The Geography of Online News Engagement

Martin Saveski[1], Daniele Quercia[2], and Amin Mantrach[2]

[1] MIT Media Laboratory, Cambridge, MA, USA
msaveski@mit.edu
[2] Yahoo Labs, Barcelona, Spain
dquercia@acm.org, amantrac@yahoo-inc.com

Abstract. Geographical processes might well impact online engagement in big countries like the USA. Upon a random sample of 200K news articles and corresponding 41M comments posted on the Yahoo! News in that country, we show that nearby individuals tend to comment and engage with similar news articles more than distant individuals do. Interestingly, at state level, topics one reads about are associated with specific socio-economic conditions and personality traits.

1 Introduction

Online actions whose geographic processes have been well-studied include not only posting status updates on Twitter [34,12,14], but also uploading pictures on Flickr [7,27], and visiting Foursquare venues [26,24].

Despite their importance, the geographic processes of online engagement on news platforms have not been widely studied. To partly fix that, we consider a dataset containing articles and user comments posted on the Yahoo! News site for more than two years, and we make two main contributions:

- We find that users engage with each other (i.e., they comment on the same articles) depending on where they live (Sections 4 and 5).
- Since one's interests have been linked to one's socio-economic conditions and personality traits, we test whether this is also the case at geographic level, and we do so by combing our online data with census data (Section 6). We find that those in states with high levels of education and well-being comment articles about research&technology but not those about politics, gossips, or sport. Instead, those in states with high levels of crime and unemployment comment on articles about sports, but not on those about economy or research&technology. Also, as for personality traits, users from states that tend to have residents low in Neuroticism (emotionally stable) comment on articles about music, those in Open and Extravert states on articles about sports, and Conscientious states on articles about economics.

2 Related Work

The main goal of this work is to study the influence of geographic processes on user engagement with online news. Next, we review work related to this topic.

L.M. Aiello and D. McFarland (Eds.): SocInfo 2014, LNCS 8851, pp. 279–289, 2014.

Influence of Time on Our Actions Online. Golder and Macy [11] examined how the use of emotion words by Twitter users changed over the course of one day, and they found that it was regularly shifted along time zones. That is similar to what Mislove *et al.* [25] independently reported when contrasting the usage of Twitter in the west coast with that in the east coast.

News in Tweets and Geographic Spread on Twitter. Kwak *et al.* [23] found that reciprocal relations on Twitter (75% of them) tend to be between users who live no more than three time zones away, hinting that the geographical distance may be related to the interest similarity. Recent studies have also examined the geographic spread of topics on Twitter by investigating the adoption of hashtags across locations around the world [20]. They found that physical distance between locations constrained the spreading of hashtags: the adoption of the same hashtag by two locations was inversely proportional to their geographical distance.

User Engagement in Online News Platforms. Jones and Altadonna [16] examined the introduction of badges (i.e., awards for users with frequent posting) to encourage user engagement on the Huffington Post website. They found that longer threads do not come from badges, but from the desirability of news articles. Diakopoulos and Naaman [8] studied the relationships between news comment topicality, temporality, sentiment, and quality. They found that some topics aroused more deleted comments (by the moderators), and correlation between the negative sentiment and the fraction of deleted comments. They also found that the frequency at which users comment is correlated with the negativity of the comments.

From this brief literature review, one concludes that we hitherto lack a detailed understanding of how geography impacts the engagement on news platforms. We thus set out to partly fix that by studying how geographical processes impact user engagement on news articles (Sect. 4).

3 Initial Analysis

3.1 Data Description

Our dataset consists of a random sample of 200K news articles and corresponding 41M comments, published from August 2010 to February 2013. Yahoo! News features articles from a variety of news publishers including: Reuters, ABC News, Associated Press, The Atlantic Wire and other. For each article, we know its publication time and comments. Each comment comes with a timestamp, the commenter's anonymous user identifier and *IP* address (which we translate into the corresponding city name using the Yahoo! Places Web service).

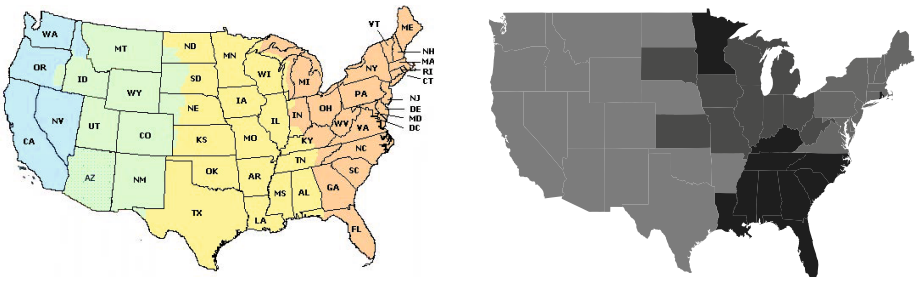

Fig. 1. (Left) USA Map of time zones. (Right) US Map of like-minded states that engage on the same articles (four different clusters of like-minded states emerge).

