Content Search Through Comparisons

Amin Karbasi[1], Stratis Ioannidis[2], and Laurent Massoulié[3]

[1] Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland
[2] Technicolor, Palo Alto, USA
[3] Technicolor, Paris, France

Abstract. We study the problem of navigating through a database of similar objects using comparisons under heterogeneous demand, a problem closely related to small-world network design. We show that, under heterogeneous demand, the small-world network design problem is NP-hard. Given the above negative result, we propose a novel mechanism for small-world network design and provide an upper bound on its performance under heterogeneous demand. The above mechanism has a natural equivalent in the context of content search through comparisons, again under heterogeneous demand; we use this to establish both upper and lower bounds on content search through comparisons.

1 Introduction

The problem we study in this paper is content search through comparisons. In short, a user searching for a target object navigates through a database in the following manner. The user is asked to select the object most similar to her target from small list of objects. A new object list is then presented to the user based on her earlier selection. This process is repeated until the target is included in the list presented, at which point the search terminates.

Searching through comparisons is typical example of *exploratory search* [14], the need for which arises when users are unable to state and submit explicit queries to the database. Exploratory search has several important real-life applications. An often-cited example [13, 12] is navigating through a database of pictures of humans in which subjects are photographed under diverse uncontrolled conditions. For example, the pictures may be taken outdoors, from different angles or distances, while the subjects assume different poses, are partially obscured, *etc.* Automated methods may fail to extract meaningful features from such photos, so the database cannot be queried in the traditional fashion. On the other hand, a human searching for a particular person can easily select from a list of pictures the subject most similar to the person she has in mind.

Users may also be unable to state queries because, *e.g.*, the are unfamiliar with the search domain, or do not have a clear target in mind. For example, a novice classical music listener may not be able to express that she is, *e.g.*, looking for a fugue or a sonata. She might however identify among samples of different musical pieces the closest to the one she has in mind. Alternatively, a user surfing the web may not know a priori which post she wishes to read;

presenting a list of blog posts and letting the surfer identify which one she likes best can steer her in the right direction.

In all the above applications, the problem of content search through comparisons amounts to determining which objects to present to the user in order to find the target object as quickly as possible. Formally, the behavior of a human user can be modeled by a so-called *comparison oracle* [5]: given a target and a choice between two objects, the oracle outputs the one closest to the target. The goal is thus to find a sequence of proposed pairs of objects that leads to the target object with as few oracle queries as possible. This problem was introduced in [5] and has recently received considerable attention [11, 12, 13].

Content search through comparisons is also naturally related to the following problem: given a graph embedded in a metric space, how should one augment this graph by adding edges in order to minimize the expected cost of greedy forwarding over this graph? This is known as the *small-world network design* problem [4, 3] and has a variety of applications as, *e.g.*, in network routing. In this paper, we consider both problems under the scenario of *heterogeneous demand.* This is very interesting in practice: objects in a database are indeed unlikely to be requested with the same frequency. Our contributions are as follows:

– We show that the small-world network design problem under general heterogeneous demand is NP-hard. Given earlier work on this problem under homogeneous demand [3, 4], this result is interesting in its own right.
– We propose a novel mechanism for edge addition in the small-world design problem, and provide an upper bound on its performance.
– The above mechanism has a natural equivalent in the context of content search through comparisons, and we provide a matching upper bound for the performance of this mechanism.
– Finally, we also establish a lower bound on any mechanism solving the content search through comparisons problem.

To the best of our knowledge, we are the first to study the above two problems in a setting of heterogeneous demand. Our analysis is intuitively appealing because our upper and lower bounds relate the cost of content search to two important properties of the demand distribution, namely its *entropy* and its *doubling constant.* We thus provide performance guarantees in terms of the *bias* of the distribution of targets, captured by the entropy, as well as the *topology* of their embedding, captured by the doubling constant.

The remainder of this paper is organized as follows. In Section 2 we provide an overview of the related work in this area. In Sections 3 and 4 we introduce our notation and formally state the two problems that are the focus of this work, namely content search through comparisons and small-world network design. We present our main results in Section 5 and our conclusions in Section 6.

2 Related Work

Content search through comparisons is a special case of nearest neighbour search (NNS) [1, 6], where it is typical to assume that database objects are embedded in a metric space with a small intrinsic dimension. Krauthgamer and Lee [10] introduce navigating nets, a data structure for NNS in doubling metric spaces. Clarkson [1] considers a similar structure for objects embedded in a space satisfying a sphere-packing property, while Karger and Ruhl [8] study NNS under growth-restricted metrics. All three assumptions have formal connections to the doubling constant we consider in this paper. However, in these works, the underlying metric space is fully observable by the search mechanism, and the demand over target objects is homogeneous. Our work assumes access only to a comparison oracle while also dealing with heterogeneous demand.

