Proceedings of the 23rd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, M. Ali, M. Gertz, Y. Huang, M. Renz, and J. Sankaranarayanan, eds., Seattle, WA, November 2015.

An Online Marketplace for Geosocial Data

Yaron Kanza
Jacobs Institute, Cornell Tech
kanza@jacobs.cornell.edu

Hanan Samet
University of Maryland
hjs@cs.umd.edu

ABSTRACT

When recording their GPS trajectories or posting geo-tagged content on social networks, people produce social spatio-temporal data that can be stored and shared, namely *geosocial data*. Much of these spatio-temporal data can be used by organizations and applications, for statistical analysis or to provide services that are based on data. By letting people sell the data they produce, to different consumers, both sides can benefit. Thus, we present here a visionary idea of a *geosocial marketplace* where people and organizations can sell, buy and exchange geosocial data, that is, trade with spatio-temporal data pertaining people. We discuss the involved challenges, such as how to define supply and demand, pricing data, privacy issues and measuring the amount of data being exchanged. We explain the importance of the approach and its applicability. We believe that the proposed vision could motivate followup research in the area of sharing and exchanging spatio-temporal data as well as determining appropriate price points.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Spatial databases and GIS*

Keywords

Geo-social; socio-spatial; marketplace; online; exchange; privacy; interoperability; personal data

1. INTRODUCTION

Smartphones and other types of GPS-enabled devices have made location data prevalent and the browsing of location data [13,30] a popular activity. This has been further fueled by the ease of recording location data. People can effortlessly collect data about where they are at different times, and share this data with others, e.g., by sharing GPS trajectories or by *checking-in* at a place, in a location-based social network. By aggregating location data of different people, applications create repositories of *geosocial data*, that is, datasets that consist of triples (l, t, u) of location l, time t and user u, specifying that the user u visited location l at time t. A geosocial dataset combines the spatio-temporal aspect with the social aspect, as data is produced by users, to reflect user activities, and it can be shared, similarly to data in online social networks. Such data can be stored locally on the device or on remote servers.

SIGSPATIAL '15, November 03-06, 2015, Bellevue, WA, USA
ACM 978-1-4503-3967-4/15/11 ...$15.00.
DOI: http://dx.doi.org/10.1145/2820783.2820881.

By using a smartphone, people can easily collect and share their location data, but it comes with a cost. First, the recording of location data requires the use of GPS, which is a large energy consumer, and this necessitates charging the smartphone more frequently, which is inconvenient. Secondly, storing the data consumes some of the storage space of the device and transmitting the data requires communication bandwidth [14]. Thirdly, the location data may reveal private information, regarding places the user visited, see [5, 21]. Thus, people should be given an incentive to record and share their location data, e.g., money.

While there is a cost to collecting and sharing location data, frequently there is also value in such data, for both organizations and other users. That is, companies and applications can utilize location data of specific individuals or of crowds, either in real time or by using historical data. The following are examples of applications that can make use of such data.

- *Recommendation systems*: By using information about the places users visit, recommendation systems can learn about the preferences and the spatial constraints of users, and provide recommendations accordingly [6].
- *Monitoring traffic*: Real-time spatio-temporal data of people's location can be used for monitoring the traffic condition and recommending travel routes to people based on real traffic data, e.g., as in [17, 25] instead of just a shortest path [20, 32].
- *Urban planning:* Location data can be used to know where people are in the city and how they move in the city, to plan the city better [4, 19, 28].
- *Public health:* Location data of crowds can be used to monitor the spread of infectious diseases and improve the utilization of public-health services [23, 24, 33].
- *Tracking celebrities:* Location data of celebrities can be of interest to some people, and in some cases can be regarded as news people may want to browse [26,29,31] while distinguishing between past and future [18], and reliability [15].

These are some examples of the many cases where organizations and applications need geosocial data. However,

Figure 1: Users upload location data to the marketplace while applications and organizations can buy location data, according to their needs and budget.

frequently it is hard to collect high-quality geosocial data. Thus, it would be useful for organizations and for service providers to be able to buy high-quality location data from people. This requires having a *marketplace* in which people could offer the data, with specified limitations, and buyers could buy data, according to their needs (Fig. 1). The incentive for sharing location data can be money, as in Amazon Mechanical Turk (`www.mturk.com`), where people receive money for executing small tasks. The incentive may also be a variant of reward points that grant access to a service, or provide some other benefit. Data may also be exchanged for other data. However, when collecting data from a crowd, exchange is problematic because most people have no need for data of other people and thus there is a need for a marketplace where people could sell and buy data.

