

Top- k Queries over Digital Traces

Yifan Li
yifanli@eecs.yorku.ca
York University

Xiaohui Yu
xhyu@yorku.ca
York University

Nick Koudas
koudas@cs.toronto.edu
University of Toronto

ABSTRACT

Recent advances in social and mobile technology have enabled an abundance of digital traces (in the form of mobile check-ins, association of mobile devices to specific WiFi hotspots, etc.) revealing the physical presence history of diverse sets of entities (e.g., humans, devices, and vehicles). One challenging yet important task is to identify k entities that are most closely associated with a given query entity based on their digital traces. We propose a suite of indexing techniques and algorithms to enable fast query processing for this problem at scale. We first define a generic family of functions measuring the association between entities, and then propose algorithms to transform digital traces into a lower-dimensional space for more efficient computation. We subsequently design a hierarchical indexing structure to organize entities in a way that closely associated entities tend to appear together. We then develop algorithms to process top- k queries utilizing the index. We theoretically analyze the pruning effectiveness of the proposed methods based on a mobility model which we propose and validate in real life situations. Finally, we conduct extensive experiments on both synthetic and real datasets at scale, evaluating the performance of our techniques both analytically and experimentally, confirming the effectiveness and superiority of our approach over other applicable approaches across a variety of parameter settings and datasets.

ACM Reference Format:

Yifan Li, Xiaohui Yu, and Nick Koudas. 2019. Top- k Queries over Digital Traces. In *2019 International Conference on Management of Data (SIGMOD '19)*, June 30–July 5, 2019, Amsterdam, Netherlands. ACM, New York, NY, USA, 18 pages. <https://doi.org/10.1145/3299869.3319857>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org. *SIGMOD '19*, June 30–July 5, 2019, Amsterdam, Netherlands
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-5643-5/19/06...\$15.00
<https://doi.org/10.1145/3299869.3319857>

1 INTRODUCTION

The prevalence of mobile devices, social media, ambient wireless connectivity, and associated positioning technologies have made it possible to record digital traces at an unprecedented rate. Such traces correspond to location sharing through social apps, handshaking with WiFi hot-spots (recording wireless chip MAC address or other device characteristics of a device in proximity of a WiFi network¹) and cellular stations via a mobile device and many other passive/active location capturing scenarios, giving rise to an abundance of *digital traces*. Such traces reveal the presence history of diverse sets of *entities* depending on the application and include humans, devices, etc. At a high level, any digital trace takes the form of a tuple, (*entity, location, timestamp*), recording that an entity (e.g., a person) was present at a physical location (e.g., a restaurant) for the indicated timestamp. Typically location corresponds to physical locations which exhibit a hierarchical structure that is known a priori (e.g., city - district - street - building), and the timestamp is discretized to a tunable atomic unit such as an hour or a minute, depending on the application. For example, the tuple (*Tom, W London, 10 a.m.*) represents the fact that Tom was at the W London hotel during the hour starting at 10 a.m.

A challenging task is to identify the entities that are closely associated with a given query entity utilizing their digital traces. Intuitively two entities are associated if there exists a large overlap on their digital traces. Numerous definitions for what constitutes an overlap are possible; for example, a large overlap in the locations followed by an overlap (proximity) of associated timestamps. Thus, if two entities were present at W London at 10 a.m., they are associated. Similarly if two entities are present at W London, one at 10 a.m. and the other at 11 a.m., they are still associated but possibly less so than the previous two entities. Alternatively, one may take into account the spatial proximity of locations to define association in addition to timestamps. Thus, if two entities are present in the same postal code at the same time, they are associated as well, but probably less so than two entities appearing at the same specific location, say a restaurant, at the same time. It is evident that there are numerous ways to define association of entities given their digital traces, which

¹The WiFi protocol reveals to the access points the MAC address of any WiFi enabled device in the vicinity of the network, even if the device is not connected to the network.

are probably application dependent. As such, we adopt a generic approach and define a class of functions that have generic properties to quantify association. All our subsequent developments in this work hold for this generic class of functions sharing such properties (see Section 2.2).

Given a suitable function to quantify association we are interested to identify the top- k associated entities to a given query entity. Supporting efficient processing of such queries over a large volume of digital traces enables a variety of applications. For example, this assists law enforcement authorities to identify individuals closely related to a person of interest. This research is motivated by our work with authorities enabling post crime investigation utilizing location data collected from mobile devices. Such information is crucial to prove the joint presence of suspects in crime scenes and also their association before and after the crime. For this specific reason, the main interest is to assess association across large sets of digital traces, corroborating the association before and after specific events. For example, in our ongoing work with a large national telecommunications provider in this problem, we will present results involving 30M individual devices with an average of 650K detections by WiFi hotspots each; in addition, each device is present on average at 500 locations during the time ranges of interest for which queries are required. In a different context, the techniques developed herein enable marketers to identify groups of individuals with related behavior in the physical world for more effective advertising. As an example, marketers may utilize associated behavior inside a shopping center (as reported by triangulated WiFi signals of mobile devices) to identify families or couples who are of prime interest for specific types of location-based marketing. Once again association across a large collection of traces enforces a closer bond between the entities involved as opposed to chance encounters.

In the target applications we engage with, the number of entities is in the multiple millions while the number of digital traces is in the billions. As such, techniques that compare the query entity to all other entities are inefficient.

Aiming to provide fast query response times, we propose a suite of indexing structures and algorithms for this problem. At a high level, we consider entities as points in a high-dimensional space with $(location, timestamp)$ pairs of each entity corresponding to a dimension. The basic idea of our approach consists of two parts: (1) transforming an entity's digital traces into a lower-dimensional space for more efficient computation; this lower-dimensional representation also allows the ordering of entities along each dimension, making it possible to build an index structure; (2) constructing an index that groups the entities in a hierarchical fashion using this lower-dimensional representation, so that associated entities tend to appear in the same group, enabling effective pruning.

We utilize MinHash techniques [7] to compute a signature for each entity. When a spatial hierarchy on locations is present we do so at each level of the spatial hierarchy, resulting in a list of signatures for each entity. The size of each signature depends on the number of hash functions utilized and can be considered as the dimensionality of this lower-dimensional space. The list of signatures for each entity are subsequently indexed with a tree structure. The guiding principal of this index is to group the entities based on their signatures at each level such that for a given query entity, only a small portion of the branches in the tree have to be explored; the remaining branches are guaranteed not to contain the top- k results and can thus be safely discarded. This is made possible by assessing a signature for each group of entities at a tree node, which serves as the basis for estimating bounds on the association between the query entity and the entities in the subtree rooted at this node. The index naturally supports incremental updates. We then develop algorithms to process top- k queries using the index.

We also develop and present a model for mobility which we validate against real data at scale. Utilizing this model, we subsequently present a thorough analysis of the pruning effectiveness of the proposed method. Our results reveal that the proposed technique has strong pruning capabilities, limiting the scope of search to only a small portion of all available entities. We also validate our model for pruning effectiveness against real mobility traces and demonstrate its accuracy both analytically and experimentally.

Experiments are conducted on both synthetic and real datasets at scale to (1) compare the performance of the proposed method against that of baseline methods, and (2) conduct a sensitivity analysis of the proposed method with respect to varying parameters of interest (e.g., number of hash functions, data characteristics). Our results demonstrate orders of magnitude performance improvement over other applicable approaches.

In summary, in this paper we make the following contributions:

- Motivated by real-life applications with telecommunications providers, we formally define the problem of Top- k query processing over digital traces. To the best of our knowledge, our work is the first to address this important problem.
- We develop a suite of novel data transformation and indexing techniques as well as the corresponding search methodologies, which demonstrate strong pruning capabilities, allowing us to focus the search only on a small portion of the space.
- We analytically and experimentally quantify the pruning effectiveness of our methods utilizing models of human mobility patterns.

- We perform extensive experiments on both real and synthetic data at scale to thoroughly study the performance of the proposed method, confirming its effectiveness and superiority over other approaches across a variety of settings.

The rest of the paper is organized as follows. Section 2 formally defines the problem of top- k query over digital traces and other assist terms. Section 3 describes the approach, including the data transformation principle, the data organization technique, and the complexity of indexing. In Section 4, we prove the early termination condition of the proposed approach and give the search algorithm. In Section 5, we analytically quantify the pruning effectiveness of the approach. In Section 6, we present the experiment results across a variety of settings. Section 7 provides an overview of related work, and Section 8 concludes this paper.

2 PRELIMINARIES

In this section, we define the terms that are required for the subsequent discussion, and formally define the problem of top- k query processing over digital traces.

2.1 Terminology

The locations we consider are spatial and thus exhibit a hierarchical structure (e.g., city - district - street - building). We assume that a description of the hierarchical structure of locations is available via a tree structure (referred to as *sp-index*) that organizes locations from coarsest to finest. Nodes in this tree are referred to as *spatial units*. We assume (as per previous work [47]) that spatial units remain unchanged over extended times periods, and thus the sp-index can be considered fixed for the period of interest.

Without loss of generality, we assume that the spatial units at the same level of the sp-index are non-overlapping. We label the levels of spatial units from 1 (for the root of the tree) to m (for the lowest level in the tree). For a spatial unit l , we use $\text{pat}(l)$ to denote the parent unit of l on the sp-index.

At the lowest level of the tree are *base spatial units*, the atomic locations in digital traces in which entities can be present. Examples of a base spatial unit include a supermarket, a restaurant, etc. All base spatial units form a set \mathcal{L} .

We assume that timestamp is discretized in base temporal units (e.g., hour). The combination of a base temporal unit and a base spatial unit is referred to as a *spatial-temporal cell* (or ST-cell). We use the associated base temporal unit and base spatial unit to denote an ST-cell, e.g., t_1l_1 . An ST-cell is the atomic unit where entities can be present. All possible such combinations form an ST-cell set \mathcal{S} .

We enhance the notion of a digital trace associated with entity e to make it suitable for a multilevel sp-index.

Definition 2.1 (Presence Instance). A presence instance (PI) p of an entity is characterized by a five attribute tuple, $p = (e, tid, level, path, pd)$, where

- e is the associated entity to p ,
- tid is the id of the sp-index where p belongs (tid is necessary when multiple sp-index trees exist),
- $level$ is the level in the sp-index where p exists,
- $path = [node_1, node_2, \dots, node_{level}]$ is the list of nodes in the sp-index on the path from the root to the node that reflects the location associated with p , and
- pd is a continuous time period associated with p ; it is in the format $[start\ time, end\ time]$.

Typically $start\ time$ is the same as $timestamp$. In some applications, such as WiFi proximity sensing of MAC addresses, $end\ time$ is obtained by the time of the last probe of the device MAC address to the WiFi network. In some other applications, such as social media check-ins, the end time is estimated based on the average time individuals spend in the venue (obtained from services such as Google Maps).

Definition 2.2 (Digital Trace). The set of PIs associated with entity e forms the digital trace of e , \mathcal{P}_e .

The overlap between the digital traces of two entities, *Adjoint Presence Instance*, is defined as follows:

Definition 2.3 (Adjoint Presence Instance). Given two PIs, $p_a = (e_a, tid_a, level_a, path_a, pd_a), p_b = (e_b, tid_b, level_b, path_b, pd_b)$, if $tid_a = tid_b, pd_a \cap pd_b \neq \emptyset$, then e_a and e_b form an adjoint presence instance (AjPI) $p_{ab} = (\{e_a, e_b\}, tid_{ab}, level_{ab}, path_{ab}, pd_{ab})$, where:

- $tid_{ab} = tid_a = tid_b$,
- $level_{ab} = |path_{ab}|$, denoting the finest level of the AjPI, which is equal to the number of common ancestors in the sp-index,
- $path_{ab} = path_a \cap path_b$, which is the set of common ancestors of the two PIs, and
- $pd_{ab} = pd_a \cap pd_b$, the intersection of the two time periods.

