2014 IEEE Fourth International Conference on Big Data and Cloud Computing

Participatory Sensor Networks as Sensing Layers

Thiago H. Silva*, Pedro O. S. Vaz de Melo*, Jussara M. Almeida*, Aline C. Viana◇ Juliana Salles†, Antonio A. F. Loureiro*
*Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
◇INRIA, France
†Microsoft Research, Redmond, WA, USA
{thiagohs, olmo, jussara, loureiro}@dcc.ufmg.br, aline.viana@inria.fr, jsalles@microsoft.com

***Abstract*—Participatory sensor networks (PSNs) regards smartphone users as consumers as well as active producers of data. A sensing layer represents a type of data, coming from a given source of data, such as web services, traditional wireless sensor networks, and PSNs. In this work, we show the usefulness and potential of having sensing layers in PSNs. We also show how we can formalize the concept of sensing layers in participatory sensor networks. Furthermore, we demonstrate how to derive and create new applications and services that are not promptly available from different sensing layers, opening up very interesting research opportunities.**

I. Introduction

In the last years, there has been a dramatic evolution of mobile phones that has trigged new ways of how we use these devices. From simple phone calls in the past, today the mobile phone is the main source of information storing our favorite songs and videos. Participatory sensor networks regards mobile phone users not only as consumers, but also as active producers of data using sensors attached to mobile phones, enabling a new and powerful source of data.

In this scenario, a sensing layer represents data, with the corresponding attributes, from a given source of data. The data represented by sensing layers have to come from a source that can be considered a sensor. Examples of data sources are web services, such as weather condition provided by "The Weather Channel"[1]; traditional wireless sensor networks; income census; and participatory sensor networks (PSN) [1]. In these examples the sensors are web service of The Weather Channel; physical sensors in a WSN; census of a city; and user & mobile devices in a PSN. In this context, the applications or organizations provide a data stream, with very different throughput. The census sensing, for instance, may be slow, e.g., data sharing every four years. These examples help to illustrate the ubiquity and diversity of data that may be available. This universe of "ubiquitous data" may be complex to understand and work with, opening up good opportunities for research studies.

Given that, we could have, for example, four different layers for a city: **Traffic alerts** layer provides traffic conditions in certain locations, such as traffic jam or accident (obtained, for example, from Waze or Bing Maps); **Check-ins** layer provides category of a certain place, such as school or pub (obtained for example, from Foursquare[2]); **Weather condition** layer provides climate conditions observed in a certain location, such as windy or rainy (obtained, for example, from Weddar[3] or The Weather Channel); and **Pictures of places** layer provides photos of a certain place, such as a monument (obtained, for example, from Instagram[4]). Each observation (at each layer) has the following attributes associated with it: time (when the observation occurred), space (geographic location), contributor sensor (e.g., user u) and specific data from a layer (*specialty data*). Note that this description illustrates two types of sensors: companies providing data through web services; and users sharing real-time data with their portable devices. Other layers could be obtained by other types of source of data, such as traffic condition provided by Bing Maps, census data, or even be derived from one or more layers, as will be exemplified latter.

We discuss the concept of sensing layers for participatory sensor networks, most of the time, because this is an emerging source of data with important characteristics, such as (near) real time and very large scalability [1]. Due to these special characteristics, the use of PSNs as sensing layers simultaneously with other layers, even derived from other sources, may bring new information about city dynamics and urban social behavior, which could enable the design of more sophisticate services (as discussed later). All the concepts discussed in this study can be used for other data sources associated to a predefined geographical region.

The main goal of this work is to formalize the concept of sensing layers in participatory sensor networks, and show how we can design sophisticated applications from different sensing layers. We demonstrate an application for identification of sights and another to perform economic-cultural analysis of regions, opening up very interesting research opportunities.

This work is organized as follows. Section II discusses the different aspects of sensing layers. Section III presents the related work. Section IV describes how we can process different sensing layers. Section V presents the design of applications based on sensing layers. Finally, Section VI presents the final remarks and future work.

II. Sensing Layers

A sensing layer consists of data describing specific aspects of a geographical location. As shown in Figure 1 by a box labeled "big raw data", these data should be collected (e.g., using an API) and processed, which also includes analysis and data standardization. The last step is the data storage. These steps do not include the extraction of context (or knowledge) from the obtained data, but organize them [2]. However, data of sensing layers could be used for context inference, generating new information.

To illustrate these processes, consider data from a PSN derived from Foursquare. In Foursquare, users can, among other activities, perform check-ins at locations and leave tips

[1]http://www.weather.com.
[2]http://www.foursquare.com.
[3]http://www.weddar.com.
[4]http://www.instagram.com.

978-1-4799-6719-3/14 $31.00 © 2014 IEEE
DOI 10.1109/BDCloud.2014.27

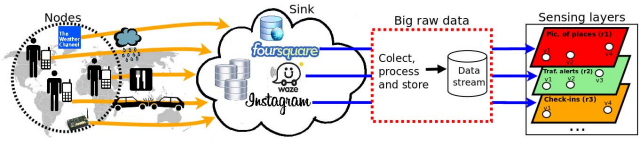

Fig. 1. Overview of participatory sensor networks with the concept of sensing layers.

on visited places. From these data, we can define at least two layers, namely: check-ins, containing the check-ins performed by users (check-ins can be used to discover popular places, for example), and tips of places, containing tips, such as "this restaurant has amazing food", provided by users about certain places. The creation of layers, as shown in Figure 1, depends on specific operations for each system. In the case of Foursquare check-ins, a possible way to get them is through Twitter. This means that we have to collect, analyze, and process tweets. The coding of tools to perform those steps varies according to the system or application. Next, we must define a structure to represent and store the data of interest associated with a given place where it was shared, thus representing a layer. Each data in a layer has the following attributes:

- t: time interval when the data was created;
- a: location (e.g, GPS coordinates, neighborhood area) where the data was generated. We represent all locations by an area. Even if the data is referred by a GPS coordinate it is error prone. For this reason it is interesting to consider an area for this point, for example, a circle with radius x, from the GPS coordinate;
- s: specialty data;
- u: one or more IDs of user(s) who generated the data;

Each layer has also a variable h, which indicates the status of the layer, where $h = 0|1$, representing the inactive and active states, respectively. The list below represents some examples of layers that are currently available:

1) check-ins (example of source: Foursquare);
2) tips of locations (example of source: Foursquare);
3) traffic alerts (example of source: Waze);
4) pictures of places (example of source: Instagram);
5) average income per area (example of source: census);
6) weather condition (example of source: The Weather Channel);
7) noise level (example of source: NoiseTube.net).

