ROYAL SOCIETY OPEN SCIENCE

rsos.royalsocietypublishing.org

Brief report

Cite this article: Botta F, Moat HS, Preis T. 2015 Quantifying crowd size with mobile phone and *Twitter* data. *R. Soc. open sci.* **2**: 150162.
http://dx.doi.org/10.1098/rsos.150162

Received: 21 April 2015
Accepted: 1 May 2015

Subject Category:
Biology (whole organism)

Subject Areas:
behaviour/environmental science/complexity

Keywords:
data science, computational social science, complex systems

Author for correspondence:
Federico Botta
e-mail: f.botta@warwick.ac.uk

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsos.150162 or via http://rsos.royalsocietypublishing.org.

THE ROYAL SOCIETY PUBLISHING

Quantifying crowd size with mobile phone and *Twitter* data

Federico Botta[1,2], Helen Susannah Moat[2] and Tobias Preis[2]

[1]Centre for Complexity Science, and [2]Data Science Lab, Behavioural Science, Warwick Business School, University of Warwick, Coventry CV4 7AL, UK

1. Summary

Being able to infer the number of people in a specific area is of extreme importance for the avoidance of crowd disasters and to facilitate emergency evacuations. Here, using a football stadium and an airport as case studies, we present evidence of a strong relationship between the number of people in restricted areas and activity recorded by mobile phone providers and the online service *Twitter*. Our findings suggest that data generated through our interactions with mobile phone networks and the Internet may allow us to gain valuable measurements of the current state of society.

2. Introduction

The ability to quickly and accurately estimate the size of a crowd is crucial in facilitating emergency evacuations and avoiding crowd disasters [1]. However, existing approaches which rely on human analysts counting samples of the crowd can be time-consuming or costly [2]. Similarly, image-processing solutions require image data to be available in which members of the crowd can be identified and counted by an algorithm [3,4]. Recent studies have provided evidence that data generated through interactions with the Internet [5–20], mobile phone networks [21–25] and other large technological systems can offer new insights into human behaviour [26–31]. Here, we investigate whether data on mobile phone usage and usage of the online social media service *Twitter* can be used to estimate the number of people in a specific area at a given time. We consider data resulting from ordinary use of smartphones, without the need for users to install specific applications on their mobile phone.

