Cardiff School of Social Sciences
Working Paper Series

www.cardiff.ac.uk/socsi/research/publications/

Working Paper 153: Social Media Analysis, Twitter and the London Olympics (A Research Note)

Authors: Pete Burnap, William Housley, Jeffrey Morgan, Luke Sloan, Matthew Williams, Nick Avis, Adam Edwards, Omer Rana and Malcolm Williams

ISBN: 978-1-908469-08-3

Social Media Analysis, Twitter and the London Olympics 2012

Pete Burnap, William Housley*, Jeffrey Morgan*, Luke Sloan*, Matthew Williams*, Nick Avis**, Adam Edwards*, Omer Rana** and Malcolm Williams***

* School of Social Science, Glamorgan Building, University of Cardiff, King Edward VII Avenue, Cardiff CF10 3WT, UK.

** School of Computer Science and Informatics, Queens Buildings, University of Cardiff, Newport Road, Cardiff CF24 3AA, UK.

Email: SloanLS@cardiff.ac.uk Tel: +44 (0)29 208 70262

Abstract

During the course of this paper we examine publically available social media data that relates to the London 2012 Olympic Games that has been harvested and analysed using the Cardiff Online Social Media ObServatory (COSMOS). Social media has matured sufficiently in terms of user uptake and incorporation into traditional media platforms and outlets that the recent London Olympics has been described as the first social media games. For example, the BBC used the Twitter stream to incorporate and mobilise audience participation into its Olympic coverage. With this in mind, this paper will explore the analysis of social media data in relation to sporting events and social media use. In doing so we identify the ways in which COSMOS can be used to identify hashtag popularity over a specific time period to identify real world events, in this case 'Super Saturday'. The paper reports on indicative evidence that links real-world sporting events to spikes in real time populations' reaction through self-reported social media updates. In turn, the paper provides an analysis of frequency and sentiment of tweets containing the most popular UK hashtag connected to the London 2012 Olympics over a specified time period. This has consequences for conceptualising the relationship between social actors, events and social media and methodological strategies for understanding the dynamic (locomotive) reactions of populations.

Keywords: new social media (NSM), demographics, COSMOS, Twitter, social media analytics and digital social science

This work is based on the project *Digital Social Research Tools, Tension Indicators and Safer Communities: a Demonstration of the Cardiff Online Social Media ObServatory (COSMOS)* which was funded by the Economic and Social Research Council under the Digital Social Research Demonstrator Programme (Grant Reference: ES/J009903/1).

Introduction

Sporting mega-events have typically been understood in relation to ‘spectacle’ (Horne 2006), the aesthetic of neo-liberalism (Boltanski and Chiappello 2005) and the tourist gaze (Urry and Larsen 2011). The London 2012 Olympiad has also been examined through the lens of sports participation, legacy and economic regeneration (Bonini, Bachert and Baena-Cagnani 2012). In addition, London 2012 was understood as the first ‘social media games’ (Silk 2012) due to the take up of social media by the wider population since 2008 and the partial integration of social media into traditional platforms e.g. BBC Sport, epitomised by the use and promotion of designated Twitter hashtags, blogs and Facebook pages. This paper focuses on the empirical relationship between social media, populations and key London 2012 events. In this sense, the paper interrogates the status of London 2012 as the first social media games and in turn reflects on how grand sporting spectacles are being mediated and responded to via networked populations and society (Castells 1996).

The Cardiff Online Social Media Observatory (COSMOS) is an information collection, archival and analysis software platform that takes advantage of freely available socially significant data from sources such as social networking sites, blogs, micro-blogs, RSS feeds and Open Data repositories (e.g. crime rates, Office for National Statistics datasets etc.). It hosts tools for analysing the harvested datasets (including sentiment analysis, tension analysis, topic detection and anomaly detection) that generate a range of social and behavioural metrics that can be digitally visualised. The wider social scientific research community will be able to engage with COSMOS via an innovative virtual research environment (VRE) when it is launched in mid-2013. Researchers will be able to use COSMOS and its datasets to pose hypothetical “what-if” questions, trying different combinations of social data analysis methods to confirm or refute an informal hypothesis and then stress-testing it further until a coherent and arguable position emerges. As an example of this process, this paper explores the relationship between real world events and social media user reaction. Identifying associations between the two is key to determining if social media data has the potential to augment traditional social scientific research methods. We use twitter as a case example of a social media feed that is open to digital social-scientific inspection and can be understood as a form of ‘digital agora’ (Housley et al. 2013).