NNS with access to a comparison oracle was first introduced by Lifshits *et al.* [5], and further explored by Lifshits and Zhang [11] and Tshopp and Diggavi [12, 13]. In contrast to [8, 10, 1], the above authors do not assume that objects are necessarily embedded in a metric space; instead, they only require that a comparison oracle can rank any two objects in terms of their similarity to a given target. To provide performance guarantees on the search cost, Lifshits *et al.* introduce a *disorder constant* [5], capturing the degree to which object rankings violate the triangle inequality. This disorder constant plays roughly the same role in their analysis as the doubling constant does in ours. Nevertheless, these works also assume homogeneous demand. Our work introduces the notion of heterogeneity while assuming that a metric embedding exists.

Small-world networks (also called navigable networks) have received a lot of attention since Kleinberg's seminal paper [9]. Our work is closest to Fraigneaud *et al.* [4], [3], who identify conditions under which graphs embedded in a doubling metric space are navigable. Again, our approach to small-world network design differs by considering heterogeneous demand, an aspect absent from earlier work.

3 Definitions and Notation

Comparison Oracle. Consider a set of objects $\mathcal{N}$, where $|\mathcal{N}| = n$, and a metric space $(\mathcal{M},d)$, where $d(x,y)$ denotes the distance between $x, y \in \mathcal{M}$. Assume that objects in $\mathcal{N}$ are embedded in $(\mathcal{M},d)$, *i.e.*, there exists a 1-to-1 mapping from $\mathcal{N}$ to a subset of $\mathcal{M}$. The objects in $\mathcal{N}$ may represent, for example, pictures in a database. The metric embedding is a mapping from the pictures to a set of attributes (*e.g.*, the person's age, her eye color, *etc.*). The distance d then represents how "similar" objects are w.r.t. these attributes. In what follows, we abuse notation and write $\mathcal{N} \subseteq \mathcal{M}$, keeping in mind that database objects (the pictures) are in fact distinct from their embedding (their attributes).

Given an object $z \in \mathcal{N}$, we write $x \preccurlyeq_z y$ if $d(x,z) \leq d(y,z)$, ordering thus objects according to their distance from z. Moreover, we write $x \sim_z y$ if $d(x,z) = d(y,z)$ and $x \prec_z y$ if $x \preccurlyeq_z y$ but not $x \sim_z y$. For a non-empty $A \subseteq \mathcal{N}$, let $\min_{\preccurlyeq_z} A$ be the set of objects in A closest to z, *i.e.*, $w \in \min_{\preccurlyeq_z} A \subseteq A$ if $w \preccurlyeq_z v$ for all $v \in A$.

A *comparison oracle* [5] is an oracle that, given two objects x, y and a target t, returns the closest object to t. More formally,

$$\text{Oracle}(x,y,t) = \begin{cases} x & \text{if } x \prec_t y, \\ y & \text{if } x \succ_t y, \\ x \text{ or } y & \text{if } x \sim_t y. \end{cases} \tag{1}$$

This oracle "models" human users: a user interested in locating, *e.g.*, a target picture t within the database, can compare two pictures with respect to their similarity to this target but cannot associate a numerical value to this similarity. When the two pictures are equidistant from t the user's decision is arbitrary.

Entropy and Doubling Constant. For any ordered pair $(s,t) \in \mathcal{N} \times \mathcal{N}$, we call s the *source* and t the *target* of the pair. We consider a probability distribution λ over all ordered pairs of objects in $\mathcal{N}$ which we call the *demand.* We refer to the marginal distributions $\nu(s) = \sum_t \lambda(s,t)$ and $\mu(t) = \sum_s \lambda(s,t)$, as the *source* and *target* distributions, respectively. Moreover, we refer to the support of the target distribution $\mathcal{T} = \mathsf{supp}(\mu) = \{x \in \mathcal{N} : \text{ s.t. } \mu(x) > 0\}$ as the *target set* of the demand.

Let σ be a probability distribution over $\mathcal{N}$. We define the *entropy* and *max-entropy* of σ, respectively, as

$$H(\sigma) = \sum_{x \in \mathsf{supp}(\sigma)} \sigma(x) \log \sigma^{-1}(x), \qquad H_{\max}(\sigma) = \max_{x \in \mathsf{supp}(\sigma)} \log \sigma^{-1}(x). \tag{2}$$

The entropy has strong connections with content search. More specifically, suppose that we have access to a so-called *membership oracle* [2] that answers queries of the following form: "Given a target t and a subset $A \subseteq \mathcal{N}$, does t belong to A?"Let t be a random target selected with distribution μ. To identify t one needs to submit at least $H(\mu)$ queries, in expectation, to a membership oracle, and there exists an algorithm (Huffman coding) that identifies t with only $H(\mu)+1$ queries, in expectation (see, *e.g.*, [2]). In the worst case, which occurs when the target is the least frequently selected object, the algorithm requires $H_{\max}(\mu)+1$ queries to identify t. Our work identifies similar bounds assuming that one only has access to a comparison oracle, as defined in (1). Not surprisingly, the entropy $H(\mu)$ also shows up in our performance bounds (Theorems 3 and 4).