A. Usefulness of Sensing Layers

The processing of a set of sensing layers may enable a large-scale study of each monitored aspect in (near) real time, and provides historical data on patterns observed over long periods. Sensing layers can be applied to several contexts of urban computing, for example, helping to better understand the dynamics of cities and urban behavior in different regions of the world, and respond quickly to unexpected changes.

The use of sensing layers currently in the literature is commonly performed independently, i.e., there is no joint analysis. The individual use of a sensing layer can still be very useful. For instance, using a sensing layer containing traffic information may enable real-time identification of highways with accidents and potholes, whose detection is difficult with traditional sensors, but it becomes more feasible when users participate in the sensing process. Such detection opens opportunities for various services, such as assist smart cars in the correct identification of problems on the road.

Despite the usefulness of using single layers only, services based on just one layer might lack of complementary data. For example, Google Flu Trends[5], a service based on Google queries, is a type of sensing layer. Very recently, a group of respected social scientists reported that Google Flu Trends not only wildly overestimated the number of flu cases in the U.S. in the 2012-13 flu season, but has also consistently overshot in the last few years [3]. According to them, the problem might be because Google Flu Trends is not using complementary information in their service. Indeed, the analysis reported in [3] shows that combining Google Flu Trends with CDC[6] data may work best. Seems that the way to save this interesting service is using multiple layers, even Matt Mohebbi, co-inventor of Google Flu Trends, agrees with that [4].

The joint analysis of multiple sensing layers can also be extremely useful in building new applications. For example, we know that a common complaint of inhabitants of large cities is traffic jam. From this, an application that naturally emerges is one that has the goal of inferring the causes of jam, an essential step for addressing the problem. This is not an easy task to accomplish, and the result may vary from place to place. However, the joint analysis of different sensing layers of the city could contribute to build a more robust application. For example, we could cross-check information provided by the following layers: traffic alerts, derived from Waze; check-ins, derived from Foursquare; and pictures of places, derived from Instagram. The first layer provides near real-time data about where traffic jams are occurring. The second one provides data about types of places located in the areas of jams. Having that, it is possible to better understand the areas of interest (for example, identifying a commercial area). Finally, by analyzing the picture of places layer, we can get visual evidence of what is happening in almost real time near the areas of jams. When analyzing data from these three layers together, we can detect, for example, cars blocking intersections, and infer the possible causes of them. Obviously, other layers may also be used, such as the weather condition, layer derived from systems such as Weddar or other traffic condition layer provide, for instance, by Bing Maps[7].

B. A Formal Model for Sensing Layers

Let $U = \{u_1, u_2 ... u_n\}$ represent a set of sensors (users & mobile device, WSN sensors, etc.), and let $P = \{p_1, p_2, ... p_n\}$ represent a set of sensing systems (e.g., WSNs or PSNs). Recall that for simplicity throughout the text the descriptions of concepts are mainly based on PSNs, but the concepts applies for other sensing processes as well. In fact, an application considering also other source of data, besides PSN, is illustrated in Section V-B.

Each user $u_i \in U$ can share unlimited data on any PSN $p_k \in P$. Each j-th data shared $d_j^{p_k}$ into a PSN p_k has the form $d_j^{p_k} = \langle t, m \rangle$, where t refers to a timestamp when user u_i has shared data in p_k, and m is a tuple containing attributes of the shared data. The tuple m is composed of the attributes

[5]http://www.google.org/flutrends.
[6]Centers for Disease Control and Prevention - https://data.cdc.gov/.
[7]www.bing.com/maps.

Timestamp (t)	Attributes (m) Area (a)	User (u)	Specialty data (s)
T1	a_1	1	"Times square"
T1	a_1	2	"Times square"
T2	a_2	1	"Fifth Av."
T3	a_4	1	"Statue of Liberty"

(a) Foursquare PSN

Timestamp (t)	Attributes (m) Area (a)	User (u)	Specialty data (s)
T1	a_1	3	"Traffic Jam"
T2	a_2	2	"Accident"
T2	a_3	3	"Police control"

(b) Waze PSN

Timestamp (t)	Attributes (m) Area (a)	User (u)	Specialty data (s)
T1	a_1	3	"photo data"
T3	a_4	1	"photo data"

(c) Instagram PSN

TABLE I
DATA STREAM DESCRIBING USERS ACTIVITY IN THREE DIFFERENT PSNS: FOURSQUARE, WAZE, AND INSTAGRAM.

present in all sensing layers data, in this case $m = (a, u, s)$, where a is the area of the location where the data was shared, s is the specialty data, and u refers to the user $u_i \in U$ who shared the data.

The data shared in $p_k \in P$ can be viewed as a data stream B^{p_k}. We define that a data stream B^{p_k} consists of all n data shared by users U in a PSN p_k in a given time. Thus, $B^{p_k} = \langle d_1^{p_k}, d_2^{p_k}, ..., d_n^{p_k} \rangle$, and B^{p_k} represents a sensing layer r_{p_k}. Table I shows examples of data present in sensing layers that have been shared in the three PSNs p_1, p_2, and p_3, illustrated in Figure 2, which represents three users sharing data in different PSNs, p_1 (red cloud), p_2 (green cloud) and p_3 (orange cloud) at three different time intervals ($T1$, $T2$ and $T3$). Note that data in the same stream can have the same time[8], since they may have been shared by multiple users simultaneously.