COSMOS and social media analysis – a workflow methodology for understanding real world events

Our key concern and question in this paper is whether we identify an empirical link between social media activity and real world events, in this case key moments in the London 2012 Olympics, thus our sampling time frame covers the whole duration of the Olympiad. The COSMOS platform generated a series of measures over that time period including popular hashtags, key word frequencies and sentiment analyses. This enabled the identification of anomalies in the social media stream within the designated timeline, e.g. spikes in sentiment levels and hashtag frequency. In this particular event the hashtag ‘#TeamGB’ was identified as a leading trending topic and was therefore selected as a suitable frame from which to initiate more granular analyses of social media traffic during the time period.

Subsequently, we visualised the daily frequency of the occurrence of #TeamGB using a histogram that clearly demonstrated a spike in traffic that contained this hashtag. Identifying this spike provided us with a good case to examine the relationship between social media traffic and a particular event. Further investigation was conducted through an increase in temporal granularity from days, to hours and ultimately to minutes. One of the advantages of social media as 'naturally occurring data' is that it provides a locomotive real-time lens on responses to events by populations (Edwards et al. 2013) and the volume of data is sufficient for minute by minute analysis. In fact, because tweets are time stamped, it is possible to perform a second by second analysis. Whilst responses separated by minutes or seconds may seem too fine-grained for an understanding of social action and response, this level of granularity remains hitherto underexplored and proves vital to understanding the events analysed in this paper.

The next step involved cross-referencing time stamps of the anomalies with reputable media feeds (e.g. BBC News). In this example, the daily spike in frequencies corresponded with what the BBC went on to label 'Super Saturday' because of the three gold medal wins for Team GB that occurred on that day. In the following section of the paper we report on data that illustrates scoping capacity, matters of temporal granularity and the relationship between human responses via social media and real world events.

Results and Data Analysis: The case of 'Super Saturday' 2012

During the course of this paper we analysed archived tweet streams harvested during July and August 2012 that included 1% of all tweets globally. As stated previously, we then drilled into the archived data by identifying topic-specific hashtags. In the time period we looked at the most frequently occurring hashtag was '#TeamGB', which we adopted as a suitable search parameter for exploring social media traffic during a key Olympic event for the hosting nation, namely 'Super Saturday'. Although this approach represents a retrospective analysis of archived materials, it provided a means of testing the relationship between social media traffic and content in relation to a recognisable event type. This approach also gave us an opportunity to test the COSMOS temporal-scoping functionality to inform future development of these tools for real time analysis.