3.2 State Commenting Graph

To understand whether any geographical process shapes user engagement, we build a graph whose nodes are US states and whose links are weighted with the number of times two users in states i and j comment on the same article. To see the extent to which different states show similar commenting patterns (whether they are like-minded, in that, they tend to engage with the same articles), we apply a community detection algorithm on the graph. We use the Louvain community detection algorithm [5], whose main advantages are both the automatic detection of the optimal number of communities (no need to set that number a priori) and its high clustering accuracy [9]. After running the algorithm, four main clusters of like-minded states are detected and mapped in Figure 1 (right). Interestingly, we see that the four detected groups are geographically clustered (i.e., cover contiguous regions). Furthermore, one readily sees a similarity between this map and the USA Map of time zones (left panel of Figure 1).

4 The Time Zone Effect

To quantify whether time zone affects engagement, we test the hypothesis:

[H1] *Users in the same time zone preferentially engage with the same articles, while users in different time zones engage with different articles.*

To this end, we perform an experiment in three steps (which we shall detail): (1) We measure the observed engagement among users in the same time-zone, 1-time, ..., k time zones apart; (2) By keeping all factors constant except the time zone that are randomly permuted, we measure again the user engagement due to chance; and (3) we compare both engagement measures to assess if the time zone affects engagement.

(1) Engagement in *k*-Time Zone Apart. To measure engagement, we associate users with their time zones[1] and count the number of times users from k-time zone apart engage in the same articles. More formally, we measure the probability p_k that two users in k time zones apart engage in the same article:

$$p_k = \frac{\sum_{i \in S} \sum_{j \in S} I_k(i,j) \cdot interaction_{ij}}{n},$$

where S is the set of all states; I_k is an indicator function that equals to 1 if states i and j are k time zones apart, or 0 otherwise; $interaction_{ij}$ is the number of times users from states i and j have engaged in the same article; and n normalizes the numerator for the total number of interactions across all time zones.

(2) Engagement due to Chance. To test whether what we observe is not due to chance, we resort to a null (random) model [30]. We reshuffle the assignment of time zones by associating each user to a random zone, and repeat this procedure 2000 times to obtain accurate estimates. The random model removes the time zone effect and keeps all other factors constant. Thus, the difference between the engagement values that are observed and those in the random model depend only on effects strictly related to time zones. If there is no difference, then what we observe does not depend on time zone. As one may expect, if the time zones associated with each user are shuffled, the probability of engagement between two users is approximately the same (i.e., $\approx$0.27) regardless of the time difference.

(3) Compare the Two Engagements. By comparing the observed engagement with the engagement under the random model (Figure 2), we find that users in the same time zone and (to a lesser extent) those one time zone away engage with the same articles (first two dark bars) more than expected by chance (light bars). By contrast, those in three and four-time zone away engage less than chance. We perform a t-test to verify whether the differences between observed values and those in the random model are statistically significant. We find that all differences are significant at p-value less than 0.001.

5 The Geography of News Engagement

We have just ascertained that users who live in the same time zone interact with each other more than what people in different time zones do. Since our null model is oversimplified, we now adopt a geographic notion that is finer grained than that of time zones. We do so by resorting to a widely-used spatial interaction model called "the gravity model" [35]. In analogy to the gravitational interaction between planetary bodies, the model posits that the interaction between two

[1] States that belong to more than one time zone are assigned to the time zone in which the majority of the territory belongs to. We considered only the continental states, Alaska and Hawaii have been excluded from the analysis.

