Reassembling the Data – How to Understand Opportunities using an Interdisciplinary Approach

Introduction

We are leaving digital traces of an increasing number of activities: emails, financial transactions, tweets, weblogs, pictures, videos etc, which are uploaded, stored, visualised, modified and, naturally, analysed. We are not only writing about or capturing what we are doing but also what we are questioning on websites such as Quora s and even our thoughts (the default message on Facebook is "*What's on your mind?*"). Activities, questions, thoughts, all this information about ourselves are produced in a such scale that it is participating to the explosion of the so-called *data deluge, exaflood* or *Big Data.* But this time, it is directly about us, not about transactions we establish with companies or administrations, which marketing researchers have been using for several years. Now it is directly about our own identity and our own existence, without intermediaries barring the computer and the Online Network Site (ONS) used.

And as everyone know, all this raw information, raises interest in lots of fields and sociology is one of them. Having a trace of all these exchanges is a gold mine for all of us, and we should be able to take these opportunities offered by the web and the digital traces that everyone is leaving behind them. All of these recorded traces , can have a fundamental impact on our approach to understanding human society. The co-occurrence of '*"deep data" about few people and "surface data" about lot of people*' (Manovich, 2011) allows us to go further than the idea of sampling or developing qualitative approach limited in its generalisation. We can, from Big Data, draw both a large and a small picture of the world by adapting its resolution to our need.

But before we can build this picture several difficulties remain. Recent debates in sociology point out difficulties and worries about Big Data uses. Several examples include the lack of demographic information, the impossibility of control over the production and the access to it, the ethical issues of analysing personal information even if they are publicly available, and the difficulty in developing skills to analyse them (Bizer, Boncz, Brodie, & Erling, 2012; Boyd & Crawford, 2011; Manovich, 2011). Even with theses difficulties, we have to take this data into account.

According to Savage and Burrows, sociologists were too slow to take advantage of this opportunity, leaving this field to computer scientists and marketing researchers (2007, 2009). For them it is a necessity to use hybrid methods, even those that sociologists are not necessarily used to, such as clustering, in order to keep sociology on the edge of new methods and methodologies. To challenge this issue, the authors point to the fact of the importance of the description in sociology, too much often left in the advantage of the causality relation. This need for descriptions in sociology is echoed in several theories, and among them, the actor-network theory (Savage, 2009). From this perspective, the description is a central concept, more important than the the causality explanation (Law, 2008).

However, we want to make several remarks about the importance of description of digital traces left by ONS users.

Methods such as clustering are useful to find relevant information in Big Data, categorising it reducing complexity of information. As pointed out by Savage, one main finding in biology was through the use clustering techniques to find useful information within Big Data produce by human DNA. But human activities are not DNA mechanisms. The description of a Web phenomenon suffers from the speed of the online activity, everything changes fast and is in constant movement. A really good description of the Web at time *t* will be outdated as soon as it is done, before it can even be published (Thelwall, 2006). Under this pressure, other techniques such as generative processes are used, which can take this dynamic nature into account to develop algorithms and really useful search engines. But what works under computer science development does not necessarily work in academic publications.

Theses questions raise some issues which are not approached by other perspectives. Our reflection is to start from the beginning and understand the structure of the data itself, not with a sociologist's eyes who try to fit this new field to their own approach but from the nature of the data itself.

Data and Social Complexity

Definition of Big Data

The phenomenon of Big Data is not necessarily new, already in the late 80s dealing with the whole US census stored in 100GB, *Big Data* was considered an issue (Jacobs, 2009). Nowadays, it is common to deal with terabytes or even petabytes of data, but the global definition of *Big Data* remains similar; *it's characterised by a huge amount of information where issues such as storing, computing and analysing cannot be answered by traditional dataset tools or traditional statistical software*. Therefore the definition is not a boundary by a typical limit of size but evolves with time, as the technology and the tools to handle it are becoming more efficient. In fact, talking only about size is an incomplete version of *Big Data.* Recently, Stapleton summarizes the *Big Data issues* from a pure computer science perspective and uses the 3Vs: *volume, variety* and *velocity (Stapleton, 2011)*.

- **Volume:** the unprecedented quantity of data leads to thinking about new approaches to deal with it. It is impossible to apply traditional algorithms to treat the data. Even the methods for storing them change e.g. the big social Web-services such as Facebook and Twitter use NoSQL database types allowing more scalable solutions; a more efficient approach toward the exponential growth of information.
- **Variety**: the data take heterogeneous forms and therefore make integration among different sources difficult; meaning, how to fit different sources of data into a database format, which implies rows

and columns.

