POLITECNICO DI MILANO
DIPARTIMENTO DI ELETTRONICA, INFORMAZIONE E BIOINGEGNERIA
DOCTORAL PROGRAMME IN INFORMATION TECHNOLOGY
COMPUTER SCIENCE AND ENGINEERING

SPATIAL ANALYSIS OF ONLINE DATA TO TRACK CITIES' SOCIO ECONOMIC INDICATORS AND URBAN LAND USE

Doctoral Dissertation of:
Carmen Karina Vaca Ruiz

Supervisor:
Prof. Piero Fraternali

Tutor:
Prof. Barbara Pernici

The Chair of the Doctoral Program:
Prof. Carlos Ettore Fiorini

2014 – Cycle XXVI

Acknowledgments

Maybe too often we see life as a tapestry backwards without appreciating the harmony of the jewels chosen to be part of it. My doctoral studies have been a journey in which the jewels of that tapestry were multiplied and enriched abundantly. A great deal of insight and experience was needed to appreciate all that diversity and richness. Those were given to me through the people with whom I have been gifted through my life. It is because of them that one of the most challenging endeavors I have faced is going to be conquered soon. At this point of that journey, I have a great need to express my gratitude to all those people for being a indelible part of my life's tapestry.

To Him, who designed the project of my life perfectly, given that the shades also add perfection to the drawings. Thanks because I saw Your ongoing creativity changing scarcity into opportunities. Thanks for Your fatherly caring expressed in the eyes and the words of my friends but also in the beauty of the ordinary. Thanks for Your being closer to me than myself. To You, to My Mother, thanks for the uncountable number of gifts received in these years. Whether it was a mountain covered with snow, a lake reflecting the sun or a colleague helping me with some mathematical reasoning, all of that just let me admire You and Your infinitude even more.

To Bojana, Eleonora, Ilio, Chiara, Ekaterina, Luca, Alysson and everyone in the office. Thanks for all the original jokes that you are able to invent, you were wonderful officemates. A Eleonora e Rosanna Corbella grazie per il vostro aiuto con la lingua italiana.

To my advisor, Piero, thanks for introducing me to the research world and giving me the opportunity to change my career in a positive way. Thanks for giving me the freedom to choose the research direction and the guidance to not get lost in the path. Thanks to the Ecuadorian government agency SENESCYT for the support with the scholarship. Thanks to my wonderful colleagues at ESPOL, you are a constant source of inspiration. Thanks to my students at ESPOL. Among other important things, I used to think in your joy and passion for what you do whenever I asked myself why I was doing a PhD.

To everyone in Yahoo Labs, Barcelona, I met there magnificent researchers, some of them co-authors of the publications mentioned here. Beyond this, I found an environment in which long hours of work were compensated by the beauty of meeting creative and bright people. Amin, Daniele and Luca should be mentioned explicitly, it was an honor for me to work with you. Thanks for the discussions, the suggested readings, the lessons about collaboration and all the professional treasure that you carry with you. Ilaria, Ruth, Paloma, Yelena, Natalia, Estefania, Nicola, Ricardo thanks for your joy and your constant desire of discovering new recipes, places, and all the possible expressions of art! Ruthy thanks for teaching me to appreciate even more the Asian food, above all, thanks for the long conversations that we had, all your love for life enriched also mine.

Cori, Gabi, Moni, Robi, Ros, Ros, Angela, Angela, Lina, Betty, Melina voi siete state la mia famiglia a Como. Grazie per gli incoraggiamenti, i giri per il lago, Moreno, la stanza perfetta dopo un viaggio lungo, la pizza fatta in casa, il risotto, tutti gli scherzi a cena, i ciocolatini pieni di alcol, i fiori, i dolci e gli arancini siciliani e soprattuto grazie per essere sempre lì per regalare il vostro tempo e il vostro sorriso.

Ana Maria, gracias por todo lo que eres. Fuiste quien creó un hogar para mí en Barcelona junto a las estupendas Montses, Merces, Susana, Fina, Sonia, Isa, Marina, Roser y todas allí. Gracias por tu sentido del humor, por hacer una fiesta de cada comida de domingo, porque todas las dificultades perdían importancia ante una ingeniosa broma, por las inolvidables lecciones que me diste en esos meses en que tuve la suerte de tenerte cerca.

Angela thanks for your way of looking at life, for remembering always that children can teach us the simplest or maybe the more complex truths. My Italian experience would have been less richer without all the trains that we took and all the jokes that we told, without the friends that we visited together, those friends from whom I received even more strength to go on. Grazie a loro e grazie a te per le sfogliatelle vere, i musei, la lasagna dal sud, la accoglienza dalla tua famiglia, il ipad ritrovato, i passaggi in aeroporto e un infinito ecc. Grazie per stare sempre lì con una gioia piena di luce che ti fa pensare quanto è immensa la sorgente di quella gioia.

Cathy, Ma. de Gracia, Ma. Leonor y todas allí, gracias por vuestra alegría y cercanía, será un regalo reecontraros. A Pilar, a Tere y a todos en Strasburgo, gracias. A Ma. Jose, Carla, Lis, Lorena, Isabel, Marita, Cristina, Krishna, Veronica, Nervo, Ma Beatriz, Katiuska, Lupe, Elenita, Prisci, Sandra, Jessi, Riju, Syamantak, Denzil, Felix, Gustavo, Socorro, Manuel, Allan, Isabel y Nicolás, Thalia, Karina, Sixifo, Carlos, Guido, Katherine, Anita, Juan Antonio, San Josemaría y Don Alvaro, Dario y Rinaldo, a todos mis amigos a quienes no he mencionado aquí gracias porque la vida de cada uno de ustedes es un gran don para mí.

A mi familia, gracias porque de ustedes aprendí las lecciones más importantes de mi vida. Gracias porque siendo aquellos que más celebran mis pequeñas conquistas, son al mismo tiempo los que, si todo saliera mal, continuarían siempre creyendo en mí. Gracias por enseñarme que se alcanzan metas altas cuando al trabajo intenso y constante se une la seguridad de un gran amor. Gracias por los trámites, los bellos recuerdos, las horas de dedicación, cada abrazo, cada sorpresa y cada llamada. Gracias a mis hermanos, Yessica, Alicia, Raúl, los admiro y agradezco a la vida tenerlos conmigo. A mis sobrinos, a Genesis, Nayib, Julia, tíos y primos, a todos en la familia, gracias por mostrarme que hay amores que la distancia no disminuirá. Raúl, gracias amado padre, tu ejemplo y tu ternura son fundamentales en mi profesión y en mi vida. Gloria, gracias mamá, te he dicho muchas veces que el doctorado deberían conferirtelo a tí. Tu fe, tu generosidad, tu corazón lleno de riqueza me han enseñado a vivir y me recuerdan constantemente cuánto vale la pena amar de verdad, darse a quien tienes al lado, sonreir y mirar a la vida siempre con la cabeza en alto.

Abstract

ONLINE geolocalized data is being massively produced as a result of both, the interactions on online social networks, and the content shared on the Internet that is annotated with geographical locations. This constitutes a rich source of information to characterize geographical places where either the people interacting reside or where the geo-tagged content is produced.

Urban resources are allocated according to socio-economic indicators, and rapid urbanization in developing countries calls for updating those indicators in a timely fashion. The prohibitive costs of census data collection make that very difficult. To avoid allocating resources upon outdated indicators, one could partly update or complement them using digital data. In this dissertation we propose methods to estimate urban indicators as well as an unsupervised learning framework to discover dynamic areas of the city using the geotagged content published by either residents or visitors.

First we conduct and analysis of online attention patterns evolution in a content sharing platform. Evolution of online social networks is driven by the need of their members to share and consume content, resulting in a complex interplay between individual activity and attention received from others. In a context of increasing information overload and limited resources, discovering which are the most successful behavioral patterns to attract attention is important to determine whether the attention attracted is random or follows some patterns instead. To shed light on the matter, we look into the patterns of activity and popularity of users in the Yahoo Meme microblogging service. We observe that a combination of different type of social and content-producing activity is necessary to attract attention and the efficiency of users, namely the average attention received per piece of content published, for many users has a defined trend in its temporal footprint. The analysis of the user time series of efficiency shows different classes of users whose activity patterns give insights on the type of behavior that pays off best in terms of attention gathering. In particular, sharing content with high spreading potential and then supporting the attention raised by it with social activity emerges as a frequent pattern for users gaining efficiency over time.

Second, we analyze a random sample of interactions in the same service but focusing on content generated in Brazil and accurately predict the GDP and the social capital of 45 Brazilian cities. To make these predictions, we exploit the sociological concept of glocality, which says that economically successful cities tend to be involved in interactions that are both local and global at the same time. We indeed show that a city's glocality, measured with social media data, effectively signals the city's economic well-

being. To this end, we aggregate the attention that the city's residents are able to attract on the platform at the level of the city and quantify it using a set of metrics that are put together in a linear model that accurately predict the GDP.

Finally, we propose an unsupervised learning framework to capture the composition of cities. To discover functional areas in a city, spatial discovery algorithms have been recently applied to social media (e.g., Foursquare) data: functional areas are often identified based on semantic annotations of places and human mobility patterns. These algorithms have, however, considered the formation of functional areas and their semantic annotation as two separate steps. As a result, the derived areas might not be the best ones to be unambiguously annotated. We propose a framework based on an objective function to maximize. By being integrated into any clustering algorithm, this function aims at finding and labeling areas such that an area's label is semantically related to the points in the area and to those in the area's neighborhood without being too general (e.g., the label 'clothing stores' is preferable to 'professional places'). We evaluate the framework with a hierarchical clustering algorithm upon Foursquare data in the cities of Barcelona, Milan, and London. We find that it is more effective than baseline methods in discovering functional areas. We complement that evaluation with a user study involving 111 participants in the three cities, and with an additional temporal segmentation of areas upon Flickr data. The results generated by our framework can benefit a variety of applications, including geo-marketing, urban planning, and social recommendations.

We summarize the results of our analysis and discuss directions for future research and applications of our work.

Sommario

I Dati online geolocalizzati sono prodotti massivamente come risultato di due fattori: l'interazione degli utenti nelle reti sociali e l'annotazione geografica del contenuto condiviso su Internet. Ci costituisce una ricca sorgente di informazione, che permette di identificare i luoghi dove gli utenti risiedono o dove il contenuto geolocalizzato viene prodotto.

Le risorse urbane vengono allocate sulla base di indicatori socio-economici e la rapida urbanizzazione nei Paesi in via di sviluppo richiede di aggiornare questi indicatori in maniera tempestiva. Tuttavia, il costo richiesto per recensire larghe collezioni di dati è proibitivo. Di conseguenza, per evitare di allocare risorse in conformità ad indicatori non aggiornati, si potrebbe pensare di aggiornare quelli esistenti o di arricchirli utilizzando fonti di dati digitali. In questa tesi proponiamo metodi per stimare indicatori urbani. Inoltre, introduciamo un framework di apprendimento non supervisionato che permette di scoprire aree dinamiche della città utilizzando contenuto geolocalizzato pubblicato da residenti e/o visitatori.

Prima di tutto, il nostro lavoro esegue un'analisi dei pattern di online attention su una piattaforma di content sharing. Levoluzione delle reti sociali online è guidata dal bisogno dei propri utenti di condividere e consumare contenuto. Come risultato, possiamo osservare una complessa interazione tra ciò che è prodotto dagli utenti e lattenzione che quel contenuto riceve. Se si considera una situazione in cui la quantità di informazione prodotta cresce e le risorse sono limitate, scoprire quali sono i comportamenti più di successo per attirare l'attenzione di altri utenti è importante, poichè ci permette di determinare se lattenzione ricevuta casuale o se segue qualche pattern. Per fare luce sulla questione, nel nostro lavoro identifichiamo pattern di attività e popolarità degli utenti nel servizio di microblogging Yahoo Meme. In questo contesto, osserviamo che combinare diversi tipi di attività sociali e di produzione di contenuto è necessario per attirare l'attenzione degli utenti. L'analisi delle serie temporali di pubblicazione di dati da parte di un utente mostra che esistono diverse classi di utenti: il rate con cui essi pubblicano dati ci suggerisce quale sia il comportamento che meglio attira l'attenzione degli utenti di una rete sociale. In particolare, condividere contenuto con alto potenziale di diffusione e supportare l'attenzione con determinati comportamenti sociali emerge come pattern frequente in utenti con alta efficienza.

In seguito, analizziamo un campione casuale di contenuto generato in Brasile sulla stessa piattaforma. Su questo campione, prediciamo il GDP e il capitale sociale di 45 città brasiliane. A questo scopo, sfruttiamo il concetto sociologico di 'glocality', secondo il quale le città con economia forte tendono ad essere coinvolte in interazioni sia globali che locali. Mostriamo quindi che la 'glocality' di una città, calcolata su dati sociali, misura in modo significativo il suo benessere. Per farlo, aggreghiamo a livello di città, l'attenzione che i cittadini riescono ad attirare, e la quantifichiamo utilizzando

una serie di metriche aggregate in un modello lineare, così da predire accuratatamente il GDP.

Infine, proponiamo un framework di apprendimento non supervisionato che cattura la composizione di ciascuna città. Recentemente alcuni algoritmi di spatial discovery sono stati applicati per scoprire le aree funzionali urbane: le aree funzionali sono spesso identificate basandosi su annotazioni dei luoghi e sui pattern di mobilità delle persone. Questi algoritmi, tuttavia, hanno considerato la formazione di aree funzionali e le loro annotazioni semantiche come due passi distinti. Come risultato, le aree estrapolate potrebbero non essere le migliori o potrebbero essere annotate in maniera ambigua. Nel nostro lavoro proponiamo un framework basato su un problema di ottimizzazione. La relativa funzione obiettivo è integrabile in diversi algoritmi di clustering e aiuta a trovare e ad etichettare aree urbane, cosicchè l'etichetta di un'area è semanticamente correlata ai punti nell'area e a quelli nel suo vicinato. In seguito, valutiamo le performance del framework con un algoritmo di clustering gerarchico sui dati di Foursquare, nelle città di Barcellona, Milano e Londra. I risultati mostrano che l'algoritmo proposto è più efficiente di altre baselines nella scoperta delle aree funzionali. Questa valutazione è arricchita da uno user study, che coinvolge 111 partecipanti in tre città, e da una segmentazione temporale delle aree su dati di Flickr. I risultati generati dal nostro framework possono favorire una grande varietà di applicazioni, tra le quali geo-marketing, urban planning e social recommendations.

Contents

List of Figures

List of Tables

The economically weaker countries,must be enabled, to make a contribution of their own to the common good with their treasures of humanity and culture, which otherwise would be lost for ever.

John Paul II, 1987

1 Introduction

Cities are complex systems that evolve across time and authorities managing cities require updated information to better allocate resources or evaluate the result of undertaken urban projects. Therefore, the dynamic nature of cities calls for continuous tracking of the forces behind urban economic development: people and places inside the city.

Flows of information among people and flows of people to the places in the city can help to model diverse phenomena in urban areas as well as describe them through estimated indicators. However, one of the main challenges that we face when modeling cities is the lack of effective methods to collect and process urban data cheaply, efficiently and systematically [77]. Traditional methods for collecting data such as surveys or census are expensive. In this work, we posit that traditional methods to characterize cities can be complemented with the analysis of online data and propose new frameworks to exploit such data.

1.1 Online sources of geolocalized data

Local governments and urban planners, who want their cities to compete globally, must project and sell them as places with *favorable conditions* to work and live, as well as places with attractive sets of amenities, themed shopping, entertainment zones among others desirable characteristics [17]. To achieve such conditions, economic and development policies are designed and must be evaluated to find out whether the city improves or fails trying to become an exciting place to inhabit, work and visit. In summary, we need to evaluate the city's socioeconomic performance and how the people, whether locals or visitors, use the different urban areas.

Measuring the city performance is of interest not only to local governments and urban planners but also to entrepreneurs searching for new markets and for companies in the tourism field that aim to discover new urban experiences for the eager travelers. Yet, the lack of data at city (or sub-city) level is a key issue when abstracting urban metropolies, both, in their economic performance and the dynamics of the functional areas where people live, work or enjoy vacations [46, 77].

Back in 2001, the United Nations in its *World cities report* enumerated some problems concerning the availability of data at city level. Among these difficulties, the report mentions that, the local economic data is usually poor, there is lack of funding for data collection and there is lack of local level skills to collect and analyze fresh data [46]. Furthermore, the UN-HABITAT report published in 2003 [77] evidences the unavailability of information for intraurban locations [77]. More recently, a comprehensive study about local governance, argues that, still, cities largely depend on the census as the main source of data despite the long lags between collection periods (five to ten years) and its high cost [74].

On the contrary, the sustained growth of geolocalized digital data at speeds never witnessed before and the diversity of devices to generate and platforms through which spread such data, opens new possibilities for urban decision makers. In spite of the recent proliferation of such rich sources of data, these have not directly led to better understanding of the urban phenomena [50]. The challenge of finding methods to effectively shape our cities from these abundant sources of data remains. We address the challenge in this dissertation from two main points of view: estimating socioeconomic indicators at city level and discovering urban functional areas as reflected by the digital content published online.

1.2 Urban profiles from online data

Cities have a growing participation,of their residents, on the creation of online content. Certainly, publishing or consuming content on the Internet, once an information channel for elites, has become part of everyday lives for ordinary people, giving rise to the *networked society*. More often than not, the content published and the interactions that take place in this *networked society* reflect not only people's lifestyles but also the characteristics of the offline communities to which they belong [122]. As a consequence, an efficient way of looking at online content for describing urban places is that of building networks upon this data and apply social network analysis to it.

Networks appear in almost every side of human life. As an illustration, take a.) the relationships among individuals reflected by online social networks friendships, b.) the arrangement of places in urban environments, e.g., stores tend to appear together with restaurants *linked* by common visitors,or c.) the information cascades that emerge when pieces of content spread virally in social media platforms. Explaining the behavior of individual components in such networks is a demanding task. In spite of the difficulty of explaining individual behavior in networked environments, the online data and interactions when observed at aggregated level, exhibit patterns that can help to explain complex phenomena, even without understanding every component of the system [118]. We ask ourselves, which is the right level of granularity for the analysis. In previous work it has been stated that data generated inside the city boundaries can be analyzed at three different levels: users, cities, urban areas [79]. Therefore, we propose methods to *apprehend* the dynamics of social media data by a.) conducting an analysis of online attention attracted by *individuals* in content sharing platforms (Chapter 3) , b.) analyzing the networks that emerge from online attention exchanged by different *cities* (Chapter 4) and c.) mining user contributed data to discover and automatically label clusters of *urban places* that are semantically related and close to each other (Chapter 5).

1.3 Contributions of the thesis

1.3.1 Research contributions

The overall contribution of this dissertation work is a collection of techniques to extract patterns from content sharing platforms. We are particularly interested on discovering patterns useful to *characterize cities* using online data from a double perspective: urban socioeconomic indicators and urban land use (Figure 1.1). Urban planners can rely on these two perspectives a.) to quantify urban problems and position the city with respect to other metropolies (indicators) as well as b.) to evaluate the way people are using the different areas of the city and where new fluxes of people are heading (land use). However, the first part presents an analysis of individual users of a social network platform in which we aim to find whether or not are there individuals more effective at attracting attention online and which are the individual patterns that pay off better on a social network.

Our methodology start with the analysis of urban theories that explain the dynamics in urban ecosystems [12, 19, 38, 41, 76, 79, 104, 108]. Using such theories, we propose methods to aggregate the online geocoded data in order to describe cities using *proxies*, i.e., quantities that are proportional to a certain city feature. The number of users in an

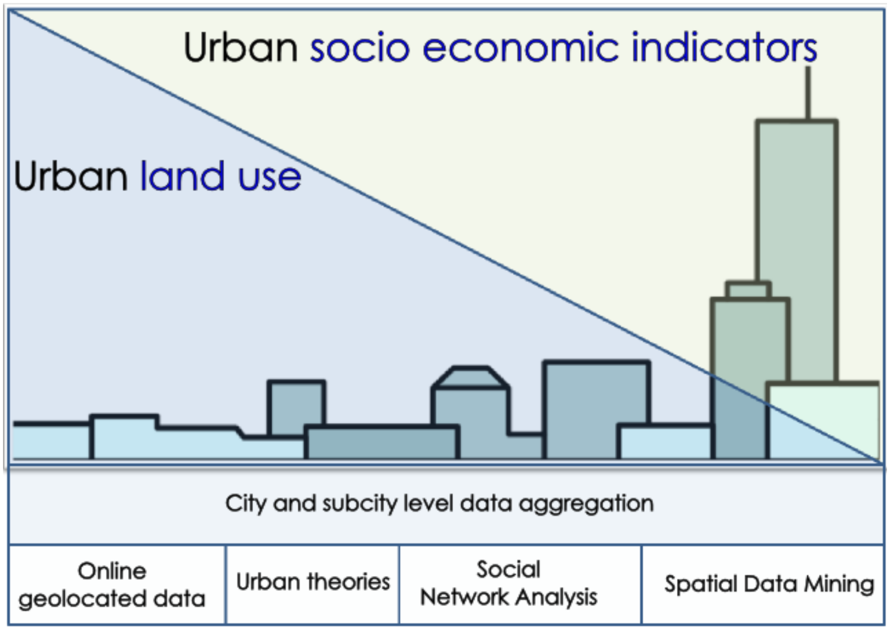


Figure 1.1: *Methodological approach to characterize cities from online geocoded data from a double perspective: urban socioeconomic indicators and urban land use. Background image adapted from www.russaimz.com*

online platform, for example, might be a good proxy for the population size given that the Internet penetration rate is high enough. Such *proxies* can be calculated with social network analysis and spatial data mining techniques applied to geolocated data. Finally, we present how to validate the city characterization obtained from online data. The presented methods as well as the validation procedure are repeatable using geographically annotated data from any sharing content platform.

Once our double path to characterize cities has been set, we disaggregate the tasks to follow such path. Thus, our first task will be to characterize online attention at individual level (without the spatial component) and at city level and our second task will consist on producing labeled urban areas after mining geolocated data. Our contributions, at solving these tasks, can be summarized as follows: First, we propose a novel framework to analyze temporal patterns of individual online attention in sharing content platforms looking at different stages of the user lifetime [111]. Next, we build upon the concept of glocality and apply it to the online attention exchanged by cities to effectively estimate urban socioeconomic [112, 112]. Finally, we propose a framework to cluster urban places labeling them automatically with categories arranged in a taxonomy (Figure 1.2).

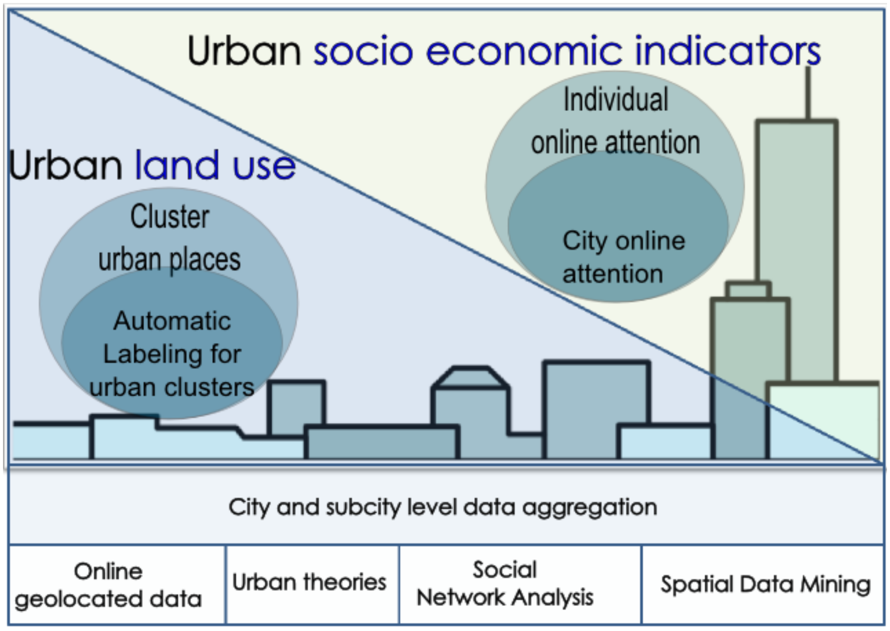

Urban socio economic indicators
Individual online attention
City online attention
Urban land use
Cluster urban places
Automatic Labeling for urban clusters
City and subcity level data aggregation
Online geolocated data
Urban theories
Social Network Analysis
Spatial Data Mining

Figure 1.2: *Methodological approach to characterize cities from online geocoded data. Detailed view of the proposed components to extract views of the city form user generated content.*

1.3.2 Thesis outline

We present the related work concerning online attention, socioeconomic indicators and urban functional areas detection in Chapter 2 where we highlight our contributions with respect to previous work. Next, we embrace in the city characterization process.

To succeed at describing urban areas from online data and interactions, we need to first to understand better the dynamics behind such interactions at *individual* level. Therefore in Chapter 3, we model the patterns of online attention gathered by individuals in a microbloging site where people publish their own ideas and/or reshare others'. Content diffusion and popularity dynamics on social media platforms have been extensively studied before [59, 60, 114] focusing on influential users. We, instead, explore patterns of attention received by *average* users where the term *average* refers to the fact that those users do not necessarily have a large number of followers or reposts.

