Chapter 9.
AI-based analysis methods in spatio-temporal data mining

Thomas Liebig

9. Introduction

Proliferation of pervasive devices capturing sensitive data streams, such as mobility records, raises concerns about individual privacy. In particular, automatic reasoning based on spatio-temporal information can cause disclosure of private information. The methods used by spatio-temporal data mining are part of the vast subject of artificial intelligence, more precisely of data mining and machine learning. Historically, the methods for spatio-temporal analysis originate from physics (focusing on moving objects), and statistics (focusing on data analysis). With the advance of probabilistic learning methods in computer science and artificial intelligence, spatio-temporal data mining itself became a mature science.

We continue this section with prerequisites for understanding spatio-temporal methods and continue in the following sections, with an overview on spatio-temporal data mining and privacy aspects of spatio-temporal analysis.

9.1.1. Time geography

The fundamental relation between space and time was formulated by Minkowski[1]. Whereas space was before considered as a three-dimensional homogeneous and isotropic extent defined by Newton[2], Minkowski firstly introduces

[1] *H. Minkowski,* Die Grundgleichungen für die elektromagnetischen Vorgänge in bewegten Körpern, *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse* (Berlin) 1908 (1908), pp. 53–111.

[2] *Newton* et al., The mathematical principles of natural philosophy by Sir Isaac Newton translated into English by *A. Motte* to which are added, Newton's system of the world a short comment on, and defence of, the Principia, by *W. Emerson*; with The laws of the moon's motion according to gravity, *by J. Machin*, A new ed. carefully rev. and corr. by *W. Davis*.(Printed for H.D. Symond, London, 1803), http://nla.gov.au/nla.cat-vn994402.

a four-dimensional extrapolation of the prior three-dimensional one. He combines time *t* with the previously introduced spatial extent *X=(x,y,z)* (or *X=(x-,y)* for two dimensional coordinate systems) by introducing *ict* as an additional dimension, with $i^2=-1$ and *c* is the maximal speed of light. Thus the sphere that a flash of light creates in three dimensional space $x^2+y^2+z^2=r^2$ becomes in four dimensions:

$$x^2+y^2+z^2 = r^2 = c^2t^2$$

$$x^2+y^2+z^2+(ict)^2 =0, \text{ with } i^2=-1 .$$

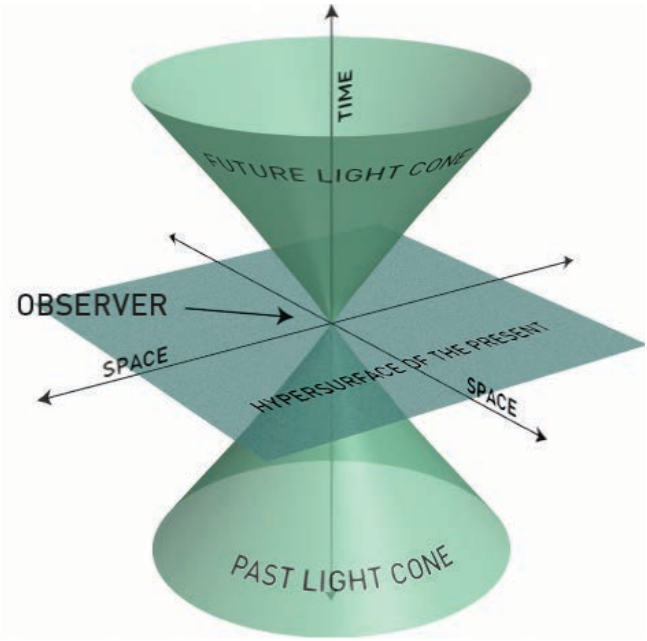

Fig. 9: Light Cone in Minkowski Space-Time[3].

In four dimensions, this equation describes a cone, depicted in Figure 9. Hence it is called the *light-cone*. The coordinates in this Minkowski space *(x,y,z,ict)* are called *Minkowski Coordinates* whereas *(x,y,z,ct)* are called *Galilei Coordinates*[4].

[3] Wikimedia Commons, *Example of a light cone, by Stib at en.wikipedia* http://creativecommons.org/licenses/by-sa/3.0/)], 2005.

[4] *E. Schmutzer*, Relativitätstheorie– aktuell: ein Beitrag zur Einheit der Physik, 4th ed. (Leipzig: Teubner, 1989), p. 180.

Making the reasonable assumption, that the speed of light is the limit for all speeds, a point at *(x,y,z,ict)* may not be affected by anything except the lower half cone, therefore called *past*. The upper half cone describes its *future*, compare Figure 9. This close relationship between space and time is also expressed in the definition of the spatial unit length *meter [m]*.

Definition [meter]: A *meter [m]* is the distance light passes in 1/299.792.458~seconds in vacuum[5].

Note that this definition implies that the speed of light *c* is given by 299.792.458 *m/s*. Whilst this space still possesses the Euclidian geometry, which describes a rectilinear space, later the curvilinear Riemann geometry was considered which allows explanation of gravity effects.

However, for most spatial modelling applications, the gravity effects and relativistic influences of mass objects are not relevant. Therefore the four-dimensional space-time with rectilinear Euclidian geometry is sufficient for this work. Whereas physicists have applied this model since 1908, in 1970 space-time was first incorporated in geography using the term *Time-Geography*[6] to visualize and analyse the motion of pedestrians. Thus the spatial coordinate system, in geography often represented by *x,y* due to the two-dimensionality of maps, is extended by time *t* as a third dimension. This results in a continuous three-dimensional line-visualization of individual movement. In physics it is common to call this line a *trajectory*[7]. Introducing a maximal speed for the motion of people (similarly to the maximum speed of light above) the transition possibilities between two points in (x_1,y_1,t_1) and (x_2,y_2,t_2) result in a so called *space-time prism*[8]. This is a volume in space-time which is formed by the intersection of the *future* of point (x_1,y_1,t_1) and the *past* of (x_2,y_2,t_2), compare Figure 10. When projecting this prism onto the *x,y* plane it defines the *possible path area*.