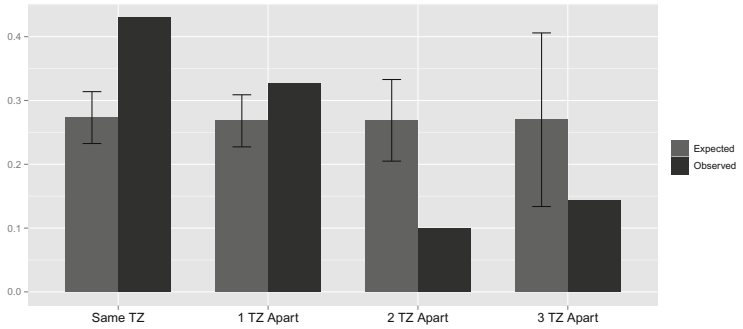

Fig. 2. The probability that two users who are k time zone (TZ) apart engage on the same article. Light bars show the expected engagement in a random model (suppressing the time zone effect), and the dark bars show the observed levels of engagement.

places (e.g., two states) is proportional to their mass (e.g., their population) and inversely proportional to their distance. Despite some criticisms [31], the model has been successfully used to describe 'macro scale' interactions (e.g., between cities, and across states), using both road and airline networks [4,18] and its use has extended to other domains, such as the spreading of infectious diseases [3,33], cargo ship movements [19], and to model intercity phone calls [22].

Here we posit that a gravity model can be used to estimate user engagement on the same articles at the *inter*-state level. The model takes the form:

$$F_{i,j}^{est} = g\frac{m_i m_j}{d_{i,j}^2} \tag{1}$$

where $F_{i,j}^{est}$ is the estimated engagement, or number of comments users living in states i and j make on the same articles, g is a scaling constant fitted to the data, and $d_{i,j}$ is the distance between the two states, for which we use the Euclidean distance between the two centroids of i and j. Engagement between areas with large number of users and at short distances are predicted to be large, whereas engagement at longer distances or between areas with low mass are predicted to be small. Overall, the correlation between the observed number of comments and gravity model estimates, measured with the Pearson Correlation Coefficient, is as high as .70, which suggests that overall the gravity model provides a good description of user engagement between states, but also that there is still a significant amount of variation not accounted for by the model. We posit that this unexplained portion is due to prevailing socioeconomic factors.

6 The Socio-economic Factors of Engagement

To begin with, we assign topics to both articles and comments. Since we need explicit topic labels (previously we just needed to compute similarity measures),

Table 1. The big five personality traits

Personality trait	High scorers	Low scorers
Openness	Imaginative	Conventional
Conscientiousness	Organized	Spontaneous
Extraversion	Outgoing	Solitary
Agreeableness	Trusting	Competitive
Neuroticism	Prone to stress and worry	Emotionally stable

we cannot use unsupervised techniques (e.g., topic modeling). Instead, we opt for studying a subset (13.8%) of the articles that have been editorially labeled with topical categories from the IPTC news subject taxonomy[2]. The taxonomy consists of 1400 topics and is organized into three levels, according to the specificity of the topics. To have the finest-grained topical view, we use the lowest level of the taxonomy. The number of labels associated with each article ranges from 1 to 25, where the average number of labels per article is 5. We aggregate these topics at state level by considering the number of times users from a given state commented on articles with a certain tag, and the number of times the tag appears in the data set (to avoid the bias of dominant topics).

The Big Five Personality Traits. The five-factor model of personality, or the big five, is the most comprehensive, reliable and useful set of personality concepts [6,10]. An individual is associated with five scores that correspond to the five main personality traits and that form the acronym of *OCEAN* (Table 1 collates a brief explanation). Imaginative, spontaneous, and adventurous individuals are high in **O**peness. Ambitious, resourceful and persistent individuals are high in **C**onscientiousness. Individuals who are sociable and tend to seek excitement are high in **E**xtraversion [2,32]. Those high in **A**greeableness are trusting, altruistic, tender-minded, and are motivated to maintain positive relationships with others [15]. Finally, emotionally liable and impulsive individuals are high in **N**euroticism [17,21].

These big five traits have been studied not only at individual level but also at geographic level [28]. Rentfrow *et al.* [29] have examined the personality scores of half a million US residents and found clear patterns of regional variation across the country, and they have also strong relationships between state-level personality and socioeconomic indicators.