Our main goal is to observe whether exist some network members who are able to sustain the attention they attract from their peers and observe which are the characteristics of their activity. Indeed, 56% of the users in the study exhibit a clearly defined trend (increasing, descreasing, peaking) in the temporal footprint of the online attention they gain. We define a method to classify noisy time series of human activity and succesfully detect the three user classes listed before and temporal changepoint. We observe that activity before and after such changepoint reveals important behavior of users that need to maintain social exchange (such as comments and new followees) to maintain the status they have acquired in the system. Effective users exhibit 23% more comments, 61% more reposts than their peers after the changepoint is detected.

Thereafter, in Chapter 4, we add the spatial dimension to move our analysis of online attention from the *individual* level to the *city* level. In order to do so, building upon solid urban theories, we design a set of metrics that quantify the city's glocality, the combination of the local and the global, and correlate it with its economic performance. Specifically, we derive a graph that depicts the attention exchanged by cities' residents and millions of cascade trees that allows us to quantify the geographical reach of a single post in a microblogging platform aggregating such information for 35 Brazilian cities. The metrics are computed using social network techniques applied to these two types of graphs. After calculating the online attention metrics for each city, we test the hypothesis that economic performance is higher for cities that are more 'glocal', in other words, cities whose residents attract attention from distant ties while maintaining strong local links. In fact, we present a linear model that predicts GDP (Gross Domestic Product) per capita with an $Adjusted\ R^2$ of 0.94. We also propose a city prestige's index to quantify the social capital of the city and found that it can be predicted from the social attention metrics with an $Adjusted\ R^2$ of 0.93.

Finally, in Chapter 5, we address the remaining level of analysis: *urban places* in the city. We aim to exploit online geolocalized data to discover clusters of urban places such that points inside the clusters are spatially close and more semantically related that points outside them. The discovered clusters represent ‘functional areas’ as they constitute areas of the city where people, for example, do shopping, go to school, visit historic places. Unlike previous work, we automatically assign a label that describes each cluster based on a taxonomy of categories extracted from the same user contributed data. The taxonomy driven approach that we adopt to build our solution results in a very flexible framework with a single parameter that allows the user to find areas at different levels of granularity:‘Shops’ if you look for more general description, opposed to ‘Clothing stores’ if you set a lower value for the user defined parameter. The problem is framed in terms of maximization of a simple objective function to be integrated into any clustering algorithm (Section 5.2). In fact, we implement a hierarchical clustering algorithm tailored adding spatial constraints. To evaluate our system, we perform a quantitative comparison against a baseline in terms of labeling accuracy using Foursquare data collected for three European cities: Barcelona, London and Milan. (Section 5.3). We find that our approach is more effective than baseline methods in discovering functional areas in those three cities. In addition to that, we conduct an user study (Section 5.4) in which we ask 111 participants to suggest the most suitable area in the city for specific tasks (e.g., where to best place a new tech startup, where to go shopping). We found that our framework matches what local residents or people who have lived in the city for years perceive.

In Chapter 6, we summarize the work of this thesis, explore the implications of the work and present future directions.

1.4 List of publications

The research work presented in this dissertation has been partially published in different venues. I also had the opportunity to contribute to some research projects that resulted in academic publications not included in this thesis. I detail these publications next:

Works related to this dissertation

Carmen Vaca Ruiz, Daniele Quercia, Luca Maria Aiello, Piero Fraternali. Taking Brazil's Pulse: Tracking Growing Urban Economies from Online Attention. *Proceedings of the 23rd World Wide Web Conference (WWW 2014 Companion volume)*, (Seoul, South Korea), April 2014.

Carmen Vaca Ruiz, Daniele Quercia, Luca Maria Aiello, Piero Fraternali. Modeling dynamics of attention in social media with user efficiency. *Collective Behaviors and Networks, EPJ Data Science, SpringerOpen Journal*, March 2014.

Carmen Vaca Ruiz, Daniele Quercia, Luca Maria Aiello, Piero Fraternali. Taking Brazil's Pulse: Tracking Growing Urban Economies from Online Attention. *Proceedings of the 23rd World Wide Web Conference (WWW 2014 Companion volume)*, (Seoul, South Korea), April 2014.

Carmen Vaca Ruiz, Daniele Quercia, Luca Maria Aiello, Piero Fraternali. Tracking Human Migration from Online Attention. *Citizen in Sensor Networks Worskhop, Springer 2014 Lecture Notes in Computer Science*, (Barcelona, Spain), September 2013.

Carmen Vaca Ruiz, Daniele Quercia, Francesco Bonchi, Piero Fraternali. Taxonomy-based Discovery and Annotation of Functional Areas in the City. Submitted for evalutation.

Other works

Carmen Vaca Ruiz, Amin Mantrach, Alejandro Jaimes and Marco Saerens. A Time-based Collective Factorization for Topic Discovery and Monitoring in News. *Proceedings of the 23rd World Wide Web Conference (WWW 2014)*, (Seoul, South Korea), April 2014.

Piero Fraternali, Andrea Castelletti,Rodolfo Soncini-Sessa, Carmen Vaca Ruiz, Andrea Rizzoli. Putting humans in the loop: Social computing for Water Resources Management.*Environmental Modelling and Software Journal, Elsevier*, (Amsterdam, Netherlands), November 2012.

Marco Brambilla, Piero Fraternali, Carmen Vaca Ruiz. Combining social web and BPM for improving enterprise performances: the BPM4People approach to social BPM.*Proceedings of the 21st World Wide Web Conference Companion volume(WWW 2012)*, (Lyon, France), April 2012.

Marco Brambilla, Piero Fraternali, Carmen Vaca. BPMN and Design Patterns for Engineering Social BPM Solutions. Business Process Management Workshops. *BPM 2011 International Workshops*, (Clermont-Ferrand, France), August 2011.

Marco Brambilla, Piero Fraternali, Carmen Vaca: A Notation for Supporting Social Business Process Modeling. *Business Process Model and Notation - Third International Workshop, BPMN*, (Lucerne, Switzerland), November 2011.

And they have a fund of wisdom and wise sayings that men have mostly never heard or have forgotten long ago.

J. R. R. Tolkien, 1937

2

Related work

2.1 Online attention

Much effort has been spent lately in measuring the effect that the activity of content production and sharing has in influencing the actions of social media participants. Depending on whether the investigation adopts the perspective of the *user* who is sharing or of the *content* being shared, emphasis has been given to the characterization of either the influential users or the process of information spreading along social connections.

Different methods to identify influentials, namely individuals who seed viral information cascades, have been proposedrecently [84], and it has been observed that simple measures such as the raw number of social connections are not good predictors of influence potential [5, 20, 92]. Instead, the ease of propagation of a piece of content is correlated with many other features, including the position of the content creator in the social network [51], demographic factors [105, 106], and the sentiment conveyed in the message [89].

For what concerns content-centered analysis, much attention has been devoted to the study of the structure and diffusion speed of information cascades in social and news media [6, 21, 124], including Yahoo Meme [53, 124]. Weng *et.al.* [124] for instance have shown that triadic closure helps to explain the link formation in early stages of the user's lifetime but later in time it is the information flow the driver for new connections. Despite the difficulty of determining whether observed cascades are generated by a real influence effect [100] (unless performing controlled experiments [7]), the role of influence in social network dynamics is widely recognized, albeit not fully understood. Factors related to influence include geolocation, visibility of the content, or exogenous factors like major geopolitical or news events for news media [16, 33, 123].

Patterns of temporal variation of popularity have been investigated previously, mostly focusing on the attention given to pieces of user-generated content. Previous work includes characterization of the peakness and saturation of video popularity on YouTube in relation to content visibility [33], crowd productivity dependence on the attention gathered by videos [52], the classification of bursty Twitter hashtags in relation to topic detection tasks [65], and the clustering of hashtag popularity histograms based on their shape [125]. Time series has been used to predict popularity in blogs, where the early reactions of the crowd to a piece of content is strongly correlated to the expected overall popularity [73, 107].

2.2 Tracking urban indicators from online data

Real-life processes and social media. Social media data has already been related to real-life outcomes. Gruhl *et al.* [44] showed that increases in blog mentions of books correspond to spikes in sales. Tweets has been used to predict the Dow Jones Industrial

Average [13], box-office revenues for movies [4], trending tickers in the stock market [94], polls' results for political opinions [81], and election outcomes [109]. This line of work has its critics. Mejova *et al.*, for example, have argued that sometimes debates on Twitter do not reflect the national preferences [75]. More recently, not only tweets but also email exchanges have been used to track migration flows among developed and developing countries [102].

Social capital. Online engagement has been shown to impact individual social capital offline. In different contexts, researchers have shown the role of Internet-mediated social interactions in supplementing and enhancing face-to-face and phone communication as well as in increasing the participation to political or voluntary organizations [119]. Ellison *et al.* found that young Facebook users strengthen offline ties through online interactions and that their primary online audience is made of people they regularly meet offline [32]. Furthermore, individual bridging and bonding social capital can be accurately predicted using Facebook interactions [103]. Previous analysis of online forums has also shown that active participation in virtual communities directly impacts the likelihood of offline exchanges among people living in the same neighborhoods [48].

Socio-economic indicators. Previous studies have also explored the connection between online interactions and socio-economic indicators of city neighbourhoods [34]. Eagle *et al.* analyzed networks derived from landline phone data and showed a strong relationship between social/geographical diversity in those networks and access to economic opportunities for city neighbourhoods across UK [31]. More recently, Quercia *et al.* have shown a correlation between the sentiment expressed in tweets originated by residents of London neighborhoods and the neighborhoods' well-being [90].

2.3 Urban areas detection

The simplest way of finding functional areas is to use a spatial clustering technique. One of the most common techniques is the Density-based Spatial Clustering of Applications with Noise (DBScan) [29]. It finds a number of clusters starting from the estimated density distribution of points, and has been recently used on Foursquare data in the three cities of New York, London, and Paris [9]. To test the hypothesis that a modern city functions as a 'social archipelago' (i.e., "a fragmented set of islands characterized by high-density social activity"), the author modeled Foursquare venues as geo-located points and found that Paris is less spatially fragmented than London; by contrast, New York is the most fragmented, twice as much as Paris.

Most of the latest research effort has gone into finding functional areas in the city. Researchers have done so in three main ways. The first way has relied on grouping together semantically-annotated points of interests. Cao *et al.* identified popular signatures (e.g., frequency distribution of different types of buildings) to find urban patches that frequently occur in different parts of the city. For example, the signature of residential areas might well be the high presence of single houses and garages [19]. Noulas *et al.* exploited semantic annotations of Foursquare venues for grouping geographic areas in New York and London [80], and saw how those changed from day to night.

The second way of finding functional areas has relied on human mobility. Mobility is derived from mobile phone traces [88] or Foursquare check-ins [69, 128]. The premise of those approaches is that people's movements signal the potential and intrinsic relations among locations. In this vein, by tracking where Foursquare users check-in, Cranshaw *et al.* were able to move beyond the politically-defined boundaries of neighborhoods and discovered areas that effectively reflect the character and life of city areas [26].

The third (and latest) way of finding functional areas is to combine semantic annotations with human mobility. Yuan *et al.* inferred the functions of each area using

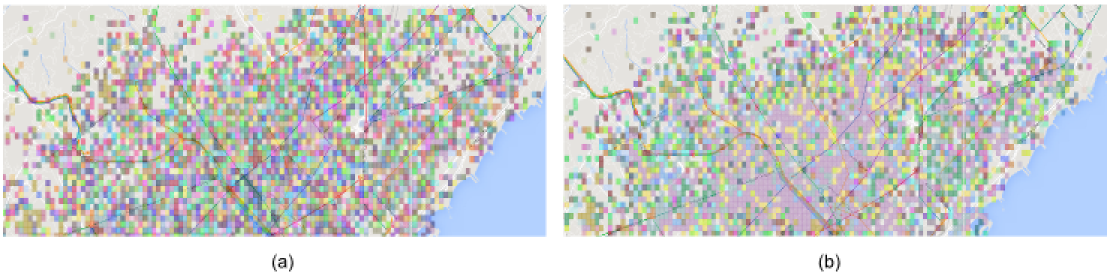


Figure 2.1: *Cells in Barcelona, each of which is labeled using: (a) the most frequent venue category; (b) the category with the highest* TF-IDF *score.*

a topic-based inference model: areas are modeled as documents, functions as topics, categories as metadata, and human mobility patterns as words [126]. They found that this way of discovering functional areas is far more effective than if one were to apply *TF-IDF* or *LDA* on the same datasets. Indeed, if one were to gather all the foursquare venues in the city of Barcelona, divide the city map into 100x100m walkable cells, and color each cell with either the most frequent venue category or the category with the highest *TF-IDF* score, then the resulting maps would be either too homogeneous (most of the cells in Figure 2.1(a) are labeled as ‘food’) or too fragmented (Figure 2.1(b)).

All the previous approaches find functional areas and, only after that, describe each of the resulting areas by either counting the categories in the area or identifying the area’s categories that are salient (those that, e.g., tend to co-occur more than chance). Understanding what an area really means from such descriptions results, however, in a considerable human effort.

2.3.1 Discussion

Online attention

In this thesis, when analyzing online attention in a social network platform, we focus on users as opposed to content and we analyze time series of a metric combining the user activity and the attention received. We do not focus on the popularity gained at a global scale, but instead we characterize temporal patterns of activity and attention of each individual.

We show that time series of individual user activity cannot be clustered accurately based on their shapes by state-of-the art methods, so we propose an algorithm to fix that. Finally, except in rare cases (e.g., [58]), previous work on network analysis has relied mostly on limited temporal snapshots. In contrast, we use the temporal data of the entire life-span of Meme, from its release date until its shutdown. A point often overlooked is the fact that users usually present a well defined temporal footprint when it comes to the amount of attention they attract from their peers in social networks. We will detail a method to detect important changepoints in such temporal footprints and the combination of activities that led to different patterns in user efficiency: increasing, decreasing and peaky.

Socio economic indicators from online attention

In the last few years, there have appeared some initiatives for measuring socio-economic conditions of city residents in developing countries using online data. For example, the United Nations and the World Bank have recently launched a program called “Data4Good”. This promotes the use of (currently untapped) digital data for, say, improving poverty measurement (“How can we measure poverty more often and more accurately?”) or dealing with corruption in international investment projects (“Can we detect fraud by looking at aid data?”). Recently, Orange released an anonymized dataset of mobile phone calls in Côte d’Ivoire, and launched a challenge in which researchers had to predict economic indicators from the activity metrics extracted from the call records [72].

In this dissertation, we complement this line of work by proposing a set of metrics that can be applied to data extracted from any data source that reflects social exchanges, including social media data. As an illustration, we are able to predict GDP per capita and social capital for 45 Brazilian cities with a content independent approach. Our framework combines the geographical properties of user online behavior with the dynamics of online attention given to content to spot the wealth of a city without imposing the need of crawling user demographics or processing the text or multimedia content published on the Internet. An alternative dataset to estimate proxies for urban wealth are the phone calls records released by telecommunication companies (anonymized and aggregated). While our method can be applied to these type of geographical information sources, our contribution relies on the fact that the huge amount of user generated content published online can be exploited, as well, provided that it contains geographical annotations.

Discovery and Automatic Annotation of Functional Areas

Finally, we propose a novel approach for clustering urban places that overcomes one of the main drawbacks of previous work: the automatic labeling of the produced clusters. Previous approaches mainly use topic detection frameworks applied to the geolocated data to produce dynamic urban areas and a second step is needed to manually assign labels to the discovered areas. In our approach, we cluster urban places and automatically label them using categories in a taxonomy. Moreover, our framework let the user to choose how general of specific the labels should be.

Previous spatial clustering algorithms work directly on geolocated points. We propose to divide the city in equally sized grid cells to aggregate points. This choice gives our method a competitive advantage since the elements to be clustered are significantly reduced. When conducting experiments, for example, DBScan, a state of the art spatial clustering algorithm, requires a large matrix to represent the distances between each of the 1M points in our Flickr dataset. Our method is applied, instead, to the cells in the city grid resulting in lower processing requirements without sacrificing semantic accuracy (Chapter 5).

Will is the engine of communication

St. Augustine, 354

3

Dynamics of online attention in microblogging platforms

3.1 Introduction

Understanding users' activities in social media platforms, in terms of the actions they take and how those actions affect the attention they receive (e.g., comments, replies, reposts of messages they post, etc.), is crucial for understanding the dynamics of social media systems as well as for designing incentives that lead to growth in terms of user activity and number of users. As expected, given the nature of such platforms, users who receive attention from their peers tend to be more engaged with the service and are less likely to churn out [52]. Insights on the kinds of actions that users take to gain more attention and become "popular" are therefore important because they can help explain how social media platforms evolve. In spite of the importance of analyzing such behavior at a large scale, the dynamics of attention are not well understood. This is largely due to two main reasons: on one hand that there are few datasets that show the evolution of a network from its very beginnings, and on the other hand, because most work has focused on the popularity of content rather than on analyzing the effects of user's behaviors on how other users react to them. For example, there have been many studies to establish the reasons behind user or item popularity in social networks (e.g., [91,107]), but the effects that the patterns of attention received have on the activity and the engagement of the "average" users have not been thoroughly explored so far.

In this chapter, we address questions that focus on social media users' behavior at different stages of their participation in social media platforms. The analysis conducted in this chapter focuses on the dynamics of attention at individual level. We are interested in discovering whether or not, individual users show patterns stable in time in the attention they receive from their peers on a content sharing platform. Later, we will extend the analysis adding the geographical component to obtain measurements at the city level. In particular, we introduce a new way to examine attention dynamics, and from this perspective perform a deep analysis of the evolution of user activity and attention in a social network from its beginning until the service ceased to exist. Analyzing the weekly efficiency, i.e. the amount of attention received in the platform normalized by the amount of content produced, we observe that 56% of the users in the dataset exhibit a footprint of their efficiency with a clearly defined trend (i.e., sharply increasing/decreasing or peaking). We are able to extract patterns of user behavior from these temporal footprints that reveal differences in the activity behavior of users of different classes. We focus our analysis on Yahoo Meme, a micro-blogging service that was launched by Yahoo in 2009 and discontinued in 2012. While the mechanisms of interaction in Yahoo Meme were similar to those found in other social media platforms, to the best of our knowledge, this is the first study that examines in detail the questions we are addressing from the perspective of user efficiency, using data from a service from its initial launch.

The main contributions presented in this chapter include:

- Study of the attention dynamics in social networks from the angle of *efficiency*, namely the ratio between attention received and activity performed. The notion of efficiency in time allows to detect patterns that could not emerge using other raw popularity or activity indicators.
- Definition of a method to classify noisy time series of user-generated events. The method is successfully used to find classes of users based on the time series of their efficiency scores, with an accuracy ranging from 0.85 to 0.93, depending on the different classes.
- Extraction of insights useful to detect and prevent user churn. For instance, exploration of the efficiency time series reveals that increase in efficiency is determined by creation of high-quality content, but the acquired attention has to be sustained with additional social activity to keep the efficiency high. If such social exchange is missing, attention received drops very quickly.

3.2 Dataset description

Meme was a microblogging service launched by Yahoo in April 2009 and discontinued in May 2012. Users could *post* messages, receive notifications of posts published by people they *follow* (follower ties are *directed* social connections), and *repost* messages of other users or *comment* on those messages. The overall number of registered users grew at a constant pace, up to almost $700K$. When neglecting uninvolved users (i.e., users who were registered, but stopped explicit activity), we observe a growing trend up to a maximum of $60K$ users around the end of the first year, and then a slow but steady decline. In Table 3.1 we report general statistics on the follower network in the

Nodes	**Edges**	**Density**	$\langle k \rangle$	$\langle k_{in} \rangle$	**GWCC$_{\%}$**	**Reciprocity**	$\langle d \rangle$	d_{max}	C
568K	20M	$6.2 \cdot 10^{-5}$	71	35	0.996	0.096	2.59	11	0.433

Table 3.1: *Yahoo Meme followers networks statistics.*Δ *= density, GWCC*$_{\%}$ *= relative size of the greatest weakly connected component,* d *= geodesic distance,* C *= clustering coefficient.*

last week of the service. The final network contains a well-connected core of users resulting in a greatest connected component covering almost the full network, with a high clustering coefficient. As already observed for other online social networks, the average path length is proportional to $\log\log(N)$, and similarly to other news media the level of social link reciprocity is very low [61].

3.3 Activity vs. attention

Activity and attention are the two dimensions we aim to examine with our study. After defining the features, we look at their relationship in terms of correlations of their raw indicators and then we study them from a novel perspective by defining a metric of user efficiency. We find that very efficient users tend to write fewer posts per week but are heavily involved in social activities such as commenting.

3.3.1 Activity and attention metrics

We define *activity* and *attention* indicators that are computed for every user. Activity indicators are measured by the number of posts (pd), reposts (rd), and comments done (cd), or by the number of new followees added ($fwee$), while attention is determined by the number of reposts (rr) ot comments received (cr) from others, and by the number of new incoming follower links (fw). Reposts received can be *direct* or *indirect* (i.e., reposting a repost). To measure attention we consider direct reposts.

The possibility of indirect reposting originates repost *cascades* that can be modeled as trees rooted in the original post and whose descendants are the direct (depth 1) and indirect (depth 2 to the leaves) reposts. Besides being another attention indicator, the cascade size (cs) is a good proxy for the *perceived interestingness* of the content because, intuitively, sharing a piece of content originated by someone who is not directly linked through a social tie, and therefore is likely to be unknown to the reposter, implies a higher likelihood that the reposter was interested in that piece content. Therefore, we consider the cascade size as a measure of content interestingness.

Even though several measure of influence, authoritativeness, or more in general importance of a user in a networked system have been developed in the past (see for instance the work by Romero *et. al* [92]), here we adopt the perspective of a single user, rather than of the whole community. Therefore, we are going to interpret the system as a black box that receives input from a user (activity) and returns some output (attention), without considering the actual effect that the input causes inside the system. Although this is a simplification, it allows us to better focus on the user dimension and to cluster users with respect to the perception they get from the interaction with the system (i.e., attention in exchange for activity).

3.3.2 Correlations

When dealing with multidimensional behavioral data, detecting causation between events can be difficult [100], but potential mechanisms driving the interactions between the different dimensions at play can be spotted through the investigation of correlations [98]. In this case, the correlations between activity and attention metrics give a first hint about the potential payoff of some user actions in terms of attention received.

In Figure 3.1, visual clues of the relationship between different metrics of activity and attention are shown in the form of heatmaps. The four plots on the left display the average values of attention indicators for users whose number of posts and comments resides in given ranges. To make sure that the trends emerging from the heatmaps are significant, we count the number of users falling in each of the range buckets. In Table 3.2 we report the average and the median number of users in each bucket of the heatmaps. As expected from the broad distributions of the activity and attention indicators, few actors have very high values for some pairs of indicators. For instance,

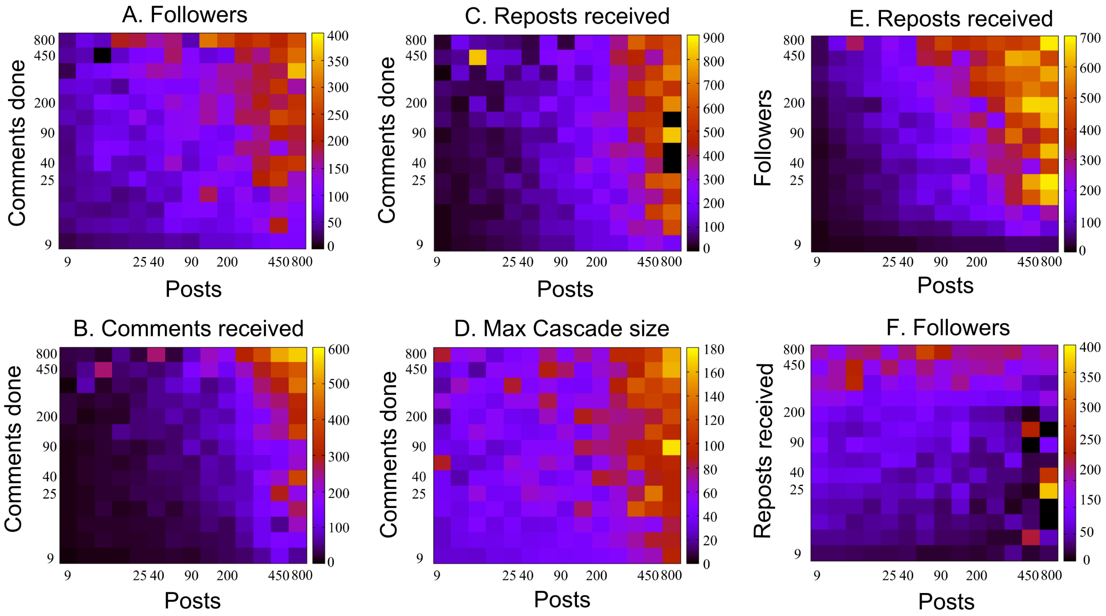


Figure 3.1: *Correlations between activity and attention. Users were grouped according to the number of x and y values (plotted on a log scale) and, for each group, the average number of the z-value was calculated and mapped to a color intensity.*

x-axis	y-axis	Average	Median
Posts	Reposts Received	73.1	39
Posts	Comments Done	77.9	32
Posts	Followers	74.3	45

Table 3.2: *Statistics for the number of users considered in each bucket of the heatmaps depicting the correlations between activity and popularity metrics (Figure 3.1). The average and median number of users per bucket in each combination of metrics is shown.*

in the heatmap 3.1.E, just 10 users are in the upper-right bucket (users with > 625 posts *and* and > 625 followers). However, in general the number of users per bucket is in average higher than 70 users, as shown by Table 3.2, thus the trends observed in the heatmaps is not obscured by cells with very low number of users.