Each pair of entities, say e_a and e_b , own zero or more AjPIs, forming set \mathcal{P}_{ab} . The definition can be naturally extended to adjoint presence instances of multiple entities.

An AjPI specifies a spatio-temporal co-occurrence of two entities, and thus reveals a potential association between the entities. Such an association is defined as a function of the corresponding presence instances and adjoint presence instances of each entity pair, as outlined next.

2.2 Problem definition

One important but challenging task is to discover all entities that are closely associated with a given entity. Since there may exist many ways to quantify association, we define

a generic class of scoring functions that share some commonly desired properties. While the particular choice of the function varies depending on the application, our approach would apply as long as the function exhibits those generic properties.

For an AjPI p_{ab} , the scoring function $f(p_{ab})$ has the following properties:

- The range of $f \in [0, 1]$,
- $\forall e_c, f(p_{ab}) \geq f(p_{ac})$ if $p_{ab}.pd.length \geq p_{ac}.pd.length \wedge p_{ab}.level \geq p_{ac}.level$.

The first property ensures that the score is properly normalized; the second property gives AjPIs at finer spatial units and for longer durations a higher score. These properties capture the intuition that the association between two entities is higher when corresponding digital traces match closely at locations (i.e., appear at finer levels of the sp-index, say at the same restaurant vs. in the same city) and their temporal co-occurrence is longer.

Let \mathcal{P}_{ab} be the set of AjPIs formed by e_a and e_b . The overall score for this set is defined as

$$F(\mathcal{P}_{ab}) = \sum_{p_{ab} \in \mathcal{P}_{ab}} f(p_{ab}), \quad (1)$$

which has to be further normalized to take into consideration the individual behaviors of e_a and e_b . It is evident that the AjPI to an entity with many PIs is less interesting than that with an entity having few PIs. Therefore, we define a scoring function for individual PI p_a , which is considered as a special case of the AjPI score, i.e., $f(p_a) = f(p_{aa})$. The score for the set of PI \mathcal{P}_a is:

$$F(\mathcal{P}_a) = \sum_{p_a \in \mathcal{P}_a} f(p_a) \quad (2)$$

Clearly, $\forall e_a, \forall e_b, F(\mathcal{P}_a) \geq F(\mathcal{P}_{ab}), F(\mathcal{P}_b) \geq F(\mathcal{P}_{ab})$.

Intuitively, closely associated entities are those who have a large presence instance overlap, i.e., more adjoint presence instances and less total presence instances for either entity. Thus we define the association degree between two entities e_a and e_b as

$$d(e_a, e_b) = G(F(\mathcal{P}_{ab}), F(\mathcal{P}_a), F(\mathcal{P}_b)) \quad (3)$$

where G can be any function satisfying the following constraints:

- $d(e_a, e_b)$ has range $[0, 1]$,
- $\forall e_c, d(e_a, e_b) \geq d(e_a, e_c)$ if $F(\mathcal{P}_b) \leq F(\mathcal{P}_c) \wedge F(\mathcal{P}_{ab}) \geq F(\mathcal{P}_{ac})$,
- $\forall e_c, d(e_a, e_b) \geq d(e_a, e_c)$ if $(\mathcal{P}_b - \mathcal{P}_c) \neq \emptyset \wedge (\mathcal{P}_b - \mathcal{P}_c) \subseteq \mathcal{P}_a$.

We let the associate degree be a generic function instead of a particular measure (e.g., Jaccard Distance), as the most suitable measure may vary in different application scenarios. A generic approach allows our Top- k algorithm to utilize the

measure that makes the most sense in each case as long as it follows set properties. Therefore, all subsequent discussions of Top- k query processing are for measures satisfying the constraints of G . We provide a recommended form of G in Section 6.1, where We also demonstrate that such a form simulates other widely-adopted measures accurately.

Let \mathcal{E} be the set of all entities. The problem of identifying the k most associated entities (entities with the highest association degree) to a given query entity is defined as:

Definition 2.4 (Top- k Query over Digital Traces). Given a query entity e_p and association degree measure d , the top- k query over digital traces is to return the set of entities Q_k such that $Q_k \subseteq \mathcal{E} - \{e_p\}$, $|Q_k| = k$ and $\forall e_q \in Q_k, \forall e_t \in (\mathcal{E} - \{e_p\} - Q_k), d(e_p, e_q) \geq d(e_p, e_t)$, where $1 \leq k < |\mathcal{E}|$.

3 OUR APPROACH

A brute-force approach to answer top- k queries involves computing the association degree between the query entity and all other entities. Clearly, the cost can be prohibitive, as the number of entities are often in the millions and the number of digital traces in billions in our target applications. As such, we introduce a data structure, called the *MinSigTree*, that indexes entities based on their presence instances, facilitating efficient pruning of entities to be examined during the search for top- k answers.

As a high-level overview, we first organize the PIs of each entity as a sequence of ST-cell sets. Then we construct a list of *signatures* for each entity which can be considered as summaries of the entity's PIs. Subsequently we construct the *MinSigTree* that groups closely associated entities together based on their signatures.

3.1 Data representation

Real-world digital traces in their raw format may require pre-processing. For instance, we may need to conduct time-zone normalization (e.g., all time-stamps normalized to GMT), build the sp-index from the longitude and latitude coordinates of places using maps (e.g., Open Street Maps) and other information (e.g., community area boundaries), and align data from different sources with varying sampling frequencies. After pre-processing, the next step is to organize the data by entity so that the presence instances of an entity at each sp-index level and the resulting association degree between entity pairs can be computed efficiently.

We build a sequence of ST-cell sets for each entity, where the length of the sequence equals the height of the sp-index, m . The sequence of sets for entity e_a is denoted as seq_a , and the i -th set in seq_a , $i \in [1, m]$, corresponding to the level i of the sp-index, is denoted as seq_a^i .

seq_a^m , for the lowest level of the sp-index, can be obtained directly from e_a 's digital trace, i.e., for an ST-cell s , $s \in seq_a^m$

iff e_a is present at s . For other levels, ST-cell set seq_a^i , $i \in [1, m]$ is built from set seq_a^{i+1} . For an ST-cell $s = t_z l_x$, $s \in seq_a^i$ iff $\exists s' = t_z l_y$, s.t. $s' \in seq_a^{i+1}$ and $l_x = pat(l_y)$.

EXAMPLE 3.1. Let L_1, L_2, L_3 , and L_4 be four base spatial units, and $pat(L_1)=pat(L_2)=L_5$, $pat(L_3)=pat(L_4)=L_6$, $m = 2$. Assume that entity e has presence in base spatial unit L_1 at time T_1 , and L_3 at time T_2 , then $seq_e^2 = \{T_1 L_1, T_2 L_3\}$. Since $T_1 L_1 \in seq_e^2$ and $L_5 = pat(L_1)$, $T_1 L_5 \in seq_e^1$, similarly, $T_2 L_6 \in seq_e^1$. Finally we have $seq_e^1 = \{T_1 L_5, T_2 L_6\}$.

The ST-cell set sequence not only records the PIs of a single entity at any level of the sp-index, but reflects the AjPI between entities as well. If entities e_a and e_b form AjPIs at level i , then $seq_a^i \cap seq_b^i \neq \emptyset$.

3.2 Data organization

Although ST-cell set sequences facilitate the direct retrieval of PIs of each entity at any level, a brute-force approach would have to explore the whole search space of all entities to identify the top- k answers, which is still too expensive. We thus propose to group entities based on their common ST-cells to allow efficient pruning of the search space. Note that the number of ST-cells in which an entity is present could vary vastly from entity to entity (e.g., one short occurrence vs. frequent and prolonged visits to multiple locations). If we consider each ST-cell as a dimension, conceptually all entities can be considered as bit vectors in a very high-dimensional space where each bit indicates whether that entity is present in the ST-cell. However, if they are physically treated as such, the storage and computation cost can be prohibitive when the number of ST-cells is large. To allow more effective indexing, we employ a family of hash functions to map ST-cell sets into a lower-dimensional space. This is achieved by assigning each entity a *signature* at each level, with each value in the signature acting as a summary of the entity's PIs, and then grouping entities by their signatures.

3.2.1 Signature. We use n_h hash functions to map an ST-cell set into a vector in a n_h -dimensional space, where each element of the vector is a hash value. Since each entity is associated with an ST-cell set sequence of length m , for an arbitrary entity e_a , we obtain m vectors, which form a list of signatures, sig_a , and we use sig_a^i to denote the i -th signature in sig_a (corresponding to level i in the sp-index), and $sig_a^i[u]$ to denote the u -th hash value in sig_a^i , where $u \in [1, n_h]$.

The way we compute signatures for each entity is similar to that for MinHash [7]. A hash function h_u maps each ST-cell to a value in the range $[0, |S| - 1]$. The u -th value in the signature sig_a^i corresponds to the minimal hash value produced by h_u across all ST-cells in seq_a^i , i.e., $sig_a^i[u] = \perp_u^i = \min_{s \in seq_a^i} \{h_u(s)\}$.

The hash functions employed above should satisfy that, for ST-cell $s = t_z l_x$ and $s' = t_z l_y$, if $l_x = pat(l_y)$, then $h_u(s) \leq h_u(s')$. Let C_x be the child spatial unit set of l_x , the above constraint is satisfied by assigning $h_u(t_z l_x) = \min_{l_c \in C_x} \{h_u(t_z l_c)\}$. The constraint guarantees the following property which makes signatures at different levels comparable:

THEOREM 3.1. For any entity $e \in \mathcal{E}$, $sig_e^i[u] \leq sig_e^{i+1}[u]$ always holds.

The proof follows from above constraint and is omitted for brevity.

EXAMPLE 3.2. Consider the following hash table:

	$T_1 L_1$	$T_2 L_1$	$T_1 L_2$	$T_2 L_2$	$T_1 L_3$	$T_2 L_3$	$T_1 L_4$	$T_2 L_4$
h_1	2	8	5	1	4	6	7	3
h_2	8	3	6	5	4	1	2	7

Assume that the four base spatial units follow relations indicated in Example 3.1. Let e_a, e_b, e_c and e_d be four entities with the following ST-cell set sequence:

e_a	$\langle \{T_1 L_5, T_2 L_5\}, \{T_1 L_2, T_2 L_1\} \rangle$
e_b	$\langle \{T_1 L_5, T_2 L_5\}, \{T_1 L_1, T_2 L_2\} \rangle$
e_c	$\langle \{T_1 L_6, T_2 L_5\}, \{T_1 L_3, T_2 L_1\} \rangle$
e_d	$\langle \{T_1 L_6, T_2 L_6\}, \{T_1 L_4, T_2 L_4\} \rangle$

We first build sig_a^2 given $seq_a^2 = \{T_1 L_2, T_2 L_1\}$. Since $h_1(T_1 L_2) = 5$, $h_1(T_2 L_1) = 8$, we have $sig_a^2[1] = 5$; similarly, since $h_2(T_1 L_2) = 6$, $h_2(T_2 L_1) = 3$, we have $sig_a^2[2] = 3$. Therefore, $sig_a^2 = \langle 5, 3 \rangle$. Subsequently, we build sig_a^1 given $seq_a^1 = \{T_1 L_5, T_2 L_5\}$. Since $L_5 = pat(L_1) = pat(L_2)$, $h_1(T_1 L_5) = \min\{h_1(T_1 L_1), h_1(T_1 L_2)\} = 2$; similarly we have $h_1(T_2 L_5) = 1$, $h_2(T_1 L_5) = 6$, $h_2(T_2 L_5) = 3$. Therefore, $sig_a^1 = \langle 1, 3 \rangle$. We build signatures for all entities and finally obtain the following signature table:

e_a	$\langle \langle 1, 3 \rangle, \langle 5, 3 \rangle \rangle$
e_b	$\langle \langle 1, 3 \rangle, \langle 1, 5 \rangle \rangle$
e_c	$\langle \langle 1, 2 \rangle, \langle 4, 3 \rangle \rangle$
e_d	$\langle \langle 3, 1 \rangle, \langle 3, 7 \rangle \rangle$

As each value in a signature is obtained by hashing all ST-cells in the corresponding set to a certain domain, it can be considered as a summary of the ST-cell set. Hash values sig_a^i enables to determine certain facts regarding the ST-cells contained in the set seq_a^i .

