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
Milano, Italy

[†]Yahoo Research
Barcelona, Spain

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 To avoid allocating resources upon outdated indicators, one could partly update or complement them using digital data.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 both economic capital and social capital of 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.

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 resources are scarce, and their allocation needs to be as targeted as possible [9]. 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In most developing countries, economic indicators at city level are often outdated [23]. 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.* [13] 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 ($r = .37$) [36]. 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 [11]: the Boston Consulting Group has classified as many as 34 Brazilian cities as emerging markets [25]. 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 *interest* raised by user-generated content (as reflected by reposting content).

To conduct our analysis, we build upon the concept of *glocality* [50, 51], 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 (Section 4).
- 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 [28]. We compute a city's prestige index, which

[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/

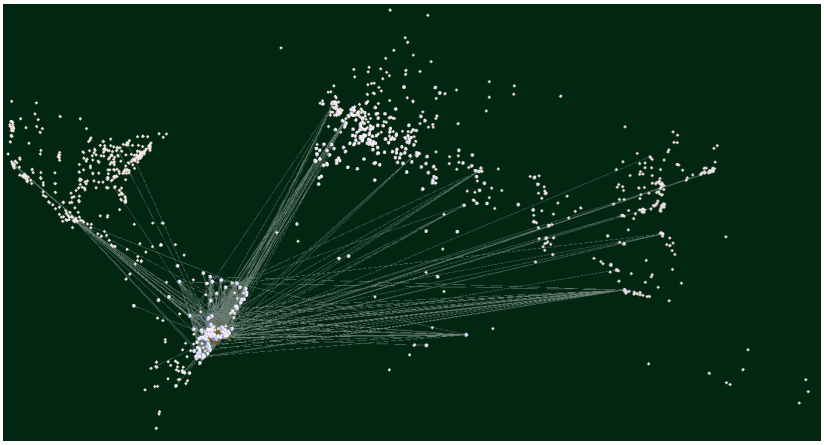

Figure 1: Graph showing the geographical locations of worlwide users who paid or received attention (i.e., reposted content) to/from cities considered in this study. Edges with low weights are not shown.

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 (Section 5).

- 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* $R^2 = 0.93$ (Section 6).

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 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 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: 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 1).

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.

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 [50, 51]. 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 [50].

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 [12]. 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

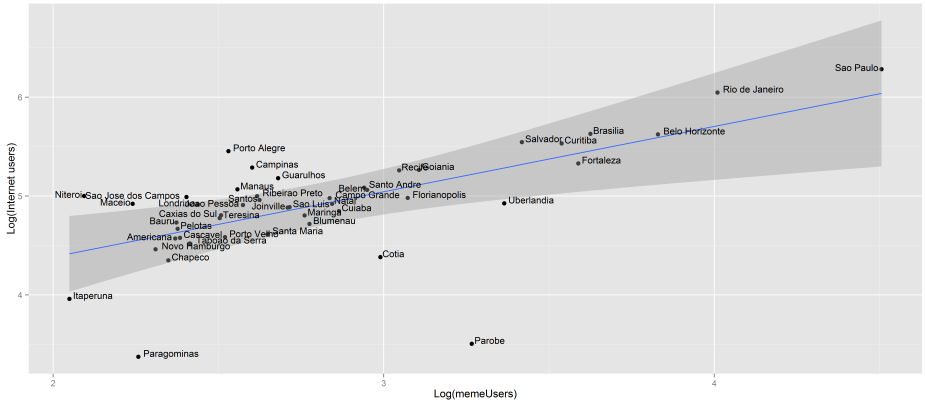

Figure 2: Number of users in our sample versus number of Internet users. Both quantities are log-transformed.

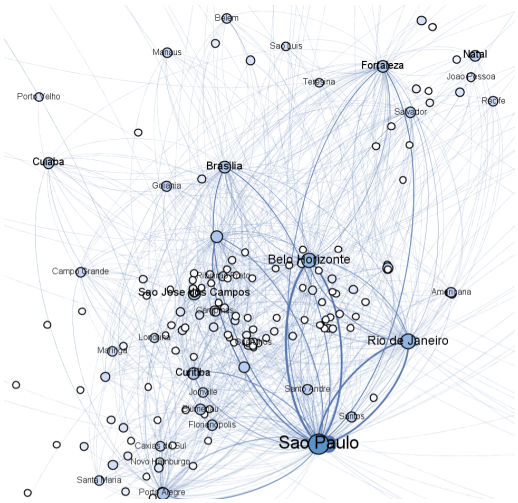

Figure 3: Attention graph whose nodes are cities and whose weighted edges reflect the intensity of reposting between cities' users.