We now correlate state-level personality scores with engagement with articles about specific topics (Figure 3, right). Economy is popular in states with conscientious residents ($r = 0.42$), and unpopular in states with residents who tend to be agreeable ($r = -0.61$) and open ($r = -0.42$). Sport is popular in states whose residents tend to be both extroverts ($r = 0.49$) and open to new experiences ($r = 0.50$). As one might expect, agreeable states avoid articles about religion ($r = -0.53$) and war&unrest ($r = -0.63$). The latter category is also avoided

[2] `http://www.iptc.org/site/NewsCodes/View_NewsCodes/`

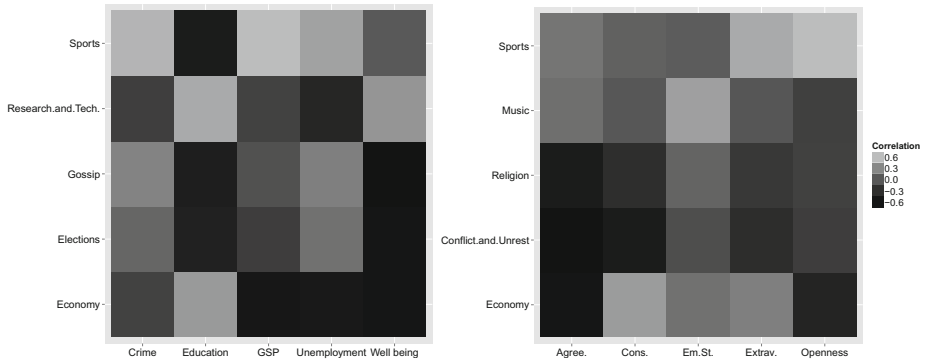

Fig. 3. Correlation between state's topics of interest and: socioeconomic indicators (left panel) and personality traits (right panel)

by conscientious states ($r = -0.49$). States with prevalence of neuroticism (emotional instability) tend to avoid article about music&theater ($r = 0.44$). Finally, states with low levels of neuroticism (i.e., emotional stability) show interest in diverse topics ($r = -0.44$).

Socioeconomic Indicators. We analyze the correlations between a state's assigned topics and the five most studied socioeconomic indicators: well-being index[3], crime level[4], rate of unemployment[5], Gross State Product[6], and education level[7] (number of people with higher education).

As illustrated in Fig. 3 (left), states with high levels of well-being (satisfaction with life) do not engage with articles about economy&business&finance ($r = -0.50$), about elections ($r = -0.53$), or about gossip&celebrities ($r = -0.53$). Economy is also not popular in states with unemployment ($r = -0.46$). Sport, instead, is popular in states with high levels of crime ($r = 0.48$), unemployment ($r = 0.39$), and low gross state product ($r = 0.52$); it is, instead, not very popular in states with high levels of education ($r = -0.43$) whose residents prefer to engage with articles about research&technology ($r = 0.43$) and avoid those on celebrities ($r = -0.40$). States with high levels of education also tend to be interested in diverse topics (i.e., those states have topical vectors with high Shannon diversity, which are correlated with education with an $r = 0.44$).

Putting All Together. In the previous section, we have found that the gravity model explains 70% of the variability of user engagement. We have now shown that socio-economic variables matter and, as a result, they might well explain

[3] http://www.thewellbeingindex.com
[4] http://www.ucrdatatool.gov
[5] http://www.bls.gov/web/laus/lauhsthl.htm
[6] http://www.usgovernmentspending.com
[7] http://www.census.gov

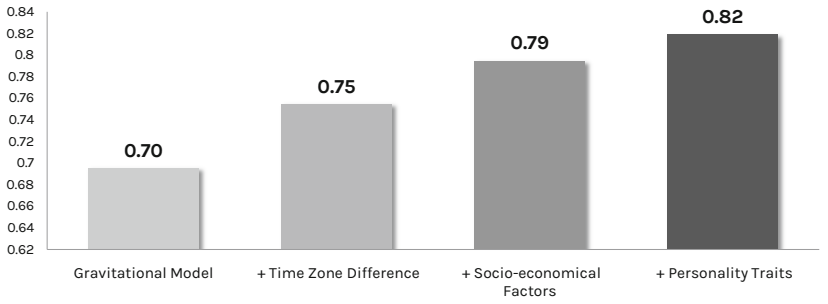

Fig. 4. Adjusted R^2 as predictors are incrementally added to the linear model

Table 2. Linear regression of comments on the same articles from different States. Significance: *** p < 0.0001, ** p < 0.001, * p < 0.01.

Variable	β	**t-value**	**p-value**	**Variable**	β	**t-value**	**p-value**
Gravitational Model	0.694	43.947	***	Bachelor	0.057	5.440	***
Time zone difference	0.855	11.893	***	SAT Scores	0.029	4.574	***
Well-being	1.181	9.220	***	Extraversion	0.002	7.383	***
Crime	-0.031	-1.365		Agreeableness	0.987	0.994	
Unemployment	0.000	0.045		Conscientiousness	-7.299	-6.038	***
GSP	-0.071	-3.734	***	Neuroticism	8.247	7.936	***
High Education	0.000	0.749		Openness	2.226	2.987	**

part of the remaining variability. To test the extent to which that is true, we build a linear regression predicting the number of user comments on the same articles from different states. By having not only the gravity model but also the socio-economic variables as predictors, the percentage of variability explained goes indeed up to 82% (Figure 4), which suggests that the linear model effectively predicts the observed user engagement (Figure 5). Table 2 reports the beta coefficients of the individual predictors in detail.