Recently there has been a growing interest in personal-data marketplaces, and a few companies have been established to satisfy this need. However, they do not focus on the specific demands of a marketplace for spatio-temporal data. For example, a user, say Alice, may only be interested in recent location data of people traveling on the roads between her home and her office, to estimate the traffic condition on these roads. Another example is of an urban planning task where the organization only needs information about the location of people who live in a certain neighborhood. The challenge is to allow sellers and buyers to specify their needs effectively and put the appropriate price tag on the data. In this paper we elaborate on these questions.

2. RELATED WORK

The need to collect, query and use geosocial data has been recognized in recent years and there is a growing interest in utilizing such data. Querying geosocial data [7,9,11] and using data to understand user preferences, in recommendation systems [3,8,10] were studied. Elbery et al. [12] studied the use of geosocial data for carpooling. Pat et al. [27] showed how geosocial data can be used to support geographic search. A marketplace of geosocial data could assist in acquiring data for such applications.

About two decades ago, the problem of how to exchange spatial data and share the data between different organizations was studied [1,2,16]. However, they focused on issues of interoperability and the need to integrate the data from heterogeneous sources. This was before the advent of smartphones and at a time when most people could not record their location history or easily share their data with others. The interoperability problems they faced can be easily solved nowadays by utilizing common exchange formats such as GeoJSON (see `http://geojson.org/`) or OSM XML, of the OpenStreetMap project.

3. FRAMEWORK

We now present our proposed framework. We consider location data gathered using a smartphone or a GPS-enabled wearable device. Therefore, data can be collected almost at all times and can be transmitted to remote servers. A *location data item* is one of the following.

- A GPS location (l, t, u) where l is a location, t is the time and u is a user identifier, specifying that user u was in location l at time t, e.g., a *check-in* in a location-based social network. An *anonymous* GPS measure (l, t) is the same, except that the user is unknown.
- A trajectory as a sequence $[(l_1, t_1, u), \ldots, (l_n, t_n, u)]$ of GPS locations of some user, sorted according to the measure time.
- A geo-tagged post (l, t, u, C) where (l, t, u) is a GPS location and C is the content (text, picture, etc.), e.g., a geo-tagged tweet or post in Twitter or Instagram.

A location dataset D is a set of location data items.

Data Sharing. *Sharing* of data is the transfer of a location dataset from one user to another user. A sharing constraint is a condition φ on the times, the locations or the context of data items, *e.g.*, limiting the locations to a certain area or limiting the times to certain hours during the day. A *constrained sharing* is a pair (D, φ) where D is a location dataset, φ is a sharing constraint, and all the data items in D satisfy φ. This is needed so that people could control the sharing of their location data.

Data Measures. When trading data, it is important to measure the amount of data or information being shared. Since there is no common way to do so, we introduce here three different ways to measure the amount of location data, in a trade. The first measure is the *point count* of a dataset D. It measures the number of GPS points in D, providing a very simple and direct way to measure the amount of data being transferred. Note that it is affected by the frequency of the GPS readings.

The second measure is the *spatio-temporal volume* of a dataset D. To measure it, suppose that each GPS point a has an area which is the disk of radius r (e.g., $r = 5$ meters) centered at a. Suppose that a has a time gap Δt which starts t seconds before its reading and ends t seconds after the reading (e.g., $t = 10$ seconds). The area of a dataset D is the union of the areas of the points in D, and the total time of D is the union of the time gaps of the points of D, both computed while considering overlapping areas and overlapping times only once. The volume of D is the product of the area of D and the total time. Therefore, the volume is a function of the area and the time *spanned by* D.

The third measure refers to the amount of information gained by the dataset. To that end, we consider the *information entropy* of D with respect to a given partition of the space, say using a grid index G, and partition of the day into time slices, e.g., partitioning the day into 24 parts by considering each hour as a slice. For each place (grid cell) x in G and time t, the probability $p(x)$ is the ratio of the number of

points of D with location in x and time t to all the points of D. That is, $p(x)$ is the probability of the event of selecting, randomly, a point from D that is located in x and has time t. The entropy of D is defined as $\sum_{x \in G} p(x) \log(\frac{1}{p(x)})$. Intuitively, the information entropy indicates how many bits are needed to represent the information in D in a compressed form. So, a user with predictable habits will produce less information than a user who visits many different places in a chaotic manner.

