Sensing Urban Social Geography Using Online Social Networking Data

Santi Phithakkitnukoon and Patrick Olivier
Culture Lab
School of Computing Science
Newcastle University, United Kingdom
santi@mit.edu, p.i.olivier@newcastle.ac.uk

Abstract

Growing pool of public-generated bits like online social networking data provides possibility to sense social dynamics in the urban space. In this position paper, we use a location-based online social networking data to sense geo-social activity and analyze the underlying social activity distribution of three different cities: London, Paris, and New York. We find a non-linear distribution of social activity, which follows the Power Law decay function. We perform inter-urban analysis based on social activity distribution and clustering. We believe that our study sheds new light on context-aware urban computing and social sensing.

Introduction

The recent explosion in deployment of pervasive systems like sensor and cellular networks, and the emergence of the user-generated information sharing systems like online social networks produce massive amount of data. This data encompasses actions of people, dynamics of cities, and how they evolve over time. It forms *digital footprints* (Girardin et al. 2008) that allow researchers to better understand human mobility (Gonzalez, Hidalgo, and Barabasi 2008), city dynamics (Reades et al. 2007), and socio-economic (Eagle, Macy, and Claxton 2010), and could ultimately help those who manage and live in urban areas to configure more liveable, sustainable, and efficient cities (Foth 2008).

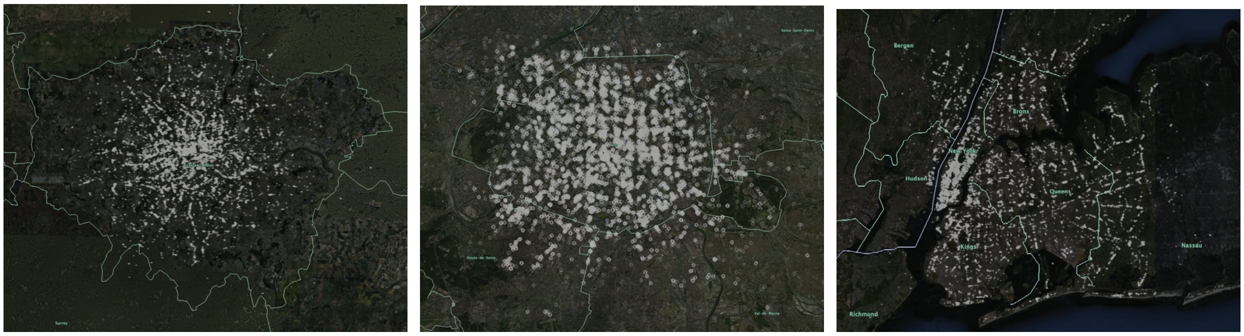
At the same time, these digital footprints create possibility of inferring social dynamics in the urban space, which in turn can provide information about how a city functions and can potentially influence many aspects of urban management by assisting local authorities, service providers, enterprises, and even citizens themselves to make more inform decisions and hence create a more efficient urban environment.

Data collected using pervasive systems has been used to sense physical dynamics such as face-to-face social network structure (Eagle, Pentland, and Lazer 2009), activity signature of the physical environment (Calabrese, Reades, and Ratti 2010), and even regional boundaries of a country (Ratti et al. 2010).

While most analyses of online social networking data have implications and applications only within the cyber world (e.g. (Leskovec 2008), (Cha et al. 2010)), here we are realizing the linkage between the cyber and physical worlds and utilizing this connection as the *crowdsensing* mechanism – particularly for socio-geographical aspect.