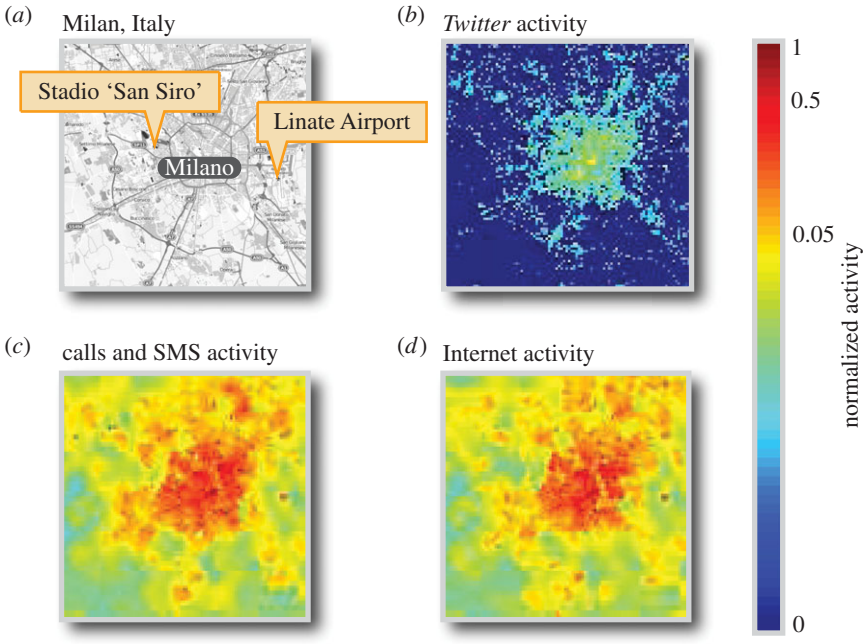

Figure 1. *Twitter*, calls and SMS, and Internet activity in Milan. (*a*) We analyse *Twitter*, calls and SMS, and Internet activity data recorded from mobile phones in the city of Milan and surroundings. The geographical area around Milan for which all these datasets are available is represented in this map, created using data from OpenStreetMap. The datasets cover the period from 1 November 2013 to 31 December 2013. We aim to determine whether such mobile phone data can be used to infer the number of people in a specific location at a specific time. To calibrate our model, we consider two case studies: San Siro football stadium and Linate Airport. (*b*) We depict the normalized number of tweets recorded during the first week of November 2013, for the geographical area shown in (*a*). Tweet counts are extracted from the full set of geolocalized tweets sent during this period. We observe a higher density of tweets in the centre of Milan. (*c*) We depict normalized data on the total number of calls made and received as well as text messages (SMS) sent and received during the time interval between 08.20 and 08.30 of 1 November 2013, for the geographical area depicted in (*a*). We again observe more activity in the centre of Milan. (*d*) We depict normalized data on the number of requests made by mobile phones to access the Internet during the time interval between 08.20 and 08.30 of 1 November 2013, for the geographical area shown in (*a*). Visual inspection of this dataset provides further evidence that more mobile phone activity is recorded in locations where greater numbers of people would be expected. Colours in (*b–d*) are normalized to the maximum recorded activity level in each dataset.

3. Data

We retrieve data on mobile phone and *Twitter* activity recorded in the city of Milan and surroundings in a period covering two months from 1 November 2013 to 31 December 2013 [32]. Both datasets describe activity in the geographical area depicted in figure 1*a*. The *Twitter* dataset consists of all messages sent via *Twitter* ('tweets'), with associated geographical coordinates located within the area shown in figure 1*a*. Tweets are also timestamped. Initial visual inspection of the *Twitter* data shows that greater numbers of tweets are recorded in the centre of Milan, where we would expect greater numbers of people to be found (figure 1*b*).

The mobile phone activity dataset describes the volume of calls made and received, SMSs sent and received and Internet connections opened, closed and maintained. Mobile phone activity measurements are provided at 10 min granularity, for cells in a discrete grid superimposed on the area of Milan. This grid has 10 000 cells of size 235 × 235 m. Further details of this dataset are provided in the electronic supplementary material. Visual inspection of the distribution of call and SMS activity (figure 1*c*) and Internet connection activity (figure 1*d*) again confirms mobile phone activity is highest in the city centre of Milan.

4. Results

We investigate whether the information present in these datasets can be used to infer the number of people in specific areas of Milan at a given time. To calibrate our model, we consider two case studies