THEOREM 3.2. For signature sig_a^i ($i \in [1, m]$) and an ST-cell s , if $\exists u \in [1, n_h]$ s.t. $sig_a^i[u] > h_u(s)$, then $s \notin seq_a^i$.

PROOF. If $s \in seq_a^i$, then $sig_a^i[u] \leq h_u(s)$. From Theorem 3.1 we know that $sig_a^i[u] \leq sig_a^m[u]$, and thus $sig_a^i[u] \leq h_u(s)$, which contradicts the condition. \square

Via Theorem 3.2, for a given signature sig , we can obtain a *pruned set* of ST-cells such that entities bearing sig are guaranteed not to have presence in those ST-cells. We use

\mathcal{PS}_a^i to denote the pruned set based on signature sig_a^i . This property will be explored in pruning the search space while computing the top- k answers.

3.2.2 MinSigTree. We design MinSigTree, an m -level tree structure, which groups entities sharing similar signatures together. Each node in the MinSigTree has at most n_h child nodes (with n_h being the number of hash functions used while computing signatures), each leaf node contains a set of entities, and each entity is contained in a single leaf node. If node N contains entity e_a , we consider all ancestor nodes of N to conceptually contain e_a as well to ease notation (but no physical storage is involved). For node N containing entity set \mathcal{E}_N , we compute a group-level signature SIG_N summarizing the PIs of all entities in \mathcal{E}_N .

Assuming that there is a virtual root node (at level 0), we use Algorithm 1 to build the MinSigTree.

Algorithm 1 Building MinSigTree

Input: Entity set \mathcal{E} , signatures of all entities
Output: MinSigTree

- 1: **Initialization:** MinSigTree root to contain all entities; root enqueued to priority queue Q ;
- 2: **for** $N : Q$ **do**
- 3: \mathcal{G} = sets of entities in N grouped by routing index;
- 4: **for** $g : \mathcal{G}$ **do**
- 5: u = routing index of g ;
- 6: \mathcal{E}_g = entities contained in g ;
- 7: SIG_g = group-level signature of \mathcal{E}_g ;
- 8: N_u = node($u, SIG_g[u], \mathcal{E}_g$);
- 9: $N.addChild(N_u)$;
- 10: **if** $i \neq m$ **then**
- 11: enqueue N_u to Q ;
- 12: **else**
- 13: insert \mathcal{E}_g to N_u ;
- 14: **end if**
- 15: **end for**
- 16: **end for**
- 17: **return** MinSigTree;

As Step 1, we fetch the level 1 signature of every entity, $(sig_1^1, sig_2^1, \dots, sig_{|\mathcal{E}|}^1)$, and divide these signatures into n_h groups. This is done in a way such that entity e_a is routed to the u -th group, if $\forall v \in [1, n_h](v \neq u), sig_a^1[u] \geq sig_a^1[v]$, i.e., u is the position of the maximal hash value in sig_a^1 (ties are broken arbitrarily). We call u the *routing index* of the u -th group (Line 3).

Step 2 involves computing a group-level signature for each node (Lines 5 - 7). Assume that node N_u contains entity set \mathcal{E}_{N_u} . Then the signature of N_u , SIG_{N_u} , can be computed by $SIG_{N_u}[v] = \min_{e \in \mathcal{E}_{N_u}} \{sig_e^1[v]\}$, where $v \in [1, n_h]$. The

newly created nodes are then inserted as the children of the root (Lines 8 - 9).

The second step computes a group-level signature for each node in a way that any hash value in SIG_{N_u} is no greater than the corresponding hash values in the signatures of entities in \mathcal{E}_{N_u} . With signatures computed this way, we can obtain a group-level pruned set,

$$\mathcal{PS}_{N_u} = \bigcap_{e \in \mathcal{E}_{N_u}} \mathcal{PS}_e^1. \quad (4)$$

All entities in \mathcal{E}_{N_u} are guaranteed not to have presence in the ST-cells contained in \mathcal{PS}_{N_u} . Note that there is no need to store the pruned set of each node, as it can be inferred from the group-level signature.

In practice, however, storing the entire signature of a node imposes space overhead. It is evident from the grouping strategy that given a group-level signature SIG_N with routing index u , $\forall v \in [1, n_h](v \neq u), SIG_N[u] \gg SIG_N[v]$. From Theorem 3.2 it follows that the pruned set of a signature is mainly decided by the large hash values in the signature. Thus one can materialize $SIG_N[u]$ only, instead of SIG_N . This greatly reduces storage costs at the expense of pruning effectiveness. We explore this further in Section 4.1.

Consider the signature table in Example 3.2. We fetch all level 1 signatures and group entities accordingly. As a result, group $N_1 = \{e_d\}$ with routing index 1, $N_2 = \{e_a, e_b, e_c\}$ with routing index 2, and $SIG_{N_1} = \langle 3, 1 \rangle$, $SIG_{N_2} = \langle 1, 2 \rangle$.

The grouping principle of Step 1 is designed in a way to prevent the group-level signature from becoming too small. For example, if e_c and e_d were to be grouped together, the group-level signature would be $\langle 1, 1 \rangle$, which would not be greater than any hash values and the pruned set would thus be empty.

Now we have grouped entities at the first level of the MinSigTree based on the level 1 signatures of all entities. However, the level 1 signatures reveal only the PI patterns at the highest/coarsest sp-index level. Intuitively, entities belonging to different groups at level 1 are guaranteed not to be strongly associated, but entities belonging to the same group may still have different PI patterns at a finer-level. For example, if two people both visited New York City, but one in Manhattan and the other in Brooklyn, their PIs are different in the district level. Therefore, we need to further partition the entities based on their finer-level signatures.

For node N_u at level i , if $i \neq m$, i.e. N_u is not at the leaf level, we fetch the level $(i + 1)$ signatures of entities in \mathcal{E}_{N_u} (Lines 10 - 11), group \mathcal{E}_{N_u} by the routing indexes, compute a signature for each new group, and add these newly created nodes as children of N_u . We repeat this process until we reach the leaf level. If an entity belongs to node N_f at the leaf level, we insert this entity to N_f (Lines 12 - 13).

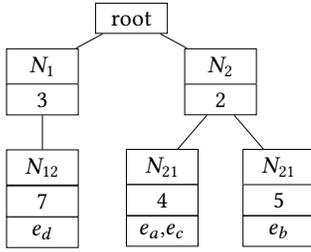


Figure 1: A sample MinSigTree

In the above example, group $N_1 = \{e_d\}$, $N_2 = \{e_a, e_b, e_c\}$. Since $sig_d^2[2] > sig_d^2[1]$, e_d belongs to the sub-group with routing index 2, i.e. $N_{12} = \{e_d\}$; similarly, we have $N_{21} = \{e_a, e_c\}$, $N_{22} = \{e_b\}$. Group level signatures are: $SIG_{N_{12}} = \langle 3, 7 \rangle$, $SIG_{N_{21}} = \langle 4, 3 \rangle$, $SIG_{N_{22}} = \langle 1, 5 \rangle$. The overall MinSigTree is given in figure 1.

By partitioning entities recursively at each level, each group will end up containing entities that are similar at all sp-index levels and thus very likely to result in high association degrees with each other. In addition, the partitioning strategy guarantees the following property.

THEOREM 3.3. *If N_a is an ancestor node of N_d , then $\mathcal{PS}_{N_a} \subseteq \mathcal{PS}_{N_d}$.*

The proof follows from the building process and is omitted for brevity.

The cost of index construction is analyzed in Appendix B.

3.2.3 Incremental update. Similar to the building process, the MinSigTree also supports incremental update. More specifically, after building the MinSigTree, we can deal with new records of entity e in four steps: (1) locate the leaf, say N_e , containing e ; (2) remove e from N_e (if N_e becomes empty, safely remove N_e from the MinSigTree); (3) compute new signatures of e ; (4) insert e to the new node N'_e based on the new signatures. During this process, only nodes in the path from root to the leaf containing e and the leaf to insert e are modified, and thus the complexity of update is linear w.r.t. the height of the MinSigTree.

Bulk updates are also naturally supported. Here we mainly introduce the steps involving signature re-computation. Given a set of entities to be updated, \mathcal{E}_u , we compute the signatures of e for each $e \in \mathcal{E}_u$, and use \mathcal{E}_N to denote the set of entities in \mathcal{E}_u to be inserted to node N . Subsequently, we compute the group level signature of \mathcal{E}_N , $SIG_{\mathcal{E}_N}$, and update the signature of N , SIG_N , as $SIG_N := \min\{SIG_N, SIG_{\mathcal{E}_N}\}$. Updating MinSigTree is discussed in detail in Section 6.8.

4 QUERY PROCESSING

The MinSigTree partitions entities to groups enabling an efficient search strategy for top- k query processing. We present an algorithm for top- k query evaluation utilizing the proposed structure.

4.1 Early termination

Given a query entity e_q with ST-cell set sequence seq_q , the basic search strategy unfolds by computing an upper bound on the association degree between e_q and each candidate node of the MinSigTree, and then progressively visit the node with the maximal upper bound until the top- k answers are identified. We outline how to compute and gradually tighten the upper bound of a node in order to prune more entities and terminate the search earlier.

We use \mathcal{S}_q to denote seq_q^m , which contains all ST-cells in which e_q is present. For each node N in the search path, we determine an upper bound, UB_N , on the association degree between e_q and entities in node N , to decide whether to continue searching or terminate.

THEOREM 4.1. *Let \mathcal{PS}_N be the pruned set of node N , and e_v be an artificial entity with ST-cell set $\mathcal{S}_v = \mathcal{S}_q - \mathcal{PS}_N$. Then $UB_N = d(e_v, e_q)$.*

PROOF. Because $\mathcal{S}_v \subseteq \mathcal{S}_q$, we have $\mathcal{P}_v \subseteq \mathcal{P}_q$, where \mathcal{P}_v and \mathcal{P}_q are the PI sets of e_v and e_q respectively. Thus, $\mathcal{P}_{vq} = \mathcal{P}_v$, where \mathcal{P}_{vq} is the set of AjPIs between e_v and e_q .

Let \mathcal{E}_N denote the set of entities contained in N . Since $\forall e_p \in \mathcal{E}_N, \mathcal{S}_p \cap \mathcal{S}_q \subseteq \mathcal{S}_v$, we have $\mathcal{P}_{pq} \subseteq \mathcal{P}_v = \mathcal{P}_{vq}$. Therefore, $F(\mathcal{P}_{pq}) \leq F(\mathcal{P}_{vq})$.

