Chapter 9

SOCIAL SENSING

Charu C. Aggarwal
IBM T. J. Watson Research Center
Yorktown Heights, NY
charu@us.ibm.com

Tarek Abdelzaher
University of Illinois at Urbana Champaign
Urbana, IL
zaher@cs.uiuc.edu

Abstract A number of sensor applications in recent years collect data which can be directly associated with human interactions. Some examples of such applications include GPS applications on mobile devices, accelerometers, or location sensors designed to track human and vehicular traffic. Such data lends itself to a variety of rich applications in which one can use the sensor data in order to model the underlying relationships and interactions. This requires the development of trajectory mining techniques, which can mine the GPS data for interesting social patterns. It also leads to a number of challenges, since such data may often be private, and it is important to be able to perform the mining process without violating the privacy of the users. Given the open nature of the information contributed by users in social sensing applications, this also leads to issues of trust in making inferences from the underlying data. In this chapter, we provide a broad survey of the work in this important and rapidly emerging field. We also discuss the key problems which arise in the context of this important field and the corresponding solutions.

Keywords: Sensor Networks, Social Sensors, Cyber-physical Networks

C.C. Aggarwal (ed.), *Managing and Mining Sensor Data*, DOI 10.1007/978-1-4614-6309-2_9,

1. Introduction

The proliferation of numerous online social networks such as *Facebook*, *LinkedIn* and *Google+* has lead to an increased awareness of the power of incorporating social elements into a variety of data-centric applications. Such networks are typically *data rich*, and contain heterogeneous data along with linkage stricture, which can be mined for a variety of purposes [39, 98, 108]. In particular, it has been observed that the use of a combination of social structure and different kinds of data can be a very powerful tool for mining purposes [136, 175, 182]. A natural way to enhance the power of such social applications is to embed sensors within such platforms in order to continuously collect large amounts of data for prediction and monitoring applications. This has lead to the creation of numerous social sensing systems such as *Biketastic* [142], *BikeNet* [55], *CarTel* [88] and *Pier* [148], which use social sensors for a variety of transportation and personal applications. The fusion of mobile, social, and sensor data is now increasingly being seen as a tool to fully enable context-aware computing [20].

A number of recent hardware platforms have extended the data-centric capabilities of social networks, by providing the ability to embed sensor data collection directly into the social network. Therefore, it is natural to explore whether sensor data processing can be tightly integrated with social network construction and analysis. For example, methods such a crowd-sourcing are a natural approach for improving the accuracy of many socially-aware search applications [168]. Some of the afore-mentioned data types on a conventional social network are static and change slowly over time. On the other hand, sensors collect vast amounts of data which need to be stored and processed in real time. There are a couple of important drivers for integrating sensor and social networks:

- One driver for integrating sensors and social networks is to allow the actors in the social network to both publish their data and subscribe to each other's data either directly, or indirectly after discovery of useful information from such data. The idea is that such collaborative sharing on a social network can increase real-time awareness of different users about each other, and provide unprecedented information and understanding about global behavior of different actors in the social network. The vision of integrating sensor processing with the real world was first proposed in [177].

- A second driver for integrating sensors and social networks is to provide a better understanding and measurement of the aggre-

gate behavior of self-selected communities or the external environment in which these communities function. Examples may include understanding traffic conditions in a city, understanding environmental pollution levels, or measuring obesity trends. Sensors in the possession of large numbers of individuals enable exploiting the crowd for massively distributed data collection and processing. Recent literature reports on several efforts that exploit individuals for data collection and processing purposes such as collection of vehicular GPS trajectories as a way for developing street maps [78], collectively locating items of interest using cell-phone reports, such as mapping speed traps using the Trapster application [190], use of massive human input to translate documents [145], and the development of protein folding games that use competition among players to implement the equivalent of global optimization algorithms [21].

The above trends are enabled by the emergence of large-scale data collection opportunities, brought about by the proliferation of sensing devices of every-day use such as cell-phones, piedometers, smart energy meters, fuel consumption sensors (standardized in modern vehicles), and GPS navigators. The proliferation of many sensors in the possession of the common individual creates an unprecedented potential for building services that leverage massive amounts data collected from willing participants, or involving such participants as elements of distributed computing applications. Social networks, in a sensor-rich world, have become inherently multi-modal data sources, because of the richness of the data collection process in the context of the network structure. In recent years, sensor data collection techniques and services have been integrated into many kinds of social networks. These services have caused a computational paradigm shift, known as *crowd-sourcing* [23, 47], referring to the involvement of the general population in data collection and processing. Crowd-sourcing, arguably pioneered by programs such as SETI, has become remarkably successful recently due to increased networking, mobile connectivity and geo-tagging [1]. We note that the phenomenon of crowd-sourcing is not exclusive to sensor data, but is also applied to other tagging and annotation processes, in which the knowledge is sourced from a social network of users. A classic example of a crowd-sourcing application is the *Amazon Mechanical Turk* [192], which allows users to submit data records for annotation at the payment of a fee for annotation purposes. Thus, the *Amazon Mechanical Turk* serves as an intermediary for crowd-sourcing of annotations for data records.

In the case of *social sensing* which is also often referred to as *people-centric sensing* [6, 26, 123] or *participatory sensing* [24], this crowd-

sourcing is generally achieved through sensors which are closely attached to humans, either in wearable form, or in their mobile phones. Some examples of integration of social and sensor networks are as follows:

- A variety of applications can be created to collect real time information from large groups of individuals in order to harness the *wisdom of crowds* in a variety of decision processes. For example, the *Google Latitude* application [184] collects mobile position data of uses, and uses this in order to detect the proximity of users with their friends. This can lead to significant events of interest. For example, proximity alerts may be triggered when two linked users are within geographical proximity of one another. This may itself trigger changes in the user-behavior patterns, and therefore the corresponding sensor values. This is generally true of many applications, the data on one sensor can influence data in the other sensors. Numerous other GPS-enabled applications such as *City sense*, *Macrosense*, and *Wikitude* [185, 195, 191] serve as gps-based social aggregators for making a variety of personalized recommendations. The approach has even been used for real-time grocery bargain hunting with the *LiveCompare* system [46].

- *Vehicle Tracking Applications:* A number of real-time automotive tracking applications determine the important points of congestion in the city by pooling GPS data from the vehicles in the city. This can be used by other drivers in order to avoid points of congestion in the city. In many applications, such objects may have implicit links among them. For example, in a military application, the different vehicles may have links depending upon their unit membership or other related data. Two classic examples of vehicular applications in the context of participatory sensing are the *CarTel* [88] and *GreenGPS* [64] systems.

- *Trajectory Tracking:* In its most general interpretation, an actor in a social network need not necessary be a person, but can be any living entity such as an animal. Recently, animal tracking data is collected with the use of radio-frequency identifiers. A number of social links may exist between the different animals such as group membership, or family membership. It is extremely useful to utilize the sensor information in order to predict linkage information and vice-versa. A recent project called *MoveBank* [186] has made tremendous advances in collecting such data sets. We note that a similar approach may be used for commercial product-tracking applications, though social networking applications are generally relevant to living entities, which are most typically people.

- *Applications to Healthcare:* In recent years, numerous medical sensor devices can be used in order to track the personal health of individuals, or make other predictions about their lifestyle [41, 65, 84, 119, 121, 122, 150]. This can be used for emergency response, long term predictions about diseases such as dementia, or other life style influence analysis of factors such as eating habits and obesity.

Social sensing applications provide numerous research challenges from the perspective of analysis. We list some of these challenges below:

- Since the collected data typically contains sensitive personal data (eg. location data), it is extremely important to use privacy-sensitive techniques [61, 133] in order to perform the analysis. A recent technique called *PoolView* [61] designs privacy-sensitive techniques for collecting and using mobile sensor data.

- Sensors, whether wearable or embedded in mobile devices, are typically operated with the use of batteries, which have limited battery life. Certain kinds of sensor data collection can drain the battery life more quickly than others (eg. GPS vs. cell tower/WiFi location tracking in a mobile phone). Therefore, it is critical to design the applications with a careful understanding of the underlying tradeoffs, so that the battery life is maximized without significantly compromising the goals of the application.

- The volume of data collected can be very large. For example, in a mobile application, one may track the location information of millions of users simultaneously. Therefore, it is useful to be able to design techniques which can compress and efficiently process the large amounts of collected data.

- Since the data are often collected through sensors which are error-prone, or may be input by individuals without any verification, this leads to numerous challenges about the *trustworthiness* of the data collected. Furthermore, the goals of privacy and trust tend to be at odds with one another, because most privacy-preservation schemes reduce the fidelity of the data, whereas trust is based on high fidelity of the data.

- Many of the applications require *dynamic* and *real time* responses. For example, applications which trigger alerts are typically time-sensitive and the responses may be real-time. The real-time aspects of such applications may create significant challenges, considering the large number of sensors which are tracked at a given time.

This chapter is organized as follows. Section 2 briefly discusses some key technological advances which have occurred in recent years, which have enabled the design of such dynamic and embedded applications. Section 3 discusses a broad overview of the key system design questions which arise in these different contexts. One of the important issues discussed in this section is privacy, which is discussed in even greater detail in a later section. Section 4 discusses some important privacy issues which arise in the context of social networks with embedded sensors. Section 5 discusses the trust-worthiness issues which arise in such crowd-sourcing systems. Section 6 introduces techniques for social network modeling from dynamic links which are naturally created by the sensor-based scenario. Since such dynamic modeling often requires trajectory mining techniques, we present methods for trajectory mining in section 7. Section 8 introduces some of the key applications associated with social sensing. Section 9 discusses the conclusions and research directions.

2. Technological Enablers of Social Sensing

A number of recent technological advances in hardware and software have enabled the integration of sensors and social networks. One such key technological advance is the development is small mobile sensors which can collect a variety of user-specific information such as audio or video. Many of the applications discussed are based on *user-location*. Such location can easily be computed with the use of mobile GPS-enabled devices. For example, most of the recent smart-phones typically have such GPS technology embedded inside them. Some examples of such mobile sensor devices may be found in [117, 100].

Sensors typically collect large amounts of data, which must be continuously stored and processed. Furthermore, since the number of users in a social network can be very large, this leads to natural scalability challenges for the storage and processing of the underlying streams. For example, many naive solutions such as the centralized storage and processing of the raw streams are not very practical, because of the large number of streams which are continuously received. In order to deal with this issue, a number of recent hardware and software advances have turned out to be very useful.

- *Development of Miniaturized Sensor Technology:* The development of miniaturized (wearable) sensors and batteries have allowed their use and deployment in a number of different social settings. For example, the development of miniaturized sensors, which can be embedded within individual attire can be helpful in a wide vari-

ety of scenarios [42, 100, 63, 33, 34]. A classic example is the spec mote, which is an extremely small sensor device that can be embedded in the clothing of a user, while remaining quite unobtrusive.

- *Advancement of smartphone technology:* In recent years, there has been considerable advancement in smartphone technology, which are now fairly sophisticated devices containing a wide array of sensors such as GPS, compass, accelerometers, bluetooth capabilities etc. In addition, these are *convergent devices*, with considerable computational capabilities, internet connectivity, and different modes of user interaction and content upload, such as social tweets, ability to record pictures and videos etc. All of these capabilities create a rich content-based and sensing environment for a wide variety of applications.

- *Increased Bandwidth:* Since sensor transmission typically requires large wireless bandwidth, especially when the data is in the form of audio or video streams, it is critical to be able to transmit large amounts of data in real time. The increases in available bandwidth in recent years, have made such real time applications a reality.

- *Increased Storage:* In spite of the recently designed techniques for compressing the data, the storage challenges for stream processing continue to be a challenge. Recent years have seen tremendous advances in hardware, which allow much greater storage, than was previously possible.

- *Development of Fast Stream Processing Platforms:* A number of fast stream processing platforms, such as the IBM System S platform [187] have been developed in recent years, which are capable of storing and processing large volumes of streams in real time. This is a very useful capability from the perspective of typical cyber-physical applications which need a high level of scalability for real-time processing.

- *Development of Stream Synopsis Algorithms and Software:* Since the volume of the data collected is very large, it often cannot be collected explicitly. This leads to the need for designing algorithms and methods for stream synopsis construction [7]. A detailed discussion of a variety of methods (such as sketches, wavelets and histograms) which are used for stream synopsis construction and analysis is provided in [7].

The sensing abilities of miniaturized devices and smartphones have also increased considerably in recent years. For example, the one of the earliest systems, which is referred to as a *sociometer* [33, 34], a small wearable device is constructed, which can detect people nearby, provide motion information and accelerometers, and also has microphones for detection of speech information. In addition, the device has the flexibility to allow for the addition of other kinds of sensors such as GPS sensors and light sensors. These sensors can be used in order to detect implicit links between people, and the corresponding community behavior. The aim of collecting a large number of such interactive behaviors is to be able to effectively model interactions, between different users, and then model the dynamics of the interaction with the use of the collected information.

Since the work in [33], much of these sensing capabilities are now available in commodity hardware such as mobile phones. For example, the *Virtual Compass* system [18] uses the sensors available in mobile phones in order to sense the interactions between different actors. Virtual Compass is a peer-based relative positioning system that uses multiple radios to detect nearby mobile devices and places them in a two-dimensional plane. It uses different kinds of scanning and out-of-band coordination to explore tradeoffs between energy consumption, and the latency in detecting movement. Methods are designed for using different kinds of sensor signals in *Virtual Compass* in order to reduce the energy footprint. More details may be found in [18].

3. Data Collection, Architectural and System Design Challenges

The aforementioned monitoring and social computing opportunities present a need for a new architecture that encourages data sharing and efficiently utilizes data contributed by users. The architecture should allow individuals, organizations, research institutions, and policy makers to deploy applications that monitor, investigate, or clarify aspects of *socio-physical* phenomena; processes that interact with the physical world, whose state depends on the behavior of humans in the loop.

An architecture for social data collection should facilitate distillation of concise actionable information from significant amounts of raw data contributed by a variety of sources, to inform high-level user decisions. Such an architecture would typically consist of components that support (i) privacy-preserving sensor data collection, (ii) data model construction, and (iii) real-time decision services. (iv) effective methods for recruitment, and (v) energy efficient design. For example, in an ap-

plication that helps drivers improve their vehicular fuel-efficiency, data collection might involve upload of fuel consumption data and context from the vehicle's on-board diagnostics (OBD-II) interface and related sensors; a model might relate the total fuel consumption for a vehicle on a road segment as a function of readily available parameters (such as average road speed, degree of congestion, incline, and vehicle weight); the decision support service might provide navigation assistance to find the most fuel-efficient route to a given destination (as opposed to a fastest or shortest route). Of course, none of these can be effectively implemented without energy-efficient data collection and participant recruitment. Below, we elaborate on the above functions.