[5] National Institute of Standards and Technology, *International System of Units (SI)* (Gaithersburg: National Institute of Standards / Technology, March 2008), http://physics.nist.gov/Pubs/SP330/sp330.pdf.

[6] *T. Hägerstrand*, What about people in Regional Science?, *Papers in Regional Science* 24, no. 1 (1970): 6–21, http://dx.doi.org/10.1007/BF01936872.

[7] *E. Schmutzer*, Relativitätstheorie– aktuell : ein Beitrag zur Einheit der Physik…

[8] *B. Lenntorp*, Paths in space-time environments: a time-geographic study of movement possibilities of individuals, Meddelanden från Lunds universitets geografiska institution (Royal University of Lund, Dept. of Geography, 1976), http://books.google.de/books?id=2UMSAQAAIAAJ.

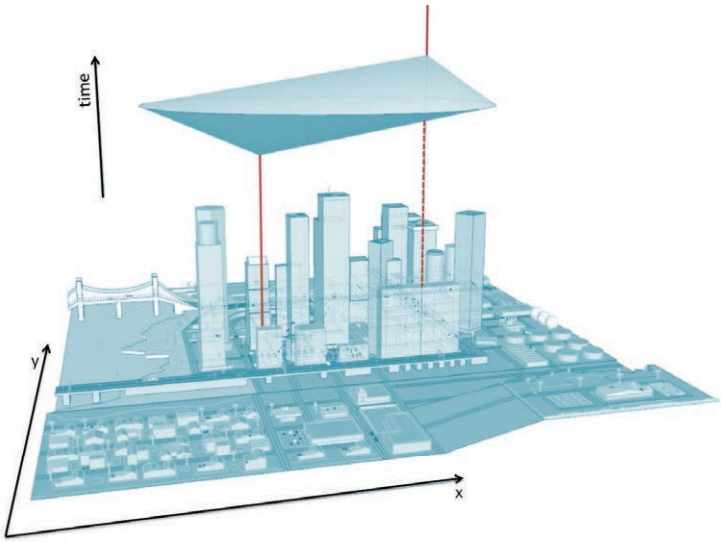

Fig. 10: Example of space-time prism representation of movement between two places in space-time. The red lines mark that the person spends some time at the left building and later on in the right building. The transition in between the two spatio-temporal locations is bounded by the space-time prism in the upper part of the picture.

A trajectory of a moving object, which represents its positions in space-time, can then be defined as:

Definition [trajectory]: A set of space-time points *(x,y,z,t)* is called *trajectory* S (or *world-line*) of a moving object, if for every contained time-stamp *t*, exactly one spatial point *(x,y,z)* is contained in *S* (*uniqueness*) and temporally subsequent points are contained in their light-cones (*continuity*).[9]

Thus, in a trajectory the spatial component *X* can be regarded as a function of time *X=f(t)*. In contrast to a moving point, the trajectory of a mass object is constrained by physical properties such as *inertia*, *impulse*, *spin* and *gravity*. This results in stronger constraints for the continuity of trajectories. Therefore, the trajectory of a mass object can be expected to be *smooth*, i.e., *f(t)* is continuously differentiable. Recent work on kinetic space-time prisms incorporates the physical behaviour of mass objects[10].

[9] *H. Minkowski*, Raum und Zeit, Vortrag, Gehalten auf der 80. Natur-Forscher-Versammlung zu Köln am 21. September 1908 (B.G. Teubner, 1909).

[10] *B. Kuijpers, H. J. Miller & W. Othman*, Kinetic space-time prisms, [in:] Proceedings of the 19th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS, GIS (ACM, 2011), 162– 170.

Time is a continuous and linear extent with the second [s] as its unit. Possible spatial reference systems for the spatial component of the space-time points are so-called geo-reference systems, *such as WGS84*[11] which locates any point on earth by its ellipsoidal coordinate triple *longitude*, *latitude* and *altitude*. As WGS84 became part of Open Geographic Consortium specifications, this geo-reference system is widely used. Another often used Cartesian coordinate system is given directly by maps, having the *x* and *y* axis perpendicular in the plane of a map image (and possibly the third *z* axis pointing orthogonally from the map surface).

9.1.2. Digital data representation

Automatic processing of trajectories and analysis of movement both require digital data storage. Two possibilities for digital storage of spatial data evolved: grid and vector representation. Grid representation aggregates spaces and straight contours are approximated by tessellations (see Figure 11), whereas vector representation preserves the spatial contours (see Figure 11).

Vectorized data needs less memory and it allows easy integration of additional dimensions. In raster space this would imply transition from pixel to voxel-space. Furthermore, vector models provide easy mapping of data between various geographic coordinate systems with different ranges or precision.