First, we observe that attention in terms of followers and comments (Figures 3.1.A-B) is correlated with both number of posts and comments done, resulting in a color gradient becoming brighter when transitioning from the lower-left corner to the upper-right one. Users who gained more followers were heavier content producers and an even more evident correlation is found when considering comments received (Figure 3.1.B), likely due to a comment reciprocity tendency (we calculated the comment reciprocity being around 24%, much higher than reciprocity in the follower network). We observe a partially similar effect when looking at content-centered indicators, namely the reposts received and the cascade size (Figures 3.1.C-D). In these cases we find a positive correlation with the number of posts, but not with the amount of comments, suggesting that social interaction, such as commenting on other people posts, does not strongly characterize content propagation.

The two plots on the right of Figure 3.1 show the relation between pairs of attention metrics with the number of posts. From Figure 3.1.E we learn that social exposure (i.e., being followed) and productivity (i.e., number of posts) are both heavily correlated with the number of reposts. However, people with moderate or heavy posting activity can reach a high level of attention even having a relatively small audience (as shown by the bright colors extending down along the right side of the map). This intuition is confirmed by the fact that swapping the axes of the two attention measures, the correlation is disrupted (Figure 3.1.F), meaning that people with high number of posts and reposts do not necessarily have a large number of followers.

3.3.3 User efficiency

The above findings support on one hand the intuitive principle about: "the more you give, the more you get" and, on the other hand, they reinforce the hypothesis that visibility is not enough to grant a wide diffusion of content (similarly to the "million follower fallacy" in the context of Twitter [20]). However, the user perception of the interaction with peers through an online system is not dependent just by the raw number of feedback actions received, but also by the amount of attention in relation with the effort spent to gain it. Given this perspective, we define the *efficiency* η of a user u in a given time frame $[t_i, t_j]$ as the amount of attention received over the amount of activity performed between t_i and t_j, for any pair of activity (Act) and attention (Att)

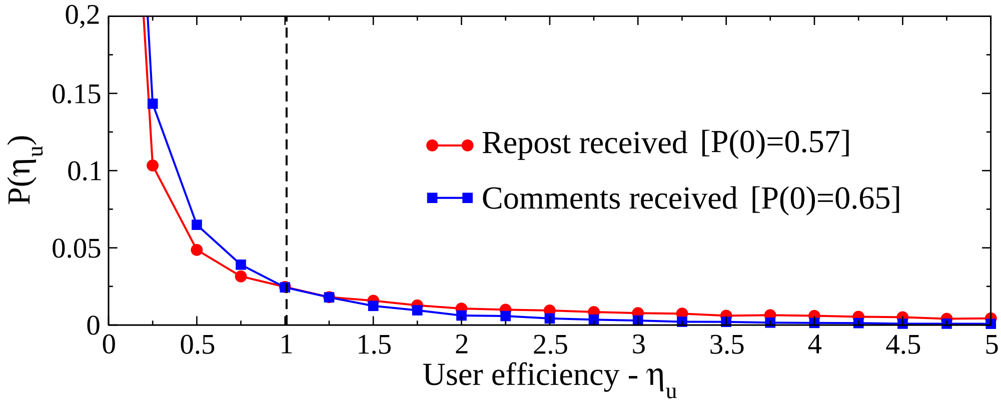


Figure 3.2: *Distribution of efficiency scores, bucketed in 0.25-wide bins. Average scores are* 0.38 *for comments and* 1.55 *for reposts.*

metrics:

$$\eta_u^{Act,Att}(t_i,t_j) = \frac{\sum_{t_i}^{t_j} Att_u}{\sum_{t_i}^{t_j} Act_u}. \tag{3.1}$$

Analogous definitions have been used in different disciplines such as physics and economics [2], and in most of the cases the efficiency is upper bounded to 1, i.e., the outcome from the system cannot exceed the energy given in input. On the contrary, in a social media setting the efficiency is unbounded and it constitutes an objective function to maximize in order to increase the engagement of the user base. Even if comments can be strong indicators of involved user participation, the main focus of the online service under study is posting and reposting, similarly to Twitter. Therefore we always consider the number of posts as the metric of activity in the efficiency formula. In the above definition (Formula 3.1) we assume that the attention that we take into account should be the one that is directly triggered by the activity considered, we use either the number of reposts ($\eta_u^{Post,Repost}$) or the number of comments ($\eta_u^{Post,Comm}$) as proxies for attention received, since other metrics such as number of followers are not necessarily responses to the posting activity.

The distribution of $\eta_u^{Post,Repost}$ and $\eta_u^{Post,Comm}$ for all the users during the complete lifetime of the network is drawn in Figure 3.2. Even if the maximum efficiency scores span up to several hundreds, the majority of users have an efficiency lower than 1, and

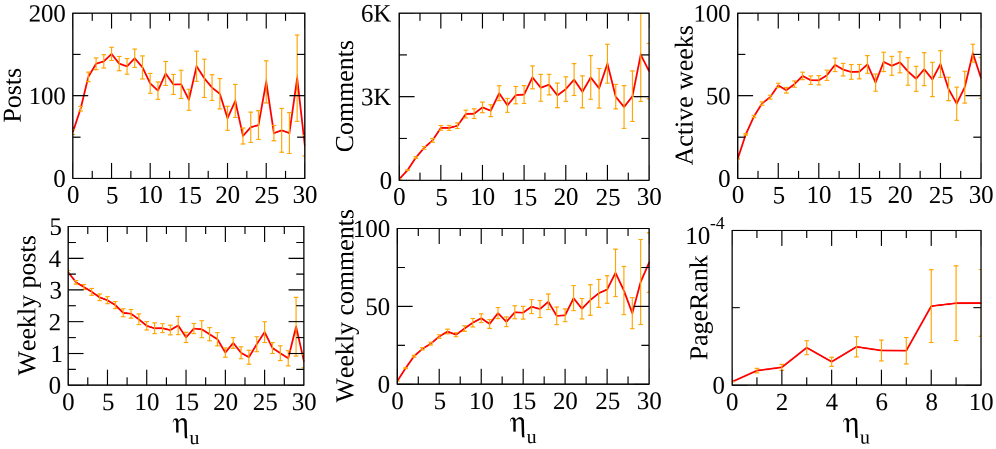


Figure 3.3: *Average values of activity and status indicators at fixed values of η_u.*

most of them have values close to zero. The average over the η_u values of all users is higher than 1 for reposts and much lower for comments. This is justified by the fact that Meme emphasized especially the repost feature. For this reason, next we consider only the efficiency of posts in relation to reposts, and we refer to it as η_u, for simplicity.

High activity is usually indicative of poor efficiency or, in other words, activity alone is not indicative of high potential of attention gain. To study more in depth the traits of efficient and inefficient users, we describe users with different η_u values according to several activity and status features, as shown in Figure 3.3.

Insightful patterns emerge. First, the higher the η_u, the lower the activity in terms of number of posts, but not in the range $0 \leq \eta_u \leq 5$ (containing most of the users), in which the number of posts grows with η_u. However, when looking at the average number of posts submitted per week instead, the trend becomes monotonic, confirming the theory about the limited attention of the audience being a barrier for attention gathering [123]. Second, the higher the η_u, the higher the amount of comments: the more efficient users are the ones who comment the most. Finally, the longevity of the profile and the prestige on the follower network (computed with standard PageRank) are also distinctive features of efficient users.

3.4 Evolution of efficiency in time

Attention attracted by users, and by consequence their efficiency, is not constant in time. It depends on the amount of activity, the position in the network and other factors. However next we show that, even if many users exhibit a oscillating but globally stable values of efficiency in time, more than half the users show sharp variations in their efficiency time series, that tell more about the activity behavior in different periods of the user lifetime. First, we give the definition of efficiency time series. Then, we explain the algorithm used to classify users efficiency traces according to the shape of their trend and discuss the properties of the four classes we found. We i) find that state-of-the-art algorithms for clustering of timeseries do not perform well on the noisy traces such the ones generated by human activity, therefore, based on the observed shapes, ii) we propose a new classification method and evaluate it against a human-curated ground truth, and iii) we analyze the differences between user behaviors in the four main user efficiency classes around the main changepoint of the efficiency curve.

3.4.1 Efficiency time series definition

By adapting the efficiency formula for a discrete-time scenario, we model the temporal efficiency evolution using weekly time series for each user u measuring the efficiency η_u after each week. The elements of the series are generated as follows:

$$\eta_u(t_i) = \frac{rr(p_{t_i})}{|p_{t_i}|}, t_i \in T_u = \{t_1, ..., t_n\},$$

where p_{t_i} represents the set of posts published by user u on week t_i, $rr(p_{t_i})$ is the total number of direct reposts received in the user's lifetime for the set of posts p_{t_i}, and T_u is the sorted list of weeks in which the user u published at least one post.

3.4.2 Time series type detection

Characterizing users based on the exhibited temporal behavior of their efficiency requires to extract automatically patterns out of the generated time series. There are two main families of state-of-the art methods for this task. The first one includes *feature-based* approaches that cluster series based on their kurtosis, skewness, trend, and chaos [116]. The latter one includes *area-under-the-curve* methods [36, 39, 117] that consist into dividing the time series into equally sized fragments, measure the area under the curve in each fragment, represent the time series as a vector of such quantities, and then apply a clustering algorithm over them (specifically, we used k-means). We first tried those state-of-the art methods to cluster the efficiency time series. As we assessed by manual inspection, both, feature-based approaches and area-under-the-curve methods, produce clusters containing highly heterogeneous curves. In addition to that, we tried also a separate approach, proposed few years ago, that transforms the curves through Piecewise Aggregate Approximation and Symbolic Aggregate Approximation (SAX) [66] and then clusters the resulting representations with k-means. The discretization of the time series was performed using the jMotif Java library that includes the SAX implementation from Pavel Senin [86]. Also this method lead to imbalanced clusters, being the 80% of the curves put in one single cluster. The main issue with those approaches is that they have been tested in the past mainly on either synthetic time series or time series without high variations due to human activity. When time series represent the activity of single actors they may have an extremely broad variety of length, shapes, and oscillation of the curve that the mentioned methods do not handle properly.

Even though the produced clusters were noisy, the area-under-the-curve method tended to group together curves in four main clusters, with a predominance of well-recognizable shapes: *increasing*, *decreasing*, *peaky* and *steady*. Some examples of time series for each class are depicted in Figure 3.4 (top). Driven by the qualitative

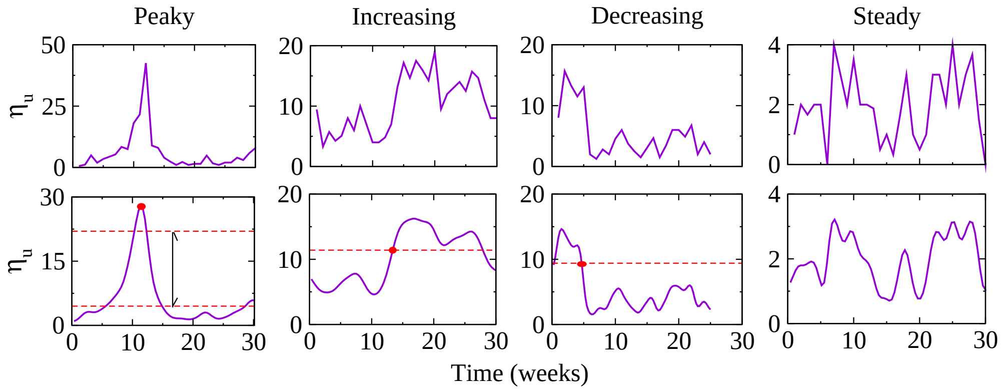


Figure 3.4: *Examples of efficiency time series for users of each class indicated by the clustering of time series. Top: raw time series, Bottom: smoothed Time series. Threshold used to detect changepoints for the first three types are reported with dashed lines*

insights that the clustering produced, we developed a tailored classification algorithm to obtain cleaner groups, based on a qualitative, discrete representation of the temporal data, Inspired by the representation of financial time series presented by Lee et al. [63]. Our algorithm executes the following steps:

1. **Smoothing**. Apply the kernel regression estimator of Nadaraya and Watson [47] to the user temporal data to obtain a smoothed time series t. The smoothing process gets rid of very sharp and punctual fluctuations, which are very frequent in human activity time series. Examples of raw curves compared to their smoothed versions are shown in Figure 3.4 (bottom).

2. **Linguistic transform**. Generate a qualitative representation of the time series t for a user u using three states: High, Medium, Low (H, M, L). We empirically set the threshold for high values to 0.6 and for medium values to 0.3 (i.e., values greater than the 60% of the maximum efficiency reached by the user are considered High). The idea of using threshold values is supported by previous work in time-series segmentation [3].

3. **Fluctuation reduction**. Search for contiguous subsequences of a given state and drop the subsequences whose length is less than the 10% of the total length. Similarly to the smoothing procedure, this step helps to eliminate noisy fluctuations in

the time series. For example, in the series $HHHMHHMMMLLL$, the fourth element, M is dropped.

4. **String collapsing**. Collapse the string representation of t by replacing subsequences of the same state with a single symbol of the same type. For instance, the resulting series from the previous example, $HHHHHMMMLLL$, is transformed to HML. The goal here is to discover those time series with a clear change in their pattern either moving from H to L or in the other direction without changing the character of the data being analyzed. Therefore, in the next step we will only consider these two types of changes.

5. **Detection of Increasing/Decreasing classes**. Look for collapsed sequences with just two groups of symbols and classify as “Increasing” a sequence transitioning from L or M to the state H and as “Decreasing” those transitioning from H to L or M. The second and third columns in Figure 3.4 show the threshold for High values as a dotted red line.

6. **Detection of Peaky class**. For the unclassified series, find those exhibiting a peaky shape by looking at outliers in the series whose value is higher than x times the average value. This method has been successfully used before in the context of Twitter, with $x = 5$ [65]. Other methods for peak detection we tested [85] find just local peaks, which are very frequent in noisy time series.

7. **Detection of changepoint**. Accurately locating the point in which a curve transitions between different levels is important to study the behavior of users in their single activity and popularity metrics around the point in time when these changes occur [8]. For the peak type curves, the changepoint is intuitively defined by the highest peak, whereas for the increasing and decreasing types the point is identified by the time in which the linguistic representation of the series transitions from H to M or L status (decreasing) or from L or M to H status (increasing). We will

elaborate on the fitness of the selected method compared to statistical ones in the next section.

8. **Detection of Steady class**. The remaining time series are classified as steady.

As in most previous work [67], in absence of an automatic way to compute the quality of the classes, two of the authors annotated a random sample of $1,000$ time series per class to assess the goodness of our algorithm. Since the expected shapes of the curves for each class are very clear (see examples in Figure 3.4) a human evaluator can decide with certainty whether the instances from the sample match the expected template. The outcome of the labeling is very encouraging, with 93% correct instances in the Decreasing class, 86% in the Increasing, and 85% in Peak, and almost perfect agreement between evaluators (Fleiss $\kappa = 0.80$). For the Steady class, where shapes can vary much, we labeled as *misclassified* any curve belonging to the other classes. We found a low portion of *misclassifed* instances (12%). We observe that the users in the steady class are around 44%, meaning that 56% of the users exhibit a temporal footprint of the efficiency curve that has a clearly defined trend. This is a finding with important implications on the applicative side, meaning that the majority of users could be accurately profiled as having consistently increasing or decreasing efficiency patterns. Companies, for example, could be more interested in targeting consumers with an increasing efficiency pattern when monitoring the communication among consumers in social media sites [70]. However, we are more interested on finding geographical patterns lying behind the social connections of efficient users.

3.4.3 Changepoint detection

Accurately locating the point in which a curve transitions between different levels is important to characterize the user behavior when his efficiency significantly increases or drops, thus allowing to study how single activity and popularity metrics vary when these changes occur [8]. Changepoint detection refers to the problem of finding time instants where abrupt changes occur [8]. Except for the steady time series, which denote a user behavior that is quite constant in time (or for which transition to higher or lower efficiency levels are much slower), all the other three types have a changepoint in which the efficiency trend changes radically in a relatively short period of time compared to the total length of the user lifetime. For the peak type curves, the changepoint is intuitively defined by the highest peak, whereas for the increasing and decreasing types the point is identified by the time in which the linguistic representation of the series transitions from H or M to L status (decreasing) or from L or M to H status (increasing). More general methods to identify changepoints relying on the changes in mean and variance have been proposed in the past. For the sake of comparison, we match our simple technique with the statistical change point analysis recently proposed by Chen et al. [23]. We find that, although for most time series the values from the two methods were very close (at most 1 or 2 weeks difference in around 80% of the cases), the statistical changepoint detection often identifies time points before or after a change of efficiency and not in the week when the actual change occurs. In fact, the generality of statistical methods is not a plus in cases in which the set of curves in input is quite homogeneous and for which ad-hoc methods result more reliable. For this reason, we use our definition of changepoint.

Once users with similar profiles in their temporal efficiency evolution have been grouped, time series are analyzed to identify meaningful changepoints.

3.5 User efficiency classes

For each detected class, we perform an analysis in aggregate over all the users first and then we characterize the evolution of the same metrics in time. We find that i) publishing interesting content helps to boost the efficiency of the subsequent posts through attention gathering and that ii) the efficiency gained in that way should be sustained by intense social activity to avoid it to drop.

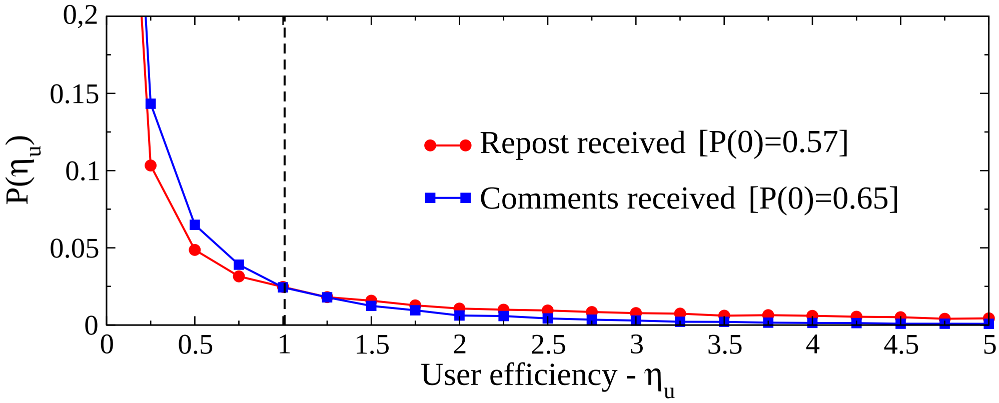


Figure 3.5: *Distribution of efficiency scores, bucketed in 0.25-wide bins. Average scores are* 0.38 *for comments and* 1.55 *for reposts.*

		Activity			Attention				Time	
Type	%users	*pd*	*cd*	*fwee*	*cr*	*fw*	*rr*	*cs*	*days*	*weeks*
Decreasing	15%	6.11	2.78	10.7	4.90	3.57	25.3	34	491	53
Increasing	16%	10.3	4.74	9.69	6.14	4.82	43.4	51	690	92
Peak	25%	8.10	2.74	6.82	4.07	3.18	9.11	32	703	85
Steady	44%	8.22	3.75	10.3	5.50	4.35	29.1	40	610	72

Table 3.3: *Activity, popularity and longevity indicators for the four user classes. Values are the median of the average weekly values. Abbreviations used are pd=posts, cd=commentsDone, $fwee$=followees, fw=followers, rr= repostsReceived, cs=cascadeSize, $days$=userLifetime, $week$=activeWeeks*

3.5.1 Static analysis of user classes

We aggregate different activity and attention indicator scores over users and weeks, for each of the four user classes. For all the indicators, we compute their average value per-week for every user and then we compute the median of all the results obtained for users of the same class. Median is used instead of average to account for the broad distribution of values.

In addition, to get a measure of the adhesion of users to the service, we measure the median number of weeks of activity and the median number of days of duration of the user account. Values for all the metrics are shown in Table 3.3, Figures3.6 and 3.7 that show a first picture of the levels of activity performed and attention attracted by users of different classes. Users in the Increasing class have the highest values for

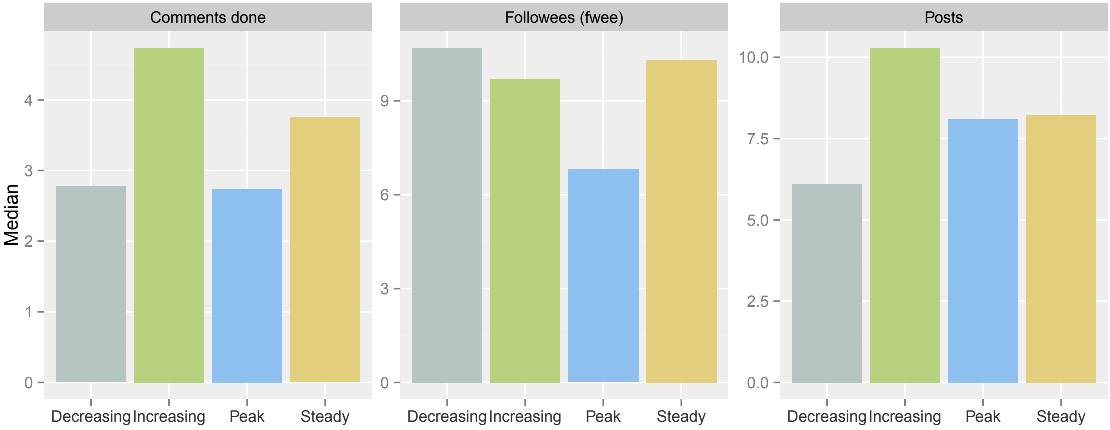


Figure 3.6: *Activity indicators for the four user classes. Values are the median of the average weekly values*

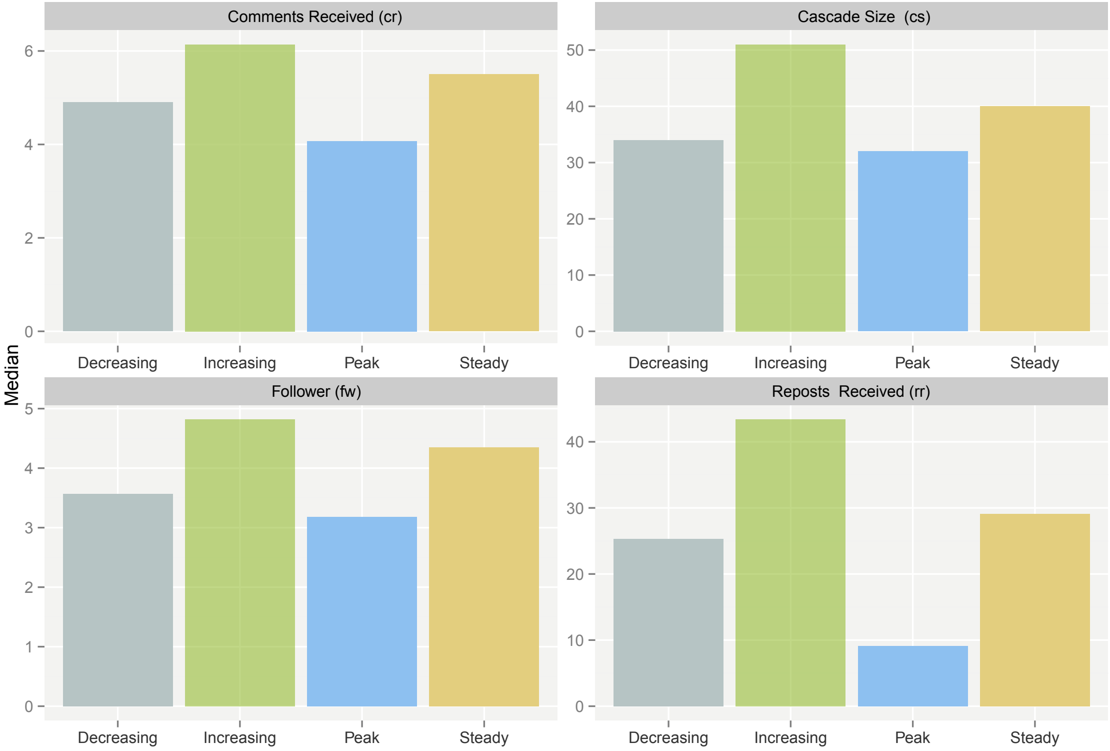


Figure 3.7: *Attention indicators for the four user classes. Values are the median of the average weekly values*

almost all the metrics compared to other groups. They are able to attract high levels of attention (fw,rr), combined with the ability of conciliating the production of content of high interestingness for the community (high cs) with social activity (high cd and $fwee$ values). As we will show later, the production of comments and addition of followees is a characteristic of this class through time. On the contrary, users belonging to the Peak class are the least active in terms of social activity (low cd and $fwee$ values) but, surprisingly, they are relatively active content publishers and have the tendency to be active for long time, exhibiting a high number of active weeks and the highest account duration. They are quite involved in posting but are not much engaged in the social interactions that complements the content production and consumption process. As we will observe next, these users do some commenting activity at the beginning of their lifetime but they reduce significantly the number of followees or comments rapidly. Users in the Decreasing and Steady classes receive both a good amount of attention and establish a high number of social links, backed up by a high content-production activity in the Steady case. Given the shorter time of involvement and knowing about their sharp efficiency drop, the users in the Decreasing class are likely people with a good level of participation who, differently from the users in Steady, reduced significantly the involvement in the service at some point.