Figure 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.

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. Thus, users who receive attention are those who produce high-quality content and enjoy advantageous positions in the social graph [24, 54].

Since attention is a scarce resource, content published in social media attracts very different levels of it [3, 10, 17, 31, 32, 37, 49, 53]. 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) [3, 10, 49]. As a result, Romero *et al.*, for example, used the amount of retweets as part of their Influence-Passivity measure [37]. 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 [39]. 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 [39], 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 [18], 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:

$$ROW_i = \frac{out_i}{n_i}, ROW'_i = \frac{out'_i}{n_i}$$

where out_i is the number of times a post originated in city i has been reposted outside it (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 [42]. 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 [5]. 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 [26]. 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 [45]. 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 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 [41], 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 ($r = 0.42$) 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

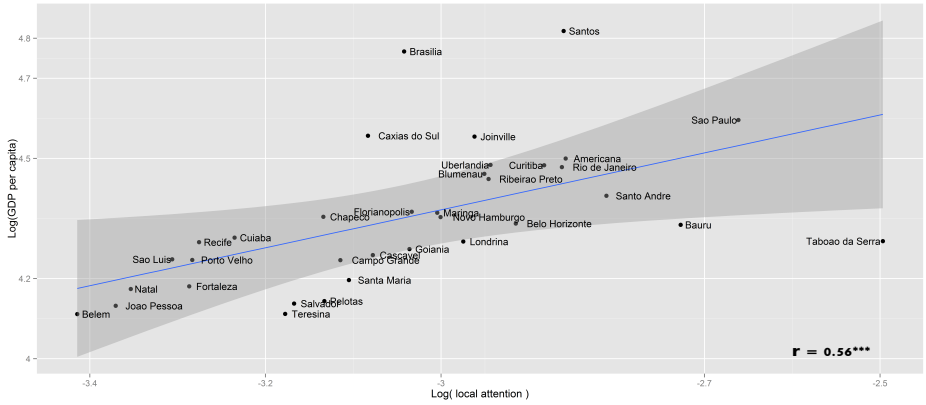

Figure 5: Local attention of a city vs. GDP per capita. In more prosperous cities, community members will devote considerable attention to content produced locally (p-value < 0.001 is expressed with *).**

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 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.

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

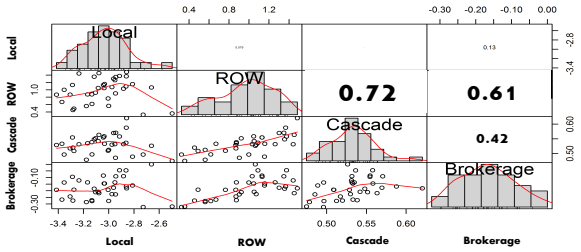

Figure 6: Correlations among the four attention metrics. We observe that the *ROW* attention metric is correlated both with *cascade* attention and *brokerage* attention. Values are log-transformed.

do social contacts affect the productivity of individuals and groups" [3].

There is no consensus on how to measure social capital. Some researchers have measured it through the analysis of participation in volunteerism [35], 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 [28] is often used[4]. This measures the range of people's social ties. Researchers ask their participants 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 [16] (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 pres-

[3] `http://en.wikipedia.org/wiki/Social_capital`

[4] There are different definitions of this specific instantiation of social capital, and Lin's has been widely accepted together with few others

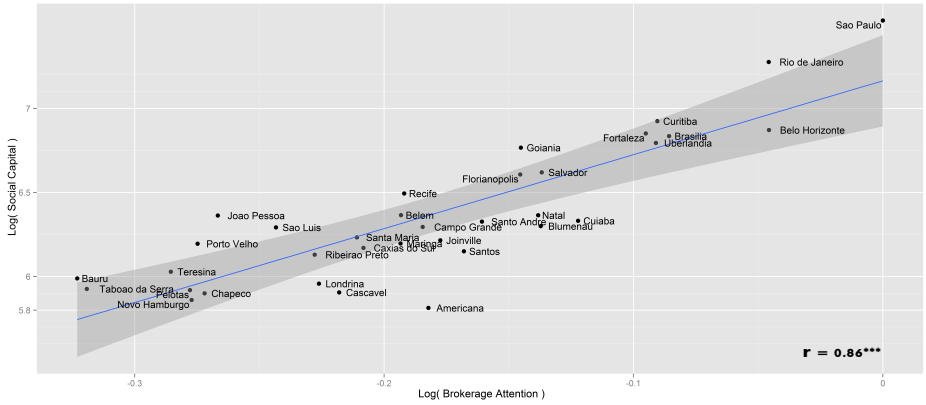

Figure 7: Brokerage attention (Eigenvector centrality) vs. social capital (highest accessed prestige). Cities with higher city prestige indexes do mediate information flow.

tige (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[5] and compute the *highest accessed prestige* for each of our cities, which we make publicly available [6]. 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 7 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).