Note that the information entropy can also be used to measure the added information when joining a new dataset D to an existing dataset $\mathfrak{DE}$. This can be done by using the same formula for entropy while the probability function p is computed based on $\mathfrak{DE}$. That is, the probability of a point a in cell x and time t is the probability that a uniformly selected point of $\mathfrak{DE}$ has a location in x and time t. The entropy is computed using the same formula as before and the probability function that is based on $\mathfrak{DE}$. The added information can also be measured using the commonly-used Kullback-Leibler divergence [22].

4. MARKETPLACE

A marketplace should allow sellers to present their wares and should provide buyers with means to define what they need and search for it. The marketplace should also allow some pricing or negotiation mechanism. Next, we present the different components and modules that an online geosocial marketplace may include.

4.1 Sellers Module

Sellers should install an app on their device. The app will collect the spatio-temporal data and will upload to the marketplace location data that comply with the seller's constraints. The app will encode the transmitted data so that users would not send fake data of locations they have not truly visited.

The goal of the seller module is to allow sellers to define the data they are willing to sell. For example, a user may only be willing to share location data collected by her phone between 10 am and 8 pm at a radius of 5 miles from her home. This can easily be done by applying a sharing constraint and discarding from the dataset all the measures that do not satisfy it. For example, the following query can be used by a seller to offer data collected in a radius of 5 miles from a specified location, between 10 am and 5 pm, along with the demographic attributes of age and gender.

```
PUBLISH location, time, age, gender
WHERE distance(location, (40.741, -74.002)) < 5
      AND time BETWEEN 10 am AND 5 pm
```

Note that a GUI can be developed, to facilitate the specification of a publishing command.

Since location data can be sensitive information, the module should include an alert or filtering function to prevent sharing places the user never shared before (e.g., places that are only seldom visited) or places that are not public, i.e., locations that very few people visit, based on the data that the marketplace application receives, unless a specific authorization is provided.

4.2 Buyers Module

Buyers should request data that comply with specifications regrading the *(1)* area of interest, i.e., only data items in the specified area should be considered; *(2)* time frame, i.e., only data items within the specified time frame should be considered; *(3)* quantity and continuity, where the buyer may only be interested in data that were collected continuously or in datasets whose size exceeds certain limits. In some cases, buyers may need demographic details of sellers and will only be interested in purchasing data from sellers who are willing to provide their demographic details.

For example, the following acquiring command specifies that the buyer wants to acquire data collected in the area of downtown Manhattan (a 2 mile radius from the World Trade Center) by people who are older than 25 and who gathered the data continuously for at least 2 hours, with at least one measure every 5 minutes.

```
ACQUIRE location, time, age, gender
WHERE distance(location, (40.712, -74.013)) < 2
      AND age > 25
      AND collection(continuous, 2 h, 5 min)
```

Additional functions constraining the location, the time and the data-collection process could be defined, to specify demands. For example, the buyer may only need data from urban places, data pertaining specific roads, or location data of people who frequently dine at a restaurant.

4.3 Browsing and Searching

Before acquiring data, buyers may want to browse the available data, to see if the data are suitable for them, or to check if data with specific attributes exist in the repository. Potential sellers may want to know what is the demand for data they produce. This will require search capabilities, to support search queries over datasets, such as the following.

```
SEARCH datasets
WHERE distance(location, (40.712, -74.013)) < 2
      AND time BETWEEN 10 am AND 2 pm
```

The result of a search may be a specimen depicted on a map. When searching for sellers, the search should specify properties of sellers and of the area containing the data.

```
SEARCH sellers
WHERE distance(location, (40.712, -74.013)) < 2
      AND age < 18
```

The challenge is to effectively answer such search queries, given that they refer to datasets or attributes of sellers and not to single data items.

