# DynamiCITY : Revealing city dynamics from citizens social media broadcasts

Vasiliki Gkatziaki, Maria Giatsoglou, Despoina Chatzakou, Athena Vakali

PII: S0306-4379(17)30065-0
DOI: 10.1016/j.is.2017.07.007
Reference: IS 1236

To appear in: *Information Systems*

Received date: 22 January 2017
Revised date: 25 June 2017
Accepted date: 25 July 2017

Please cite this article as: Vasiliki Gkatziaki, Maria Giatsoglou, Despoina Chatzakou, Athena Vakali, DynamiCITY : Revealing city dynamics from citizens social media broadcasts, *Information Systems* (2017), doi: 10.1016/j.is.2017.07.007

**Highlights**

- The problem of modeling cities into dynamic areas is addressed
- The methodology proposed extends and compares Dynamic Area Extraction Approaches
- A hybrid representation of POIs that combines several features is introduced
- The usefulness of the proposed approach is demonstrated through DynamiCITY platform

# DynamiCITY: Revealing city dynamics from citizens social media broadcasts

Vasiliki Gkatziaki[a], Maria Giatsoglou[a], Despoina Chatzakou[a], Athena Vakali[a]

[a] *Informatics Department, Aristotle University of Thessaloniki, Greece*

**Abstract**

Social media and mobile devices have revolutionized the way people communicate and share information in various contexts, such as in cities. In today's "smart" cities, massive amounts of multiple forms of geolocated content is generated daily in social media, out of which knowledge for social interactions and urban dynamics can be derived. This work addresses the problem of detecting urban social activity patterns and interactions, by modeling cities into "*dynamic areas*", i.e., coherent geographic areas shaped through social activities. Social media users provide the information on such social activities and interactions in cases when they are on the move around the city neighborhoods. The proposed approach models city places as feature vectors which represent users visiting patterns (*social activity*), the time of observed visits (*temporal activity*), and the context of functionality of visited places *category.* To uncover the dynamics of city areas, a clustering approach is proposed which considers the derived feature vectors to group people's activities with respect to location, time, and context. The proposed methodology has been implemented on the DynamiCITY platform which demonstrates neighborhood analytics via a Web interface that allows end-users to explore neighborhoods dynamics and gain insights for city cross-neighborhood patterns and inter-relationships.

*Keywords:* Data Mining, Crowdsourcing, Urban Dynamics, Social Data Mining, Smart City Applications

## 1. Introduction

Social media ubiquitous dominance has drastically reshaped the ways with which people interact and communicate. Today's online social media platforms are constantly accessed through multiple 'always on' (e.g., mobile) devices capable of capturing people's geolocation, time of activity, and other attribute declarations. In such city contexts, people act both as city reporting "producers" (via their constant mobile broadcasting on social media platforms), and as city service "consumers" (by the information pushed to them via multiple social media-driven applications).

Social media platforms allow users to broadcast their presence (i.e., *check-in*) at a specific place, or Point on Interest (POI), in real time. Such places are characterized by a specific category (e.g., café, office, restaurant, etc.), while users can collectively "form" their characteristics by providing comments, related photos, and tags. Thus, the role of "citizen prosumer" turns humans into "sensors" and enables them to impact the scale of city's social media information production and sharing, augmented with declarations of specific location and time of activity. The abundance of such geolocated user generated content (UGC) offers a valuable information repository which can be largely exploited to uncover urban dynamics and inter-relationships among people, places, and timelines [1, 2, 3, 4].

This work is motivated by the challenges brought from cities transformation into live and evolving "data engines" which can serve as testbeds of social urban interactions and dynamics detection. To deal with cities large scales and areas, the cities division into sub-regions (e.g., municipalities, neighborhoods, etc.) is considered as an important source of information to facilitate data's scope and visualization analytics. In this work, a methodology is proposed for the segmentation of a city into dynamic areas based on multiple urban space usage characteristics extracted from social media user activities (e.g., check-ins). The proposed approach is inspired by earlier efforts which uncovered areas within a city based on the clustering of its enclosed POIs, by leveraging their spatial distribution, combined with the distribution of their checked-in users *social characteristics* [3], or check-in timestamps *temporal characteristics* [4]. In the present work, state of the art is advanced by progressing on the assumption that city dynamics are not shaped by only a single dimension or type of characteristic, but rather on a set of dynamics 'forces'. A novel, *hybrid* city segmentation approach is proposed at which social, temporal, and spatial characteristics are integrated, as well as their context of functionality. The discovered dynamic areas reflect the stories that are shaped by their people, places, and activity. To this end, we present an area-level profiling approach that enables the comprehension of area's functionality and usage profile. The proposed methodology has been tested on multiple city cases, and here, due to lack of space, five real-world datasets (for New York, London, and San Francisco) are presented. The adopted area-level profiling approach allows comparisons among the dynamic areas between the different cities and discovered areas that present similar activity. The discovered dynamic areas are also demonstrated through the DynamiCITY Web interface that offers a comprehensive and insightful city analytics exploration.

The main contributions of this work are the following:

- *Extending and Comparing Dynamic Area Extraction Approaches:* a hybrid representation approach for POIs based on vectors that combine several types of features for capturing more effectively the POI properties and visiting patterns. These representations (profiles) characterize a given POI based on the observed user activities on it (*social* and *temporal* characteristics), as well as based on its *geographic location* and *context of functionality.* Dynamic, coherent areas are discovered based on a generic spectral clustering algorithm and we comparatively examine the results derived based on representations using only social or temporal characteristics, with the proposed hybrid representation approach.

- *Facilitating City Exploration:* we demonstrate the usefulness of the proposed approach in the context of complex urban environments through DynamiCITY. DynamiCITY is a platform comprising: (i) the Dynamic Area Extractor (the back-end component) which handles the data collection, analysis, and dynamic area extraction and summarization; and (ii) an online application which visualizes and presents the results to end-users.

The remainder of the paper is structured as follows. The next section reviews related research work to justify the proposed city segmentation methodology for the multi-featured discovery and profiling of dynamic regions, outlined in Section 3. Section 4 presents the DynamiCITY platform, which implements data collection, analysis, and dynamic area extraction and visualization. Finally, experimental results for three case studies on different large scale cities are demonstrated at Section 5, with conclusions and future work highlighted at Section 6.

## 2. Related Work

This work is relevant with existing research efforts in the areas of *city segmentation* approaches, *urban mobility pattern detection*, and *POI recommendation*, as summarized next.

**Urban mobility pattern detection.** Several studies have explored ways to incorporate geo-tagged data derived from social media within a geographic region such as a city area, in order to discover urban dynamics, mobility patterns, and **functional regions**. Noulas et al. [5] investigated whether and how the check-ins of Foursquare users can provide useful insights about their activities at a given time and place. Ferrari et al.[6] analyzed urban mobility patterns extracted from geo-tagged Twitter posts to explore user behavior within a city. Based on such mobility patterns, they organized different locations in clusters and examined the presence of patterns, i.e., locations that tend to be visited together (geographic topic-based models), in relation to different days of a week. Finally, Long et al.[7] used Foursquare check-in data to examine the relations between different geographic areas in terms of their co-appearance in the user trajectories.

**POI recommendation.** Geographic topic-based models, described in the previous paragraph, can be further leveraged in relation to a given POI along with the interests of citizens/city visitors for providing recommendations, predicting events, and building marketing strategies with respect to the city. Noulas et al. [8] proposed a random walk model for personalized new place recommendation in Foursquare, based on the exploitation of both the users' existing social ties and check-ins. Similarly, Long et al.[9] proposed a HITS-based (Hypertext Induced Topic search) approach for recommending POIs to social media users, by taking into account their social relations and promoting POIs with high authority scores (based on the users' hubiness). In addition to the users' social ties, the recommendation approach proposed by [10] also considered the users' opinions, as expressed in their tips on Foursquare POIs, and the POI categories. Capdevila et al. [11] developed GEOSRS, a POI recommender system that leverages user's reviews on places. More specifically, they built a recommender system that combines efficiently the text review content and the sentiment to produce personalized POI suggestions. In [12], they developed a personalized location-based recommendation approach by combining the geo-tags of photos uploaded by a given user in Flickr, with the geo-tags of the most similar users. Bao et al. [13] proposed a POI recommender for Foursquare by extracting candidate local experts for each POI category and calculating the similarity between users based on their social media check-in histories and preferences. Van et al. [14] identified POIs within a city (based on geo-tagged Flickr photos) and proposed a location-based recommender which considers both the overall popularity of POIs and the time that a given user visited the city.