3.5.2 Variation around the changepoint

Here we investigate deeper how users in each class distribute the amount of activity in time. We perform an analysis around the changepoint of the efficiency curve, and see if the different temporal patterns can explain *why* their efficiency level changed over time. We decompose the timeseries into different *phases* and study the relations between them in terms of the activity and attention indicators, specifically, for all the users belonging to the classes where the changepoint is given (i.e., all but the Steady class).

Let us define three user-dependent time steps: the week in which the user activity

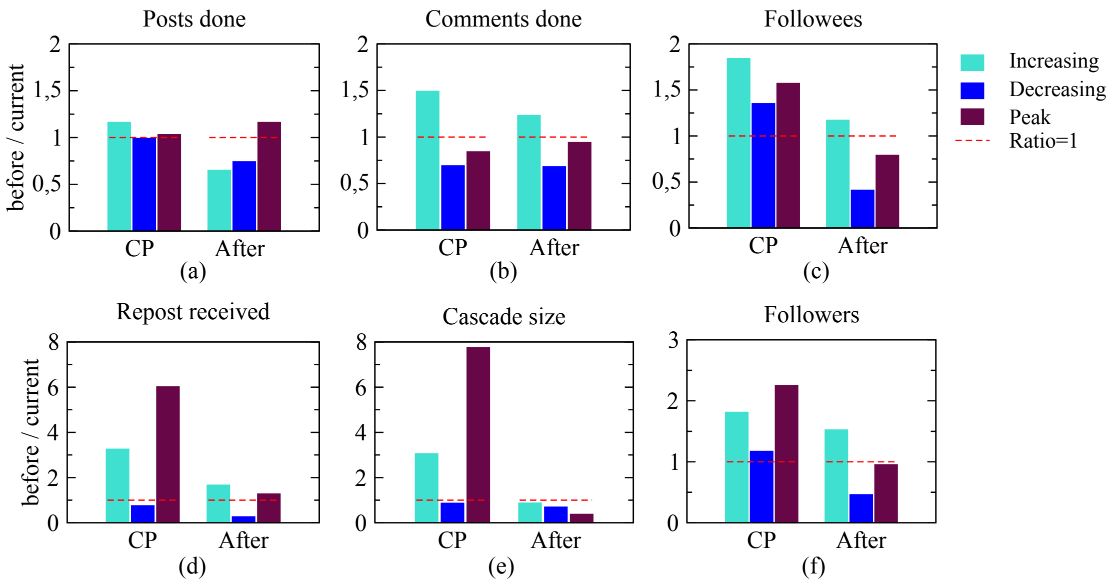


Figure 3.8: *Ratio of activity and attention metrics between the Before phase and later phases (Change Point and After), for the 3 user classes.*

started w_{start}, the week of the changepoint of the efficiency curve w_{cp}, and the week of the end of the activity w_{end}, after which no other action is performed by the user. Accordingly, we define three *phases* of the user lifespan referred as *Before*, *CP*, *After*,which represent, respectively: the weeks in the [w_{start},w_{cp}) interval, the changepoint week w_{cp}, and the weeks in the (w_{cp},w_{end}] interval. We calculate the average weekly amount of activity and attention metrics during these three macro-aggregates of weeks. The three values obtained for each indicator capture the variation of activity and attention when approaching the critical point in which a consistent change of efficiency is detected. To implement such calculation we used a Hadoop Streaming pipeline whose input constitutes all the user activity and the week of the changepoint. The pipeline aggregates all the activity and attention values for each time step defined before and returns three values for each

To detect the variation of the values in the three phases we compute two ratios for each user: a) RatioCP= activity-or-attention metric measured in w_{cp} divided by the same metric computed in $[w_{start}, w_{cp})$, and b) RatioAfter= activity-or-attention metric measured in $(w_{cp}, w_{end}]$divided by the same metric during $[w_{start}, w_{cp})$. Ratios are then averaged over all the users of each class. Comparison of ratios between different user classes reveals the key differences between them: values above 1 mean that the value of the indicator grew in *CP* of in *After* phases compared to the *Before* phase. Final results for different values of activity and attention are reported in Figure 3.8. For instance, in Figure 3.8(a), we observe that RatioAfter is above 1 just for users in the Peaky class. It means that the users in that class have published more posts after the changepoint than they did before it. We can summarize our findings as follows:

- **Activity and attention at *CP***. Users of all classes maintain a similar trend in the number of posts done in *CP* with a slight increase in the case of the Increasing class (Figure 3.8(a)). The three attention metrics at the bottom of Fig. 3.8 reveal more about trend changes. For Peak and Increasing classes, the number of reposts received, cascade size and followers increases significantly in *CP* compared to *Before* (Figure 3.8(d),(e),(f) respectively). Since reposts received and cascade size are proxies for content interestingness, this indicates the production of content that attracts the attention of a much higher number of users. For both classes, this is the most likely cause of the rise of their efficiency at *CP*. For the Decreasing class the attention values start dropping instead. Finally, differently from other classes, users in the Increasing class produce a higher number of comments in *CP* (Figure 3.8(b)).

- **Social activity after *CP***. In the *After* phase, social interaction such as the number of comments and the addition of new followees considerably increase compared to *Before* for the Increasing class (Fig. 3.8(b,c)), while they remain stable or in slight decrease for the Peak class. Decreasing class values drop also in this case.

- **Content production activity after *CP***. The reverse scenario is found when looking at the posting activity. In the *After* phase, Peak post messages at a higher rate than Before (Figure 3.8(a), while *Increasing* posting activity drops in favor of a higher attention to social interaction.

The main lesson learned from the above findings is that the submission of pieces of "interesting" content, namely posts that attract the attention of a wider audience than usual, is the trigger to transition to higher efficiency levels. However, efficiency cannot be maintained without cost. Increasing engagement in social activity and expanding the potential audience turns out to be an effective strategy not to lose efficiency. Conversely, producing more content without reinforcing the social relationships with the potential consumers of the content results in a rapid drop of efficiency to the original levels. The difference between the Increasing and Peaky classes is particularly striking, having the Increasing-type users fully exploiting social activity with 17% more followees, 23% more comments and 61% reposts after their changepoint, while Peaky-type users keep their activity approximately stable (except for an increase of reposts done). Moreover, as expected, when a status of equilibrium between attention received and activity is disrupted by an arbitrary reduction of productivity and social interactions, the efficiency is destined to fade quickly.

3.6 Conclusions

We explored the interplay between activity and attention in Yahoo Meme by defining the notion of user *efficiency*, namely the amount of attention received in relation to the content produced. We find that, unlike the raw attention measures, efficiency has strong negative correlation with the amount of user activity and users who are involved in social activities such as commenting, have higher centrality in the social network than average, but are not necessarily heavy content producers.

However, if we consider commenting as a form of content creation, we observe that comment takes less effort than creating a post but, frequently, it can be more effective. It is so because the reciprocity plays a role and the comments network exhibits a higher reciprocity than that of the follower network. Users can, thus, benefit from the visibility of a post whenever they comments on it.

We classify into four main classes (sharp increasing/decreasing steps, peaks or stable trend) the time series of user efficiency with a novel algorithm that overcomes limitations of previous approaches and we find four main clusters. By analyzing the variation of activity and attention around the changepoints of the timeseries, we find evidences that user efficiency is boosted by a particular combination of production of interesting content and constant social interactions (e.g., comments). In these cases, users gather the attention from a wider audience by publishing content with higher spreading potential and then they manage to keep the attention high through regular and intensified social activity.

A clear temporal footprint of the 'efficiency' has been observed for more than half of our users, thus, the attention online is not given to random people. We have shown that some individuals are more effective at social interactions in the network. Such effectiveness follows stable patterns across time. Given that the increase on efficiency is usually associated with the individual publishing engaging content, we now ask ourselves, is there any spatial pattern associated with the people who produces such interesting content and have the ability to attract attention with continuous social interactions? Moreover, sociologist theories about social network connections named 'social capital' say that productivity in a society increases when people connect among themselves giving attention to each other [87]. Therefore, we will look at the correlation between the attention received by individuals in the city and the city's economic performance. In the next chapter, we propose methods to quantify the same attention metric, i.e. content reposted by others, at the urban level. Our online attention analysis moves from the individual level presented here to the city level presented next, adjusting to the different levels of granularity suggested by Neal *et.al.* [79] for analyzing city data: individual, intra-city, inter-city.

What is the city but the people

Shakespeare, 1607

4

Tracking socio economic indicators from online attention

4.1 Introduction

The exponential growth of online sharing content platforms has been followed by numerous studies of the individual users' ability to attract attention online [5, 20, 40, 78, 93]. However, the spatial patterns of online attention have been overlooked. The geographic annotations of online content open up new possibilities: interactions on the Internet can be exploited to characterize not only individual users but also geographical places by aggregating the attention received within a given city's limits. Therefore, online data together with appropriate metrics can be used to estimate indicators specially in developing countries where the quality of government digital data might be very poor [77].

Developing countries are experiencing increasing rates of urbanization. The 1.4 billion people living in the developing world's cities are expected to increase by 96 percent by 2030, according to the report published by the *World Bank and International Monetary Fund* this year[1]. Urbanization will exacerbate the problem of intra-urban data unavailability on those countries. Consequently, it is important to propose new methods to profile city areas with new sources of data.

In most developing countries, economic indicators at city level are often outdated [49]. A way to solve this problem is to estimate cities' indicators from online data. Previous studies have shown that one could partly track socio-economic indicators from digital data, and do so in a timely fashion. Eagle *et al.* [31] analyzed a mobile phone calls network in UK showing that user's network diversity is associated with economical advantage. More recently, researchers showed that the sentiment extracted from tweets is correlated with the economic well-being of London neighbourhoods ($r = .37$) [90]. Yet, those studies have been conducted only in developed countries such as USA and UK.

Cities in emerging markets, most of which are in developing countries, are overlooked. These cities are of increasing interest since they will account for nearly 40 percent of the global growth in the next 15 years [28]: the Boston Consulting Group has classified as many as 34 Brazilian cities as emerging markets [55]. We thus focus on Brazil, a fast growing developing country that has become the second biggest market, outside US, for social media sites such as Twitter[2]. We get hold of social media data from Yahoo Meme, a social media platform extremely popular in Brazil, and examine the relationship between socio-economic indicators and levels of attention paid to content produced by the residents of different cities, where attention is defined as the

[1] http://www.worldbank.org/

[2] http://thenextweb.com/twitter/2013/01/16/twitter-to-open-office-in-brazil-its-second-biggest-market-after-the-us-in-accounts/

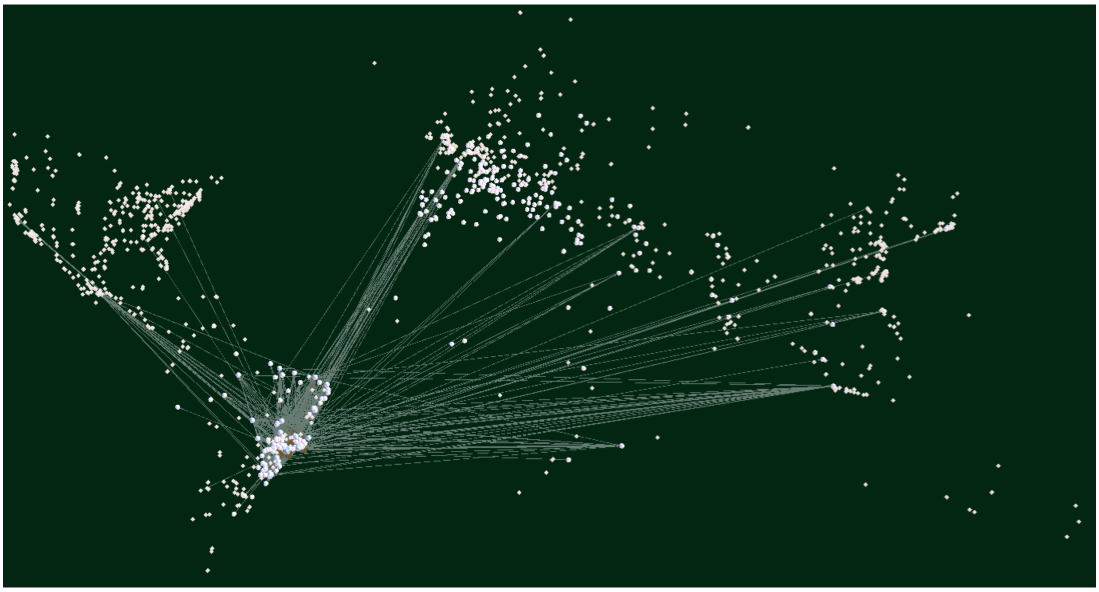


Figure 4.1: *Graph showing the geographical locations of worlwide users who paid or received attention (i.e., reposted content) to/from cities considered in this study. Edges with low weights are not shown.*

interest raised by user-generated content (as reflected by reposting content).

To conduct our analysis, we build upon the concept of *glocality* [120, 121], the combination of global and local interactions in which a city is involved. We propose indicators to estimate the *glocality* of a city by studying interactions between global and local users. In particular, we instantiate the concept of interactions going beyond simple activity measures by considering the *attention* received, collectively, by a city's residents on the platform (attention on individual posts is aggregated at the level of city). In so doing, we make three main contributions:

- We propose a set of online attention metrics that act as a proxy of the city *glocality* by quantifying the ability of its residents to interact globally while maintaining strong local links. We correlate a city's *GDP* per capita with the attention received by their residents and find that it is correlated with local attention and global attention.

- We test the hypothesis that people with access to higher levels of social capital obtain more attention because they have better access to economic growth opportunities [68]. We compute a city's prestige index [68], which is a measure of the mixture of people with different occupations and has been previously used to measure a city's social capital. We build this index from census data and find that it is positively correlated with one of the attention indicators.

- Finally, we put together the proposed online indicators to predict the economic and social capital of cities. We find that our models fit well the data and predict GDP per capita with $Adjusted\ R^2$ = 0.47 and social capital with $Adjusted\ R^2$ = 0.93.

4.2 Dataset

Yahoo Meme was a microblogging platform, similar to Twitter, with the exception that users can post content of any length or type (text, pictures, audio, video), being text and pictures the more frequently posted content. In addition to posting, users could also *follow* other users, *repost* others' content, and *comment* on it. In this study, we use a random sample of interactions on Yahoo Meme from its birth in 2009 until the day it was discontinued in 2012 (Table 4.1). Despite its moderate popularity in USA, Yahoo Meme was popular in Brazil, as witnessed by the fact that the top 45 cities in terms of number of interactions are all located there. Reposting was the main activity in the service (22M sample records) compared to comments (4M). We extract the users who posted the content in our sample and georeference them based on their IP addresses using a Yahoo service. We remove the users for whom we did not obtain results at city level (e.g. users employing proxy servers to connect to the Internet) obtaining 80K users. For this set of users and their respective posts, we extract all the repost *cascades*.

To attain geographic representability, we ascertain that the number of users in the top Brazilian cities in our dataset is significantly correlated with the number of Internet users (Figure 4.2). As a result, we conduct a correlation analysis with the set of 35 cities (without outliers) and we report the results of the predictive analysis for the two sets:

Property	Value
Number of users	80K
Number of posts	13.1M
Number of reposts	22M
Number of comments	4M
Number of reposts cascades	1.4M
Number repost edges between cities	25K

Table 4.1: *Yahoo Meme dataset statistics. Number of records sample from the Yahoo Meme dataset for different types of interactions.*

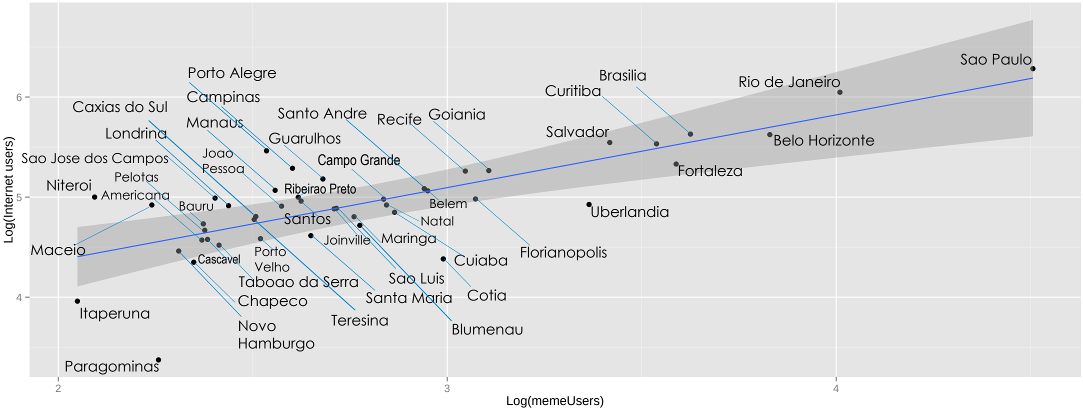


Figure 4.2: *Number of users in our sample versus number of Internet users. Both quantities are log-transformed.*

the 35 cities inside the confidence area and the complete sample of 45 cities (including outliers). We will see that such a number grants statistical significant results. That is because we are left with 1.4M repost cascades whose original content was produced in the 35 cities and was consumed across the world (Figure 4.1).

4.3 Glocal: Global+Local

Glocalization is a concept that refers to the combination of local and global interactions as two sides of the same coin. Barry Wellman used the term *glocal* to qualify communication patterns observed over interactions through the Internet [120, 121]. Wellman states that the Internet influences the way we interact with, or obtain resources from, other people, enabling changes in our ‘network capital’. Online interactions enrich this *network capital* by strengthening *local* links and providing access to *global* information and to distant circles: people who use more the Internet both know better their neighbours and have a higher number of distant ties [120].

Glocality not only characterizes people with a strong online presence but also successful cities. Prosperous cities are associated with rich local and global interactions: London, for example, has been characterized as a city where the interweaving of local and global is intense [30]. In our case, interactions between people take place online and consist of generating and reposting content: one user is publishing content and another user is paying *attention* to it. We use the ability of attracting *attention* to derive metrics that characterize cities. In this section, we present the definition of attention, justify attention as the base of our metrics, and propose metrics for the global and local dimensions of attention received by a city’s users.

From interactions to city-level attention. Attention is the currency used by members of social media platforms to either reward the effort of producing new content or manifest interest in what is published. Due to the ever increasing volume of content and the cognitive/time limits of the information consumers, attention has become a scarce and valued resource. Quantifying the amount of attention attracted by the city’s residents will help to characterize the way in which geographical communities wealth influences the online interactions of the individual users.

In a similar way than we did in Chapter 3, we focus on content transmission reflected by the act of reposting. The choice of attention over activity will prove to be fruitful: we will show that GDP indeed correlates with attention metrics but not with activity ones (Section 4.4). To quantify the attention received by users, two graphs are built:

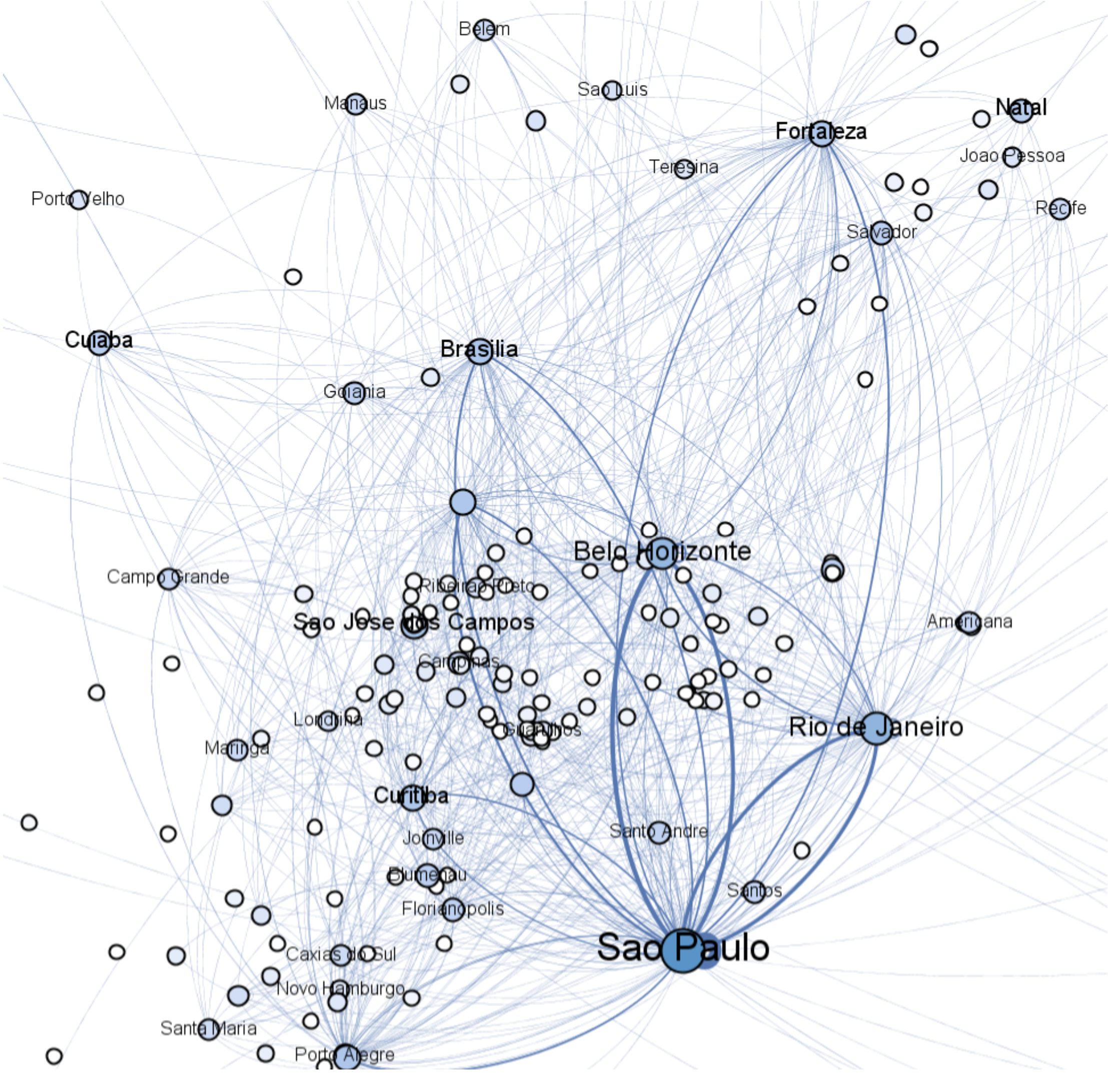


Figure 4.3: *Attention graph whose nodes are cities and whose weighted edges reflect the intensity of reposting between cities' users.*

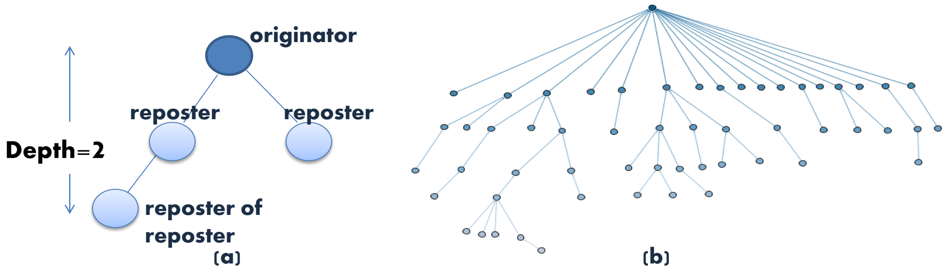


Figure 4.4: *Tree-like repost cascade. On the left, there is an example cascade, which is rooted at the content originator and connects those who, in turn, repost the content. On the right, a real repost cascade from our dataset.*

Attention graph. The city *attention graph* is built using reposts interactions. This is a weighted directed graph where nodes are cities, and directed weighted edges (i, j, w) represent the volume w of reposts between city j where the *reposter* lives, and city i where the original *poster* lives. In this graph, self-edges are allowed as many reposts occur between users living in the same city. The resulting *attention graph* has 1,310 nodes and 25K weighted edges. Figure 4.3 has been generated using Gephi .

Cascade graph. A tree for each post is also built(Figure 4.4). The tree's root represents the original poster and its edges connect those who have reposted that content at different points in time. We analyze 1.4M trees with average depth of 3.41.

These two graphs are used to quantify attention with metrics described next.

4.3.1 Global attention

Cities that enable global information flow are key actors in the world economy [95]. Such cities are, for example, the chosen place for the headquarters of international firms, or the destination of mass tourism. These cities do not exist in isolation [95], as they have strong connections to other cities. In this section, we elaborate on the importance of world-class cities to connect with each other and broker information.

Rest of the World. In his book titled "The triumph of the city", Glaesser showed that Brazil, China and India are very likely to become far richer over the next fifty years [41], and this wealth will be created by cities that are connected to the rest of the world and not by those that are isolated. Cities are connected to the rest of the world through the flow of people (e.g., migration, tourism, business), goods (trade), information (e.g., news) and knowledge (e.g., scientific collaboration). These types of flow foster economic growth in different ways: transactions between immigrants and their home towns, international markets for local products, cultural exchange or improvement of business practices through the transfer of scientific knowledge to local industries. To paraphrase these intuitions in our context, we define our first global attention metric for city i. This is called Rest of the World's attention paid to city i (ROW_i) and is defined as the number of reposts that city i has attracted from the rest of the world or from other Brazilian cities, normalized with respect to the total number n_i of users in that city:

$$ROW_i = \frac{out_i}{n_i}$$

,

$$BR_i = \frac{out'_i}{n_i}$$

, where out_i is the number of times a post originated in city i has been reposted outside it (the world excluding Brazil); out'_i, instead, counts the reposts received outside the city but inside Brazil.