3.1 Privacy-Preserving Data Collection

In a grassroots application that is not managed by a globally trusted authority, an interesting challenge becomes ensuring the privacy of data shared. Anonymity is not a sufficient solution because the data themselves (such as GPS traces) may reveal the identity of the owner even if shared anonymously. One interesting direction is to allow individuals to "lie" about their data in a way that protects their privacy, but without degrading application quality. For example, in a traffic speed monitoring application reconstruction of community statistics of interest (such as average traffic speed on different streets) should remain accurate, despite use of perturbed data ("lies" about actual speed of individual vehicles) as input to the reconstruction process. This is possible thanks to deconvolution techniques that recover the statistical distribution of the original signals, given the statistical distribution of perturbed data and the statistical distribution of noise. Solutions to this and related problems can be found in literature on privacy-preserving statistics [9]. Recently, special emphasis was given to perturbing time-series data [61], since sensor data typically comprise a correlated series of samples of some continuous phenomenon. Perturbing time-series data is challenging because correlations among nearby samples can be exploited to breach privacy. Recent results demonstrate that the frequency spectrum of the perturbation signal must substantially overlap with the frequency spectrum of the original data time-series for the latter to be effectively concealed [61]. Generalizations to perturbation of correlated multi-dimensional time-series data were proposed in [133]. The main challenge addressed in this work was to account for the fact that data shared by different sensors are usually not independent. For example, temperature and location data can be correlated, allowing an attacker to make inferences that breach privacy by exploiting cross-sensor correlations.

A related interesting problem is that of perturbation (i.e., noise) energy allocation. Given a perturbation signal of a particular energy budget (dictated perhaps by reconstruction accuracy requirements), how to allocate this energy budget across the frequency spectrum to optimally conceal an original data signal? A recent technique defines privacy as the amount of mutual information between the original and perturbed signals. Optimality is defined as perturbation that minimizes the upper bound on such (leaked) mutual information. The technique describes how optimal perturbation is computed, and demonstrates the fundamental trade-off between the bound on information leak (privacy) and the bound on reconstruction accuracy [132]. We note that the privacy protection issues for social sensing data arise both during trajectory data collection, and trajectory data management [38]. Since this section is focussed only on the data collection and system design issues, we will discuss this issue in a more holistic and algorithmic way in a later section of this chapter.

3.2 Generalized Model Construction

Many initial participatory sensing applications, such as those giving rise to the above privacy concerns, were concerned with computing community statistics out of individual private measurements. The approach inherently assumes richly-sampled, low-dimensional data, where many low-dimensional measurements (e.g., measurements of velocity) are redundantly obtained by individuals measuring the same variable (e.g., speed of traffic on the same street). Only then can good statistics be computed. Many systems, however, do not adhere to the above model. Instead, data are often high-dimensional, and hence sampling of the high-dimensional space is often sparse. The more interesting question becomes how to generalize from high-dimensional, sparsely-sampled data to cover the entire input data space? For instance, consider a fuel-efficient navigation example, where it is desired to compute the most fuel-efficient route between arbitrary source and destination points, for an arbitrary vehicle and driver. What are the most important generalizable predictors of fuel efficiency of current car models driven on modern streets? A large number of predictors may exist that pertain to parameters of the cars, the streets and the drivers. These inputs may be static (e.g., car weight and frontal area) or dynamic (e.g., traveled road speed and degree of congestion). In many cases, the space is only sparsely sampled, especially in conditions of sparse deployment of the participatory sensing service. It is very difficult to predict *a priori* which parameters will be more telling. More importantly, the key predictors

might differ depending on other parameters. For example, it could be that the key predictors of fuel efficiency for hybrid cars and gas-fueled cars are different. It is the responsibility of the model construction services to offer not only a general mechanism for applications to build good models quickly from the data collected, but also a mechanism for identifying the scope within which different predictors are dominant. A single "one-size-fits-all" prediction model, computed from all available data, is not going to be accurate. Similarly computing a model for each special case (e.g., a model for each type of car) is not going to be useful because, as stated above, the sampling is sparse. Hence, it is key to be able to generalize from experiences of some types of vehicles to predictions of others. Recent work combined data mining techniques based on regression cubes and sampling cubes to address the model generalization problem for sparse, high-dimensional data [64].

3.3 Real-time Decision Services

Ultimately, a generalized model, such as that described above, may be used as an input to an application-specific optimization algorithm that outputs some *decisions* for users in response to user queries. For example, estimates of fuel consumption on different roads on a map can be input to Dijkstra's algorithm to find the minimum fuel route between points specified by the user. This route constitutes a decision output. Hence, support for real-time stream processing and decision updates must be provided as part of the social sensing architecture.

A key property of real-time decision services is the involvement of humans in the loop. A significant challenge is therefore to design appropriate user interfaces. End-user devices will act as data custodians who collect, store, and share user data. The level at which these custodians interact with the user, as well as the nature of interactions, pose significant research problems with respect to minimizing inconvenience to the user while engaging the user appropriately. Context sensing, collaborative learning, persuasion, and modeling of socio-sensing systems (with humans in the loop) become important problems. Participation incentives, role assignment, and engagement of users in modeling and network learning become important application design criteria that motivate fundamental research on game theoretic, statistical, machine learning, and economic paradigms for application design.

3.4 Recruitment Issues

The quality of the social experience gained from a sensor-based framework is dependent on the ability to recruit high quality participants for

sensor collection and sharing. Many sensing systems such as those in geo-tagging applications user completely *open* frameworks in which any participant who wishes to contribute is allowed to join the sensing environment. This may of course result in numerous issues in terms of the quality of the final results:

- The participants who join may not be sufficiently trustworthy. This may impact the quality of the results. We will formally discuss the issue of trust in a later section.
- The act of participants *choosing to join* may bias the final variables which are being tracked by the application. For example, when the sensing is used in order to obtain feedback of a particular type, urban participants are more likely to join because of greater prevalence of smart phones. This kind of skew may also affect the quality of the final results.

We note that when the recruitment is performed at the initiative of the designer of the sensing system, a greater amount of control is achieved. This tends to reduce the self-selection bias, which is naturally inherent in a purely voluntary system. The work in [143] observes that the process of recruiting volunteers for participatory sensing campaigns is analogous to recruiting volunteers or employees in non-virtual environments. This similarity is used in order to create a three stage process for recruitment:

- **Qualifier:** This refers to the fact that the participants must meet minimum requirements such as availability and reputation. This ensures that high quality responses are received from the participants.
- **Assessment:** Once participants that meet minimum requirements are found, the recruitment system then determines which candidates are most appropriate based on both diversity and coverage. This ensures that bias is avoided during the recruitment process.
- **Progress Review:** Once the sensing process starts, the recruitment system must check participants' coverage and data collection reputation to determine if they are consistent with their base profile. This check can occur periodically, and if the similarity of profiles is below a threshold, this is used as a feedback to an additional recruitment process.

We note that the above process is quite similar to that of an employee hiring process in an organization, which is designed to maximize diversity, reduce bias, and maximize quality of the volunteers.

3.5 Energy Efficient Design

Since sensors often have limited battery power, it is critical to design the participatory sensing systems in order to maximize the life of the system. Many critical sensing components such as GPS tend to consume a lot of power, and can therefore result in a short life-cycle for the system. A number of tricks can be used in order to reduce the energy consumption, which may involve a reduction in the sampling rate at which the data is collected. This reduction in the sampling rate can often be achieved with the use of side-information which is collected through more efficient means. Some examples of common tricks which are used in order to improve the energy efficiency are as follows:

- In the *Sensloc* system proposed in [97], a GPS device, an accelerometer, and a WiFi scanner are simultaneously used in order to detect particular variables such as location with good accuracy. We note that the energy requirement of different kinds of sensors may vary, and the accuracy of the different sensors may also vary in a time dependent way. The typical tradeoff is that while GPS is extremely power hungry, it is also more accurate. Therefore, at a given time, the data is selectively sampled at varying rates from different sensors in order to fuse the measurements together, and provide an accurate estimation of the desired variable without too much energy consumption.

- A rate-adaptive positioning system known as *RAPS* is proposed in [129]. The system is based on the observation that GPS is generally less accurate in urban areas, and therefore it makes sense to turn it on only as often as necessary to achieve this accuracy. The *RAPS* system uses the location-time history of the user to estimate user velocity and adaptively turn on GPS only if the estimated uncertainty in position exceeds the accuracy threshold. It also estimates user movement using a duty-cycled accelerometer, and utilizes Bluetooth communication to reduce position uncertainty among neighboring devices. In addition, the system employs celltower-RSS blacklisting to detect GPS unavailability, in which case it is not turned on at all. Thus, the use of context-sensitive information in order to adaptively turn on GPS results in a considerable about of power savings [129].

- Often, a large amount of rich context sensitive information is available, which can be used to improve the accuracy of the sensing measurements without spending additional energy. For example, if the last known GPS location is overlaid on a map, then the fu-

ture location within a specific time period may be limited both by the time elapsed and the layout of that location. This can be used to estimate the location more accurately with the use of a smaller number of samples [40].

- One major issue, which has been observed in [183] is that the requests for GPS may come from Location Based Applications (LBA), and therefore, it is critical to design the energy saving strategies in the context of the LBA requests, which may not necessarily be synchronized with one another. The framework uses four design principles corresponding to *substitution*. Substitution makes use of alternative location-sensing mechanisms, such as network-based location sensing, which consume less power than GPS. Suppression uses less power-intensive sensors such as an accelerometer to suppress unnecessary GPS sensing for a statitionary user. Piggybacking synchronizes the location sensing requests from multiple running LBAs. Adaptation aggressively adjusts system-wide sensing parameters such as time and distance, when battery level is low.

- In addition to software solutions, it is also possible to implement hardware solutions. For example, simple operations can be directly performed in main memory with dedicated hardware, without actually using the (more energy-intensive) main processor [135].

In addition to the power efficiency of the *sensing* process, issues also often arise about the power-efficiency of the *data transmission* process. Typically, data transmission is significantly more expensive, as compared to the sensing process itself. For example, many applications are enabled by the ability to capture videos on a smartphone and to have these videos uploaded to an internet connected server. This capability requires the transfer of large volumes of data from the phone to the infrastructure. Typically smartphones have multiple transfer interfaces such as 3G, Edge, and Wifi, all of which vary considerably in terms of availability, data transfer rates and power consumption. In many cases, the underlying applications are naturally delay-tolerant, so that it is possible to delay data transfers until a lower-energy WiFi connection becomes available. This tradeoff is explored in some detail in the *SALSA* system proposed in [137]. An online algorithm is proposed, which can automatically adapt to channel conditions and it requires only local information to decide whether and when to defer a transmission. Such an approach has been shown to result in considerable power savings without significantly affecting the operation of the underlying system.

3.6 Other Architectural Challenges

Proper design of the above system components gives rise to other important challenges that must be solved in order to enable development and deployment of successful mobile sensing applications that adequately meet user needs. The following relates challenges described in a recent NSF-sponsored report[1] on social sensing.

From the application perspective, mobile sensing applications depend significantly on social factors (user adoption, peer pressure, social norms, social networks, etc) as well as the nature of physical phenomena being monitored or controlled. Exciting interdisciplinary research challenges exist in describing the properties of distributed socio-physical applications. For example, what are the dynamics of information propagation in such systems? What are closed-loop properties of interaction involving social and physical phenomena? What are some fundamental bounds on capacity, delivery speed, and evolution of socio-sensing systems? Answering such questions is fundamental to informed design and performance analysis of sensing applications involving crowd-sourcing.

From the underlying physical network perspective, mobile sensing applications herald an era where many network clients are embedded devices. This motivates the investigation of a network architecture, where the main goal from networking shifts from offering a mere communication medium to offering *information distillation services*. These services bridge the gap between myriads of heterogeneous data feeds and the high-level human decision needs. In a network posed as an information service (as opposed to a communication medium), challenges include division of responsibilities between the end-device (e.g., phone) and network; paradigms for data collection on mobile devices, architectural support for data management, search, and mining; scalability to large-scale real-time information input and retrieval; improved context-awareness; support for predictability; and investigation of network and end-system support for reduction of cognitive overload of the information consumer. Other challenges in the design of network protocols for mobile sensing include energy management, integration of network storage, personalized search and retrieval, support for collaborative sensing, and exploitation of a rich realm of options in information transfer modalities and timing, including deferred information sharing and delay-tolerant communication.

[1] National Science Foundation Workshop Report on Future Directions in Networked Sensing Systems: Fundamentals and Applications, The Westin Arlington Gateway, Arlington, VA, November 12-13, 2009.

While several social sensing applications are already deployed, exciting research opportunities remain in order to help understand their emergent behavior, optimize their performance, redesign the networks on which they run, and provide guarantees to the user, such as those on bounding unwanted information leakage.

4. Privacy Issues in Social Sensing

Social sensing offers interesting new challenges pertaining to privacy assurances on data. General research on privacy typically focuses on electronic communication as opposed to ramifications of increasing sensory instrumentation in a socio-physical world. In contrast, traditional embedded systems research typically considers computing systems that interact with physical and engineering artifacts and belong to the same trust domain. A need arises to bridge the gap in privacy research by formulating and solving privacy-motivated research challenges in the emerging social sensing systems, where users interact in the context of social networks with embedded sensing devices in the physical world.

Sharing sensor data creates new opportunities for loss of privacy (and new privacy attacks) that exploit physical-side channels or a priori known information about the physical environment. Research is needed on both privacy *specification* and *enforcement* to put such specification and enforcement on solid analytic foundations, much like specification and enforcement of safety requirements of high-confidence software.

Specification calls for new physical privacy specification interfaces that are easy to understand and use for the non-expert. *Enforcement* calls for two complementary types of privacy mechanisms; (i) *protection mechanisms from involuntary physical exposure*, and (ii) *control of voluntary information sharing*. The former enforce *physical privacy*. They are needed to prevent "side-channel" attacks that exploit physical and spatio-temporal properties, characteristic of embedded sensing systems, to make inferences regarding private information. Control of voluntary information sharing must facilitate privacy-preserving exchange of time-series data. A predominant use of data in social sensing applications is for aggregation purposes such as computing statistical information from many sources. Mathematically-based data perturbation and anonymization schemes are needed to hide user data but allow fusion operations on perturbed or partial data to return correct results to a high degree of approximation.

While privacy-preserving statistics and privacy-preserving data mining are mature fields with a significant amount of prior research, sharing of sensor data offers the additional challenge of dealing with *cor-*

related multi-dimensional time-series data represented by sensory data streams. Correlations within and across sensor data streams and the spatio-temporal context of data offer new opportunities for privacy attacks. The challenge is to perturb a user's sequence of data values such that (i) the individual data items and their trend (i.e., their changes with time) cannot be estimated without large error, whereas (ii) the distribution of the data aggregation results at any point in time is estimated with high accuracy. For instance, in a health-and-fitness social sensing application, it may be desired to find the average weight loss trend of those on a particular diet or exercise routine as well as the distribution of weight loss as a function of time on the diet. This is to be accomplished without being able to reconstruct any individual's weight and weight trend without significant error.