6. PREDICTING SOCIO ECONOMIC CAPITALS

We separately model GDP per capita and social capital as a linear combination of the attention metrics plus terms to account for pairwise interactions between indicators (i.e., interaction effects). For example, for GDP per capita, we have the model 1 that uses, as the predictors, the four attention metrics calculated for the city i without taking into account any census data:

$$log(GDPpc_i) = \alpha + \beta_1 \cdot log(local_i) + \beta_2 \cdot log(ROW_i) + \beta_3 \cdot log(brokerage_i) + \beta_4 \cdot log(cascade_i) + \epsilon_i \quad (1)$$

where $GDPpc_i$ is the GDP per capita of city i and ϵ_i is the error term. To account for the skewness of the data, we log-transformed each variable.

Next, we calculate the correlation among each pair of the proposed metrics above (Figure 6) and observe that the group of indexes to measure *global attention* are correlated among each other, as one can expect: *ROW* is correlated with *brokerage* ($r = 0.65$) and *cascade*($r = 0.72$); also, *brokerage* and *cascade* show a positive correlation ($r = 0.42$). To account for such correlations among the predictors, we build the model 2 adding, to the model 1, what we call $Interactions_{im}$, i.e., all possible pairwise product terms among the four attention predictors, for instance, $log(local_i) \cdot log(ROW_i)$ would be one of the products computed and added to the predictors.

Model 3 takes into consideration the census data. Our goal at building this model is to compare the performance of the online attention-based models and that which also includes city size as it is known that socio-economic indicators are correlated with it [4, 34].

$$log(GDPpc_i) = \alpha + \beta_m \cdot Attention_{im} + \rho \cdot log(Population_i) + \mu \cdot Internet_i + \epsilon_i \quad (2)$$

where $Attention_{im}$ is the log-value for each of the four at-

Model	Predictors	$Adj.R^2$ 35	$Adj.R^2$ 45
1	$\{Attention_{im}\}$	**0.47**	**0.51**
2	$\{Attention_{im}\} + \{Interactions_{im}\}$	0.44	0.46
3	$\{Attention_{im}\} + Internet_i + Population_i$	0.49	0.51
4	$ROW_i + local_i$	0.43	0.45
5	$brokerage_i + local_i$	0.40	0.44
6	$cascade_i + local_i$	**0.47**	0.45

Table 2: $Adj.\ R^2$ for different models predicting city i's GDP per capita. $Attention_{im}$ represents the four attention metrics. $Interactions_{im}$ represents all the pairwise product terms among the attention predictors. Model 1's predictors are the four attention metrics m, model 2 adds their interaction effects, model 3 controls for the city's Internet penetration rates and population, and Models 4-6 test (pairwise) attention metrics separately. p-values are < 0.001.

[5]http://www.ibge.gov.br
[6]anonymized-link

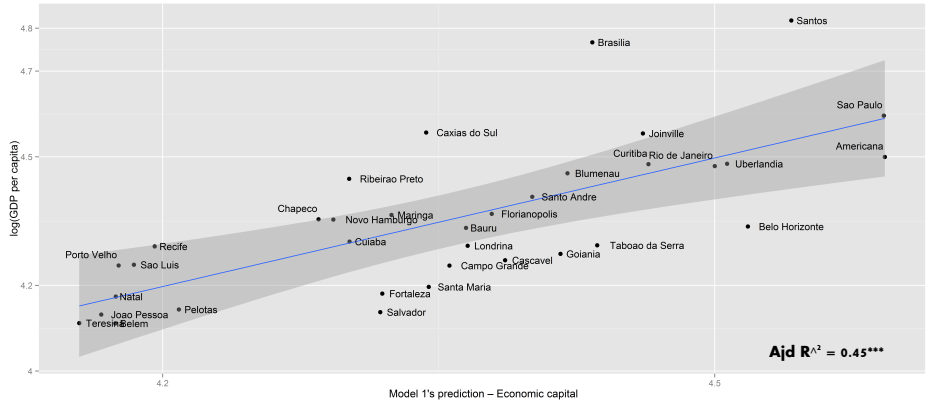

Figure 8: The model 1's performance for predicting GDP per capita with *Adj.* R^2=0.45. Despite the strong correlation, our model understimates or overstimates the value for some cities (shown outside the confidence region). The model's prediction error is low: its Mean Absolute Error is 0.09.