Fig 1: Frequency of #TeamGB from 01/07/12 to 31/08/12 by Day

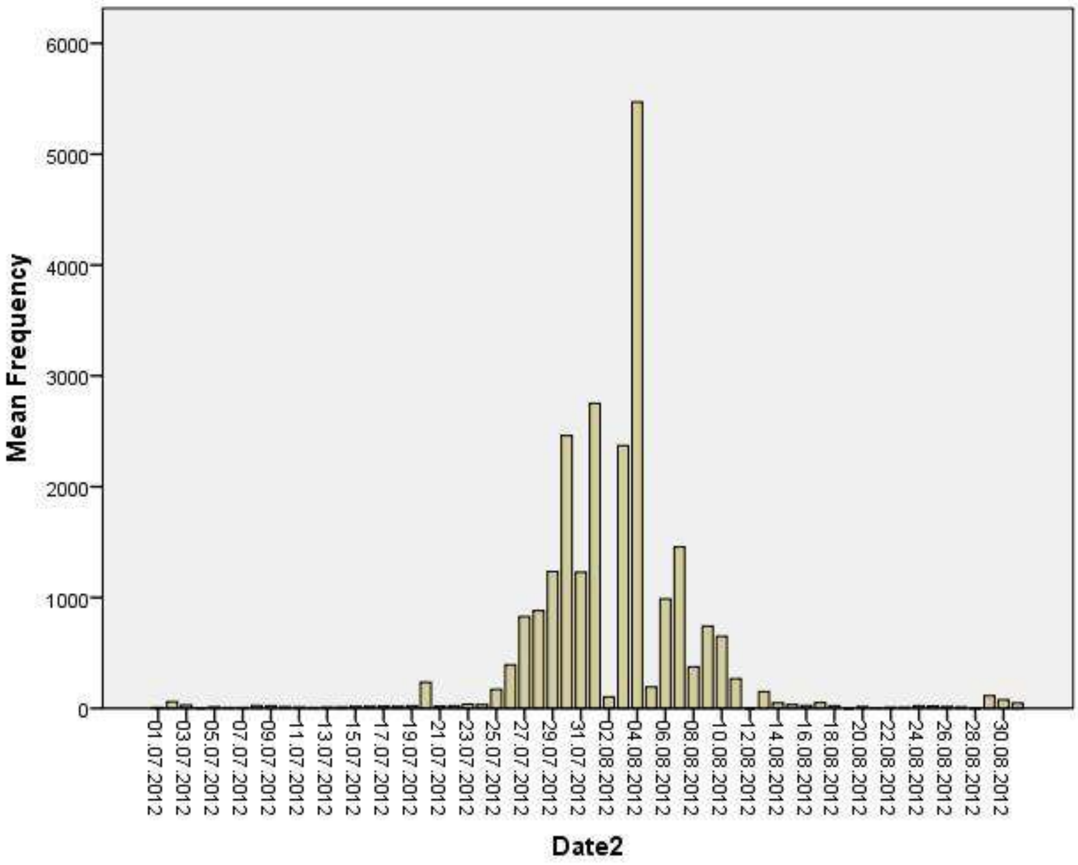

Figure 1 illustrates the frequency with which the hashtag '#TeamGB' appears on a day-by-day basis between 1st July and 31st August. This 62-day period includes both the run up to the Olympics and its aftermath, both of which can be identified on figure 1 by the lack of hashtag mentions either side of the peaks that cover the actual sporting events. No specialist anomaly detection software is required to identify the 4th August as displaying an unusually high number of tweets in which #TeamGB is used and a cursory glance at media sources for that day report that this was 'Super Saturday' on which Team GB won 3 gold medals.

To reiterate, we are interested in whether a disproportionate increase in the frequency of any hashtag is a function of a real world event. More specifically for this case study we know that 'Super Saturday' was characterised by three key events - when Jessica Ennis won gold in the Heptathlon event with a win in the 800m, Greg Rutherford won gold in the long jump, and Mo Farah won gold in the 10,000m. This leads us to refine the research question:

RQ1 = Is an increase in the frequency of the most popular hashtag a function of socially significant moments in time during the event - in this case, the #TeamGB hashtag and GB success at the Olympics?

Demographic differences are an important feature of social science research. As an example we have chosen the gender of the tweeter as an important demographic characteristic in social media traffic (Morgan et al. forthcoming). This leads us to our second research question:

RQ2 = Are there differences in sentiment between men and women in relation to the events above?

One of the advantages of the COSMOS platform is its ability to 'zoom in' on virtual phenomena through disaggregating data into smaller units of time. Figure 2 shows the hourly breakdown of #TeamGB mentions during Super Saturday in addition to the days before and after for comparative purposes. Now we can see that the spike in usage identified in figure 1 is due to high levels of activity on the 4th August between 20.00h and 23.00h, which correspond with the evening athletics session in which Ennis (21:02), Rutherford (21:24) and Farah (21:46) won gold in their relevant events. This provides positive evidence for our first research question (RQ1) as we can see the co-occurrence of socially significant moments in time (Great British success at the Olympics) and increased hashtag usage in relation to the event, suggesting a relationship between the two factors. .