$\forall e_p \in \mathcal{E}_N$, if $\mathcal{P}_v \subseteq \mathcal{P}_p$, then $F(\mathcal{P}_p) \geq F(\mathcal{P}_v)$, and thus $d(e_v, e_q) \geq d(e_p, e_q)$; otherwise, we have $(\mathcal{P}_v - \mathcal{P}_p) \neq \emptyset$ and $(\mathcal{P}_v - \mathcal{P}_p) \subseteq \mathcal{P}_q$, and thus $d(e_v, e_q) \geq d(e_p, e_q)$ according to the definition of d given in Equation (3). \square

In practice, it is not required to compute the entire pruned set of a node; instead, it can be conducted in a more efficient way. Let u be the routing index of the node N . For an ST-cell $s \in \mathcal{S}_q$, if $h_u(s) < SIG_N[u]$, it is guaranteed that $s \in \mathcal{PS}_N$. All such ST-cell s form the partial pruned set \mathcal{PPS}_N , which can be used to create the artificial entity e'_v with ST-cell set $\mathcal{S}'_v = \mathcal{S}_q - \mathcal{PPS}_N$. Evidently, \mathcal{S}'_v may be slightly larger than \mathcal{S}_v , which leads to a larger upper bound UB'_N . However, as the hash value at the routing index is in general far larger than the other hash values in the group-level signature, UB'_N is expected to be very close to UB_N . In the experiment we utilize partial pruned sets to evaluate performance; details are given in Section 6.

As discussed in Theorem 3.3, the pruned set of a descendant node contains that of its ancestor nodes. Therefore, in a specific branch of the MinSigTree, the upper bound can be gradually tightened before we reach the leaf nodes and check the contained entities.

4.2 Search algorithm

The search algorithm based on the early termination condition is given in Algorithm 2. We initialize the result as a priority queue sorted by association degree (from high

to low), and start the search from the root of MinSigTree (Line 1) whose upper bound is set to 1. We fetch the node with maximal UB in the candidate list (Line 3) and insert all its child nodes into the candidate list (Line 8). Once we reach a leaf node, we calculate the exact association degree between the query node and all entities in this node, and update the result accordingly (Lines 10 - 13). The process terminates when either (1) we have identified k entities, and the association degree between any of these k entities and the query entity is no less than the maximal UB of the remaining candidates (Lines 4 - 5), or (2) all leaves have been explored (Line 16). It is worth noting that the algorithm is applicable to all association degree measures as long as they satisfy the constraints of d specified in Section 2.2.

Algorithm 2 Top- k query processing

Input: MinSigTree T , k , query entity e , measure d
Output: k most associated entities to e

- 1: **Initialization:** Result = $\{\}$, Candidate = $\{\text{root of } T\}$;
- 2: **while** Candidate $\neq \emptyset$ **do**
- 3: N = node with maximal UB in Candidate;
- 4: **if** Result.minKey $\geq N.UB$ and Result.size $= k$ **then**
- 5: return Result;
- 6: **end if**
- 7: **if** N is not leaf **then**
- 8: Candidate = Candidate \cup $\{\text{all child nodes of } N\}$;
- 9: **else**
- 10: \mathcal{E}_N = entities contained in N ;
- 11: **for** $e' : \mathcal{E}_N$ **do**
- 12: $s = d(e, e')$;
- 13: Result.update(e', s)
- 14: **end for**
- 15: **end if**
- 16: **end while**
- 17: **return** Result;

EXAMPLE 4.1. Let us again consider the MinSigTree in Figure 1 as an example. We use a Dice similarity-based function as the measure of association degree: $d(e_i, e_j) = 0.1 \times \frac{|seq_i^1 \cap seq_j^1|}{|seq_i^1| + |seq_j^1|} + 0.9 \times \frac{|seq_i^2 \cap seq_j^2|}{|seq_i^2| + |seq_j^2|}$. Let e_c be the query entity, and the Top-1 result is desired. As indicated in Example 3.2, $seq_c^2 = \{T_1L_3, T_2L_1\}$, $h_1(T_1L_3) = 4$, $h_2(T_1L_3) = 4$, $h_1(T_2L_1) = 8$, $h_2(T_2L_1) = 3$. We start the search from the root. For node N_1 , as $3 < h_1(T_1L_3)$ and $3 < h_1(T_2L_1)$, we have $\mathcal{PPS}_{N_1} = \emptyset$, and thus the upper bound of N_1 , $UB_{N_1} = 1$; similarly $UB_{N_2} = 1$. The candidate queue is $(1 : \{N_1, N_2\})$. Since there is no remaining node at level 1, we dequeue N_1 , the only child of which is N_{12} . As $7 > h_2(T_1L_3)$ and $7 > h_2(T_2L_1)$, we have $\mathcal{PPS}_{N_{12}} = \{T_1L_2, T_2L_1\}$, $UB_{N_{12}} = 0.1 \times 1 + 0.9 \times 0 = 0.1$, where 1 is the UB of the

parent node of N_{12} , and 0 corresponds to the fact that both query ST-cells are contained in $\mathcal{PPS}_{N_{12}}$. We then dequeue N_2 . The first child of N_2 is N_{21} . As $4 < h_1(T_1L_3)$ and $4 < h_1(T_2L_1)$, $UB_{N_{21}} = 1$. For node N_{22} , as $5 > h_2(T_1L_3)$ and $5 > h_2(T_2L_1)$, $UB_{N_{22}} = 0.1 \times 1 + 0.9 \times 0 = 0.1$. The candidate queue becomes $(1 : \{N_{21}\}, 0.1 : \{N_{12}, N_{22}\})$. We then dequeue N_{21} . Since N_{21} is a leaf node, we calculate the actual association degree between e_a and entities contained in N_{21} , and obtain $d(e_a, e_c) = 0.5$. Since $d(e_a, e_c) > 0.1$, the algorithm returns e_a .

5 PRUNING EFFECTIVENESS ANALYSIS

In this section, we introduce a hierarchical mobility model significantly extending a well-established single-level individual mobility (IM) model [42]. In addition, we theoretically analyze the pruning effectiveness of our algorithms using the proposed model.

5.1 Individual mobility model

In the ensuing discussion, β , ρ , γ , α , ζ , μ , and ν are all model parameters.

For an entity e , the duration Δt of each PI follows

$$P(\Delta t) \sim |\Delta t|^{-1-\beta}, \quad (5)$$

which indicates that the duration of each PI follows a power law distribution, i.e., entities tend to stay for a short duration at each base spatial unit than for a long period.

When e leaves the current base spatial unit, it will either take an exploratory jump to a new base spatial unit, or return to somewhere it has previously visited. The probability of taking an exploratory jump is

$$P_{new} = \rho S^{-\gamma}, \quad (6)$$

where S is the number of base spatial units visited. As e visits more base spatial units, i.e., when S increases, the probability of e taking an exploratory jump decreases.

The direction of an exploratory jump is selected randomly, and its displacement follows

$$P(\Delta r) \sim |\Delta r|^{-1-\alpha}, \quad (7)$$

which stipulates that an entity tends to jump to some base spatial unit near its current position.

When taking a returning jump, the probability of returning to l is proportional to the number of e 's previous visits to l . The visit frequency of e to its y -th most visited base spatial unit follows

$$f_y \sim y^{-\zeta}, \quad (8)$$

which indicates that most visits of an entity are to the few top-ranked base spatial units.

Given a duration t , the total number of distinct base spatial units visited by e is

$$S(t) \sim t^\mu, \quad (9)$$

and the mean squared displacement follows

$$\langle \Delta x^2(t) \rangle \sim t^\nu, \quad (10)$$

which indicates that the longer the duration, the further e will drift away from its starting position.

5.2 Hierarchical individual mobility model

The IM model in Section 5.1 describes human mobility patterns at the finest spatial level. However, AjPIs may occur at multiple levels. In this section, we give the general spatial units distribution patterns and aggregate the mobility pattern at the finest level into patterns at higher levels.

To ease analysis, we assume that the area of interest is a square with side length L , and that it is equally divided into a grid of non-overlapping cells where each cell is a square with side length L_{bsu} . Each base spatial unit corresponds to a cell in this grid. Therefore, there are $(\frac{L}{L_{bsu}})^2$ base spatial units in total. For the sp-index, the size of each spatial unit (i.e., the number of base spatial units contained therein) and the structure of the tree depend on two parameters:

- *Width*, i.e., the number of nodes at each level; and
- *Relative density*, i.e., the relative sizes of nodes at the same level.

Intuitively, there are more spatial units at a finer level in the tree. Therefore, we assume that the width parameter follows a power law distribution w.r.t. level, i.e.,

$$W_l = Q \cdot l^a, \quad (11)$$

where $l \in [1, m]$ is the level, a is a tunable parameter, and $Q = (\frac{L}{L_{bsu}})^2 / m^a$ serves as a normalization factor.

In most cases the nodes at the same level have varying sizes, e.g. business districts usually have more buildings than rural areas. Therefore, we use the following power law distribution to model the relative sizes of nodes at level l :

$$D_l^i = W_l \cdot R \cdot i^b, \quad (12)$$

where $i \in [1, W_l]$ is the index of nodes at level l , b is a tunable parameter, and $R = 1 / \sum_{i=1}^{W_l} i^b$ is a normalization factor.

With parameters L , L_{bsu} , a and b , we can obtain the number of spatial units and also the size of each spatial unit at any level. Next we demonstrate how distributions introduced in Section 5.1 modeling mobility at the finest level can be extended and supplemented with other necessary distributions to derive a hierarchical mobility model.

Let U be a spatial unit at level l which contains a set of base spatial units \mathcal{S}_U . An exploratory jump of an entity takes place when (1) the entity jumps to a new base spatial unit; and (2) the new base spatial unit is contained in a spatial unit previously not visited, at level l . The probability of the first condition is given in Equation (6); the probability of the

second condition, referred to as P_{out} , can be computed by

$$P_{out}(U) = \frac{n_{visited}^U}{n_{reachable}^U} \sum_{s \in \mathcal{S}_U} \frac{1}{|\mathcal{S}_U|} H(s), \quad (13)$$

where $n_{reachable}^U$ denotes the number of spatial units within one jump's distance from U , $n_{visited}^U$ denotes the number of spatial units visited among these reachable ones, s is a base spatial unit in \mathcal{S}_U , and $H(s)$ denotes the probability of jumping outside U from s . It is evident that $H(s)$ is a function of the distance from s to the boundary of U as well as the jump distance distribution given in Equation (7). Therefore, the probability of taking an exploratory jump to a new spatial unit, P'_{new} , is

$$P'_{new}(U) = P_{new} \times P_{out}(U) \quad (14)$$

Since spatial units at higher levels have varying sizes and ranges, it is essential to derive the probability of an entity having visited unit U (the size of which is $|\mathcal{S}_U|$) within time t , $P_U(t)$.

$$P_U(t) = \frac{|\mathcal{S}_U|}{|\mathcal{S}|} + \sum_{U'} M(U, U', t), \quad (15)$$

where \mathcal{S} denotes the set of base spatial units, U' denotes some other spatial unit at level l . To derive this probability we consider two cases: the starting position of the entity is within U or it is not. The probability of the former case is $\frac{|\mathcal{S}_U|}{|\mathcal{S}|}$. For the latter case, the starting position can be within any other spatial unit, U' . $M(U, U', t)$ describes the probability of an entity starting from unit U' having visited U after time t , which can be inferred by the mean square displacement distribution given in Equation (10).

The visit frequency of an entity to its y -th most visited base spatial unit is given in Equation (8). At higher levels, the visit frequency rank, y , reflects not only personal preference, but also unit characteristics: spatial units containing more base spatial units are likely to be top-ranked. Therefore, we can safely assume that the visit frequency follows the same distribution at higher level, where y now describes the rank of the visit frequency to a particular spatial unit.