For $x \in \mathcal{N}$, we denote by $B_x(r) = \{y \in \mathcal{M} : d(x,y) \leq r\}$ the closed ball of radius $r \geq 0$ around x. Given a probability distribution σ over $\mathcal{N}$ and a set $A \subset \mathcal{N}$ let $\sigma(A) = \sum_{x \in A} \sigma(x)$. We define the *doubling constant* $c(\sigma)$ to be the minimum $c > 0$ for which $\sigma(B_x(2r)) \leq c \cdot \sigma(B_x(r))$, for any $x \in \mathsf{supp}(\sigma)$ and any $r \geq 0$. As we will see, search trough comparisons depends not only on the entropy $H(\mu)$ but also on the topology of μ, as captured by $c(\mu)$.

4 Problem Statement

We now formally define the two problems we study. The first is *content search through comparisons* and the second is the *small-world network design* problem.

4.1 Content Search Through Comparisons

Consider the object set $\mathcal{N}$. Although its embedding in $(\mathcal{M}, d)$ exists, we are constrained by not being able to directly compute object distances; instead, we only have access to a comparison oracle. In particular, we define *greedy content search* as follows. Let t be a target and s an object serving as a starting point. The greedy content search algorithm proposes an object w and asks the oracle to select among s and w the object closest to t, *i.e.*, it evokes $\text{Oracle}(s, w, t)$. This is repeated until the oracle returns something other than s, say w'. If $w' \neq t$, the algorithm repeats these steps, now from w'. If $w' = t$, the search terminates.

Formally, for $k = 1, 2, \dots$, let x_k, y_k be the k-th pair of objects submitted to the oracle: x_k is the *current object*, which greedy content search is trying to improve upon, and y_k is the *proposed object*, submitted to the oracle for comparison with x_k. Let $o_k = \text{Oracle}(x_k, y_k, t) \in \{x_k, y_k\}$ be the oracle's response, and define the *history* of the search up to and including the k-th access as $\mathcal{H}_k = \{(x_i, y_i, o_i)\}_{i=1}^{k}$.

The source object is always one of the first two objects submitted to the oracle, *i.e.*, $x_1 = s$. Moreover, $x_{k+1} = o_k$, *i.e.*, the current object is always the closest to the target so far. The selection of the proposed object y_{k+1} is determined by the history $\mathcal{H}_k$ and the object x_k. In particular, given $\mathcal{H}_k$ and the current object x_k there exists a mapping $(\mathcal{H}_k, x_k) \mapsto \mathcal{F}(\mathcal{H}_k, x_k) \in \mathcal{N}$ such that $y_{k+1} = \mathcal{F}(\mathcal{H}_k, x_k)$, where here we take $x_0 = s \in \mathcal{N}$ (the source/starting object) and $\mathcal{H}_0 = \emptyset$ (*i.e.*, before any comparison takes place, there is no history).

We call the mapping $\mathcal{F}$ the *selection policy* of the greedy content search. In general, we allow the selection policy to be randomized; in this case, the object returned by $\mathcal{F}(\mathcal{H}_k, x_k)$ is a random variable, whose distribution $\Pr(\mathcal{F}(\mathcal{H}_k, x_k) = w)$ for $w \in \mathcal{N}$ is fully determined by $(\mathcal{H}_k, x_k)$. Observe that $\mathcal{F}$ depends on the target t only indirectly, through $\mathcal{H}_k$ and x_k; this is because t is only "revealed" when the search terminates. We say that a selection policy is *memoryless* if it depends on x_k but not on the history $\mathcal{H}_k$.

Our goal is to select an $\mathcal{F}$ that minimizes the number of accesses to the oracle. In particular, given a source object s, a target t and a selection policy $\mathcal{F}$, we define the search cost $C_{\mathcal{F}}(s, t) = \inf\{k : y_k = t\}$ to be the number of proposals to the oracle until t is found. This is a random variable, as $\mathcal{F}$ is randomized; let $\mathbb{E}[C_{\mathcal{F}}(s, t)]$ be its expectation. We thus define the following problem.

> Content Search Through Comparisons (CSTC): Given an embedding of $\mathcal{N}$ into $(\mathcal{M}, d)$ and a demand distribution $\lambda(s, t)$, select $\mathcal{F}$ that minimizes the expected search cost $\bar{C}_{\mathcal{F}} = \sum_{(s,t) \in \mathcal{N} \times \mathcal{N}} \lambda(s, t) \mathbb{E}[C_{\mathcal{F}}(s, t)]$.

Note that, as $\mathcal{F}$ is randomized, the free variable in the above optimization problem is the distribution $\Pr(\mathcal{F}(\mathcal{H}_k, x_k) = w)$.

4.2 Small-World Network Design

In the small-world network design problem the objects in $\mathcal{N}$, embedded in $(\mathcal{M}, d)$, are connected to each other. The network formed by such connections is represented by a directed graph $G(\mathcal{N}, \mathcal{L} \cup \mathcal{S})$, where $\mathcal{L} \cap \mathcal{S} = \emptyset$, $\mathcal{L}$ is the set of *local* edges and $\mathcal{S}$ is the set of *shortcut* edges. The edges in $\mathcal{L}$ satisfy the following property:

Property 1. For every pair of distinct objects $x, t \in \mathcal{N}$ there exists an object u such that $(x, u) \in \mathcal{L}$ and $u \prec_t x$.