Brokerage. Cities that foster global flows not only have good network connectivity but they may also connect other cities with each other. Sao Paulo, for example, is a strategic place for firms that want to join the Brazilian emerging market. Short *et al.* have named these cities 'gateways cities'. Such places foster globalization while taking advantage of their position for their own growth [101]. We quantify the gateway capacity of a city with brokerage attention. This captures the extent to which a city mediates the

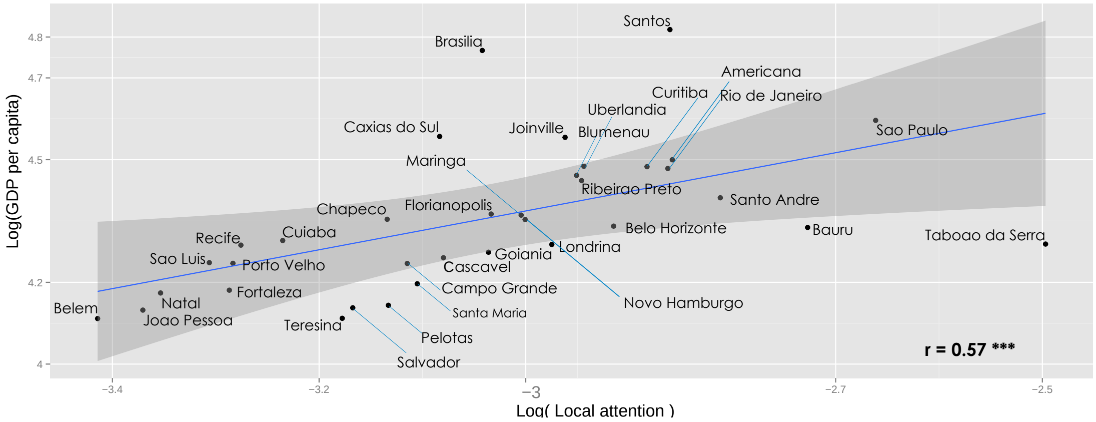


Figure 4.5: *Local attention of a city vs. GDP per capita. In more prosperous cities, community members will devote considerable attention to content produced locally (p-value* < 0.001 *is expressed with* $***$*).*

flow of information to other cities. One way of quantifying such a tendency is to take the city *attention graph* G as defined earlier (Figure 4.3), and compute centrality measures: $brokerage_i = centrality(G, i)$, where *centrality* is a function that returns one of these three metrics : *eigenvector centrality*, *betweenness centrality* and *PageRank* for the graph G and the city i. The selected centrality measures extend the concept of the degree centrality according to which the most important actors in a directed graph would be simply the ones with a higher number of incoming links. Table 4.2 shows a brief definition of the measures and their meaning in this context.

Cascade. The last metric quantifies the ability of a city's users to produce content that spreads far away in the social graph. We take all the posts originated in city i and, for each post k of those, we build a cascade graph (described in the previous Section) and compute the longest direct path in it (max_depth_k). Depth of the diffusion tree of a post indicates multiple levels of exchange and constitutes a signal of successful information diffusion for that post [11]. Given the skewness of the distributions, we

Centrality metric	Definition	Meaning for the city graph
Eigenvector [14]	It is calculated by finding the eigenvector of the adjacency matrix of the graph G with the largest eigenvalue.	It quantifies the importance of a city i depending on the importance of the cities it connects to.
Betweenness [35]	It is the ratio of the number of shortest paths that pass through node i.	It quantifies the importance of city i by the number of times that it mediates the flow of information with other cities.
Pagerank [83]	Pagerank scores the node i in the graph using the probability of visiting i in a random walk of the graph,i.e. it corresponds to the stationary distribution of a random walk on the city graph.	It gives higher scores to the cities with either many nodes or some important ones pointing to them.

Table 4.2: *Social network features to quantify the city i's Brokerage.*

use the geometric average to aggregate the depth values at the level of city.

$$cascade_i = \left(\prod_{k \in P_i} max\_depth_k\right)^{1/|P_i|}$$

where P_i is the set of posts whose producers live in city i.

4.3.2 Local attention

Successful cities not only offer their residents opportunities for global connections but also foster local connections by, for example, having a variety of 'third places' (e.g., coffee places, gyms) where people gather and enjoy the company of neighbors or even strangers [62]. More generally, the intervening opportunities (number of opportunities available in a geographic place) hypothesis [104] states that the number of persons moving to a given distance is inversely proportional to the number of intervening opportunities. Thus, places with dense population (such as cities) offer a considerable number of intervening opportunities and thus encourage interactions at limited distance (local interactions). In the context of attention given to online content, this theory translates into saying that community members will devote considerable attention to content produced

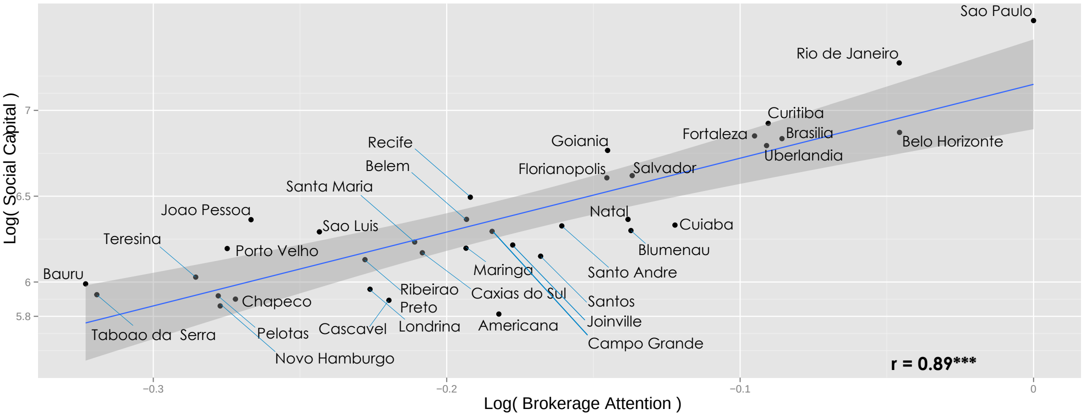


Figure 4.6: *Brokerage attention (Eigenvector centrality) vs. social capital (highest accessed prestige). Cities with higher city prestige indexes do mediate information flow.*

close to the city where they live, if that city offers considerable intervening opportunities, that is, is socio-economically prosperous. To quantify this intuition, given a post originated in city i, we consider its producer and all the actors who expressed interest in it (i.e., reposted it). We compute the average geographic distance between the producer and the consumers, using the Haversine formula that accounts for the spherical shape of the Earth [97], and define the *geographical reach* of a post k as:

$$geo\_reach_k = \frac{1}{|R_k|} \sum_{j \in R_k} d_{ij}$$

where R_k is the set of reposts of k and, for each repost, d_{kj} is the distance between city i (where post k was originated) and city j (where the repost was generated). We compute these values upon the *complete* traces of the reposting cascade, avoiding any data bias. Then, we take all the posts originated in the city i, and aggregate their geographical reach values geo_reach_k using the geometric average. This indicator is considered inversely: the lower the average distance, the more local the attention received.

$$local_i = \left(\left(\prod_{k \in P_i} geo\_reach_k \right)^{1/|P_i|} \right)^{-1}$$

where P_i is the set of posts whose producer lives in i.

We also consider a simpler local metric defined as the number of reposts in_i that the city i has attracted from its residents, normalized with respect to the total number n_i of users in that city:

$$intra\_city_i = \frac{in_i}{n_i}$$

4.4 Attention and GDP

Based on the literature (Section 4.3), we test the hypothesis that *GDP* (wealth creation) positively correlates with the following features of cities:

H1. Attention from ROW. We correlate GDP per capita with *global attention* and find that it positively correlates with ROW, the attention received from the rest of the world ($r = 0.42$) and with ROW', the attention received from other Brazilian cities ($r = 0.38$).

H2. Brokerage Attention. We correlate GDP per capita with each of our three centrality measures, obtaining $r = 0.41$ for eigenvector centrality, $r = 0.39$ for betweenness centrality and $r = 0.30$ for Pagerank.

H3. Attention Cascades. We select the cascades with diameter greater than 1 (i.e., successful propagations at least two hops away) and correlate *cascade attention* with the GDP per capita and obtain a correlation of $r = 0.41$.

H4. Attention from local users. GDP per capita and attention from residents (*local*) are expected to exhibit a positive correlation: indeed they display a positive correlation coefficient of $r = 0.56$. However, Figure 4.5 shows that Brasilia and Santos, for example, perform way better than expected if one were to consider only *local attention* suggesting that other processes explain their success (e.g., Brasilia is the capital of Brazil and Santos is a major port).

> We also correlate GDP per capita with $intra\_city$ and find that they are positively correlated ($r = 0.41$). Thus, the metric $local$ captures better the extent to which a city attracts attention from people in close locations. It is so because $local$ reflects the geographical span of the entire repost cascade whereas $intra\_city$ is limited to the reposts attracted inside the city. Thus, we will use the metric $local$ for the predictions in the next section.

To account for skewness, all the attention metrics are log-transformed before calculating the correlations with each of the 35 cities' GDP. The results obtained are statistically significant, at least with p-value < 0.05.

Why attention and not simply activity. Previous studies have shown that simple activity metrics might not fully capture the production of *quality* content, and that is why we opted for metrics capturing attention. Indeed, if we were to consider the simplest activity measure (i.e., number of posts per capita in a city) and correlate it with the city's GDP, we would find no correlation at all ($r = 0.061$), experimentally supporting our initial theoretical choice.

4.5 Attention and Social Capital

The main idea behind social capital is that social networks have value. "Just as a screwdriver (physical capital) or a university education (cultural capital or human capital) can increase productivity (both individual and collective), so do social contacts affect the productivity of individuals and groups" [87].

There is no consensus on how to measure social capital. Some researchers have measured it through the analysis of participation in volunteerism [87], while others through access to people who might offer diverse opportunities. In this (latter) vein, the 'position generator' developed by the American sociologist Nan Lin in 2001 [68] is often used[3]. This measures the range of people's social ties. Researchers ask their par-

[3] There are different definitions of this specific instantiation of social capital, and Lin's has been widely accepted

ticipants whether they know anyone in 37 different occupations, and consider, for each individual, the occupation with the highest prestige: this is the individual's *highest accessed prestige*. The prestige is measured with the International Socio-Economic Index (ISEI), which scores an occupation based on the level of education required for it and the income it results into [37] (the ISEI score was calculated considering the context of Brazil among other 16 countries). The individual-level definition of *accessed prestige* was previously used at the level of city too [1]: one simply takes the number of city's residents in the different occupations from census data (e.g., X residents being doctors in one city), multiplies each number with the corresponding occupation prestige (e.g., the highest prestige score is indeed for doctors and is 85), and considers the highest multiplication ($85 \cdot X$): this is the city's *highest accessed prestige*. We consider the 2010 data provided by the Brazilian census bureau[4] and compute the *highest accessed prestige* for each of our cities. In a way similar to GDP, we test the following hypothesis regarding social capital:

H5. City's social capital positively correlates with ROW, $cascade$, $brokerage$ and $local$ attention metrics.

We find that a city's social capital does not correlate with the attention that its residents received, quantified with ROW, $cascade$, $local$ metrics (r is not statistically significant). By contrast, as the Figure 4.6 shows, we find that it does correlate with the extent to which its residents mediate the information flow among other cities (r for *brokerage* is 0.86 using *eigenvector centrality*, 0.89 using *betweenness*, and 0.89 using PageRank). It confirms the concept stated by Lin *et. al*, that the social capital of a community is a function of its brokerage opportunities.

together with few others

[4] http://www.ibge.gov.br

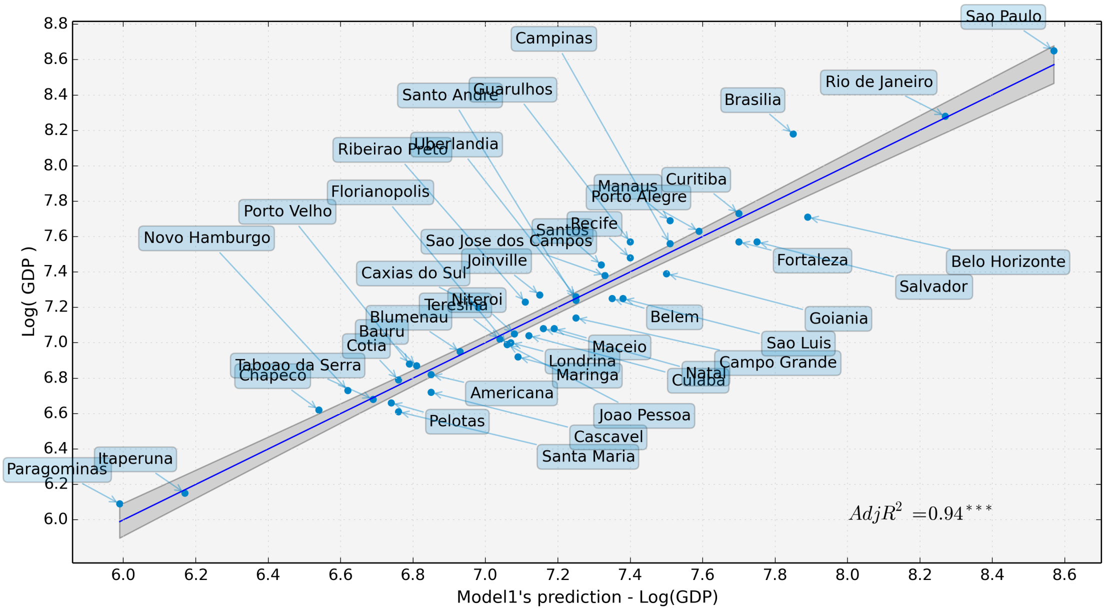


Figure 4.7: *The model 1's performance for predicting GDP per capita with Adj. R^2=0.94. Despite the strong correlation, our model understimates or overstimates the value for some cities (shown outside the confidence region). The model's prediction error is low: its Mean Absolute Error is 0.09.*

4.6 Predicting Socio Economic Capitals

We separately model GDP per capita and social capital of $city_i$ as a linear combination of the four attention metrics plus terms to account for pairwise interactions between indicators (i.e., interaction effects). We control for the city's Internet penetration rate and population (with data provided by the Brazilian census bureau in 2010[5]). We control for those two variables as Internet penetration is associated with online activity, and larger cities tend to be economically prosperous as they enjoy "increasing returns to scale": a city becomes more attractive and productive as it grows [10, 41]. Model1 is defined as follows:

$$\begin{aligned} log(GDP_i) = \alpha + \beta_1 \cdot log(local_i) + \beta_2 \cdot log(ROW_i) + \\ \beta_3 \cdot log(brokerage_i) + \beta_4 \cdot log(cascade_i) + \\ +\rho \cdot log(Population_i) + \mu \cdot Internet_i + \epsilon_i \end{aligned} \tag{4.1}$$

[5] http://www.ibge.gov.br

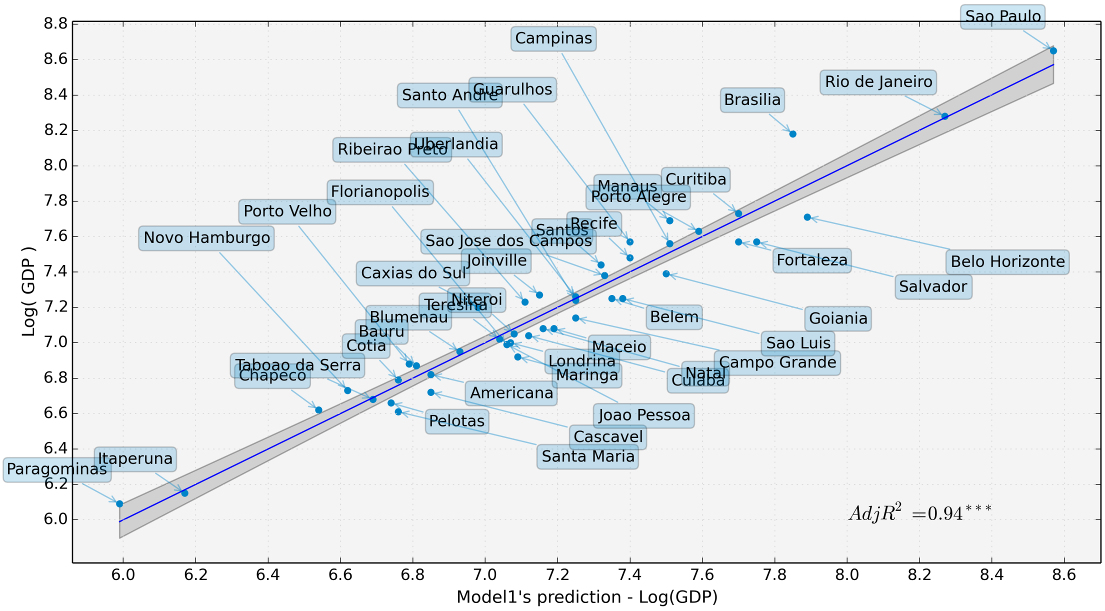


Figure 4.8: *The Model1's performance predicting GDP with $Adj.\ R^2$=0.94 for the set of 35 Brazilian cities.*

where $Internet_i$ is the city's Internet's penetration rate, $Population_i$ is the city's population, and ϵ_i is the error term. To account for the skewness of the data, we log-transformed each variable.

We also build the Model2 adding the $Interactions_{im}$ (Table 4.3) that accounts for all possible pairwise product terms among the four attention predictors. We find similar results for the two sets of cities. Thus, for simplicity, the remaining part of the section describes the results obtained for 35 cities (without outliers).

The models have been built for the two sets of cities in our study and the $Adj\ R^2$ is similar using either of these sets. We observe that, in Model1, the four attention metrics complement the predictive power of the census data, $Adj\ R^2$ is 0.94 (Table 4.3) with a 47.90% of the variance explained by the census data and the remaining 52.10% by the attention metrics.

By analyzing the beta coefficients of Model1, the one with the best performance, we find that *local attention* accounts for 6.86% of the models' explanatory power while the aggregated β values of the three *global attention* metrics contribute with 45.24%.

Model	Predictors	*Adj.* R^2	*Adj.* R^2
		35	45
1	$Population_i + Internet_i + \{Attention_{im}\}$	**0.93**	**0.940**
2	$Population_i + Internet_i + \{Attention_{im}\} + \{Interactions_{im}\}$	0.927	0.939

Table 4.3: *Adj.* R^2 *for different models predicting city* i*'s GDP on the two sets of cities. Model1's uses the four attention metrics* m*, Model2 adds their interaction effects. All models control for the city's Internet penetration rates and population.* $p\text{-}values$ *are* < 0.001.

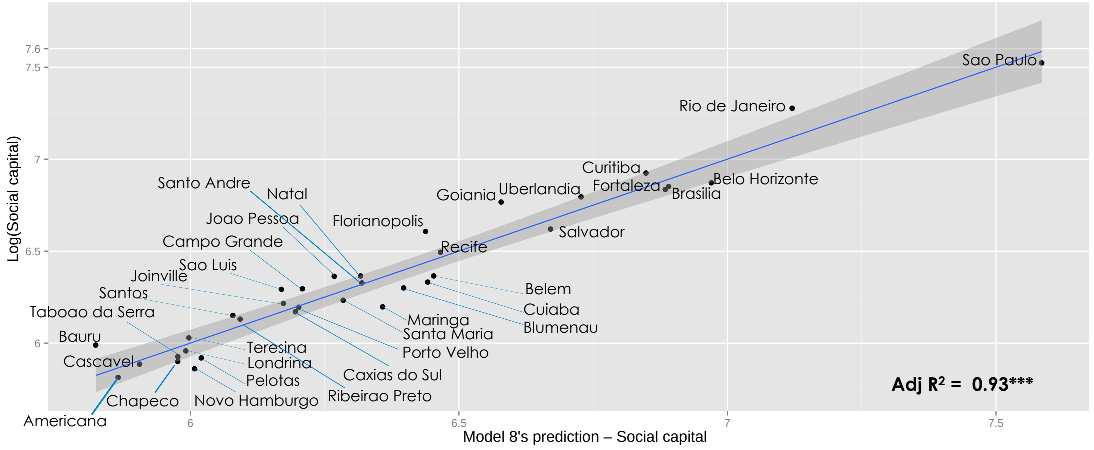


Figure 4.9: *The model3's performance for predicting social capital. Its Mean Absolute Error is as low as 0.08.*

Out of the three global attention metrics, *cascade* attention has the highest impact as it explains 24.50% of GDP's variance. As for Model1's accuracy, the model achieves a Mean Absolute Error (MAE) of 0.09 on a logarithmic scale, where the minimum value is 6.09 and maximum is 8.65, meaning that, on average, the model predicts GDP within 1.48% of its true value, meaning that, on average, the model predicts GDP within 1.48% of its true value. Figure 4.8 plots predicted values against actual ones. The outlier for which GDP is higher than expected is Brasilia (the capital of the country).

We repeated the same linear modeling to predict a city's social capital (M3-5). We found that the model 3, that includes attention predictors plus interaction effects, has the highest *Adj.* R^2, which is equal to 0.93 (Table 4.4), achieving a Mean Absolute

Model	Predictors	$Adj.\ R^2$ 35	$Adj.\ R^2$ 45
3	$\{Attention_{im}\}$ + $\{Interactions_{im}\}$	**0.93**	0.85
4	$\{Attention_{im}\}$ + $Population_i + Internet_i$	0.91	0.86
5	$brokerage_i$ + $local_i$	0.72	0.66

Table 4.4: *$Adj.\ R^2$ for different models predicting city i's social capital. Model3's predictors are the four attention metrics m and their interaction effects, Model4 controls for the city's Internet penetration rates and population, and Model5 tests brokerage and the local attention metric separately. p-$values$ are* < 0.001.

Error of 0.08 on a logarithmic scale, where the minimum value is 5.8 and maximum is 7.5. This means that, on average, the model predicts social capital within 1.33% of its true value. By computing the beta coefficients of the Model3 we find that the β value for *local attention* accounts for 8% of the explanatory power, while the aggregated β values of the three *global attention* metrics (the most important of which is brokerage) contribute with 31% and the γ values (pairwise interactions coefficients) account for the remaining 61%. Figure 4.9 compares the model's predictions against the actual values, clearly showing the accuracy of the linear model. The model for predicting social capital using *local* attention in combination with *brokerage* (model 5) variables reports the lower performance.

We report the results obtained for models that consider the metric *local* for *local attention*, *ROW* for *rest of the world attention* and *eigenvector centrality* for *brokerage attention* as they exhibit higher correlation coefficients (Section 4.4). However, we repeated the linear modeling considering the alternative metrics(*ROW'*,*intra_city*,*betweeness-Pagerank*) and obtained similar performance.

4.7 Discussion

4.7.1 Theoretical Implications

Our results complement previous studies that correlated economic status with social media data at the level of urban *neighborhoods* [31, 90]. We find consistent results at the level of city too. This is done by considering, for the first time, *attention* exchanged between cities as a predictor of their economic wealth upon an urban sociological framework.

We also show that attention is affected by real-world geographic proximity, thus confirming previous studies on the role of physical distance on online interactions [96]. Additionally, we find that receiving attention from actors residing far, both geographically and in the social graph, positively signals economic well-being. This confirms that users are becoming 'glocalized' [121] taking advantage of the Internet to communicate with both local and long-range ties.

The strong correlation between social capital and brokerage is also of theoretical interest for social network researchers. The result confirms that social opportunities come from diverse social connections not only for individuals (as the strength of weak ties [43] and the structural hole theory [18] would suggest) but also for cities.

4.7.2 Practical Implications

Our work shows evidence that, with online attention metrics, one is able to effectively and cheaply predict economic and social capital indicators (respectively, $Adj.\ R^2 = 0.45$, and $Adj.\ R^2 = 0.93$ in the case of Yahoo!Meme in Brazil). To quantify online attention, we define a set of general and easily interpretable metrics: volume of reposts coming from local users, from global users, and from those far away in the social graph. We also observe the brokerage ability of a city in spreading information. These metrics can be computed from aggregated data, made publicly available by social media companies, without the need for researchers of accessing potentially privacy sensitive data.

The possibility of tracking socio-economic well-being of communities at scale supports the vision behind ‘smart cities’: new information and communication technologies will be needed to promote healthy and socially sustainable communities and, more generally, to better manage complex urban systems. In the spirit of ‘smart cities’, predictions derived from social media data could help city planners in taking the pulse of the economic prosperity without waiting years for census data to be collected. This holds not only for cities but also for countries: studies of the global interconnectedness are often based on how international corporations are linked [108] and can now be informed by how countries connect online as well.

4.7.3 Limitations

This study has three main limitations that call for further work. The first is demographic bias: users of the platform are usually young people and might represent a more affluent segment of the general population.

The second limitation is about language independent features to quantify attention, which were chosen for their generalizability. In the future, one can also analyze the actual content being shared.

Third, our results do not speak for causality, so analyzing different temporal snapshots to potentially observe causal relationships is in order.

4.8 Conclusion

Before it can be used effectively, large-scale data needs to be processed somehow. In line with the emerging discipline of web/data science, we opted for a methodology that makes use of well-established theories in urban sociology to produce actionable data analytics. We have shown how those theories could be put to use to take the pulse of developing urban economies. We have determined which online attention metrics are useful proxy indicators of economic capital and social capital. This contribution is just the tip of the iceberg when it comes to exploring the uses of large-scale data for social good. There is a growing interest in using digital data for development opportunities, since the number of people using social media are growing rapidly in developing countries as well. Local impacts of recent global shocks - food, fuel and financial - have proven to not be immediately visible and trackable, often unfolding “beneath the radar of traditional monitoring systems” [110]. To tackle that problem, policymakers are looking for new ways of monitoring local impacts, and tracking online attention might well be one such way.