Examples of data perturbation techniques can be found in [14, 13, 59]. The general idea is to add random noise with a known distribution to the user's data, after which a reconstruction algorithm is used to estimate the distribution of the original data. Early approaches relied on adding independent random noise. These approaches were shown to be inadequate. For example, a special technique based on random matrix theory has been proposed in [95] to recover the user data with high accuracy. Later approaches considered hiding individual data values collected from different private parties, taking into account that data from different individuals may be correlated [86]. However, they do not make assumptions on the model describing the *evolution* of data values from a given party over time, which can be used to jeopardize privacy of data streams. Perturbation techniques must specifically consider the data evolution model to prevent attacks that extract regularities in correlated data such as spectral filtering [95] and Principal Component Analysis (PCA) [86]. In addition to data perturbation, numerous *group-based anonymization* methods have been proposed such as k-anonymity and ℓ-diversity [9]. In k-anonymity methods, the data features are perturbed, so that adversarial attacks always retain an ambiguity level over k-different participants. In ℓ-diversity, criteria are imposed over a group to ensure that the values of the sensitive attributes are sufficiently diverse within a group. This is motivated by the observation that k-anonymity may sometimes not preserve the truth about individual sensitive values, when all sensitive values within an anonymized group are the same.

In work discussed earlier in this chapter [61], it was shown that privacy of time-series data can be preserved if the noise used to perturb the data is itself generated from a process that approximately models the measured phenomenon. For instance, in the weight watchers example, we may have an intuitive feel for the time scales and ranges of weight

evolution when humans gain or lose weight. Hence, a noise model can be constructed that exports realistic-looking parameters for both the direction and time-constant of weight changes. The resulting perturbed stream can be aggregated with that of others in the community. Since the distributions of noise model parameters are statistically known, it is possible to estimate the sum, average and distribution of added noise (of the entire community) as a function of time. Subtracting that known average noise time series from the sum of perturbed community curves will thus yield the true community trend. The distribution of community data at a given time can similarly be estimated (using de-convolution methods) since the distribution of noise (i.e., data from virtual users) is known. The estimate improves with community size.

The approach preserves individual user privacy while allowing accurate reconstruction of community statistics. Several research questions arise that require additional work. For example, what is a good upper bound on the reconstruction error of the data aggregation result as a function of the noise statistics introduced to perturb the individual inputs? What are noise generation techniques that minimize the former error (to achieve accurate aggregation results) while maximizing the noise (for privacy)? How to ensure that data of individual data streams cannot be inferred from the perturbed signal? What are some bounds on minimum error in reconstruction of individual data streams? What noise generation techniques maximize such error for privacy? Privacy challenges further include the investigation of attack models involving corrupt noise models (e.g., ones that attempt to deceive non-expert users into using perturbation techniques that do not achieve adequate privacy protection), malicious clients (e.g., ones that do not follow the correct perturbation schemes or send bogus data), and repeated server queries (e.g., to infer additional information about evolution of client data from incremental differences in query responses). For example, given that it is fundamentally impossible to tell if a user is sharing a properly perturbed version of their real weight or just some random value, what fractions of malicious users can be accommodated without significantly affecting reconstruction accuracy of community statistics? Can damage imposed by a single user be bounded using outlier detection techniques that exclude obviously malicious users? How does the accuracy of outlier detection depend on the scale of allowable perturbation? In general, how to quantify the tradeoff between privacy and robustness to malicious user data? How tolerant is the perturbation scheme to collusion among users that aims to bias community statistics? Importantly, how does the *time-series* nature of data affect answers to the above questions com-

pared to previous solutions to similar problems in other contexts (e.g., in relational databases)?

Furthermore, how can the above perturbation techniques, defense solutions, and bounds be extended to the sharing of multiple correlated data streams, or data streams with related context? For example, consider a social sensing application where users share vehicular GPS data to compute traffic speed statistics in a city. In this case, in order to compute the statistics correctly as a function of time and location, each vehicle's speed must be shared together with its current GPS location and time of day. Perturbing the speed alone does not help privacy if the correct location of the user must be revealed at all times. What is needed is a perturbation and reconstruction technique that allows a user to "lie" about their speed, location, and time of day, altogether, in a manner that makes it impossible to reconstruct their true values, yet allow an aggregation service to average out the added multi-dimensional noise and accurately map the true aggregate traffic speed as a function of actual time and space. This problem is related to the more general concern of privacy-preserving classification [158, 176], except that it is applied to the challenging case of aggregates of time-series data. Other methods for centralized and distributed privacy preservation in time series include the methods discussed in [130, 141], though these methods are generally *offline*, and cannot easily perform the privacy preservation in real time, as would be needed for a typical social sensing application.

In many participatory sensing applications, users may upload different kinds of data such as images, text, or other feeds to the system. Such data are often tagged with location (WiFi or GPS) and the time-stamp, which can have serious consequences in terms of location privacy. Alternatively, the users may have to continuously provide their location to an untrusted service provider, or provide responses to queries which may compromise their privacy. Some of the earliest work on location privacy [152] focusses only on user identity suppression, while preserving the full fidelity of the location data. This approach of course suffers from the well known problem of adversarial attacks with background information about approximate location. The work in [66, 75, 94, 131] avoids this pitfall by using a k-anonymity approach for the spatio-temporal scenario. The work in [94] proposed a technique called *tessellation*, in which a point location is enlarged to a tile which contains at least k users. This is essentially a spatio-temporal version of the *generalization* technique which is often used in k-anonymity applications. It was observed in [87], that tessellation is not useful in applications where the large tiles do not provide the fine grained information about the location for a particular user (such as the road information). Therefore, the work in [87]

uses a clustering (micro-aggregation) approach, which is able to preserve more fine grained information about the location. In this context, the method of [2] also treats the trajectory of an object as a cylinder in 3-dimensional space, where the radius of the cylinder is non-zero because of the uncertainty in the GPS position of the object. The key here is to understand is that the uncertainty is inherent to the method of collecting the data, since all GPS collection methods have a certain level of error associated with them. In this context, the work in [2] defines the concept of (k, δ)-anonymity, which is a set S of at least k trajectories, such that all of these trajectories lie within a distance of at most $\delta/2$ of the average position of these different trajectories. We note that it may not be possible to create $(k.\delta)$-anonymized groups from the original data set, if some of the trajectories are somewhat isolated. Therefore, the work in [2] proposes the *Never Walk Alone (NWA)* algorithm, in which the positions of some of the objects is distorted with *space translation*, so that it is possible to construct such (k, δ)-anonymized groups from the data. The approach constructs these anonymized groups while minimizing the total distortion in the data.

Many mobile applications can infer the *context* of a user from GPS (e.g. whether a user is at home or work). It has become increasingly common for many mobile applications to aggressively collect such context data [56] for a variety of applications. Such context can sometimes be very sensitive from a release perspective. For example, a user may not wish anyone to know whether they are currently in a hospital. The afore-mentioned k-anonymization does not necessarily help protect the sensitivity of context, if all of the k users within a group are at the same sensitive location. A number of methods use full suppression techniques [83, 157] in which the location or context of the user is suppressed when they are at a sensitive location. However, it has been observed in [74] that the *fact of the suppression itself* can be sensitive information, in the presence of a powerful adversary with greater background knowledge.

Another issue with mobile sensing applications is that considerable temporal correlations exist between the different locations of a single or multiple users. Such correlations can be used in order to perform privacy attacks which can infer the sensitive locations of different users. In this context, a number of methods [30, 68, 69, 76, 126] have been designed which utilize the temporal correlations in the privacy preservation process. The work in [76] observes that one can use linear interpolation to infer suppressed locations. Therefore, the work in [76] works by constructing zones which contain multiple sensitive locations, and the anonymization process introduces a sufficient amount of uncertainty in each zone. It has been observed in [30] that information about the veloc-

ity of a user can be used in order to infer their location during successive time instants. For example, for two successive zones containing a user, the velocity of the user provides implicit limits on where they may or may not be found at any given time. The work in [30] protects against such kinds of privacy attacks. The work in [68] improves these methods by introducing temporal delays. However, none of these methods can provably protect privacy, when an adversary knows the system that is used for anonymization. The work in [74] designs a scheme which can preserve the privacy of sensitive user locations in the presence of such powerful background knowledge.

Location privacy systems can also be understood in terms of *Quality of Service (QoS)* models in response to user location queries. Such models consider the fact that the use of generalization (eg. spatial and temporal cloaking) and suppression (eg. dropping a trajectory from query output) for privacy preservation reduces the accuracy of responses to user-queries. Therefore, a significant amount of research has also been focussed on performing the privacy-preservation with a focus on maintaining certain levels of QoS for privacy preservation [17, 67, 125, 149]. These methods generally work with optimizing common models for k-anonymity and ℓ-diversity, with a specific focus on improving the QoS for user queries.

Finally, it has been recognized, that in many mobile sensing applications, it is not required to collect the individual sensor streams, but one may only desire to compute the aggregate statistics from these sensors. For example, many location-based vehicular services are designed into the national transportation infrastructure in many countries. These include usage- or congestion-based road pricing, traffic law enforcement, traffic monitoring, and vehicle safety systems. Such applications often require the computation of aggregate statistics, but poorly chosen implementations can result in violations of privacy. For example, the GPS monitoring of cars as they arrive, or the use of surveillance cameras and toll transponders can result in privacy violations.

In the context of such applications, the following functionalities need to be provided:

- In many applications, some centralized server needs to compute a function of a car's *path*, which is essentially a list of time-position tuples. A system called *VPriv* [134] provides a protocol to compute path functions in a way, such that it does not reveal anything more than the result of the function to the server. In addition, an enforcement mechanism is provided (using random spot checks) that allows the server and application to handle misbehaving cars.

- The *PrivStats* system computes aggregate statistics on user location information and guarantees location privacy even in the face of side information about user location and movement patterns. It is also resistant to large amounts of spurious data upload by users.

Many applications require the computation of a specific function on the data, and therefore, it is critical to design methods for computing the function accurately on the perturbed data. For example, the problem of privacy-preserving regression modeling of sensor data has been discussed in [3].

5. Trust in Social Sensing

At the broadest level, social sensing systems can be considered multi-agent systems, that interact with one another and provide a variety of data-centric services to one another. Therefore, a number of issues of trust arise in the context of such large-scale social-centric applications, which are common to many traditional peer-to-peer applications [138]. Such issues typically deal with the the aspect of designing trustworthy protocols for interactions between different agents, both in terms of the choice of interactions, and the time of these interactions. A detailed survey of the (more traditional) literature along this direction may be found in [138]. The more recent social sensing work has focussed on the *data-centric* aspects of trust, rather than the *interaction-centric* aspects.

The openness of participatory sensing systems provides them with a tremendous amount of power in collecting information from a wide variety of sources, and distilling this information for data mining purposes. However, it is this very openness in data collection, which also leads to numerous questions about the quality, credibility, integrity, and trustworthiness of the collected information [45, 51, 71, 72]. Furthermore, the goals of privacy and trust would seem to be at odds with one another, because all privacy-preservation mechanisms *reduce* the fidelity of the data for the end-user, whereas the end-user trust is dependent on *high* fidelity of the data. Numerous questions may arise in this respect:

- How do we know that the information available to the end user is correct, truthful and trustworthy?
- When multiple sources provide conflicting information, how do we know who to believe?
- Have errors been generated in the process of data collection, because of inaccuracy or hardware errors?

The errors which arise during hardware collection are inherent to the device used, and their effect can be ameliorated to some extent by care-

ful design of the underlying application. For example, the *LiveCompare* [46] application (described in detail in the application section), which is used for comparison shopping of grocery products, works by allowing individuals to transmit photographs taken in stores of grocery products, and then presents similar pictures of products taken in nearby stores. The approach allows the transmitting of product photos taken by individual users of competing products, but does not automatically try to extract the pricing information from the price tags in the photograph. This is because the extraction process is known to be error-prone, and this design helps avoid the inaccuracy of reporting the pricing of competing products. It also avoids manual user input about the product which reduces error and maximizes trustworthiness.

For the case of specific kinds of data such as *location data*, a variety of methods can be used in order to verify the truthfulness of the location of a mobile device [107]. The key idea is that time-stamped location certificates signed by wireless infrastructure are issued to co-located mobile devices. A user can collect certificates and later provide them to a remote party as verifiable proof of his or her location at a specific time. The major drawback of this approach is that the applicability of these infrastructure based approaches for mobile sensing is limited as cooperating infrastructure may not be present in remote or hostile environments of particular interest to some applications. Furthermore, such an approach can be used only for particular kinds of data such as location data.

In the context of participatory sensing, where raw sensor data is collected and transmitted, a basic approach for ensuring the integrity of the content has been proposed in [51], which guards whether the data produced by a sensor has been maliciously altered by the users. Thus, this approach relies on the approach of *platform attestation* which vouches that the software running on the peripheral has not been modified in an unintended manner. This kind of approach is more useful for sensors in which the end data is produced by the device itself, and an automated software can be used for detection of malicious modification. In essence, the approach allows the trusted sensing peripherals to sign their raw readings, which allows the remote entity to verify that the data was indeed produced by the device itself and not modified by the user.

An additional challenge which naturally arises in the context of data trustworthiness is that the goals of data integrity and authenticity run contrary to the goals of user privacy. Almost all privacy-preserving data mining algorithms reduce the data fidelity in some way in order to reduce the ability to identify sensitive information about the user. Clearly,

such an approach will not work in the context of systems such as those proposed in [51].

Trusted Platform Module (TPM) hardware [71], commonly provided in commodity PCS, can be leveraged to help provide this assurance. To address the problem of protecting the privacy of data contributors, techniques such as requiring explicit authorization for applications to access local resources and formulating and enforcing access control policies can be used. A TPM is a relatively inexpensive hardware component used to facilitate building trusted software systems. It is possible to leverage the TPM functionality of attesting to the integrity of software running on a device to a remote verifier. The TPM can attest to the software platform running on the machine by providing a signed quote of its PCR(s) in response to a challenge from a remote verifier.

In many cases, user actions may change the data (such as the cropping of an image), but this may not actually affect the trust of the underlying data. The work in [72] proposes *YouProve*, which is a partnership between a mobile device's trusted hardware and software that allows un-trusted client applications to directly control the fidelity of data they upload and services to verify that the meaning of source data is preserved. The approach relies on trusted analysis of derived data, which generates statements comparing the content of a derived data item to its source. For example, the work in [72] tests the effectiveness of the method on a variety of modifications on audio and photo data, and shows that it is possible to verify which modifications may change the meaning of the underlying content.

A more critical question about trustworthiness arises when the data is collected through the actions of end users. In such cases, the user responses may have an inherent level of errors which may need to be evaluated for their trustworthiness. The issue of truthfulness and trust arises more generally in any kind of application, *where the ability to contribute information is open.* Such openness is a double-edged sword, in that it greatly increases information availability at the expense of trust. Aside from social and participatory sensing platforms, any *web-enabled platforms* which allow the free contribution of information may face such issues.

In this context, the problem of trustworthiness has been studied for resolving multiple, conflicting information provision on the web. The earliest work in this regard was proposed in [170], where the problem of studying conflicting information from different providers was studied [170]. Subsequently, the problem of studying trustworthiness in more general dynamic contexts was studied in [48, 49].

A number of recent methods [103, 159–162] address this issue, in which a consistency model is constructed in order to measure the trust in user responses in a participatory sensing environment. The key idea is that untrustworthy responses from users are more likely to be different from one another, whereas truthful methods are more likely to be consistent with one another. This broad principle is used in order to model the likelihood of participant reliability in social sensing with the use of a Bayesian approach [159, 160, 162]. A system called *Apollo* [103] has been proposed in this context in order to distill the likely truth from noisy social streams.