I.e., for any x and t, x has a local edge leading to an object closer to t.

A comparison oracle can be used to route a message from s to t over the edges in graph G. In particular, given graph G, we define *greedy forwarding* [9] over G as follows. Let $\Gamma(s)$ be the neighborhood of s, *i.e.*, $\Gamma(s) = \{u \in \mathcal{N} \text{ s.t. } (s, u) \in \mathcal{L} \cup \mathcal{S}\}$. Given a source s and a target t, greedy forwarding sends a message to neighbor w of s that is as close to t as possible, *i.e.*, $w \in \min_{\preccurlyeq_t} \Gamma(s)$. If $w \neq t$, the above process is repeated at w; if $w = t$, greedy forwarding terminates. Property 1 guarantees that greedy forwarding from any source s will eventually reach t: there is always a neighbor closer to t than the object/node forwarding the message. Moreover, if the message is at x, the closest neighbor w can be found using $|\Gamma(x)|$ queries to a comparison oracle.

The edges in $\mathcal{L}$ are called "local" because they are typically determined by object proximity. For example, in the classic paper by Kleinberg [9], objects are arranged uniformly in a rectangular k-dimensional grid—with no gaps—and d is taken to be the Manhattan distance on the grid. Moreover, there exists an $r \geq 1$ such that

$$\mathcal{L} = \{(x, y) \in \mathcal{N} \times \mathcal{N} \text{ s.t. } d(x, y) \leq r\}. \tag{3}$$

Assuming every position in the rectangular grid is occupied, such edges indeed satisfy Property 1. In this work, we *do not* require that edges in $\mathcal{L}$ are given by any locality-based definition like (3); our only assumption is that they satisfy Property 1. Nevertheless, for consistency, we also refer to edges in $\mathcal{L}$ as "local".

Our goal is to select the shortcut edges in $\mathcal{S}$ so that greedy forwarding is as efficient as possible. In particular, assume that we can select no more than β shortcut edges, where β is a positive integer. For S a subset of $\mathcal{N} \times \mathcal{N}$ such that $|S| \leq \beta$, we denote by $C_S(s, t)$ the cost of greedy forwarding, in message hops, for forwarding a message from s to t given that $\mathcal{S} = S$. We allow the selection of shortcut edges to be random: the set $\mathcal{S}$ can be a random variable over all subsets S of $\mathcal{N} \times \mathcal{N}$ such that $|S| \leq \beta$. We denote the distribution of $\mathcal{S}$ by $\Pr(\mathcal{S} = S)$ for $S \subseteq \mathcal{N} \times \mathcal{N}$ such that $|S| \leq \beta$. Given a source s and a target t, let $\mathbb{E}[C_{\mathcal{S}}(s, t)] = \sum_{S \subseteq \mathcal{N} \times \mathcal{N} : |S| \leq \beta} C_S(s, t) \cdot \Pr(\mathcal{S} = S)$ be the expected cost of forwarding a message from s to t with greedy forwarding, in message hops.

We consider again a heterogeneous demand: a source and target object are selected at random from $\mathcal{N} \times \mathcal{N}$ according to a demand probability distribution λ. The *small-world network design problem* can then be formulated as follows.

> Small-World Network Design (SWND): Given an embedding of $\mathcal{N}$ into $(\mathcal{M}, d)$, a set of local edges $\mathcal{L}$, a demand distribution λ, and an integer $\beta > 0$, select a r.v. $\mathcal{S} \subset \mathcal{N} \times \mathcal{N}$, where $|\mathcal{S}| \leq \beta$, that minimizes $\bar{C}_{\mathcal{S}} = \sum_{(s,t) \in \mathcal{N} \times \mathcal{N}} \lambda(s, t) \mathbb{E}[C_{\mathcal{S}}(s, t)]$.

In other words, we wish to select $\mathcal{S}$ so that the cost of greedy forwarding is minimized. Note again that the free variable of SWND is the distribution of $\mathcal{S}$.

5 Main Results

We now present our main results with respect to SWND and CSTC. Our first result is negative: optimizing greedy forwarding is a hard problem.

Theorem 1. *SWND is NP-hard.*

The proof of this theorem can be found in our technical report [7]. In short, the proof reduces DominatingSet to the decision version of SWND. Interestingly, the reduction is to a SWND instance in which (a) the metric space is a 2-dimensional grid, (b) the distance metric is the Manhattan distance on the grid and (c) the local edges are given by (3). Thus, SWND remains NP-hard even in the original setup considered by Kleinberg [9].

The NP-hardness of SWND suggests that this problem cannot be solved in its full generality. Motivated by this, as well as its relationship to content search through comparisons, we focus our attention to the following version of the SWND problem, in which we place additional restrictions to the shortcut edge set $\mathcal{S}$. First, $|\mathcal{S}| = |\mathcal{N}|$, and for every $x \in \mathcal{N}$ there exists *exactly one* shortcut edge $(x, y) \in \mathcal{S}$. Second, the object y to which x connects is selected independently at each x, according to a probability distribution $\ell_x(y)$. *I.e.*, for $\mathcal{N} = \{x_1, x_2, \ldots, x_n\}$, the joint distribution of shortcut edges has the form:

$$\Pr(\mathcal{S} = \{(x_1, y_1), \ldots (x_n, y_n)\}) = \prod_{i=1}^{n} \ell_{x_i}(y_i). \tag{4}$$

We call this version of the SWND problem the *one edge per object* version, and denote it by 1-SWND. Note that, in 1-SWND, the free variables are the distributions ℓ_x, $x \in \mathcal{N}$.