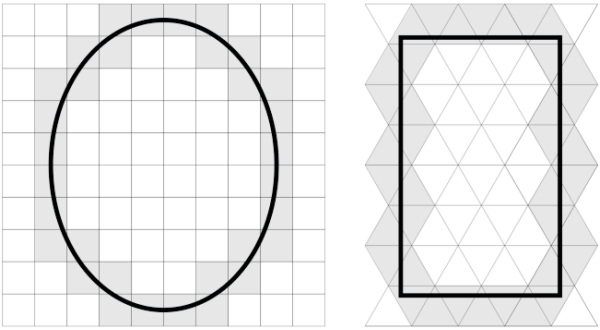

Fig. 11: Tessellated Spatial Objects (image credits[12]*).*

[11] National Imagery and Mapping Agency, *Department of Defense World Geodetic System 1984: its definition and relationships with local geodetic systems,* technical report TR8350.2 (St. Louis, MO, USA: National Imagery and Mapping Agency, January 2000), http://earth-info.nga.mil/GandG/publications/tr8350.2/tr8350_2.html.

[12] *T. Liebig*, Spatio -- Temporal Data Mining with Bayesian Networks (Diploma Thesis, Chemnitz University of Technology, 2007).

Primitive spatial vector object types are point, line, area, network and their compound[13] shown in Figure 12.

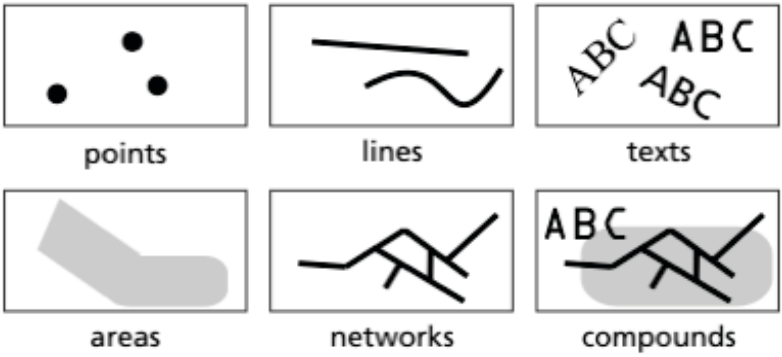

Fig. 12: Spatial Vector Objects (image credits[14])

For the exchange and storage of spatio-temporal data the Open Geographic Consortium defined open file format standards, protocols and interfaces. The most popular ones are the Keyhole Markup Language (KML), an XML notation for description of spatial vector objects in WGS84, and Geographic Markup Language (GML) which can model arbitrary geographic data and not just its visualization. The interfaces and their associated protocols define the connections between software modules or devices. For example, the transmission of sensor readings that are located at a specific location should adhere to the *Sensor Observation Service Protocol* (SOS), whereas the transmission of map information from a server to a map-service could be either done in vector format (using the Web Feature Service protocol) or in raster format. The creation of these open communication standards led to a modularization of previously proprietary (closed) geographical information systems.

9.2. Spatio-temporal data mining

Although not all spatially-related data contain coordinates (e.g. some Twitter messages) they mostly contain information on locations or moving objects.

In both dimensions, space and time, data items have limited validity. For example, a message containing the weather information at a particular spatio-temporal coordinate is invalid in future or at large distance. The GPS information of a moving object (e.g. a vehicular position) loses validity immediately after being recorded.

[13] *N. Bartelme*, Geoinformatik: Modelle, Strukturen, Funktionen (Springer, Berlin, 1995),
[14] *T. Liebig*, Spatio -- Temporal Data…

Thus, analysis models developed on spatio-temporal measurements have to incorporate the latest data samples and need to perform in real-time. This does not preclude learning from historic data samples in order to compare current situations with the past and project it into the future.

For a comprehensive introduction to spatio-temporal data mining we refer to the book[15], which resulted from the GeoPKDD project funded by the European Commission under the Sixth Framework Programme, IST-6FP-014915.

Spatio-temporal data comes in a variety of forms and representations, depending on the domain, the observed phenomenon and the observation method. In principle, there are three types of spatio-temporal data: spatial time series, events and trajectories.

A spatial time series consists of tuples (attribute, object, time, location).

An event of a particular type eventi is triggered from a spatial time series under certain conditions and contains the tuples verifying these conditions (eventi, objectn, timen, locationn).

A trajectory is a spatial time series for a particular objecti. It contains the location at a given time and is a series of tuples (objecti, timen, locationn). Every timestamp timen is contained at most once.

9.2.1. Frequent patterns

The challenge of frequent pattern mining is the identification of frequently co-occurring sets of items or more complex patterns (that describe relations between space and time). Input items can be elements of spatial time series, events, or the tuples of trajectories. Outputs are frequent sets of these items. A common algorithm for mining these data sets for frequent item sets is the apriori algorithm that generates candidates for frequent item sets as unions of smaller frequent item sets. A common parameter for frequent item mining is the minimum support which is a threshold to distinguish between frequent and infrequent sets of items. As the coordinates in trajectories may be too fine-grained to identify frequently co-visited places, the T-pattern algorithm[16] extracts spatial regions from the trajectories which are frequently visited and returns frequent visit patterns between them.

9.2.2. Classification, regression, prediction

For spatio-temporal data, the group of regression and prediction tasks originates in geostatistics. The idea is to formulate a model of the data and use the

[15] *F. Giannotti & D. Pedreschi*, Mobility, Data Mining and Privacy -Geographic Knowledge Discovery (Springer, 2008).

[16] *F. Giannotti* et al., Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, [in:] *KDD* (ACM, 2007), pp. 330–339, http://doi.acm.org/10.1145/1281192.1281230.

model to infer unknown values. While classification works on discrete values, regression is for continuous ones. Prediction infers a label (class membership, target value) using a decision or regression function that was learned from complete instances. Classification, Regression and Prediction are applicable to all three data types: spatio-temporal time series, events and trajectories. A characteristic of spatio-temporal data (if constrained by space and time e.g. traffic flow in a street network) is the autocorrelation between the values, whereby close values are more related than distant ones[17]. A commonly used regression method from geo-statistics, Kriging[18], models the autocorrelation with variograms that describe the correlation between spatio-temporal values at different positions as a function of their distance.