Such social streams are also often used in the context of applications, where *alarms* are raised in response to specific events. The nature of the alarm may vary with the application scenario. For example, in a military network, the alarm may be raised because of enemy threats, whereas in a patient monitoring application, the alarm may be raised because of a medical emergency. Such applications are inherently error prone and raise many false alarms because of technology limitations. For example, errors in the collection of the sensor readings, or an innocuous activity may trigger a false alarm. In [153], the problem of trustworthiness of such alarms has been studied, and a number of methods have been proposed in order to provide more accurate and trustworthy alarms.

6. Implied Social Networks: Inference and Dynamic Modeling

In the case of an explicitly linked social network, the relationships between different entities are quite clear, and therefore the dynamics of the interaction can be modeled relatively easily. However, in the case of a participatory sensing environment, the links between different entities may change *rapidly and dynamically.* Furthermore, such links may either be *explicit* or *implicitly derived* based on the dynamic interactions between participants. For example, the *Google Latitude* application allows for explicit links between different agents. On the other hand, in many social applications [52, 34], the links and communities between different agents may need to be derived based on their location and behavior. In such cases, the structure of the social network itself and the underlying communities [53, 35, 36, 163] can be derived directly from the details of the underlying interaction. This is a challenging problem, especially when the number of agents are large, and the number of interactions between them is even larger and dynamically evolving. Furthermore, a variety of context-specific information such as organizational rhythms, socially significant location and daily activity patterns may need to be

simultaneously derived and used [52] for inferring the significant links. The work in [44] derives the links between users based on their mobility patterns from GPS trajectories. In order to achieve this goal, the work in [44] divides the spatial regions into a grid, and constructs nodes for each cell. An edge exists between a pair of nodes, if a trajectory exists which starts at one cell and ends at another. By performing the discretization at varying levels of granularity, it is possible to analyze different characteristics of the underlying users. The work in [44] specifically shows how the approach can be used for effective community detection.

An interesting work in [173] examines the common patterns in the activities of different geo-tracked users, and makes friendship or linkage recommendations on the basis of significant overlaps in activity patterns. It has also been observed in [115] that different kinds of sharing in activity patterns may have different significance for different users. For example, it is possible that two individuals that are friends may not spend a lot of time together, but only a couple of hours on a Saturday night. On the other hand, a pair of co-workers who are not friends may share a lot of time together. Thus, it is critical to be able to learn the importance of different kinds of commonality in patterns in the prediction process [115]. Such trajectory analysis is useful not just for determining useful relationships, but also interesting places, travel sequences or activities which are relevant to such relationships [27, 181]. In particular, an interesting *authority-based* model for relating social behavior and location behavior has been proposed in [27]. The essential idea is to construct a graph which models relationships of the trajectories of the different users to the different locations. The idea is that authoritative users are also likely to visit authoritative places and vice-versa. This is used in order to construct a page-rank like model in order to determine both the authoritative users and authoritative locations simultaneously.

Many sensing platforms such as those discussed in [33], yield sensor data which is varied, and is of a multi-modal nature. For example, the data could contain information about interactions, speech or location. It is useful to be able analyze such data in order to summarize the interactions and make inferences about the underlying interactions. Such multi-modal data can also be leveraged in order to make predictions about the nature of the underlying activities and the corresponding social interactions. This also provides a virtual technique to perform link inferences in the underlying network.

The collection of activity sensing data is not very useful, unless it can be leveraged to summarize the nature of the activities among the different participants. For example,in the case of the techniques discussed in [34], the IR transceiver is used to determine which people are in prox-

imity of one another. However, this cannot necessarily be used in order to determine whether the corresponding people are interacting with another. A knowledge of such interactions can be determined with the use of *speech segmentation* techniques in which it is determined which participants are interacting with one another. The speech portions are segmented out of the ambient noise, and then segmented into conversations. The knowledge of such face-to-face interactions can be used to build dynamic and virtual links among the different participants.

We note that a dynamically linked social network can be modeled in two different ways:

- The network can be modeled as a group of dynamic interacting agents. The stochastic properties of these agents can be captured with the use of hidden markov models in order to characterize various kinds of behaviors. This is the approach used for community modeling as discussed in [15, 36].
- The interactions of the participants can be modeled as links which are continuously created or destroyed depending upon the nature of the underlying interactions. as a graph stream, in which the nodes represent the participants, and the edges represent the interactions among these different participants. Recently, a number of analytical techniques have been designed in order to determine useful knowledge-based patterns in graph streams [8]. These include methods for dynamically determining shortest-paths, connectivity, communities or other topological characteristics of the underlying network.

The inherently dynamic nature of such interactions in an evolving and dynamic social network leads to a number of interesting challenges from the perspective of social network analysis. Some examples of such challenges are discussed below.

(1) Determination of dynamic communities in graph streams: Communities are defined as dense regions of the social network in which the participants frequently interact with one another over time. Such communities in a dynamically evolving social network can be determined by using agent-based stochastic analysis or link-based graph stream analysis. Methods for modeling such a social network as a group of dynamically evolving agents are discussed in [15, 36]. In these techniques, a hidden markov model is used in conjunction with an influence matrix in order to model the evolving social network.

A second approach is to model the underlying face-to-face interactions as dynamic links. This creates an inherently dynamic network scenario in which the structure of the communities may continuously evolve over

time. Therefore, a key challenge is to determine such communities in dynamic networks, when the clustering patterns may change significantly over time. Methods for determining evolving clusters and communities in networks have been discussed in [10, 12, 31, 28, 79, 151]. Many of these methods determine communities in the underlying data by incorporating concepts of *temporal smoothness*, wherein the structure of the communities is allowed to evolve only in a smooth way over time. On the other hand, when the data is of very high volume (such as a graph stream), it is also critical to design very efficient methods for community maintenance. Graph streams pose a special challenge because of the rapid nature of the incoming edges, and their use for determination of evolving communities.

(2) Mining Structural Patterns in Time-Evolving Social Networks: Aside from the common problem of community detection, another interesting problem is that of mining structural patterns of different kinds in time evolving graphs. Some common methods for finding such patterns typically use matrix and tensor-based tools, which are comprehensively described in a tutorial in [60]. Common problems in time-evolving graphs include those of frequent pattern determination, outlier detection, proximity tracking [156], and subgraph change detection [118].

(3) Modeling spatio-temporal dynamics: Many of the approaches discussed above model the dynamics of the interactions as dynamic links. While this provides greater generality, it does not capture the spatio-temporal nature of the underlying agents. For example, the data received in a GPS application often contains spatio-temporal information such as the positions of different agents, and their underlying interactions. Therefore, an interesting and important challenge is to model the aggregate spatio-temporal dynamics in order to determine the underlying patterns and clusters. Such spatio-temporal dynamics can be used in order to make interesting *spatial predictions* such as future regions of activity or congestions. Many methods for clustering, community detection, classification, and outlier detection from such data have been proposed in [104, 105, 112–115, 109, 110] and are discussed in some detail in the application section of this chapter. In many cases, such data may even be combined with other content-based data such as GPS-tagged images and documents in order to further improve the quality of the underlying inference [172].

(4) Modeling Influential Community Members: This problem is essentially that of determining the members of the participatory sensor network, who have the greatest influence on their peers in the community. Alternatively, it may also be interesting to trace back the spread of

rumors or other information in the community. In a static network such as *Facebook*, the problem of influence analysis is much more straightforward, because it depends upon the static connections between the different communities [96]. In a dynamic network, the underlying network structure may change rapidly over time, depending upon the interactions between the underlying entities. Some recent work on dynamic influence analysis addresses this scenario of interactions between dynamic and evolving entities [11]. This method can determine either influential nodes or determine the most likely points of release, based on a given influence pattern and also a given pattern of interactions. A classic example of a dynamic network in the context of social sensing is the *face-to-face interaction network*, in which it may be desirable to determine the influence of such interactions on specific behaviors. For example, the work in [122] used a mobile phone-based sensing platform to examine the influence of face-to-face interactions in the life-style choices of participants such as obesity, eating and exercise habits. It was shown that the use of sensing platforms can be very effective at modeling the influence effects of such interactions (which turned out to be significant for this scenario).

As discussed earlier, the determination of dynamic interactions can sometimes require the real-time modeling of *implied interactions* (such as face-to-face interactions), which are hard to infer from sensor data can also sometimes be sensitive information. This also leads to numerous privacy challenges, especially since the interactions between the participants may be considered personal information. As mentioned earlier, privacy continues to be an important issue for such social sensing applications. A number of privacy-sensitive approaches for face-to-face activity modeling and conversation segmentation have been discussed in [164–167].

The dynamic modeling of social sensing applications, naturally lead to a lot of trajectory data in real applications. Therefore, significant amount of research has been devoted to determining spatio-temporal patterns from such trajectories. Such patterns may be derived with or without additional content information. A number of these methods will be discussed in the next section.

7. Trajectory Mining for Social Sensing

Social sensing applications have naturally lead to the collection of trajectory database from the rich GPS data, which is collected in a wide variety of applications. The increasing popularity and availability of mobile phones also enables the collection of trajectory data from willing

participants with the use of widely downloadable mobile applications, as long as appropriate privacy-protection mechanisms are in place. A classic example of such a data set is the well known *GeoLife* data set [194]. Such data sets are not just collected for humans, but even from animals for tracking purposes. An example of such an animal tracking database is the *Movebank* database [186], which contains detailed data about animal trajectories in the data. Finally, many recent document and image creation hardware such as GPS-enabled cameras and cellphones automatically stamp the content with GPS locations. This creates a very rich data set containing *both* content and (implicit) trajectories. The availability of such data makes it important to design more effective and efficient methods for trajectory mining.

Trajectory data is particularly useful from the perspective of mining aggregate community movement patterns. A variety of interesting patterns can be mined in such trajectory data sets, which provide insights into the aggregate movements. The aggregate movements are best represented by clusters, which are variously referred to as *flocks*, *convoys*, or *swarms* [22, 77, 91, 102, 104, 113], depending upon the model which is used to characterize these clusters. Typically, the goal is to either determine objects with trajectories of similar shape, or objects which move together in clusters. The major difference between these different kinds of moving clusters are as follows:

- *Flocks:* These correspond to groups of objects which move within a fixed disc of a particular size over consecutive time-stamps [77, 102]. As a result, the underlying trajectories will often have a similar geometric shape.

- *Moving Cluster:* This refers to a group of objects which have considerable overlap between successive time-stamps [93]. As in the previous cases, the constraint on the objects moving together in successive time stamps leads to trajectories of similar shape.

- *Convoys:* In this case, we again find groups of objects which move together, except that the concept of density is used in order to define the objects that move together. As before, the objects need to move together over consecutive time stamps [90, 91]. In many scenarios, the use of density provides a flexible way of modeling the movement of significant masses of objects together.

- *Swarms:* In the case of swarms, the objects are required to move together as before, except that we do not impose the requirement that the objects should be together over consecutive time stamps

[113]. In such a case, the shapes of the trajectories of the different objects may sometimes be quite different. The approach discussed in [113] first uses an off-the-shelf spatial clustering algorithm to partition the objects into clusters at each time-stamp. This transforms the spatial trajectories into data which represents membership of objects in clusters. Subsequently, a frequent pattern mining-like approach is applied to this transformed data into order to determine those objects which belong to the same spatial cluster for a significant number of time-stamps. An Apriori-like approach is used for this purpose, in combination with a number of additional pruning tricks, which use the temporal characteristics of the data. Since the consecutiveness of the membership information is not used in the pattern-mining phase, the swarms are based on significant levels of co-location at any period in time.

The problem of clustering is particularly useful from the perspective of trajectory mining, because it provides summary information which can be used for other applications. For example, the *TraClass* method proposed in [105] uses two kinds of clustering in order to provide additional summary information, which enables more effective classification. One kind uses the characteristics of different regions in the clustering, but it does not use the movement patterns. The other kind uses the characteristics of different trajectories in the clustering. The two kinds of clusters provide useful complementary information in the classification process. It has been shown in [105], how this additional information can be leveraged for a more effective classification process.

While clustering determines the *typical* movement patterns, a related problem is that of determining unusual (or *atypical*) movement patterns [105, 109, 110]. Such movement patterns are also referred to as outliers. Another variation on the problem is the determination of periodic patterns [114], which we wish to determine common patterns of movement which repeat periodically in the trajectory data, or hot routes in road networks [111]. A comprehensive range of trajectory mining techniques have been developed in the context of the *MoveMine* project at UIUC [115].

While much of this work has been performed in the context of animals, similar techniques can be generalized to the case of humans. Human movements are of course somewhat more complex, because of the greater complexity of social interactions as compared to animals. Some recent work has been performed on studying the trajectory patterns of humans, which were collected from mobile phones [73]. It was shown that human trajectories show a high degree of temporal and spatial regularity, and each individual shows a highly time-independent charac-

teristic travel distance and a significant probability to return to a few highly frequented locations. On further simplification, it was shown that individual travel patterns collapse into a single spatial probability distribution. This suggests that humans follow simple reproducible patterns. This simple observation has consequences for all phenomena driven by human mobility, such as epidemic prevention, emergency response, urban planning and agent-based modeling.

A key area of research for mobile trajectory analysis is to determine frequent and repetitive trajectories in the data. The most basic analysis from this perspective is to determine similar trajectories to a given target trajectory. A variety of methods on the topic of indexing moving object databases may be found in [80]. The problem has also been studied in the context of the gps trajectories created by mobile phones [43, 174]. A method for performing user-oriented trajectory search for trip recommendations has been proposed in [147].

More generally, the work in [70] explores the sequential pattern mining problem in the context of trajectory pattern mining. The idea is to determine sequences of places in the data, which occur together frequently in the data, and with similar transition times. The sequential pattern mining paradigm can be extended to this case by incorporating temporal constraints into successive elements of the sequence.

Trajectory patterns can also be derived from geo-tagged photos, in which users utilize gps-enabled mobile phones to take photos and upload them. Since the user location and time is recorded, when they take the photo, this provides natural way to derive the trajectory of the user. For example, the work in [171] mines frequent sequential trajectory patterns from such geo-tagged social media. However, the number of patterns may be too large to be informative to a user. Therefore, a ranking mechanism is introduced in order to determine the importance of the different reported patterns. The relationships between users, locations and patterns and their importance are utilized for ranking purposes. For example, trajectories are considered important, if they are followed by important users, and contain important locations. The vice-versa relationships also hold in this case. These importance relationships are modeled in [171] with the use of matrices representing the pairwise relationships between users, locations and patterns. A system of equations is set up with these matrices and solved in order to determine the importance values of the different trajectories. In addition, a *diversification criterion* is introduced in order to ensure that trajectories with large segments in common are not reported simultaneously. This is done in order to maximize the amount of useful in information in a small number of presented results. The GPS data can also be used in order to determine

interesting locations, trajectories, or even the transportation modes of the different users [180, 181].