One way to work with sensing layers is to represent them in the same structure, what we call here *work plan*, containing one or more layers. This work plan represents the resulting plan composed by data combined after applying appropriate algorithms to the corresponding layers we are interested in. How to perform this combination depends on the functionality of the layer(s) that it captures. The abstraction used to represent a combination of data from one or more layers is a data dictionary M, which is a collection of pairs $\{key : value\}$. This structure was chosen because of its simplicity, which helps to ease the concepts understanding. Keep in mind that other structures could be used, as long as they respect the principles represented here.

We define that the operation responsible for the work plan creation is called $\boldsymbol{COMBINATION}(\mathcal{F}, relation())$, where $\mathcal{F}$ is a subset of $\mathcal{B} = \{B^{p_1}, B^{p_2}, ..., B^{p_n}\}$, or $\mathcal{F} \subseteq \mathcal{B}$, and $relation()$ is a function that defines the relationship between data from the streams B^{p_k} contained in $\mathcal{F}$. The function $relation()$ defines the keys of the work plan M, and the data that these keys refer to, which are other observations of the data $d_i^{p_k}$ not used as key. The operation

[8]This model faces the clock synchronization problem. Therefore, "same time" means close times accepted to be considered equivalent.

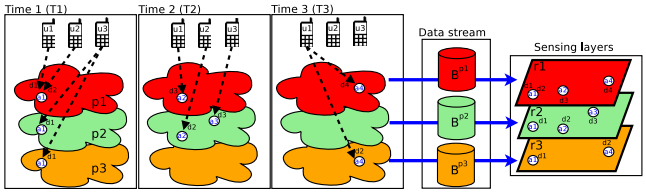

Fig. 2. Illustration of sharing data in three PSNs throughout the time, resulting in layers.

$\boldsymbol{COMBINATION}$ results in the work plan M.

To demonstrate the operation $\boldsymbol{COMBINATION}$, we illustrate here two types of relations used to combine data: (1) by location and (2) by users (sensors). To demonstrate a work plan containing combined data by location, consider the activity shown in Figure 2. In this case, $\mathcal{F} = \{B^{p_1}, B^{p_2}, B^{p_3}\}$. The work plan M_1 represents this activity, and it is illustrated in Figure 3. Observe that the work plan represents data that have been shared across all considered layers. The color of the symbol representing a given data d_i' indicates from which layer it was extracted. The data shared in the same location are grouped and indexed by the key that represents the location. In the work plan M_1, one key k_i is represented by a_i, which is a unique area among all areas of all data shared in the considered layers: r_1, r_2, and r_3. The $d_i^{r_j'}$ refers to the observations not used as key of the data d_i from the layer r_j, or $\langle t, u, s \rangle$. Thus, each unique areas become a key in work plan M_1. Work plan M_1, as described, presents the following structure: $M_1 = \{a_1 : \{d_1^{r1'}, d_2^{r1'}, d_1^{r2'}, d_1^{r3'}\}, a_2 : \{d_3^{r1'}, d_2^{r2'}\}, a_3 : \{d_3^{r2'}\}, a_4 : \{d_4^{r1'}, d_2^{r3'}\}\}$.

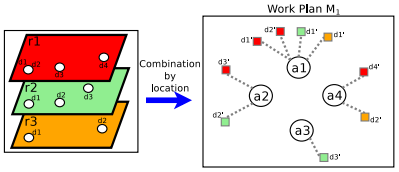

Fig. 3. Combination by location.

Figure 4 illustrates the combination by user. In this case, a work plan is build containing keys that represent user ids. The figure shows the work plan M_2, which was created considering the activities shown in Figure 2. The content of the work plan is: $M_2 = \{u_1 : \{d_1^{r1'}, d_3^{r1'}, d_4^{r1'}, d_2^{r3'}\}, u_2 : \{d_2^{r1'}, d_2^{r2'}\}, u_3 : \{d_1^{r2'}, d_1^{r3'}, d_3^{r2'}\}\}$. As we can see, each unique user has become a key in M_2. This work plan grouped all attributes by the same user in different layers.

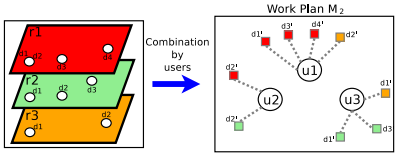

Fig. 4. Combination by users.

C. Issues of Data from Multiple Layers

There are issues when dealing with data from several layers simultaneously. For instance, in order to perform data combination, such as by location or user, as exemplified, we have to make sure that the data is consistent in all layers. This is a mandatory condition for correct functioning.

Consider that we want to combine two layers A and B by locations. The format of data location in Layer A is expressed as latitude and longitude, and as street address in Layer B. One way to solve this inconsistency is performing a geocoding process, using, for example, the Yahoo! geocoding tool[9]. In this way the street address will be transformed in a geographic coordinate (latitude and longitude).

Another issue that might happen when combining data by location is regarding to areas that overlap each other. How to define a key in this case? One possibility is consider several keys, one for the intersection between those areas, and one or two[10] as the area(s) not overlapped. Another option is to define just one key, this might be interesting when one area is inside another, so the key becomes the bigger area.

The combination by users is specially an issue when our sensor is an user, as in PSNs, because the same user may participate in different layers. Let's suppose we want to combine data by users using the check-ins layer (obtained from Foursquare) and the picture of places layer (obtained from Instagram). Since we are dealing with independent systems, users (sensors) have different identification. One way to try to bypass this issue is verifying other networks in order to match the user ID of one layer in another. For example, users of Foursquare and Instagram tend to be also users of Twitter [5]. In this way, the key in the combination process could be the twitter ID.

Note that if the combination by user desired is between a PSN layer and other layers that doesn't have users as sensors, such as WSNs, the inconsistency does not exist, because every sensor has its unique ID. Although it is necessary to evaluate if a combination by users (sensors) between those layers lead to the desired information.