5.3 Analysis of pruning effectiveness

The model proposed in Section 5.2 enables us to simulate the movements of entities, estimate the overlap between the digital traces of any entities at all levels, with which we can calculate the expected association degree between an entity and its k most associated entities, d_e . Thus we can discard all branches whose upper bound is smaller than d_e . With more branches discarded, answering the query will be more efficient. Here we formally define pruning effectiveness (PE):

Definition 5.1 (Pruning Effectiveness). Given a set of entities \mathcal{E} , a query entity e , an association degree measure d , and

a searching strategy S , if S accurately answers a top- k query w.r.t \mathcal{E} , e , and d by checking entities in set \mathcal{E}' ($\mathcal{E}' \in \mathcal{E}$) only, then the pruning effectiveness of S is $\frac{|\mathcal{E}'| - k}{|\mathcal{E}|}$.

The average PE of is obtained averaging the PE of the top- k query answers over multiple entities.

Evidently, the UB of a child node on the MinSigTree cannot be larger than that of its parent nodes. Therefore, we can estimate PE by computing the percentage of leaf nodes on MinSigTree whose UBs are larger than d_e .

Suppose that the total number of base spatial units is n and the duration is t , the range of hash functions is thus $[0, n \times t - 1]$. For entity e_a with ST-cell set sequence seq_a and signatures sig_a , the probability of $sig_a^m[u] = i$ is

$$p(sig_a^m[u] = i) = \sum_{x=1}^{|seq_a^m|} C_{|seq_a^m|}^x \left(\frac{1}{n \times t}\right)^x \left(\frac{n \times t - i}{n \times t}\right)^{|seq_a^m| - x} \quad (16)$$

The condition of $sig_a^m[u] = i$ is that, $\exists \mathcal{S}_a \subset seq_a^m$, $\mathcal{S}_a \neq \emptyset$, s.t. $\forall s \in \mathcal{S}_a$, $h_u(s) = i$, and $\forall s' \in seq_a^m - \mathcal{S}_a$, $h_u(s') > i$. We assume that $|\mathcal{S}_a| = x$, the probability of which is $C_{|seq_a^m|}^x \left(\frac{1}{n \times t}\right)^x$, then all remaining ST-cells take hash values larger than i , the probability of which is $\left(\frac{n \times t - i}{n \times t}\right)^{|seq_a^m| - x}$. By grouping entities with the MinSigTree, the signature of a node N , SIG_N , satisfies $p(SIG_N[u] = i) \approx p(sig_a^m[u] = i)$ (equal when N only contains e_a).

Let r be the routing index of N , then the probability of $SIG_N[r] = i$ is

$$p(SIG_N[r] = i) = \sum_{x=1}^{n_h} C_{n_h}^x p(SIG_N[u] = i)^x p(SIG_N[u] < i)^{n_h - x},$$

$$p(SIG_N[u] < i) = \sum_{x=0}^{i-1} p(SIG_N[u] = x) \quad (17)$$

With the knowledge of $p(SIG_N[r] = i)$ we can estimate the value distribution of all leaves. Assume that range $[0, n \times t - 1]$ is divided into n_r consecutive equal-sized sub-ranges R , then we use $V[j]$ to denote the percentage of leaves whose value on the routing index is bounded by $R[j]$, $0 \leq j < n_r$.

Let n_c be the minimal number of ST-cells shared by entities with association degree larger than d_e . For node N , if $\exists \mathcal{S}_{ap} \in seq_a^m$, $|\mathcal{S}_{ap}| \geq n_c$, s.t. $\forall s \in \mathcal{S}_{ap}$, $s \notin \mathcal{P}S_N$, then N cannot be discarded.

Since hash functions are selected randomly, the hash values of all ST-cells are independent. Suppose that $SIG_N[r]$ is bounded by $R[j]$, then the probability that N cannot be discarded is

$$q(R[j]) = \sum_{x=n_c}^{|seq_a^m|} C_{|seq_a^m|}^x \left(\frac{n \times t - 1 - R[j]}{n \times t - 1}\right)^x \left(\frac{R[j]}{n \times t - 1}\right)^{|seq_a^m| - x} \quad (18)$$

PE can thus be calculated with the following equation:

$$PE = \sum_{j=0}^{n_r} V[j] q(R[j]) \quad (19)$$

6 EXPERIMENTS

In this section, we present a thorough experimental evaluation of our approach utilizing synthetic and real datasets, varying parameters of interest to explore the sensitivity of our proposal as well as PE trends.

6.1 Settings

Environment. The experiments are conducted on an Amazon Web Service EC2 instance, with a 30 core 2.3GHz Xeon CPU, 120GB of RAM, and ITB EBS Throughput Optimized HDD (maximal throughput 1,750MiB/s). The programming language is Java (version 1.8.1).

Datasets. We employ both real and synthetic datasets in our evaluation. Synthetic data are used as it is easy to vary parameters for sensitivity analysis. The synthetic dataset (referred to as SYN in the sequel) is generated by the hierarchical IM model in Section 5 with varying values of the parameters α , β , γ , ζ , ρ , a , b and m . Unless otherwise specified, we set $\alpha = 0.6$, $\beta = 0.8$, $\gamma = 0.2$, $\zeta = 1.2$, $\rho = 0.6$, which correspond to the normal mobility pattern (as per [42]), and $a = 2$, $b = 2$, $m = 4$ (a and b usually take values in the range $[1, 2]$ in real datasets², and 4 is the typical hierarchical level in a city). The sensitivity to these parameters governing data characteristics is evaluated in Section 6.4. The locations in the data are drawn from a set of 9 equal-sized sp-indexes with 250K locations in total. The data consists of the digital traces of 100M entities for a period of 30 days. The real dataset (referred to as REAL) is a WiFi hotspot handshaking data set provided to us by a large telecommunications provider and includes 30 million mobile devices and 76,739 WiFi hotspots. The hotspots are organized into a 4-level sp-index.

The data distribution is depicted in Appendix C.

Association degree measure. There are two properties any association degree measure (ADM in sequel) must possess (as discussed in Section 2.2), namely monotonicity with respect to AjPI level and duration. For purposes of exposition we utilize the following extensible function as the ADM:

$$d(e_a, e_b) = \frac{\sum_{l=1}^m l^u \left(\frac{|\mathcal{P}_{ab}^l|}{|\mathcal{P}_a^l| + |\mathcal{P}_b^l|}\right)^v}{max}, \quad (20)$$

where max is a normalization factor guaranteeing the score falls into the range $[0, 1]$, $|\mathcal{P}_{ab}^l|$ denotes the total duration of all level l AjPIs in set \mathcal{P}_{ab} , and $u > 0$ and $v > 0$ are parameters that can be tuned.

²<https://data.cityofnewyork.us/City-Government/Points-Of-Interest/rxuy-2muj>

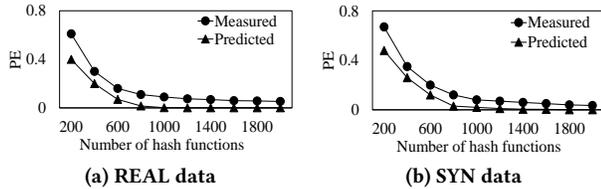


Figure 2: PE vs. the number of hash functions

Note that measures such as Jaccard, Dice and Cosine Similarity can be readily applied as well as they all share the two properties required by our association degree measure. We compare the ranking results of the proposed ADM to those of several widely-adopted set similarity measures in Appendix D. The results indicate that the proposed ADM compares favourably in terms of its ranking results to Jaccard, Dice and Cosine Similarity, when $v \in [0.5, 2]$ in Equation (20), and thus we utilize values in this range during experiments.

6.2 Baseline approach

We consider the following approach based on locality as the baseline approach for comparison purposes. At each level, we treat an ST-cell set of an entity as a transaction, each ST-cell as an item, and utilize frequent pattern mining techniques to find those frequently co-occurring ST-cells. As a result, ST-cells are partitioned into clusters, where each cluster is expected to contain ST-cells that are close to one another temporally and spatially. If there are n clusters in total, then we can assign each entity an n -bit vector, where the i -th bit in the vector of e equals 1 if e has presence in at least one ST-cell contained in cluster i , and 0 otherwise. We can thus use a bit-map to organize all entities. Given a query entity e_q , we compute an ADM upper bound between e_q and all bit-vectors. We start searching from the entities indexed by the vector with the highest UB, and continue until k entities are found where the minimal ADM of the k entities is already greater than the UBs between e_q and all remaining vectors.

The major drawback of such an approach is that in practice ST-cells show low degrees of locality, e.g., people living in the same neighborhood may work in different companies spread across the city, which makes it very difficult to identify frequently co-occurring ST-cells. The direct consequence is that the clusters obtained demonstrate strong coupling and the bit vectors cannot capture the PI patterns of entities well. Therefore, the upper bound is loose as will be discussed later in Section 6.7.

6.3 Sensitivity to the number of hash functions

PE is closely related to the number of hash functions utilized to compute the signatures, n_h . We thus evaluate the PE of the proposed approach by varying n_h , and compare the measured

PE in the experiment with the one predicted for the model of Section 5.3. The results are presented in Figure 2.

From the result one can observe that the MinSigTree provides high PE with more hash functions. The reason is that, compressing the large number of ST-cells into a low-dimensional space makes entities less unique, or even indistinguishable. With more hash functions employed, signatures can better summarize the PIs of entities and thus only closely associated entities will be placed in the same group. Diminishing returns occur when the number of hash functions reaches 1,000, as each entity has become unique enough that further employment of hash functions does not change the grouping.

As Figure 2 shows, the predicted PE is slightly better than measured, primarily for the following reasons:

- Spatial units in the hierarchical IM model are assumed to be rectangles for analysis purposes, while in practice units can be in any shapes. As a result, the mobility patterns at higher levels diverge from the model;
- It is assumed that the hash values are uniformly distributed on the range, which is not always the case in practice.

6.4 Sensitivity to data characteristics

We evaluate the PE under different mobility patterns and location distributions by varying all parameters in the hierarchical IM model. The hierarchical IM model involves a large number of parameters, each controlling different aspects of human mobility or location distribution. As such we vary one parameter each time and fix other parameters to the value associated with normal patterns (as per [42]) to investigate the individual influence of different parameters on performance. The results of answering Top-1, Top-10, and Top-50 queries with 2,000 hash functions under different data characteristics are presented in Figure 3.

One can observe that curves in Figure 3(a) show a descending trend, as α controls the movement locality in the following way: as α increases, an entity is more likely to jump to locations in proximity when it leaves the current position. A higher level of locality will produce more closely associated entities, and thus lead to better performance.

Curves in Figure 3(b) demonstrate little variation, which indicates that the approach is not sensitive to the expected duration of each presence instance. This is because we partition PI into ST-cells, and consider the digital traces of an entity as a set of ST-cells. As a result, whether these ST-cells are consecutive in time or not has no influence on PE.

Parameters ρ and γ together control the tendency of an entity to return to some previously visited location. With smaller ρ and larger γ , entities visit fewer locations in total, which increases the locality. Therefore, Figure 3(c) depicts an

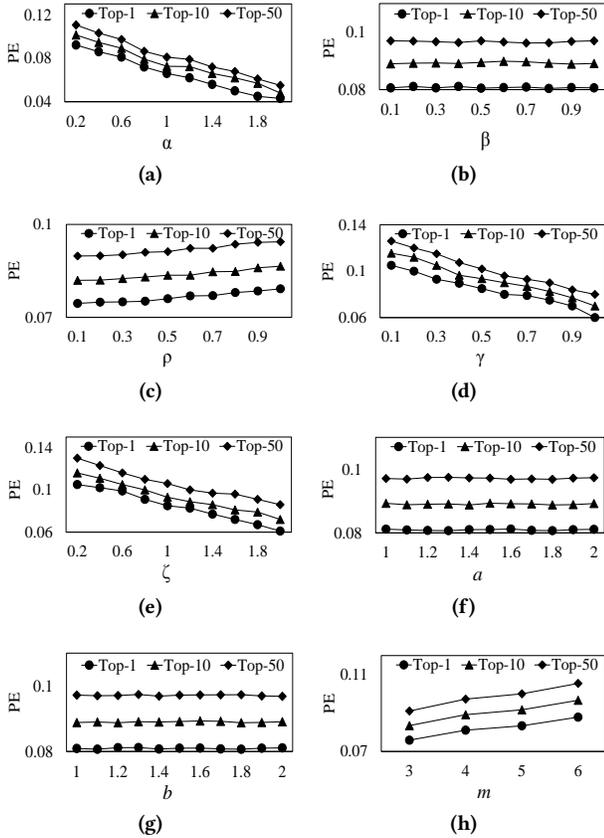


Figure 3: PE vs. data characteristics

ascending trend and Figure 3(d) a descending trend. ρ acts as a linear parameter, while γ is on the exponent; therefore curves in Figure 3(d) appear steeper than in Figure 3(c).