Geographically weighted regression[19] is another commonly used method which models an unknown value as a linear combination of observed values; the weights of the observed values vary for different locations. Spatial k-nearest neighbour algorithm[20] infers a data point as a weighted sum of the k nearest points. Classification of the tuples in spatio-temporal time series is important for outlier detection – a possible method is usage of 1-class support vector machines. They describe the subspace of normal observations by a minimum enclosing ball, with outliers outside the ball. As the split decision cannot necessarily be described spherical in the attributes of the observations, they are transferred to a feature space. Instead of computing the transformation for all incoming data, an inner product in feature space is defined which can be computed directly using the observations. The inner productmaps two observations to a real number. Core vector machines compute an approximation of the minimum enclosing ball with constant space and time requirements[21] which contains all observations when scaled by a factor of $(1+e)$, with $e>0$. Prediction of future values in a spatio-temporal time series has to respect Tobler's law, whereby close values correlate more than distant ones[22]. This autocorrelation can be directly reflected by so-called graphical models. Every observation at a location for a given time is assigned to a random variable. In a graphical model the conditional dependencies of the probability distributions for the random variables are denoted by edges. A recently developed

[17] *W. Tobler*, A Computer Movie Simulating Urban Growth in the Detroit Region, *Economic Geography* 46, no. 2 (1970), pp. 234–240.

[18] *D. G. Krige*, A Statistical Approach to Some Mine Valuation and Allied Problems on the Witwatersrand (1951).

[19] *A.S. Fotheringham, Ch. Brunsdon & M. Charlton*, Geographically Weighted Regression: The Analysis of Spatially Varying Relationships (Wiley, 2002).

[20] *X. Gong & F. Wang*, "Three Improvements on KNN-NPR for Traffic Flow Forecasting," in *Proceedings of the 5th International Conference on Intelligent Transportation Systems* (IEEE Press, 2002), 736–740.

[21] *M. Badoiu & K.L. Clarkson*, Optimal core-sets for balls, *Comput. Geom.* 40, no. 1 (2008): 14–22.

[22] *Tobler*, A Computer Movie Simulating Urban Growth in the Detroit Region.

method[23] uses Markov Random Fields where every random variable for a particular time slice is connected with its direct neighbours and their ancestors from a previous time slice. This method becomes efficient through the regularization of the optimization step, which saves computation when the measurements remain about the same.

9.2.3. Clustering and similarity search

Clustering focuses on the identification of groups of objects (clusters) where the elements of a group are similar and the elements of different clusters are dissimilar. For non-overlapping clusters, the result is a partitioning of the data. Clustering can be applied to all three spatio-temporal data types: events, spatio-temporal time series and trajectories. A commonly applied spatio-temporal clustering method is the density-based algorithm DBSCAN[24] that computes the spatio-temporal density of the data points and extracts clusters as highly dense sets of points. DBSCAN defines similarity between points based on their spatio-temporal distance and follows Tobler's law. Other similarity measures e.g. between time series, between the properties of spatio-temporal events, or between trajectories can be defined. The well-known k-Means algorithm has recently been turned into an algorithm for streaming data[25]. Another method that can be applied for cluster analysis is OPTICS[26]. The Voronoi tessellation method[27] partitions space based on a set of spatial points. Every spatial point in this set is associated with a surrounding polygon comprising all spatial locations that are not closer to any other point contained in the set. Zeinalipour-Yazti et al. introduced the distributed spatio-temporal similarity search problem[28]: given a query trajectory Q, the purpose of the proposed algorithm was to find the trajectories that follow a motion

[23] *N. Piatkowski, S. Lee & K. Morik*, "Spatio-temporal random fields: compressible representation and distributed estimation,"*Machine Learning*, 2013, 1–25.

[24] *M. Ester* et al., A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, in Second International Conference on Knowledge Discovery and Data Mining (AAAI Press, 1996), pp. 226–231.

[25] *H. Fichtenberger* et al., BICO: Birch meets Coresets for k-means, in *Algorithms– ESA 2013* (Springer Berlin / Heidelberg, 2013).

[26] *M. Ankerst* et al., OPTICS: ordering points to identify the clustering structure, *SIGMOD Rec.* (New York, NY, USA) 28, no. 2 (June 1999), pp. 49–60.

[27] *G.F. Voronoï*, Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Deuxième mémoire. Recherches sur les parallélloèdres primitifs., *Journal für die reine und angewandte Mathematik (Crelle's Journal)*, no. 134 (December 1908), pp. 198–287, http ://dx.doi.org/10.1515/crll.1908.134.198.

[28] *D. Zeinalipour-Yazti, S. Lin & D. Gunopulos*, Distributed spatio-temporal similarity search, [in:] *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06 (Arlington, Virginia, USA: ACM, 2006), pp. 14–23, doi:10.1145/1183614.1183621, http://doi.acm.org/10.1145/1183614.1183621.

similar to Q, when each of the target trajectories is segmented across a number of distributed nodes. Two algorithms were proposed, UB-K and UBLB-K, which combine local computations of lower and upper bounds on the matching between the distributed subsequences and Q. The approach generates the desired result without pulling together all the distributed subsequences over the fundamentally expensive communication medium. The described problem finds applications in a wide array of domains, such as cellular networks, wildlife monitoring, and video surveillance.