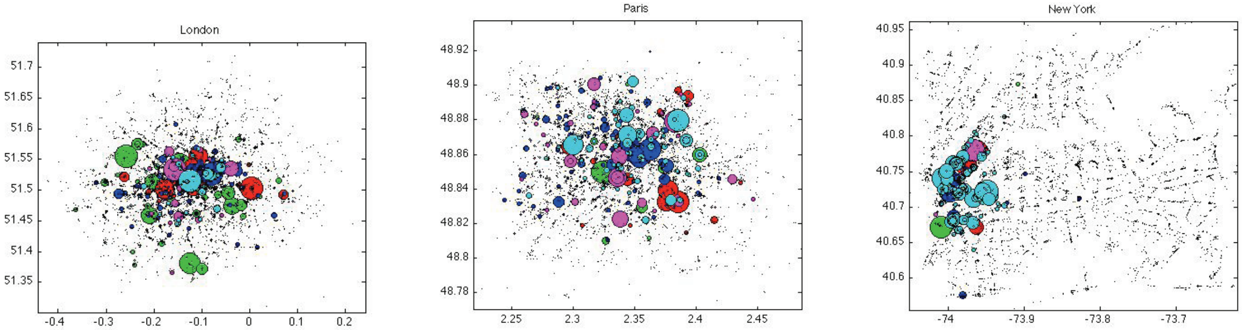
Online Social Networking Data

In this work, we use the data collected from a location-based online social network called *Foursquare* (Foursquare 2011). The 4sq network allows GPS-enabled mobile phone users to interact by communicating their whereabouts to their friend group. The user can *check in* at a *venue*, which can be a site of social encounter either located in a building (e.g. restaurant, bar, art gallery) or an open space (e.g. park). Its venue information provides dynamic geo-social aspect that distinguishes it from that of generic online social network e.g. Twitter, Facebook, Linkedin. Checking in at a venue can also earn points and rewards for the user. The typical reasons for checking-in include tracking friends, reporting a new venue, recommending a place, sharing experience, playing game (competing to become the ‘major’ of a place), and looking for nearby venues. The data acquired from 4sq thus provides us with the social interaction density and its geographical relation to the physical landscape. In this study, we analyze 800,00 data records from the 4sq network consisting of *checkins* and corresponding *venues* along with the venue’s categories from three different cities: London, Paris, and New York City, over about 16 months (March 2009 to July 2010). The data has been collected by a systematic crawl of the 4sq public search API (Bawa-Cavia 2010).This allows us to capture a snapshot of aggregated geo-social interaction and reveals how social activity is distributed in the urban space. Figure 1 shows the overall geographical density of social activity (checkins), which creates the *signature* of the three cities. With the venue’s category information along with the volume of checkins of each venue, we can capture a richer signature of the city as shown in Fig. 2 where each dot represents a venue, the dot’s size corresponds to the amount of checkins, and the color indicates different categories: green, red, blue, magenta, and cyan represent *Shop*, *Arts*, *Food*, *Park*, and *Nightlife*, respectively. Therefore the data allows individual as well as cross analysis of the urban space.



(a) London (b) Paris (c) New York

Figure 1: Geographical distribution of social interaction (*checkins*)



(a) London (b) Paris (c) New York

Figure 2: Geographical distribution and density of social activity

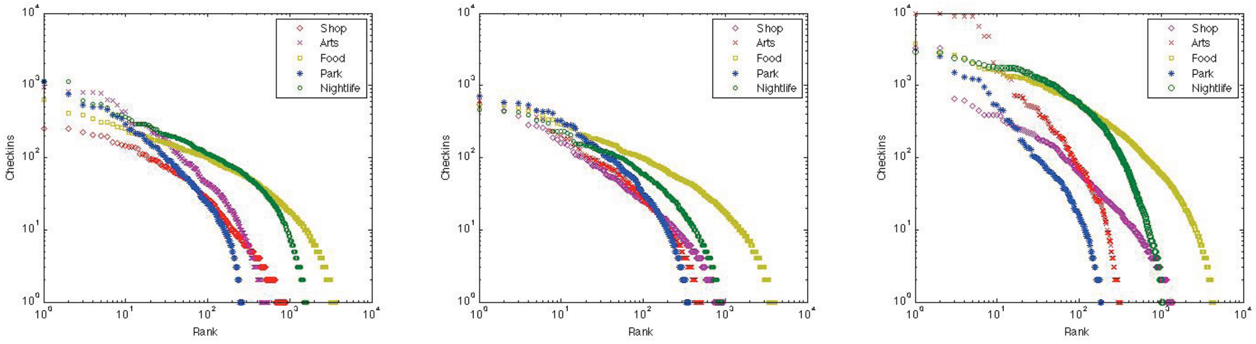
Social Activity Distributions

Clearly, one can observe a non-uniform distribution of social activity for each city. This different social activity volume across different areas suggests different *social hubs* of different activity categories. This motivates us to explore further in terms of how each social activity is distributed (statistically) and how social hubs are distributed geographically.