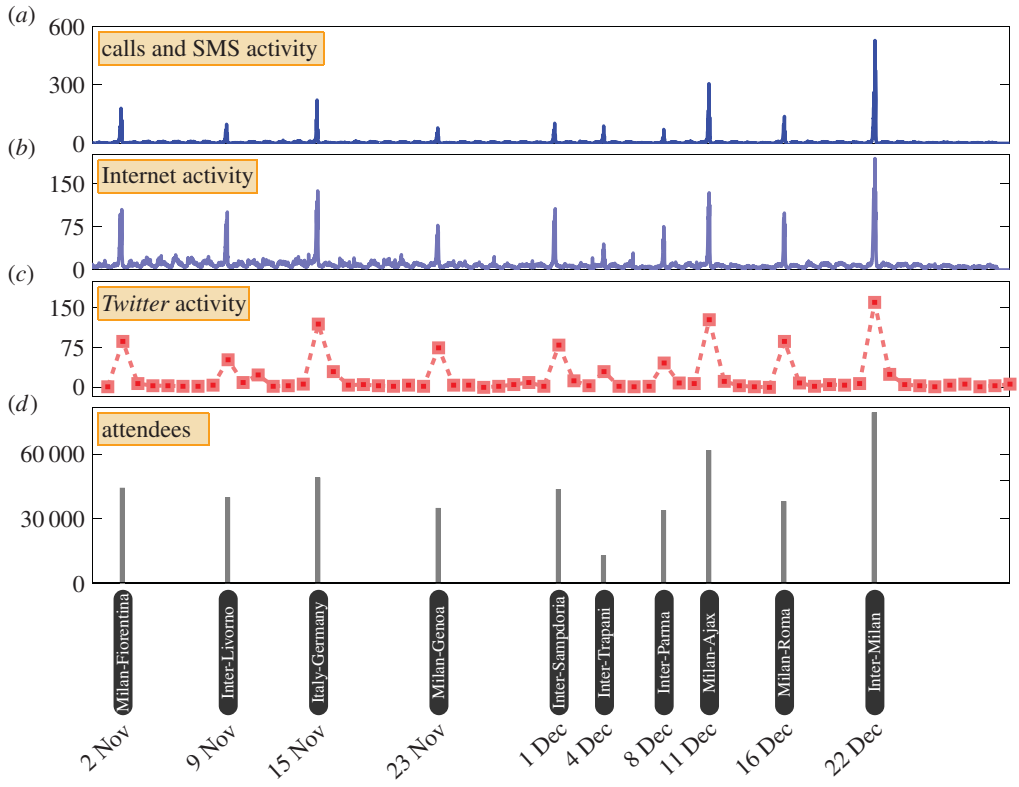

Figure 2. Mobile phone and *Twitter* activity in football stadium Stadio San Siro. (*a*) We depict the time series of mobile phone call and SMS activity recorded in the cell in which the football stadium is located, during the period of analysis between 1 November 2013 and 31 December 2013. The time series is plotted at 10 min granularity. (*b*) Similarly, we depict the time series of Internet connection activity in the cell in which Stadio San Siro is located, at 10 min granularity. (*c*) Finally, we depict the daily counts of tweets recorded within the vicinity of the stadium. (*d*) We determine the dates of football matches which took place during this period, and plot the number of attendees which were recorded at each of these matches. Visual inspection reveals a remarkable alignment between the spikes that can be observed in the communication activities and the dates on which football matches took place. The heights of the spikes bear a strong similarity to the number of attendees at each match.

of access restricted areas for which relevant data exist: San Siro football stadium, for which we have attendance counts for 10 football matches which took place during the period of analysis, and Linate Airport, for which we use flight schedule data to create a proxy indicator for the number of people present in the airport at any given time.

We examine the time series of call and SMS activity (figure 2*a*), Internet activity (figure 2*b*) and *Twitter* activity (figure 2*c*) recorded in the vicinity of the football stadium Stadio San Siro during the period of analysis between 1 November 2013 and 31 December 2013. The coordinates of the area for which data were analysed are given in the electronic supplementary material, tables S3 and S4. In all three time series, we observe 10 distinct spikes, which occur at the same times across all time series. We find that the dates on which these spikes occur coincide exactly with the dates on which the 10 football matches took place in the stadium during this period (figure 2*d*). Furthermore, we note that the relative sizes of the spikes in the mobile phone and *Twitter* activity time series (figure 2*a–c*) bear a strong similarity to the relative sizes of the attendance counts for these matches, as depicted in figure 2*d*.

We extract the maximum values of the spikes in calls and SMS activity, Internet activity and *Twitter* activity. We observe a linear relationship between the number of people attending the football matches and the volume of incoming and outgoing phone calls and SMS messages (adjusted $R^2 = 0.771$, $N = 10$, $p < 0.001$, ordinary least-squares regression; figure 3*a*). We find similar relationships between the number of attendees and both Internet activity (adjusted $R^2 = 0.937$, $N = 10$, $p < 0.001$, ordinary least-squares regression; figure 3*b*) and *Twitter* activity (adjusted $R^2 = 0.855$, $N = 10$, $p < 0.001$, ordinary least-squares regression; figure 3*c*). While figure 3*a–c* suggest a strongly linear relationship between mobile phone activity data and the number of attendees, we note that this relationship holds in a non-parametric analysis too (calls and SMS activity: Spearman's $\rho = 0.927$, $N = 10$, $p < 0.001$;