For a given demand λ, recall that μ is the marginal distribution of the demand λ over the target set $\mathcal{T}$, and that for $A \subset \mathcal{N}$, $\mu(A) = \sum_{x \in A} \mu(x)$. Then, for any two objects $x, y \in \mathcal{N}$, we define the *rank* of object y w.r.t. object x as follows:

$$r_x(y) \equiv \mu(B_x(d(x, y))) \tag{5}$$

where $B_x(r)$ is the closed ball with radius r centered at x.

Suppose now that shortcut edges are generated according to the joint distribution (4), where the outgoing link from an object $x \in \mathcal{N}$ is selected according to the following probability:

$$\ell_x(y) \propto \frac{\mu(y)}{r_x(y)}, \tag{6}$$

for $y \in \mathsf{supp}(\mu)$, while for $y \notin \mathsf{supp}(\mu)$ we define $\ell_x(y)$ to be zero. Eq. (6) implies the following appealing properties. For two objects y, z that have the same distance from x, if $\mu(y) > \mu(z)$ then $\ell_x(y) > \ell_x(z)$, *i.e.*, y has a higher probability of being connected to x. When two objects y, z are equally likely to be targets, if $y \prec_x z$ then $\ell_x(y) > \ell_x(z)$. The distribution (6) thus biases both towards objects close to x as well as towards objects that are likely to be targets. Finally, if the metric space $(\mathcal{M}, d)$ is a k-dimensional grid and the targets are uniformly distributed over $\mathcal{N}$ then $\ell_x(y) \propto (d(x,y))^{-k}$. This is the shortcut distribution used by Kleinberg in [9]; (6) is thus a generalization of this distribution to heterogeneous targets as well as to more general metric spaces.

Our next theorem, whose proof is in Section 5.1, relates the cost of greedy forwarding under (6) to the entropy H, the max-entropy $H_{\max}$ and the doubling parameter c of the target distribution μ.

Theorem 2. *Given a demand λ, consider the set of shortcut edges $\mathcal{S}$ sampled according to* (4)*, where $\ell_x(y)$, $x, y \in \mathcal{N}$, are given by* (6)*. Then*

$$\bar{C}_{\mathcal{S}} \leq 6c^3(\mu) \cdot H(\mu) \cdot H_{\max}(\mu).$$

Note that the bound in Theorem 2 depends on λ only through the target distribution μ. In particular, it holds for *any* source distribution ν, and *does not require* that sources are selected independently of the targets t. Moreover, if $\mathcal{N}$ is a k-dimensional grid and μ is the uniform distribution over $\mathcal{N}$, the above bound becomes $O(\log^2 n)$, retrieving thus Kleinberg's result [9].

Exploiting an underlying relationship between 1-SWND and CSTC, we can obtain an efficient selection policy for greedy content search. In particular,

Theorem 3. *Given a demand λ, consider the memoryless selection policy* $\Pr(\mathcal{F}(\mathcal{H}_k, x_k) = w) = \ell_{x_k}(w)$ *where ℓ_x is given by* (6)*. Then*

$$\bar{C}_{\mathcal{F}} \leq 6c^3(\mu) \cdot H(\mu) \cdot H_{\max}(\mu).$$

The proof of this theorem is almost identical, *mutatis mutandis*, to the proof of Theorem 2, and can be found in our technical report [7]. Like Theorem 2, Theorem 3 characterises the search cost in terms of the doubling constant, the entropy and the max-entropy of μ. This is very appealing, given (a) the relationship between $c(\mu)$ and the topology of the target set and (b) the classic result regarding the entropy and accesses to a membership oracle, as outlined in Section 3.

The distributions ℓ_x are defined in terms of the embedding of $\mathcal{N}$ in $(\mathcal{M}, d)$ and the target distribution μ. Interestingly, however, the bounds of Theorem 3 can be achieved if *neither* the embedding in $(\mathcal{M}, d)$ *nor* the target distribution μ are a priori known. In our technical report [7] we propose an adaptive algorithm that asymptotically achieves the performance guarantees of Theorem 3 only through access to a comparison oracle. In short, the algorithm learns the ranks $r_x(y)$ and the target distribution μ as searches through comparisons take place.

A question arising from Theorems 2 and 3 is how tight these bounds are. Intuitively, we expect that the optimal shortcut set $\mathcal{S}$ and the optimal selection policy $\mathcal{F}$ depend both on the entropy of the target distribution and on its doubling constant. Our next theorem, whose proof is in Section 5.2, establishes that this is the case for $\mathcal{F}$.

