Taking Brazil's Pulse: Tracking Growing Urban Economies from Online Attention

Carmen Vaca Ruiz[†‡][*]   Daniele Quercia[+]   Luca Maria Aiello[+]   Piero Fraternali[‡]

cvaca@fiec.espol.edu.ec, {dquercia,alucca}@yahoo-inc.com, piero.fraternali@polimi.it

[‡]Politecnico di Milano   [+]Yahoo Labs   [†]Escuela Superior Politecnica del Litoral
Milan, Italy   Barcelona, Spain   Guayaquil,Ecuador

ABSTRACT

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 To avoid allocating resources upon outdated indicators, one could partly update or complement them using digital data. It has been shown that it is possible to use social media in developed countries (mainly UK and USA) for such a purpose. Here we show that this is the case for Brazil too. We analyze a random sample of a microblogging service popular in that country and accurately predict the GDPs 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.

Categories and Subject Descriptors

H.4 [**Information Sytems Applications**]: Miscellaneous

Keywords

microblogging; online attention; urban indicators; social network analysis; social media

1. INTRODUCTION

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 inequality. To partly fix inequality, financial resources need to be invested, yet those
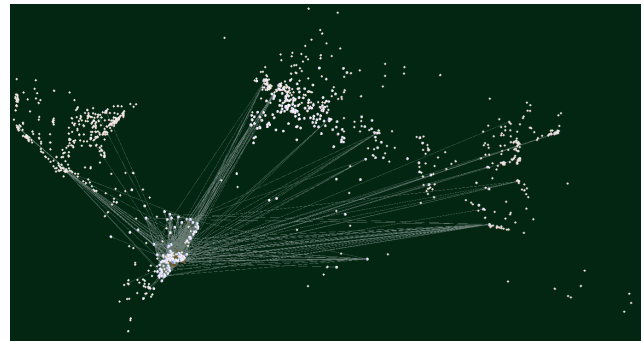
Figure 1: Graph showing the geographical locations of worlwide users who paid or received attention (i.e., reposted content) to/from cities considered in this study. Edges with low weights are not shown.

resources are scarce, and their allocation needs to be as targeted as possible [4]. To do so, one needs to profile city areas using socio-demographic factors.

In most developing countries, economic indicators at city level are often outdated [10]. 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.* [7] 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 ($r = .37$) [14]. Yet, those studies have been conducted only in developed countries such as USA and UK. We, thus, focus on Brazil, a 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 and examine the relationship between economic indicators and levels of attention paid to content produced by the residents of different cities, where attention is defined as the *interest* raised by user-generated content (as reflected by reposting content).

To conduct our analysis, we build upon the concept of *glocality* [27, 28], the combination of global and local interactions in which a city is involved. We propose indicators
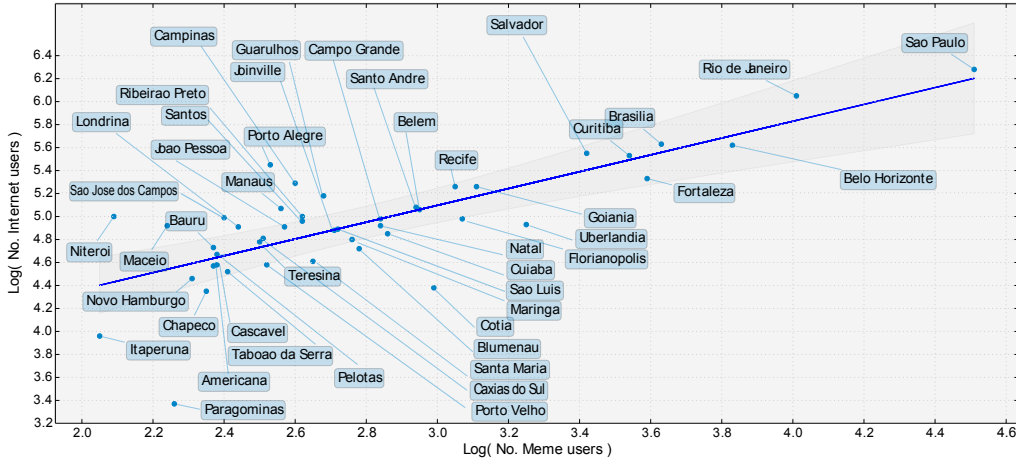
Figure 2: Number of users in our sample versus number of Internet users. Both quantities are log-transformed.