While social sensing applications are generally defined for the case of people, a similar analysis can be applied to the case of online tracking of animals. For example, animals which are drawn from the same community or family may be considered to have implicit links among them. Such links can be utilized for the perspective of detailed understanding of how community and family membership affects geographical patterns. Such information can be very useful for a variety of applications, such as building disease propagation models among animals.

7.1 Integrating Sensor Data with Heterogeneous Media for Enhanced Mining and Inference

Many of the devices (such as mobile phones), which enable social sensing applications are *convergent devices*, which provide multiple functionalities in recording different kinds of media data. For example, most mobile phones today provide the capability to record photos, videos, text blogging and tweets, and upload them directly in real time. Thus, such media data automatically becomes *geo-tagged*, and this additional information provides a rich source of information for improving the mining process.

For example, the problem of providing location and activity recommendations on the basis of user contributed comments and their GPS trajectories has been studied in [179]. The user comments provide deeper insights into their activity histories, which can be leveraged for a better mining process. The collective wisdom of the trajectories and comments of different users can be leveraged in order to provide answers to questions such as the following:

- For a particular activity, what are the most appropriate places to visit?
- For a particular location, which a user has already visited, what are the other activities that can be performed at that location?

In order to achieve this goal, the user location and activity histories are used as the input. We note that the activity histories can only be indirectly derived from user comments, by mining the relevant words in the comments, which are related to specific activities. The location features and activity-activity correlations are mined in order to obtain additional knowledge. A collective matrix factorization methods was applied in [179] in order to mine interesting locations and activities and recommend them to users. Location information is also useful for rec-

ommending specific non-spatial products or items on the basis of spatial history, as discussed in [101].

This general principle can also be applied for geographical topic discovery and comparison from GPS-associated documents [172]. While topic modeling of documents is widely known, the use of geographic information in the process provides rich opportunities for adding additional insights into the process. Many interesting concepts, including cultures, scenes, and product sales, correspond to specialized geographical distributions. The goal of geographical topic discovery is to discover such interesting concepts. The two main questions in this context are as follows:

- What are the coherent topics of interest in the different geographical regions?
- How can the different topics be compared across the different geographical regions?

The work in [172] proposes and compares three different models which use pure location, pure text, and a joint model of location and text, which is referred to as LGTA (Latent Geographical Topic Analysis). The approach is used on several data sets from the *Flickr* web site. It is shown that the first two methods work in some data sets but fail in others, whereas LGTA works well in all data sets at finding regions of interest and also providing effective comparisons of the topics across different locations. This suggests that geographical data and content data provide complimentary information to one another for the mining process. Further work along this direction in the context of *topic evolution* is proposed in [169]. From a real-time perspective, it is often useful to utilize location information for providing context-sensitive newsfeeds to users [19].

An interesting application in [32] shows that the latent information in user trajectories, which are extracted from the GPS data in photos can even be used to generate travel itineraries. For example, the media sharing site *Flickr*, allows photos to be stamped by the time of when they were taken and be mapped to points of interest with the use of geographical and tag meta-data. This information can be used to construct itineraries with a two-step approach. First, the photo streams of individual users are extracted. Each photo stream provides estimates on where the user was, how long he stayed at each place, and what was the transit time between places. In the second step, all user photo streams are aggregated into a Point of Interest (POI) graph. Itineraries are then automatically constructed from the graph based on the popularity of the POIs, and subject to the user time and destination constraints.

8. Social Sensing Applications

In this section, we will discuss a number of recent applications which have been designed in the context of sensors and social networks. Many of these applications are related to storage and processing of mobile data which is continuously collected over time. Such mobile data can be used in order to provide real time knowledge of the different users to one another, trigger alerts, provide an understanding of social trends, and enable a variety of other applications. In this section, we will discuss a number of social-centric applications, which have been developed in recent years. These include specific systems which have been designed by companies such as Google, Microsoft, and SenseNetworks, as well as a number of generic applications, which have not yet been fully commercialized.

8.1 CrowdSourcing Applications for User-Centered Activities

A number of crowdsourcing applications have recently been designed for providing feedback in a number of user-centered activities such as buying behavior, location trends, or other miscellaneous user activity. Examples of such applications include *Google Latitude*, *CitySense*, *Macrosense* and *Wikitude* applications. The Citysense and Macrosense applications both collect real-time data from a variety of GPS-enabled cell phones, cell phone tower triangulation, and GPS-enabled cabs. The two applications share a number of similarities in terms of the underlying methodology, but they have different features which are targeted towards different kinds of audiences. We describe them below:

8.1.1 The Google Latitude Application. The Google Latitude Application uses GPS data which is collected from Google map users on mobile cell phones. It is also possible to collect more approximate data with the use of cell phone tower location data (in case the mobile phones are not GPS enabled), or with the use of IP addresses of a computer which is logged into the personalized google page called *iGoogle*. The Latitude application enables the creation of *virtual friends*, who are essentially other users that carry the same location-enabled device, or use other devices such as personal computers which can transmit approximate location data such as IP-addresses. A number of other applications which enabled by the Google Latitude master application are as follows:

- **Location Alerts:** The application allows the triggering of alerts when someone is near their latitude friends. The alerts are trig-

gered only when something interesting is being done. This is done on the basis of both time and location. For example, an alert could be triggered when two friends are at a routine place, but an unusual time. Alternatively, it could be triggered when two friends are at a routine time but unusual place.

- **Public Location Badge:** It is possible to post one's location directly on blog or social network. This in turn increases the visibility of one's information to other users of the site.
- **Use with Chat Applications:** The mobile location can also be used in conjunction with the *Google Talk* application which allows users to chat with one another. Users who are chatting with one another can see each other's location with the use the embedded latitude functionality.

It is clear that all of the above techniques change the nature and dynamics of social interactions between users. For example, the triggering of alerts can itself lead to a changed pattern of interaction among the different users. The ability to mine the dynamics of such interactions is a useful and challenging task for a variety of applications.

While *Google Latitude* is perhaps the most well known application, it is by no means the only one. A number of recent applications have been designed which can track mobile devices on the internet through GPS tracking. Some of these applications have been designed purely for the purpose of tracking a device which might be lost, whereas others involve more complex social interactions. Any software and hardware combination which enables this has the potential to be used for social sensing applications. Some examples of such applications are as follows:

- **Navizon Application:** This application [188] uses GPS in order to allow social interactions between people with mobile phones. It allows the tracking of mobile friends, coverage of particular areas, and trails followed by a particular user.
- **iLocalis Application:** This application [189] is currently designed only for particular mobile platforms such as the iPhone, and it allows the tracking of family and friends. In addition, it is also designed for corporate applications in which a group of mobile employees may be tracked using the web. Once friendship links have been set up, the application is capable of sending a message to the friends of a particular user, when they are nearby.

8.1.2 CitySense Application. The citysense application is designed for the broad consumer base which carries mobile cell phones.

The Citysense application is designed to track important trends in the behavior of people in the city. For example, the application has been deployed in San Francisco, and it can show the busiest spots in the city on a mobile map.

The CitySense application also has a social networking version of a collaborative filtering application. The application stores the personal history of each user, and it can use this personal history in order to determine where other similar users might be. Thus, this can provide recommendations to users about possible places to visit based on their past interests.

A very similar application is the WikiCity project [25] which collects real time information with the use of GPS and mobile devices. These are then used to collect the location patterns of users, and their use in a variety of neighborhoods.

8.1.3 MacroSense Application. The MacroSense application [195] is similar in terms of the data it collects and kind of functionality it provides; however it is focussed towards the commercial segment in predicting consumer behavior. The application can predict the behavior of customers based on their location profile and behavior. The application can predict what a particular customer may like next. The broad idea is to segment and cluster customers into marketing groups based on their behavior, and use this information in order to make predictions. For example, the popularity of a product with users who are most like the target can be used for predictive purposes. Thus, this approach is somewhat like collaborative filtering, except that it uses the behavior of customers rather than their feedback. The effectiveness of particular behaviors which predict the interests are also used. This analysis can be performed in real time, which provides great value in terms of predictive interactions. The analytics can also be used in order to predict group influences for the behaviors of the underlying subjects.

8.1.4 LiveCompare for Grocery Bargain Hunting. A system called *LiveCompare* [46] has been proposed for grocery bargain hunting with the use of participatory sensing. *LiveCompare* works with participating grocery store shoppers with camera-enabled mobile phones with internet access. Virtually, all smart phones today are camera- and internet-enabled, and therefore this requirement is quite a reasonable one, which does not require any additional expenditure from participating users. The camera phones are used in order to snap a picture of the product's price tag. We note that this price tag, typically contains a UPC bar code, from which information about the product can be ex-

tracted. The barcodes can be decoded from the photograph with the use of barcode libraries such as *ZXing*[193]. At this point, the numerical UPC value and the just-taken photograph are transferred to *LiveCompare*'s central server. This data is stored in *LiveCompare*'s database for use in future queries, and the UPC value determines the unique product for which price comparisons are requested. The client also sends its GPS or GSM cell information to the server so that the current store can be identified. This location information allows *LiveCompare* to limit query results to include only nearby stores. Results include store information and the option to view the time-stamped photographs associated with the specific product in question at each store. Users are not required to manually input pricing data in order to improve trustworthiness; this low burden of participation improves the ability to recruit participants during deployment. In any participatory system, it is recognized that to contribute data, users give up their time, attention, and mobile device's battery power. Therefore, it is critical to ensure that users have sufficient incentive to participate. *LiveCompare* directly addresses this challenge through its query protocol. When a user submits a query from a grocery store, he identifies the product for which he wants price comparison information by snapping a photograph of the product's price tag (including bar code). The server *appends the photograph submitted during the query to its database.* Thus, by requiring that a geotagged photograph be uploaded as part of a query, *LiveCompare* automatically populates its database whenever a user initiates a query. Thus, the principle of increasing incentive and participation is: "*To use, you must contribute.*"

The problem of sharing consumer prices with the use of mobile phones has started gaining attention recently. For example, the *Mobishop* system for sharing consumer prices with mobile phones has been proposed in [146]. Methods for sharing fuel prices with the help of a network of mobile phone cameras has been proposed in [50].

8.1.5 Location-Aware Search, Feedback, and Product Recommendation. Virtually all mobile phones have applications which enable GPS-based searches for popular businesses such as restaurants, coffee shops, gas stations, or department stores. For example, the $YellowPages$ application on most mobile phones is now GPS enabled. Furthermore, many social review systems (which allow users to share their opinions about businesses) such as *Yelp* [199] integrate the social reviews with GPS-enabled search. This allows a user to not only search for business of interest, but even businesses which have positive reviews associated with them. These applications also allow users to

enter their feedback about their own experiences into the system. This unique combination of user-based text feedback and mobile sensing is powerful combination, which provides unprecedented information and flexibility in terms of combining location information with the social opinions of other users.

For shopping applications, the ability to perform *recommendations* is a useful functionality in a wide variety of scenarios. Since spatial location is highly corrected to user-buying behavior, it is natural to use GPS information for such applications. An important observation in this work is that some items or products (eg. restaurants) are spatial in nature, whereas others (eg. movies) are non-spatial in nature, since the user-experience with the product in not locality dependent. Similarly, ratings of a user may sometimes be spatial in nature, when some locations (eg. *FourSquare*) allow location-based check in and ratings. The work in [101], which uses location-based ratings for the recommendation process. The *LARS* [101] suppots a taxonomy of three classes of location-based ratings– (i) spatial ratings for non-spatial items, (ii) non-spatial ratings for spatial items, and (iii) spatial ratings for spatial items. *LARS* uses spatial partitioning in order to utilize spatially closer users for the recommendation process. This maximizes system scalability without affecting the recommendation quality. Furthermore, since users prefer closer locations for the purpose of their buying behavior, the spatial nature of items is used in order to recommend items which are closer to querying users. This is modeled with the use of a *travel penalty*. It has been shown in [101], that these features can be used either separately or together in order to maximize the effectiveness of the recommendation process.

8.1.6 Wikitude Augmented Reality Application. The wikitude application [191] is designed for mobile phones (such as *Blackberry* and *iPhone*, and uses the GPS location and the compass within mobile phones in order to provide an "augmented reality" experience from the mobile phone, by pointing it in different directions. The application is connected with social networking application such as *Facebook* and *Twitter*, and can collect messages, tweets and events from users within a particular neighborhood, and can be made available to the user. In addition, by pointing the device in a particular direction, it may be possible to find useful points of interest such as restaurants, shopping places, or movie theaters. It is even possible to determine mobile coupons and discounts from shops within a particular neighborhood with this kind of application.

8.1.7 Microsoft SensorMap. Most of the applications discussed above are based on location data, which is automatically collected based on user behavior. The SensorMap project [127] at Microsoft allows for a more general framework in which users can *choose to publish any kind of sensor data*, with the understanding that such shared knowledge can lead to interesting inferences from the data sets. For example, the sensor data published by a user could be their location information, audio or video feeds, or text which is typed on a keyboard. The goal of the SensorMap project is to store and index the data in a way such that it is efficiently searchable. The application also allows users to index and cache data, so that users can issue spatio-temporal queries on the shared data.

The SensorMap project is part of the SenseWeb project, which allows sharing and exploring of sensor streams over geo-centric interfaces. A number of key design challenges for managing such sensor streams have been discussed in [120]. Other key challenges, which are associated with issues such as the privacy issues involved with continuously collecting and using the sensors which are only intermittently available is discussed in [99].

8.2 RFID Technology: The Internet of Things

The general idea of social sensing can also be extended to applications which use RFID technology to track objects, as opposed to "social" sensing paradigms, which track people. This technology is also transformative for social sensing, because of the close relations between people and objects in many scenarios, and the social inferences, which may be possible with the use of such tracking technology. The idea is that radio frequency tags are attached to commercial products or other objects to be tracked, and these tags do little more than provide their unique *Electronic Product Code (EPC)* to nearby sensor readers. Thus, the movements of objects of interest can be identified by appropriate receivers at checkpoints where the object movement is tracked. Furthermore, these readers can be connected to the internet, where they can publish the data about the objects, and enable effective search, querying, and indexing of these objects with the use of the semantic web framework [82].

Animals, commercial products, baggage and other high volume items are often tracked with the use of Radio Frequency Identification (RFID) tags. For example, RFID technology has been used to track the movement of large animals such as whales with chips embedded in them. Such chips may sometimes even have transmitters embedded in them,

which can be picked up by satellite. RFID technology has even found application in a number of medical applications, in which RFID chips are embedded in patients in order to track their case history. RFID Technology has lead to the general vision of *the internet of things* [16], in which uniquely identifiable objects can be continuously tracked over time. In the case of commercial applications, the products may have implicit links among them which correspond to shared batches or processes during the production and transportation process. Such tracking data can be used in conjunction with linkage analysis in order to determine the causality and origin of tainted products. It can also be used to track the current location of other products which may be tainted. Such data is typically quite noisy, error-prone, incomplete, and massive in volume. Thus, this leads to numerous challenges in data compression, storage and querying. A detailed tutorial on RFID methods may be found in [81]. The technology is also discussed in some detail in a later chapter of this book [4, 5].

