster edges only between the cluster heads. This will reduce the distortion to 5 while keeping the same number to total edges. The degree of each node in S' will increase by at most a constant, λ^4. Adjusting Theorem 8 respectively to support metric spanners (and only a subgraph spanner) will enable us to use S' instead of S.

7 Conclusion

This paper initiated the study of a fundamental network design problem. While our work is motivated in particular by emerging technologies for more flexible datacenter interconnects as well as by peer-to-peer overlays, we believe that our model is very natural and of interest beyond this specific application domain considered in this paper. For example, the design of a sparse, bounded-degree and distance-preserving network can also be understood from the perspective of graph sparsification [27]: the designed network can be seen as a compact representation of the original network.

The subject of bounded network design offers several interesting avenues for future research. In particular, while we presented asymptotically optimal network design algorithms for a wide range of distributions and while we believe that the entropy is the right measure to assess network designs, there remain several (dense) distributions for which the quest for optimal network designs remains open, perhaps also requiring us to explore alternative flavors of graph entropy.

Acknowledgments. We would like to thank Michael Elkin for many inputs and discussions.

References

1 Chen Avin, Michael Borokhovich, Bernhard Haeupler, and Zvi Lotker. Self-adjusting grid networks to minimize expected path length. In *Proc. SIROCCO*, pages 36–54, 2013.

2 Chen Avin, Michael Borokhovich, and Stefan Schmid. OBST: A self-adjusting peer-to-peer overlay based on multiple BSTs. In *Peer-to-Peer Computing (P2P), 2013 IEEE Thirteenth International Conference on*, pages 1–5. IEEE, 2013.

3 Chen Avin, Kaushik Mondal, and Stefan Schmid. Demand-aware network designs of bounded degree. In *arXiv report no. 1705.06024*, 2017.

4 Hubert T-H Chan, Michael Dinitz, Anupam Gupta, et al. Spanners with slack. In *ESA*, volume 6, pages 196–207. Springer, 2006.

5 T-H Hubert Chan and Anupam Gupta. Small hop-diameter sparse spanners for doubling metrics. *Discrete & Computational Geometry*, 41(1):28–44, 2009.

6 T.-H. Hubert Chan, Anupam Gupta, Bruce M. Maggs, and Shuheng Zhou. On hierarchical routing in doubling metrics. *ACM Trans. Algorithms*, (4):55:1–55:22, 2016.

7 Moses Charikar, Mohammad Taghi Hajiaghayi, Howard Karloff, and Satish Rao. l_2^2 spreading metrics for vertex ordering problems. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1018–1027. Society for Industrial and Applied Mathematics, 2006.

8 C. Chen, R. Vitenberg, and H. A. Jacobsen. Scaling construction of low fan-out overlays for topic-based publish/subscribe systems. In *2011 31st International Conference on Distributed Computing Systems*, pages 225–236, June 2011.

9 Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

10 Erik D Demaine, Dion Harmon, John Iacono, Daniel Kane, and Mihai Patraşcu. The geometry of binary search trees. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 496–505. SIAM, 2009.

11 Uriel Feige and James R Lee. An improved approximation ratio for the minimum linear arrangement problem. *Information Processing Letters*, 101(1):26–29, 2007.

12 Pierre Fraigniaud, Emmanuelle Lebhar, and Zvi Lotker. A doubling dimension threshold θ (log log n) for augmented graph navigability. In *ESA*, pages 376–386. Springer, 2006.

13 Anupam Gupta, Robert Krauthgamer, and James R Lee. Bounded geometries, fractals, and low-distortion embeddings. In *Proc. IEEE FOCS*, pages 534–543, 2003.

14 Sariel Har-Peled and Manor Mendel. Fast construction of nets in low-dimensional metrics and their applications. *SIAM Journal on Computing*, 35(5):1148–1184, 2006.

15 Lawrence H Harper. Optimal assignments of numbers to vertices. *Journal of the Society for Industrial and Applied Mathematics*, 12(1):131–135, 1964.

16 Su Jia, Xin Jin, Golnaz Ghasemiesfeh, Jiaxin Ding, and Jie Gao. Competitive analysis for online scheduling in software-defined optical wan. In *Proc. IEEE INFOCOM*, 2017.

17 M. Ghobadi et al. Projector: Agile reconfigurable data center interconnect. In *Proc. ACM SIGCOMM*, pages 216–229, 2016.

18 Kurt Mehlhorn. Nearly optimal binary search trees. *Acta Inf.*, 5:287–295, 1975.

19 Adam Meyerson and Brian Tagiku. Minimizing average shortest path distances via shortcut edge addition. In *Proc. APPROX/RANDOM*, pages 272–285, Berlin, Heidelberg, 2009.

20 Melih Onus and Andréa W Richa. Minimum maximum-degree publish–subscribe overlay network design. *IEEE/ACM Transactions on Networking*, 19(5):1331–1343, 2011.

21 Melih Onus and Andréa W Richa. Parameterized maximum and average degree approximation in topic-based publish-subscribe overlay network design. *Computer Networks*, 94:307–317, 2016.

22 David Peleg and Alejandro A. Schäffer. Graph spanners. *Journal of Graph Theory*, 13(1):99–116, 1989.

23 Arjun Roy, Hongyi Zeng, Jasmeet Bagga, George Porter, and Alex C Snoeren. Inside the social network's (datacenter) network. In *ACM SIGCOMM Computer Communication Review*, volume 45, pages 123–137. ACM, 2015.

24 Stefan Schmid, Chen Avin, Christian Scheideler, Michael Borokhovich, Bernhard Haeupler, and Zvi Lotker. Splaynet: Towards locally self-adjusting networks. *IEEE/ACM Trans. Netw.*, 24(3):1421–1433, June 2016.

25 Ankit Singla. Fat-free topologies. In *Proc. 15th ACM Workshop on Hot Topics in Networks (HotNets)*, pages 64–70, 2016.

26 Daniel Dominic Sleator and Robert Endre Tarjan. Self-adjusting binary search trees. *J. ACM*, 32(3):652–686, July 1985.

27 Daniel A Spielman and Shang-Hua Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proc. ACM STOC*, pages 81–90, 2004.