Theorem 4. *For any integer K and D, there exists a metric space $(\mathcal{M}, d)$ and a target measure μ with entropy $H(\mu) = K\log(D)$ and doubling constant $c(\mu) = D$ such that the average search cost of any selection policy $\mathcal{F}$ satisfies*

$$\bar{C}_{\mathcal{F}} \geq H(\mu)\frac{c(\mu)-1}{2\log(c(\mu))}. \tag{7}$$

Hence, the bound in Theorem 3 is tight within a $c^2(\mu)\log(c(\mu))H_{\max}$ factor.

5.1 Proof of Theorem 2

According to (6), the probability that object x links to y is given by $\ell_x(y) = \frac{1}{Z_x}\frac{\mu(y)}{r_x(y)}$, where $Z_x = \sum_{y\in\mathcal{T}}\frac{\mu(y)}{r_x(y)}$ is a normalization factor bounded as follows.

Lemma 1. *For any $x \in \mathcal{N}$, let $x^* \in \min_{\preccurlyeq_x}\mathcal{T}$ be any object in $\mathcal{T}$ among the closest targets to x. Then $Z_x \leq 1 + \ln(1/\mu(x^*)) \leq 3H_{\max}$.*

Proof. Sort the target set $\mathcal{T}$ from the closest to furthest object from x and index objects in an increasing sequence $i = 1, \ldots, k$, so the objects at the same distance from x receive the same index. Let A_i, $i = 1, \ldots, k$, be the set containing objects indexed by i, and let $\mu_i = \mu(A_i)$ and $\mu_0 = \mu(x)$. Furthermore, let $Q_i = \sum_{j=0}^{i}\mu_j$. Then $Z_x = \sum_{i=1}^{k}\frac{\mu_i}{Q_i}$. Define $f_x(r) : \mathbb{R}^+ \to \mathbb{R}$ as $f_x(r) = \frac{1}{r} - \mu(x)$. Clearly, $f_x(\frac{1}{Q_i}) = \sum_{j=1}^{i}\mu_j$, for $i \in \{1, 2\ldots, k\}$. This means that we can rewrite Z_x as $Z_x = \sum_{i=1}^{k}(f_x(1/Q_i) - f_x(1/Q_{i-1}))/Q_i$. By reordering the terms involved in the sum above, we get $Z_x = f_x(\frac{1}{Q_k})/Q_k + \sum_{i=1}^{k-1}f_x(1/Q_i)(\frac{1}{Q_i} - \frac{1}{Q_{i+1}})$. First note that $Q_k = 1$, and second that since $f_x(r)$ is a decreasing function, $Z_x \leq 1 - \mu_0 + \int_{1/Q_k}^{1/Q_1}f_x(r)dr = 1 - \frac{\mu_0}{Q_1} + \ln\frac{1}{Q_1}$. This shows that if $\mu_0 = 0$ then $Z_x \leq 1 + \ln\frac{1}{\mu_1}$ or otherwise $Z_x \leq 1 + \ln\frac{1}{\mu_0}$. □

Given the set $\mathcal{S}$, recall that $C_{\mathcal{S}}(s,t)$ is the number of steps required by the greedy forwarding to reach $t \in \mathcal{N}$ from $s \in \mathcal{N}$. We say that a message at object v is in phase j if $2^j\mu(t) \leq r_t(v) \leq 2^{j+1}\mu(t)$. Notice that the number of different phases is at most $\log_2 1/\mu(t)$. We can write $C_{\mathcal{S}}(s,t)$ as

$$C_{\mathcal{S}}(s,t) = X_1 + X_2 + \cdots + X_{\log\frac{1}{\mu(t)}}, \tag{8}$$

where X_j are the hops occurring in phase j.Assume that $j > 1$, and let $I = \left\{w \in \mathcal{N} : r_t(w) \leq \frac{r_t(v)}{2}\right\}$. The probability that v links to an object in the set I, and hence moving to phase $j-1$, is $\sum_{w\in I}\ell_{v,w} = \frac{1}{Z_v}\sum_{w\in I}\frac{\mu(w)}{r_v(w)}$. Let $\mu_t(r) = \mu(B_t(r))$ and $\rho > 0$ be the smallest radius such that $\mu_t(\rho) \geq r_t(v)/2$. Since we assumed that $j > 1$ such a $\rho > 0$ exists. Clearly, for any $r < \rho$ we have $\mu_t(r) < r_t(v)/2$. In particular, $\mu_t(\rho/2) < \frac{1}{2}r_t(v)$. On the other hand, since the doubling parameter is $c(\mu)$ we have $\mu_t(\rho/2) > \frac{1}{c(\mu)}\mu_t(\rho) \geq \frac{1}{2c(\mu)}r_t(v)$. Therefore,