tention metrics m for city i, $Population_i$ is the city's population and $Internet_i$ is the city's Internet's penetration rate. This is what we will call model 3 in Table 2, which describes all the six models and report the *Adj.* R^2 obtained on the two sets of cities described on Section 2 (35 and 45 cities).

Local attention dimension does not correlate with any other dimension, and this speaks to its complementary predictive power. Thus, we build models using *local attention* with one of the *global attention* metrics(models 4,5,6).

To sum up, models 1 and 2 are built using the four online attention metrics derived in Section 3 (model 2 additionally accounts for pairwise interactions between those metrics). In model 3, we control for the city's Internet penetration rate and population (with values extracted from the same census data's source). 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 [19, 4, 34]. Bettencourt *et al.* have shown that when population increases in a 100%, there can be a 15% of improvement in economic indicators and Pan *et al.* observes that this relationship doesn't always verify in developing countries. We observe that, after controlling for those two variables, the predictive power of the model 1 remains almost the same, *Adj* R^2 goes from 0.47 to 0.49 for the set of 35 cities and 0.51 in both cases for the set of 45 cities(Table 2) with a 25% of the variance explained by the census data and the remaining by the attention metrics as shown in Table 3. After controlling for census data variables, the attention metrics keep their predictive value.

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).

By computing the beta coefficients of model 1, the one with the best performance (without census data), we find that *local attention* accounts for 18% of the models' explanatory power while the aggregated β values of the three *global attention* metrics contribute with 82%. Out of the three global attention metrics, *cascade* attention has the highest impact as it explains 65% of GDP's variance. As for model 1's accuracy, the model achieves a Mean Absolute Error (MAE) of 0.09 on a logarithmic scale, where the minimum value is 4.11 and maximum is 4.81, meaning that, on average, the model predicts GDP per capita within 1.22% of its true value. Figure 8 plots predicted values against actual ones. The two outliers for which GDP is higher than expected are Brasilia (the capital of the country) and Santos (the biggest seaport in Latin America).

Predictor	β value	%
$log(ROW_i)$	0.0404	02%
$log(local_i)$	0.26738	11%
$log(cascade_i)$	1.52596	62%
$log(brokerage_i)$	0.04589	02%
$log(Internet_i)$	0.59617	24%
$log(Population_i)$	0.02027	01%

Table 3: β coefficients for each of the terms on model 3 that predicts GDP per capita for 35 cities.

We repeated the same linear modeling to predict a city's social capital (M7-10). We observe that after controlling for Internet penetration and population, the predictive power of the model 7 goes from 0.88 to 0.91 with a 5.7% of the variance explained by the census data. We found that the model 8, that includes attention predictors plus interaction effects, has the highest *Adj.* R^2, which is equal to 0.93 (Table 4), achieving a Mean Absolute 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 model 8 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 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 10) variables reports the lower performance. We consider the *brokerage* metric in the model 10 as it is the unique

Model	Predictors	$Adj.R^2$ 35	$Adj.R^2$ 45
7	$\{Attention_{im}\}$	0.88	0.84
8	$\{Attention_{im}\} + \{Interactions_{im}\}$	**0.93**	0.85
9	$\{Attention_{im}\} + Internet_i + Population_i$	0.91	0.86
10	$brokerage_i + local_i$	0.72	0.66

Table 4: $Adj.\ R^2$ for different models predicting city i's social capital. Model 7's predictors are the four attention metrics m, model 8 adds their interaction effects, model 9 controls for the city's Internet penetration rates and population, and models 10 test brokerage and the local attention metric separately. p-$values$ are < 0.001.

global attention metric that correlated well with the social capital index.

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). However, we repeated the linear modeling considering the alternative metrics(*ROW'*,*intra_city*,*betweeness-Pagerank*) and obtained similar performance.