Another issue is that different layers might refer to data valid for different interval of times. This is natural because some data sources provides near real time data, others not. For example, an alert in a Waze PSN refers to a traffic situation that may not exist five minutes later. However, a census data usually is valid for a big interval of time, months or years, until the next census is released. We have to be aware of all those issues when designing new applications and define a way to treat them.

There might be other issues. For example, issues related to the volume of data. If we do not have significant data for a certain layer, its utilization may not lead to the correct information extraction. Different data sources may present different characteristics for this issue. For instance, in a PSN many factors influence the volume of data, for example, geographical, cultural and economical aspects. The granularity of areas may also influence other data sources. If we consider, for example, data from WSN as a layer we may not have data for an entire metropolis, because of scalability problem.

In summary, note the importance of a characterization processes, as we shown in [1]. We have to know the properties of the layers we want to use, in order to verify if their simultaneous use may lead to the intended information extraction. The $relation()$ informed to the $\boldsymbol{COMBINATION}$ encapsulates the solution chosen for dealing with heterogeneous data, which is application dependent. If there is no solution to eliminate the inconsistency between data from two layers, then they can not be used together.

[9]https://developer.yahoo.com/boss/geo.

[10]If one area is not completely inside another.

III. Related Work

The use of layered (multi-layer) models to extract new information or design new applications is not new [6], [7], [8]. Very recent studies focused on a particular type of multi-layer network, the multiplex, where each agent can be networked in different ways, and with different intensity, on several multiple layers simultaneously. This model is useful, for example, to study links that the same user has in different social networks (layers), for instance, to better understand the information spreading or to measure social tie strength (the case studied by Hristova et al. [7]). Another example is the study of transportation in a city. The network of bus routes and stops (layer 1) is different from a subway network (layer 2) in the same city, but a user can use both networks to reach its destination [8].

In the same direction, Xin et al. [9] proposed a layered graph to model to develop routing and interface assignment algorithms. Laura et al. [10] proposed a layered model for the Web network, aiming to design a model that resembles better the complex nature of the Web. A GIS (geographic information system) is another example, because it often utilizes a layered model for characterizing and describing our world. It uses maps to visualize and work with geographic information in several layers [11]. GIS is related to the ideas proposed here, in fact, some GIS tools could be used to support the proposed framework, for example, in the combination process. Our proposal differ from a simple implementation of a GIS because it is not driven by jurisdictional (such as a city), purpose, or application requirements. We focus on the discussion of a sensing layer framework. Besides that we envision demonstrate the potential of simultaneous use of sensing layers derived from PSNs, for the extraction of new information related to the study of city dynamics and urban social behavior.

Recently, there is an evident interest in understanding how users behave across different Web systems [12], [13], [14], [15], [16], [17]. More related to our proposal, there are studies that consider different sources of data simultaneously to better understand the dynamics of cities. For example, Bollen et al. [18] investigated whether collective mood states derived from Twitter feeds are correlated to the value of the Down Jones Industrial Average (DJIA) over time. Ribeiro et al. [19] correlate data from Foursquare and Instagram with traffic conditions reported by Bing Maps. Gomide et al. [20] analyzed how Dengue epidemic is reflected on Twitter. Martani et al. [21] examine the underlying drivers of energy consumption through several sensed data. Sagl et al. [22] analyzed the collective human behavior based on mobile data, and correlated it with meteorological data from weather stations.

In sum, this work differs from all previous studies because: (i) define the concept of PSNs as sensing layers; (ii) propose a framework that enables integration of the analysis and exploration of multiple layers simultaneously; and (iii) present

applications that use the proposed framework and illustrate the potential of using multiple sensing layers.

IV. Processing Sensing Layers

This section discusses how to process one or more sensing layers. To that end a number of example operations are proposed. Section IV-A presents examples of such operations and Section IV-B presents some strategies to process layers using the proposed operations.

A. Operations

In Section II-B, we illustrate how to represent sensing layers in a work plan, for example, by location (M_1) or users (M_2). The general purpose of work plans is to be basic structures that can be easily manipulated. Recall that the structure chosen here to represent a work plan is a data dictionary. Having a work plan, as the M_1 or M_2 shown in Figures 3 and 4, we can apply operations to derive other structures and also extract new information. The list below provides examples of some generic operations:

- **dGRAPH** (directed graph): It expects as input a work plan M, and the result is a directed graph $G = (V, E)$. This operation builds a directed graph $G = (V, E)$, where each key k_i in the work plan represents a node $v_i \in V$, and the data indexed by k_i are attributes of v_i. An edge $e = (v_i, v_j)$ is added depending on the desired analysis, which is expressed through some specific operations, as we describe below. Initially, $E = \varnothing$. All variables of the work plan are incorporated in the graph;
- $\boldsymbol{CNG}$ (change): It expects a work plan M, a layer identification (ID), and a status (0 or 1). It results in the alteration of the variable h of the informed layer, i.e., it changes the status of a layer through the variable h. If the informed status is 0, then $h = 0$ and the work plan are adjusted accordingly, i.e., this particular layer of the work plan is disabled. The layer disabled can be enabled again with the same data at the disabling time;
- $\boldsymbol{RESET}$: It expects a directed or undirected G graph, and results in a work plan M. It is extracted all the necessary information from the graph to build a corresponding work plan. All variables of the graph are incorporated in the work plan;
- $\boldsymbol{dEDGE}$ (directed edges): It expects a directed graph G resulted from a work plan combined by locations, and results in a graph G' containing directed edges. This operation creates a directed edge from node v_i to node v_j if and only if at least one user shared data, in any layer, in the location represented by the node v_j right after sharing data, also in any layer, in a location represented by the node v_i. The weight of an edge represents the total number of transitions performed from v_i to v_j considering transitions of all users. Note that it is possible to have more than one transition for the same user;
- $\boldsymbol{DEL}$ (delete): It expects a graph G and an integer t. The result is a subset graph G_{subset} derived from G. This operation deletes edges $e_i \in E$ (E is a set of edges of G), with weight $w_i < t$;
- $\boldsymbol{rdGRAPH}$ (random directed graph): It expects a directed graph $G(V, E)$. The result is a random directed graph $G_R(V, E_R)$. The random graph G_R is constructed keeping the same nodes of G and uses the same number of individual transitions of G. However, instead of considering the real transition $v_i \rightarrow v_j$ performed by an individual, the operation randomly choose two nodes to replace v_i and v_j, simulating random transitions performed by users;
- $\boldsymbol{MERGE}$: It expects a work plan M_1, a work plan M_2, and a data relation $relation()$. M_1 and M_2 have to be produced following the same data relation, for example, by locations as explained above in the process **COMBINATION**. This operation results in a work plan $M_{merged}(V, E)$ representing the merge of the sensing layers represented by M_1 and M_2. This operation merge information of M_1 and M_2, respecting the data relation informed $relation()$.