**City segmentation.** Several factors influence the residents of a city to visit a place and can be leveraged for the **dynamic segmentation** of the city into regions. Unlike previously discussed approaches in relation to urban mobility patterns (such as [6]), city segmentation approaches derive regions based on the people's activities which are also **geographically coherent**. Intuitively, such approaches reflect more accurately the existing dynamics and behavioral/activity patterns, compared to a static city segmentation (e.g., based on population demographics, or fixed limits established by the municipality). In this sense, Cranshaw et al. [3] proposed Livehoods, a clustering methodology for segmenting a city into dynamic areas, based on the check-in activity of Foursquare users on the city's POIs, by exploiting both their spatial (i.e., geographic) and social (i.e., based on the distribution of their checked-in users) proximity. Also, Rosler and Liebig [4] proposed an approach for segmenting a city dynamically, based on temporal (i.e., temporal distributions of check-ins in a given region) and spatial characteristics. Zhang et al.[15] developed a methodology for generating boundaries of coherent neighborhoods dynamically by taking into consideration spatial characteristics and temporal activity (of tourists and locals) around Foursquare POIs, as well as their type. At first, the OPTICS clustering algorithm is applied on POIs for identifying activity hotspots, then the city is segmented into grid cells which are represented in terms of their enclosed hotspots, and neighborhoods are detected based on the homogeneity of nearby cells. Falher et al. [16] proposed a methodology for unveiling similar areas across different cities by harnessing their pair-wise *earth mover's* distances. Each area is represented as a feature vector that describes the activity in the corresponding area and the type of POIs it includes. Rizzo et al. [17] proposed a methodology for the automatic creation of thematic maps by leveraging geo-tagged data derived from social media. Specifically, they proposed a clustering algorithm, named GEOSUBCLU that detects homogeneous areas that are described by similar POIs in terms of the representative categories. Finally, Frias et al. [18] proposed a methodology for automatically detecting the use of different regions within a city based on user activity in Twitter, and also presented a POI recommendation approach.

**Advancing previous efforts.** In this paper, we propose a generic methodology for the dynamic segmentation of a city into areas by jointly considering various POI characteristics based on social media user activity, under the assumption that a city is formulated by a set of forces. More specifically, the revealing of dynamic areas within a city context is conducted by considering both social and temporal characteristics in conjuction to their context of functionality (i.e., the type of social activity, such as event venue, food spots, nightlife, etc.) either separately, or under a hybrid scheme, in contrast to the already existing works that see such characteristics solely as distinct modes. Additionally, here, we advance the way that the social and temporal attributes are incorporated to the hybrid mode in an effort to further improve the results obtained by the dynamic segmentation process, e.g., during the extraction of the social characteristics up to now efforts consider only the absolute number of times that a place has been visited from a user by neglecting in this way the overall activity patterns of a user. Finally, to better establish the suitability of the proposed characteristics, we also propose the DYNAMICITY platform which offers an insightful view on the different latent city areas and their unique characteristics.

## 3. Dynamic City Segmentation

This section outlines the problem addressed, which involves the segmentation of a given city into dynamic areas, provides the necessary background, and then presents the proposed methodology.

### *3.1. Problem Definition and Background*

It is a commonplace, municipalities, to divide a city (or a geographic region) into static areas based on certain characteristics. Nevertheless, cities and residential geographic regions change dynamically over time and behave as living organisms. An area can be born, grow, merge, split, and die over time. To this end, cities and geographic regions are shaped and evolved dynamically over time based on certain characteristics (social, temporal, and functional). Consequently, a *dynamic area* can be defines as 'a geographic region that is shaped based on certain characteristics and can grow, merge, split, born, or even die through time'. The division of a city into *dynamic areas* could be beneficial to researchers, citizens, authorities, and companies to understand better how a city functions and its needs, to gain useful insights, and finally to even help in decision making. For instance, an advertising agency could leverage such insights to conclude to the most qualified area to perform an event considering also the most appropriate day of week and time of day by exploiting areas' temporal characteristics in an effort to maximize the chance to conduct a successful event.

The main research problem we address in this work is the segmentation of a given geographic region (here, we consider a given city) into coherent dynamic areas by leveraging and analyzing the geo-tagged activities of social media users, and in particular, their check-ins to various POIs. The concept of a check-in and a POI are, thus, principal in this problem, and are defined as follows.

**Definition 1 (Point of Interest).** *A Point of Interest (POI) is a place which is characterized by a specific functionality/use. Each POI $v^i$ is represented as a tuple comprising its set of geo-coordinates $g^i$ (latitude and longitude) and category $ct^i$ (e.g., café, office, restaurant, etc.) $v_i =< g^i, ct^i >$.*

**Definition 2 (Check-in).** *A check-in is an action performed by a user to self-report her position to a given POI. Each check-in $c_i$ is represented as a tuple comprising the user who checked-in $u^i$, the POI $v^i$ and timestamp of the action $t^i$, $c_i =< u^i, v^i, t^i >$.*

Hence, the *Dynamic City Segmentation Problem* can be defined as follows.

**Problem 1 (*Dynamic City Segmentation*).**

***Given:*** *a city, a set of POIs within its span, and a set of check-ins into them,*

***Identify:*** *a set of areas that are coherent in terms of the users' visiting patterns with respect to their enclosed POIs and their characteristics, and*

***Extract:*** *profiles (descriptions) for each area that assist their interpretation and their further leveraging in applications.*

In this work, we examine existing approaches for the Dynamic City Segmentation problem, and further extend them to address the problem in a more global manner, by jointly taking into consideration different forces that shape urban mobility and POI visiting. We extract features that characterize POIs based on the observed activities within their context and then, the city segmentation is implemented as a POI clustering process. In general, the goal of Cluster Analysis is to split a set of objects into groups so that objects belonging to the same group are as similar as possible and as dissimilar as possible with objects of different groups. Therefore, the clustering of POIs has been selected as a well-suited approach for correlating POIs that are similar in terms of their various characteristics, and then each cluster is mapped to a distinct area spanning the convex hull of its assigned POIs. Table 1 summarizes the main notation followed in this paper.

Table 1: Notation

| ***Symbol*** | ***Explanation*** |
|---|---|
| $V = \{v_1, ..., v_{n_V}\}$ | Set of POIs |
| $U = \{u_1, ..., u_{n_U}\}$ | Set of Users |
| $C = \{c_1, ..., c_{n_c}\}$ | Set of Check-ins |
| $Ct = \{ct_1, ..., ct_{n_{ct}}\}$ | Set of Categories that POIs can belong to |
| $d(i, j)$ | Geographic distance of POI $v_i$ from POI $v_j$ |
| $sim(i, j)$ | Similarity between POIs $v_i$, $v_j$ |
| $k_{min}/k_{max}$ | Minimum/Maximum number of clusters |
| $A = (a_{i,j})_{i,j=1,...,n_V}$ | Affinity Matrix |
| $G(A)$ | Graph derived from the Affinity Matrix |
| $N_i = \{n_1, ..., n_m\}$, $n_j \in V$ | $m$ nearest neighbors of POI $v_i$ |

### *3.1.1. POI Spectral Clustering Algorithm*

Spectral clustering techniques make use of the spectrum of the affinity matrix to reduce the dimensionality of the dataset before applying clustering. The Affinity Matrix is provided as input to the algorithm and offers a quantitative assessment of the pairwise similarity of the datasets instances.

The Dynamic City Segmentation problem has been addressed before by using spectral clustering for separating POIs into groups [3]. In this approach, the proposed algorithm requires as input the set of POIs $V$ (as in Definition 1), the affinity matrix $A = (a_{i,j})$ for $V$, parameters $k_{min}$ and $k_{max}$ corresponding to the minimum and maximum desired number of clusters - areas, and parameter $\tau$, which controls the maximum area that a given cluster can span. Each POI is considered as a vertex in the graph $G(A)$, in which two POIs are connected with an edge when their similarity is greater than zero and the edge's weight represents their similarity. Affinity matrix contains the weights of graph $G(A)$, and its construction process will be elaborated in the next paragraph.

*Affinity Matrix Construction.* The information conveyed in the affinity matrix obviously impacts the outcome of the POI clustering algorithm (presented above); therefore it is of great significance to define appropriate POI similarity measures. In the earlier approach of [3], the proposed affinity matrix formulation considered both the POIs' feature representation and their spatial proximity (in order to build geographically cohesive areas). In specific, the use of *social activity* features was proposed for the representation of POIs (referred to as *social profiles* in this work and defined in Definition 4), but as it will be discussed in the next section, in general any type of POI vector-based profile can be used instead. Next, we provide the definition of the POI affinity matrix according to [3] for the readers' convenience.