$$\frac{1}{2c(\mu)}r_t(v) < \mu_t(\rho/2) < \frac{1}{2}r_t(v). \tag{9}$$

Let $I_\rho = B_t(\rho)$ be the set of objects within radius $\rho/2$ from t. Then $I_\rho \subset I$, so $\sum_{w \in I} \ell_{v,w} \geq \frac{1}{Z_v} \sum_{w \in I_\rho} \frac{\mu(w)}{r_v(w)}$. By triangle inequality, for any $w \in I_\rho$ and y such that $d(y,v) \leq d(v,w)$ we have $d(t,y) \leq \frac{5}{2} d(v,t)$. This means that $r_v(w) \leq \mu_t(\frac{5}{2} d(v,t))$, and consequently, $r_v(w) \leq c^2(\mu) r_t(v)$. Therefore, $\sum_{w \in I} \ell_{v,w} \geq \frac{1}{Z_v} \frac{\sum_{w \in I_\rho} \mu(w)}{c^2(\mu) r_t(v)} = \frac{1}{Z_v} \frac{\mu_t(\rho/2)}{c^2(\mu) r_t(v)}$. By (9), the probability of terminating phase j is uniformly bounded by

$$\sum_{w \in I} \ell_{v,w} \geq \min_v \frac{1}{2c^3(\mu) Z_v} \overset{\text{Lem. 1}}{\geq} \frac{1}{6c^3(\mu) H_{\max}(\mu)} \tag{10}$$

As a result, the probability of terminating phase j is stochastically dominated by a geometric random variable with the parameter given in (10). This is because (a) if the current object does not have a shortcut edge which lies in the set I, by Property 1, greedy forwarding sends the message to one of the neighbours that is closer to t and (b) shortcut edges are sampled independently across neighbours. Hence, given that t is the target object and s is the source object,

$$\mathbb{E}[X_j | s, t] \leq 6c^3(\mu) H_{\max}(\mu). \tag{11}$$

Suppose now that $j = 1$. By the triangle inequality, $B_v(d(v,t)) \subseteq B_t(2d(v,t))$ and $r_v(t) \leq c(\mu) r_t(v)$. Hence, $\ell_{v,t} \geq \frac{1}{Z_v} \frac{\mu(t)}{c(\mu) r_t(v)} \geq \frac{1}{2c(\mu) Z_v} \geq \frac{1}{6c(\mu) H_{\max}(\mu)}$ since object v is in the first phase and thus $\mu(t) \leq r_t(v) \leq 2\mu(t)$. Consequently,

$$\mathbb{E}[X_1 | s, t] \leq 6c(\mu) H_{\max}(\mu). \tag{12}$$

Combining (8), (11), (12) and using the linearity of expectation, we get $\mathbb{E}[C_{\mathcal{S}}(s,t)] \leq 6c^3(\mu) H_{\max}(\mu) \log \frac{1}{\mu(t)}$ and, thus, $\bar{C}_{\mathcal{S}} \leq 6c^3(\mu) H_{\max}(\mu) H(\mu)$. □

5.2 Proof of Theorem 4

Our proof amounts to constructing a metric space and a target distribution μ for which the bound holds. Our construction will be as follows. For some integers D, K, the target set $\mathcal{N}$ is taken as $\mathcal{N} = \{1, \ldots, D\}^K$. The distance $d(x,y)$ between two distinct elements x, y of $\mathcal{N}$ is defined as $d(x,y) = 2^m$, where

$$m = \max\{i \in \{1, \ldots, K\} : x(K-i) \neq y(K-i)\}.$$

We then have the following

Lemma 2. *Let μ be the uniform distribution over $\mathcal{N}$. Then (i) $c(\mu) = D$, and (ii) if the target distribution is μ, the optimal average search cost C^* based on a comparison oracle satisfies $C^* \geq K \frac{D-1}{2}$.*

Before proving Lemma 2, we note that Thm. 4 immediately follows as a corollary.

Proof (of Lemma 2). Part (i): Let $x = (x(1), \ldots x(K)) \in \mathcal{N}$, and fix $r > 0$. Assume first that $r < 2$; then, the ball $B(x,r)$ contains only x, while the ball

$B(x, 2r)$ contains either only x if $r < 1$, or precisely those $y \in \mathcal{N}$ such that $(y(1), \ldots, y(K-1)) = (x(1), \ldots, x(K-1))$ if $r \geq 1$. In the latter case $B(x, 2r)$ contains precisely D elements. Hence, for such $r < 2$, and for the uniform measure on $\mathcal{N}$, the inequality

$$\mu(B(x, 2r)) \leq D\mu(B(x, r)) \tag{13}$$

holds, and with equality if in addition $r \geq 1$.

Consider now the case where $r \geq 2$. Let the integer $m \geq 1$ be such that $r \in [2^m, 2^{m+1})$. By definition of the metric d on $\mathcal{N}$, the ball $B(x, r)$ consists of all $y \in \mathcal{N}$ such that $(y(1), \ldots, y(K-m)) = (x(1), \ldots, x(K-m))$, and hence contains $D^{\min(K,m)}$ points. Similarly, the ball $B(x, 2r)$ contains $D^{\min(K,m+1)}$ points. Hence (13) also holds when $r \geq 2$.

Part (ii): We assume that the comparison oracle, in addition to returning one of the two proposals that is closer to the target, also reveals the distance of the proposal it returns to the target. We further assume that upon selection of the initial search candidate x_0, its distance to the unknown target is also revealed. We now establish that the lower bound on C^* holds when this additional information is available; it holds a fortiori for our more resticted comparison oracle.