8.3 Vehicular Participatory Sensing

In vehicular participatory sensing, a variety of sensor data from vehicles such as mobile location, or other vehicular performance parameters may be continuously transmitted to users over time. Such data may be shared with other users *in the aggregate* in order to preserve privacy. This is the social aspect of such applications, since they enable useful individual decisions based on global patterns of behavior. In addition, vehicular participatory sensing may be used in order to enable quick responses in case of emergencies involving the vehicle operation. We note that much of the work discussed above for animal and moving object trajectory mining [104, 105, 112–115, 109, 110] are also applicable to the case of vehicular data. In addition, vehicular data poses unique challenges in terms of data collection, sensing, transmission and privacy issues. Classic examples of vehicular participatory sensing include the *CarTel* [88] and *GreenGPS* systems [64]. While we will focus on a detailed discussion of these systems as the most well known representatives of vehicular participatory sensing, a number of other sensing systems have been designed for different applications such as traffic monitoring and road conditions [124], cyclist experience mapping [55, 142], and the determination of transportation modes [144].

The problem of sharing bike track paths by different users has been explored in [142]. The problem of finding bike routes is naturally a trial-and-error process in terms of finding paths which are safe and enjoyable. The work in [142] designs *Biketastic*, which uses GPS-based sensing on

a mobile phone application in order to create a platform which enables rich sharing of biker experiences with one another. The microphone and the accelerometer embedded on the phone are sampled to infer route noise level and roughness. The speed can also be inferred directly from the GPS sensing abilities of the mobile phone. The platform combines this rich sensor data with mapping and visualization in order to provide an intuitive and visual interface for sharing information about the bike routes. A different application uses the time-stamped location information in order to determine the mobility profiles of individuals [144]. Next, we will discuss the *Cartel* and *GreenGPS* systems.

8.3.1 CarTel System. The *CarTel* project at MIT [88] is designed for mining and managing large amounts of sensor data, which are derived from vehicular participatory sensing. The most common data is vehicular position data, from which large amounts of information about road congestion, conditions, and other violations may be determined. The project focusses on the collection and use of such data in an efficient and privacy-preserving way. The actual data may be collected either from mobile phones in the car or from embedded devices within the car itself. For example, the Onboard Diagnostics Interface (OBD-II) equipped on modern cars can be used to collect tremendous amounts of useful data in this context. The OBD-II is a diagnostic system that monitors the health of the automobile using sensors that measure approximately 100 different engine parameters. Examples of monitored measurements include fuel consumption, engine RPM, coolant temperature and vehicle speed. Vehicles that have been sold in the United States after 1996 are mandatorily equipped with a sensing subsystem called the On-Board Diagnostic (OBD-II) system. A number of key components of the *CarTel* system are as follows:

Traffic Mitigation: In this case, two systems *VTrack* and *CTrack* [154, 155] have been proposed for processing error-prone position streams for estimating trajectory delays accurately. Since the location data is typically error-prone as a result of transmission errors, or outages, the technique is designed to be resistant to errors. In particular, the *CTrack* system [154] can work with the position data from cellular base stations, in which the location error is much higher than GPS data. The system continuously collects the data, and combines real-time and historic delay estimates to produce predictions of future delays at various points in time in the future. The results of the predictive model are sent to a commute portal where users can view the data along with appropriate traffic routing strategies.

Road Conditions: The idea in this approach [58] is to use the oppor-

tunistic mobility of sensor-equipped vehicles to detect and report the surface conditions of roads. Each car in the system carries 3-axis acceleration and GPS sensors, gathering location-tagged vibration data. The system uses uses *CarTel*'s opportunistic wireless protocols to deliver the data over whatever wireless network is available to a back-end server (discussed in detail below). The server processes this vibration data using machine learning techniques in order to predict the surface conditions.

Data Muling and Networking: The data collected in a vehicle (such as information about the road surface conditions) may sometimes need to be routed to a back-end server, even in cases where a continuous mobile connection is not available. In such cases, intermittent wifi access points may be available along the route of the vehicle. should use wireless networks opportunistically [57, 29]. The idea is to use a combination of WiFi, Bluetooth, and cellular connectivity, using whatever mode is available, while being completely transparent to underlying applications. In some cases, cars may be used as mules in order to carry the data, when direct connectivity is not available [29].

Query Processing of Intermittently Connected Data: Participatory sensing sensor network applications must cope with a combination of node mobility and high data rates when media-rich data such as audio, video or images are being captured by a sensors. As a result of the mobility, the sensor networks may display intermittent and variable network connectivity, and often have to deliver large quantities of data relative to the bandwidth available during periods of connectivity. In order to handle this challenge, a system known as *ICEDB (Intermittently Connected Embedded Database)* [178] was proposed, which incorporates a delay-tolerant continuous query processor, coordinated by a central server and distributed across the mobile nodes. The system contains algorithms for prioritizing certain query results to improve application-defined utility metrics.

Privacy Protection: The process of tracking the position of individual vehicles is fraught with numerous challenges from a privacy perspective. Therefore, techniques are needed to be able to compute appropriate functions on the location data, without violating individual privacy. The *CarTel* system provides excellent privacy protection of user location data, while being able to compute aggregate functions on the location statistics. This is called the *VPriv* system [134]. More details on this system are discussed in the section on privacy in this chapter.

8.3.2 Green GPS. Green GPS [64] is a participatory sensing navigation service that allows drivers to find the most fuel-efficient routes

for their vehicles between arbitrary end-points. Green GPS relies on data collected by individuals from their vehicles as well as on mathematical models to compute fuel efficient routes.

The most fuel efficient route may depend on the vehicle and may be different from the shortest or fastest route. For example, a fast route that uses a freeway may consume more fuel because fuel consumption increases non-linearly with speed. Similarly, the shortest route that traverses busy city streets may be suboptimal because of downtown traffic. The data collected by the different drivers can be used in conjunction with mathematical models in order to make effective predictions. A natural question arises as to the nature of the data which can be collected by the different individuals for this purpose. The service exploits measurements of standard vehicular sensor interfaces that give access to most gauges and engine instrumentation.

To build its fuel efficiency models, Green GPS utilizes a vehicle's OBD-II system and a typical scanner tool in conjunction with a participatory sensing framework. The team is collecting data from vehicles driven by research participants to determine what factors influence fuel consumption. The data collected by the participants is driving the creation of a mathematical model that enable computing fuel consumption of different cars on different road segments. Early studies have shown that a 13% reduction in consumer gas consumption is possible over the shortest path and 6% over the fastest path.

8.4 Participatory Sensing in Healthcare

A variety of *participatory sensing techniques* can be used for enabling real-time services. In participatory sensing, users agree to allow data about them to be transmitted in order to enable a variety of services which are enabled in real time. The ability to carry such devices allows its use for a variety of healthcare applications involving the elderly. For example, elderly patients can use this in order to call for care when necessary. Similarly, such sensing devices can be utilized for a variety of safety and health-care related applications.

Several companies such as *Vivometrics*, *Bodymedia*, and *Mini-mitter* have [196–198] have designed enhanced versions of the Holter ECG monitoring device [85], which is commonly used for ambulatory services. These enhanced devices are able to monitor a patient's ECG for longer periods of time, and transmit them remotely to the physician. Such a concept is very useful for high-risk populations (such as elderly patients), because it allows quick and time-critical responses, which has the potential to save lives. While *inpatient mobile sensing* is quite common in

medical domains, the advancement of this natural concept to more proactive applications such as round-the-clock monitoring has only been a recent development.

A method called *LiveNet* is proposed in [150], in which a flexible distributed mobile platform that can be deployed for a variety of proactive healthcare applications that can sense one's immediate context and provide feedback. This system is based on standard PDA hardware with customized sensors and a data acquisition hub, which provides the ability for local sensing, real-time processing, and distributed data streaming. This integrated monitoring system can also leverage off-body resources for wireless infrastructure, long-term data logging and storage, visualization/display, complex sensing, and computation-intensive processing. The *LiveNet* system also allows people to receive real-time feedback from their continuously monitored and analyzed health state. The system can also communicate health information to caregivers and other members of an individual's social network for support and interaction. One of the attractive features of this system is that it can combine general-purpose commodity hardware with specialized health/context sensing within a networked environment. This creates a multi-functional mobile healthcare device that is at the same time a personal real-time health monitor, which provides both feedback to the patient, the patient's social network, and health-care provider.

We note that a significant number of predictions can also be made without collecting data which is clinical in nature. In particular, the daily activities of an individual can provide key insights into their health. Smartphones have now become sophisticated enough that the data from the different sensors can be fused in order to infer the daily activities of an individual [65]. For example, the presence of illness and stress can affect individuals in terms of their total communication, interactions with respect to the time of day, the diversity and entropy of face-to-face communications and their movement. In order to achieve this goal, the work in [121] uses mobile phone based co-location and communication sensing to measure different attributes about the daily activity of an individual. It has been shown in [121], that the collection of even simple day-to-day information has a powerful effect on the ability to make an accurate diagnosis. Methods have also been proposed for finding sequential patterns from human activity streams, in order to determine the key activity trends over time. Furthermore, such activity monitoring cane be used to model the influence of different individuals on each other in terms of their daily activities. The work in [122] used a mobile phone platform to examine how individuals are influenced by face-to-face interactions in terms of their obesity, exercise and eating habits. It was shown that

such interactions do have a significant influence over individuals, which may propagate in the social network over time. Such an approach [54, 139, 140] has also been applied to the problem of geriatric care. This is because medical conditions such as dementia in older patients show up as specific kinds of activity patterns over time. It has been shown in [54], how such activity recognition methods can be used in the context of geriatric care.

From a predictive modeling perspective, a key challenge which arises is that a large amount of data may potentially need to be collected simultaneously from a large number of patients in order to make accurate real time predictions. This requires the design of fast data stream processing algorithms [7]. A recent paper [89] proposes a number of real-time data stream mining methods for fast and effective predictive modeling from sensor data. This kind of approach can be used for a wide variety of medical conditions, though the nature of the data collected and the predictive modeling would depend upon the nature of the disease modeling at hand. For example, the work in [84] discusses a variety of methods which can be used for diabetes monitoring with the use of collected data. Another interesting method for health and fitness monitoring has been developed in [119], in which modern mobile phones are used in order to both sense and classify the activities of an individual in real time. It has been shown that such machine learning algorithms can be used in conjunction with the collected data in order to provide effective monitoring and feedback. A discussion of some of the challenges in selecting sensors for health monitoring with the use of participatory sensing may be found in [41].

9. Future Challenges and Research Directions

In this chapter, we examined the emerging area of integrating sensors and social networks. Such applications have become more commonplace in recent years because of new technologies which allow the embedding of small and unobtrusive sensors in clothing. The main challenges of using such technologies are as follows:

- Such applications are often implemented on a very large scale. In such cases, the database scalability issues continue to be a challenge. While new advances in stream processing have encouraged the development of effective techniques for data compression and mining, mobile applications continue to be a challenge because of the fact that *both* the *number* of streams and rate of data collection may be extremely large.

- A major challenge in sensor-based social networking are the privacy issues inherent in the underlying applications. For example, individuals may not be willing to disclose their locations [66] in order to enable applications such as proximity alerts. In many cases, such constraints can greatly reduce the functionality of such applications. A major challenge in such applications is to provide individual hard guarantees on their privacy level, so that they become more willing to share their real time information.

- The trust issues continue to be a challenge for such applications, because of the openness of such systems in allowing participants to contribute information. Furthermore, the goals of privacy and trust seem to be at odds with one another, because the former is achieved by lowering data fidelity, and the latter requires higher data fidelity.

- Battery life continues to be a severe constraint in such applications. Therefore, it is critical to tailor application design to work efficiently in power-constrained scenarios.

- The architectural challenges for such systems continue to be quite extensive. For example, the use of centralized processing methods for such large systems does not scale well. Therefore, new methods [120, 127] have moved away from the centralized architecture for stream collection and processing.

The future challenges of such research include the development of new algorithms for large scale data collection, processing and storage. Some advancements [7, 120, 127] have already been made in this direction.

Acknowledgements

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] T. Abdelzaher, Y. Anokwa, P. Boda, J. Burke, D. Estrin, L. Guibas, A. Kansal, S. Madden, J. Reich. Mobiscopes for Human Spaces. *IEEE Pervasive*, 6(2), pp. 20–29, April 2007.

[2] O. Abul, F. Bonchi, M. Nanni. Never Walk Alone: Uncertainty for Anonymity in Moving Object Databases, *ICDE Conference*, 2008.

[3] H. Ahmadi, N. Pham, R. Ganti, T. Abdelzaher, S. Nath, J. Han. Privacy-aware regression modeling of participatory sensing data. *SenSys*, 2010.

[4] C. C. Aggarwal, J. Han. A Survey of RFID Data Processing, *Managing and Mining Sensor Data*, Springer, 2013.

[5] C. C. Aggarwal, N. Ashish, A. Sheth. The Internet of Things: A Survey from the Data-Centric Perspective, *Managing and Mining Sensor Data*, Springer, 2013.

[6] C. C. Aggarwal, T. Abdelzaher. Integrating Sensors and Social Networks, *Social Network Data Analytics*, Springer, 2011.

[7] C. C. Aggarwal (ed.) Data Streams: Models and Algorithms, *Springer*, 2007.

[8] C. C. Aggarwal, H. Wang (ed.) Managing and Mining Graph Data, *Springer*, 2010.

[9] C. C Aggarwal, P. Yu (ed.) Privacy-Preserving Data Mining: Models and Algorithms, *Springer*, 2008.

[10] C. C. Aggarwal, P. S. Yu. Online Analysis of Community Evolution in Data Streams, *SIAM Conference on Data Mining*, 2005.

[11] C. C. Aggarwal, S. Lin, P. Yu. On Influential Node Discovery in Dynamic Social Networks, *SDM Conference*, 2012.

[12] C. C. Aggarwal, Y. Zhao, P. Yu. On Clustering Graph streams, *SIAM Conference on Data Mining*, 2010.

[13] D. Agrawal, C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the 20th ACM SIGMOD Symposium on Principles of Database Systems*, pages 247–255, 2001.

[14] R. Agrawal, R. Srikant. Privacy preserving data mining. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, pages 439–450, Dallas, TX, May 2000.

[15] C. Asavathiratham, The lnfluence Model: A Tractable Representation for the Dynamics of Networked Markov Chains, *TR, Dept. of EECS, MIT*, Cambridge, 2000.

[16] K. Ashton. That 'Internet of Things' Thing. In: *RFID Journal*, 22 July, 2009.

[17] B. Bamba, L. Liu, P. Pesti, T. Wang. Supporting Anonymous Location Queries in Mobile Environments with PrivacyGrid, *World Wide Web Conference*, 2008.

[18] N. Banerjee, S. Agarwal, P. Bahl, R. Chandra, A. Wolman, M. Corner. Virtual Compass: Relative Positioning to Sense Mobile Social Interactions. *Pervasive*, 2010.

[19] J. Bao, M. Mokbel, C.-Y. Chow. GeoFeed: A Location Aware News Feed System. *ICDE Conference*, pp. 54–65, 2012.