4.4 Pricing Mechanism

One of the most challenging tasks in the development of a marketplace is how to effectively match buyers and sellers and how to create a mediator that determines the price of the data. A naive approach is to set the price according to the amount of data, e.g., based on the measures discussed in Section 3. However, this does not take into account geographical and social heterogeneity, where there are places for which data is scarce or where the demand is high, in comparison to other places. Also, there may be populations for which data is deficient, and this should affect the value of data collected by a person who belongs to such a population.

To value geosocial data, while taking heterogeneity into account, the system should estimate matches of a buyer and a seller pertaining a certain area. Let a *potential match* (b, s, a) be a triple of buyer b, seller s and area a. This potential match can be compared to all the other triples of

buyer, seller and area. For such a comparison, the system should *(i)* rank places according to the number of people who share data about them, *(ii)* rank places according to the demand for data about them, *(iii)* rank sellers according to the size of the population with their demographic attributes, and *(iv)* rank sellers according to the demand for data from people having their attributes. In these rankings, higher supply (more potential sellers) reduces the score and higher demand increases the score. The above four ranking scores are combined using some monotone function, say addition. The value of each potential match (b, s, a) is given relatively to the other potential matches based on the combined rank. To each data request by buyer b' regarding area a', the system could match the request to the sellers S such that $\{(b', s, a') \mid s \in S\}$ are the triples with the lowest price among all the triples of buyer b' and area a'.

The system could also support a simpler approach where buyers specify how much they are willing to pay for the data and sellers who comply with the required conditions receive a notification. Then, if the sellers agree to the offered price, they may share the data with the buyer. The first suitable sellers who agree to the price get to sell their data, until the buyer gets the amount of data she desires. When the buyer does not get enough data, she can offer a higher price.

5. CONCLUSION

We presented a vision of a geosocial marketplace that facilitates the sharing and selling of location data of people in a controlled way, while taking into account privacy issues. The marketplace should allow organizations to buy data in a managed way, based on specifications of their needs. The tasks of specifying needs and constraints, matching sellers and buyers, price negotiation and privacy control are discussed. Such a marketplace will assist organizations and researchers to acquire spatio-temporal data of individuals and crowds to support their research and applications. How to effectively regulate the marketplace is an open question.

6. ACKNOWLEDGMENTS

The work of Yaron Kanza was supported in part by the Israel Ministry of Science and Technology (Grant 3-9617) and by the Israel Science Foundation (Grant 1467/13). The work of Hanan Samet was supported in part by the National Science Foundation (Grants IIS-12-19023 and IIS-13-20791).

7. REFERENCES

[1] D. Abel. Spatial internet marketplaces: A grand challenge? In *Advances in Spatial Databases*, pages 1–8. Springer Berlin, Germany, 1997.

[2] D. J. Abel, V. J. Gaede, K. L. Taylor, and X. Zhou. SMART: Towards Spatial Internet Marketplaces. *GeoInformatica*, 3(2):141–164, 1999.

[3] J. Bao, Y. Zheng, and M. F. Mokbel. Location-based and preference-aware recommendation using sparse geo-social networking data. In *SIGSPATIAL '12*, pages 199–208, Redondo Beach, CA, 2012.

[4] R. A. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing*, 10(4):18–26, 2011.

[5] A. R. Beresford and F. Stajano. Location privacy in pervasive computing. *IEEE Pervasive computing*, 2(1):46–55, 2003.

[6] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez. Recommender systems survey. *Knowledge-Based Systems*, 46:109–132, 2013.

[7] A. Croitoru, A. Crooks, J. Radzikowski, and A. Stefanidis. Geosocial gauge: A system prototype for knowledge discovery from social media. *IJGIS*, 27(12):2483–2508, 2013.

[8] V. de Graaff, M. van Keulen, and R. A. de By. Towards geosocial recommender systems. In *WI&C '12*, pages 8:1–8:4, Lyon, France, 2012.

[9] Y. Doytsher, B. Galon, and Y. Kanza. Querying geo-social data by bridging spatial networks and social networks. In *LBSN '10*, pages 39–46, San Jose, CA, 2010.

[10] Y. Doytsher, B. Galon, and Y. Kanza. Storing routes in socio-spatial networks and supporting social-based route recommendation. In *LBSN '11*, pages 49–56, Chicago, 2011.

[11] Y. Doytsher, B. Galon, and Y. Kanza. Querying socio-spatial networks on the world-wide web. In *WWW '12*, pages 329–332, Lyon, France, 2012.