Fig 2: Frequency of #TeamGB from 00.00 on 03/07/12 to 22.00 on 05/08/12 by Hour

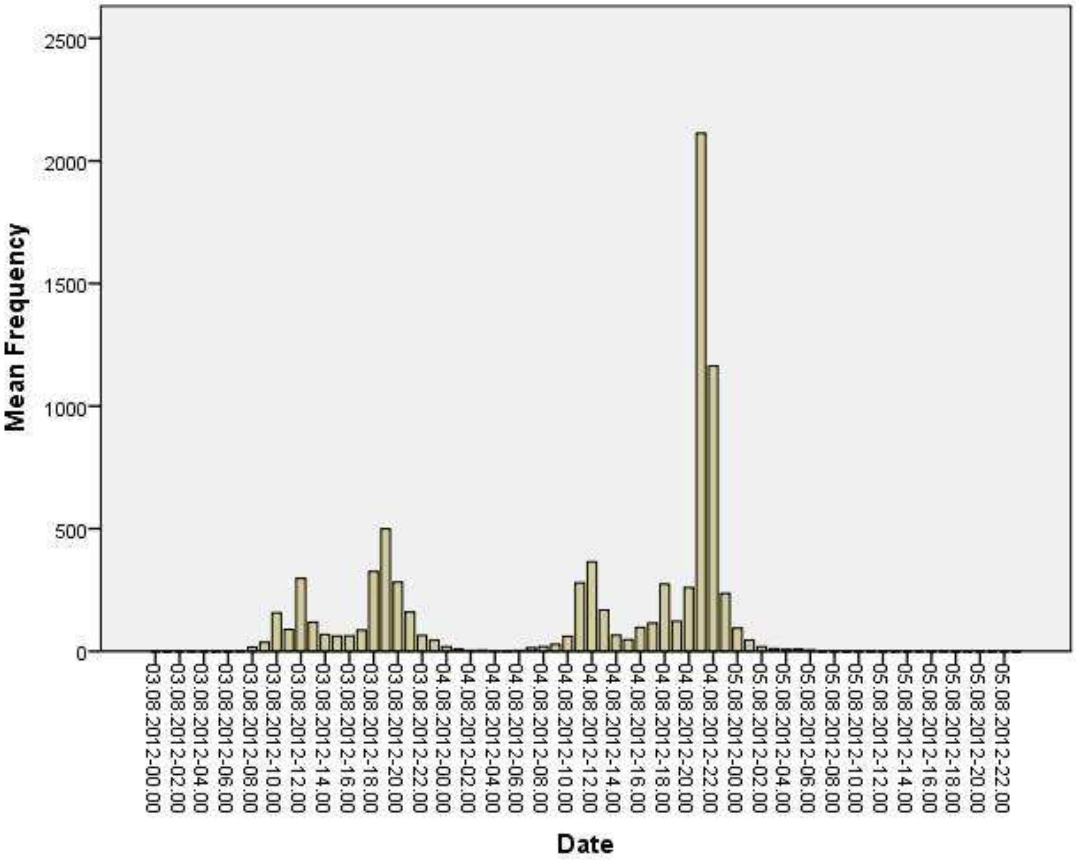

Figure 3 provides even greater granularity through a minute by minute breakdown of #TeamGB frequency between 20.00h and 23.00h. Between this period there are clearly three peaks in activity which correspond directly with events in the athletics stadium:

- The first peak around 21.04 directly corresponds to Jessica Ennis' win in the 800m
- The second peak around 21.24 occurs when Rutherford was declared the winner of the long jump
- The third and final peak around 21.48 is when Mo Farah crossed the finish line and won the 10,000m

Through collecting tweets in real time and analysing the data at such a low level of granularity, we can begin to tentatively suggest causal explanations for the increase in #TeamGB frequency which is directly linked to specific events on the evening of the 4th August, providing evidence in answer to our second question. The locomotive nature of Twitter data is much more suitable for indicating causation on a detailed temporal scale than even punctiform longitudinal survey data (Edwards et al. 2013).