[20] A. Beach, M. Gartrell, X. Xing, R. Han, Q. Lv, S. Misra, K. Seada. Fusing Mobile, Sensor and Social Data to Fully Enable Context Aware Computing, *Hotmobile*, 2010.

[21] A. Beberg, V. S. Pande. Folding@home: lessons from eight years of distributed computing. *IEEE International Parallel and Distributed Processing Symposium*, pp. 1–8, 2009

[22] M. Benkert, J. Gudmundsson, F. Hubner, T. Wolle. Reporting flock patterns. *COMGEO*, 2008.

[23] D. Brabham. Crowdsourcing as a model for problem solving: An introduction and cases. *The Journal of Research into New Media Technologies*, 14(1), pp. 75-90, 2008.

[24] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, M. B. Srivastava. Participatory Sensing, *WWW Conference*, 2006.

[25] F. Calabrese, K. Kloeckl, C. Ratti. Wikicity: Real-Time Urban Environments. *IEEE Pervasive Computing*, 6(3), 52-53, 2007. http://senseable.mit.edu/wikicity/rome/

[26] A. T. Campbell et al. The Rise of People Centric Sensing, *IEEE Internet Computing*, 12(4), 2008.

[27] X. Cao, G. Cong, C. S. Jensen. Mining Significant Semantic Locations From GPS Data. *VLDB Conference*, 2010.

[28] D. Chakrabarti, R. Kumar, A. Tomkins. Evolutionary clustering. *KDD Conference*, 2006.

[29] K. Chen. CafNet: A Carry-and-Forward Delay-Tolerant Network. *MEng Thesis, MIT EECS*, Feb. 2007.

[30] R. Cheng, Y. Zhang, E. Bertino, S. Prabhakar. Preserving user location privacy in mobile data management infrastructures. *Privacy Enhancing Technologies*, 2006.

[31] Y. Chi, X. Song, D. Zhou, K. Hino, B. Tseng. Evolutionary spectral clustering by incorporating temporal smoothness. *ACM KDD Conference*, 2007.

[32] M. Choudhury, M. Feldman, S. Amer-Yahia, N. Golbandi, R. Lempel, C. Yu. Automatic construction of travel itineraries using social breadcrumbs. *HT*, 2010.

[33] T. Choudhury A. Pentland. The Sociometer: A Wearable Device for Understanding Human Networks. *International Sunbelt Social Network Conference*, February 2003.

[34] T. Choudhury, A. Pentland. Sensing and Modeling Human Networks using the Sociometer, *International Conference on Wearable Computing*, 2003.

[35] T. Choudhury, A. Pentland. Characterizing Social Networks using the Sociometer. *North American Association of Computational Social and Organizational Science*, 2004.

[36] T. Choudhury, B. Clarkson, S. Basu, A. Pentland. Learning Communities: Connectivity and Dynamics of Interacting Agents. *International Joint Conference on Neural Networks*, 2003.

[37] T. Choudhury, M. Philipose, D. Wyatt, J. Lester. Towards Activity Databases: Using Sensors and Statistical Models to Summarize People's Lives. *IEEE Data Engineering Bulletin*, Vol. 29 No. 1, March 2006.

[38] C.-Y. Chow, M. F. Mokbel. Privacy of Spatial Trajectories. *Computing with Spatial Trajectories*, pp. 109–141, 2011.

[39] A. Clauset, M. E. J. Newman, C. Moore. Finding community structure in very large networks. *Phys. Rev. E 70, 066111*, 2004.

[40] I. Constandache, R. Choudhury, I. Rhee. Towards mobile phone localization without war-driving, *INFOCOM Conference*, 2010.

[41] D. Cook, L. Holder. Sensor selection to support practical use of health-monitoring smart environments. *Wiley Interdisc. Rev.: Data Mining and Knowledge Discovery* 1(4): pp. 339–351, 2011.

[42] D. Cook, W. Song. Ambient Intelligence and Wearable Computing: Sensors on the Body, in the Home and Beyond. *JAISE*, 1(2): pp. 83–86, 2009.

[43] C. Costa, C. Laoudias, D. Zeinalipour-Yazti, D. Gunopulos. SmartTrace: Finding similar trajectories in smartphone networks without disclosing the traces. *ICDE Conference*, 2011.

[44] M. Coscia, S. Rinzivillo, F. Giannotti, D. Pedreschi. Optimal Spatial Resolution for the Analysis of Human Mobility, *ASONAM Conference*, 2012.

[45] L. P. Cox. Truth in Crowdsourcing, *IEEE Journal on Security and Privacy*, Volume 9, Issue 5, September 2011.

[46] L. Deng, L. P. Cox. Livecompare: grocery bargain hunting through participatory sensing. *HotMobile*, 2009.

[47] A. Doan, R. Ramakrishnan, A. Halevy. Crowdsourcing systems on the World-Wide Web. *Communications of the ACM*, 54(4), pp. 86–96, 2011.

[48] X. Dong, L. Berti-Equille, D. Srivastava. Truth discovery and copying detection in a dynamic world. *VLDB*, 2(1): pp. 562–573, 2009.

[49] X. Dong, L. Berti-Equille, D. Srivastava. Integrating conflicting data: the role of source dependence. *Proc. VLDB Endow.*, 2: pp. 550–561, August 2009.

[50] Y. Dong, S.S. Kanhere, C.T. Chou, N. Bulusu. Automatic collection of fuel prices from a network of mobile cameras, *IEEE DCOSS*, 2008.

[51] A. Dua, N. Bulusu, W. Feng, W. Hu. Towards Trustworthy Participatory Sensing. *Proceedings of the Usenix Workshop on Hot Topics in Security*, 2009.

[52] N. Eagle, A. Pentland. Reality Mining: Sensing Complex Social Systems, *Personal and Ubiquitous Computing*, 10(4), 2006.

[53] N. Eagle, A. Pentland, D. Lazer. Inferring social network structure using mobile phone data. *Proceedings of the National Academy of Sciences (PNAS)*, 106:15274–15278, 2009.

[54] M. Schmitter-Edgecombe, P. Rashidi, D. Cook, L. Holder. Discovering and Tracking Activities for Assisted Living, *The American Journal of Geriatric Psychiatry*, In Press, 2011.

[55] S. Eisenman, E. Miluzzo, N. Lane, R. Peterson, G. Ahn, A. Campbell. The Bikenet Mobile Sensing System for cyclist experience mapping, *ACM Sensys*, 2007. Extended version appears in *ACM TOSN*, 6(1), 2009.

[56] W. Enck, P. Gilbert, B. Chun, L. P. Cox, J. Jung, P. McDaniel, A. Sheth. Taintdroid: An information-flow tracking system for realtime privacy monitoring on smartphones. *OSDI*, 2010.

[57] J. Eriksson, H. Balakrishnan, S. Madden, Cabernet: Vehicular Content Delivery Using WiFi. *ACM MOBICOM Conference*, 2008.

[58] J. Eriksson, L. Girod, B. Hull, R. Newton, S. Madden, H. Balakrishnan. The Pothole Patrol: Using a Mobile Sensor Network for Road Surface Monitoring. *MobiSys*, 2008.

[59] A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the SIGMOD/PODS Conference*, pages 211–222, 2003.

[60] C. Faloutsos, T. Kolda, J. Sun. Mining Large Time-Evolving Data using Matrix and Tensor Tools, *ICDM Conference*, 2007.

[61] R. K. Ganti, N. Pham, Y.-E. Tsai, T. F. Abdelzaher. PoolView: Stream Privacy in Grassroots Participatory Sensing, *SenSys*, 2008.

[62] R. K. Ganti, Y.-E. Tsai, T. F. Abdelzaher. SenseWorld: Towards Cyber-Physical Social Networks. *IPSN*, pp. 563–564, 2008.

[63] R. K. Ganti, P. Jayachandran, T. F. Abdelzaher, J. A. Stankovic, SATIRE: A Software Architecture for Smart AtTIRE. *Mobisys*, 2006.

[64] R. K. Ganti, N. Pham, H. Ahmadi, S. Nangia, T. Abdelzaher. GreenGPS: A Participatory Sensing Fuel-Efficient Maps Application. *Mobisys*, San Francisco, CA, June 2010.

[65] R. K. Ganti, S. Srinivasan, A. Gacic. Multi-sensor fusion in smartphones for lifestyle monitoring, *International Conference on Body Sensor Networks*, 2010.

[66] B. Gedik, L. Liu. Location Privacy in Mobile Systems: A Personalized Anonymization Model. *ICDCS Conference*, 2005.

[67] G. Ghinita, P. Kalnis, S. Skiadopoulos. PRIVE: Anonymous Location-Based Queries in Distributed Mobile Systems. *WWW Conference*, 2007.

[68] G. Ghinita, M. Damiani, C. Silvestri, E. Bertino. Preventing velocity-based linkage attacks in location-aware applications. *GIS*, 2009.

[69] G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi, K.-L. Tan. Private queries in location based services: anonymizers are not necessary. *SIGMOD Conference*, 2008

[70] F. Giannotti, M. Nanni, F. Pinelli, D. Pedreschi. Trajectory pattern mining. *ACM KDD Conference*, 2007.

[71] P. Gilbert, L. Cox, J. Jung, D. Wetherall. Towards Trustworthy Mobile Sensing, *Hotmobile*, 2010.

[72] P. Gilbert, J. Jung, K. Lee, H. Qin, D. Sharkey, A. Sheth, L. Cox. YouProve: Authenticity and Fidelity in Mobile Sensing, *ACM Sensys*, 2011.

[73] M. Gonzlez, Hidalgo, A.-L. Barabasi. Understanding Individual Human Mobility Patterns, *Nature*, 453: pp. 779–782, 2008.

[74] M. Gotz, S. Nath, J. Gehrke. MaskIt: Privately releasing user context streams for personalized mobile applications. *ACM SIGMOD Conference*, 2012.

[75] M. Gruteser, D. Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. *MobiSys*, 2003.

[76] M. Gruteser, X. Liu. Protecting privacy in continuous location-tracking applications. *IEEE Security and Privacy*, 2: pp. 28–34, 2004.

[77] J. Gudmundsson, M. van Kreveld. Computing longest duration flocks in trajectory data. *GIS*, 2006.

[78] T. Guo, K. Iwamura, M. Koga. Towards high accuracy road maps generation from massive GPS traces data. *Proc. of IGARSS*, pp. 667–670, 2007.

[79] M. Gupta, C. Aggarwal, J. Han, Y. Sun. Evolutionary Clustering and Analysis of Heterogeneous Bibliographic Networks, *ASONAM Conference*, 2011.

[80] R. Guting, M. Schneider. Moving Objects Databases, *Morgan Kaufmann*, 2005.

[81] J. Han, J.-G. Lee, H. Gonzalez, X. Li. Mining Massive RFID, Trajectory, and Traffic Data Sets (Tutorial). *ACM KDD Conference*, 2008.
Video of Tutoral Lecture at: http://videolectures.net/kdd08_han_mmrfid/

[82] O. Hassanzadeh, A. Kementsietsidis. Data Management Issues for the Semantic Web, *ICDE Conference*, 2012.

[83] Y. He, S. Barman, D. Wang, J. Naughton. On the complexity of privacy-preserving complex event processing. *PODS*, 2011.

[84] A. Helal, D. Cook, M. Schmatz. Smart Home-Based Health Platform for Behavioral Monitoring and Alteration of Diabetes Patients. *Journal of Diabetes Science and Technology*, 3(1): pp. 141–148, January 2009.

[85] N. J. Holter, J. A. Generelli. Remote recording of physiologic data by radio. *Rocky Mountain Medical Journal*, pp. 747–751, 1949.

[86] Z. Huang, W. Du, B. Chen. Deriving private information from randomized data. In *Proceedings of the 2005 ACM SIGMOD Conference*, pages 37–48, Baltimore, MD, June 2005.

[87] K. Huang, S. Kanhere, W. Hu. Preserving Privacy in Participatory Sensing Systems, *Computer Communications*, 33, pp. 1266–1280, 2010.

[88] B. Hull, V. Bychkovsky, K. Chen, M. Goraczko, A. Miu, E. Shih, Y. Zhang, H. Balakrishnan, S. Madden, CarTel: A Distributed Mobile Sensor Computing System, *ACM SenSys*, 2006.

[89] V. Jakkula, D. Cook, G. Jain. Prediction Models for a Smart Home based Health Care System. *Advanced Information Networking and Applications Workshops*, 2007.

[90] H. Jeung, H. T. Shen, X. Zhou. Convoy queries in spatio-temporal databases. *ICDE Conference*, 2008.

[91] H. Jeung, M. L. Yiu, X. Zhou, C. Jensen, H. Shen, Discovery of Convoys in Trajectory Databases, *VLDB Conference*, 2008.

[92] Y. Jing, S. Baluja. Pagerank for product image search. *WWW Conference*, pages 307–316, 2008.

[93] P. Kalnis, N. Mamoulis, S. Bakiras. On discovering moving clusters in spatio-temporal data. *SSTD*, 2005.

[94] A. Kapadia, N. Triandopoulos, C. Cornelius, D. Peebles, D. Kotz. AnonySense: opportunistic and privacy-preserving context collection, *Proceedings of Sixth International Conference on Pervasive Computing (Pervasive)*, 2007.

[95] H. Kargupta, S. Datta, Q. Wang, K. Sivakumar. On the privacy preserving properties of random data perturbation techniques. In *Proceedings of the IEEE International Conference on Data Mining*, pages 99–106, 2003.

[96] D. Kempe, J. Kleinberg, E. Tardos, Maximizing the Spread of Influence in a Social Network, *ACM KDD Conference*, 2003.

[97] D. Kim, Y. Kim, D. Estrin, M. B. Srivastava. Sensloc: Sensing everyday places and paths using less energy, *ACM Sensys Conference*, 2010.

[98] R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins. Trawling the web for emerging cyber-communities. *WWW Conference*, 1999.

[99] A. Krause, E. Horvitz, A. Kansal, F. Zhao. Toward Community Sensing. *IPSN*, pp. 481–492, 2008.

[100] M. Laibowitz, N.-W. Gong, J. A. Paradiso, Wearable Sensing for Dynamic Management of Dense Ubiquitous Media. *BSN*, 2009.

[101] J. Levandoski, M. Sarwat, A. Eldawy, M. Mokbel. LARS: A Location-Aware Recommender System. *ICDE Conference*, pp. 450–461, 2012.

[102] P. Laube, S. Imfeld. Analyzing relative motion within groups of trackable moving point objects. *GIS*, 2002.

[103] H. K. Le, J. Pasternack, H. Ahmadi, M. Gupta, Y. Sun, T. Abdelzaher, J. Han, D. Roth, B. Szymanski, S. Adali. Apollo: Towards factfinding in participatory sensing. *IPSN Conference*, 2011.

[104] J.-G. Lee, J. Han, K.-Y. Whang, Trajectory Clustering: A Partition-and-Group Framework, *ACM SIGMOD Conference*, 2007.

[105] J.-G. Lee, J. Han, X. Li. Trajectory Outlier Detection: A Partition-and-Detect Framework, *ICDE Conference*, 2008.