We can have also specific operations to produce new information (which could be represented in a new layer), using one or more existing layers, such as the following operations:

- $\boldsymbol{fPOIS}$ (find POIs): It expects a work plan M representing a layer such as *check-ins* and *pictures of places* combined by locations. Other layers might also be used, but previous verification of feasibility is needed, for example, data might not be available for the geographical region of interest. This operation results in a work plan of a new layer containing popular areas, or points of interest (POI), based on the number of activities performed on them. This operation identifies POIs applying the algorithm specified in [23], to select geographic areas;
- $\boldsymbol{fSIGHTS}$ (find sights): It expects a work plan M^{POIs} containing POIs. The result is a graph G^{SIGHTS} containing sights. This operation identifies sights from a work plan M^{POISs}, where keys are the areas a of POIs identified in a particular pre-defined geographic region. This algorithm is described in [23]. More details of this operation are presented in Section V-A.

We chose specifically those operations because they are used in the applications presented in the next sections. Note that other operations can be proposed. For instance, another operation to create edges differently from $\boldsymbol{dEDGE}$. This new operation, called for example $\boldsymbol{uEDGE}$, could be suitable for a graph G produced from a work plan combined by users. The operation $\boldsymbol{uEDGE}$ could create an undirected edge between v_i and v_j, if and only if user u_i, represented by node v_i, shared data in the same location (layer independent) that user u_j, represented by node v_j. The weight of an edge represents the total number of locations that nodes v_i and v_j have in common. Other operations could be designed to add (directed or undirected) edges with different way to assign weights.

B. Processing Strategies

As shown in the previous section, our framework provides several operations useful to process sensing layers in several manners. To give an example of the results we can obtain in processing sensing layers using those operations, we demonstrate how to obtain: flow graphs, graphs that map the locations where the same user shared data, thus capturing the movements or transitions in a geographical area; and also points of interest and sights. It is particularly interesting to illustrate the creation of flow graphs because its is a fundamental piece of some operations, for instance $\boldsymbol{fSIGHTS}$.

Consider the data sharing of the situation illustrated in Figure 2. After a certain time, we can process the data in order to extract knowledge in different ways. Take for instance the

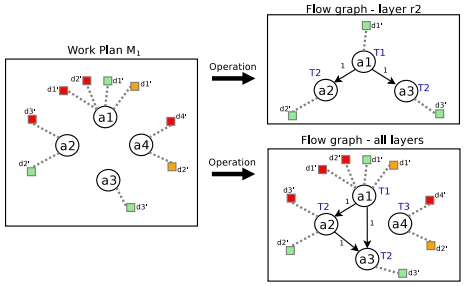

Fig. 5. Illustration of flow graph creation from one single layer, and also from multiple layers.

Algorithm 1: Generation of a flow graph for one single layer.

input : work plan M combined by locations
output: flow graph $G_{r_2}^{flow}$ that represents data from layer r_2

```
M ← M_1 ;        // M_1 is the work plan created previously
M' ← CNG(M, r1, 0);
M'' ← CNG(M', r3, 0);
G ← dGRAPH(M'');
G_{r2}^{flow} ← dEDGE(G);
```

flow graph labeled "flow graph - layer r2", shown in Figure 5. The Algorithm 1 describe the steps necessary to generate this graph, referred to as $G_{r_2}^{flow}$ (built from layer r_2). In this algorithm we consider the work plan M_1 as explained above (combined by locations). We initially apply the operation $\boldsymbol{CNG}$ hiding layers r_1 and r_3. After that, we have to generate a directed graph G using $\boldsymbol{dGRAPH}$ and apply the operation $\boldsymbol{dEDGE}$ in G, obtaining $G_{r_2}^{flow}$. In this case, we have a flow graph that represents data from a single layer. With this graph we can extract many valuable information, for example, regular trajectories in a city.

Another possible analysis is to consider different layers simultaneously. In Figure 5, the part named the "flow graph – all layers" shows a graph, which we call G_{all}^{flow}. The Algorithm 2 describes the steps necessary to generate G_{all}^{flow}. This algorithm also consider the work plan M_1 created above. As we can see in the algorithm, in order to obtain G_{all}^{flow} we need to apply the operation $\boldsymbol{dEDGE}$ in G. In the resulting graph, the nodes represent data shared in the same location at any layer. Edges connect nodes $v_i \rightarrow v_j$ if at least one user shared data in the location represented by node v_j, right after sharing a data in the location represented by the node v_i.

Algorithm 2: Generation of a flow graph for multiple layers.

input : work plan M combined by locations
output: flow graph G_{all}^{flow} that represents data from multiple layers

```
M ← M_1 ;        // M_1 is the work plan created previously
G ← dGRAPH(M);
G_{all}^{flow} = dEDGE(G);
```

New information could be obtained by processing data available from one or more sensing layers. Points of interest (POI) in a city, identified from data shared in Instagram and obtained using operation $\boldsymbol{fPOIS}$, represent an example. To identify a sight it is necessary the POIs, according to the operation $\boldsymbol{fSIGHTS}$. This is demonstrated in Figure 6.