9.2.4. Geo-coding and map matching

The transformation of geo-locations is subject to geo-coding and map matching. As geo-coding aims at identification of a location for a spatio-temporal event without direct reference to an identifier (e.g. a text message that mentions a street name), map matching transfers the coordinates of events or trajectories from one reference system to another one. Map matching tasks are common for GPS trajectories which are recorded in the WGS84 reference system and have to be mapped to a discrete street network graph. The spatial extents of the street segments are used for distance calculations between the street network and a particular point. The algorithm in Lou et al.[29] uses these distances to generate a set of closest segments for every point of a trajectory (the segment candidates). For the identification of the most likely street segment among these candidates a routing algorithm is used which makes assumptions about individual mobility. In the result every point of the trajectory is matched to a street segment.

9.3. Privacy threats in spatio-temporal data analysis

From a business perspective, mobility data with sufficiently precise location estimation are often valuable for enabling various location-based services; from the perspective of privacy advocates, such insights are often deemed a privacy threat or a privacy risk. Location privacy risks can arise if a third-party acquires a data tuple (user ID, location), which proves that an identifiable user has visited a certain location. In most cases, the datum will be a triple that also includes a time field describing when the user was present at this location. Although in theory there are no location privacy risks if the user cannot be identified or if the location cannot be inferred from the data, in practice it is difficult to determine when identification and such inferences are possible. In the following we reflect an overview on privacy aware learning by Wainwright et al.[30]:

[29] *Y. Lou* et al., Map-matching for low-sampling-rate GPS trajectories, [in:] Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS '09 (New York, NY, USA: ACM, 2009), pp. 352–361.

[30] *M.J. Wainwright,, M.I. Jordan & J.C. Duchi*, Privacy aware learning, *Advances in Neural Information Processing Systems*. 2012

There is a long history of research at the intersection of privacy and statistics, going back at least to the 1960s, when Warner[31] suggested privacy-preserving methods for survey sampling, and to later work related to census-taking and presentation of tabular data[32]. More recently, there has been a large amount of computationally-oriented work on privacy[33].

In this section, we provide a brief overview on the subject. We refer the interested reader to the comprehensive survey by Dwork[34].

Most work on privacy attempts to limit disclosure risk: the probability that some adversary can link a released record to a particular member of the population or identify that someone belongs to a dataset that generates a statistic[35]. In statistical literature, work on disclosure limitation and so-called linkage risk, for example as in the framework of Duncan and Lambert[36], has yielded several techniques for maintaining privacy, such as aggregation, swapping features or responses between different data points, or perturbation of data. Other authors have proposed measures for measuring the utility of released data (e.g.,[37]). The currently standard measure of privacy is differential privacy, due to Dwork et al.[38], which roughly states that the answer to a data

[31] *S.L. Warner*, Randomized response: A survey technique for eliminating evasive answer bias, *Journal of the American Statistical Association* 60, no. 309 (1965), pp. 63–69.

[32] *S.R. Ganta, S.P. Kasiviswanathan & A. Smith*, Composition attacks and auxiliary information in data privacy, [in:] Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (ACM, 2008), pp. 265–273.

[33] *C. Dwork* et al., Calibrating noise to sensitivity in private data analysis, in *Theory of cryptography* (Springer, 2006), 265–284; *C. Dwork*, Differential privacy: A survey of results, in *Theory and applications of models of computation* (Springer, 2008), pp. 1–19; *S. Zhou, L. Wasserman & J.D. Lafferty*, Compressed regression [in:] Advances in Neural Information Processing Systems (2008), pp. 1713–1720; *L. Wasserman & S. Zhou*, A statistical framework for differential privacy, *Journal of the American Statistical Association* 105, no. 489 (2010), pp. 375–389; *R. Hall, A. Rinaldo & L. Wasserman*, Random differential privacy, *arXiv preprint arXiv:1112.2680*, 2011, I. Dinur & K. Nissim, Revealing information while preserving privacy, [in:] Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (ACM, 2003), pp. 202–210; *A. Blum, K. Ligett & A. Roth*, A learning theory approach to noninteractive database privacy, *Journal of the ACM (JACM)* 60, no. 2 (2013), p. 12; *K. Chaudhuri, C. Monteleoni & A. D. Sarwate*, "Differentially private empirical risk minimization,"

[34] *Dwork*, Differential privacy: A survey of results.

[35] *G. T. Duncan & D. Lambert*, Disclosure-limited data dissemination, *Journal of the American statistical association* 81, no. 393 (1986), pp. 10–18; J.P. Reiter, Estimating risks of identification disclosure in microdata, *Journal of the American Statistical Association* 100, no. 472 (2005), pp. 1103–1112; *A. F. Karr* et al., A framework for evaluating the utility of data altered to protect confidentiality, *The American Statistician* 60, no. 3 (2006), pp. 224–232.

[36] *Duncan & Lambert*, Disclosure-limited data dissemination.

[37] *Karr* et al., "A framework for evaluating the utility of data altered to protect confidentiality*"; L.H. Cox, A. F. Karr & S. K. Kinney*, Risk-Utility Paradigms for Statistical Disclosure Limitation: How to Think, But Not How to Act, *International Statistical Review* 79, no. 2 (2011): 160–183.

[38] *Dwork* et al., Calibrating noise to sensitivity in private data analysis.

query must not depend too much on the samples, and it should be difficult, given the answer to a query, to ascertain whether a vector is contained in the used dataset.