We decompose the search procedure into phases, depending on the current distance to the destination. Let L_0 be the integer such that the initial proposal x_0 is at distance 2^{L_0} of the target t, i.e. $(x_0(1), \ldots, x_0(K-L_0)) = (t(1), \ldots, t(K-L_0))$, $x_0(K-L_0+1) \neq t(K-L_0+1)$. No information on t can be obtained by submitting proposals x such that $d(x, x_0) \neq 2^{L_0}$. Thus, to be useful, the next proposal x must share its $(K-L_0)$ first components with x_0, and differ from x_0 in its $(K-L_0+1)$-th entry. Now, keeping track of previous proposals made for which the distance to t remained equal to 2^{L_0}, the best choice for the next proposal consists in picking it again at distance 2^{L_0} from x_0, but choosing for its $(K-L_0+1)$-th entry one that has not been proposed so far. It is easy to see that, with this strategy, the number of additional proposals after x_0 needed to leave this phase is uniformly distributed on $\{1, \ldots D-1\}$, the number of options for the $(K-L_0+1)$-th entry of the target.

A similar argument entails that the number of proposals made in each phase equals 1 plus a uniform random variable on $\{1, \ldots, D-1\}$. It remains to control the number of phases. We argue that it admits a Binomial distribution, with parameters $(K, (D-1)/D)$. Indeed, as we make a proposal which takes us into a new phase, no information is available on the next entries of the target, and for each such entry, the new proposal makes a correct guess with probability $1/D$. This yields the announced Binomial distribution for the numbers of phases (when it equals 0, the initial proposal x_0 coincided with the target).

Thus the optimal number of search steps C verifies $C \geq \sum_{i=1}^{X}(1+Y_i)$, where the Y_i are i.i.d., uniformly distributed on $\{1, \ldots, D-1\}$, and independent of the random variable X, which admits a Binomial distribution with parameters $(K, (D-1)/D)$. Thus using Wald's identity, we obtain that $\mathbb{E}[C] \geq \mathbb{E}[X]\mathbb{E}[Y_1]$, which readily implies (ii). □

Note that the lower bound in (ii) has been established for search strategies that utilize the entire search history. Hence, it is *not* restricted to memoryless search.

6 Conclusions

In this work, we initiated a study of CTSC and SWND under heterogeneous demands, tying performance to the topology and the entropy of the target distribution. Our study leaves several open problems, including improving upper and lower bounds for both CSTC and SWND. Given the relationship between these two, and the NP-hardness of SWND, characterizing the complexity of CSTC is also interesting. Also, rather than considering restricted versions of SWND, as we did here, devising approximation algorithms for the original problem is another possible direction.

Earlier work on comparison oracles eschewed metric spaces altogether, exploiting what where referred to as *disorder inequalities* [5, 11, 12]. Applying these under heterogeneity is also a promising research direction. Finally, trade-offs between space complexity and the cost of the learning phase vs. the costs of answering database queries are investigated in the above works, and the same trade-offs could be studied in the context of heterogeneity.

References

[1] Clarkson, K. L. Nearest-neighbor searching and metric space dimensions. In *Nearest-Neighbor Methods for Learning and Vision: Theory and Practice*, G. Shakhnarovich, T. Darrell, and P. Indyk, Eds. MIT Press, 2006, pp. 15–59.

[2] Cover, T. M., and Thomas, J. *Elements of Information Theory.* Wiley, 1991.

[3] Fraigniaud, P., and Giakkoupis, G. On the searchability of small-world networks with arbitrary underlying structure. In *STOC* (2010).

[4] Fraigniaud, P., Lebhar, E., and Lotker, Z. A doubling dimension threshold $\theta(\log \log n)$ for augmented graph navigability. In *ESA* (2006).

[5] Goyal, N., Lifshits, Y., and Schutze, H. Disorder inequality: a combinatorial approach to nearest neighbor search. In *WSDM* (2008).

[6] Indyk, P., and Motwani, R. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *STOC* (1998), pp. 604–613.

[7] Karbasi, A., Ioannidis, S., and Massoulie, L. Content search through comparisons. Tech. Rep. CR-PRL-2010-07-0002, Technicolor, 2010.

[8] Karger, D., and Ruhl, M. Finding nearest neighbors in growth-restricted metrics. In *SODA* (2002).

[9] Kleinberg, J. The small-world phenomenon: An algorithmic perspective. In *STOC* (2000).

[10] Krauthgamer, R., and Lee, J. R. Navigating nets: simple algorithms for proximity search. In *SODA* (2004).

[11] Lifshits, Y., and Zhang, S. Combinatorial algorithms for nearest neighbors, near-duplicates and small-world design. In *SODA* (2009).

[12] Tschopp, D., and Diggavi, S. N. Approximate nearest neighbor search through comparisons. 2009.

[13] Tschopp, D., and Diggavi, S. N. Facebrowsing: Search and navigation through comparisons. In *ITA workshop* (2010).

[14] White, R., and Roth, R. *Exploratory Search: Beyond the Query-Response Paradigm.* Morgan & Claypool, 2009.