Similarly, Figure 3(e) demonstrates a descending trend, as ζ influences the locality by controlling the visit frequency distribution of an entity to locations. With higher ζ , most visits are to a few most frequently visited locations, while with lower ζ , visits are more uniformly distributed.

Curves in Figure 3(f) and (g) depict little variation, indicating that good PE can be achieved under any spatial distribution patterns. As is clear from the search algorithm, we touch the records of entity e only if the PI patterns of e resembles the PI patterns of the query entity at all sp-index levels. Although the values of a and b influence spatial units distribution at higher levels, base spatial unit numbers and distributions in the explored area are always constant, which means that the PI patterns of entities at the finest level do not change. As a result, groupings at level m of the MinSigTree remain unchanged under different values of a and b .

From Figure 3(h) we observe that the approach performs better with smaller m , i.e., fewer levels in the hierarchy. The reason is that with more spatial levels, more entities form AjPIs with each other, and thus the search space grows. As

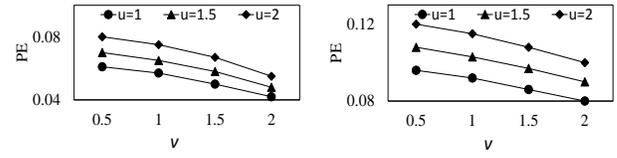


Figure 4: PE vs. ADM parameters

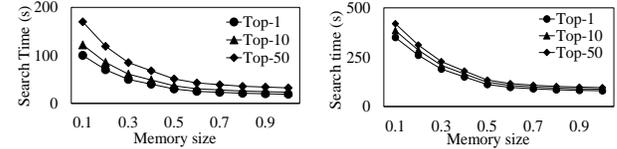


Figure 5: Search time vs. memory size

an example, if we assume that the spatial hierarchy is city-street-district-building, then if $m = 1$, we only consider AjPIs at the building level, while with $m = 2$ we consider AjPIs at both the building level and the street level, etc.

6.5 Sensitivity to ADM parameters

Values of u and v defined in the ADM of Section 6.1 provide different weights to AjPI level and duration when selecting associated entities. The PE under different ADM parameter values are presented in Figure 4.

As is clear from Figure 4, smaller u (level parameter) and larger v (duration parameter) yield high PE in both data sets. The reason is that, while ST-cells contain timestamps, they do not contain level information. Since signatures are computed based on ST-cells, the AjPI level is not encoded into the signature. As a result, entities sharing AjPIs for longer duration are more likely to have similar signatures than entities sharing AjPIs at finer levels. The results reveal that the approach performs better in cases where duration is the dominant factor of the association degree between entities.

6.6 Sensitivity to memory size

If more data can be stored in memory, the time spent to fetch records from disk is reduced. Therefore, the allocated memory size has an impact on query time. Figure 5 depicts the time required to answer Top-1, Top-10, and Top-50 queries with 2,000 hash functions under different memory sizes.

The horizontal axis in Figure 5 denotes the allocated memory size (relative size compared to raw data). It is evident that the curves in Figure 5 depict a descending trend as expected. The curve drops super-linearly with respect to the allocated memory size. The reason is that, the relative position of entities in the MinSigTree is not always guaranteed to be correlated to their association degrees, especially when the

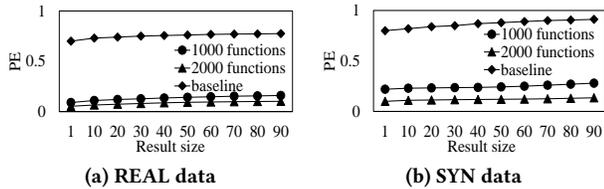


Figure 6: PE vs. result size (k)

number of hash functions is small (as discussed in Section 6.3). As a result, although we organize records by their relative position in the MinSigTree, closely associated entities are not always placed in adjacent disk blocks. However, as the memory size reaches 40% – 50% of the dataset size, the curves exhibit only small variation. We also experimented with different measures besides the ADM in Equation (20). Our results indicate that the choice of the measure does not impact the runtime performance of the index and the overall trends remain the same.

6.7 Sensitivity to result size

We also evaluate our approach as k (number of results desired in top- k) increases compared to the baseline method in Figure 6. PE on both SYN and REAL decreases slightly with increased result size, which is the consequence of both the ADM distribution and the nature of the branch-bound technique. Let e_q be the query entity, e_a be the i -th most associated entity to e_q , and e_b be the $(i + 1)$ -th most associated one. Let $df(i) = d(e_q, e_a) - d(e_q, e_b)$ denote the ADM difference between e_a and e_b . As Figure 10 indicates, the association degree distribution ranges for entities are denser when the association degree is small, i.e., $df(i) > df(j)$ if $i < j$ and $df(i) \rightarrow 0$ as i increases. Since the number of hash functions used to compute the signature is far less than the number of ST-cells, the UB of a node is not always guaranteed to be very tight. Let UB_b be the upper bound of the node containing e_b , then $d(e_q, e_a) < UB_b$ may occur, especially when $df(i) \approx 0$, i.e., i is large, which means we always need to check e_b before returning e_a . As a result, more entities are checked when the value of k , i.e., result size, is large, which implies the trend of the curves in Figure 6.

The baseline method, as argued in Section 6.2, is based on the existence of clusters among ST-cells, which is not typical in real-life digital traces. Consequently, the PE of the approach is greatly limited, which explains the results in Figure 6 showing that MinSigTree outperforms the baseline approach by large factors.

6.8 Indexing and update cost

The pre-processing cost to build the MinSigTree is depicted in Figure 7. Pre-processing time grows almost linearly with the number of hash functions (n_h), as the most expensive

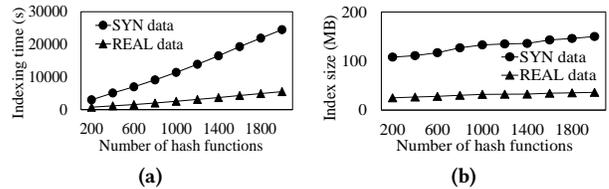


Figure 7: Indexing cost

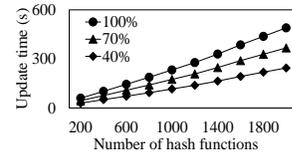


Figure 8: Update cost

step in the index construction process is the computation of signatures for each entity, which requires n_h hash operations for each ST-cell where the entity has a presence.

The size of the MinSigTree is provided in Figure 7(b). Generally, each node in the MinSigTree contains two integers, one indicating its routing index, and the other recording the hash value of the routing index. A leaf node also includes a pointer to the entities contained in this node. With more hash functions, each entity becomes more unique and thus a node with entity set \mathcal{E}_N may split into several new nodes, each containing a subset of \mathcal{E}_N . Therefore, the size of the MinSigTree increases with the number of hash functions. However, the overhead is quite small compared to the data size.

Figure 8 illustrates the time required for dynamic index updates. In particular, the figure depicts the time required to update records for 1 million entities in an already built MinSigTree. Since the update process is independent from the data distribution, we present experiments on SYN data. As discussed in Section 2.1, we assume that the sp-index remains unchanged over an extended period (as per [47]). Thus we focus on two common update operations in practice: inserting new entities and updating existing entities. As presented in Section 3.2.3, when updating existing entities, new digital traces and changes in existing digital traces are processed in the same fashion. Therefore, we do not distinguish between the two cases in this experiment. We report the time required under conditions when 100%, 70%, and 40% of the entities updated are existing entities, respectively. The time to update grows linearly with the number of hash functions as in the case of building the index. In addition, one can observe that inserting new entities requires less time than modifying the records for existing entities. The reason is that, when updating an existing entity we have to perform the following steps: (1) locate the entity’s position in the MinSigTree, (2) remove it from the corresponding leaf node in the index, (3) compute its new signature, and (4) insert it

to the proper node. In contrast, for a new entity only steps (3) and (4) are required.

The only parameter that can be tuned by users for performance/cost trade-off is the number of hash functions. In different scenarios one can decide on the suitable number of hash functions utilizing the pruning effectiveness curves of Figure 2 and cost curves of Figures 7 and 8.

7 RELATED WORK

We are not aware of any work that directly addresses the processing of digital traces as defined herein. Work on query processing over trajectories can be considered related, which focuses on the movement of entities. However the bulk of the work in this area deals with spatial proximity or shape of trajectories and is applicable to moving objects mainly. There are two branches of existing research on querying trajectories, namely, nearest neighbor queries [1, 4, 5, 10–12, 18, 23, 27, 28, 30, 33, 35, 40, 45], and top- k queries [3, 14, 32, 34, 39, 41, 44, 48, 49, 51, 52]. Nearest neighbor queries retrieve spatially close trajectories [11] or entities [10, 28] under particular distance measures [1, 5, 23] with various constraints [4, 12, 18, 27, 30, 33, 35, 40, 45], e.g., uncertain trajectories [33] or road networks [35]. Top- k queries define metrics [32, 45] on specific trajectory types [3, 14, 34, 39, 41, 44, 49, 51], e.g., activity trajectory [52] or semantic trajectory [51], to quantify the similarity between trajectories and retrieve trajectories or entities most similar to a given trajectory [32], entity [51], set of locations [52], etc. One representative piece of work related to the problem studied herein is Frentzos et al. [20], which proposes a set of metrics for k -Most Similar Trajectory (k -MST) search over moving object databases, and designs an approximation method for efficient search with R-tree-like structures. These works however along with work on similarity search over trajectories mainly focus on spatial closeness or trajectory shape, without considering the influence of spatial topology, such as hierarchy, on measuring the association among trajectories and the corresponding entities. Consequently they lack the ability to infer the association degree between entities from their trajectories. In addition, as the metrics used therein are based on sequence distance (e.g., Longest Common Sub-Sequence [46]), or Time Series distance (e.g., Dynamic Time Warping [37]), which either ignore the time dimension or assume trajectories are aligned in time, they are not guaranteed to satisfy the monotonic properties of the association degree measure of Equation (3).

A few pieces of work in recent years also deal with digital traces of human beings [25, 36]. Digital traces in their context, however, mainly refer to the records produced by digital devices on the Internet, such as emails, twitter posts, etc, which are not associated with spatial-temporal presences and

thus share little semantic similarity with the digital traces proposed in this paper.

Existing top- k query processing techniques, including sorted-list based approaches [15, 17], layer based approaches [8, 50], R-tree based approaches [6, 9], are primarily designed to address the problem for low-dimensional data. Since the dimensionality dealt with in this paper is extremely high (as each ST-cell is one dimension, and there are millions of ST-cells), these approaches are not applicable to this problem. On the contrary, the approach proposed in this paper utilizes hashing functions for dimensionality reduction in a fashion that preserves presence instance patterns and thus exact top- k queries can still be supported.

Frequent pattern mining algorithms [2, 24] discover frequently co-occurring items from transaction databases. Such algorithms have also been proposed to identify communities among populations [19, 29]. These approaches are not effective to answer top- k queries in our problem domain as digital traces do not typically contain such patterns.