It turns out that social activity is distributed in a non-linear fashion as shown in Fig. 3. Each social activity appears to follows the *Power Law* distribution, which means that only a few places attract high social activity while the majority of places show very low social interaction – hence those few places represent the 'social hubs'. Note that plots in Fig. 3 are in log-log scales and each social activity distribution can be described by a power-law decay function $y = kx^{-\alpha}$ where y denotes checkins, k is a constant, and x represents the rank with the decay term α. From Fig. 3, Food and Nightlife appear to form the strongest activity hubs – most clustered.

To quantify this non-linear characteristic for further inter-urban analysis, we estimate the power-law decay term α according to the maximum-likelihood fitting methods with goodness-of-fit tests based on the Kolmogorov-Smirnov statistic and likelihood ratios (Clauset, Shalizi, and Newman 2009). The estimated values of α are plotted in Fig. 4 and it confirms our previous observation of Food and Nightlife being the top clustered social activities. The decay values also reveal a surprisingly low variation among social activities in New York City, which implies that the social interaction of each activity category is distributed in a similar fashion. Paris appears to also have this similar characteristic but with higher decay rates – stronger social hubs. Among the three cities, London seems to be different than others in the way that Food and Nightlife activity form much stronger social hubs than other social activity types – e.g. there are major restaurants and night clubs that seem to attract majority of people.

To get a sense of how *social centers* are distributed geographically, we cluster each social activity category based on the geo-location using the k-means clustering technique. With the number of clusters $k = 4$ (chosen arbitrarily), Fig. 5, 6, and 7 show resulting clusters (with centroids) of each social activity for London, Paris, and New York, respectively. This provides to some extent a geographical distribution of social centers of each social activity category. With our arbitrary choice of number of clusters, London's Park and Shop show a similar cluster distribution while Food and Nightlife are relatively correlated, and Arts is showing a lone different social clustering. For Paris, Arts' and Nightlife's social clusters are geographically distributed similarly while Food, Park, and Shop are sharing a similar pattern. NewYork, on the other hand, is having different so-