7. DISCUSSION

7.1 Theoretical Implications

Our results complement previous studies that correlated economic status with social media data at the level of urban *neighborhoods* [13, 36]. 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 [40]. 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' [51] 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 [20] and the structural hole theory [8] would suggest) but also for cities.

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

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.

8. RELATED WORK

Real-life processes and social media. Social media data has already been related to real-life outcomes. Gruhl *et al.* [21] showed that increases in blog mentions of books correspond to spikes in sales. Tweets has been used to predict the Dow Jones Industrial Average [6], box-office revenues for movies [2], trending tickers in the stock market [38], polls' results for political opinions [33], and election outcomes [47]. 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 [30]. More recently, not only tweets but also email exchanges have been used to track migration flows among developed and developing countries [43].

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 [52]. 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 [14]. Furthermore, individual bridging and bonding social capital can be accurately predicted using Facebook interactions [44]. 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 [22].

Socio-economic indicators. Previous studies have also explored the connection between online interactions and socio-

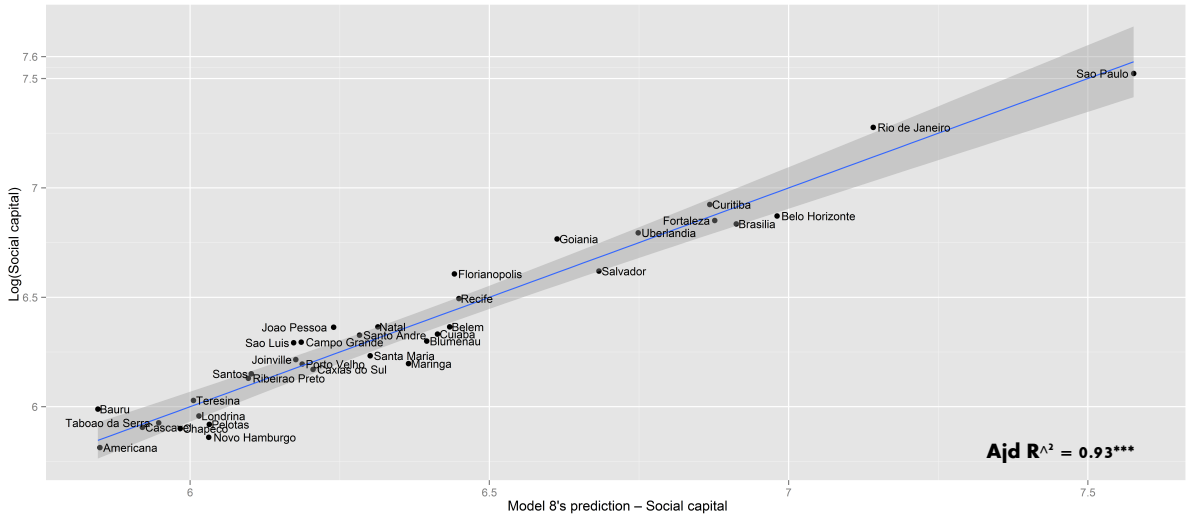

Figure 9: The model 8's performance for predicting social capital. Its Mean Absolute Error is as low as 0.08.

economic indicators of city neighbourhoods [15]. 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 [13]. 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 [36].

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 [29]. 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.

Online attention. As for online attention, instead, research has focused on the graph perspective, analyzing the dynamics of popularity and information diffusion of user-generated content [3, 10, 17, 31, 32, 37, 49, 53], and proposing models to reproduce them [54, 27]. Brodersen *et al.* investigated the relationship between popularity and geographic spreading of online YouTube videos and observed that, despite the virality of the most popular items, online video consumption appears heavily constrained by geographic locality [7].

9. 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" [48]. To tackle that problem, policymakers are looking for new ways of monitoring local impacts, and tracking online attention might well be one such way.

10. REFERENCES

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

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

[3] Asur, S., Huberman, B. A., Szabo, G., and Wang, C. Trends in social media: Persistence and decay. In *Proceedings of the 5th AAAI Conference on Weblogs and Social Media (ICWSM)* (2011).

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

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

in Advances in Social Networks Analysis and Mining (ASONAM) (2012).

[6] Bollen, J., Mao, H., and Zeng, X. Twitter mood predicts the stock market. *Journal of Computational Science 2*, 1 (2011), 1–8.

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

[8] Burt, R. S. Structural Holes and Good Ideas. *The American Journal of Sociology* (2004).

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

[10] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the 4th AAAI Conference on Weblogs and Social Media (ICWSM)* (2010).