There is no stone in the street and no brick in the wall that is not actually a deliberate symbol a message from some man, as much as it if were a telegram or a post card.

G.K. Chesterton, 1923

5

Discovery and Automatic Annotation of Functional Areas in the City

5.1 Introduction

By 2025, the majority of the people world-wide will have access to mobile phones [99] that facilitate the geographic annotation of the content published online. Given that part of those annotations are done at a intra-urban level, the process of characterizing the city can be extended to find out the way in which the residents or the visitors use predominantly the urban areas. The focus of our study moves now from the inter-city to the intra-city level [79].

A quick understanding of a complex city might be provided by new ways of discovering functional areas. Traditional urban-planning categories contain broad terms such as ‘commercial, residential, educational’ or more specific ones such as ‘bank, school, grocery store’ [64]. Automatic functional areas discovery might benefit a variety of stakeholders: tourists who look for historical sites; locals who are after niche shopping; retail analysts who have to recommend where new brick-and-mortar shops are best placed [56].

After automatically discovering functional areas, one needs to annotate them. However, as we shall see in Chapter 2, area annotation is still “a very challenging problem in traditional urban planning” [126]. That is because existing approaches rely on topic-based inference models or segmentation techniques that can describe an area as a frequency distribution of its representative categories at best. By representative, we mean categories that are most frequent in the area or that occur in that area more than chance.

To partly fix that, we propose a framework for discovering functional areas. Our framework exploits the fact that categories are often (or can be) arranged in a taxonomy and, as such, areas might well be annotated with any node in the taxonomy. In putting forward this framework, we make the following main contributions:

- We frame the area discovery problem in terms of maximization of a simple objective function to be integrated into any clustering algorithm (Section 5.2). This function aims at finding and labeling area such that an area’s label is semantically related to the points in the area and in the area’s neighborhood without being too general (e.g., the label ‘clothing stores’ is preferable to ‘professional places’).

- We evaluate the framework with a hierarchical clustering algorithm upon Foursquare data in the cities of Barcelona, Milan, and London (Section 5.3). We find that it is more effective than baseline methods in discovering functional areas in those three cities.

- We complement our quantitative evaluation with a qualitative one (Section 5.4). We ask 111 participants to suggest areas where to carry out specific tasks (e.g., where to best place a new tech startup, where to go shopping). We then compare the areas they suggested with those automatically returned by our framework.

We conclude by discussing desirable properties of the framework, including its flexibility in defining any type of taxonomy such as a temporal one for Flickr pictures (Section 5.5).

5.2 Problem Statement

5.2.1 Definitions

We are given a map, represented as a graph $G = (A, w)$ with vertex set A and a system of edge weights $\{w_{i,j}\}_{a_i,a_j \in A}$. We refer to the elements of A as "cells". We are also given a taxonomy represented as a tree $T = (V, E)$; we denote $L \subseteq V$ the set of leaves of T. Finally, we are given an initial labeling function $\ell : A \to L$ that assigns a leaf of the tree to each cell.

Our goal is to find a labelling $\ell^* : A \to V$ of the cells to any node in the taxonomy tree such that:

- the labeling generalizes the initial labeling; that is, the new label $\ell^*(a)$ of each cell $a \in A$ should be an ancestor of $\ell(a)$ in T;
- adjacent cells have, to the maximum possible extent, the same label in ℓ^*;
- we do not generalize too much, that is to say, labels closer to the leaves of T are preferable.

There is a natural tradeoff between these objectives. Consider the extreme case in which we label all the cells with the root of the tree: on the one hand we would have perfect homogeneity of labeling, but on the other hand we would have over-generalized. At the other extreme, setting $\ell^* = \ell$ incurs no generalization cost, but adjacent cells will generally have distinct labels.

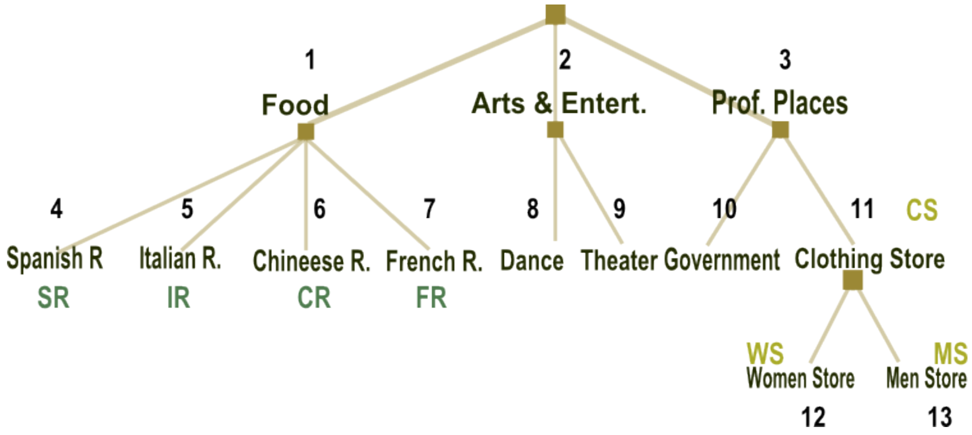
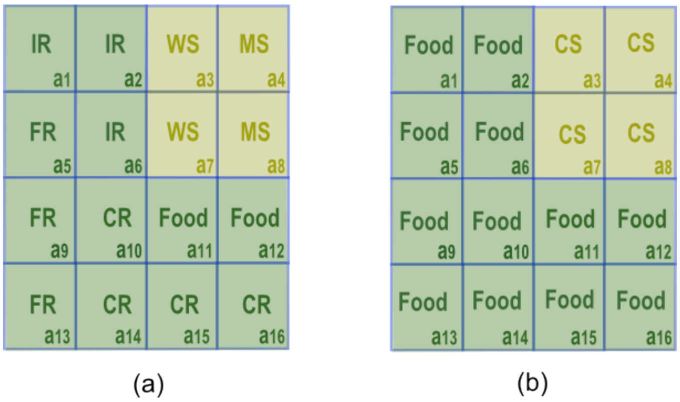


Figure 5.1: *Example taxonomy used to cluster the sixteen city cells shown in Figure 5.2*



IR a1	IR a2	WS a3	MS a4
FR a5	IR a6	WS a7	MS a8
FR a9	CR a10	Food a11	Food a12
FR a13	CR a14	CR a15	CR a16

(a)

Food a1	Food a2	CS a3	CS a4
Food a5	Food a6	CS a7	CS a8
Food a9	Food a10	Food a11	Food a12
Food a13	Food a14	Food a15	Food a16

(b)

Figure 5.2: *Labeling of sixteen cells using the taxonomy shown in Figure 5.1, the initial labeling is shown in (a) and the final labeling using LCA as semantic distance is show in (b)*

One simple solution would be computing a labeling assignment $\ell_0 : A \rightarrow L$ picking the highest frequency label in each cell and clustering cell pairs (a_i,a_j) using, as semantic distance, the distance of their assigned label to their lowest common ancestor (LCA) in T. We illustrate such an approach in Figure 5.2 where sixteen cells are clustered using the taxonomy in Figure 5.1. Although this summarises the idea behind our problem, the simplest solution would tend to assign labels that are too general. We propose a different approach next.

We formalize our problem as follows. Let $adv : V \times A \rightarrow \mathbb{R}$ denote a function representing the advantage of assigning each label to each cell. Given a user-defined parameter $\lambda \in [0, 1]$, we want to find the labeling $\ell^* : A \rightarrow V$ that maximizes:

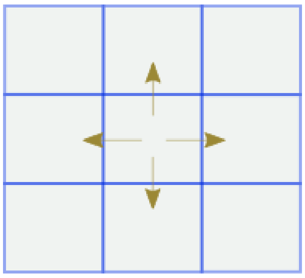


Figure 5.3: *Definition of adjacent cells using the 'rook case'*

$$\sum_{a \in A} \left(\lambda \sum_{b \in A} w_{a,b} \, \mathbb{I}[\ell^*(a) = \ell^*(b)] + (1 - \lambda) adv(\ell^*(a), a) \right) \tag{5.1}$$

For a given cell $a \in A$, the first term measures how well the proposed labeling $\ell^*(a)$ covers a's neighborhood; the second term quantifies how well the proposed labeling covers the cell itself without being too general (i.e., it does not incur into overgeneralization).

Note that we may assume that the graph G is undirected and $w_{a,b} = w_{b,a}$ for all pairs $a, b \in A$; otherwise, simply define a new weight function w' by $w'_{a,b} = \frac{w_{a,b}+w_{b,a}}{2}$.

Choice of edge weights w For our purposes, we define $w_{a,b} = \dfrac{1}{k_a}$ if a and b are adjacent, and 0 otherwise, where k_i is the number of neighbors of a. In our experiments, we use the 'rook case' notion of adjacency (Figure 5.3), in which the four surrounding cells (above,below,left,

right) are considered adjacent.

Choice of labeling function adv First, we define some*coverage functions*.

For a label $l \in V$ and a cell $a_i \in A$, we define $cov(l, a_i) = 1$ if l is an ancestor of $\ell(a)$ in T, and zero otherwise. The *average* coverage of l for the entire map is defined by

$$cov(l, A) = \frac{log10 \left(10 + \sum_{i=1}^{n} cov\left(l, a_i\right)\right)}{log10(n)}$$

We log-transform both numerator and denominator to account for the skewness of the numerator and for the large number at the denominator. Then we define $adv(l,a) = 1 - cov(l,A)$ if $cov(l,a) = 1$, and $-\infty$ otherwise.

Extension to fractional labelings If we assume that each cell is not initially associated to a unique label, but instead has a distribution over the nodes of T, then we are given $\ell : L \times A \to [0,1]$ such that $\forall a \in A$, $\sum_{l \in L} \ell(l,a) = 1$. Given a node of the taxonomy $l \in V$, let $L(l) \subseteq L$ denote the set of leaves contained in the subtree rooted at l. We define the coverage of l for a given cell $a \in A$ by $cov(l,a) = \sum_{l \in L(v)} \ell(l,a)$. Now we define $cov(l,A)$ as before, and

$$adv(l,a) = cov(l,a) - cov(l,A)$$

5.2.2 Relationship to metric labeling

Our problem specializes the *uniform metric labeling* problem, formulated by Kleinberg and Tardos [57]. A series of papers further study the complexity of this problem and its variants [22, 24, 45, 71], as well as its applications to computer vision and image segmentation (see [15, 113] and the references therein).

The goal of the uniform metric labeling problem is again to maximize the objective function in Equation (5.1), but there is no underlying taxonomy tree. Thus, existing solutions to metric labeling apply in our setting, although hardness results do not translate immediately.

Finding an exact solution to the uniform metric labeling problem is NP-hard [15], and remains so even if the graph G is planar [113]. Kleinberg and Tardos [57] give a polynomial-time factor-2 approximation algorithm based on rounding a linear programming relaxation. Boykov *et al.* [15] show a connection to minimum multicut and prove that a greedy approach based on local improvements also achieves a 2-approximation.[1]

Both approaches are computationally intensive. In an instance with n vertices, m

[1] Strictly speaking, their technique does not necessarily stop after a polynomial number of iterations, but can be easily adapted to do so and find a $2 + \varepsilon$ approximation in polynomial time, for any $\varepsilon > 0$.

Algorithm 1 Hierarchical clustering - pseudocode

```
procedure HAC($T, A, \lambda, contr()$)
    for each cell $a_i$ in A do
        Assign $a_i$ to a newly created cluster $C_k$
        Assign label $l$ to $C_k$: $contr(C_k, \ell^*(C_k))$ is max
    end for
    for each pair of adjacent clusters $C_k, C_h$ do
        checkEnqueuePair($C_k, C_h, \lambda, T$)
    end for
    while ($priorityqueue$ is not empty) do
        Get next tuple $(M_{kh}, l_{ij}, contr(M_{kh}, \ell^*(M_{kh})))$
        mergePair($C_k, C_h, l_{ij}, contr(M_{kh}, \ell^*(M_{kh})$)
        Replace references to $C_k$ and $C_h$ with $M_{kh}$
        Update $contr()$ for each neighbor of $M_{kh}$
    end while
end procedure
```

edges and L labels, [57] require solving an LP with $O((m+n)L)$ constraints, while in [15] each iteration requires a maximum flow computation in a suitably defined graph with $O(m+nL)$ edges.

Under certain complexity-theoretic conjectures [71], the best possible approximation factor is $2-2/L$. We propose next a greedy approach and we prove that it is efficient at finding the representative areas of the city.

5.2.3 Our algorithm

To find the areas and labeling for the map that maximize the objective function, we would need to test all possible labels assigned to all possible areas and select the configuration for which the function is maximum. Since that would be computationally prohibitive, we need to find an efficient way of finding a satisfactory area division and labeling. To this end, we modify the hierarchical clustering algorithm. This allows us to merge candidate cluster pairs in incremental ways: by incremental, we mean that each potential merge is independently evaluated and it takes place only if the objective function increases as a result.

We start with the initial labeling ℓ in which each cell is assigned the most popular label inside it. Next, we apply the hierarchical clustering that works as follows (Algorithm 1). Each cell in the map is initially assigned to a new cluster (*line 3*), resulting in as many clusters as cells. For each cluster C_k, since we can select any of the *candidate labels*: the labels present in it and their ancestors in T, we need to compute the contribution to the objective function for each of those labels and select the one that results in the maximum (*line 4*):

$$contr(C_k, \ell^*(C_h)) = \sum_{a_i \in C_k} \Bigg(\lambda \sum_{j \in [1,n]} w_{i,j} \, \mathbb{I}(\ell^*(a_i) = \ell^*(a_j)) + \\ (1 - \lambda)\big(cov(\ell^*(a_i), a_i) - cov(\ell^*(a_i), A)\big) \Bigg)$$

To start merging those clusters, we augment the original algorithm (not tailored to spatial applications) with a simple spatial notion: only *adjacent* clusters can be merged [2]. By testing which clusters are adjacent and which are not, we have a set of cluster pairs that could be potentially merged (*line 6*).

In the $checkEnqueuePair$ procedure (*line 7*), we test whether we are better off in merging the two clusters or in keeping them separate. For each candidate pair, we compute the first cluster's contribution to the objective function, and the second cluster's contribution. The two contributions are computed considering the two clusters' current labels. The contribution of the first cluster C_k is computed with the previous formula over all C_k's cells, and the contribution of the second cluster is computed summing over all cells in C_h. Having those two individual contributions, we are now able to decide whether to merge the two clusters or not. We merge them only if that merging operation contributes to the objective function equally or more than the sum of the two individual contributions; otherwise, the two clusters are best left separate. The contribution of the newly merged cluster M_{kh} is computed with the previous formula: the

[2] One desirable by-product of this restriction is that the algorithm's search space is greatly reduced.

only difference is that the sum is done over all the cells in *both* clusters. That contribution $contr(M_{kh}, \ell^*(M_{kh}))$ changes depending on the label assigned to the newly merged cluster. Since we can assign any of the *candidate labels* (i.e. the intersection of C_k and C_h's *candidate labels*), we need to compute the contribution for each of those labels and select the ones that result in a non-negative merging benefit: that is,

$$\big(contr(M_{kh}, \ell^*(M_{kh})) - \\ (contr(C_k, \ell^*(C_k)) + contr(C_h, \ell^*(C_j))) \big) \geq 0$$

By selecting C_k and C_h, we mean that we put them in a priority queue in which cluster pairs are ordered by their merging benefits.

After putting all cluster pairs with non-negative merging benefits in the queue, we visit the queue by performing ordered merging operations starting with those with highest benefits (*line 10*). At each merging operation (*line 11*), the queue is partly updated (*line 12*): after combining, say, C_k and C_h, we refresh the queue by replacing all references to either C_k and C_k with M_{kh} and updating the contributions to the objective function of M_{kh}'s neighbors. The merging operations end when the queue is empty.

5.3 Evaluation

The goal of our algorithm is to cluster points in a map. The points in the same cluster are geographically close and semantically related. To ascertain whether our proposal meets this goal, we ought to answer two main questions:

(*Area Distinctiveness*) To which extent is our proposal able to group points that are related to each other in the same areas?

(*Labeling Accuracy*) Does the label assigned to each area well describe the area's points?

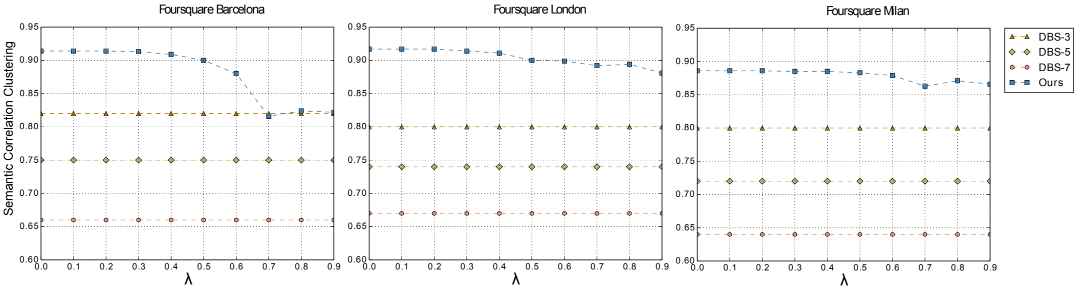


Figure 5.4: *Semantic correlation clustering values for DBSscan and for our proposal as a function of λ in: (a) Barcelona; (b) London; and (c) Milan. The parameters of DBScan are $minpts\{3,5,7\}$= and $\epsilon = 100$ meters.*

Baseline. To answer those questions in a comparative fashion, we need to resort to a baseline algorithm. DBScan is widely used for spatially clustering points but it does not return any label for its clusters. We thus augment it by labeling DBScan's clusters with the most popular label in each cluster. As DBScan has no notion of semantic labeling, its results should be considered to be a lower bound. The algorithm works by iteratively aggregating spatial points into clusters based on a threshold distance ϵ and a minimum cluster size $cmin$. We try $\epsilon = 400m$ and $cmin = \{3, 4, 5\}$, as they are commonly used values [9]. Points that cannot be assigned to any cluster are marked as 'noise.

Foursquare dataset. We use the Foursquare REST Public API and crawl 22K, 60K and 37K venues located within the bounding box of the three cities of Barcelona, London and Milan, respectively. We then consider only the venues with at least 10 check-ins, leaving us with 14K, 30K, 18K venues. Each venue comes with its unique identifier, name, latitude, longitude and category. Those three cities have been chosen for their very different characteristics. From a geodemographic perspective, with such a choice, we explore different population sizes (3,2M for Barcelona, 8,3M for London, and 1,3M for Milan) and population densities (5,060 inhabitants per square kilometer in Barcelona, 4,542 in London, and 7,536 in Milan). From a Foursquare (or, more

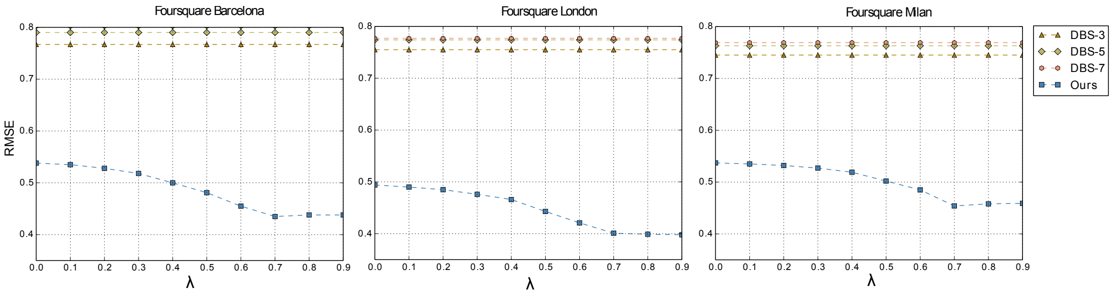


Figure 5.5: *Labeling error (RMSE) for DBScan and for our proposal as a function of λ in: (a) Barcelona; (b) London; and (c) Milan.*

generally, social media) perspective, both Barcelona and Milan are less 'mature' than London, having a smaller user base and smaller number of check-ins. For our experiments, we crawl the entire tree of categories used by Foursquare to categorize venues and divide the bounding box of each city into walkable cells, each of which is initially labeled with its most frequent category and is roughly 100x100 meters in size. Previous research has established that 200m tends to be the threshold of walkable distance in urban areas [25, 82] (in dense parts, this is equivalent to 2.5-minute walk), making our choice of 100m sufficiently conservative.

5.3.1 Distinctiveness

An objective function that measures the quality of the clustering, considered both in the area of theoretical computer science and data mining, is related to the correlation-clustering problem. The values of the correlation clustering CC are experimentally less affected by the number of clusters produced by a clustering method than other clustering measures (e.g., normalized cut):

$$CC = \sum_{\substack{p_i \in C_k \\ p_j \in C_k}} (1 - dist(p_i, p_j)) + \sum_{\substack{p_i \in C_k \\ p_j \in C_h}} dist(p_i, p_j)$$

Since points in the same area should ideally be closer to each other than points in different areas, this measure reflects the quality of the clustering as it increases with the points' closeness $(1 - dist(p_i, p_j))$, if the two points are in the same area; and with the distance $dist(p_i, p_j)$,if the two points are in different areas. To define the notion of semantic distance in a taxonomy, we resort to Jiang and Conrath [54]'s definition:

$$dist(p_i, p_j) = 2\, log(Pr(lca(\ell^*(p_i), \ell^*(p_j)))) - \\ (log(Pr(\ell^*(p_i))) + log(Pr(\ell^*(p_j))))$$

where $Pr(\ell^*(p_i))$ is the occurrence probability of the label assigned to p_i in the city map (i.e., number of points labeled with $\ell^*(p_i)$ over the total number of points), and $lca(\ell^*(p_i), \ell^*(p_j))$ is the lowest common ancestor of the two labels assigned to p_i and p_j.

We compare the semantic correlation clustering values in the three cities of Barcelona, London, and Milan (Figure 5.4). In all cities, we observe the same two main results. First, our proposal consistently performs better than the best DBScan, which is the one using a minimum number of points equal to 3. The values of correlation clustering for our framework are all above 0.90. Second, as λ increases, the semantic correlation clustering stays flat for DBScan (as it does not depend on λ) and slightly decreases for our proposal. In Barcelona, that decrease become noticeable only for $\lambda > 0.6$: yet, after that value, the correlation clustering is still above the best DBScan's values. This suggests that enforcing homogeneity with high values of λ does not impact the distinctiveness of the resulting functional areas in both Milan and London but, to a limited extent, impacts that in Barcelona.

5.3.2 Labeling Accuracy

After testing the distinctiveness of the functional areas, we need to test whether the areas are properly labeled. To assess whether a label assigned to each area well describe the points inside it, we use the Root Mean Squared Error (*RMSE*) and refer to it as labeling error:

$$RMSE(A, \ell^*(A)) = \sqrt{\sum_{C_k \in A} \frac{1}{|C|} \sum_{p_i \in C_k} \frac{1}{|C_k|} dist(\ell^*(p_i), \ell^*(C_k))^2}$$

In words, over all clusters (areas) C_k in the map, and over all points in each area, we measure the *semantic* distance between the label $\ell^*(p_i)$ assigned to the point and the label $\ell^*(C_k)$ assigned to the corresponding area. The higher the map's *RMSE*, the lower the accuracy of the labels (i.e., the higher the mismatch between the labels of the individual points and those of the corresponding areas).

We compare the labeling error values in the three cities of Barcelona, London, and Milan (Figure 5.5). By choosing the most popular label in each DBScan's cluster, the error is above 0.75 in the three cities. For our proposal, the error is always below 0.5. As λ increases, the error decreases (as high values for lambda enforce labeling homogeneity), and for $\lambda > 0.7$, the error is minimum.

By combining the two sets of results presented so far, we conclude that values of λ in the range $[0.35, 0.55]$ strike the right balance between area distinctiveness and labeling accuracy.

5.4 User Study

To complement our quantitative results and evaluate whether the found functional areas and their labels actually align with what residents perceive about their city, we conduct a mixed method user study.

5.4.1 Experimental setup and execution

We recruit 111 study participants in the three cities. Each participant is presented with a map of the city containing 5 (Milan) or 6 (Barcelona and London) functional areas that are highlighted and sequentially numbered (e.g., the bottom panel of Figure 5.6). The missing area for Milan corresponds to electronics - there is no specific area where to buy them in Milan, unlike Barcelona and London. The areas are identified by our framework when it is run with $\lambda = 0.4$.

We start our study by asking each of our participants to read a consent form and optionally provide age, gender, years living in the city, and email address. We recruit 40 (Barcelona), 40 (London), 31 (Milan) participants. Among them, the percentage of female-male is 75%-25% for London, 53%-47% for Barcelona, and 58%-42% for Milan. The most common age band is that of 30-35 (London) and 24-29 (Barcelona and Milan), and all our participants have lived in their city for more than four years.

We then provide brief instructions about the study and present six different tasks, one at a time. The tasks are chosen based on the area labels returned by our framework, and each task corresponds to one label. Of course, the area labels are unknown to the participants. A participant has to imagine the following six tasks and rate the extent to which each of the areas is suitable for each of the tasks. The ratings are expressed on a Likert scale (i.e., strongly disagree, disagree, neither agree nor disagree, agree and strongly agree).

Task-Office. A young entrepreneur is looking for a location where to base his new tech startup. You would recommend ...

Task-Electronics. A newly arrived student needs to buy electronics. You would recommend ...