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 1: Yahoo Meme dataset statistics.

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 two 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 (Section 4).

- We put together the proposed online indicators to predict the economic capital of cities. We find that our models fit well the data and predict GDP with *Adjusted* $R^2 = 0.94$ (Section 5).

2. DATASET

Yahoo Meme was a microblogging platform, similar to Twitter, with the exception that users could 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 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) leaving us with 80K users. For this set of users and their respective posts, we extract all the repost *cascades*.

To attain geographic representativeness, 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 2).We build predictions models for two sets: the first consists only of cities that are within the confidence interval; and the second consists of all the 45 cities. We will see the results will not change and that, in both cases, they are statistically significant. That is because we are left with 1.4M repost cascades whose original content was produced in the 45 cities and was consumed across the world (Figure 1).

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 [27, 28]. Wellman states that online interactions enrich our *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 [27].

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 [6]. 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.
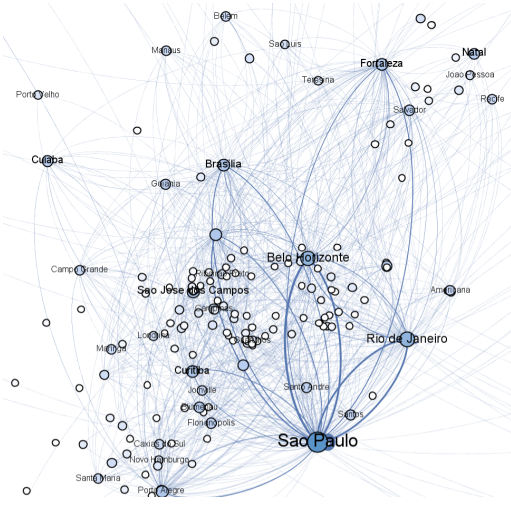
Figure 3: Attention graph whose nodes are cities and whose weighted edges reflect the intensity of reposting between cities' users.

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.

Since attention is a scarce resource, content published in social media attracts very different levels of it [1, 5, 12, 16]. Previous studies have shown that attention is not fully captured by simple metrics such as activity (number of posts) or popularity (number of followers, PageRank in the follower network) [1, 5]. As a result, Romero *et al.*, for example, used the amount of retweets as part of their Influence-Passivity measure [16]. In a similar way, 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). To quantify the attention received by users, two graphs are built:

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 3).

Cascade graph. A tree for each post is also built (Figure 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.

3.1 Global attention

Cities that enable global information flow are key actors in the world economy [17]. 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 [17], as they have strong connections to other cities.

Figure 4: Tree-like repost cascade. On the left, there is an example cascade, which is rooted at the content originator. On the right, a real repost cascade from our dataset.

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 [8], and this wealth will be created by cities that are connected to the rest of the world and not by those that are isolated. 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, normalized with respect to the total number n_i of users in that city: $ROW_i = \frac{out_i}{n_i}$, where out_i is the number of times a post originated in city i has been reposted outside it.

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 [20]. We quantify the gateway capacity of a city with brokerage attention. This captures the extent to which a city mediates the 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 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.

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 [2]. Given the skewness of the distributions, we 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.

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 [11]. More generally, the intervening opportunities hypothesis states that the number of persons moving to a given distance is inversely proportional to the number of intervening opportunities [22]. In the context of attention given to online content, this theory translates into saying that community members will devote considerable attention to content produced close to the city where they live, if that city offers considerable intervening opportunities, that is, is 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 [19], 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. ATTENTION AND GDP

Based on the literature (Section 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 ($r = 0.42$).

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 $r = 0.56$.

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 near locations. It is so because *local* reflects the geographical span of the entire re-

post cascade whereas *intra_city* is limited to the reposts attracted inside the city.