**Definition 3 (POI affinity matrix).** *Given a set of $n_V$ POIs $V$, a geographical distance function $d(\cdot, \cdot)$, a POI similarity function $sim(\cdot, \cdot)$ and a number of nearest neighbors $m$, the POI affinity matrix $A = (a_{i,j})_{i,j=1,...,n_V}$ can be defined as follows:*

1. *For each POI $v$, we calculate the $m$ nearest neighbors of $v$, according to $d(v, \cdot)$ and assign them to set $N_m(v)$.*

2. *Then, each element of the affinity matrix $a_{i,j}$ is assigned the similarity between POIs $v_i, v_j \in V$ as follows:*

$$a_{i,j} = \begin{cases} sim(i,j) + \alpha & \textit{if } i \in N_m(j) \\ 0 & \textit{if otherwise} \end{cases}$$

*where $\alpha$, a small constant, ensuring connectivity of a given POI with all its neighbors.*

The above definition is generic in the sense that it allows the selection of the desired similarity and geographical distance functions.

### *3.2. Methodology*

In this section we present the proposed methodology for the segmentation of a city into dynamic areas based on the activity of social media users. Based on the POI Spectral Clustering Algorithm (Section 3.1.1), we propose the use of alternative POI profiles that represent different forces driving the dynamic area formation process. We also present a meta-analysis process that can be followed after the discovery of the dynamic areas for revealing and interpreting their latent characteristics. The proposed methodology is summarized in Algorithm 1.

**Algorithm 1:** City Segmentation in Dynamic Areas

```
Data: Check-in history
Result: Dynamic Areas, Dynamic Areas Profiles
Select representation type of POIs:
    (I) Represent each POI as a "bag of users”
    (II) Represent each POI with an hourly activity vector
    (III) Represent each POI with a “hybrid” vector which combines different types of features
Create POI profiles based on the selected type of representation
Find the $k$-closest neighbors per POI
Create POI Affinity Matrix
Apply POI Spectral Clustering Algorithm
Extract dynamic areas profiles
```

In the description of the algorithm, the first step is to build the POI representations (i.e., profiles) based on the users' check-ins at various places within the city. We consider three sets of features for the POI profiling:

- **SOCIAL** representation(ex: Table 2a), which is extracted based on the visiting patterns of social media users in city POIs, where each POI is represented as a “bag of users”.
- **TEMPORAL** representation (ex: Table 2b), which is based on the temporal activity observed on each POI. Thus, each POI is represented by two temporal activity profiles, which are essentially 24-dimensional vectors. The first represents the POI hourly activity during week days ($d^0..d^{23}$), and the second during weekends ($w^0..w^{23}$).
- **HYBRID** representation(ex: Table 2c), which integrates features of its social and temporal profile with its context of functionality (such as food, profession, education, etc.).

The SOCIAL profile and a simpler version of the TEMPORAL profile have already been explored in previous works. Since in real life different kinds of ‘forces’ impact cities and their areas, here we propose the HYBRID POI representation as an approach to shape areas by jointly leveraging the variety of different features that can be inferred from social media users' geotagged activities. In Table 2, we present examples of the SOCIAL, TEMPORAL and HYBRID representations of a POI. The proposed profiles (and corresponding types of affinity matrices) are presented in detail in Section 3.2.1.

(a) Social Representation of a POI

| $u_0$ | $u_1$ | $u_2$ | ... | $u_{n_U}$ |
|---|---|---|---|---|
| 0 | 5 | 10 | ... | 8 |

(b) Temporal Representation of a POI

| $d^0$ | $d^1$ | $d^2$ | ... | $d^{23}$ | $w^0$ | $w^1$ | $w^2$ | ... | $w^{23}$ |
|---|---|---|---|---|---|---|---|---|---|
| 28 | 18 | 5 | ... | 39 | 55 | 40 | 28 | ... | 125 |

(c) Hybrid Representation of a POI

| $u_0$ | $u_1$ | $u_2$ | ... | $u_{n_U}$ | $d^0$ | $d^1$ | $d^2$ | ... | $d^{23}$ | $w^0$ | $w^1$ | $w^2$ | ... | $w^{23}$ | $ct_0$ | $ct_1$ | ... | $ct_{n_{ct}}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 5 | 10 | ... | 8 | 28 | 18 | 5 | ... | 39 | 55 | 40 | 28 | ... | 125 | 0 | 1 | ... | 0 |

Table 2: Examples of social/temporal/hybrid representation of POIs.

### *3.2.1. POI Profiling*

As described in Section 3.1.1, the POI Spectral Clustering Algorithm requires the provision of an affinity matrix that encodes the POIs' pairwise similarities. Next, we present three different types of POI representation profiles that consider the SOCIAL, TEMPORAL and HYBRID features, respectively, and provide the corresponding POI similarity functions that employ them.

*Social POI Profile.* The main factor that shapes the different areas of a city is the people that live and act within its span. The social activity that takes place around the various POIs plays an important role in the formulation and definition of a city's latent areas. Under this concept, a POI can be represented by its social profile, as below.

**Definition 4 (Social POI profile).** *Given a set $U$ of $n_U$ users, a set of POIs $V$, and a set $C$ containing check-ins of users in $U$ to POIs in $V$, each POI $v_j$ is treated as a* ***"bag of users"****. Thus, it is represented as an $n_U$-dimensional vector, where the i-th component holds the number of times user $u_i$ checked-in to $v_j$ (the Social POI profile).*

The definition given above was proposed in [3] and assigns the absolute number of check-ins for each user to the POI representation vector. The intuition behind this type of profile is that similar POIs are visited by the same people. However, by simply considering the number of check-ins made by each user to each given POI, we do not account for the differences that exist in the users' overall check-in frequency. For instance, a certain restaurant may be the most visited place for two users $u_i$ and $u_j$, but $u_i$ might check-in at a much sparser rate than $u_j$. This is mainly indicative of the engagement level of each user with the application, whereas, in reality, this POI might be the favorite restaurant of both users $u_i$ and $u_j$. Furthermore, there are also users that present excessive (spam-like) activity in social media, such as very large numbers of (probably fake) check-ins in the case of social media with geolocation capabilities, therefore the POI profiling approach described in Definition 4 would affect the resulted similarity dramatically. To alleviate this effect we apply a feature normalization process on the POI profiles and derive the normalized social POI profiles, defined next.

**Definition 5 (Normalized Social POI profile).** *Considering each POI as a document and each user as a term of a document, we transform each Social POI profile into the Normalized Social POI profile, $v^S$, by calculating $tf - idf$ weights per term (user) for each document (POI).*

As described in Definition 5, we use the $tf - idf$ (term frequency inverse document frequency) weighting function for assigning values (weights) for each element (user) in the POI representation vectors. Modeling each POI as a document, each unique user as a document's term and the collection of POIs as the corpus. Hence,

**Definition 6 (Social similarity).** *The social similarity $sim(i, j)$ between two POIs $v_i$ and $v_j$ is defined as the cosine similarity of the vectors of their normalized social profiles $v_i^S$, $v_j^S$.*

*Temporal POI Profile.* An important feature of city areas is the variation of the level of activity within their span in terms of time (e.g., day of week, time of day). Typically, areas tend to exhibit some sort of periodicity in terms of the peoples visit patterns: for instance, areas with a high proportion of entertainment spots will be probably mostly visited during weekends or in the afternoon, whereas shopping areas will be mainly visited during the day time. Rosler and Liebig [4] proposed a POI profile generation process based on the temporal activity observed on them. Here, we try to take the concept of temporal POI profiling one step further and represent each POI with a 48-dimensional vector, which separates the hourly check-in activity that took place during the weekdays from check-ins that happened on weekends. The temporal profile of a given POI can then be defined as follows.

**Definition 7 (Temporal POI profile).** *Given a set $U$ of $n_U$ users and a set $C$ that contains check-ins, each POI $v_j$ can be represented by two vectors $dv_j$ and $wv_j$, with the first vector representing the* ***hourly***

Table 3: Similarity of POIs $v_1$, $v_2$ based on their activity shifts $q_1$, $q_2$ from $t$ to $t+m$

| $aff(q_1, q_2)$ | | $q_1$ | | |
|---|---|---|---|---|
| | | Increase | No Change | Decrease |
| $q_2$ | Increase | 1 | 0.5 | 0 |
| | No Change | 0.5 | 1 | 0.5 |
| | Decrease | 0 | 0.5 | 1 |

***activity*** *on the POI during* ***weekdays****, and the second the* ***hourly activity*** *on* ***weekends****. Each vector has 24 dimensions, with each dimension representing an hour of the day and containing the total number of check-ins made to this POI on the corresponding hour during weekdays and weekends, accordingly. The concatenation of these two vectors is the Temporal POI profile* $v_j^T$.

The similarity between two POIs based on their temporal profiles can be derived by comparing their hourly activity distribution "shapes", similarly to [19]. In particular, if POIs $v_i$ and $v_j$ are considered to be similar, then the change of their activity from hour $t$ to hour $t+m$ should also be similar in their weekday and weekend temporal profiles, for all possible values of $t$ and $m$. More precisely, the activity change for POI $v_i$ from $dv_i^t$ to $dv_i^{t+m}$ and $wv_i^t$ to $wv_i^{t+m}$ should be similar to the activity change of POI $v_j$ from $dv_j^t$ to $dv_j^{t+m}$ and $wv_j^t$ to $wv_j^{t+m}$. The shift $q(v_i^t, v_i^{t+m})$ of the activity for $v_i$ from $t$ to $t+m$ can have three discrete tags:

- "increase" if $v_i^t < v_i^{t+m}$,
- "decrease" if $v_i^t > v_i^{t+m}$,
- "no change" in case $v_i^t = v_i^{t+m}$.