[11] Dobbs, R., Remes, J., and Smit, S. The world's new growth frontier: Midsize cities in emerging markets. *McKinsey Quarterly, http://www.mckinsey.com* (2011).

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

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

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

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

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

[17] Ghosh, R., and Lerman, K. Predicting influential users in online social networks. In *Proceedings of the 4th AAAI Conference on Weblogs and Social Media (ICWSM)* (2010).

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

[19] Glaeser, E. L., and Kohlhase, J. E. Cities, regions and the decline of transport costs. *Regional Science* (2004).

[20] Granovetter, M. The strength of weak ties: A network theory revisited. *Sociological theory* (1983).

[21] Gruhl, D., Guha, R., Kumar, R., Novak, J., and Tomkins, A. The predictive power of online chatter. In *Proceedings of the eleventh ACM conference on Knowledge discovery in data mining (KDD)* (2005).

[22] Harris, K., and Flouch, H. Online neighbourhood networks study: Social capital and cohesion. Tech. rep., The Networked Neighbourhoods Group, 2010.

[23] Henderson, J. V., Storeygard, A., and Weil, D. N. Measuring economic growth from outer space. Tech. rep., National Bureau of Economic Research, 2009.

[24] Huberman, B. A., Romero, D. M., and Wu, F. Crowdsourcing, attention and productivity. *Journal of Information Science* (2009).

[25] Jin, D., Michael, D., Foo, P., Guevara, J., Peña, I., Tratz, A., and Verma, S. 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).

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

[27] Lerman, K., Jain, P., Ghosh, R., Kang, J.-H., and Kumaraguru, P. Limited attention and centrality in social networks. In *Proceedings of Conference on Social Intelligence and Technology (SOCIETY)* (2013).

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

[29] Mao, H., Shuai, X., Ahn, Y.-Y., and Bollen, J. Mobile communications reveal the regional economy in côte dŠivoire. In *Proceedings of the 3rd Conference on the Analysis of Mobile Phone Datasets (NetMob)* (2013).

[30] Mejova, Y., Srinivasan, P., and Boynton, B. 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).

[31] Naaman, M., Becker, H., and Gravano, L. Hip and trendy: Characterizing emerging trends on Twitter. *Journal of the American Society for Information Science and Technology* (2011).

[32] Naveed, N., Gottron, T., Kunegis, J., and Che Alhadi, A. Bad News Travels Fast: A Content-based Analysis of Interestingness on Twitter. In *Proceedings of the Web of Science Conferece* (2011).

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

[34] Pan, W., Ghoshal, G., Krumme, C., Cebrian, M., and Pentland, A. Urban characteristics attributable to density-driven tie formation. *Nature communications 4* (2013).

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

[36] Quercia, D., Ellis, J., Capra, L., and Crowcroft, J. Tracking gross community happiness from tweets. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)* (2012).

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

[38] Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A., and Jaimes, A. Correlating financial time series with micro-blogging activity. In *Proceedings of the fifth ACM international conference on Web search and data mining (WSDM)* (2012).

[39] Sassen, S. The global city: Introducing a concept. *Brown Journal of World Affairs* (2004).

[40] Scellato, S., Mascolo, C., Musolesi, M., and Latora, V. Distance matters: geo-social metrics for online social networks. In *Proceedings of the 3rd conference on Online social networks (WOSN)* (2010).

[41] Scellato, S., Noulas, A., Lambiotte, R., and Mascolo, C. Socio-spatial properties of online location-based

social networks. In *Proceedings of the 5th AAAI Conference on Weblogs and Social Media (ICWSM)* (2011).

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

[43] State, Bogdan, W. I., Zagheni, E., et al. Studying inter-national mobility through ip geolocation. In *Proceedings of the sixth ACM Conference on Web search and data mining (WSDM)* (2013).

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

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

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

[47] Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. Predicting elections with twitter: What 140 characters reveal about political sentiment. *Proceedings of the 4th AAAI Conference on Weblogs and Social Media (ICWSM)* (2010).

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

[49] Ver Steeg, G., and Galstyan, A. Information transfer in social media. In *Proceedings of the 21st ACM Conference on World Wide Web (WWW)* (2012).

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

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

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

[53] Weng, J., Lim, E., Jiang, J., and He, Q. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM conference on Web search and data mining (WSDM)* (2010).

[54] Weng, L., Flammini, A., Vespignani, A., and Menczer, F. Competition among memes in a world with limited attention. *Scientific reports* (2012).