(a) London (b) Paris (c) New York

Figure 3: Power-law distribution of each activity category

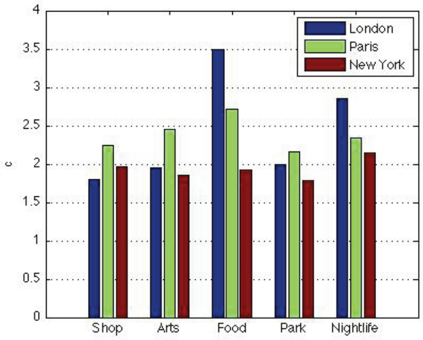


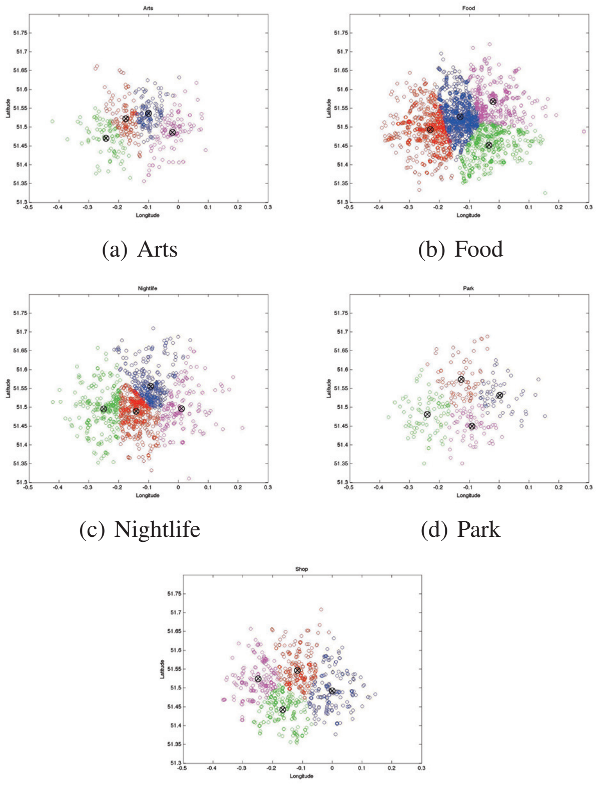
Figure 4: Power-law decay rate (α) comparison

cial clustering patterns among Park, Shop, and Arts while Food and Nightlife are showing a relative similarity as also observed in London.

Conclusion

Urban spaces are being blanketed by streams of digital data generated by inhabitants. This large pool of bits creates a unique opportunity for harvesting and analyzing behavioral data to better understand about the city and people. In this position paper, we analyze the aggregated data over about 16 months from a location-based online social network called *Foursquare*. The data has been collected from London, Paris, and New York. With the unique characteristic of the Foursquare network that allows the users to interact (*check in*) with the physical landscape (*venue*), we are able perform a geo-social and inter-urban analyses. With different venue's categories in our data, we find that social activity is distributed in a non-linear fashion and is following a Power Law distribution with Food and Nightlife social activity being the strongest social hubs across the three cities. Statistically we show that New York has a low variation in social distributions among different activity types compared with other two cities. Moreover, we observe a similar social clustering between Food and Arts in London and New York, but Paris on the other hand appears to have Arts and Nightlife social activity clustered comparably.

There are however some limitations in this study. The demographics and penetration of the Foursquare network in the

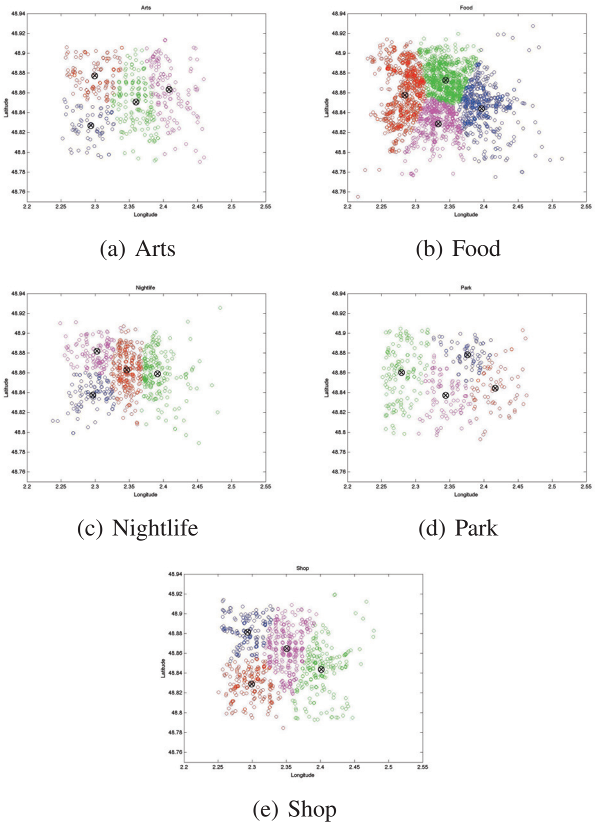


(a) Arts (b) Food

(c) Nightlife (d) Park

(e) Shop

Figure 5: Social clusters of each activity category in London



(a) Arts

(b) Food

(c) Nightlife

(d) Park

(e) Shop

Figure 6: Social clusters of each activity category in Paris

cities of study can also impact to the results. The arbitrary selection of number of centroids in our k-means clustering may not be a representative for social centers. Nonetheless we believe that to some extent this study helps us realize the usefulness of online social network data that can be utilized to better understand physical space and sociality. As our future direction, we will continue to investigate on the sole use of this data to understand the city as well as integrating it with data from other sources e.g. transportation, telecommunication to enrich context inference.