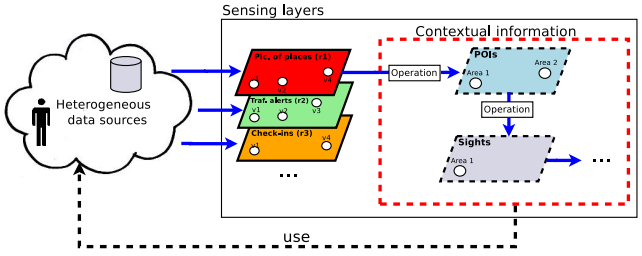

Fig. 6. Illustration of new layers creation from the picture of places layer.

In this figure, the new information obtained is expressed as new layers. Note that these new layers are represented in the box labeled "Contextual information". Basically, new information generated from other sensing layers are contextual information. Recall that contextual information might have the power to influence the data generation. For example, once users know where are the points of interest they may tend to share more data in those places instead of others.

V. Applications Using the Sensing Layers Framework

In this section, we discuss two applications that illustrate the potential of the proposed framework for working with sensing layers, in Sections V-A and V-B.

A. Identification of Sights

We discuss an application that identifies sights considering multiple layers simultaneously, highlighting the improvements on the strategy presented in [23], which considers only one layer. In this new analysis, we consider a Instagram and a Foursquare dataset, which were collected directly from Twitter[11], since Instagram photos and Foursquare check-ins are not publicly available by default. Those datasets have the time of collection in common (11/May/2013 – 25/May/2013). Each content (photo or check-in) consists of GPS coordinates (latitude and longitude) and the time when it was shared.

The picture of places layer (r_1) is represented by the Instagram dataset, and the layer check-ins (r_2) by Foursquare dataset. Our goal is to obtain results using both layers. To that end, we first combine the data by location, producing a work plan M_1. First we want to identify sights for the layer r_1. With that in mind, we disable layer r_2 from M_1 using the operation $\boldsymbol{CNG}$ obtaining M_{r_1}. After that, we apply $\boldsymbol{fPOIS}$ in M_{r_1} to generate $M_{r_1}^{pois}$, work plan containing the POIs. In the resulting work plan $M_{r_1}^{pois}$, the keys are the areas of the identified POIs. In the scenario illustrated in Figure 6, we have only two keys for a work plan representing POIs, represented by Area 1 and Area 2.

Each POI in $M_{r_1}^{pois}$ represents a popular area a in a given geographical region, e.g., a city. Popularity is identified through the volume of shared data made available by users u. That is, a POI represents the activity performed by a group of users u in a time interval t. Note that, the specialty data s, in this particular case, is the POI area itself. We use the work plan $M_{r_1}^{pois}$ for the extraction of sights with the help of the operation $\boldsymbol{fSIGHTS}$. First, the operation $\boldsymbol{fSIGHTS}$ creates a directed graph (in the example $G_{r_1}^{pois}$), from the

[11]http://www.twitter.com.

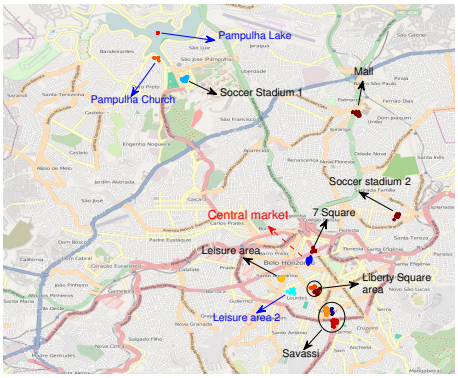

Fig. 7. All identified sights with Foursquare and Instagram datasets.

received work plan (in the example $M_{r_1}^{pois}$), using operation $\boldsymbol{dGRAPH}$. Next it maps the flow of users performed between POIs. For this, it applies the operation $\boldsymbol{dEDGE}$ in $G_{r_1}^{pois}$ obtaining $G_{r_1}^{pois-flow}$. After that, popular transitions that connect two nodes $v_i \rightarrow v_j$ are selected. For that, it uses the operation $\boldsymbol{DEL}$ in the graph $G_{r_1}^{pois-flow}$ using a parameter t. The parameter t in this case is calculated in the way presented in [23]. According to the conjecture considered in the algorithm of the operation $\boldsymbol{fSIGHTS}$, the popular transitions selected connects the sights, which are represented in the graph $G_{r_1}^{sights} = (V', E')$. In this graph, nodes $v_i \in V'$ are the areas a of the identified sights.

Next, we identify the sights $G_{r_2}^{sights}$ for r_2. First, we enable r_2 and disable layer r_1 from M_1. The next steps are performed similarly to the way it was presented for r_1. After that we merge the contextual layers containing the sights for layers r_1 and r_2, $M_{r_1}^{sights}$ and $M_{r_2}^{sights}$, respectively, in the work plan M_{total}^{sigths}. This work plan contains all identified sights, which are shown by the Figure 7.

The sight indicated by a red arrow (Central Market) was identified only by Foursquare. Sights pointed by a blue arrow (Pampulha Church, Pampulha Lake, and leisure Area 2) were not identified with Foursquare. All sights are very relevant. It is important to observe the potential for complementary results using both layers.

B. Economic-Cultural Analysis of Regions

The application described in this section allows various economic-cultural analyses. In this paper, we focus on two, which complement previous studies that correlated economic status with social media data [24]. The objective of the first analysis is to correlate the general sentiment expressed in the tips for all locations in a given census tract a_i (geographic region defined for the purpose of taking a census), with the median income of the inhabitants of this tract. On the other hand, the aim of the second analysis is to study the movement of users in the considered tracts, taking into account the typical income of these tracts. This second analysis aims to identify possible social segregation in a city.

