Scalable Linkage across Location Enhanced Services

Fuat BASIK
Supervised By: Hakan Ferhatosmanoğlu and Buğra Gedik
Department of Computer Engineering, Bilkent University, Turkey
fuat.basik@bilkent.edu.tr

ABSTRACT

In this work, we investigate methods for merging spatio-temporal usage and entity records across two location-enhanced services, even when the datasets are semantically different. To address both effectiveness and efficiency, we study this linkage problem in two parts: *model* and *framework*. First we discuss models, including *k-l* diversity— a concept we developed to capture both spatial and temporal diversity aspects of the linkage, and probabilistic linkage. Second, we aim to develop a framework that brings efficient computation and parallelization support for both models of linkage.

1. INTRODUCTION

An important portion of digital footprint left behind by entities interacting with online services contains spatio-temporal references. This footprint is a fertile resource for business intelligence applications [11]. We refer to the services that create spatio-temporal records of their usage as *Location Enhanced Services* (LES). For instance, Foursquare/Swarm[1] — a popular social networking service, records the locations of users when they check-in at a point-of-interest (POI) registered in the system. Similarly, mobile phone service providers generate a record every time a call is made, which includes the cell tower whose coverage area contains the user's location.

Records with similar location and time naturally observe similar phenomena. The data analyst can gather such data from multiple sources, which are typically anonymized due to privacy concerns. These sources could generate semantically different datasets, or the semantic link between the sources could have been lost due to anonymization. As most data science tasks require large amount of data for accurate training with higher confidence, scientists need to combine data from multiple sources to produce accurate aggregate patterns. For example, spatio-temporal usage records belonging to the same real-world user can be matched across records from two different location-enhanced services, even when the datasets are semantically different. Another example would be linkage of the sensor data from different vendors that are embedded to the same moving system, i.e. self-driving cars. This linkage enables data scientists and service providers to obtain information that they cannot derive by mining only one set of usage records. Consider a LES provider who combines user segmentation results derived from its own usage records with social segmentation results derived from the publicly available Swarm records. There are several algorithmic and systems challenges to merge information from multiple sources of anonymized spatio-temporal data that are collected with necessary permissions. To cover both effectiveness and efficiency, we divide this linkage problem into two parts: *model* and *framework*.

To develop effective models, one needs to define a similarity or probabilistic measure for linkage, which considers time, location, and the relationship between the two. This is relatively simpler for many record linkage tasks [4], where linkage is defined based on a similarity measure defined over records (such as Minkowski distance or Jaccard similarity). In spatio-temporal linkage, for a pair of users from two different datasets to be considered as matching, their usage history must contain records that are close both in space and time; and there must not be *negative matches*, such as records that are close in time, but far in distance. We call such negative matches, *alibi*s. To address these challenges, we introduce two *linkage models*. The first one is based on *k-l diversity* — a new concept we have introduced to capture both spatial and temporal diversity aspects of the linkage. A pair of entities, one from each dataset, is called *k-l* diverse if they have at least k co-occurring records (both temporally and spatially) in at least l different locations, and, such pairs of entities must not have any alibis. The second model we aim to develop is based on *probabilistic* linkage — in which we seek to model the matching probability of two entities based on their spatio-temporal history. A pair of entities are called *match, or linked* with probability P, which is proportional to their common events aggregated on grids, and timestamps. P is inversely proportional to number of all other entities simultaneously acting at the same grid.

Considering that location-based social networks get millions of updates every day, linkage over hundreds of days of data would take impractically long amount of time. Naïve record linkage algorithms that compare every pair of records take $\mathcal{O}(n^2)$ time [6], where n is the number of records. The generic entity matching tools do not provide the necessary optimization for scalability and efficiency of spatio-temporal

[1] www.foursquare.com / www.swarmapp.com

Proceedings of the VLDB 2017 PhD Workshop, August 28, 2017. Munich, Germany.

linkage [3]. In order to merge data sets in a reasonable time, we will develop a scalable *framework* that takes advantage of the spatio-temporal structure of the data. The *ST-Link* algorithm we have recently modeled to realize the *k-l* diversity model in real world, uses two filtering steps before pairwise comparisons of candidate users, and makes use of spatial index trees, temporal sliding windows and log-structured merge trees [7]. In addition to effective indexing techniques, we believe efficiency could benefit from parallelization of computation.