[106] J.-G. Lee, J. Han, X. Li, H. Gonzalez. TraClass: trajectory classification using hierarchical region-based and trajectory-based clustering. *PVLDB*, 1(1): pp. 1081–1094, 2008.

[107] V. Lenders, E. Koukoumidis, P. Zhang, M. Martonosi. Location-based trust for mobile user-generated content: applications, challenges and implementations. *Hotmobile*, 2008.

[108] J. Leskovec, K. J. Lang, A. Dasgupta, M. W. Mahoney. Statistical properties of community structure in large social and information networks. *WWW Conference*, 2008.

[109] X. Li, Z. Li, J. Han, J.-G. Lee. Temporal Outlier Detection in Vehicle Traffic Data. *ICDE Conference*, 2009.

[110] X. Li, J. Han, S. Kim, H. Gonzalez. ROAM: Rule- and Motif-Based Anomaly Detection in Massive Moving Object Data Sets. *SDM Conference*, 2007.

[111] X. Li, J. Han, J.-G. Lee, H. Gonzalez. Traffic Density-Based Discovery of Hot Routes in Road Networks. *SSTD*, 2007.

[112] Y. Li, J. Han, J. Yang. Clustering Moving Objects, *ACM KDD Conference*, 2004.

[113] Z. Li, B. Ding, J. Han, R. Kays. Swarm: Mining Relaxed Temporal Moving Object Clusters. *VLDB Conference*, 2010.

[114] Z. Li, B. Ding, J. Han, R. Kays, Mining Hidden Periodic Behaviors for Moving Objects, *ACM KDD Conference*, 2010.

[115] Z. Li, J. Han, M. Ji, L. A. Tang, Y. Yu, B. Ding, J.-G. Lee, R. Kays. MoveMine: Mining moving object data for discovery of animal movement patterns. *ACM TIST*, 2(4): 37, 2011.

[116] Z. Li, C. X. Lin, B. Ding, J. Han. Mining Significant Time Intervals for Relationship Detection. *SSTD Conference*, 2011.

[117] J. Lifton, M. Feldmeier, Y. Ono, C. Lewis, J. A. Paradiso. A platform for ubiquitous sensor deployment in occupational and domestic environments. *IPSN*, 2007.

[118] Z. Liu, J. Xu Yu, Y. Ke, X. Lin, L. Chen. Spotting Significant Changing Subgraphs in Evolving Graphs, *ICDM Conference*, 2008.

[119] B. Longstaff, S. Reddy, D. Estrin. Improving activity classification for health applications on mobile devices using active and semi-supervised learning, *ICST Conference on Pervasive Computing Technologies for Healthcare* , 2010.

[120] L. Luo, A. Kansal, S. Nath, F. Zhao. Sharing and exploring sensor streams over geocentric interfaces, *ACM SIGSPATIAL international conference on Advances in geographic information systems*, 2008.

[121] A. Madan, D. Cebrian, D. Lazer, A. Pentland. Social Sensing for Epidemiological Behavior Change, *Ubicomp*, 2010.

[122] A. Madan, S. Moturu, D. Lazer, A. Pentland. Social Sensing: Obsesity, Healthy Eating and Exercise in Face-to-Face Networks, *Wireless Health*, 2010.

[123] E. Miluzzo, N. Lane, K. Fodor, R. Peterson, S. Eisenman, H. Lu, M. Musolesi, X. Zheng, A. Campbell. Sensing meets mobile social networks: the design, implementation and evaluation of the CenceMe Application, *SenSys*, 2008.

[124] P. Mohan, V. Padmanabhan, R. Ramjee, Nericell: rich monitoring of road and traffic conditions using mobile smartphones, *ACM SenSys*, 2008.

[125] M. Mokbel, C. Chow, W. G. Aref. The New Casper: Query Processing for Location-based Services without Compromising Privacy, *VLDB Conference*, 2006.

[126] A. Narayanan, N. Thiagarajan, M. Lakhani, M. Hamburg, D. Boneh. Location privacy via private proximity testing. *NDSS*, 2011.

[127] S. Nath, J. Liu, F, Zhao. SensorMap for Wide-Area Sensor Webs. *IEEE Computer*, 40(7): 90-93, 2008.

[128] M. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 2006.

[129] J. Paek, J. Kim, R. Govindan. Energy-efficient rate-adaptive gps-based positioning for smartphones, *MobiSys*, 2010.

[130] S. Papadimitrioum F. Li, G. Kollios, P. Yu. Time series compressibility and privacy. In *VLDB Conference*, 2007.

[131] L. Pareschi, D. Riboni, A. Agostini, C. Bettini. Composition and generalization of context data for privacy preservation. *PerComm*, 2008.

[132] N. Pham, T. Abdelzaher, S. Nath. On Bounding Data Stream Privacy in Distributed Cyber-physical Systems, *IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (IEEE SUTC)*, Newport Beach, CA, June, 2010.

[133] N. Pham, R. K. Ganti, Y. S. Uddin, S. Nath, T. Abdelzaher, Privacy preserving Reconstruction of Multidimensional Data Maps in Vehicular Participatory Sensing, *EWSN*, 2010.

[134] R. A. Popa, H. Balakrishnan, A. Blumberg. VPriv: Protecting Privacy in Location-Based Vehicular Services. *USENIX Security Symposium*, 2008.

[135] B. Priyantha, D. Lymberopoulos, J. Liu. Enabling energy-efficient continuous sensing on mobile phones with Littlerock, *Information Processing in Sensor Networks*, 2010.

[136] G. Qi, C. Aggarwal, T. Huang. Community Detection with Edge Content in Social Media Networks, *ICDE Conference*, 2012.

[137] M.-R. Ra, J. Paek, A. B. Sharma, R. Govindan, M. H. Krieger, M. J. Neely. Energy-delay tradeoffs in smartphone applications, *MobiSys*, 2010.

[138] S. Ramchurn, D. Huynh, N. Jennings. Trust in Multi-agent Systems, *Knowledge Engineering Systems*, 19(1), pp. 1–25, 2004.

[139] P. Rashidi, D. Cook. Mining Sensor Streams for Discovering Human Activity Patterns over Time. *ICDM Conference*, 2010.

[140] P. Rashidi, D. Cook, L. Holder, M. Schmitter-Edgecombe. Discovering Activities to Recognize and Track in a Smart Environment, *IEEE TKDE*, 23(4), pp. 527–539, 2011.

[141] V. Rastogi, S. Nath. Differentially Private Aggregation of Distributed Time-Series with Transformation and Encryption, *ACM SIGMOD Conference*, 2010.

[142] S. Reddy, K. Shilton, G. Denisov, C. Cenizal, D. Estrin, M. B. Srivastava. Biketastic: sensing and mapping for better biking. *CHI*, pp. 1817–1820, 2010.

[143] S. Reddy, D. Estrin, M. B. Srivastava. Recruitment Framework for Participatory Sensing Data Collections.*Pervasive*, pp. 138–155, 2010.

[144] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, M. Srivastava. Using mobile phones to determine transportation modes, *ACM Transactions on Sensor Networks*, 6(3), pp. 1–27, 2010.

[145] M. Romaine, J. Richardson. State of the Translation Industry. *Translation Industry Report*, My Gengo, 2009.

[146] S. Sehgal, S.S. Kanhere, C.T. Chou. Mobishop: using mobile phones for sharing consumer pricing information, (Demo paper), *IEEE DCOSS*, 2008.

[147] S. Shang, R. Ding, B. Yuan, K. Xie, K. Zheng, P. Kalnis. User Oriented Trajectory Search for Trip Recommendation, *EDBT Conference*, 2012.

[148] K. Shilton, D. Estrin, R. Govindan, J. Kang. Designing the Personal Data Stream: Enabling Participatory Privacy in Mobile Personal Sensing. In *Research Conference on Comunication, Information, and Internet Policy (TPRC)*, 2009.

[149] L. Stenneth, P. S. Yu, O. Wolfson. Mobile Systems Location Privacy: "MobiPriv" A Robust k-anonymous System. *IEEE International Conference on Wireless and Mobile Computing, Networking and Communications*, 2010.

[150] M. Sung, C. Marci, A. Pentland. Wearable Feedback Systems for Rehabilitation, *Journal of Neuroengineering and Rehabilitation*, 2:17, 2005.

[151] J. Sun, S. Papadimitriou, P. Yu, C. Faloutsos. Graphscope: Parameter-free Mining of Large Time-Evolving Graphs, *KDD Conference*, 2007.

[152] K.P. Tang, J. Fogarty, P. Keyani, J.I. Hong. Putting people in their place: an anonymous and privacy-sensitive approach to collecting sensed data in location-based applications, *SIGCHI*, 2006.

[153] L. Tang, X. Yu, S. Kim, J. Han, C. Hung, W Peng. Tru-Alarm: Trustworthiness Analysis of Sensor Networks in Cyber-Physical System. *ICDM Conference*, 2010.

[154] A. Thiagarajan, L. Ravindranath, H. Balakrishnan, S. Madden, L. Girod. Accurate, Low-Energy Trajectory Mapping for Mobile Devices. *Proceedings of the NSDI*, 2011.

[155] A. Thiagarajan, L. Ravindranath, K. LaCurts, S. Toledo, J. Eriksson, S. Madden, H. Balakrishnan. VTrack: Accurate, Energy-Aware Road Traffic Delay Estimation Using Mobile Phones. *ACM SenSys*, 2009.

[156] H. Tong, S. Papadimitriou, P. Yu, C. Faloutsos. Proximity-Tracking on Time-Evolving Bipartite Graphs, *SDM Conference*, 2008.

[157] J. Tsai, P. Kelley, L. Cranor, N. Sadeh. Location sharing technologies: Privacy risks and controls. I/S: A *Journal of Law and Policy for the Information Society*, 6(2), 2010.

[158] Z. Yang, S. Zhong, R. N. Wright. Privacy-preserving classification without loss of accuracy. In *Proceedings of the Fifth SIAM International Conference on Data Mining*, pages 92–102, 2005.

[159] D. Wang, T, Abdelzaher, H. Ahmadi, J. Pasternack, D. Roth, M. Gupta, J. Han, O. Fatemieh, H. Le, C. Aggarwal. On Bayesian Interpretation of Fact-finding in Information Networks. *Fusion*, 2011.

[160] D. Wang, T. Abdelzaher, L. Kaplan, C. Aggarwal. On Quantifying the Accuracy of Maximum Likelihood Estimation of Participant Reliability in Social Sensing. *DMSN*, 2011.

[161] D. Wang, L. Kaplan, H. Le, T. F. Abdelzaher. On truth discovery in social sensing: a maximum likelihood estimation approach. *IPSN Conference*, 2012.

[162] D. Wang, T. Abdelzaher, L. Kaplan, C. Aggarwal. On Scalability and Robustness Limitations of Real and Asymptotic Confidence Bounds in Social Sensing, *SECON Conference*, 2012.

[163] M. Wirz, D. Roggen, G. Troster, Decentralized detection of group formations from wearable acceleration sensors, in *Proceedings IEEE SocialCom*, 2009

[164] D. Wyatt, T. Choudhury, J. Bilmes. Creating Social Network Models from Sensor Data, *NIPS Network Workshop*, 2007.

[165] D. Wyatt, T. Choudhury, J. Bilmes. Conversation Detection and Speaker Segmentation in Privacy Sensitive Situated Speech Data. *Proceedings of Interspeech*, 2007.

[166] D. Wyatt, T. Choudhury, H. Kautz. Capturing Spontaneous Conversation and Social Dynamics: A Privacy-Sensitive Data Collection Effort. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[167] D. Wyatt, T. Choudhury, J. Bilmes. Learning Hidden Curved Exponential Random Graph Models to Infer Face-to-Face Interaction Networks from Situated Speech Data. *Proceedings of AAAI*, 2008.

[168] T. Yan, V. Kumar, D. Ganesan. Crowdsearch: Exploiting crowds for accurate real-time image search on mobile phones, *MobiSys*, 2010.

[169] H. Yang, S. Chen, M. Lyu, I. King. Location-based Topic Evolution, *1st International Workshop on Mobile Location-based Services*, 2011. http://dl.acm.org/citation.cfm?id=2025894

[170] X. Yin, J. Han, P. S. Yu. Truth discovery with multiple conflicting information providers on the web. *IEEE TKDE*, 2008.

[171] Z. Yin, L. Cao, J. Han, J. Luo, T. Huang. Diversified Trajectory Pattern Ranking in Geo-tagged Social Media. *SDM Conference*, 2011.

[172] Z. Yin, L. Cao, J. Han, C. Zhai, T. Huang. Geographical topic discovery and comparison. *WWW Conference*, 2011.

[173] X. Yu, A. Pan, L. A. Tang, Z. Li, J. Han. Geo-Friends Recommendation in GPS-based Cyber-physical Social Network. *ASONAM Conference*, 2011.

[174] D. Zeinalipour-Yazti, C. Laoudias, M. Andreou, D. Gunopulos. Disclosure-Free GPS Trace Search in Smartphone Networks. *Mobile Data Management*, 2011.

[175] T. Zhang, A. Popescul, B. Dom. Linear prediction models with graph regularization for web-page categorization. In *KDD*, pages 821–826, 2006.

[176] N. Zhang, S. Wang, W. Zhao. A new scheme on privacy-preserving data classification. In *ACM KDD Conference*, pages 374–383, 2005.

[177] F. Zhao, L. Guibas. *Wireless Sensor Networks: An Information Processing Approach*, Morgan Kaufmann, 2004.

[178] Y. Zhang, B. Hull, H. Balakrishnan, S. Madden. IceDB: Continuous Query Processing in an Intermittently Connected World. *ICDE Conference*, 2007.

[179] V. W. Zheng, Y. Zheng, X. Xie, Q. Yang. Collaborative location and activity recommendations with gps history data. *WWW Conference*, 2010.

[180] Y. Zheng, L. Liu, L. Wang, X. Xie. Learning transportation mode from raw gps data for geographic applications on the web, *WWW Conference*, 2008.

[181] Y. Zheng, L. Zhang, X. Xie, W.-Y. Ma. Mining interesting locations and travel sequences from GPS trajectories. *WWW Conference*, 2009.

[182] Y. Zhou, H. Cheng, J. X. Yu. Graph clustering based on structural/attribute similarities. *PVLDB*, 2(1): pp. 718–729, 2009.

[183] Z. Zhuang, K.-H. Kim, J. P. Singh. Improving energy efficiency of location sensing on smartphones, *MobiSys*, 2010.

[184] http://latitude.google.com

[185] http://www.citysense.com

[186] http://www.movebank.org

[187] http://www-01.ibm.com/software/data/infosphere/streams/

[188] http://www.navizon.com

[189] http://ilocalis.com

[190] http://www.trapster.com

[191] http;//www.wikitude.com

[192] https://www.mturk.com/mturk/welcome

[193] http://code.google.com/p/zxing

[194] http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/

[195] https://www.sensenetworks.com/products/macrosense-technology-platform/

[196] Vivometrics, http://www.vivometrics.com

[197] SenseWear, BodyMedia Corporation http://www.bodymedia.com

[198] Mini Logger, Minimitter Corporation http://www.minimitter.com

[199] Yelp Business Review Site http://www.yelp.com