Subsequently, the similarity between POIs $v_i$ and $v_j$ based on their activity shifts from $t$ to $t+m$ is represented as $aff(q(v_i^t, v_i^{t+m}), q(v_j^t, v_j^{t+m}))$ and defined as in Table 3.

**Definition 8 (Temporal similarity).** *Given two POIs* $v_i$, $v_j$ *and their temporal profiles* $v_i^T$, $v_j^T$, *their temporal similarity can be defined as follows:*

$$sim_{temp}(i,j) \quad = \quad \frac{\sum\limits_{t<t'} aff(q(dv_i^t, dv_i^{t'}), q(dv_j^t, dv_j^{t'}))}{NC} \quad + \quad \frac{\sum\limits_{t<t'} aff(q(dw_i^t, dw_i^{t'}), q(dw_j^t, dw_j^{t'}))}{NC}$$

where $NC$ the number of all possible comparisons.

*Hybrid POI Profile.* Rather than considering separately the effects of social and temporal POI profiling to reveal dynamic city areas, we propose an hybrid approach that assumes that city areas are characterized and formulated based on their social and temporal activity characteristics, as well as from the type of POIs they contain. Thus, a given POI can be characterized based on a *hybrid* profile that incorporates its social and temporal profiles (as given in Definition 5 and 7), as well as its context of functionality. For example, in Fourquare, the functionality of a POI is defined in terms of a predefined set of categories – types, such as: food spots, professional places, nightlife spots, event venues, etc. Throughout the paper, we use this concept of categories for describing the POI context of functionality, although the approach can be easily extended to other types of such information.

Having a set $V$ containing $n_V$ POIs, a set $Ct$ is defined that contains the $n_{ct}$ different categories a POI can belong to. Hence, each POI can be represented as a vector of $n_{ct}$ dimensions (i.e., the *category profile*) which has zero values in all dimensions except of those corresponding to categories the POI belongs to, which have a value of 1 (e.g., if the $l$th vector dimension is 1, then the POI belongs to category $l$). In several scenarios this vector has only one non-zero dimension, under the constraint that a POI can belong only to a single category. However this is not necessarily always the case: e.g., a given POI can be characterized by the music venue and food categories or another one by the art and shop categories. Moreover, in other cases, instead of the fixed-category set approach of describing the POI context of functionality, different kinds of information can be used such as tags assigned on POIs by users to characterize them in free form (in this case a given POI may have several tags). The proposed hybrid POI profile is then defined as follows:

**Definition 9 (Hybrid POI profile).** *The Hybrid POI profile of POI $v_i$ is defined as the vector $v_i^H$ with $(n_U + n_t + n_{ct})$ dimensions that comprises the POI's normalized social, temporal, and category profile. The Hybrid POI profile can be enriched with more features in case they are available.*

Therefore we can consider that two POIs are similar if they have similar hybrid profiles, meaning that they have similar social and temporal features, as well as categories. Then, hybrid similarity between POIs $v_i, v_j \in V$ is calculated based on their hybrid profiles $v_i^H$, $v_j^H$ using the cosine similarity measure.

**Definition 10 (Hybrid similarity).** *Given two POIs $v_i$ and $v_j$ and their hybrid profiles, the hybrid similarity between them is given by applying the cosine similarity on their hybrid vectors $v_i^H$ and $v_j^H$ and after weighting the features based on the contribution we wish each feature to have in the similarity calculation.*

The weights that represent the impact of the different types of feature in the calculation process are empirically set based on the application scenario (e.g., emphasis on user mobility patterns in the formulation of the dynamic areas).

*3.2.2. Dynamic Area Profiling*

The last step of the proposed algorithm is the *dynamic area profiling*, where we represent each area in terms of two types of features: the *composition*-based and *activity*-based features. Composition-based features include:

- The **Size** of each extracted area in $km^2$.
- The **Density** of each extracted area in POIs per $km^2$.
- The **Number of POIs** the area contains.
- The **POI Category Distribution** in each given area $i$, which is a $n_{ct}$-dimensional vector over all $n_{ct}$ available categories, with dimension $m$ holding the value:

  $$vc_{i,m} = \frac{vn_{i,m}}{\sum_{j=1}^{n_{ct}} vn_{i,j}}$$

  where $vn_{i,m}$ is the number of POIs in area $i$ that belong to category $m$.
- The **Check-in Category Distribution** for each area $i$, which is an $n_{ct}$-dimensional vector, with the dimension corresponding to category $m$ having the value:

  $$cc_{i,m} = \frac{tc_{i,m}}{\sum_{j=1}^{n_{ct}} tc_{i,j}}$$

  where $tc_{i,m}$ is the number of check-ins to POIs in area $i$ that belong to category $m$.

On the other hand, activity-based features contain characteristics of the areas in relation to activity observed within their span across time. More precisely, the following two features are considered:

- The **weekday temporal profile** which is represented by a 24-dimensional vector, with each dimension corresponding to an hour of the day. Activity within area $l$ in the $i$-th hour is described by:

$$wd_{i,l} = \frac{wdc_{i,l}}{\sum_{h=1}^{24} wdc_{h,l}}$$

  where $c_{i,l}$ is the number of check-ins performed during weekdays in the $i$-th hour of the day in area $l$.

- The **weekend temporal profile** which is defined similarly for temporal activity observed during the weekends on each hour of the day (represented by $we_{i,l}$ for area $l$ and the $i$-th hour of day).

At last, dynamic area profiling can combine any of the above features, but for the unbiased comparison of areas within different cities, we propose the following representation which comprises features that are *normalized.*

**Definition 11 (Dynamic area activity profile).** *The dynamic area activity profile is a composite feature vector comprising the weekday temporal, weekend temporal, POI category distribution and check-in category distribution profiles. For instance the profile of area $t$ would be:*

$$AP_t =< wd_t, we_t, vc_t, cc_t >$$

**Definition 12 (Dynamic area similarity).** *The similarity of dynamic areas $i$ and $j$ is computed as the cosine similarity between their profiles $AP_i$ and $AP_j$.*

## 4. DynamiCITY: A City Clustering Social Media-driven Platform

This work aims at identifying urban dynamic coherent areas based on people's social and temporal activity within a city, as reported in social media, and more specifically in social media emphasizing location sharing, such as Foursquare. The implementation of the proposed methodology is enabled though the DynamiCITY platform, which also allows exploring a city in an engaging and insightful manner through the presentation of the discovered dynamic areas and their characteristics.

DynamiCITY consists of two main components:

- A backend component (Figure 1), which is responsible for collecting publicly available Foursquare check-ins within a predefined geographic region (in this case, a city). Furthermore, it is responsible for analyzing users' activity around POIs and for separating the geographic region into dynamic coherent areas.
- A frontend component, which is a Web-based application that visualizes the discovered city areas through interactive maps, diagrams, and analytics. The Web application can be leveraged from the local administration, citizens, as well as from visitors, in order to understand how the city is organized and functions.

## 5. Results

This section presents the results of our methodology applied on three city scenarios: New York, London, and San Francisco. Then, the results of the application of the proposed methodology are discussed, comparing the different segmentation types and identifying extracted areas with similar characteristics in different cities. The evolution of New York dynamic areas through time is also analyzed, identifying similarities and changes across the different time periods under investigation. Finally, a comparison between "static" New York areas, i.e., fixed neighborhood borders, and extracted dynamic areas (uncovered by DynamiCITY) is performed.