To illustrate this application, we consider two datasets derived from Foursquare and one derived from the census of NY. The first, named CHECKINS-NY, consists of 34,677 check-ins performed in New York City, in a week of April 2012. CHECKINS-NY is a subset of a Foursquare dataset used in [1]. The second dataset, named TIPS-NY, contains all the tips contributed by users up to January 2013 in all unique locations of the dataset CHECKINS-NY. The tips were collected through the Foursquare API. Each tip contains a location, a user ID, a time, and the textual content of the tip. We consider only tips in English. We define that a tip is in the English language if at least half of the words of the tip is listed in a dictionary containing key words in English. This resulted in 157,197 tips (2,531 discarded). The last dataset, named CENSUS-NY, contains information of the census of New York City, and it refer to the 2006-2010 American Community Survey. The area of each tract is pre-defined by the census of New York. It contains, among other information, the median income per tract (information we are interested here).

Group	Mean Sent. (std)	(+3,+4)%	(+1,+2)%	(0)%	(-1,-2)%	(-3,-4)%
<25000	0,46 (0,67)	0	73,08	21,15	5,77	0
$\geqslant 25000$ and <50000	0,73 (0,63)	1,23	84,31	12,92	1,23	0,31
$\geqslant 50000$ and <75000	0,81 (0,46)	0,40	93,28	5,93	0,39	0
$\geqslant 75000$ and <100000	0,9 (0,36)	0	96,97	3,03	0	0
$\geqslant 100000$	0,87 (0,28)	0,96	98,08	0,96	0	0

TABLE II
GENERAL SENTIMENT PER GROUPS OF TRACTS

The TIPS-NY dataset is used to represent a sensing layer called tips of locations (r_1). The layer r_1 is composed of a data stream B_i. Each data stream has the form: $\langle t, (a, u, s)\rangle$. An example of the specialty data s of this layer is: "This place is awesome, I recommend the burger.". The income layer (r_2), derived from CENSUS-NY, is composed of a dataset d_j for different tracts of New York. Each specialty data in d_j has the median income of the inhabitants of a particular tract. The form of d_1 is t =2006-2010, a = [area of Tract 1], u = "USA Census", s = "median income in US$ for the Tract 1"$\rangle$. Note that, this is an example of layer obtained from a different source than PSNs. This illustrates the use of other sources of data about predefined geographical regions.

For the first analysis we combine the data from the layers r_1 and r_2. The chosen method is the combination by location, method described in Section II-B. This combination process consider the keys as the areas of the tracts. Each key k_i combines, among other data, the tips of all the places that are located within the area of a tract, and the median income information of the tract. The combination process results in a work plan M_1.

Thus, we use M_1 to calculate the general sentiment about all locations in each tract. For this analysis, we used the program SentiStrength [25][12], to classify the sentiment expressed in the tips. SentiStrength computes the sentiment of a tip in a scale from -4 (strongly negative) to 4 (strongly positive), 0 indicates a neutral sentiment. This program is applied to each tip and then combined by location, and finally by tract.

Then we calculate the average sentiment for all locations in a given tract. Next, we group the tracts in five income groups: less than US$25,000; between US$25,000 and US$50,000; between US$50,000 and US$75,000; between US$75,000 and US$100,000; and over US$100,000. Finally, we calculate the average value of sentiment for each of the five income groups, considering all tracts that belong to each group.

Table II presents this result, and also for each income group, the percentage of average sentiment that falls in one of five range of sentiment: (+3, +4), (+1, +2), (0), (-1, -2), and (-3, -4). As we can observe, the result suggests that poor tracts

[12]We used the tool IFeel [26] to help in the selection of this sentiment analysis program.

tend to have the worst sentiment expressed by users. This may be associated with low quality services in these tracts. With the tract income increasing, opinions tend to be more positive. Although the average sentiment for the richest tracts group (over US$100,000) is slightly lower than the second richest (between US$75,000 and US$100,000), this group still has a larger number of positive tips compared to all other groups, and does not have negative tips. Note the potential of this analysis for social studies, e.g., for the study of inequalities in the quality of services in cities.

For the second analysis the dataset CHECKINS-NY is used to represent a sensing layer called check-ins (r_3). We combine layers r_2 (as defined above) and r_3 by location on the work plan M_2. We then create a graph G_2, and use it to generate a flow graph G_2^{flow}, where the edges are the transitions performed by the same user in different tracts (nodes in the graph). We exclude loops, i.e., visits from the same user on the same tract, generating then $G_2^{flow'}$. To gather evidence of the existence of segregation, we study the assortativity related to the median income by tract in $G_2^{flow'}$. This is a way to try to observe the existence of segregation.

The assortativity measures the similarity of connections in the network relative to a particular attribute, and ranges from -1 to +1 [27]. In an *assortative network* (with positive assortativity), vertices with similar values for a given attribute (e.g., the same income) tend to be connected (be similar) to each other, whereas in a *disassortative network* (negative assortativity), the opposite happens. All tracts were associated with a class based on the median income of the tract: Class A for median incomes up to US$75,000; and Class B for higher median incomes. The assortativity considering these two classes as attributes of $G_2^{flow'}$ is 0.14. Thus, the network for this attribute is assortative, indicating a trace of segregation, i.e., users tend to share content (or attend) in tracts that have the same class of income.

After that, we create ten random graphs $G_{Ri}(V, E_{Ri})$, where $i = 1, \ldots, 10$, using the operation $\boldsymbol{rdGRAPH}$. For each graph G_{Ri} is also randomly associated a class of a node, A or B. The number of nodes of class A and B are also consistent with the one observed in $G_2^{flow'}$. After that, we calculate the assortativity for all random graphs $G_{R1..10}$. The assortativity for all graphs, with 95% confidence level, are in the range is: $[-0.0084, -0.0014]$. As we can see, these random networks do not indicate segregation. Obviously, in order to draw any conclusion in this sense, a more detailed investigation is needed. However, this result shows the potential for joint analysis of multiple layers.

Note also the potential of considering the same layers to generate a work plan M_3 combined by users. Besides identifying users' preferences, we can also try to infer their social class studying the income of the tracts that the user visits. This can be useful for social studies, and for more effective advertising.

VI. Final Remarks

This work formalizes the concept of sensing layers in participatory sensor networks, and shows how we can design sophisticated applications from different sensing layers. This study also presented applications that illustrate the potential of sensing layers.