7 Discussion

Our study suffers from two main limitations. First, we have used the users' IP addresses to localize them. So users on the move might be associated with different IP addresses and consequently with different locations. While it might happen to associate the same user to different cities, we found that it had been extremely rare to associate them to different states. Second, our study does not establish any casual relationship. To that end, one would need to apply our methodology to different snapshots over a long period of time.

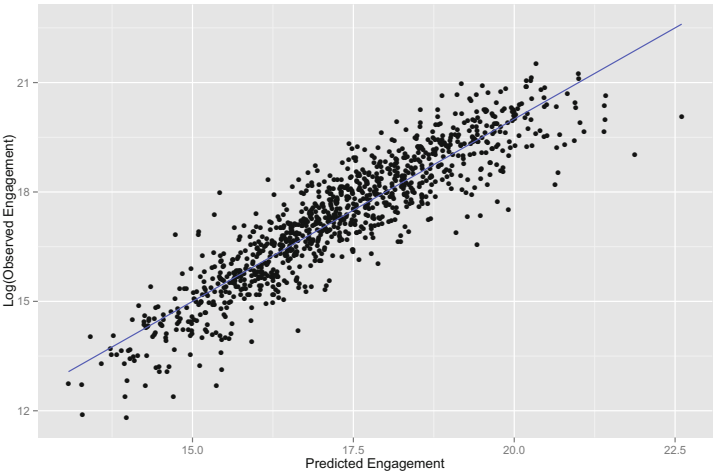

Fig. 5. Observed engagement versus the linear model's predictions

Based on our results, one might well wonder whether like-minded users comment on the same articles, creating fertile ground for group polarization [13]: as a by-product of commenting together (i.e., of engaging with each other), those like-minded users, the theory goes, might develop views that are more extreme than their initial inclinations [1]. For the future, it might be beneficial to explore how geo-temporal patterns of news engagement impact a country's opinion formation.

References

1. An, J., Quercia, D., Crowcroft, J.: Partisan Sharing: Facebook Evidence and Societal Consequences. In: ACM Conference on Online Social Networks (COSN) (2014)
2. Anderson, C., John, O.P., Keltner, D., Kring, A.M.: Who attains social status? effects of personality and physical attractiveness in social groups. Journal of personality and social psychology 81(1), 116 (2001)
3. Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J.J., Vespignani, A.: Multiscale mobility networks and the spatial spreading of infectious diseases. Proceedings of the National Academy of Sciences 106(51), 21484–21489 (2009)
4. Barrat, A., Barthélemy, M., Pastor-Satorras, R., Vespignani, A.: The architecture of complex weighted networks.. Proceedings of the National Academy of Sciences of the United States of America 101(11), 3747–3752 (2004)
5. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008(10), 10008 (2008)
6. Costa, P.T., McCrae, R.R.: The revised neo personality inventory (neo-pi-r). The SAGE handbook of personality theory and assessment 2, 179–198 (2008)
7. Cox, A.M., Clough, P.D., Marlow, J.: Flickr: a first look at user behaviour in the context of photography as serious leisure. Information Research 13(1), 5 (2008)