Hashing techniques are widely adopted in set duplicate detection tasks [22, 43], among which MinHash demonstrates excellent performance [21, 31, 38, 53, 54]. Hashing approaches, however, are always used as approximation rather than to calculate exact similarity. We modify MinHash approaches in this paper to support exact top- k queries.

8 CONCLUSIONS AND FUTURE WORK

The proliferation of ambient connectivity for certain entity types gives rise to query processing problems of the resulting digital traces. In this paper, we initiated the study and formally defined the problem of top- k query over digital traces, and developed a suite of techniques to efficiently process such queries. We proposed a hash-based indexing structure and combined it with a given spatial hierarchy to answer exact top- k queries, which is a combination that has not been investigated before. We generalized a well-established mobility model to a hierarchical spatial environment and analytically quantified the pruning effectiveness of the proposed method. We also presented extensive experiments on both synthetic and real data sets demonstrating the practical utility of our proposal.

This study introduces several directions for further work. Although top- k query is a natural query to study in this context, several other interesting query processing questions exist. Extending the proposed techniques to other operators such as approximate top- k and joins as well as studying alternate embedding with diverse properties are important directions. Natural extensions to identifying outlier digital traces as well as related data mining questions are worthy of further investigation.

ACKNOWLEDGMENTS

This work was supported in part by the NSERC Discovery Grants. We thank Dr. Parke Godfrey and the anonymous reviewers for their valuable comments and helpful suggestions.

REFERENCES

- [1] Tenindra Abeywickrama, Muhammad Aamir Cheema, and David Taniar. 2016. K-nearest neighbors on road networks: a journey in experimentation and in-memory implementation. *Proceedings of the VLDB Endowment* 9, 6 (2016), 492–503.
- [2] Charu C Aggarwal and Jiawei Han. 2014. *Frequent pattern mining*.
- [3] Pritom Ahmed, Mahbub Hasan, Abhijith Kashyap, Vagelis Hristidis, and Vassilis J Tsotras. 2017. Efficient Computation of Top-k Frequent Terms over Spatio-temporal Ranges. In *Proceedings of the 2017 ACM SIGMOD International Conference on Management of Data*. 1227–1241.
- [4] Ahmed M Aly, Walid G Aref, and Mourad Ouzzani. 2012. Spatial queries with two kNN predicates. *Proceedings of the VLDB Endowment* 5, 11 (2012), 1100–1111.
- [5] Senjuti Basu Roy and Kaushik Chakrabarti. 2011. Location-aware type ahead search on spatial databases: semantics and efficiency. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*. 361–372.
- [6] Norbert Beckmann, Hans-Peter Kriegel, Ralf Schneider, and Bernhard Seeger. 1990. The R⁺-tree: an efficient and robust access method for points and rectangles. In *Acm Sigmod Record*, Vol. 19. 322–331.
- [7] Andrei Z Broder. 1997. On the resemblance and containment of documents. In *Compression and complexity of sequences 1997. proceedings*. 21–29.
- [8] Yuan-Chi Chang, Lawrence Bergman, Vittorio Castelli, Chung-Sheng Li, Ming-Ling Lo, and John R Smith. 2000. The onion technique: indexing for linear optimization queries. In *ACM Sigmod Record*, Vol. 29. 391–402.
- [9] Lu Chen, Yunjun Gao, Gang Chen, and Haida Zhang. 2016. Metric all-k-nearest-neighbor search. *IEEE Transactions on Knowledge and Data Engineering* 28, 1 (2016), 98–112.
- [10] Lei Chen, M Tamer Özsu, and Vincent Oria. 2005. Robust and fast similarity search for moving object trajectories. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. 491–502.
- [11] Zaiben Chen, Heng Tao Shen, Xiaofang Zhou, Yu Zheng, and Xing Xie. 2010. Searching trajectories by locations: an efficiency study. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. 255–266.
- [12] Farhana M Choudhury, J Shane Culpepper, Timos Sellis, and Xin Cao. 2016. Maximizing bichromatic reverse spatial and textual k nearest neighbor queries. *Proceedings of the VLDB Endowment* 9, 6 (2016), 456–467.
- [13] Richard Cole. 1988. Parallel merge sort. *SIAM J. Comput.* 17, 4 (1988), 770–785.
- [14] Tobias Emrich, Maximilian Franzke, Hans-Peter Kriegel, Johannes Niedermayer, Matthias Renz, and Andreas Züfle. 2014. An extendable framework for managing uncertain spatio-temporal data. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. 1087–1090.
- [15] Ronald Fagin, Ravi Kumar, Mohammad Mahdian, D Sivakumar, and Erik Vee. 2004. Comparing and aggregating rankings with ties. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 47–58.
- [16] Ronald Fagin, Ravi Kumar, and Dakshinamurthi Sivakumar. 2003. Comparing top k lists. *SIAM Journal on discrete mathematics* 17, 1 (2003), 134–160.
- [17] Ronald Fagin, Amnon Lotem, and Moni Naor. 2003. Optimal aggregation algorithms for middleware. *Journal of computer and system sciences* 66, 4 (2003), 614–656.
- [18] Yixiang Fang, Reynold Cheng, Wenbin Tang, Silviu Maniu, and Xuan Yang. 2016. Scalable algorithms for nearest-neighbor joins on big trajectory data. *IEEE Transactions on Knowledge and Data Engineering* 28, 3 (2016), 785–800.
- [19] Zhenni Feng and Yanmin Zhu. 2016. A survey on trajectory data mining: techniques and applications. *IEEE Access* 4 (2016), 2056–2067.
- [20] Elias Frentzos, Kostas Gratsias, and Yannis Theodoridis. 2007. Index-based most similar trajectory search. In *IEEE 23rd International Conference on Data Engineering (ICDE), 2007*. 816–825.
- [21] Junhao Gan, Jianlin Feng, Qiong Fang, and Wilfred Ng. 2012. Locality-sensitive hashing scheme based on dynamic collision counting. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. 541–552.
- [22] Jinyang Gao, Hosagrahar Visvesvaraya Jagadish, Wei Lu, and Beng Chin Ooi. 2014. DSH: data sensitive hashing for high-dimensional k-nnsearch. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. 1127–1138.
- [23] Ralf Hartmut Güting, Thomas Behr, and Jianqiu Xu. 2010. Efficient k-nearest neighbor search on moving object trajectories. *The VLDB Journal The International Journal on Very Large Data Bases* 19, 5 (2010), 687–714.
- [24] Jiawei Han, Jian Pei, and Yiwen Yin. 2000. Mining frequent patterns without candidate generation. In *ACM sigmod record*, Vol. 29. 1–12.
- [25] Cheng-Kang Hsieh, Longqi Yang, Honghao Wei, Mor Naaman, and Deborah Estrin. 2016. Immersive recommendation: News and event recommendations using personal digital traces. In *Proceedings of the 25th International Conference on World Wide Web*. 51–62.
- [26] Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika* 30, 1/2 (1938), 81–93.
- [27] Jae-Gil Lee, Jiawei Han, and Kyu-Young Whang. 2007. Trajectory clustering: a partition-and-group framework. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*. 593–604.
- [28] Zhenhui Li, Ming Ji, Jae-Gil Lee, Lu-An Tang, Yintao Yu, Jiawei Han, and Roland Kays. 2010. MoveMine: mining moving object databases. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. 1203–1206.
- [29] Yunhao Liu, Yiyang Zhao, Lei Chen, Jian Pei, and Jinsong Han. 2012. Mining frequent trajectory patterns for activity monitoring using radio frequency tag arrays. *IEEE Transactions on Parallel and Distributed Systems* 23, 11 (2012), 2138–2149.
- [30] Jiaheng Lu, Ying Lu, and Gao Cong. 2011. Reverse spatial and textual k nearest neighbor search. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*. 349–360.
- [31] Qin Lv, William Josephson, Zhe Wang, Moses Charikar, and Kai Li. 2007. Multi-probe LSH: efficient indexing for high-dimensional similarity search. In *Proceedings of the 33rd international conference on Very large data bases*. 950–961.
- [32] Chunyang Ma, Hua Lu, Lidan Shou, and Gang Chen. 2013. KSQ: Top-k similarity query on uncertain trajectories. *IEEE Transactions on Knowledge and Data Engineering* 25, 9 (2013), 2049–2062.
- [33] Johannes Niedermayer, Andreas Züfle, Tobias Emrich, Matthias Renz, Nikos Mamoulis, Lei Chen, and Hans-Peter Kriegel. 2013. Probabilistic nearest neighbor queries on uncertain moving object trajectories. *Proceedings of the VLDB Endowment* 7, 3 (2013), 205–216.

- [34] Julien Pilourdault, Vincent Leroy, and Sihem Amer-Yahia. 2016. Distributed evaluation of top-k temporal joins. In *Proceedings of the 2016 ACM SIGMOD International Conference on Management of Data*. 1027–1039.
- [35] Michalis Potamias, Francesco Bonchi, Aristides Gionis, and George Kollios. 2010. K-nearest neighbors in uncertain graphs. *Proceedings of the VLDB Endowment* 3, 1-2 (2010), 997–1008.
- [36] Tobias Preis, Helen Susannah Moat, Steven R Bishop, Philip Treleaven, and H Eugene Stanley. 2013. Quantifying the digital traces of Hurricane Sandy on Flickr. *Scientific reports* 3 (2013), 3141.
- [37] Hiroaki Sakoe and Seibi Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing* 26, 1 (1978), 43–49.
- [38] Venu Satuluri and Srinivasan Parthasarathy. 2012. Bayesian locality sensitive hashing for fast similarity search. *Proceedings of the VLDB Endowment* 5, 5 (2012), 430–441.
- [39] Zhou Shao, Muhammad Aamir Cheema, David Taniar, and Hua Lu. 2016. Vip-tree: an effective index for indoor spatial queries. *Proceedings of the VLDB Endowment* 10, 4 (2016), 325–336.
- [40] Mehdi Sharifzadeh and Cyrus Shahabi. 2010. Vor-tree: R-trees with voronoi diagrams for efficient processing of spatial nearest neighbor queries. *Proceedings of the VLDB Endowment* 3, 1-2 (2010), 1231–1242.
- [41] Jieming Shi, Dingming Wu, and Nikos Mamoulis. 2016. Top-k relevant semantic place retrieval on spatial RDF data. In *Proceedings of the 2016 ACM SIGMOD International Conference on Management of Data*. 1977–1990.
- [42] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. 2010. Modelling the scaling properties of human mobility. *Nature Physics* 6, 10 (2010), 818.
- [43] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Heng Tao Shen. 2013. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. 785–796.
- [44] Bo Tang, Shi Han, Man Lung Yiu, Rui Ding, and Dongmei Zhang. 2017. Extracting top-k insights from multi-dimensional data. In *Proceedings of the 2017 ACM SIGMOD International Conference on Management of Data*. 1509–1524.
- [45] Lu-An Tang, Yu Zheng, Xing Xie, Jing Yuan, Xiao Yu, and Jiawei Han. 2011. Retrieving k-nearest neighboring trajectories by a set of point locations. In *International Symposium on Spatial and Temporal Databases*. 223–241.
- [46] Michail Vlachos, George Kollios, and Dimitrios Gunopulos. 2002. Discovering similar multidimensional trajectories. In *IEEE 18th International Conference on Data Engineering (ICDE), 2002*. 673–684.
- [47] Hongjian Wang, Daniel Kifer, Corina Graif, and Zhenhui Li. 2016. Crime rate inference with big data. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 635–644.
- [48] Haozhou Wang, Kai Zheng, Xiaofang Zhou, and Shazia Sadiq. 2015. Sharkdb: An in-memory storage system for massive trajectory data. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. 1099–1104.
- [49] Sheng Wang, Zhifeng Bao, J Shane Culpepper, Timos Sellis, Mark Sanderson, and Xiaolin Qin. 2017. Answering top-k exemplar trajectory queries. In *IEEE 33rd International Conference on Data Engineering (ICDE), 2017*. 597–608.
- [50] Dong Xin, Chen Chen, and Jiawei Han. 2006. Towards robust indexing for ranked queries. In *Proceedings of the 32nd international conference on Very large data bases*. 235–246.
- [51] Bolong Zheng, Nicholas Jing Yuan, Kai Zheng, Xing Xie, Shazia Sadiq, and Xiaofang Zhou. 2015. Approximate keyword search in semantic trajectory database. In *IEEE 31st International Conference on Data*

Engineering (ICDE), 2015. 975–986.