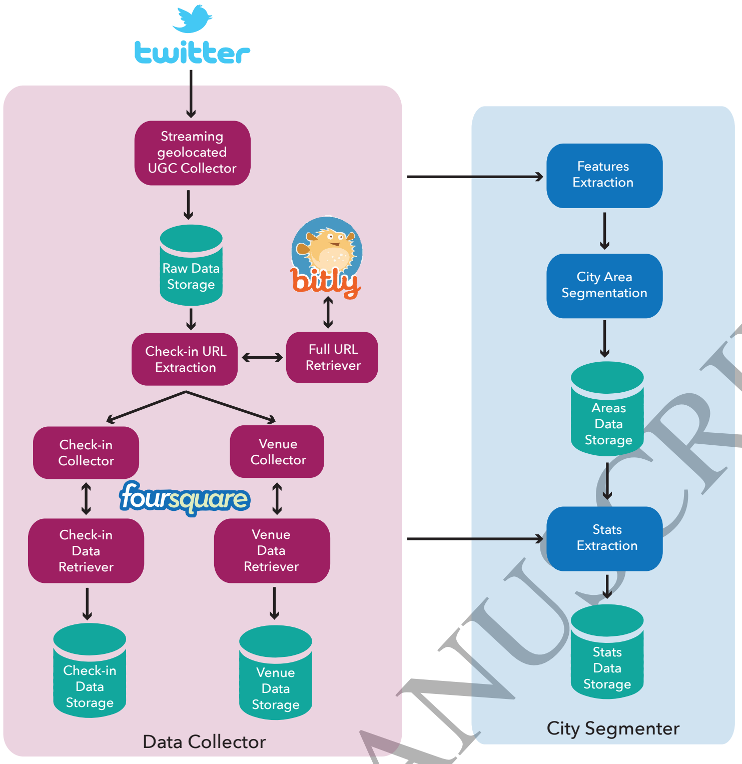


Figure 1: Architecture of the back-end component

### *5.1. Data Collection*

Foursquare is one of the most popular social media applications offering geolocation capabilities, with its community enumerating more than 50 million members – users that have made over 10 billion check-ins in numerous POIs, with millions of check-ins happening every day[1]. Foursquare allows its users to share their activity with their friends in other social networks such as Twitter. Due to Foursquare's privacy policy, which does not provide access to the real-time check-in stream, we used Twitter Streaming API to collect, indirectly, Foursquare check-ins that users publicly share to their Twitter profiles. Data collection was mainly focused on three different metropolitan cities: New York, San Francisco, and London.

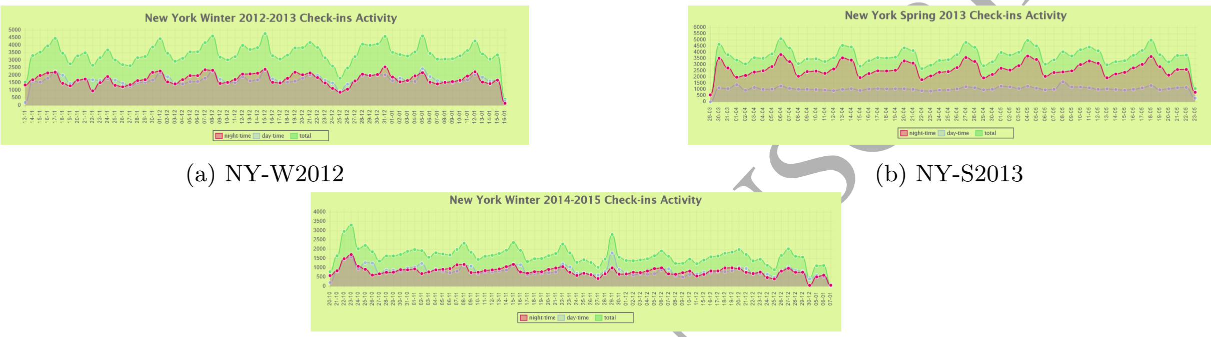
Table 4, summarizes some basic statistics about each dataset, including the time period for the data collection, the number of unique users and POIs in the dataset, and the total number of check-ins. For New York City three datasets were collected, reflecting user activities in different time periods, specifically: winter 2012-2013, spring 2013, and winter 2014-2015. The motivation for this long-term data collection was to demonstrate how DynamiCITY can assist studying the evolution of a city and its areas across time. It is notable that during winter, check-ins within New York were equally shared between day-time and night-time, whereas during spring only one fourth of check-ins were made during the day. Figures 2a–2c depict the check-in timelines for each New York dataset and time period of day.

An interesting observation based on the collected data, is that the day of week with the highest activity in terms of check-in is Saturday in all three cities, followed by Sunday in New York and San Francisco. However, in the case of London, Sunday appears to be one of the least active days of the week (only 13% of the total check-ins were made during Sundays). Moreover, 'Food' was the most dominant POI category in

[1] https://foursquare.com/about

Table 4: Dataset statistics.

| Dataset | City | Time Period | #Users | #Venues | #Check-ins |
|---|---|---|---|---|---|
| NY-W2012 | New York | 13/11/2012 - 16/01/2013 (Winter 2012-2013) | 26,183 | 34,829 | 221,869 |
| NY-S2013 | New York | 29/03/2013 - 23/05/2013 (Spring 2013) | 24,669 | 33,519 | 204,951 |
| NY-W2014 | New York | 20/10/2014 – 07/01/2015 (Winter 2014-2015) | 14,717 | 24,601 | 118,008 |
| SF-W | San Francisco | 20/10/2014 - 26/10/2015 (Winter 2014-2015) | 4,953 | 6,773 | 27,821 |
| LN-W | London | 20/10/2014 - 19/01/2015 (Winter 2014-2015) | 9,434 | 14,377 | 59,190 |



(a) NY-W2012

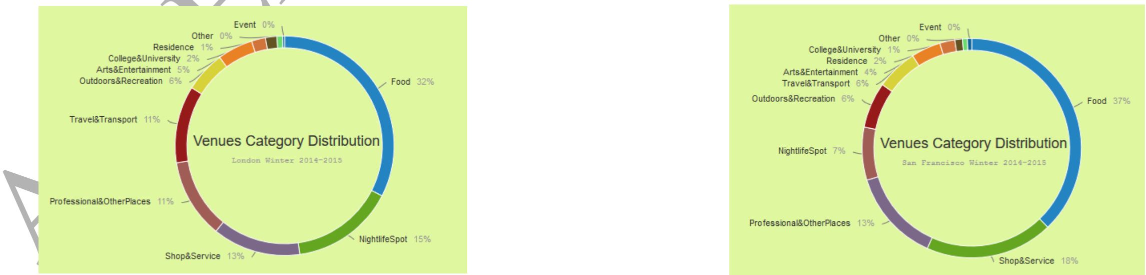
(b) NY-S2013

(c) NY-W2014

Figure 2: Daily number of check-ins in New York during different time periods

all datasets followed by 'Shops and Services', in most of the cases. Figure 3 indicatively presents the POI category distribution in London and San Francisco datasets.

Comparison of the three New York datasets revealed that out of 34,829 POIs of NY-W2012, ˜55% were active in NY-S2013, and only ˜32% in the NY-W2014 dataset. It is noteworthy that ˜4% of NY-W2012 POIs reappear in NY-W2014, but not in NY-S2013 (indicating probably places that operate or are popular during the winter). In terms of users, out of 26,183 users who checked-in during NY-W2012, ˜42% were also active in NY-S2013, but only 15.7% of these users appear in NY-W2014. Tables 5a–5b depict the common POIs and users, accordingly, per dataset.



(a) LN-W

(b) SF-W

Figure 3: Pie charts of POI category distribution in London (left chart) and San Francisco (right chart) datasets

(a) Common POIs

| | | *1* | *2* | *3* |
|---|---|---|---|---|
| 1 | NY-W2012 | 26,183 | | |
| 2 | NY-S2013 | 10,986 | 24,669 | |
| 3 | NY-W2014 | 11,079 | 11,361 | 24,601 |

(b) Common Users

| | | *1* | *2* | *3* |
|---|---|---|---|---|
| 1 | NY-W2012 | 34,829 | | |
| 2 | NY-S2013 | 19,028 | 33,519 | |
| 3 | NY-W2014 | 4,105 | 4,498 | 14,717 |

Table 5: Tables comparing NY datasets

### *5.2. City Segmentation Results and Comparison between Different Representation Types*

The methodology of Section 3 was applied on each of the collected datasets for extracting social, temporal, and hybrid dynamic areas. Specifically, the analysis of:

- NY-W2012 check-ins resulted in: 208 social areas, 81 temporal areas, and 77 hybrid areas.
- NY-S2013 check-ins resulted in: 185 social, 196 temporal, and 150 hybrid areas.
- NY-W2014 check-ins resulted in: 207 social, 178 temporal, and 168 hybrid areas.
- LN-W check-ins resulted in: 178 social, 140 temporal, and 191 hybrid areas.
- SF-W check-ins resulted in: 169 social, 193 temporal, and 183 hybrid areas.

Table 6 presents the most popular extracted hybrid area in each dataset in terms of the check-ins made within their span. Starting with LN-W, area 85[2] was the most popular one, in terms of absolute check-in number, corresponding to Harmondsworth village in London which contains Heathrow Airport. A total of 3,341 check-ins performed in Harmondsworth, while the most popular POIs were: London Heathrow Airport and Terminal 2 "The Queens Terminal". Next, in SF-W, the area that gathered the most check-ins corresponds to the south part of the South of Market neighborhood of San Francisco. On average, the area presented high activity on Thursdays and Saturdays. Next, in NY-W2012, area 61, corresponding to the Diamond District, emerged as the most popular one, and mostly active on Saturdays from 18:00 till 21:00. Likewise, in NY-S2013, the area with the most check-ins was again Diamond District but even though Saturday was still the day with the highest activity, there was a shift in the activity's peak to 21:00 till 00:00 (probably due to the spring period). Finally, in NY-W2014, area 156, which contains a part of Springfield Gardens, a neighborhood of Queens, was the most popular one. Kennedy International Airport, Terminal 5 and JetBlue Airways are some of the most popular POIs of the area.