2. LINKAGE MODELS

Datasets. We denote the two spatio-temporal usage record datasets from the two LES across which the linkage is to be performed as $\mathcal{I}$ and $\mathcal{E}$.

Entities and events. *Entities*, or users, are real-world systems or people who use LES. We use the terms user and entity interchangeably. They are represented in the datasets with their ids, potentially anonymized, which are different for the two LES. *Events* correspond to usage records generated by a LES as a result of users interacting with the service. For an event $e \in \mathcal{E}$ (or $i \in \mathcal{I}$), $e.u$ (or $i.u$) represents the entity associated with the event. We use $U_\mathcal{E}$ and $U_\mathcal{I}$ to denote the set of entity ids in the datasets $\mathcal{E}$ and $\mathcal{I}$, respectively. We have $U_\mathcal{E} = \{e.u : e \in \mathcal{E}\}$ and $U_\mathcal{I} = \{i.u : i \in \mathcal{I}\}$.

Location and time. Each event in the dataset contains location and time information. The location information is in the form of a region, denoted as $e.r$ for event e. We do not use a point for location, as for most LES the location information is in the form of a region. We assume the time information is a point in time.

2.1 k-l Diversity

The core idea behind the *k-l* diversity model is to locate pairs of entities whose events satisfy *k-l* diversity. Furthermore, such pairs of entities must not have any alibis.

Co-occurrence. Two events from different datasets are called co-occurring if they are close in space and time. For two records $i \in \mathcal{I}$ and $e \in \mathcal{E}$, closeness is defined in terms of intersection of regions. To capture closeness in time, we use a parameter α, and call two events are close in time if they are within a window of α time units of each other.

Alibi. While a definition of similarity is necessary to link events from two different datasets, a definition of dissimilarity is also required to rule out pairs of entities as potential matches in our linkage. Such *negative matches* enable us to rule out incorrect matches and also reduce the space of possible matches throughout the linkage process. We refer to these negative matches as *alibi*s. In this work, we use alibi to define events from two different datasets that happened around the same time but at different locations, such that it is not possible for a user to move from one of these locations to the other within the duration defined by the difference of the timestamps of the events.

Entity linkage. Let $x \in U_\mathcal{I}$ and $y \in U_\mathcal{E}$ be two entities. In order to be able to decide whether two entities are the same, we search for k co-occurring event pairs and at least l of them are at diverse locations. However, each co-occurring event pair does not count as 1, since each of these events could co-occur with many other events. Let $C(i, e)$ be the function to represent aforementioned co-occurrence relation of records i, and e, We weight these co-occurring event pairs as:

$$w(i,e) = |\{i_1.u : C(i_1, e) \wedge i_1 \in \mathcal{I}\}|^{-1} \cdot |\{e_1.u : C(i, e_1) \wedge e_1 \in \mathcal{E}\}|^{-1} \quad (1)$$

Given a co-occurring event pair between two entities, we check how many possible entities' events could be matched to these events. For instance, in Figure 1, consider the solid line at the top with the weight 1/6. The event on its left could be matched to events of 2 different entities, and the event on its right could be matched to events of 3 different entities. To compute the weight of a co-occurring pair, we multiply the inverse of these entity counts, assuming the possibility of matching from both sides are independent. As such, in the Figure 1, we get $1/2 \cdot 1/3 = 1/6$.

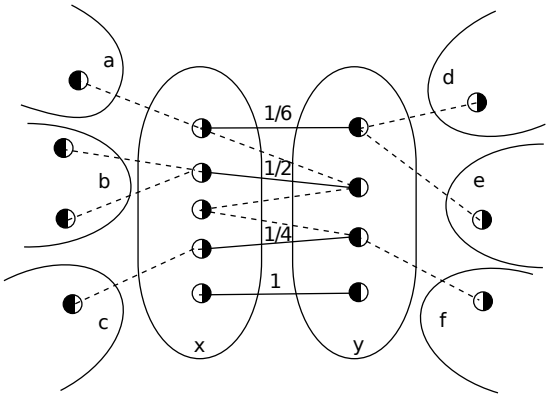

Figure 1: The co-occurring event pairs are shown using dashed lines. Events from a given entities are shown within circles. Entities a, b, c, and y are from one LES, and the entities d, e, f, and x are from the other LES.