Figure 3 aggregates the sentiment by minute and displays averaged trend data on the sentiment of the tweets containing #TeamGB (Sentistrength 2012). Although the pattern appears erratic (due to low frequency of tweeters having a disproportionate effect on the average sentiment), we can note the substantial decrease in negative sentiment (below the 'Neutral Sentiment' line) that begins just before Ennis' 800m win and continues until around 23.00h. This suggests that the success of Team GB in the athletics stadium encouraged more positive Twitter traffic.

Fig 3: Frequency of #TeamGB from 20.00 to 23.00 on 04/08/12 by Minute and Overall Sentiment

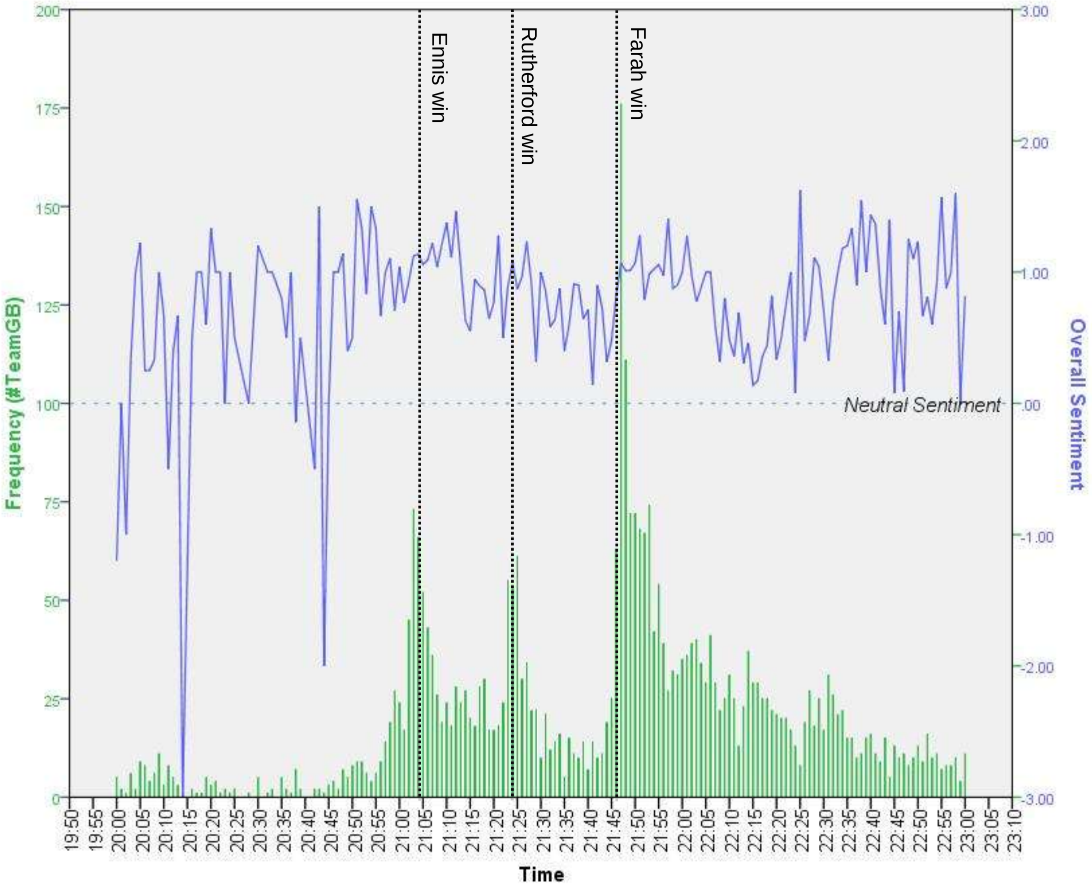

Fig 4: Male/female sentiment of tweets containing #TeamGB between 20.00 to 23.00 on 04/08/12