Normalized Mutual Information (NMI) was used for the comparison of the different types of segmentation of a city into dynamic areas. NMI measures the similarity between two distributions and is often used in cluster analysis for evaluating the quality of the extracted clusters. Here, NMI is used for comparing the different types of city segmentation into social, temporal, hybrid, and geo-areas. Geo-area extraction was performed by considering only the spatial proximity of POIs.

NMI between the different types of clustering is expected to be relatively high as the key feature in the calculation of the affinity matrix is the POI spatial distribution. Tables 7a–7e present the NMI values between the different types of clustering per dataset. In general, it can be noticed that geo-areas are most similar to temporal, indicating that nearby places tend to have similar temporal profiles. Moreover, the dynamic areas with the lowest similarity with the resulted geo-areas are, in most cases, the hybrid. This indicates that the more available features we have for building POI profiles, the more interesting / unexpected the resulting clustering will be.

Based on the NMI value, the extracted hybrid clusters of NY-S2013 and of NY-W2014 seem to be equally affected by the temporal and social profiles of POIs. On the other hand, the hybrid clusters of NY-W2012

[2] From here on we will use numeric format for area references that correspond to the numbers that are used in DynamiCITY for the reader's convenience.

Table 6: Popular Hybrid Areas per dataset.

| dataset | #area | name | #check-ins | most popular POI | most popular categories |
|---|---|---|---|---|---|
| NYW-2012 | 61 | Diamond District | 11,747 | Time Square | 7.9% Arts & Entertainment |
| NY-S2013 | 96 | Diamond District | 6,172 | Times Square | 4.8% Arts & Entertainment |
| NY-W2014 | 156 | Springfield Gardens | 3,868 | J.F. Kennedy International Airport | 5.09% Travel & Transport |
| SF-W | 114 | South of Market | 729 | San Francisco Marriot Marquis | 7.14% Event |
| LN-W | 23 | Harmondsworth village | 3,341 | London Heathrow Airport | 4.4% Travel & Transport |

seemed to be more affected by the social POI profiles. The extracted hybrid areas of London and San Francisco are also equally affected by the temporal and social POI profiles.

Finally, NMI tends to be higher in all types of segmentation as the number of the collected check-ins decreases. Thus, more information about users' activity with respect to the various POIs results into more interesting segmentation.

### *5.3. Similar Areas between Different Cities*

This section compares the extracted hybrid areas between different cities. Motivated by the challenge of understanding cities in depth, we applied the methodology of [16] for the discovery of similar city areas between the different cities under investigation. To this end, we used the generated area profiles, and calculated the cosine similarity between areas in different cities. The next paragraphs present the pairs of cross-city areas that were found to have the highest similarity, and discuss their common characteristics.

Comparing the extracted hybrid areas of London and New York (NY-W2014), we noticed that the popular area of Oxford circus (area 42) in London is matched to (i.e., has the highest similarity with) Union Square-University Village (area 130) in New York. Both areas contain POIs related to Colleges & Universities and to Shops & Services, while they both have low activity on Thursdays. Hyde Park (area 98) in London is most similar to Bryant Park (area 103) in New York. The areas that contain airports in both cities also appear to have high similarity. More specifically, Heathrow (area 23) in London is matched with La Guardia (area 158) in New York. The second most similar area to Heathrow in New York is the area of Springfield Gardens that contains the John F. Kennedy International Airport, followed by Essex County, which contains the Newark Liberty International Airport. Furthermore, the hipster area of Dalston[3] (area 14) in London is matched to Noho[4] (area 78) in New York. These areas mostly contain POIs related to Arts & Entertainment, Nightlife spots, and Shops & Services, while they both appear to be mostly active on Saturdays.

Comparing the extracted hybrid areas of London and San Francisco, Oxford circus (area 42) in London is matched to Nob hill[5] (area 110) in San Francisco. Again, these areas are characterized by POIs related to Colleges & Universities, as well as to Shops & Services, while in both cases the day of week with the lowest activity is Thursday. The city center of London (area 140) corresponds to Bayview District (area 11) in San Francisco, with both areas having several Event-related and Professional places. High similarity is

[3] http://www.dailymail.co.uk/travel/travel_news/article-2837916/The-hipster-guide-world-poor-inner-city-districts-achingly-cool-beards-optional.html

[4] http://www.businessinsider.com/maps-show-how-hipsters-have-taken-over-new-york-2014-10

[5] http://www.thebolditalic.com/articles/7182-this-map-of-sfs-hipster-neighborhoods-shows-how-much-the-city-has-changed

(a) NY-W2012

| | | *1* | *2* | *3* | *4* |
|---|---|---|---|---|---|
| 1 | social | 1 | | | |
| 2 | temporal | 0.832 | 1 | | |
| 3 | hybrid | 0.821 | 0.854 | 1 | |
| 4 | geo | 0.832 | 0.908 | 0.849 | 1 |

(b) NY-S2013

| | | *1* | *2* | *3* | *4* |
|---|---|---|---|---|---|
| 1 | social | 1 | | | |
| 2 | temporal | 0.878 | 1 | | |
| 3 | hybrid | 0.874 | 0.872 | 1 | |
| 4 | geo | 0.877 | 0.893 | 0.876 | 1 |

(c) NY-W2014

| | | *1* | *2* | *3* | *4* |
|---|---|---|---|---|---|
| 1 | social | 1 | | | |
| 2 | temporal | 0.887 | 1 | | |
| 3 | hybrid | 0.88 | 0.877 | 1 | |
| 4 | geo | 0.879 | 0.916 | 0.87 | 1 |

(d) LN-W

| | | *1* | *2* | *3* | *4* |
|---|---|---|---|---|---|
| 1 | social | 1 | | | |
| 2 | temporal | 0.887 | 1 | | |
| 3 | hybrid | 0.88 | 0.877 | 1 | |
| 4 | geo | 0.879 | 0.916 | 0.87 | 1 |

(e) SF-W

| | | *1* | *2* | *3* | *4* |
|---|---|---|---|---|---|
| 1 | social | 1 | | | |
| 2 | temporal | 0.908 | 1 | | |
| 3 | hybrid | 0.902 | 0.909 | 1 | |
| 4 | geo | 0.899 | 0.922 | 0.895 | 1 |

Table 7: NMI between different types of clusterings

also observed between Dalston (area 14) in London and Nob hill (area 9) in San Francisco, with both areas containing POIs belonging to the categories of Arts & Entertainment, Nightlife spots, and Shops & Services.

Further on, Union Square - University Village hybrid area (area 130) in New York is matched to Mission District (area 109) in San Francisco. Both areas contain places related to Colleges & Universities as well as Shop and Services. Union Square - University Village has also high similarity with other university areas in San Francisco, such as North of Panhandle (area 141) and Mission Street (area 178). Greenpoint (area 131) in New York is most similar to Bayview District (area 11), and they are both characterized by Event-related POIs. These areas are also mostly active on weekends. The hipster area of Noho (area 78) in New York is matched to areas North Beach (area 104) and Western Addition (area 91) in San Francisco. Finally, the most popular area (in terms of total check-in number) in San Francisco, South of Market (area 114), is most similar to Korea Town – Madison Square in New York, one of the most popular areas in New York (3rd in terms of check-ins).

The above analysis demonstrates the usefulness of DynamiCITY as a city exploration tool that can be used for offering area recommendations in new cities visited for the first time, based on areas on cities the user is familiar with. Moreover, area profiling and cross-city similarity can be used for transfering best practices across cities, from a city administation and/or urban planning perspective.

### *5.4. New York Areas Evolution through Time*

Under the assumption that cities and urban space usage change through time and that this is reflected in the digital footprints of social media users, we studied three different datasets for the city of New York that span different periods of time. To identify the evolution of the extracted areas through time, we leveraged the methodology that [20] proposed for detecting the evolution of clusters. Having two sets of clusters extracted for sucessive time spans, there are six different scenarios about their evolution from the earliest time span to the latest. A cluster area can, thus: grow (i.e., more POIs become attached to it overall), contract (i.e., the number of POIs it contains decreases), merge with another cluster, split in two or more clusters, be ”born” (i.e., most of its POIs appear for the first time in the latest time span), or “die” (i.e., most of the POIs appearing in the earliest time span disappear in the latest one).