- **Velocity**: the constant renewal and flow of these data raise difficulties with regards to storing and treating them in real time. The algorithms need to analyse vast quantity of data but without the traditional delay. The importance of the speed of execution covers two different needs. People (especially in business) need to have data as quickly as possible. The second reason, more important for us, is that often the stream of data has a lack of memory. The information which is not processed in real time is lost.

Data and Social Complexity

It is the combination of these three characteristics which creates the complex task of *Big Data* for computer science and what we can call *data complexity* - the difficulty to recall and analyse vast and different amounts of information in real time. When Jacobs is talking about the US census as the Big Data issue, this is only in the context of *volume* and it is not about the complexity of the data itself - its *variety* in regard to the analysis we can apply on it. Census create data for a specific goal, measuring demographic variables. Surveys are similar as they are a way to answer questions by producing data (often inferred information) in order to answer the hypothesis of the research. As the goal is clearly specified and the process of analysis is known in advance, the complexity is, therefore, simplified during the conception of the survey and resolved before the collection/creation of data. The *variety* of inputs can be reduced into categories; similarly the difficulty of geography is solved by creating statistical areas and we can reduce a potentially long and passionate debate about any question into a 5-points Likert scale. With Big Data, it is different; the difficulty is not during the production of data itself, but before the analysis. Data is more ecological as it is not about the inferred comportments but the actual behaviour. However, they also come in a less practical format and several steps are needed before we are able to compute them and extract useful information - *data reduction*. This is because the dataset is usually too large to be analysed so some useless information needs to be removed: *data cleansing* deals with missing and inconsistent data, *data integration,* combines several datasets and finally *data transformation*, is the normalizing of the data for more efficient analysis (Apeh & Gabrys, 2011; Kirmse, Udeshi, Shuma, & Bellver, 2011).

However this is not the only type of complexity *Big Data* raises; for example, beside the *variety* and volume of the data itself, it is also what the data represents, and which behaviour they measure, such as buying items, expressing political interest, websurfing history and social bonding. Every behaviour is recorded and therefore the data represents complex, social and individual actions, showing a varied and complex virtual environment. Data collected on Amazon will differ from a status on Facebook which will also respond to different purposes than a status on Linkedin. Furthermore, the meaning of each behaviour or each bit of information will not necessarily be the same, even within the same kind of sources of production. The *social complexity* represents these different behaviours and therefore present the difficulty of knowing what we are studying and how to study it. Both types of complexity are studied in different disciplines. *Data complexity*

presents a computer science task and s*ocial complexity* finds answers in social-science.

However; when we are looking on studies on ONS behaviour, even if it can represent vast arrays of different social situations; only one complexity seems to be important. Researchers are using social network analysis to study relations between ONS users, and Information Retrieval to categorise information and assign topics to the discussion within the ONS. But even if there is a necessity to use techniques and algorithms that are more adapted to the nature of Big Data, researchers are still predicting the behaviour of people through the ONS as if the volume of information was the main issue to resolve, or the only advantage to take from it. They are doing it successfully; we can find predictions about almost everything, from internal behaviour, such as the success on Twitter (Zaman, Herbrich, Van Gael, & Stern, 2010), on Youtube (Wattenhofer, Wattenhofer, & Zhu, 2012), to external behaviours of the network such as for election results (Tumasjan, Sprenger, Sandner, & Welpe, 2010), the spread of flu with Google research (Ginsberg et al., 2008), financial markets and mood on a national level (Bollen, Mao, & Zeng, 2011) or even earthquakes (Sakaki, Okazaki, & Matsuo, 2010). But to cite Callebaut : "*Science, Woese suggested, is impelled by two main factors: technological advance and 'a guiding vision (overview).' Successful scientific change requires a properly balanced relationship between the two*" (Callebaut, 2012).

What is important here in the development of an interdisciplinary perspective is that constraints imposed by the need of data complexity reduction will impact the comprehension of what the data represent. We will see the limits dictated by the common type of analysis used in Big Data and see how we can cross it with sociological perspective to surpass them.

Consequence of Analysis

1. Correlation versus Interpretation

Twitter is heavily studied within ONS research. A practice on Twitter which is often used to understand users behaviour and predict activity is the retweet. But even if the retweet itself is well understood, no one seem to agree on how retweeting works and by what is influenced. Also, all studies measured an activity, an actual behaviour but this activity has not been clearly defined itself and so we do not know what it represents. Sometime it appears to represent (or be influenced by) homophily (J. Weng, Lim, Jiang, & He, 2010) or transitivity phenomenon (Golder & Yardi, 2010) if it is studied with Social Network Analysis. In other occasions it is used as a variable measuring attention (L. Weng, Flammini, Vespignani, & Menczer, 2012) if the RT is using it as measure to study other phenomenona such as the content of the tweet (Nagarajan, Purohit, & Sheth, 2010; Naveed, Gottron, Kunegis, & Alhadi, 2011). All these studies successfully reduce the data complexity into a manageable form by finding information in millions of tweets and users, and giving good descriptions of a specific social activity (sharing content over the web). They provide a strong correlation with different perspectives and definitions of the retweet, but the problem is more than just having different hypotheses in competition. The lack of link between the activity measured and what it is supposed to represent is equal to displacing the issue from *data complexity* toward *social*

complexity.