[12] A. Elbery, M. ElNainay, F. Chen, C.-T. Lu, and J. Kendall. A carpooling recommendation system based on social vanet and geo-social data. In *SIGSPATIAL'13*, pages 556–559, Orlando, FL, 2013.

[13] C. Esperança and H. Samet. Experience with SAND/Tcl: a scripting tool for spatial databases. *JVLC*, 13(2):229–255, 2002.

[14] R. Gotsman and Y. Kanza. A dilution-matching-encoding compaction of trajectories over road networks. *GeoInformatica*, 19(2):331–364, 2015.

[15] N. Gramsky and H. Samet. Seeder finder - identifying additional needles in the Twitter haystack. In *LBSN'13*, pages 44–53, Orlando, FL, 2013.

[16] O. Günther and R. Müller. From GISystems to GIServices: Spatial Computing on the Internet Marketplace. In *Interoperating Geographic Information Systems*, pages 427–442. Springer, 1999.

[17] I. Hefez, Y. Kanza, and R. Levin. Tarsius: A system for traffic-aware route search under conditions of uncertainty. In *SIGSPATIAL'11*, pages 517–520, Chicago, 2011.

[18] A. Jackoway, H. Samet, and J. Sankaranarayanan. Identification of live news events using Twitter. In *LBSN'11*, pages 25–32, Chicago, 2011.

[19] Y. Kanza, E. Kravi, and U. Motchan. City nexus: Discovering pairs of jointly-visited locations based on geo-tagged posts in social networks. In *SIGSPATIAL'14*, pages 597–600, Dallas, TX, 2014.

[20] Y. Kanza, E. Safra, Y. Sagiv, and Y. Doytsher. Heuristic algorithms for route-search queries over geographical data. In *SIGSPATIAL'08*, pages 11:1–11:10, Irvine, CA, 2008.

[21] J. Krumm. A survey of computational location privacy. *Personal and Ubiquitous Computing*, 13(6):391–399, 2009.

[22] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 03 1951.

[23] R. Lan, M. D. Adelfio, and H. Samet. Spatio-temporal disease tracking using news articles. In *HealthGIS'14*, pages 31–38, Dallas, TX, 2014.

[24] R. Lan, M. D. Lieberman, and H. Samet. The picture of health: map-based, collaborative spatio-temporal disease tracking. In *HealthGIS'12*, pages 27–35, Redondo Beach, CA, 2012.

[25] R. Levin and Y. Kanza. Tars: Traffic-aware route search. *Geoinformatica*, 18(3):461–500, 2014.

[26] M. D. Lieberman and H. Samet. Supporting rapid processing and interactive map-based exploration of streaming news. In *SIGSPATIAL'12*, pages 179–188, Redondo Beach, CA, 2012.

[27] B. Pat, Y. Kanza, and M. Naaman. Geosocial search: Finding places based on geotagged social-media posts. In *WWW '15*, pages 231–234, Florence, Italy, 2015.

[28] C. Ratti, S. Williams, D. Frenchman, and R. Pulselli. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B Planning and Design*, 33(5):727, 2006.

[29] H. Samet, M. D. Adelfio, B. C. Fruin, M. D. Lieberman, and B. E. Teitler. Porting a web-based mapping application to a smartphone app. In *SIGSPATIAL'11*, pages 525–528, Chicago, 2011.

[30] H. Samet, H. Alborzi, F. Brabec, C. Esperança, G. R. Hjaltason, F. Morgan, and E. Tanin. Use of the SAND spatial browser for digital government applications. *CACM*, 46(1):63–66, 2003.

[31] H. Samet, J. Sankaranarayanan, M. D. Lieberman, M. D. Adelfio, B. C. Fruin, J. M. Lotkowski, D. Panozzo, J. Sperling, and B. E. Teitler. Reading news with maps by exploiting spatial synonyms. *CACM*, 57(10), 2014.

[32] J. Sankaranarayanan, H. Samet, and H. Alborzi. Path oracles for spatial networks. *PVLDB*, 2(1):1210–1221, 2009.

[33] L. A. Waller and C. A. Gotway. *Applied Spatial Statistics for Public Health Data*, volume 368. John Wiley & Sons, Hoboken, New Jersey, 2004.