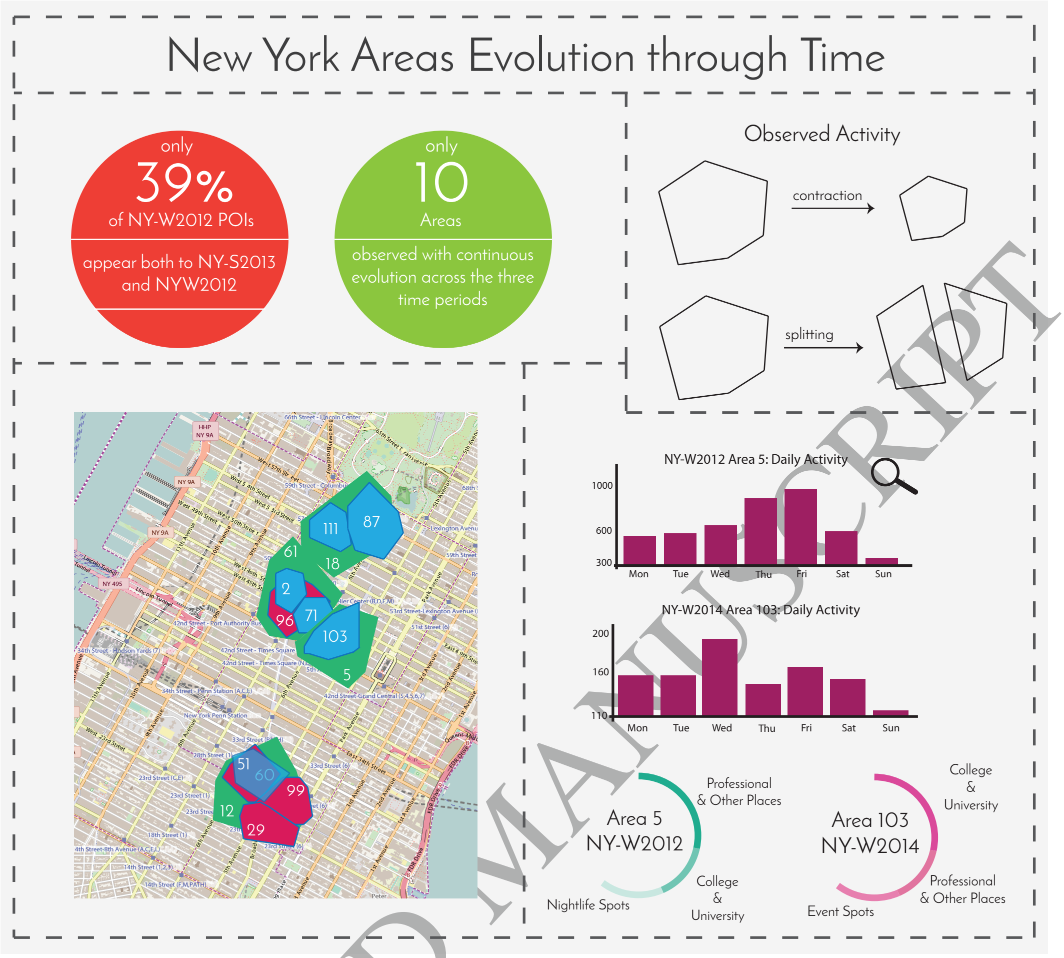
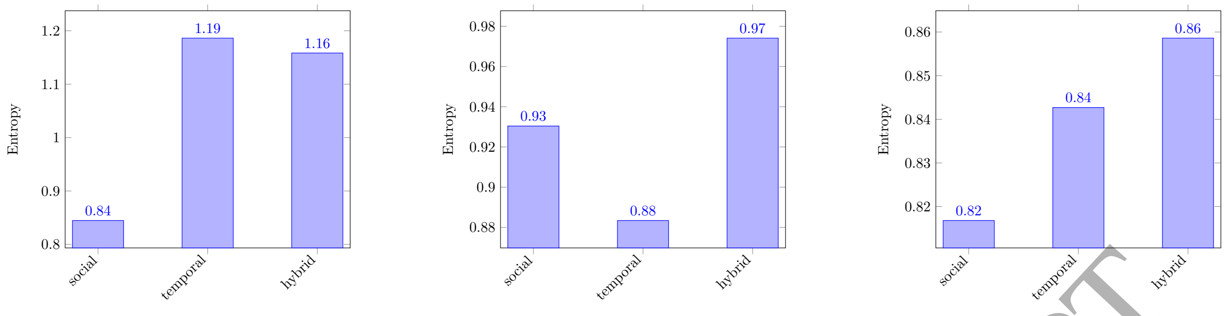


Figure 4: Infograpic depicting the evolution of New York dynamic areas through time. Areas corresponding to NY-W2012, NY-S2013 and NY-W2014 in the map are depicted in green, pink and blue, respectively

In the New York datasets, there are 77, 150, and 162 hybrid areas in NY-W2012, NY-S2013, and NY-S2013, respectively. It should be noted here that the shaping of areas is not only attributed to users' activities, but also to other factors such as changes in the Fourquare application itself. Specifically, in May of 2014, Foursquare divided its core activity into two applications: Foursquare and Swarm, the first being more focused on recommendations, whereas the second on making and sharing check-ins. This event caused Foursquare to lose several users which led to a decrease in the total number of check-ins. The derived hybrid areas in NY-W2012 cover wider geographic areas and contain a large number of POIs. Thus, the area evolutions in NY-S2013 and NY-W2014 are mainly expressed as splits or contractions.

As already mentioned, check-ins collected in different time spans may not refer to the same set of POIs. For instance, it was observed that only 39% of the total POIs appear both in NY-S2013 and NY-W2012. Thus, there were only a few areas that they had an observable evolution starting in NY-W2012, continuing to NY-S2013, and lasting until NY-W2014. Precisely, there were 10 areas with a continuous evolution across the three time periods, either by their split into smaller areas or by their contraction to areas with fewer POIs. Figure 4 depicts some of these areas. Area 12 (Chelsea) in NY-W2012, for example, splits into areas 99, 60, and 29 in NY-S2013, and then area 60 contracts to area 51 in NY-W2014. Likewise, area 61 (Diamond District) in NY-W2012, contracts to area 96 (NY-S2013), and then splits into areas 2 and 71 (NY-W2014). This area (represented by area 61 in NY-W2012 and area 96 in NY-S2013) is mostly characterized by POIs that are related to Arts and Entertainment, while it is most active on Saturdays. The two areas into which

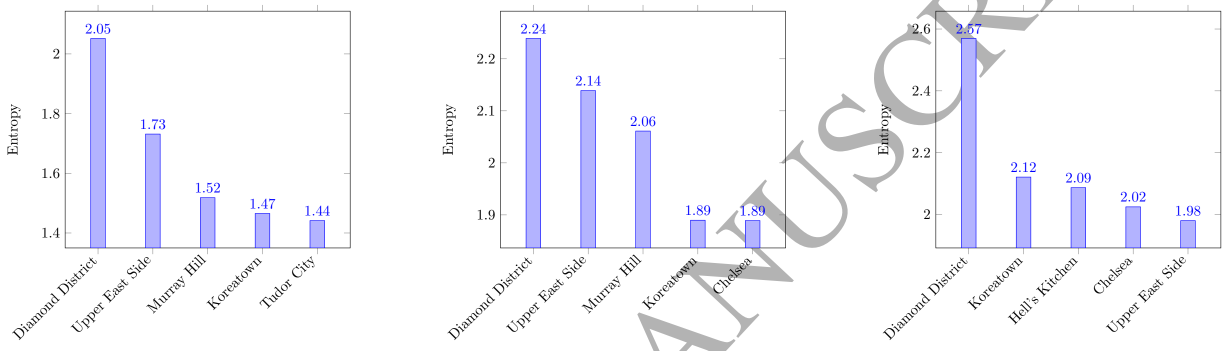


(a) NY-W2012

(b) NY-S2013

(c) NY-W2014

Figure 5: Mean entropy of dynamic areas vs static



(a) NY-W2012

(b) NY-S2013

(c) NY-W2014

Figure 6: Static areas with the highest entropy

it splits in NY-W2014 however, seem to have different characteristics. Specifically, area 2 (NY-W2014) is characterized mostly by POIs related to Arts and Entertainment and Saturdays and Sundays are the days with the highest observable activity. On the other hand, area 71 (NY-W2014) is mostly characterized by POIs related to Events, mostly visited on Tuesdays and Fridays. In NY-W2012 and NY-S2013, areas 61 and 96, respectively, did not contain event-related POIs and their highest activity was observed during the weekend. Thus, the split of area 96 (NY-S2013) into the two areas in NY-W2014, was probably a result of the shift of the category distribution of POIs the area contained, and due to the change observed in the temporal pattern of POI visiting in area 71.