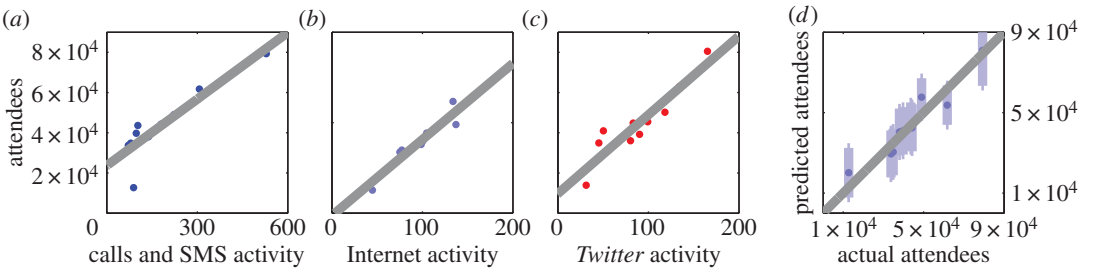

Figure 3. Comparing football match attendance figures to mobile phone and *Twitter* activity. (*a*) We investigate whether there is a relationship between the number of people attending each football match and the recorded mobile phone call and SMS activity inside the stadium. We find a linear relationship between these two variables (adjusted $R^2 = 0.771$, $N = 10$, $p < 0.001$, ordinary least-squares regression). (*b*) Similarly, we find a pattern consistent with a linear relationship between Internet connection activity in the stadium and the number of attendees at each match (adjusted $R^2 = 0.937$, $N = 10$, $p < 0.001$, ordinary least-squares regression). (*c*) We also observe a linear relationship between *Twitter* activity in the stadium and the number of match attendees (adjusted $R^2 = 0.855$, $N = 10$, $p < 0.001$, ordinary least-squares regression). (*d*) We explore whether this relationship could be exploited to infer the number of attendees from communication data if no other measurements were available. Using data on Internet activity, we build a linear regression model using only nine out of the 10 football matches and then predict the attendance at the 10th match. We then repeat this leaving a different match out every time. Here, we plot the resulting estimates and their 95% prediction intervals. We find that the actual number of attendees falls within the 95% prediction interval for all 10 matches.

Internet activity: Spearman's $\rho = 0.976$, $N = 10$, $p < 0.001$; *Twitter* activity: Spearman's $\rho = 0.924$, $N = 10$, $p < 0.001$).

We investigate the possibility of using the information present in communication data to infer the number of attendees in situations where no other measurements are easily accessible. As an example, we consider data on Internet activity, for which the relationship with the number of recorded attendees was strongest. We carry out a *leave-one-out cross-validation* analysis as follows: for each of the 10 attendance figures, we build a linear regression model based on the remaining nine attendance figures and the corresponding Internet activity data. We then use this model to generate an estimate of the attendance figure which was removed from the recorded Internet activity data. In figure 3*d*, we plot the resulting estimates and their 95% prediction intervals. We find that the actual attendance figure is always within the 95% prediction interval of our estimate.

We note that our analysis of mobile phone activity data may be affected by capacity constraints, such as signal truncation, on mobile phone communication in the stadium. Should data on such constraints become available in the future, the influence of these constraints on the relationship between communication data and crowd size may merit further analysis.

We perform a parallel analysis of the relationship between mobile phone and *Twitter* data and the number of passengers at Linate Airport. To estimate the number of people in Linate Airport at any given hour during the analysis period, we assume that passengers may arrive at the airport up to 2 h before a departing flight, and depart within an hour following a flight arrival. For each hour, we therefore calculate the number of flights departing in the following 2 h or arriving in the previous hour, and use this as a proxy indicator for the number of passengers in the airport. We base our calculations on one week of flight schedule data from May 2014, as explained in the electronic supplementary material, and assume that weekly flight schedules are relatively constant. Our proxy indicator is therefore calculated for each of the 168 hours in a week. We omit the three initial days and two final days of the analysis period to create a period of exactly eight weeks.