8. Diakopoulos, N., Naaman, M.: Topicality, time, and sentiment in online news comments. In: CHI 2011 Extended Abstracts on Human Factors in Computing Systems, pp. 1405–1410. ACM (2011)
9. Fortunato, S.: Community detection in graphs. Physics Reports 486(3), 75–174 (2010)
10. Goldberg, L.R., Johnson, J.A., Eber, H.W., Hogan, R., Ashton, M.C., Cloninger, C.R., Gough, H.G.: The international personality item pool and the future of public-domain personality measures. Journal of Research in Personality 40(1), 84–96 (2006)
11. Golder, S.A., Macy, M.W.: Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. Science 333(6051), 1878–1881 (2011)
12. Huberman, B.A., Romero, D.M., Wu, F.: Social networks that matter: Twitter under the microscope. arXiv preprint arXiv:0812.1045 (2008)
13. Isenberg, D.J.: Group polarization: A critical review and meta-analysis. Journal of Personality and Social Psychology 50(6), 1141 (1986)
14. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: An analysis of a microblogging community. In: Zhang, H., Spiliopoulou, M., Mobasher, B., Giles, C.L., McCallum, A., Nasraoui, O., Srivastava, J., Yen, J. (eds.) WebKDD 2007. LNCS, vol. 5439, pp. 118–138. Springer, Heidelberg (2009)
15. Jensen-Campbell, L.A., Graziano, W.G.: Agreeableness as a moderator of interpersonal conflict. Journal of personality 69(2), 323–362 (2001)
16. Jones, J., Altadonna, N.: We don't need no stinkin'badges: examining the social role of badges in the huffington post. In: Conference on Computer Supported Cooperative Work, pp. 249–252 (2012)
17. Jong, G.D., Sonderen, E.V., Emmelkamp, P.: A comprehensive model of stress: the roles of experience stress and Neuroticism in explaining the stress- distress relationship. Psychotherapy and Psychosomatics 68 (1999)
18. Jung, W., Wang, F.: Gravity model in the Korean highway. EPL (Europhysics Letters) 81 (2008)
19. Kaluza, P., Kölzsch, A., Gastner, M.T., Blasius, B.: The complex network of global cargo ship movements.. Journal of the Royal Society, Interface the Royal Society 7(48), 1093–103 (2010)
20. Kamath, K.Y., Caverlee, J., Lee, K., Cheng, Z.: Spatio-temporal dynamics of online memes: a study of geo-tagged tweets. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 667–678. International World Wide Web Conferences Steering Committee (2013)
21. Karney, B.R., Bradbury, T.N.: The longitudinal course of marital quality and stability: A review of theory, methods, and research. Psychological bulletin 118(1), 3 (1995)
22. Krings, G., Calabrese, F., Ratti, C., Blondel, V.D.: Urban gravity: a model for inter-city telecommunication flows. Journal of Statistical Mechanics: Theory and Experiment 2009(07), L07003 (2009)
23. Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or a news media? In: Proceedings of the 19th international conference on World wide web, pp. 591–600. ACM (2010)
24. Lindqvist, J., Cranshaw, J., Wiese, J., Hong, J., Zimmerman, J.: I'm the mayor of my house: examining why people use foursquare-a social-driven location sharing application. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2409–2418. ACM (2011)

25. Mislove, A., Lehmann, S., Ahn, Y.Y., Onnela, J.P., Rosenquist, N.: Pulse of the Nation: U.S. Mood Throughout the Day inferred from Twitter (2010), www.ccs.neu.edu/home/amislove/twittermood/
26. Noulas, A., Scellato, S., Mascolo, C., Pontil, M.: An empirical study of geographic user activity patterns in foursquare. ICWSM 11, 70–573 (2011)
27. Nov, O., Naaman, M., Ye, C.: Analysis of participation in an online photo-sharing community: A multidimensional perspective. Journal of the American Society for Information Science and Technology 61(3), 555–566 (2010)
28. Quercia, D.: Don't worry, be happy: The geography of happiness on facebook. In: Proceedings of the 5th Annual ACM Web Science Conference, pp. 316–325. ACM (2013)
29. Rentfrow, P.J., Gosling, S.D., Potter, J.: A theory of the emergence, persistence, and expression of geographic variation in psychological characteristics. Perspectives on Psychological Science 3(5), 339–369 (2008)
30. Sheskin, D.J.: Handbook of Parametric and Nonparametric Statistical Procedures, 4th edn. Chapman and Hall (2007)
31. Simini, F., González, M.C., Maritan, A., Barabási, A.L.: A universal model for mobility and migration patterns. Nature, 8–12 (2012)
32. Swickert, R.J., Rosentreter, C.J., Hittner, J.B., Mushrush, J.E.: Extraversion, social support processes, and stress. Personality and Individual Differences 32(5), 877–891 (2002)
33. Viboud, C., Bjornstad, O.N., Smith, D.L., Simonsen, L., Miller, M.A., Grenfell, B.T.: Synchrony, waves, and spatial hierarchies in the spread of influenza. Science 312(5772), 447–451 (2006)
34. Zhao, D., Rosson, M.B.: How and why people twitter: the role that micro-blogging plays in informal communication at work. In: Proceedings of the ACM 2009 International Conference on Supporting Group Work, pp. 243–252 (2009)
35. Zipf, G.K.: The P 1 P 2/D hypothesis: On the intercity movement of persons. American sociological review 11(6), 677–686 (1946)