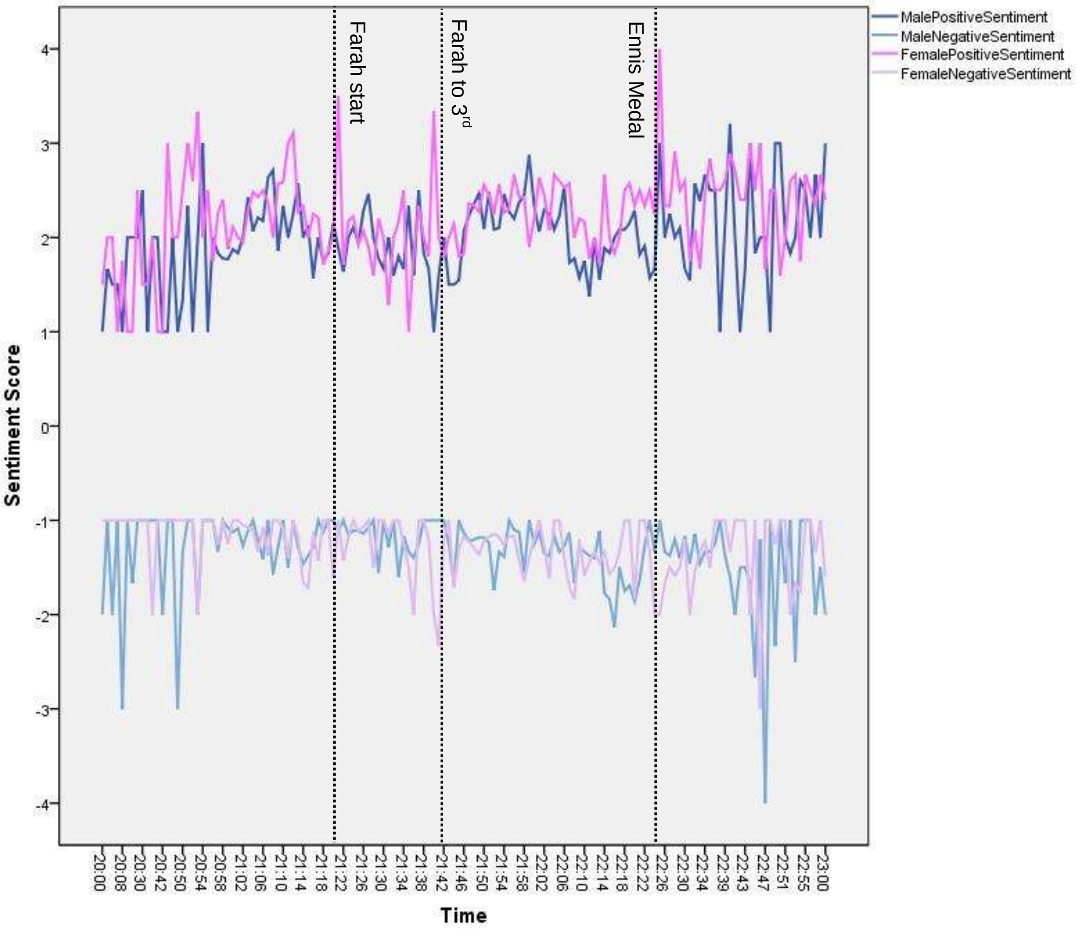

Figure 4 illustrates the differences in sentiment between male and females tweeters over the same time period. Sentiment analysis gives each tweet positive and negative scores based on emotional content (Burnap et al. 2013). The figure above charts the scores separately by gender, which allows scrutiny of a key variable of interest. We have presented the positive and negative sentiment aggregated by minute separately to avoid regression to the mean. For example, a tweet might be given both a negative score of -4 and a positive score of +4 that would result in an average score of 0, thus occluding the natural variance and therefore extremes in opinion. Separating positive and negative sentiment is therefore a more fruitful approach for observing differences between groups.

Although both male and female positive sentiment increases after Ennis' win it is interesting to note that the major positive peaks are attributed to female tweeters whilst the major negative dips are primarily a result of male tweeters. Closer inspection of figure 4 reveals that the positive peaks are not actually correlated with wins but actually with key information flows from traditional media sources. For example, the three highest positive peaks on the graphs are:

- 21.20: when Mo Farah started the race
- 21.42: when he moved up from the back of the field to the front ("The crowd reacts to Farah's surge to third place with a thunderous roar" BBC 2012)
- 22.25: when Jessica Ennis is awarded her gold medal (the biggest peak)

This is an interesting finding as it indicates that extremes of sentiment are not simply correlated with winning, but also with mass media reports. Clearly sampled qualitative analysis of the dataset is necessary in order to provide an empirically grounded interpretive frame for reading this form of quantitative data and visualisation. COSMOS provides access to the qualitative content of social media data for interpretative inspection, affording the analyst with the functionality to further scope into the dataset to interrogate patterns of interest, however this stage of the analysis is beyond the scope of this paper.