References

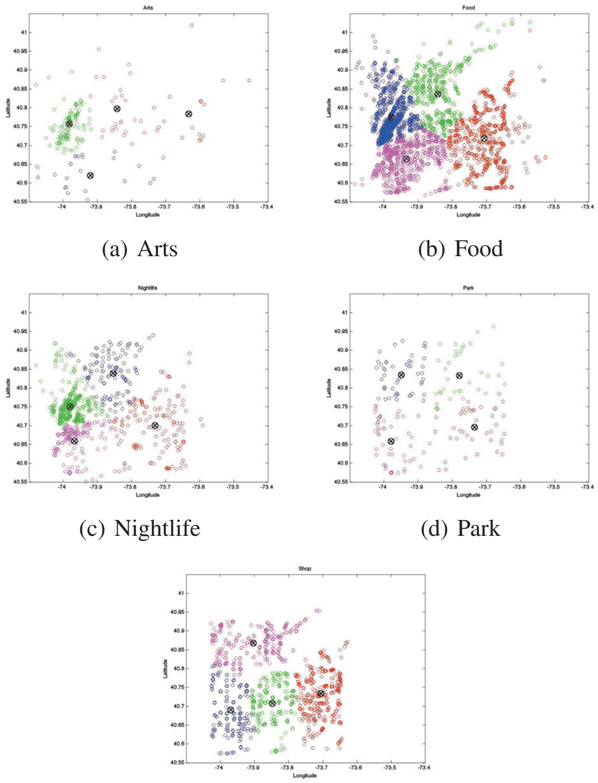
Bawa-Cavia, A. 2010. The city as social archipelago. http://www.urbagram.net/archipelago.

Calabrese, F.; Reades, J.; and Ratti, C. 2010. Eigenplaces: Segmenting space through digital signatures. *IEEE Pervasive Computing* 9:78–84.

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, K. P. 2010. Measuring user influence in Twitter: The million follower fallacy. In *Proc. of AAAI Conf. on Weblogs and Social*.

Clauset, A.; Shalizi, C. R.; and Newman, M. E. J. 2009. Power-law distributions in empirical data. *SIAM Rev.* 51:661–703.

Eagle, N.; Macy, M.; and Claxton, R. 2010. Network diversity and economic development. *Science* 328(5981):1029–1031.



(a) Arts

(b) Food

(c) Nightlife

(d) Park

(e) Shop

Figure 7: Social clusters of each activity category in New York

Eagle, N.; Pentland, A.; and Lazer, D. 2009. Inferring social network structure using mobile phone data. *PNAS* 106(36):15274–15278.

Foth, M. 2008. *Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City*. Hershey, PA: Information Science Reference.

Foursquare. 2011. http://foursquare.com.

Girardin, F.; Calabrese, F.; Fiore, F. D.; Ratti, C.; and Blat, J. 2008. Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive Computing* 7:36–43.

Gonzalez, M. C.; Hidalgo, C. A.; and Barabasi, A.-L. 2008. Understanding individual human mobility patterns. *Nature* 453(7196):779–782.

Leskovec, J. 2008. *Dynamics of large networks*. Ph.D. Dissertation, Carnegie Mellon University, Pittsburgh, PA, USA. AAI3340652.

Ratti, C.; Sobolevsky, S.; Calabrese, F.; Andris, C.; Reades, J.; Martino, M.; Claxton, R.; and Strogatz, S. H. 2010. Redrawing the map of great britain from a network of human interactions. *PLoS ONE* 5(12):e14248.

Reades, J.; Calabrese, F.; Sevtsuk, A.; and Ratti, C. 2007. Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing* 6(3):30–38.