To account for skewness, all the attention metrics are log-transformed before calculating the correlations with each of the 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 [5, 12, 16] that simple activity metrics 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. The absence of correlations is caused by cities with low GDPs that happen to have high activity but low attention.

5. PREDICTING ECONOMIC CAPITAL

We model GDP of $city_i$ as a linear combination of the four attention metrics. We control for the city's Internet penetration rate and population (with data provided by the Brazilian census bureau in 2010[3]). 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 [8]. Model 1 is defined as follows:

$$\begin{aligned} log(GDP_i) = {}& \alpha + \beta_1 \cdot log(local_i) + \beta_2 \cdot log(ROW_i) + \\ & \beta_3 \cdot log(brokerage_i) + \beta_4 \cdot log(cascade_i) + \quad (1) \\ & + \rho \cdot log(Population_i) + + \mu \cdot Internet_i + \epsilon_i \end{aligned}$$

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 Model 2 adding the $Interactions_{im}$ (Table 2) that accounts for all possible pairwise product terms among the four attention predictors .

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 Model 1, the four attention metrics complement the predictive power of the census data, $Adj\ R^2$ is 0.94 (Table 2) 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 Model 1, 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%. 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 8.647, meaning that, on average, the model predicts GDP within 1.48% of its true value. Figure 5 plots predicted values against actual ones. The outlier for which GDP is higher than expected is Brasilia (the capital of the country).

We report the results obtained for models that consider the metric *local* for *local attention* and *eigenvector central-*
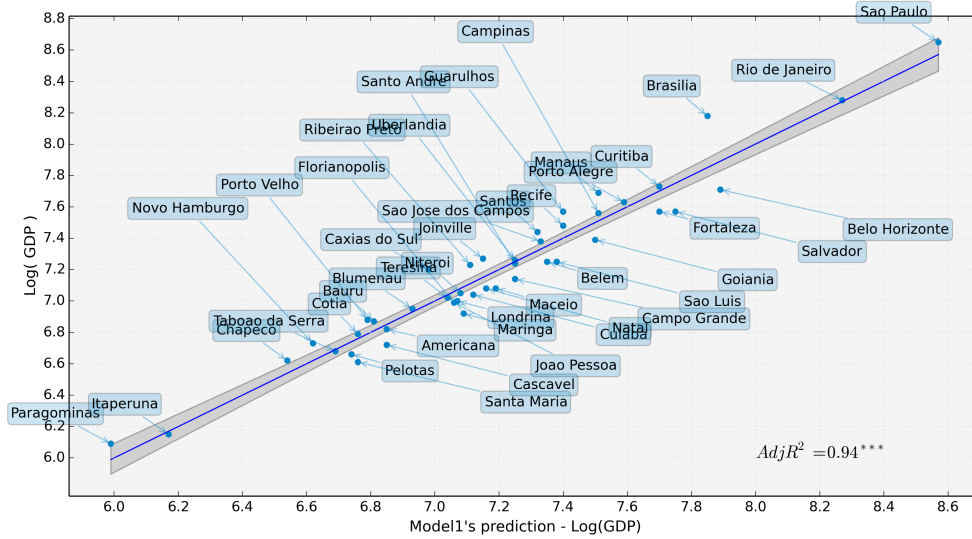
---

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

Figure 5: The Model1's performance predicting GDP with *Adj.* R^2=**0.94 for the set of 45 Brazilian cities.**

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

Table 2: *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-values* **are** < 0.001**.**

ity for *brokerage attention* as they exhibit higher correlation coefficients (Section 4). However, we repeated the linear modeling considering the alternative metrics(*intra_city, betweeness,Pagerank*) and obtained similar performance.

6. DISCUSSION

Theoretical Implications. Our results complement previous studies that correlated economic status with social media data at the level of urban *neighborhoods* [7, 14]. 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 [18]. 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' [28] taking advantage of the Internet to communicate with both local and long-range ties.

Practical Implications. Our work shows evidence that, with online attention metrics, one is able to effectively and cheaply predict the economic capital of a city (*Adj.* $R^2 = 0.94$). 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.