Many other applications could be proposed. For example, in any city is likely to find many places where people perform more often a particular activity, for example an area of bars and restaurants where people meet to socialize. These locations could be identified with the help of the check-ins layer. The information provided by other layers could help users choosing the best areas of interest at the moment. For example, a user could use the information provided by the traffic alerts layer to identify among all the options, the area with the lowest number of traffic problems at the time, and use the picture of places layer to view the style of the establishments in those areas and the people who frequent them.

As future work, we plan to design more sophisticated applications that show the potential of sensing layers and consider other aspects in those layers that could explore such data mining.

References

[1] T. Silva, P. Vaz De Melo, J. Almeida, and A. Loureiro, "Large-scale study of city dynamics and urban social behavior using participatory sensing," *Wireless Communications, IEEE*, vol. 21, no. 1, pp. 42–51, Feb 2014.

[2] A. K. Dey and G. D. Abowd, "Towards a Better Understanding of Context and Context-Awareness," in *CHI Workshop*, The Hague, The Netherlands, 2000.

[3] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "Google flu trends still appears sick: An evaluation of the 2013-2014 flu season," 2014.

[4] S. Lohr, *Google Flu Trends: The Limits of Big Data*, New York Times, March 2014.

[5] M. Duggan and A. Smith, *Social Media Update 2013*, Pew Research, Jan 2014.

[6] M. Kivela, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer Networks," 2013. [Online]. Available: arxiv.org/abs/1309.7233

[7] D. Hristova, M. Musolesi, and C. Mascolo, "Keep Your Friends Close and Your Facebook Friends Closer: A Multiplex Network Approach to the Analysis of Offline and Online Social Ties," in *Proc. of ICWSM'14*, Ann Arbor, USA, 2014.

[8] M. D. Domenico, A. Sole-Ribalta, E. Cozzo, M. Kivela, Y. Moreno, M. A. Porter, S. Gomez, and A. Arenas, "Mathematical formulation of multi-layer networks," *Physical Review X 3, 041022*, 2013.

[9] C. Xin, B. Xie, and C.-C. Shen, "A novel layered graph model for topology formation and routing in dynamic spectrum access networks," in *Proc. of DySPAN'05*, Baltimore, USA, 2005, pp. 308–317.

[10] L. Laura, S. Leonardi, G. Caldarelli, P. De, and P. D. L. Rios, "A multi-layer model for the web graph," in *Proc. of Workshop on Web Dynamics*, Honolulu, USA, 2002.

[11] K.-t. Chang, *Introduction to geographic information systems*. McGraw-Hill New York, 2010.

[12] B. Hecht and M. Stephens, "A Tale of Cities: Urban Biases in Volunteered Geographic Information," in *Proc. of ICWSM*, Ann Arbor, USA, 2014.

[13] M. Rowe and H. Alani, "Mining and comparing engagement dynamics across multiple social media platforms," in *Proc. of WebSci*, Bloomington, Indiana, USA, 2014, pp. 229–238.

[14] F. Figueiredo, J. M. Almeida, F. Benevenuto, and K. P. Gummadi, "Does Content Determine Information Popularity in Social Media? ," in *Proc. of CHI*, Toronto, Canada, 2014.

[15] A. Lima, L. Rossi, and M. Musolesi, "Coding Together at Scale: GitHub as a Collaborative Social Network," in *Proc. of ICWSM*, Ann Arbor, USA, 2014.

[16] C. Wagner, S. Asur, and J. Hailpern, "Religious politicians and creative photographers: Automatic user categorization in twitter," in *Proc. of SocialCom*, 2013, pp. 303–310.

[17] R. Ottoni, D. de Las Casas, J. ao Paulo Pesce, W. M. Jr, C. Wilson, A. Mislove, and V. Almeida, "Of Pins and Tweets: Investigating how users behave across image-and text-based social networks," in *Proc. of ICWSM*, Ann Arbor, USA, June 2014.

[18] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[19] A. I. J. T. Ribeiro, T. H. Silva, F. Duarte-Figueiredo, and A. A. Loureiro, "Studying traffic conditions by analyzing foursquare and instagram data," in *Proc. of PE-WASUN*, Montreal, Canada, 2014, pp. 17–24.

[20] J. Gomide, A. Veloso, W. M. Jr., V. Almeida, F. Benevenuto, F. Ferraz, and M. Teixeira, "Dengue surveillance based on a computational model of spatio-temporal locality of twitter," in *Proc. of WebSci*, Evanston, USA, 2011.

[21] C. Martani, D. Lee, P. Robinson, R. Britter, and C. Ratti, "Enernet: Studying the dynamic relationship between building occupancy and energy consumption," *Energy and Buildings*, vol. 47, pp. 584–591, 2012.

[22] G. Sagl, T. Blaschke, E. Beinat, and B. Resch, "Ubiquitous geo-sensing for context-aware analysis: Exploring relationships between environmental and human dynamics," *Sensors*, vol. 12, no. 7, pp. 9800–9822, 2012.

[23] T. H. Silva, P. O. S. Vaz de Melo, J. M. Almeida, J. Salles, and A. A. F. Loureiro, "A picture of Instagram is worth more than a thousand words: Workload characterization and application," in *Proc. of DCOSS'13*, Cambridge, USA, 2013, pp. 123–132.

[24] C. Vaca Ruiz, D. Quercia, L. M. Aiello, and P. Fraternali, "Taking brazil's pulse: Tracking growing urban economies from online attention," in *Proc. of WWW*, Seoul, Korea, 2014, pp. 451–456.

[25] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment in short strength detection informal text," *J. Am. Soc. Inf. Sci. Technol.*, vol. 61, no. 12, pp. 2544–2558, Dec. 2010.

[26] P. Gonalves, M. Arajo, F. Benevenuto, and M. Cha, "Comparing and combining sentiment analysis methods," in *Proc. of COSN'13*, Boston, USA, 2013.

[27] M. E. Newman, "Assortative mixing in networks," *Phy. rev. let.*, vol. 89, no. 20, p. 208701, 2002.