Conclusions

The event detection process and consequent scoping through greater temporal granularity using social media as a data source can provide the basis for further mixed methods work on the relationship between key variables of interest (e.g. gender) and the responses of online populations to key events. In the context of this paper we have presented empirical examples that shed light on some of the characteristics of the first Olympic games to be characterised by the ubiquitous use of social media. Clearly this provides the grounds for further development and replication of this type of analysis for future sporting mega events. However, it is also clear that the relationship between key variables of interest or demographic groups in relation to events via naturally mediated behaviour is of significant value to the emerging field of digital social science generally. In this sense it provides the means of beginning to treat social media data (and its analysis) as a social scientific measure of the pulse of the world.

The future development of this type of work via COSMOS will involve the exploration of causal links between social media and real world events that might include civil disorder, health scares, natural disasters and neighbourhood tension (Williams et al. 2013). The locomotive nature of social media data (i.e. greater temporal granularity) represents a fecund opportunity for empirical digital social science that departs from the constraints of traditional cross-sectional and longitudinal designs that are punctiform in nature (Edwards et al. 2013).

Key to the use of social media data in the social sciences will be the mobilisation of a range of digital tools spanning the methodological divide between qualitative and quantitative analysis. This will provide an opportunity for framing and interpreting social media as big and broad data and exploring its relationship with forms of official curated data in order to open up new pathways of analysis. In this sense the analysis of social media data may provide a useful resource in dealing with the challenges and opportunities presented by digital society in the 21st Century.

References

BBC Olympics Coverage http://www.bbc.co.uk/sport/0/olympics/2012/ <accessed 12/12>

Boltanski, L. and Chiapello, E. (2005) The New Spirit of Capitalism, London: Verso

Bonini, M., Bachert and Baena-Cagnani (2012) What we should learn from the London Olympics, *Current Opinion in Allergy and Clinical Immunology*

Burnap, P., Rana, O., Avis, N., Williams, M., Housley, W., Edwards, A. (2013) Detecting Tension in Online Communities with Computational Twitter Analysis, *Technological Forecasting and Social Change*, vol. to be assigned

Castells, M. (1996) The rise of the network society, Oxford: Blackwell Publishers

Edwards, A., Housley, W., Sloan, L., Williams, M.L. and Williams, M. (2013) 'Digital Social Research and the Sociological Imagination: Surrogacy, Augmentation and Re-orientation', *International Journal of Social Research Methodology*, Computational Social Science: Research Strategies, Design and Methods, Housley, W. Williams, M.L. Williams, M. and Edwards, A. (Eds.) Special Issue, Volume 16:2

Horne, J. (2006) Sport in consumer culture, Basingstoke: Palgrave Macmillan

Housley, W. Williams, M.L. Williams, M. and Edwards, A. (Eds.) (2013) Computational Social Science: Research Strategies, Design and Methods, *International Journal of Social Research Methodology*, Special Issue, Volume 16:2

Morgan, J., Sloan, L., Housley, W., Williams, M., Edwards, A., Burnap, P. and Rana, O. (*under review*) Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter

Urry, J. and Larsen, J. (2011) The tourist gaze 3.0., Sage Publications Limited

Sentistrength http://sentistrength.wlv.ac.uk/ <accessed 12/12>

Silk, M. (2011) Towards a Sociological Analysis of London 2012, *Sociology*, 45, 5 , 733-748

Williams, M.L. Edwards, A., Housley, W., Burnap, P., Rana, O., Avis, N., Morgan, J. and Sloan, L. (2013), 'Policing Cyber-Neighbourhoods: Tension Monitoring and Social Media Networks', *Policing & Society* (special issue, forthcoming)