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 [23] and can now be informed by how countries connect online as well.

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, similar to what [15] did for London. Third, we have predicted a specific economic indicator. Currently we are expanding this work by predicting migration levels [26] and social capital for our Brazilian cities.

7. RELATED WORK

Real-life processes and social media. Social media data has already been related to real-life outcomes. Gruhl *et al.* [9] showed that increases in blog mentions of books correspond to spikes in sales. Tweets has been used to predict the Dow Jones Industrial Average [3], polls' results for political opinions [13], and election outcomes [24]. More recently, not only tweets but also email exchanges have been used to track migration flows among developed and developing countries [21].

Socio-economic indicators. 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 [7]. More recently, Quercia *et al.*

have shown a correlation between the sentiment expressed in tweets originated by residents of London neighborhoods and the neighbourhoods' well-being [14]. Our research complements 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. A key benefit of our approach lies in that it is language independent and it exploits the massive amount of information available on sharing-content platforms.

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. 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" [25]. To tackle that problem, policymakers are looking for new ways of monitoring local impacts, and tracking online attention might well be one such way.

9. ACKNOWLEDGMENTS

10. REFERENCES

[1] S. Asur, B. Huberman, G. Szabo, and C. Wang. Trends in social media: Persistence and decay. In *Proc. of the 5th AAAI ICWSM*, 2011.

[2] D. Bhattacharya and S. Ram. Sharing News Articles Using 140 Characters: A Diffusion Analysis on Twitter. In *Proc. of the IEEE/ACM ASONAM*, 2012.

[3] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2011.

[4] L. Capra and D. Quercia. Middleware for social computing: a roadmap. *Journal of Internet Services and Applications*, 2012.

[5] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proc. of the 4th AAAI ICWSM*, 2010.

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

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

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

[9] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *Proc. of the eleventh ACM KDD*, 2005.

[10] J. Henderson, A. Storeygard, and D. Weil. Measuring economic growth from outer space. Technical report, National Bureau of Economic Research, 2009.

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

[12] N. Naveed, T. Gottron, J. Kunegis, and A. Che Alhadi. Bad News Travels Fast: A Content-based Analysis of Interestingness on Twitter. In *Proc. of the 3rd ACM WebSci*, 2011.

[13] B. O'Connor, R. Balasubramanyan, B. Routledge, and N. A. Smith. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proc. of the 4th AAAI ICWSM*, 2010.

[14] D. Quercia, J. Ellis, L. Capra, and J. Crowcroft. Tracking gross community happiness from tweets. In *Proc. of the ACM CSCW*, 2012.

[15] D. Quercia, D. Ó Séaghdha, and J Crowcroft. Talk of the City: Our Tweets, Our Community Happiness. In *Proc. of 6th AAAI ICWSM*, 2012.

[16] D. M Romero, W. Galuba, S. Asur, and B. A Huberman. Influence and passivity in social media. In *Machine Learning and Knowledge Discovery in Databases*. Springer, 2011.

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

[18] S. Scellato, C. Mascolo, M. Musolesi, and V Latora. Distance matters: geo-social metrics for online social networks. In *Proc. of the 3rd WOSN*, 2010.

[19] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. In *Proc. of the 5th AAAI ICWSM*, 2011.

[20] JR. Short, C. Breitbach, S. Buckman, and J. Essex. From world cities to gateway cities: extending the boundaries of globalization theory. *City*, 2000.

[21] B. State, I. Weber, and E. Zagheni. Studying international mobility through IP geolocation. In *Proc. of the sixth ACM WSDM*, 2013.

[22] S. A. Stouffer. Intervening opportunities: a theory relating mobility and distance. *American Sociological Review*, 1940.

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

[24] A. Tumasjan, T Oliver Sprenger, P. Sandner, and I. Welpe. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *Proc. of the 4th AAAI ICWSM*, 2010.

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

[26] C. Vaca Ruiz, D. Quercia, L. M. Aiello, and P. Fraternali. Tracking Human Migration from Online Attention. In *Citizen in Sensor Networks*. Springer, 2014.

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

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