- [52] Kai Zheng, Shuo Shang, Nicholas Jing Yuan, and Yi Yang. 2013. Towards efficient search for activity trajectories. In *IEEE 29th International Conference on Data Engineering (ICDE), 2013*. 230–241.
- [53] Yuxin Zheng, Qi Guo, Anthony KH Tung, and Sai Wu. 2016. LazyLsh: Approximate nearest neighbor search for multiple distance functions with a single index. In *Proceedings of the 2016 ACM SIGMOD International Conference on Management of Data*. 2023–2037.
- [54] Erkang Zhu, Ken Q Pu, Fatemeh Nargesian, and Renée J Miller. 2017. Interactive navigation of open data linkages. *Proceedings of the VLDB Endowment* 10, 12 (2017), 1837–1840.

A NOTATIONS AND ABBREVIATIONS

The notations and abbreviations are given in Table 1.

B COST OF INDEX CONSTRUCTION

THEOREM B.1. *The maximal I/O cost in the indexing process is $2n_p \times [1 + \lceil \log_{n_b} \lceil n_p/n_b \rceil \rceil] + n_p$, where n_p is the total number of pages storing the digital traces and n_b is the number of buffer pages in memory.*

Methods proposed in Section 3 require digital traces to be organized by entities, but real world data have varying formats. In a system with adequate memory we can load all records into memory and directly fetch the digital traces of a specific entity. However, when memory becomes the bottleneck, sorting the digital traces by entity becomes a necessity. We employ the well-known B -way external merge sort [13] algorithm to do the ordering. Following the sorting principle, the I/O cost in the sorting process is thus $2N \times [1 + \lceil \log_B \lceil N/B \rceil \rceil]$. After the sorting, we need to read the records of each entity once for signature computation.

THEOREM B.2. *The total processor cost in the indexing process is $\Theta(|\mathcal{E}|Cmn_h)$, where C is the average number of ST-cells in which an entity has presence.*

With access to the digital traces organized by entity, we can compute the signature list for each entity and build the MinSigTree. As described in Section 3, for each entity, we fetch its m ST-cell sets, employ a family of n_h functions to map each set to a signature, and then build an m -level MinSigTree based on the signatures of all entities. Therefore, the total processor cost in the indexing process is $O(|\mathcal{E}|Cmn_h)$.

THEOREM B.3. *The minimum memory required in the indexing process is $\min\{(n_h)^m, |\mathcal{E}| \times m\} + n_h + C$.*

Since the signatures of each entity are computed independently, we can fetch one entity into memory at a time and update the MinSigTree incrementally. In order to avoid extra I/O cost, we need to keep the MinSigTree and hash functions in memory. Theoretically, the size of the MinSigTree is $n_h + (n_h)^2 + \dots + (n_h)^m \approx (n_h)^m$. However, since the total number of entities is $|\mathcal{E}|$, the number of leaves in the tree is bounded by $|\mathcal{E}|$. Since each node has one and only one

Table 1: Notations and Abbreviations

Notation/Abbreviation	Definition
\mathcal{E}	the set of all entities
\mathcal{S}	the set of all ST-cells
PI	presence instance
p_a	a PI of entity e_a
\mathcal{P}_a	the digital traces of entity e_a
AjPI	adjoint presence instance
p_{ab}	an AjPI between e_a and e_b
\mathcal{P}_{ab}	all AjPIs between entity e_a and entity e_b
seq_a	the ST-cell set sequence of entity e_a
sig_a	the signature list of entity e_a
SIG_N	the signature of node N
IM model	individual mobility model
α	parameter in IM model controlling the displacement of consecutive PIs
β	parameter in IM model controlling the duration of PI
ρ, γ	parameters in IM model controlling the probability of exploratory jump
ζ	parameter in IM model controlling visit frequency
m	level of sp-index
a	width parameter of sp-index
b	relative density parameter of sp-index
\mathcal{PS}	pruned set
\mathcal{PPS}	partial pruned set
PE	pruning effectiveness
ADM	association degree measure
u	parameter in ADM controlling the weight of level
v	parameter in ADM controlling the weight of duration

parent node, the number of nodes at other level of the tree is also bounded by $|\mathcal{E}|$. Therefore, the size of the MinSigTree is $\min\{(n_h)^m, |\mathcal{E}| \times m\}$. The minimal memory required is thus $(\min\{(n_h)^m, |\mathcal{E}| \times m\} + n_h + C)$, storing the MinSigTree, n_h hash functions and the ST-cells of one entity.

C DATA DISTRIBUTION

The data distribution is depicted in Figure 9, demonstrating both data distribution across levels as well as distribution of AjPI duration. Note that the vertical axes in all these plots are in log scale. Figure 9(a) depicts the number of entities forming AjPIs with a particular entity at each level on REAL. Given an entity e , as shown in Figure 9(a), roughly 22 million entities form AjPIs with e at level 1 (two entities forming an AjPI at a finer level also form an AjPI at the coarser levels), etc. Figure 9(b) illustrates the same distribution on SYN. Figure 9(c) provides the duration distribution of AjPI at each level: roughly 20 million entities form AjPI with e at level 1 for durations shorter than 100 hours, etc.

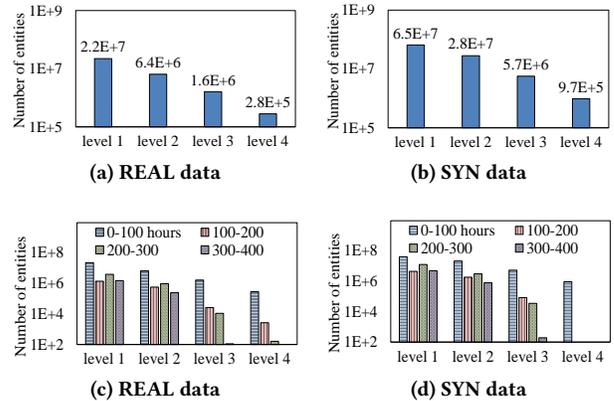


Figure 9: Data distribution

Figure 10 provides the association degree distribution under $u = 1$ and $v = 1$, where the horizontal axes are association degree ranges, and the the height of a bar denotes the number of entities falling in the corresponding ADM range

with the query entity. From Figure 10, it is evident that most entities bear low association degrees with a particular entity.

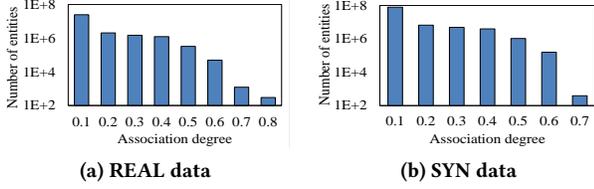


Figure 10: Association degree distribution

D MEASURE COMPARISON

Intuitively, a similarity measure, M_p , simulates another similarity measure, M_q , on a data set D , if the ranking order and the association degrees by M_p and M_q on D are close. Formally, assume that the top- k associated entities to a query entity e measured by M_p form a sequence R_p (sorted by association degree), R_p^i ($i \in [1, k]$) is the i -th entity in the ranking; assume that $R_p^i.deg$ denotes the association degree of R_p^i to e , then the simulation effectiveness of M_p to M_q is quantified by the metrics in Equation (21).

$$K_{avg}(M_p, M_q) = E(K(R_p \widehat{\langle \mathcal{R}_q - \mathcal{R}_p \rangle}, R_q \widehat{\langle \mathcal{R}_p - \mathcal{R}_q \rangle})),$$

$$ADDiff(M_p, M_q) = \frac{\sum_{i=1}^k |R_p^i.deg - R_q^i.deg|}{k}, \quad (21)$$

where K_{avg} is a generalized form of Kendall's tau distance [26] to measure the distance between top- k lists [16], \mathcal{R}_p is the set of entities in R_p , $\langle * \rangle$ is an operator transferring a set to a sequence in any order, $\widehat{}$ denotes the concatenation of two sequences, E is the expectation, and K denotes Kendall's tau distance [26] introduced below.

Given two ranked lists, τ_1 and τ_2 , both of size n , the Kendall's tau distance between τ_1 and τ_2 , $K(\tau_1, \tau_2)$, is computed with Equation (22).

$$K(\tau_1, \tau_2) = \frac{|\{(i, j) : i < j, P(i, j) \vee Q(i, j)\}|}{n(n-2)/2},$$

$$P(i, j) = \tau_1(i) < \tau_1(j) \wedge \tau_2(i) > \tau_2(j),$$

$$Q(i, j) = \tau_1(i) > \tau_1(j) \wedge \tau_2(i) < \tau_2(j), \quad (22)$$

where $\tau_1(i)$ and $\tau_2(i)$ are the ranking orders of element i in lists τ_1 and τ_2 respectively. The Kendall's tau distance between two identical lists is 0, and the Kendall's tau distance between two reverse lists is 1. Kendall's tau distance is commonly used in measuring the ordinal correlation between ranked lists.

Given a data set, $K_{avg}(M_p, M_q)$ describes the consistency of the ranking orders, which corresponds to the effectiveness of the results, and $ADDiff$ depicts the deviation of the association degrees, which influences the performance of the approach. M_p simulates M_q if both $\tau(M_p, M_q)$ and $ADDiff(M_p, M_q)$ are low.

Table 2: Simulation effectiveness

(a) Average Kendall's tau distance			
	Top-1	Top-10	Top-50
Dice	0.0	0.0	0.0
Jaccard	0.0	0.0	0.0
Cosine	2.0E-3	6.7E-3	1.1E-2
(b) Association degree difference			
	Top-1	Top-10	Top-50
Dice	0.0	0.0	0.0
Jaccard	1.1E-2	6.7E-3	5.0E-3
Cosine	3.2E-5	4.0E-5	5.5E-5

In order to apply set similarity measures to hierarchical spatial environment, at each spatial level we use Dice, Jaccard, or Cosine metric to compute the similarity between the digital traces of two entities, and use the weighted summation of similarities at all levels as the final association degree. Since weights are independent from the measures and have no influence on the simulation, we simply let the weight of level i , w_i , take value $\frac{i}{Z}$ (Z is a normalization factor), which corresponds to $u = 1$ in Equation (20), and vary the value of v to evaluate the simulation effectiveness of the ADM to other similarity measures.

Table 2 gives the simulation effectiveness of the ADM to other measures. The best simulation to Dice and Cosine Similarity is obtained when $v = 1$, and the best simulation to Jaccard Similarity is obtained when $v = 1.2$. We can observe from Table 2 that the ADM simulates other measures accurately, especially when the result size (k) is small. It is worth noting that the ADM exactly takes the form of Dice Similarity when $v = 1$. As is clear from Equation (20), varying the value of v only changes the association degree, but has no influence on the ranking order. Our experiments indicate that when v is in the range of $[0.5, 2]$ the association degree computed by the ADM is close to those of the other measures, and thus we utilize values in this range during experiments.