l **diverse event pairs.** For the same entity pair to be considered l-diverse, there needs to be at least l unique locations for the co-occurring event pairs in it. However, for a location to be counted towards these l locations, the weights of the co-occurring event pairs for that location must be at least 1. Here, one subtle issue is defining a unique location. Intuitively, when datasets have different granularities for space, using the higher granularity ones to define uniqueness would give more accurate results. This could simply be a grid-based division of the space.

Entities x and y could be linked to each other, if they have k co-occurring event pairs in l diverse locations and their datasets do not contain alibi event pairs. Moreover, we only consider entity pairs for which there is no ambiguity, i.e. no two pairs (x, y) and (x, z) that are *k-l* diverse. Setting too low *k-l* values would lead many ambiguous pairs while too high values would lead many false negatives. To find the balance in between these two, we apply *elbow detection* techniques on k, and l distributions.

2.2 Probabilistic Linkage

Besides k-l diversity, we aim to model the spatio-temporal linkage problem using a probabilistic model. For consistency, we try to use the same notation with k-l diversity model as much as possible.

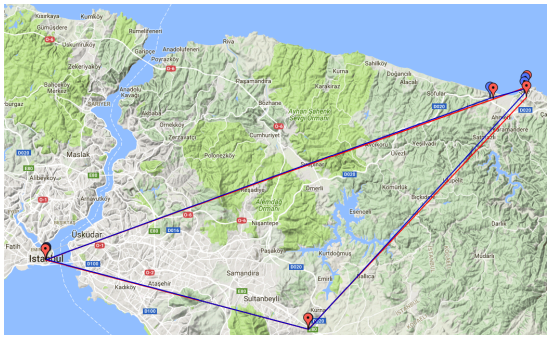

Figure 2: An example of 5 co-occurring events in 3 diverse locations from real world data. All weights are assumed to be 1

Probabilistic model starts by aggregating all of the entities on a common grid using the spatio-temporal features of the datasets. G_k denotes the set of entities in a specific cell in the grid and $|G_k^{\mathcal{I}}(t)|$ denotes the number of entities from dataset $\mathcal{I}$ in cell G_k at some time interval $[t - t + \alpha]$. Let $x \in U_{\mathcal{I}}$ and $y \in U_{\mathcal{E}}$ be two entities, and set of entities co-located with entity x in $U_{\mathcal{E}}$, and set of entities co-located with entity y in $U_{\mathcal{I}}$ at time $[t - t + \alpha]$, in the grid is given as $G_x(t)$, and $G_y(t)$ respectively.

Assuming two sets, S_1 and S_2, where $|S_1| = |S_2| = n$, the number of possible different complete matches (CM) (each element in S_1 has a partner in S_2) between the elements of these sets is $n!$, using trivial combinations without repetition. If the number of elements in the sets are not equal, i.e. $|S_1| = n \neq |S_2| = m$ then the problem turns into choosing m out of n (where $n \geq m$) and calculating complete match with m elements, i.e. $CM(n, m) = \binom{n}{m} m!$.

As the user set of one *LES* is typically not a subset of the second, we define the *partial match (PM)* where only k out of the m elements in S_2 match with k out of n elements in S_1. In this case we also need to choose k out of m and use the complete match, i.e. $PM(n, m, k) = \binom{m}{k} CM(n, k) = \binom{m}{k}\binom{n}{k} k!$.

Let $x \equiv y$ represents entities x, and y are the same real-world entity (match), the probability of a pair of specific two items to match each other is calculated by the number of events where x and y match divided by the number of events in the universal set of all possibilities.