We compare this proxy indicator to the average mobile phone call and SMS activity and to the average Internet activity recorded for each hour in a week, in the cells in which the airport is located, as detailed in the electronic supplementary material, table S5. We find that greater phone call and SMS activity corresponds to a greater estimated number of passengers (adjusted $R^2 = 0.175$, $N = 168$, $p < 0.001$, ordinary least-squares regression; figure 4*a*). Similarly, we find that greater Internet activity relates to a higher estimated number of passengers (adjusted $R^2 = 0.143$, $N = 168$, $p < 0.001$, ordinary least-squares regression; figure 4*b*). The relationships we find are weaker than those found in the previous case study, but remarkable given the coarse nature of our estimate of the number of passengers.

We analyse *Twitter* activity in the area of the airport detailed in the electronic supplementary material, table S6. In this case, we observe a stronger relationship between the estimated number of passengers and

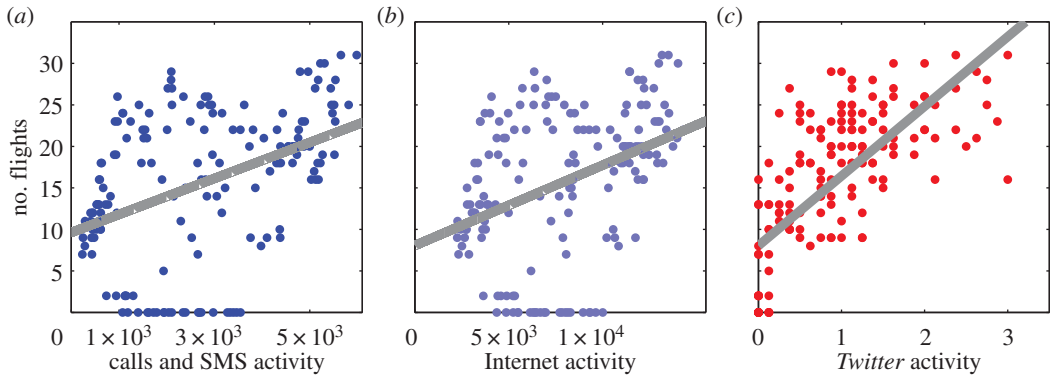

Figure 4. Parallel analysis of the relationship between mobile phone and *Twitter* data and the number of passengers at Linate Airport. (*a*) We create a proxy indicator for the number of passengers at Linate Airport in each hour by calculating the number of flights departing in the following 2 h or arriving in the previous hour. We compare this proxy indicator to the average mobile phone call and SMS activity recorded for each hour in a week, in the cells in which the airport is located. We find that greater activity corresponds to a greater estimated number of passengers (adjusted $R^2 = 0.175$, $N = 168$, $p < 0.001$, ordinary least-squares regression). The relationship we find is weaker than that found for the football attendance figures, but remarkable given the coarse nature of our estimate of the number of passengers. (*b*) We then explore the relationship between the proxy indicator of the number of passengers and Internet connection activity recorded in the cells in which the airport is located. Again, we find that greater Internet activity corresponds to a higher number of passengers (adjusted $R^2 = 0.143$, $N = 168$, $p < 0.001$, ordinary least-squares regression). (*c*) As a final example, we consider *Twitter* activity recorded in the cells in which the airport is located. Again, we consider the average number of tweets recorded during each of the 168 hours in a week, over the eight week period of our analysis. Here, we find a stronger relationship between the estimated number of passengers and activity on *Twitter* (adjusted $R^2 = 0.510$, $N = 168$, $p < 0.001$, ordinary least-squares regression).

activity on *Twitter* (adjusted $R^2 = 0.510$, $N = 168$, $p < 0.001$, ordinary least-squares regression; figure 4*c*). We observe that mobile phone, SMS and Internet activity is still recorded when no flights take place, generally during night-time periods. By contrast, few tweets are logged at these times, potentially explaining the greater strength of this relationship.

We note that roughly 58% of the passengers travelling to and from Linate Airport are Italian [33]. Given the current costs of using mobile phone networks abroad, the mobile phone activity analysed here may reflect the behaviour of Italian passengers more strongly than the behaviour of international passengers.