We identified five NYC dynamic areas which evolved from NY-W2012 to NY-W2014, but which were dissolved in NY-S2013. These areas seem to retain their hybrid profiles through the wintertime but not through springtime. For instance, area 5 (Diamond District) in NY-W2012 contracts to area 103 in NY-W2013, with both areas being characterized by POIs related to College & Universities, as well as Professional & Other places. The hourly profiles of these areas exhibit similar activity, but area 5 (NY-W2012) appears to have higher activity on Thursdays and Fridays, while area 103 on Wednesdays and Fridays. Additionally, area 18 (San Juan Hill) NY-W2012 splits into areas 111 and 87 NY-W2014. These areas are mostly described by venues related to Arts & Entertainment but they appear to have quite dissimilar temporal profiles.

### *5.5. Static vs. Dynamic Areas in New York*

Intuitively, a segmentation of a city into dynamic areas based on specified features such as its social and temporal activity as well as the category of POIs it includes, is expected to reflect more accurately the functional use of city spaces compared to a static city segmentation. A question that arises in relation to the segmentation of a city into dynamic areas is how much the derived areas deviate from a typical static city devision. A static city segmentation is typically established by the municipality, and is mainly based on demographics, Here, focusing once again on New York, we adopt a static division of the city in 289 distinct areas (neighborhoods), provided by the OpenStreetMap API[6]. To evaluate the diversity of the derived dynamic areas, we calculate the entropy of each derived dynamic area in terms of the number of different static areas the POIs it contains belong to.

Figures 5a–5c depict the mean entropy of the extracted dynamic areas per type of segmentation and dataset, in bar charts. It appears that social areas have the highest similarity with static areas during winter, while during spring people tend to socialize in wider geographic regions and the diversity of the extracted social areas is increased. Almost in all datasets, hybrid areas appear to be the most dissimilar with the static areas of New York. In the case of NY-W2014, the extracted dynamic areas appear to be less diverse compared to the other datasets. Subsequently, the more information we have about the activity around POIs the more interesting the resulted segmentation of the city will be (i.e., the effect of user activities on dynamic area shaping will be significant, and thus not overshadowed by the geographic distance effect).

Finally, figures 6a–6c present the static areas with the highest entropy in terms of the number of dynamic areas their POIs belong to. Diamond District is the static area with the highest entropy in all datasets, containing 11 hybrid areas in NY-W2012, 15 in NY-S013, and 21 in NY-W2014. Upper East Side is also a static area with high entropy in all datasets, containing POIs from 9, 8, and 11 hybrid areas in NY-W2012, NY-S2013, and NY-W2014, respectively. Another static area with high entropy is Koreatown that contains POIs from 7, 6, and 12 hybrid areas in NY-W2012, NY-S2013, and NY-W2014. These areas are characterized by numerous, diverse POIs and check-in activity, and thus the proposed segmentation methodology manages to identify smaller dynamic sub-areas within them that have different uses and characteristics.

## 6. Conclusions

In this work we examined how geolocated UGC from social media within a city can be harnessed to discover urban dynamics. A literature review on this subject was presented, while our research interest focused on the separation of a certain geographical region (e.g., a city) into coherent areas based on the geolocated activities of social media users. Initially, we relied on existing work to separate a city into social and temporal areas. Then, based on the assumption that areas of a city are formed due to a set of forces, we proposed a methodology for extracting dynamic areas (hybrid areas) shaped by users visiting patterns, the time of observed visits, and the type of the visited POIs *category* (e.g., food, entertainment, arts, etc.) and their spatial distribution. Then, we introduced the architecture of DynamiCITY, the platform that was developed for data analysis, dynamic areas extraction and their presentation along with statistics in relation to the extracted dynamic areas.

The analysis of five check-in datasets from Foursquare corresponding to the cities of London, San Francisco, and New York indicated that the extracted *hybrid dynamic areas* tend to be more diverse than *social dynamic areas* and *temporal dynamic areas*. Additionally, the evolution of the New York *hybrid dynamic areas* across time showed that, in most cases, the initial dynamic areas tended to become smaller in size with time, since the population of POIs, as well as the preferences of the users, did not remain stable, but varied in the different time-periods. Lastly, a comparison between the static and dynamic *hybrid areas* of New York suggested that usually the areas that present the highest activity in social media tend to have the largest variance in terms of functionality, visiting patterns, etc.

Future extension of this work could be focused on the exploitation of geographically tagged data from different social networks and for a broader time window. More specifically, we would like to collect data across all seasons in the course of one year and analyze them in order to detect differences on users' activity during the aforementioned period, as well as study how the dynamic areas are shaped. Combining geographically tagged content from different social media services will result in the increase of observations, and thus the analysis results will be more interesting and accurate. Additionally, the proposed platform could be extended

[6] https://wiki.openstreetmap.org/wiki/API

towards combining information about a city not only from social media, but also from sensors positioned within the city reporting information about air pollution, noise, traffic, etc., for offering citizens and local authorities a better, unified view about the city. Finally, the developed tool can be further enhanced by providing suggestions about areas that a citizen could visit based on her current visiting patterns in other cities.

### References

### References

[1] C. A. Davis Jr, G. L. Pappa, D. R. R. de Oliveira, F. de L Arcanjo, Inferring the location of twitter messages based on user relationships, Transactions in GIS 15 (6) (2011) 735–751.

[2] T. H. Silva, P. O. S. Vaz de Melo, J. M. Almeida, J. Salles, A. A. F. Loureiro, A comparison of foursquare and instagram to the study of city dynamics and urban social behavior, in: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, UrbComp '13, ACM, 2013, pp. 4:1–4:8.

[3] J. Cranshaw, R. Schwartz, J. I. Hong, N. M. Sadeh, The livehoods project: Utilizing social media to understand the dynamics of a city., in: ICWSM, 2012.

[4] R. Rösler, T. Liebig, Using Data from Location Based Social Networks for Urban Activity Clustering, Springer, 2013, pp. 55–72.

[5] A. Noulas, S. Scellato, C. Mascolo, M. Pontil, An empirical study of geographic user activity patterns in foursquare, in: In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, 2011, pp. 570–573.

[6] L. Ferrari, A. Rosi, M. Mamei, F. Zambonelli, Extracting urban patterns from location-based social networks, in: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, LBSN '11, ACM, 2011, pp. 9–16.

[7] X. Long, L. Jin, J. Joshi, Exploring trajectory-driven local geographic topics in foursquare, in: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, ACM, 2012, pp. 927–934.

[8] A. Noulas, S. Scellato, N. Lathia, C. Mascolo, A random walk around the city: New venue recommendation in location-based social networks, in: Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust, IEEE Computer Society, 2012, pp. 144–153.

[9] X. Long, J. Joshi, A hits-based poi recommendation algorithm for location-based social networks, in: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM '13, ACM, 2013, pp. 642–647.

[10] D. Yang, D. Zhang, Z. Yu, Z. Wang, A sentiment-enhanced personalized location recommendation system, in: Proceedings of the 24th ACM Conference on Hypertext and Social Media, HT '13, ACM, 2013, pp. 119–128.

[11] J. Capdevila, M. Arias, A. Arratia, Geosrs: A hybrid social recommender system for geolocated data, Information Systems 57 (2016) 111–128.

[12] M. Clements, P. Serdyukov, A. P. de Vries, M. J. Reinders, Using flickr geotags to predict user travel behaviour, in: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '10, ACM, 2010, pp. 851–852.

[13] J. Bao, Y. Zheng, M. F. Mokbel, Location-based and preference-aware recommendation using sparse geo-social networking data, in: Proceedings of the 20th International Conference on Advances in Geographic Information Systems, SIGSPATIAL '12, ACM, 2012, pp. 199–208.

[14] S. Van Canneyt, S. Schockaert, O. Van Laere, B. Dhoedt, Time-dependent recommendation of tourist attractions using flickr, in: BNAIC: Belgian/Netherlands Artificial Intelligence Conference, 2011.

[15] A. X. Zhang, A. Noulas, S. Scellato, C. Mascolo, Hoodsquare: Modeling and recommending neighborhoods in location-based social networks, in: 2013 International Conference on Social Computing (SocialCom), IEEE, 2013, pp. 69–74.

[16] G. Le Falher, A. Gionis, M. Mathioudakis, Where is the soho of rome? measures and algorithms for finding similar neighborhoods in cities, in: 9th AAAI Conference on Web and Social Media - ICWSM 2015, 2015.

[17] G. Rizzo, R. Meo, R. G. Pensa, G. Falcone, R. Troncy, Shaping city neighborhoods leveraging crowd sensors, Information Systems.

[18] V. Frias-Martinez, V. Soto, H. Hohwald, E. Frias-Martinez, Characterizing urban landscapes using geolocated tweets, in: Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust, SOCIALCOM-PASSAT '12, IEEE Computer Society, 2012, pp. 239–248.

[19] L. co Todorovski, B. Cestnik, M. Kline, Qualitative clustering of short time-series: A case study of firms reputation data, IDDM2002 (2002) 141.

[20] G. Palla, A.-L. Barabási, T. Vicsek, Quantifying social group evolution, Nature 446 (7136) (2007) 664–667.