If we are given a single snapshot of the grid at time t the probability of a randomly chosen pair of co-located entities in the different services being the same entity (we are assuming a complete match case only) can be found as following:

$$P(x \equiv y) = \frac{PM(n-1, m-1, m-1)}{PM(n, m, m)} \tag{2}$$

$$= \frac{CM(n-1, m-1)}{CM(n, m)} = \frac{1}{n} \tag{3}$$

$$= \begin{cases} \frac{1}{max(|G_k^{\mathcal{I}}(t)|, |G_k^{\mathcal{E}}(t)|)}, & \text{if } G_x(t) = G_y(t) = G_k \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

This is an intuitive result since a random entity from the smaller set can be equal to any element in the larger set with an equal probability.

If we have more than one sample of the entity (for time slots t_0 to time slot t_T where we don't necessitate the slots to be sequential in time) and the grid we can then use the history of the entity. The probability is similar except taking the tracks of the entities into account:

$$P(x \equiv y) = \begin{cases} \frac{\sum_{k=1}^{m} PM(n-1, m-1, k-1)}{\sum_{k=1}^{m} PM(n, m, k)}, & G_x(t_i) = G_y(t_i)\ \forall t_i \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $n > m$.

An important issue is the decision on the values of n and m. If the entities x and y have l events sharing the same cell and time interval, one must look for all possible pairs that satisfy this property. Since the number of users in the intersection are smaller and is expected to decrease rapidly for large l values the probability of two users in respective services being the same user with the same track will be considerably high.

It is fair to assume that both the systems have considerable amount of common entities which will match. This commonality needs to be calculated empirically by ground truth values. After this value is found we can use this value to limit the k values (e.g. $k \in [k_1 : k_2]$) when calculating the probabilities and the probability in Theorem turns to:

$$P(x \equiv y) = \frac{\sum_{k=k_1}^{k=k_2} PM(n-1, m-1, k-1)}{\sum_{k=k_1}^{k=k_2} PM(n, m, k)} \tag{6}$$

One can relax the condition of equality for the tracking based on the location accuracy of the services and the chosen grid sizes. Moreover, similar to the *diversity* concept of the k-l diversity model, the distance among the shared cells could be used to distinguish between multiple pairs sharing a common user, with close matching probabilities, i.e. $P(x \equiv y) = P(x \equiv z)$.

3. FRAMEWORK

The second component of this doctoral work is the *framework* to perform the linkage efficiently. Naïve record linkage algorithms that compare every pair of records take $\mathcal{O}(n^2)$ time [6], where n is the number of records. Therefore, there are number of techniques implemented, i.e. indexing, blocking, to prune search space of linkage. To perform the linkage in reasonable time, we take advantage of the spatio-temporal structure of the data. To realize effectiveness of the k-l diversity model, we develop an algorithm called *ST-Link* [2]. Our implementation for the probabilistic model is still ongoing.

The *ST-Link* algorithm uses two filtering steps before pairwise comparisons of candidate entities are performed to compute the final linkage. It first distributes entities (users) over coarse-grained geographical regions that we call *dominating grid cells*. Such grid cells contain most of the activities of their users. For two users to link, they must have a common dominating grid. Once this step is over, the linkage is independently performed over each dominating grid cell. To identify the dominating grids, we make a sequential scan over all records, and utilize a quad-tree based index, which limits the area of the smallest grid from below. During the temporal filtering step, *ST-Link* uses a sliding window based scan to build candidate user pairs, while also pruning this

list as alibis are encountered for the current candidate pairs. Finally, our complete linkage model is evaluated over candidate pairs of users that remain following the spatial and temporal filtering steps. During this linkage step, we will need the time sorted events of the users at hand. For that purpose, during the forward scan, we also create a disk-based index sorted by the user id and event time. This index enables us to quickly iterate over the events of a given user in timestamp order, which is an operation used by the linkage step. Also, if one of the datasets is more sparse than the other, it performs the linkage by iterating over the users of the dense datasets first, making sure their events are loaded only once. This is akin to the classical join ordering heuristic in databases.