Task-Hotels. A guy working in the hotel service industry has to visit as many hotels as possible in a short time. You would recommend ...

Task-Clothing. A friend of yours wants to visit as many clothing stores as possible. You would take her to ...

Task-University. A newly arrived foreign student wishes to experience university life. You would take her to ...

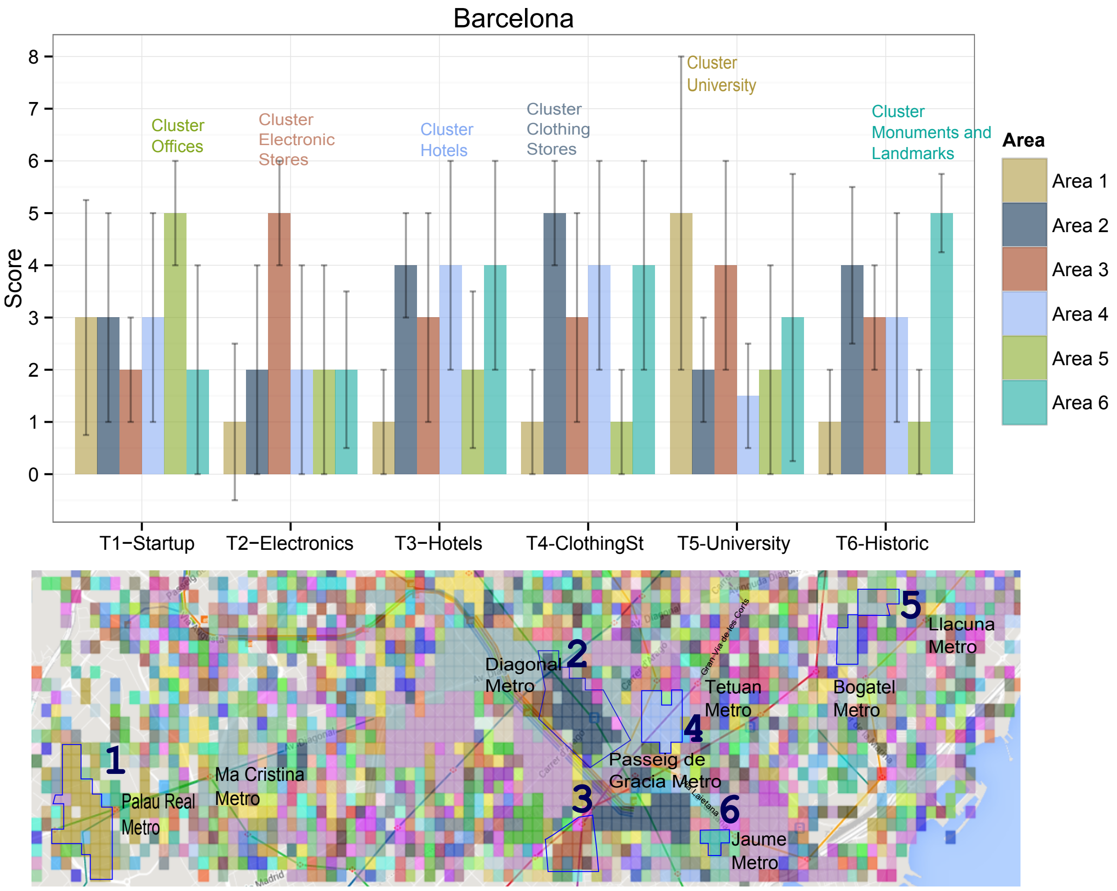


Figure 5.6: *Barcelona areas our framework identified to be best for each hypothetical task (bottom panel), and the extent to which our respondents suggested the very same areas (top panel).*

Task-Monuments. A friend of yours is visiting the city for the first time and wishes to see historic places and monuments. You would take her to . . .

After providing answers for each task, the participant is asked to motivate his/her answers in free-text form and eventually indicate the name of the area that (s)he would have recommended for that task (it could but does not have to be one of the six areas).

5.4.2 Quantitative and Qualitative results

In Barcelona, the functions proposed by our framework (unknown to our respondents) match those suggested by our respondents for the office, electronics, clothing, and historic tasks (top panel of Figure 5.6): the area labeled with a given category (e.g., elec-

Question	Area1	Area2	Area3	Area4	Area5	Area6
T1-Office	3.00	3.00	2.00	3.00	**5.00**	2.00
T2-Electronics	1.00	2.00	**5.00**	2.00	2.00	2.00
T3-Hotels	1.00	4.00	3.00	**4.00**	2.00	4.00
T4-ClothingSt	1.00	**5.00**	3.00	4.00	1.00	4.00
T5-UniversityLife	**5.00**	2.00	4.00	1.50	2.00	3.00
T6-HistoricMonuments	1.00	4.00	3.00	3.00	1.00	**5.00**

Table 5.1: *Extent to which each area in Barcelona is best in each task. Median values in a Likert scale are reported. Highlighted values correspond to the areas labeled to be best for each task.*

tronics) is associated with the corresponding category's task (e.g., 'task-electronics') with a 'strongly agree' assessment (i.e., with a median score of 5). These results are also reflected in our respondents' comments. For example, the majority of our respondents correctly consider area 5 to be best for the 'startup-task' and identify it to be the "22@ area", which is known, as one respondent puts it, "as the innovation & entrepreneurship area in Barcelona". By contrast, for the two remaining tasks (i.e., hotels and university), the variability of the answers is high. For the 'hotels-task', our framework has identified area 4 to be best. Some respondents agree with that, while others add that areas 2 and 6, being located in the central part of town, have many hotels too. For the 'university-task', our respondents indicate area 1 (identified by our framework) to indeed host university buildings yet prefer to suggest area 3 for 'university life': this is where most students live.

As for London, the functions proposed by our framework match those suggested by our respondents for the office, electronics, clothing, university and historic tasks (top panel of Figure 5.7): for those tasks, the median scores are all 5. Once again, for the 'hotels-task', our respondents identified multiple areas (mainly central ones) to be best (not only the one identified by our framework).

Finally, in Milan, the functions proposed by our framework match those suggested by our respondents for the clothing and university tasks (top panel of Figure 5.7), and, to a certain extent, for the hospital task (the answer variability is lowest for the area

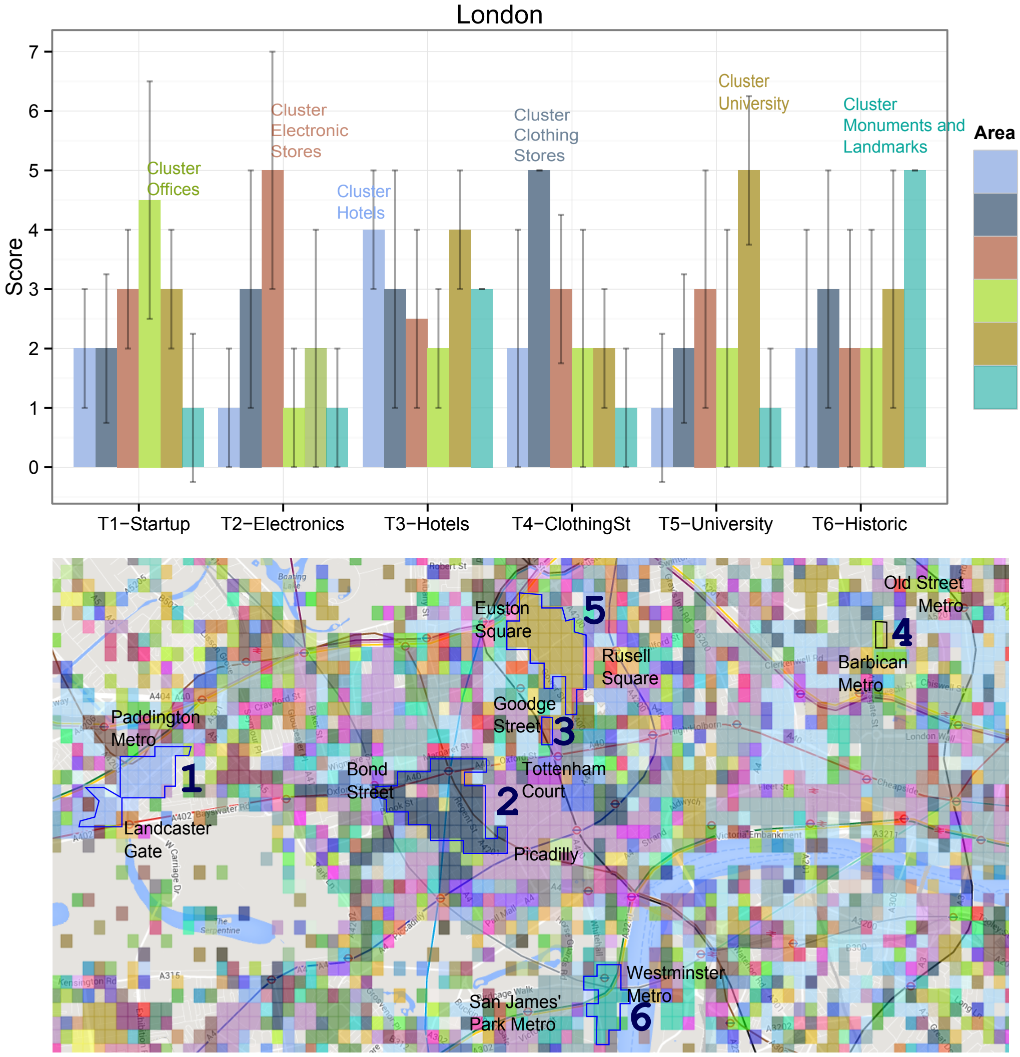


Figure 5.7: *London areas our framework identified to be best for each hypothetical task (bottom panel), and the extent to which our respondents suggested the very same areas (top panel).*

identified by our framework). Again, for the ‘hotels-task’, a central area in addition to the one proposed by our framework is often mentioned.

For the remaining ‘startup-task’, our respondents offer a variety of answers: some suggest the area labeled by our framework, while others suggest the area near the main technical university (Politecnico). They motivate this latter answer by saying that, if placed near Politecnico, the startup could benefit from technology transfer and could avoid the problems of more central areas of the city, which “are not specialized in technology and are simply crowded and expensive”.

Taken together, the user study’s quantitative and qualitative results both suggest that our framework is able to identify areas in the three cities and effectively label them with their functions.

5.5 Discussion

We now dwell on some of our framework’s desirable properties and discuss some open questions.

5.5.1 Flexible framework

Previous approaches for discovering functional areas have modeled contextual factors such as time. For example, Yuan *et al.* not only derived Beijing’s functional areas but also their evolution from 2010 to 2011 [126]. Next, we briefly show that considering a temporal taxonomy allows for studying how the city fabric changes during different seasons.

To this end, we gather a random sample of more than 1M geo-referenced pictures within the bounding box of Barcelona from the Flickr public API, 400K of which are generated by distinct users in distinct locations. For each picture, we gather its unique identifier, latitude, longitude, the owner’s identifier, the date of creation and the date of upload. We then consider the temporal taxonomy in Figure 5.9(a) and accordingly label each picture with one of the taxonomy categories depending on the picture’s date

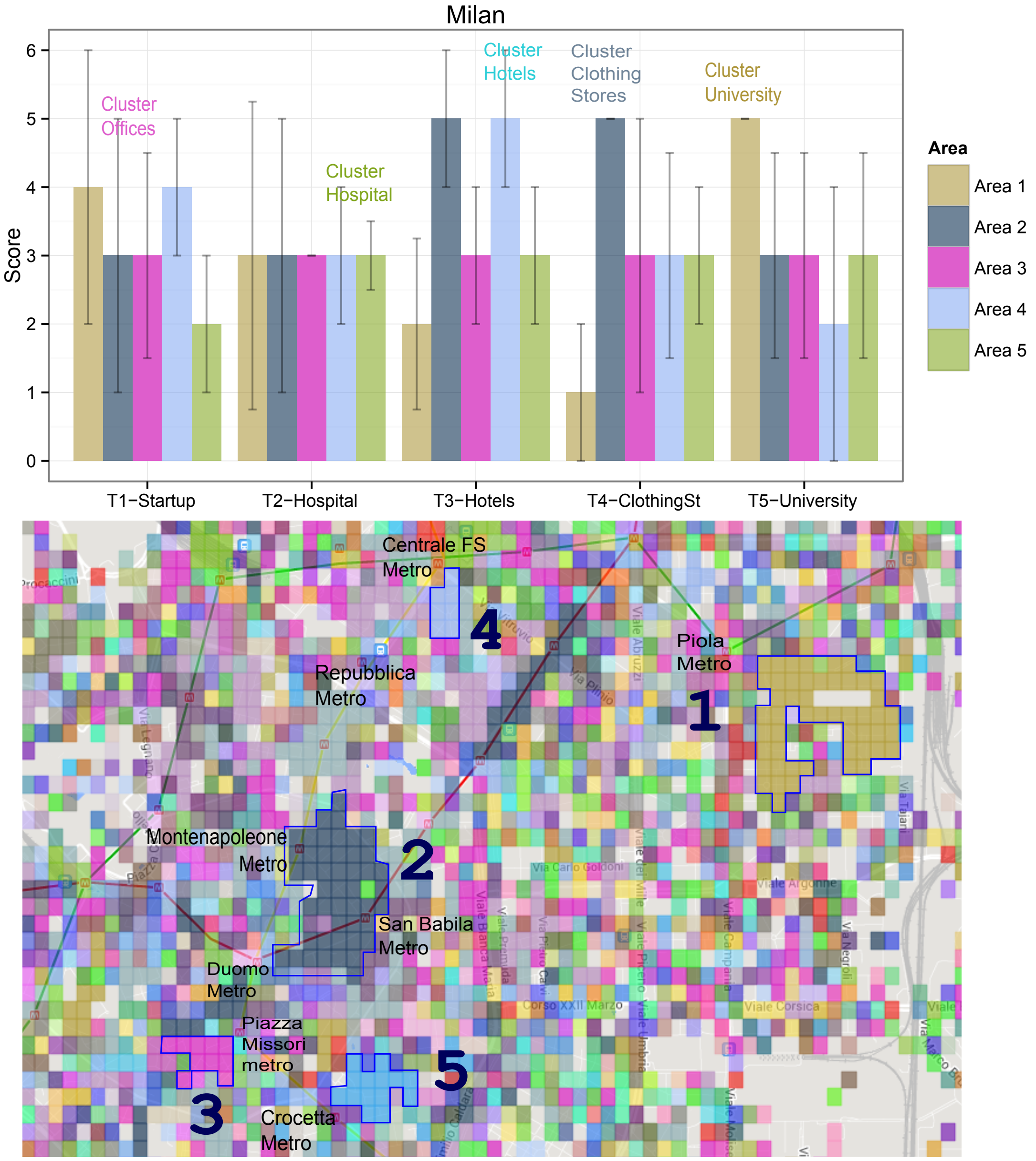
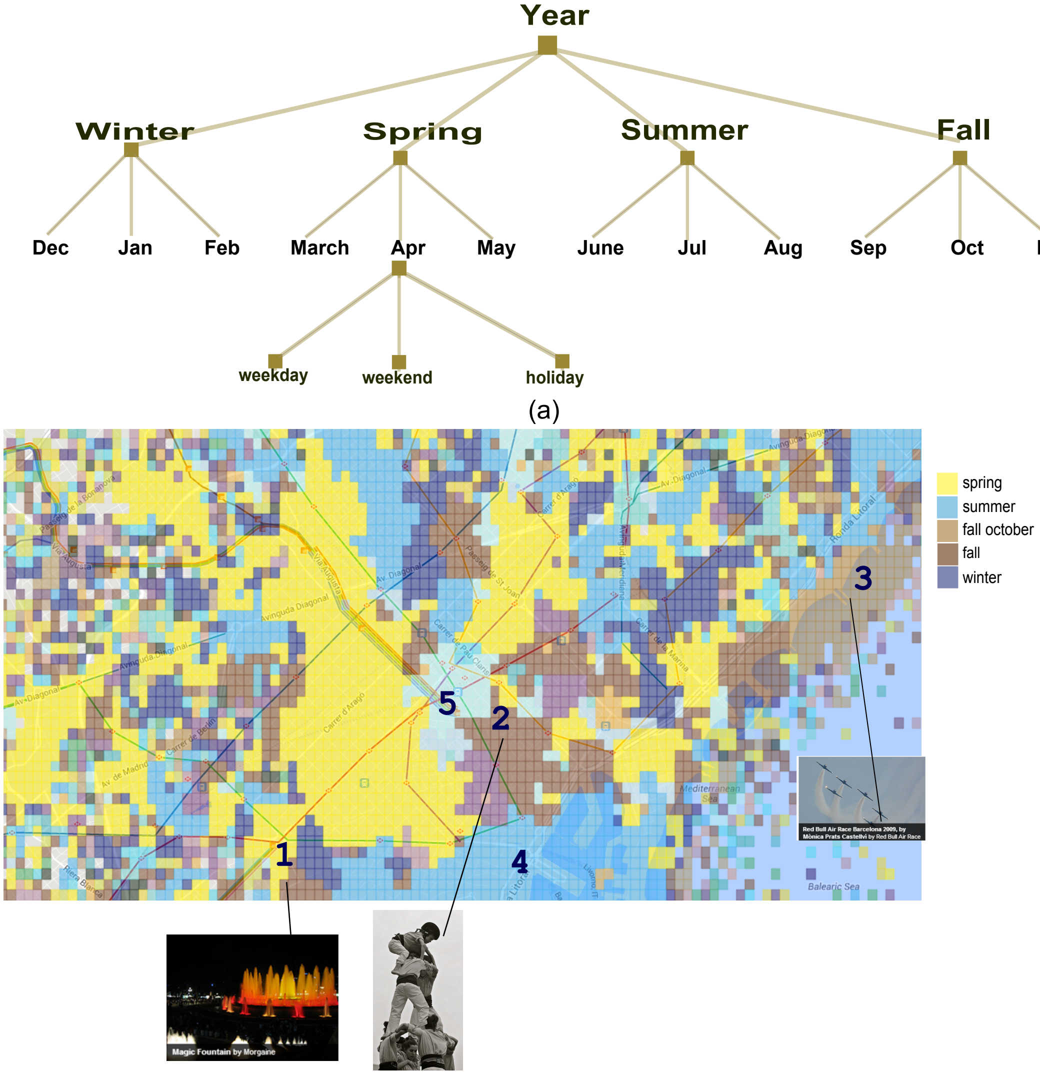


Figure 5.8: *Milan areas our framework identified to be best for each hypothetical task (bottom panel), and the extent to which our respondents suggested the very same areas (top panel).*



(a)

(b)

Figure 5.9: *Temporal analysis of Flickr pictures in Barcelona. The areas reported in (b) are determined by our framework using the temporal taxonomy in (a) with* $\lambda = 0.4$.

of creation. We then run our framework with, again, $\lambda = 0.4$ and obtain the areas in Figure 5.9(b), which we number from 1 to 5:

Area 1. The framework categorizes it with the ‘fall’ label. The area mainly covers a road called ‘Avinguda de la Reina Maria Cristina’, which is close to Plaza d’Espana. By looking at the actual pictures, we find that most of them depict the Magic Fountain (pictured at the bottom of Figure 5.9) during ‘La Merce’ festival. This celebrates the ‘La Virgen de La Merce’, patron of Barcelona, towards the end of September, and it is one of the most important festivities for the residents of Barcelona.

Area 2. This area is marked with ‘fall’ and is related with the Merce festival too. It mainly includes the road of ‘Via Laietana’, in which, during the festival, the world-renowned ‘castellets’ (towers of individuals on top of each other) are built. They are world-renowned since they receive coverage from major international news outlets.

Area 3. This area is marked with the very specific label of ‘fall-october’. This label comes as a surprise as the area includes beaches popular during the summer. By looking at the pictures, we find out that they are about the ‘Festa al cel’, which we now learn is the greatest air showcase in the country. The label fall-october is not the most frequent among the pictures in that area but understandably is the most discriminative compared to nearby areas, which happen to be beaches.

Areas 4 and 5. They contain the most famous beaches and central squares (e.g., Placa de Catalunya) popular among summer tourists. As one expects, the areas are categorized as ‘summer’.

Those results suggest that, with a temporal taxonomy, one is able to identify events key in specific seasons. More importantly, those results speak to our framework’s flexibility: different definitions of taxonomy result into very distinct notions of functional areas. Take a demographic taxonomy, which segments users into different socio-demographic classes according to age, gender, and profession. With it, our framework could potentially discover areas that serve similar or distinct functions for a variety of lifestyles [127]. Another example is a weather taxonomy. By using such a taxonomy, our framework could discover areas that are visited preferentially when, say, it is sunny or raining.

In practice, our framework's flexibility enables key applications. For example, a number of mobile personalized services are trying to figure out how to make geofencing a reality. The term geofencing refers to the use of geofences in combination with mobile services. The idea is that notifications are sent to mobile users whenever they cross a geofence (a geographic boundary). Many common geofencing scenarios are based on a simple radius around a point of interest, like a shop. One of the most important challenges is that geospatial calculations are complex and, since they require the use of GPS chips, they tend to drain the battery in a few hours. Our proposal partly fixes that as it offers a scalable solution: one could imagine to download the shape files of different functional areas on the phone and cheaply support background mobile applications that send tourist information or personalized shopping offers. To see how, imagine a user saving books and electronics on her electronic wishlist while at work. When traveling back home, her phone could generate alerts with the list of functional areas in which she could stop by and acquire some of the items on the list.

5.5.2 Usable framework

The proposed framework aims at being not only flexible but also usable. We believe it is so for two main reasons: i) it requires to fine tune a single parameter; and ii) it supports current visualization paradigms.

Single parameter. To discover functional areas, only one parameter has to be set. This

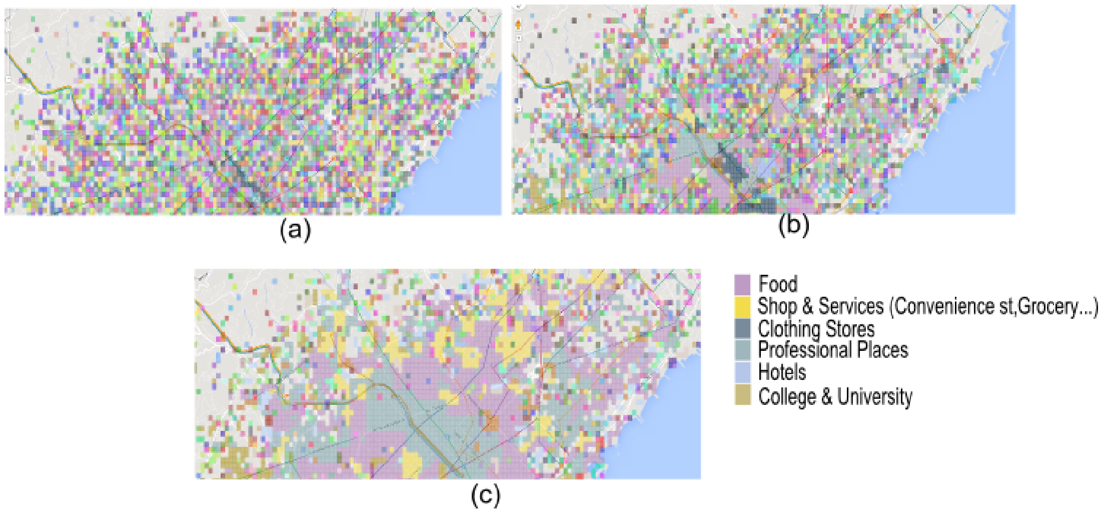


Figure 5.10: *Areas in Barcelona identified by our framework from Foursquare venues with: (a)* $\lambda = 0.1$*; (b)* $\lambda = 0.4$*; and (c)* $\lambda = 0.9$*.*

parameter is intuitive and goes from 0 to 1: 0 corresponds to the finest-grained division of functional areas, while 1 corresponds to the most homogeneous division. The value of this parameter depends on what the algorithm's user is after. The difference between retail analysts' needs and tourists' is a case in point. Retail analysts might wish to emphasize areas that do not have a single function but have compound functions, so they would set a low value for the parameter (Figure 5.10(a) reports the Barcelona map for $\lambda = 0.1$). By contrast, tourists might be after a quick and digestible snapshot of the city fabric and, a such, would set a high value for it (Figure 5.10(c) reports the Barcelona map for $\lambda = 0.9$).

Supporting current visualization paradigms. The labels with which we annotate areas are not only human interpretable but also part of a taxonomy. That makes it possible to show such annotations simply as a list of categories next to a map, which is what people nowadays are used to and, as such, avoids pushing them outside their comfort zone. Having the map of a city and, next to it, the list of categories in our taxonomy, a user could click on a category and see the areas annotated with that category highlighted on the map. This was not readily possible with the most popular approaches of discovering functional areas (i.e., topic-based inference models, segmentation techniques). That is because those approaches express area annotations as category distributions, which are not easily translatable in a drop-down menu or, for that matter, in any visualization paradigm individuals are used to.

5.6 Conclusion

We have proposed a taxonomy-driven framework for discovering functional areas and have extensively tested it in three cities. By changing the type of taxonomy under study, we have shown that our framework offers flexibility on the types of areas that could be potentially discovered: for example, we have shown that it discovers not only functional areas but also seasonal ones. Based on those positive results, our framework promises to partly overcome a wide range of challenges, including: recent industry efforts in the area of mobile and personalization like geo-fencing; spatio-temporal studies of how the city is effectively used by social media users; and urban socio-cultural investigations such as that of how, given a language taxonomy, ethnic groups geographically sort themselves. To collectively meet those challenges, in the near future, we will make the framework's code publicly available and make it suitable for consumption of researchers in different fields such as marketing, computer science, and urban sociology.