9.3.1. Threats from moving spatial sensors

Recently, several location privacy incidents were reported in the media. A famous incident regards the case of Apple[39], where 3G Apple iOS devices were reported to store the location of their mobile users' in unencrypted form for a period of over one year. This precise location information was stored without the knowledge of the users and was transmitted to the iTunes application during the synchronization of the device. According to Apple, the stored location information was not used to track the users but was attributed to a programming error which was later fixed with a software update. Google was also reported to be using precise location data, collected from users' mobile devices, to improve the accuracy of its navigation services[40], while Microsoft[41] recently admitted that their camera application in Windows Phone 7 ignored the users' privacy settings to disable transmitting their location information to Microsoft. In response to this incident, the company issued a software update. Although the above-mentioned privacy incidents did not lead to actual harm caused to the individuals due to the lack of location privacy, the continual flurry of such breaches is worrying as it becomes evident that sensitive location information may easily fall into the wrong hands[42]. In the following subsections, we elaborate on different types of privacy risk that can lead to user identification or disclose sensitive location.

9.3.2. Collection of location information with assigned user ID

This is the most trivial case, as long as the location of the user is estimated with sufficient accuracy for providing the intended location based service (LBS). In cases where the location is not yet precise enough, various techniques (e.g. fusion of several raw location data from various sensors) allow for improved accuracy.

Example 3.2.1: A cellular mobile network operator (MNO) routinely stores tuples of the form (cell ID and sector ID, user ID), e.g. within the call detail records data (CDR) for billing purposes.

[39] *N. Bilton*, 3G Apple iOS Devices Are Storing Users' Location Data, The New York Times, Published: April 20, 2011, 2011.

[40] *M. Helft*, Apple and Google Use Phone Data to Map the World, The New York Times, Published: April 25, 2011, 2011.

[41] *D. McCullagh*, Microsoft collects locations of Windows phone users, CNet News, Published: April 25, 2011, 2011.

[42] *N. Bilton*, Holding Companies Accountable for Privacy Breaches, The New York Times, Published: April 27, 2011.

Example 3.2.2: A smartphone app gets the GPS-location for a user who has already been identified, e.g. by his/her login to the application or by a payment transaction.

Example 3.2.3: From a smartphone, a smartphone application provider receives the IDs and signal strengths of several nearby transmitters (base stations, WiFi devices, etc.). Based on previously established maps of these transmitters, the application provider is able to estimate a more precise location.

Additionally, application providers may have direct access to a variety of publicly available spatial and temporal data such as geographical space and inherent properties of different locations and parts of the space (e.g. street vs. park) various objects existing or occurring in space and time: static spatial objects (having particular constant positions in space), events (having particular positions in time), and moving objects (changing their spatial positions over time). Such information either exists in explicit form in public databases like OpenStreetMap, WikiMapia or in smartphone application providers' data centers, or can be extracted from publicly available data by means of event detection or situation similarity assessment[43]. Combining such information with positions and identities of users allows deep semantic understanding of their habits, contacts, and lifestyle.

9.3.3. Collection of anonymous location information

When location data is collected without any obvious user identifiers, privacy risks are reduced and such seemingly anonymous data is usually exempted from privacy regulations. It is, however, often possible to re-identify users based on quasi-identifying data that have been collected. Therefore, the aforementioned risks can apply even to such anonymous data. The degree of difficulty in re-identifying anonymized data depends on the exact details of the data collection and anonymization scheme as well as on the adversaries' access to background information. Consider the following examples: Re-identifying individual samples. Individual location records can be re-identified through observation attacks[44]. The adversary knows that user Alice was the only user in location (area) *l* at time *t*, perhaps because the adversary has seen the person at this location or because records from another source prove it. If the adversary now finds an anonymous datum (*l, t*) in the collected mobility data, the adversary can infer that this datum could only

[43] *G.L. Andrienko* et al., From movement tracks through events to places: Extracting and characterizing significant places from mobility data, in *IEEE VAST* (2011), 161–170; *G.L. Andrienko* et al., Identifying Place Histories from Activity Traces with an Eye to Parameter Impact, *IEEE Trans. Vis. Comput. Graph.* 18, no. 5 (2012): 675–688.

[44] *C. Y. T. Ma* et al., Privacy vulnerability of published anonymous mobility traces, in *Proceedings of the sixteenth annual international conference on Mobile computing and networking*, MobiCom '10 (New York, NY, USA: ACM, 2010), 185–196.

have been collected from Alice and has therefore re-identified the individual. In this trivial example, there is actually no privacy risk from this re-identification because the adversary knew a priori that Alice was at location *l* at time *t*, so the adversary has not learned anything new. There are, however, three important variants of this trivial case that can pose privacy risks. First, the anonymous datum may contain a more precise location *l'* or a more precise time *t'* than the adversary knew about a priori. In this case, the adversary learns this more precise information. Second, the adversary may not know that Alice was at *l* but simply know that Alice is the only user who has access to location *l*. In this latter case, also referred to as restricted space identification, the adversary would learn when Alice was actually present at this location. Third, the anonymous datum may contain additional fields with potentially sensitive information that the adversary did not know before. Note, however, that such additional information can also make the re-identification task easier.

Re-identifying time-series location data. Re-identification can also become substantially easier when location data is repeatedly collected and time series location traces are available. We refer to time series location traces, rather than individual location samples, when it is clear which set of location samples was collected from the same user (even though the identity of the user is not known). For example, the location data may be stored in separate files for each user or a pseudonym may be used to link multiple records to the same user.