5. Conclusion

Our results provide evidence that accurate estimates of the number of people in a given location at a given time can be extrapolated from mobile phone or *Twitter* data, without requiring users to install further applications on their smartphones. As well as being of clear practical value for a range of business and policy stakeholders, our findings suggest that data generated through our interactions with mobile phone networks and the Internet may allow us to gain valuable measurements of the current state of society.

Data accessibility. Datasets used in this study are available via the Dryad Repository (doi:10.5061/dryad.1rk60).

Authors' contributions. F.B., H.S.M. and T.P. performed analyses, discussed the results and contributed to the text of the manuscript.

Competing interests. We declare we have no competing interests.

Funding. F.B. acknowledges the support of UK EPSRC EP/E501311/1. H.S.M. and T.P. acknowledge the support of the Research Councils UK grant EP/K039830/1.

References

1. Helbing D, Farkas I, Vicsek T. 2000 Simulating dynamical features of escape panic. *Nature* **407**, 487–490. (doi:10.1038/35035023)
2. Yip PSF *et al.* 2010 Estimation of the number of people in a demonstration. *Aust. New Zeal. J. Stat.* **52**, 17–26. (doi:10.1111/j.1467-842X.2009.00562.x)
3. Chan AB, Liang ZSJ, Vasconcelos N. 2008 Privacy preserving crowd monitoring: counting people without people models or tracking. *IEEE Conf. on Computer Vision and Pattern Recognition*. (doi:10.1109/CVPR.2008.4587569)
4. Kong D, Gray D, Tao H. 2006 A viewpoint invariant approach for crowd counting. *18th Int. Conf. on Pattern Recognition*, vol. 3, pp. 1187–1190. (doi:10.1109/ICPR.2006.197)
5. Choi H, Varian H. 2012 Predicting the present with Google Trends. *Econ. Rec.*

88, 2–9. (doi:10.1111/j.1475-4932.2012.00809.x)