The cross comes before the crown and tomorrow is a Monday morning.

C.S. Lewis, 1942

6
Conclusions and reflections

Citizens in both, rich and poor cities, are generating enormous amount of geolocated data and a great deal of social relationships in the offline world are reflected by exchanges either on social networks platforms or mobile phones activity. At the same time, local governments and urban planners suffer from the lack of systematic ways of collecting and analyzing data at city level to inform their decisions and fight inequity when allocating resources to different areas of the city.

In this dissertation, we have presented three framework that can be used to better understand online interactions and to aggregate geolocated content to extract meaningful characteristics of metropolis where the data is generated. Along this path, we have designed a framework to model the attention received by users in a content sharing platform across time and discovered useful insights about the activity patterns that successful users perform to maintain their social ties (Chapter 3). We then exploit the location of online interactions to show that effective ways of aggregating the data at city level, together with solid urban theories, can help us to effectively predict the wealth of a city (Chapter 4). Specifically we have proposed a set of content independent features that can be measured in any content sharing platform and we have shown that those metrics act as effective proxies for the city economic capital. Interestingly, we have observed that quantifying the attention received by locals in combination with the attention received from users outside the city, can accurately describe urban socioeconomic indicators.

Next, we characterize urban land use from online data. To this end, a clustering framework that detects and automatically label functional urban areas (Chapter 5) is designed and implemented. . The clustering framework was evaluated on three different cities (Barcelona, London and Milan) both quantitatively and qualitatively showing that it effectively discover characteristic areas around the city. An important contribution of the clustering framework is the approach driven by a taxonomy of categories. Specifically, the taxonomy gives to the framework the ability of finding clusters and label them using more abstract or more specific labels as specified by a single and intuitive user parameter. Moreover, different taxonomies can be applied to the same dataset of geolocated points obtaining different views of the city that might reflect, for instance, temporal patterns, geographic distribution of visitors among others.

The results strongly suggest that online interactions and geolocated content, in general, have a large potential to let us describe different urban phenomena. We propose directions for further work next.

6.1 Future work

The work presented can be extended in many ways towards a comprehensive platform to early observe trends on urban indicators as well as disruptive changes on the way people use different areas of the city. Building digital integrated tools that translate the activity in online platforms into numerical urban indicators constitute a paradigm shift for local governments specially in developing countries where the amount of official digital data is very low. We analyze future directions of our work next.

6.1.1 Urban indicators

Social network analysis techniques applied to information spread processes on online social networks have been shown effective to build metrics that quantify the city wealth (Chapter 4). The proposed framework can be extended with content dependent features designed with the same methodology, i.e., analyze urban theories that refers to city issues such as transport, unemployment, slums and from those theories design a set of keywords to track social media platforms in order to spot problems in each of these areas. A set of common diffusion patterns for these keywords can be derived and compared across time to early detect important changes letting people to be real city sensors. As a result a integrated visual interface showing the pulse of the city with respect to different indicators can be built.

Furthermore, the online activity can be exploited not only to observe critical issues but also to evaluate the impact of different types of events in the city. How does the number of people talking about a future event on a city refers to the amount of pollution, number of tourist or traffic congestion? Does the city global or local attention increased after a official-promoted event took place in the city? How many non local people is talking about the event? Answering these questions can help local authorities not only to plan ahead but also to evaluate new policies or small projects designed to promote the city in terms of, for example, increases in the number of national visitors talking about public spectacles.

As we have said before, the urban analysis can be conducted at different scales. An important direction for the work on urban indicators consists on applying the *glocality* concept to the level of neighborhoods where the absence of available data is even more certain. A small number of cities around the world have indicators available at the level of neighborhoods and this information can be exploited to validate the approach that later can be applied to a larger number of cities. In fact, we have already conducted some experiments in this direction and we have found that emigrants communities can be easily spotted from digital data applying such a method.

6.1.2 Urban dynamic areas

A common trend on cities is the urban sprawl: demand for land on the city center decreases and at the same time, demand for land on peripheral areas increases [76]. Tracking the dynamics behind such people movements from one area to another for living or opening new offices constitutes an important application of our work. We plan to develop methods to track, across time, the changes in the way people are using different areas in the city. A set of techniques, that include topic detection to choose starting categories and timeseries analysis, can be integrated here to detect when the 'function' of urban areas present an important departure from what it has been common for a certain period of time. In this way, we can help to redefine the official boundaries of cities based on the data produced by their residents.

Currently the unsupervised learning framework proposed on Chapter 5 considers only adjacent cells as neighbors. We can extend such proposal and analyze the impact of including weights that decline with distance as it has been done by geographers when detecting different types of urban hotspots [38]. Moreover, an important extension of the framework should allow us to label urban areas in a mixed land use scenario. Urban planners acknowledge that mixed land use can foster urban economies by promoting greater density, reducing cars dependency and traffic congestion, promoting walking and bicycle use [27]. Therefore, the clustering framework presented in this work needs to be extended to consider areas represented with multiple labels.

Places are both physical locations and a product of the interactions among people who use them [115]. Therefore, a future development in the work related to the detection of urban areas, i.e., interconnected places on space, should take into account the social nature of physical locations. Social media can reveal a great deal about people preferences and we will look into those to add this information to our urban areas detection framework. As a result, we foresee a platform in which urban planners as well as final users can navigate a city from different perspectives: how locals and foreigners live the city, how the ethnic food places are dispersed around the city, which are the most important events that disrupt the way in which people use different urban areas are just three examples of what we can observe through online geolocalized digital data together with the appropriated methods to mine and interpret it.

Bibliography

[1] J Abel and T Gabe. Human capital and economic activity in urban america. *Regional Studies*, 2011.

[2] S. Arthur and S. M. Sheffrin. *Economics: Principles in action*. 2003.

[3] J. Assfalg, H.P. Kriegel, P. Kroger, P. Kunath, A. Pryakhin, and M. Renz. Similarity search on time series based on threshold queries. *Advances in Database Technology-EDBT*, pages 276–294, 2006.

[4] Sitaram Asur and Bernardo A Huberman. Predicting the future with social media. In *Proc. of IEEE/WIC/ACM Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 2010.

[5] Sitaram Asur, Bernardo A Huberman, Gabor Szabo, and Chunyan Wang. Trends in social media: Persistence and decay. In *Proceedings of the 5th AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[6] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web Search and Data Mining (WSDM)*, pages 65–74. ACM, 2011.

[7] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web (WWW)*, pages 519–528. ACM, 2012.

[8] M. Basseville and I.V. Nikiforov. *Detection of abrupt changes: theory and applications*. 1993.

[9] Anil Bawa-Cavia. Sensing The Urban: Using location-based social network data in urban analysis. In *Pervasive Urban Applications (PURBA)*, 2011.

[10] L. Bettencourt, J. Lobo, D. Helbing, C. Kühnert, and G. West. Growth, innovation, scaling, and the pace of life in cities. *Proc. of the National Academy of Sciences*, 2007.

[11] D. Bhattacharya and S. Ram. Sharing News Articles Using 140 Characters: A Diffusion Analysis on Twitter. In *Proc. of the IEEE/ACM Conference in Advances in Social Networks Analysis and Mining (ASONAM)*, 2012.

[12] Orna Blumen. The spatial distribution of occupational prestige in metropolitan tel aviv. *Area*, 30(4):343–357, 1998.

[13] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[14] Phillip Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2(1):113–120, 1972.

[15] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(11):1222–1239, 2001.

[16] A. Brodersen, S. Scellato, and M. Wattenhofer. Youtube around the world: geographic popularity of videos. In *Proceedings of the 21st conference on World Wide Web (WWW)*, pages 241–250. ACM, 2012.

[17] Stanley D Brunn, Jack Francis Williams, and Donald J Zeigler. *Cities of the world: world regional urban development*. Rowman & Littlefield, 2003.

[18] Ronald S. Burt. Structural Holes and Good Ideas. *The American Journal of Sociology*, 2004.

[19] Zechun Cao, Sujing Wang, Germain Forestier, Anne Puissant, and Christoph F. Eick. Analyzing the Composition of Cities Using Spatial Clustering. In *Proceedings of the 2^{nd} ACM SIGKDD International Workshop on Urban Computing (UrbComp)*, 2013.

[20] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. *AAAI Conference on Weblogs and Social Media(ICWSM)*, 10:10–17, 2010.

[21] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World Wide Web (WWW)*, pages 721–730. ACM, 2009.

[22] Chandra Chekuri, Sanjeev Khanna, Joseph Naor, and Leonid Zosin. A linear programming formulation and approximation algorithms for the metric labeling problem. *SIAM Journal on Discrete Mathematics*, 18(3):608–625, 2004.

[23] J. Chen and A.K. Gupta. *Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance*. 2011.

[24] Julia Chuzhoy and Joseph Naor. The hardness of metric labeling. *SIAM Journal on Computing*, 36(5):1376–1386, 2007.

[25] Partnership C.L. and TfL. Legible London wayfinding study. In *TfL*, 2006.

[26] Justin Cranshaw, Raz Schwartz, Jason I. Hong, and Norman M. Sadeh. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. In *Proceedings of the 6th AAAI Conference on Weblogs and Social Media (ICWSM)*, 2012.

[27] Barry J Cullingworth, J Barry Cullingworth, and Roger Caves. *Planning in the USA: policies, issues, and processes*. Routledge, 2013.

[28] Richard Dobbs, Jaana Remes, and Sven Smit. The world's new growth frontier: Midsize cities in emerging markets. *McKinsey Quarterly, http://www.mckinsey.com*, 2011.

[29] Duan, Lian and Xu, Lida and Guo, Feng and Lee, Jun and Yan, Baopin. A Local-density Based Spatial Clustering Algorithm with Noise. *Information Systems*, 32(7), November 2007.

[30] John Eade. *Living the global city: Globalization as local process*. Routledge, 2003.

[31] N. Eagle, M. Macy, and R. Claxton. Network diversity and economic development. *Science*, 2010.

[32] N. B Ellison, Charles S., and C. Lampe. The benefits of Facebook "friends:" Social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication*, 2007.

[33] F. Figueiredo, F. Benevenuto, and J. M. Almeida. The tube over time: characterizing popularity growth of youtube videos. In *Proceedings of the fourth ACM international conference on Web search and data mining (WSDM)*, pages 745–754. ACM, 2011.

[34] R. Forrest and A. Kearns. Social Cohesion, Social Capital and the Neighbourhood. *Urban Studies*, 2001.

[35] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[36] Tak-Chung Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, 2011.

[37] H. Ganzeboom and D. Treiman. Internationally comparable measures of occupational status for the 1988 international standard classification of occupations. *Social science research*, 1996.

[38] Arthur Getis and Jared Aldstadt. Constructing the spatial weights matrix using a local statistic. In *Perspectives on Spatial Data Analysis*, pages 147–163. Springer, 2010.

[39] P. Geurts. Pattern extraction for time series classification. *Principles of Data Mining and Knowledge Discovery*, 2001.

[40] Rumi Ghosh and Kristina Lerman. Predicting influential users in online social networks. In *Proc. of the 4th AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.

[41] Edward Glaeser. *Triumph of the city: How our greatest invention makes US richer, smarter, greener, healthier and happier*. Macmillan, 2011.

[42] Edward L. Glaeser and Janet E. Kohlhase. Cities, regions and the decline of transport costs. *Regional Science*, 2004.

[43] Mark Granovetter. The strength of weak ties: A network theory revisited. *Sociological theory*, 1983.

[44] Daniel Gruhl, Ramanathan Guha, Ravi Kumar, Jasmine Novak, and Andrew Tomkins. The predictive power of online chatter. In *Proc. of the eleventh ACM conference on Knowledge Discovery in Data Mining (KDD)*, 2005.

[45] Anupam Gupta and Éva Tardos. A constant factor approximation algorithm for a class of classification problems. In *Proceedings of the 32nd annual ACM symposium on Theory of computing*, pages 652–658, 2000.

[46] UN Habitat. Cities in a globalizing world: global report on human settlements 2001. *United Nations-Habitat, London*, 2001.

[47] W. Härdle and P. Vieu. Kernel regression smoothing of time series. *Journal of Time Series Analysis*, 13(3):209–232, 1992.

[48] Kevin Harris and Hugh Flouch. Online neighbourhood networks study: Social capital and cohesion. Technical report, The Networked Neighbourhoods Group, 2010.

[49] J Vernon Henderson, Adam Storeygard, and David N Weil. Measuring economic growth from outer space. Technical report, National Bureau of Economic Research, 2009.

[50] Martin Herold, Helen Couclelis, and Keith C Clarke. The role of spatial metrics in the analysis and modeling of urban land use change. *Computers, Environment and Urban Systems*, 29(4):369–399, 2005.

[51] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. In *Proceedings of the 20th conference companion on World Wide Web (WWW)*, pages 57–58. ACM, 2011.

[52] Bernardo A Huberman, Daniel M Romero, and Fang Wu. Crowdsourcing, attention and productivity. *Journal of Information Science*, 35(6):758–765, 2009.

[53] D. Ienco, F. Bonchi, and C. Castillo. The meme ranking problem Maximizing microblogging virality. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pages 328–335. IEEE, 2010.

[54] J. Jiang and D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of the Conference on Research in Computational Linguistics*, 1997.

[55] D Jin, DC Michael, P Foo, J Guevara, I Peña, A Tratz, and S Verma. Winning in emerging-market cities. A guide to the world's largest growth opportunity. *Boston, MA: The Boston Consulting Group,* $http://www.bcg.com$, 2011.

[56] Dmytro Karamshuk, Anastasios Noulas, Salvatore Scellato, Vincenzo Nicosia, and Cecilia Mascolo. Geo-spotting: Mining Online Location-based Services for Optimal Retail Store Placement. In *Proceedings of the* 19^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2013.

[57] Jon M. Kleinberg and Éva Tardos. Approximation algorithms for classification problems with pairwise relationships: metric labeling and Markov random fields. *Journal of the ACM (JACM)*, 49(5):616–639, 2002.

[58] F. Kooti, H. Yang, M. Cha, K. P. Gummadi, and W. A. Mason. The emergence of conventions in online social networks. In *AAAI Conference on Weblogs and Social Media(ICWSM)*. AAAI, 2012.

[59] Gueorgi Kossinets and Duncan J Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, 2006.

[60] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In *Link mining: models, algorithms, and applications*, pages 337–357. Springer, 2010.

[61] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World Wide Web (WWW)*, pages 591–600. ACM, 2010.

[62] Charles Landry. *The creative city: A toolkit for urban innovators*. Earthscan, 2008.

[63] C.H. L. Lee, A. Liu, and W.S. Chen. Pattern discovery of fuzzy time series for financial prediction. *Knowledge and Data Engineering, IEEE Transactions on*, 18(5):613–625, 2006.

[64] Chanam Lee and Anne Vernez Moudon. The 3ds+ r: Quantifying land use and urban form correlates of walking. *Transportation Research Part D: Transport and Environment*, 11(3):204–215, 2006.

[65] J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on World Wide Web (WWW)*, pages 251–260. ACM, 2012.

[66] J. Lin, E. Keogh, L. Wei, and S. Lonardi. Experiencing sax: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, 15(2):107–144, 2007.

[67] J. Lin and Y. Li. Finding structural similarity in time series data using bag-of-patterns representation. In *Scientific and Statistical Database Management*, pages 461–477. Springer, 2009.

[68] Nan Lin. *Social capital: A theory of social structure and action*. Cambridge University Press, 2002.

[69] Xuelian Long, Lei Jin, and James Joshi. Exploring Trajectory-driven Local Geographic Topics in Foursquare. In *Proceedings of the ACM Conference on Ubiquitous Computing (UbicComp)*, 2012.

[70] W Glynn Mangold and David J Faulds. Social media: The new hybrid element of the promotion mix. *Business horizons*, 52(4), 2009.

[71] Rajsekar Manokaran, Joseph Naor, Prasad Raghavendra, and Roy Schwartz. Sdp gaps and ugc hardness for multiway cut, 0-extension, and metric labeling. In *Proceedings of the 40th annual ACM symposium on Theory of Computing*, pages 11–20, 2008.

[72] Huina Mao, Xin Shuai, Yong-Yeol Ahn, and Johan Bollen. Mobile communications reveal the regional economy in côte divoire. In *Proceedings of the 3rd Conference on the Analysis of Mobile Phone Datasets (NetMob)*, 2013.

[73] M. Mathioudakis, N. Koudas, and P. Marbach. Early online identification of attention gathering items in social media. In *Proceedings of the third ACM international conference on Web Search and Data Mining (WSDM)*, pages 301–310. ACM, 2010.

[74] Patricia L Mccarney and Klaus Segbers. *Cities and Global Governance: New Sites for International Relations*. Ashgate Publishing, Ltd., 2013.

[75] Yelena Mejova, Padmini Srinivasan, and Bob Boynton. GOP primary season on Twitter: popular political sentiment in social media. In *Proceedings of the sixth ACM conference on Web search and data mining (WSDM)*, 2013.

[76] Vittorio Gargiulo Morelli and Luca Salvati. *Ad hoc urban sprawl in the Mediterranean city: Dispersing a compact tradition?* Edizioni Nuova Cultura, 2010.

[77] Eduardo López Moreno. *Slums of the World: The Face of Urban Poverty in the New Millennium?* UN-HABITAT, 2003.

[78] Mor Naaman, Hila Becker, and Luis Gravano. Hip and trendy: Characterizing emerging trends on Twitter. *Journal of the American Society for Information Science and Technology*, 2011.

[79] Zachary P Neal. *The connected city: How networks are shaping the modern metropolis*. Routledge, 2012.

[80] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks. In *Proceedings of 3rd Workshop Social Mobile Web (SMW11). Colocated with ICWSM 2011.*, 2011.

[81] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proc. of the 4th AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.

[82] S. O'Sullivan and J. Morrall. *Walking Distances to and from Light Rail Transit Stations*. Transportation Research Board, 1996.

[83] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

[84] Aditya Pal and Scott Counts. Identifying topical authorities in microblogs. In *Proceedings of the fourth ACM international conference on Web Search and Data Mining (WSDM)*, pages 45–54. ACM, 2011.

[85] G. Palshikar. Simple algorithms for peak detection in time-series. In *Proc. 1st Int. Conf. Advanced Data Analysis, Business Analytics and Intelligence (ADABAI)*, 2009.

[86] Senin Pavel. j-Motif, SAX implementation in Java. Last accessed on October 2014.

[87] R.D. Putnam. *Bowling Alone: The Collapse and Revival of American Community*. Simon & Schuster, 2001.

[88] D. Quercia, N. Lathia, F. Calabrese, G. Di Lorenzo, and J. Crowcroft. Recommending Social Events from Mobile Phone Location Data. In *Proceedings of the* 10^{th} *IEEE International Conference on Data Mining (ICDM)*, 2010.

[89] Daniele Quercia, Jonathan Ellis, Licia Capra, and Jon Crowcroft. In the mood for being influential on twitter. In *Privacy, security, risk and trust (passat), 2011 IEE third international conference on and 2011 IEEE third international conference on Social Computing (SocialCom)*, pages 307–314. IEEE, 2011.

[90] Daniele Quercia, Jonathan Ellis, Licia Capra, and Jon Crowcroft. Tracking gross community happiness from tweets. In *Proc. of the ACM Conference on Computer Supported Cooperative Work (CSCW)*, 2012.

[91] J. Ratkiewicz, S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani. Characterizing and modeling the dynamics of online popularity. *Physical review letters*, 105(15), 2010.

[92] Daniel M. Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A. Huberman. Influence and passivity in social media. In *WWW'11: Proceedings of the 20th International Conference Companion on World Wide Web*, pages 113–114, New York, NY, USA, 2011. ACM.

[93] Daniel M Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A Huberman. Influence and passivity in social media. In *Machine learning and Knowledge Discovery in Databases*, pages 18–33. Springer, 2011.

[94] Eduardo J Ruiz, Vagelis Hristidis, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes. Correlating financial time series with micro-blogging activity. In *Proceedings of the fifth ACM international conference on Web search and data mining (WSDM)*, 2012.

[95] Saskia Sassen. The global city: Introducing a concept. *Brown Journal of World Affairs*, 2004.

[96] Salvatore Scellato, Cecilia Mascolo, Mirco Musolesi, and Vito Latora. Distance matters: geo-social metrics for online social networks. In *Proc. of the 3rd conference on Online social networks (WOSN)*, 2010.

[97] Salvatore Scellato, Anastasios Noulas, Renaud Lambiotte, and Cecilia Mascolo. Socio-spatial properties of online location-based social networks. In *Proc. of the 5th AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[98] R. Schifanella, A. Barrat, C. Cattuto, B. Markines, and F. Menczer. Folks in folksonomies social link prediction from shared metadata. In *Proceedings of the third ACM international conference on Web Search and Data Mining*, pages 271–280. ACM, 2010.

[99] Eric Schmidt and Jared Cohen. *The New Digital Age: Transforming Nations, Businesses, and Our Lives*. Random House LLC, 2013.

[100] Cosma Rohilla Shalizi and Andrew C Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, 40(2):211–239, 2011.

[101] John Rennie Short, Carrie Breitbach, Steven Buckman, and Jamey Essex. From world cities to gateway cities: extending the boundaries of globalization theory. *City*, 2000.

[102] Bogdan State, Ingmar Weber, Emilio Zagheni, et al. Studying inter-national mobility through IP geolocation. In *Proc. of the sixth ACM Conference on Web Search and Data Mining (WSDM)*, 2013.

[103] C. Steinfield, N. Ellison, and C. Lampe. Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. *Journal of Applied Developmental Psychology*, 2008.

[104] S. A. Stouffer. Intervening opportunities: a theory relating mobility and distance. *American sociological review*, 1940.

[105] Thorsten Strufe. Profile popularity in a business-oriented online social network. In *Proceedings of the 3rd Workshop on Social Network Systems(SNS)*, page 2. ACM, 2010.

[106] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 177–184. IEEE, 2010.

[107] Gabor Szabo and Bernardo A Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.

[108] Peter J Taylor, P Ni, B Derudder, M Hoyler, J Huang, F Lu, K Pain, F Witlox, X Yang, D Bassens, et al. Measuring the world city network: New developments and results. *GaWC Research Bulletin*, 300, 2009.

[109] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *Proc. of the 4th AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.

[110] UN. Big Data for Development: A Primer. *United Nations, Global Pulse*, 2013.

[111] Carmen Vaca Ruiz, Luca M Aiello, and Alejandro Jaimes. Modeling dynamics of attention in social media with user efficiency. *EPJ Data Science*, 3(1):5, 2014.

[112] Carmen Vaca-Ruiz, Daniele Quercia, Luca Maria Aiello, and Piero Fraternali. Tracking human migration from online attention. In *Citizen in Sensor Networks*, pages 73–83. Springer, 2014.

[113] Olga Veksler. *Efficient graph-based energy minimization methods in computer vision*. PhD thesis, Cornell University, 1999.

[114] Alessandro Vespignani. Modelling dynamical processes in complex socio-technical systems. *Nature Physics*, 8(1):32–39, 2012.

[115] Michael A Von Hausen. *Dynamic Urban Design*. iUniverse, 2012.

[116] Xiaozhe Wang, Kate Smith, and Rob Hyndman. Characteristic-based clustering for time series data. *Data Mining and Knowledge Discovery*, 13(3):335–364, 2006.

[117] T Warren Liao. Clustering of time series data - a survey. *Pattern Recognition*, 38(11):1857–1874, 2005.

[118] Duncan J Watts. *Six degrees: The science of a connected age*. WW Norton & Company, 2004.

[119] B. Wellman, A. Haase, J. Witte, and K. Hampton. Does the Internet Increase, Decrease, or Supplement Social Capital? Social Networks, Participation, and Community Commitment, 2001.

[120] Barry Wellman. Little boxes, glocalization, and networked individualism. In *Digital cities II: Computational and sociological approaches*. Springer, 2002.

[121] Barry Wellman. The glocal village: Internet and community. *Idea&s: The Arts & Science Review*, 2004.

[122] Barry Wellman and Caroline Haythornthwaite. *The Internet in everyday life*. John Wiley & Sons, 2008.

[123] L. Weng, A. Flammini, A. Vespignani, and F. Menczer. Competition among memes in a world with limited attention. *Scientific Reports*, 2, 2012.

[124] Lilian Weng, Jacob Ratkiewicz, Nicola Perra, Bruno Gonçalves, Carlos Castillo, Francesco Bonchi, Rossano Schifanella, Filippo Menczer, and Alessandro Flammini. The role of information diffusion in the evolution of social networks. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 356–364, 2013.

[125] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM conference on Web Search and Data Mining (WSDM)*, pages 177–186. ACM, 2011.

[126] Jing Yuan, Yu Zheng, and Xing Xie. Discovering Regions of Different Functions in a City Using Human Mobility and POIs. In *Proceedings of the* 18^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2012.

[127] Nicholas Jing Yuan, Fuzheng Zhang, Defu Lian, Kai Zheng, Siyu Yu, and Xing Xie. We Know How You Live: Exploring the Spectrum of Urban Lifestyles. In *Proceedings of the* 1^{st} *ACM Conference on Online Social Networks (COSN)*, 2013.

[128] Zhang, Amy X. and Noulas, Anastasios and Scellato, Salvatore and Mascolo, Cecilia. Hoodsquare: Modeling and Recommending Neighborhoods in Location-based Social Networks. In *Proceedings of IEEE Social Computing Conference (SocialCom)*, 2013.