Example 3.3.1: A partner of the mobile network operator (MNO) has obtained anonymized traces of a user, e.g. as a sequence of CDRs where all user IDs have been removed. While this looks like anonymous location data, various approaches exist to re-identify the user associated with these mobility traces. One approach is to identify the top 2 locations where the user spent most time. This corresponds in many cases to the user's home and work locations. Empirical research has further observed that the pair (home location, work location) is often already sufficient to identify a unique user[45]. A recent empirical study[46] explains various approaches for re-identification of a user. Another paper has analyzed the consequences for privacy law and its interpretation of increasingly strong re-identification methods[47]. Further re-identification methods for location data rely on various inference and data mining techniques.

[45] *P. Golle & K. Partridge*, On the Anonymity of Home/Work Location Pairs, in *Pervasive Computing*, ed. *H. Tokuda* et al., vol. 5538, Lecture Notes in Computer Science (Springer Berlin / Heidelberg, 2009), 390–397.

[46] *H. Zang & J. Bolot*, "Anonymization of location data does not work: a large-scale measurement study," in *Proceedings of the 17th annual international conference on Mobile computing and networking*, MobiCom '11 (New York, NY, USA: ACM, 2011), 145–156.

[47] *Ohm P.*, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, *UCLA Law Review, Vol. 57, p. 1701, 2010*, 2009.

9.3.4. Collection of data without location

Even in the absence of actual location readings provided by positioning devices, location disclosures may occur by means of other modern technologies. Recent work by Jun et al. demonstrated that the complete trajectory of a user can be revealed with 200m accuracy by using accelerometer readings, even when no initial location information is known[48]. What is even more alarming is that accelerometers, typically installed in modern smartphones, are usually not secured against third-party applications, which can easily obtain such readings without requiring any special privileges. Acceleration information can thus be transmitted to external servers and be used to disclose user location even if all localization mechanisms of the mobile device are disabled.

Another example of privacy disclosures in mobile devices regards the monitoring of user screen taps through the use of accelerometer and gyroscope readings. Recent work by Miluzzo et al.[49] demonstrated that user inputs across the display, including the on-screen keyboard, of a mobile device can be silently identified with high precision through the use of motion sensors and machine learning analysis. Their prototype implementation achieved tap location identification rates of as high as 90% in accuracy, practically demonstrating that malevolent applications installed in mobile devices may severely compromise the privacy of the users.

Last but not least, several privacy vulnerabilities may arise through the various resource types that are typically supported and communicated by modern mobile phone applications. Hornyack, et al.[50] examined several popular Android applications which require both internet access and access to sensitive data, such as location, contacts, camera, microphone, etc. for their operation. Their examination showed that almost 34% of the top 1100 popular Android applications required access to location data, while almost 10% of the applications required access to the user contacts. As can be anticipated, access of third-party applications to such sensitive data sources may lead both to user re-identification and to sensitive information disclosure attacks, unless privacy enabling technology is in place.

Example 3.4.1: During a vacation, a user has taken many photographs, which are all tagged with a time-stamp but not geo-coded. There are, however, techniques

[48] *H. Jun* et al., "ACComplice: Location inference using accelerometers on smartphones," in *Communication Systems and Networks (COMSNETS),* 2012 Fourth International Conference on (2012), pp. 1–9.

[49] *E. Miluzzo* et al., "Tapprints: your finger taps have fingerprints," [in:] *Proceedings of the 10th international conference on Mobile systems, applications, and services*, MobiSys '12 (New York, NY, USA: ACM, 2012),pp. 323–336.

[50] *P. Hornyack* et al., "These aren't the droids you're looking for: retrofitting android to protect data from imperious applications," [in:] Proceedings of the 18th ACM conference on Computer and communications security, CCS '11 (New York, NY, USA: ACM, 2011), pp. 639–652.

to assign a geo-location to most images, as long as these they contain some unique features. Similarly, there are techniques to assign real names to most persons in the photographss, e.g. by using tools or crowdsourcing as provided e.g. by a social network or other platforms to store photos. Having times and places for a photo stream one might reconstruct precise trajectories.

Example 4.4.2: An app is able to continuously read the accelerometer of a handset. This enables it to reconstruct a 3D trace of the user's movements.

9.3.5. Episodic movement data

Most of the data collected by mobile phone network operators are referred to as "Episodic Movement Data": data about spatial positions of moving objects where the time intervals between the measurements may be quite large and therefore the intermediate positions cannot be reliably reconstructed by means of interpolation, map matching, or other methods. Three main types of uncertainty distinguish episodic from continuous movement data and these were identified in[51]. First, the most common type of uncertainty is the lack of information about the spatial positions of the objects between the recorded positions (*continuity*), which is caused by large time intervals between the recordings and by missed recordings. Second, a frequently occurring type of uncertainty is low granularity of the recorded positions (*accuracy*). Due to these two types of uncertainty, episodic movement data cannot be treated as continuous trajectories, i.e., unbroken lines in the spatio-temporal continuum such that some point on the line exists for each time moment. Third, the number of recorded objects (*coverage*) may also be uncertain due to the usage of a service or due to the utilized sensor technology. For example, one individual may carry two or more devices, which will be registered as independent objects. Some recording techniques only capture devices which are turned on. The activation status may change as a device carrier moves. As discussed above, the information encoded in episodic data is much smaller than in continuous movement data. Many of the existing data analysis and privacy preservation methods designed for dealing with movement data are explicitly or implicitly based on the assumption of continuous objects movement between the measured positions and are therefore not suitable for episodic data. However, due to the increased availability of mobile phone data, analysis methods for episodic movement data and the retrieval of data for unobserved locations are rapidly evolving. Though such techniques pose a privacy risk, they also help us understand what sensitive information can be extracted from location traces.