6. Goel S, Hofman JM, Lahaie S, Pennock DM, Watts DJ. 2010 Predicting consumer behavior with web search. *Proc. Natl Acad. Sci. USA* **107**, 17 486–17 490. (doi:10.1073/pnas.1005962107)
7. Mocanu D, Baronchelli A, Gonçalves B, Perra N, Vespignani A. 2013 The twitter of babel: mapping world languages through microblogging platforms. *PLoS ONE* **8**, e61981. (doi:10.1371/journal.pone.0061981)
8. Moat HS, Preis T, Olivola CY, Liu C, Chater N. 2014 Using big data to predict collective behavior in the real world. *Behav. Brain Sci.* **37**, 92–93. (doi:10.1017/S0140525X13001817)
9. Preis T, Reith D, Stanley HE. 2010 Complex dynamics of our economic life on different scales: insights from search engine query data. *Phil. Trans. R. Soc. A* **368**, 5707–5719. (doi:10.1098/rsta.2010.0284)
10. Curme C, Preis T, Stanley HE, Moat HS. 2014 Quantifying the semantics of search behavior before stock market moves. *Proc. Natl Acad. Sci. USA* **111**, 11 600–11 605. (doi:10.1073/pnas.1324054111)
11. Preis T, Moat HS, Stanley HE. 2013 Quantifying trading behavior in financial markets using Google trends. *Sci. Rep.* **3**, 1684. (doi:10.1038/srep01684)
12. Moat HS, Curme C, Avakian A, Kenett DY, Stanley HE, Preis T. 2013 Quantifying Wikipedia usage before stock market moves. *Sci. Rep.* **3**, 1801. (doi:10.1038/srep01801)
13. Preis T, Moat HS, Stanley HE, Bishop SR. 2012 Quantifying the advantage of looking forward. *Sci. Rep.* **2**, 350. (doi:10.1038/srep00350)
14. Mestyán M, Yasseri T, Kertész J. 2013 Early prediction of movie box office success based on Wikipedia activity big data. *PLoS ONE* **8**, e71226. (doi:10.1371/journal.pone.0071226)
15. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. 2009 Detecting influenza epidemics using search engine query data. *Nature* **457**, 1012–1014. (doi:10.1038/nature07634)
16. Preis T, Moat HS. 2014 Adaptive nowcasting of influenza outbreaks using *Google* searches. *R. Soc. open sci.* **1**, 140095. (doi:10.1098/rsos.140095)
17. Georgiev P, Noulas A, Mascolo C. 2014 The call of the crowd: event participation in location-based social services. *Proc. 8th Int. AAAI Conf. on Weblogs and Social Media, Ann Arbour, MI, USA, June 2014*, pp. 141–150.
18. Sui D, Elwood S, Goodchild M. (eds) 2012 *Crowdsourcing geographic knowledge: volunteered geographic information in theory and practice*. Berlin, Germany: Springer.
19. Wirz M, Franke T, Roggen D, Mitleton-Kelly E, Lukowicz P, Troster G. 2012 Inferring crowd conditions from pedestrians' location traces for real-time crowd monitoring during city-scale mass gatherings. *IEEE 21st Int. Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises* (*WETICE*), 367–372. (doi:10.1109/WETICE.2012.26)
20. Weppner J, Lukowicz P. 2013 Bluetooth based collaborative crowd density estimation with mobile phones. *Proc. 11th Annual IEEE Int. Conf. on Pervasive Computing and Communications* (*Percom 2013*), pp. 193–200. (doi:10.1109/PerCom.2013.6526732)
21. Quercia D, Lathia N, Calabrese F, Di Lorenzo G, Crowcroft J. 2010 Recommending social events from mobile phone location data. *IEEE 10th Int. Conf. on Data Mining*, pp. 971–976. (doi:10.1109/ICDM.2010.152)
22. Gonzalez MC, Hidalgo CA, Barabasi AL. 2008 Understanding individual human mobility patterns. *Nature* **453**, 779–782. (doi:10.1038/nature06958)
23. Song C, Qu Z, Blumm N, Barabasi A. 2010 Limits of predictability in human mobility. *Science* **327**, 1018–1021. (doi:10.1126/science.1177170)
24. Onnela JP, Saramaki J, Hyvonen J, Szabó G, Lazer D, Kaski K, Kertész J, Barabasi AL. 2007 Structure and tie strengths in mobile communication networks. *Proc. Natl Acad. Sci. USA* **104**, 7332–7336. (doi:10.1073/pnas.0610245104)
25. Lathia N, Pejovic V, Rachuri KK, Mascolo C, Musolesi M, Rentfrow PJ. 2013 Smartphones for large-scale behavior change interventions. *IEEE Pervasive Comput*. **12**, 66–73. (doi:10.1109/MPRV.2013.56)
26. Vespignani A. 2009 Predicting the behavior of techno-social systems. *Science* **325**, 425–428. (doi:10.1126/science.1171990)
27. Lazer D *et al.* 2009 Computational social science. *Science* **323**, 721–723. (doi:10.1126/science.1167742)
28. Michel JB *et al.* 2011 Quantitative analysis of culture using millions of digitized books. *Science* **331**, 176–182. (doi:10.1126/science.1199644)
29. Petersen A, Tenenbaum J, Havlin S, Stanley HE. 2012 Statistical laws governing fluctuations in word use from word birth to word death. *Sci. Rep.* **2**, 313. (doi:10.1038/srep00313)
30. King G. 2011 Ensuring the data rich future of the social sciences. *Science* **331**, 719–721. (doi:10.1126/science.1197872)
31. Watts DJ. 2007 Connections: a twenty-first century science. *Nature* **445**, 489. (doi:10.1038/445489a)
32. Botta F, Moat HS, Preis T. 2015 Data from 'Quantifying crowd size with mobile phone and Twitter data.' Dryad Repository. (doi:10.5061/dryad.1rk60)
33. ENAC. 2012 Dati di Traffico. See www.enac.gov.it/repository/ContentManagement/information N1503236759/Dati_ traffico_2012_al28032013.pdf (accessed in January 2015).