Our experimental evaluation shows that k-l diversity model is effective (up to 89% precision and 61% recall), yet the efficiency could benefit from a distributed approach. However, distributed processing is challenging due to mobility of users, and the scale of the data. First, distributing records based on their spatio-temporal features would spread records of a single user to multiple processing nodes, hence lead to high inter-machine communications cost. While the concept of dominating grid cells addresses this issue, scalability would still suffer from spatial skew of real data (in our experiments %18 of all records were residing on a single grid out of 120 grids). Since the temporal filtering techniques requires at least one batch of data to reside at the same machine (this issue exists in both models either for filtering or aggregating), records cannot be written to machines in parallel which would lead to low write performance. With these challenges identified, we are going to focus on optimizations of both models to create a single optimized *framework* which could efficiently perform linkage for both models. Such framework would be beneficial for both industry and academia when performing aggregation of semantically different datasets for social good applications, and when benchmarking the linkage research.

4. RELATED WORK

Record Linkage. One of the earliest appearances of the term *record linkage* is by Newcombe et al. [9]. In the literature, it is also referred to as entity resolution (ER), deduplication, object identification, and reference reconciliation, discussed in [4]. Most of the work in this area focus on a single type of databases and define the linked records with respect to a similarity metric. To the best of our knowledge, linking the users of the usage records, specifically targeted at spatio-temporal datasets is novel.

Spatial Record Linkage and Spatial Joins. Many join algorithms are proposed in the literature for spatial data [8]. Spatial record linkage and join algorithms are not directly applicable for spatio-temporal data as they are based on intersection of minimum bounding boxes, one-sided nearest join, or string similarity. Spatio-temporal joins have constraints on both spatial and temporal domains [1]. [10] is a recent work with similar motivation in which calculates weights of matching between users and applies maximum weight partitioning techniques. Their experiments validate the accuracy of this approach, but they do not focus on scalability.

User Identification. Our work has commonalities with the work done in the area of user identification. For instance, de Montjoye et al. [5] have shown that, given a spatio-temporal dataset of call detail records, one can uniquely identify the 95 % of the population by using 4 randomly selected spatio-temporal points. However, linking users is different from identification, as identification leaves whose data to aggregate question unanswered.

5. CONCLUSIONS & RESEARCH PLAN

In this paper, we introduced two linkage models for matching users across location enhanced services, and discussed implementation techniques. We have already realized a single machine implementation of the k-l diversity model with *ST-Link* algorithm. We are now working on validation and implementation of the probabilistic model, and aim to compare these two models with each other. Our single machine implementations showed that both models could benefit from a parallelized distributed implementation. Therefore, we set the development of a distributed and generic framework as the future goal of this doctoral work.

6. REFERENCES

[1] P. Bakalov and V. Tsotras. Continuous spatiotemporal trajectory joins. In *GeoSensor Networks*, volume 4540 of *Lecture Notes in Computer Science*, pages 109–128. Springer Berlin Heidelberg, 2008.

[2] F. Basik, B. Gedik, C. Etemoglu, and H. Ferhatosmanoglu. Spatio-temporal linkage over location-enhanced services. *IEEE Trans. on Mobile Computing*, PP(99):1–1, 2017.

[3] O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, and J. Widom. Swoosh: A generic approach to entity resolution. *The VLDB Journal*, 18(1):255–276, Jan. 2009.

[4] P. Christen. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media, 2012.

[5] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3, 2013.

[6] L. Getoor and A. Machanavajjhala. Entity resolution: Theory, practice & open challenges. In *VLDB Conference (PVLDB)*, 2012.

[7] S. Ghemawat and J. Dean. LevelDB. https://github.com/google/leveldb, 2015.

[8] E. H. Jacox and H. Samet. Spatial join techniques. *ACM Trans. Database Syst.*, 32(1), Mar. 2007.

[9] H. B. Newcombe, J. M. Kennedy, S. J. Axford, and A. P. James. Automatic linkage of vital records: Computers can be used to extract "follow-up" statistics of families from files of routine records. *Science*, 130(3381):954–959, 1959.

[10] C. Riederer, Y. Kim, A. Chaintreau, N. Korula, and S. Lattanzi. Linking users across domains with location data: Theory and validation. In *Proc. of the 25th Int. Conf.on WWW*, pages 707–719, 2016.

[11] A. Skovsgaard, D. Sidlauskas, and C. Jensen. Scalable top-k spatio-temporal term querying. In *IEEE Int. Conference on Data Engineering (ICDE)*, pages 148–159, March 2014.