[51] *N. Andrienko* et al., Visual Analytics for Understanding Spatial Situations from Episodic Movement Data, *KI– Künstliche Intelligenz*, 2012, 241–251, http://dx.doi.org/10.1007/s13218-012-0177-4.

9.3.6. Threats from stationary spatial sensors

Smartphones became a convenient way to communicate and access information. With the integration of GPS sensors, mobility mining was pushed forward[52]. The mobility information of multiple devices is usually stored on a server which performs analysis in order to extract knowledge on movement behaviour. In the easiest case this is the number of visitors to specific places. The processing of the data streams became infeasible for large use cases, where millions of people are monitored and massive data streams have to be processed. In such Big Data scenarios, the expensive computation (matching and counting in individual, continuous GPS streams) is split among the parties and only the aggregation step remains on the server (In contrast Boutsis and Kalogeraki[53] present a method that distributes the query). Thus, continuous movement records (GPS) are reduced to episodic movement data[54] consisting of geo-referenced events and their aggregates: number of people visiting a certain location, number of people moving from one location to another one, and so on. The preprocessing of the GPS data streams is then performed locally on the location based devices and the aggregation is subject to crowd sourcing. Recent work focusses on in-situ analysis to monitor location based events (visits[55], moves[56]) or even more complex movement patterns[57] in GPS streams. In all cases a database with the locations or patterns of interest is provided in advance, and the mobile device computes event-histograms for succeeding time-slices. These histograms are much smaller and may be aggregated by the server in order to achieve knowledge on current movement behaviour. However, the transmission of such individual movement behaviour still poses privacy risks[58]. Even access by third parties compromises individual privacy as recent disclosures on the NSA PRISM program reveal. The devices monitor daily behaviour and thus

[52] *Giannotti & Pedreschi*, Mobility, Data Mining and Privacy – Geographic Knowledge Discovery.

[53] *I. Boutsis & V. Kalogeraki*, Privacy preservation for participatory sensing data, *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)* (Los Alamitos, CA, USA), 2013, pp. 103– 113.

[54] *Andrienko* et al., Visual Analytics for Understanding Spatial Situations from Episodic Movement Data.

[55] *C. Kopp, M. Mock & M. May*, Privacy-preserving distributed monitoring of visit quantities, [in:] Proceedings of the 20th International Conference on Advances in Geographic Information Systems, SIGSPATIAL '12 (New York, NY, USA: ACM, 2012), pp. 438–441, doi:10.1145/2424321. 2424384.

[56] *B. Hoh* et al., Enhancing Privacy and Accuracy in Probe Vehicle-Based Traffic Monitoring via Virtual Trip Lines, *IEEE Trans. Mob. Comput.* 11, no. 5 (2012), pp. 849–864.

[57] *S.-C. Florescu* et al., "Efficient Mobility Pattern Detection on Mobile Devices," in *Proceedings of the ECAI'12 Workshop on Ubiquitous Data Mining* (2012), pp. 23–27.

[58] *G. Andrienko* et al., "Report from Dagstuhl: the liberation of mobile location data and its implications for privacy research," *ACM SIGMOBILE Mobile Computing and Communications Review* 17, no. 2 (2013), pp. 7–18.

reveal workplace and working hours, the place where users spent the night and other locations indicating information on sensitive subjects as health, religion, political opinions, sexual orientation, etc. Thus, the transferred episodic movement data may even lead to re-identification. The protection of the individual histogram in such a data stream of locally aggregated mobility events is therefore an important task. The adversary model is a compromised server that utilizes the received individual histogram for inference of identities and other sensitive data. Existing methods either act at the network layer[59] or, inspired by the differential privacy paradigm, they add random noise[60]. The work in[61] denotes a protocol for secure aggregation among multiple parties, but their algorithm requires extensive communication between the parties and is infeasible in a single server scenario; also their encryption can be broken after several computation cycles. Recently, Liebig[62] proposed usage of homeomorphic encryption for secure aggregation of distributed mobility histograms.

9.4. Discussion and final remarks

In this work we provided an introduction to spatio-temporal data mining and highlighted popular analysis methods. Afterwards privacy threats of these analysis methods were discussed and examples were presented. With the advent of Big Data systems, we see a trend towards real-time and distributed data analysis. While these systems provide great utility e.g. for crisis response or intelligent traffic systems, the protection of vulnerable data is difficult in these scenarios. Established privacy measures, e.g. differential privacy, are not directly applicable to streaming data and novel privacy measures are required. Moreover legislation has to provide corridors for legal data handling that allow for innovative applications and protect individual data and identities. Future research will focus on privacy preserving analysis in these setting.

Acknowledgements

This research received funding from the European Union's Seventh Framework Programme under grant agreement number FP7-318225, INSIGHT and from the European Union's Horizon 2020 Programme under grant agreement number H2020-ICT-688380, VaVeL. Additionally, this work was supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876, project A1.

[59] *Kopp, Mock & May*, Privacy-preserving distributed monitoring of visit quantities.

[60] *A. Monreale* et al., Privacy-Preserving Distributed Movement Data Aggregation, [in:] *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography (Springer International Publishing, 2013), pp. 225–245, doi:10.1007/978-3-319-00615-4_13.

[61] *C. Clifton* et al., "Privacy-preserving data integration and sharing," in *DMKD* (2004), pp. 19–26.

[62] *T. Liebig*, "Privacy Preserving Centralized Counting of Moving Objects" in *AGILE 2015, F. Bacao, M. Y. Santos & M. Painho*, (eds.), Springer International Publishing, 2015, pp. 91-103.