With a survey, if we want to measure an intention or a thought, we are building different scales, testing the validity and fidelity of them to know if we are effectively measuring what we are supposed to measure. Here the problem is in the reverse, we have the behaviour and we want to know what they represent. But the transactional perspective is not building to answer such a question, clustering and any Information Retrieval approach fail to fill this gap in understanding. We could argue that this diversity of potential processes the studies highlight, is only different hypothesis in competition. But it is not only several potential latent variables they are measuring, it is different environments within the same world.

2. Fragmented World

We all know that the difference between virtual and physical reality could not be pre-supposed anymore. It is similar with ONS, we cannot pre-suppose their equivalence anymore, as our knowledge about them increase and we know they are creating distinct realms within each other and different ONS have different purposes and a different usage and audience (Hargittai & Hsieh, 2010; Lerman & Ghosh, 2010; Skeels & Grudin, 2009). But within the same ONS huge differences can occur and assuming a homogeneous phenomenon within one ONS is a mistake.

The most obvious is the cultural difference measured by different language uses. A study found a significant difference in the RT/tweets rates between Indonesian (39%), Japanese (7%) and English speakers (13%) (Hong, Convertino, & Chi, 2011). Interestingly, the authors started to recognize the multiple possible roles of the RT; they mention it could be used as information sharing or as social bonding. However, they do not given any explanation as to how to decide which role it is or if it is both but conclude that the Indonesian bond more than Japanese[1].

Also, some studies clustered different characteristics with respect to the type of event. For instance, when an external event is discussed on Twitter, the global activity around the daily peak will differ if an expected event, a long debate/events or a sudden phenomenon occur (Lehmann, Gonçalves, Ramasco, & Cattuto, 2011). Not only will the retweet rate differ between different events, but will also differ with respect to the time of day ("Bitly blog - You just shared a link. How long will people pay attention?," 2011). We know there are differences, but as the retweet will be used to explain this global difference, it is impossible to know the impact of it on anactivity and to know if the tweets are different or not.

3. Global and Individual Perspectives

The problem of the fragmented world is mainly a consequence of the method chosen. All the studies described above about RT (and globally about ONS analysis) deploy methods which understand the data as global and homogeneous and using only global metrics and high scale of understanding. Beside the facts

1 On contrary, confronted to a potential dual role but for hashtag (there it was indicator of content or membership), Yang et al. build metrics to test which role is responsible of the acceptance of an hashtag. They found out that both reason are used and can predict hashtag adoption (Yang, Sun, Zhang, & Mei, 2012).

ONS need a more granularity perspective to take into account all variables, the macro-level of statistics prevents us from understanding an individual and develop a better understanding of the underlying process. It seems that, as there is a vast volume of data, we do not need to understand local phenomenon; we need prediction about huge activities such as financial market or mood on a national level and to know more about what is predicting human activity on world scale, we need to understand RT with a global perspective. In summary, we need an intermediary position, in the middle of the scale.

This goal is shared in a study which showed how it is possible to study local interaction between people on Twitter and how it is possible to differentiate RT behaviour and discussion with a local, micro-level perspective on large scale data. They compare this discussion, usually between two people, with normal tweets and RTs. They found out that the "chat-tweet" type are shorter than normal tweets and RTs (Macskassy, 2012). Another study used global analysis of the network in order to construct several network metrics and associate them to the users (Quercia, Capra, & Crowcroft, 2012). They are measuring three metrics, the ratio between following/followers, the reciprocity (proportion of bi-directional relations) and proportion of triadic relations. Beside these network metrics, they are analysed the content of the tweets with semantic analysis and then built a measure of topic diversity for users, and found out that users who tweet about different topics have greater access to structural holes and have higher network status (rates followers/following).

This perspective seems for us the best example of integrating local and global perspectives beside a network and content analysis but also testing hypothesis from theory, here the *imagined communities* (Anderson, 1991). This research tries to understand behaviour behind the data and which concepts can explain it instead of being only focused on *data complexity*. However, we think we can go further than this tentative approach and think about a way to establish causality relations. To do that we need to stop concentrating on the best algorithms or on the size of the dataset and go further from a sociological perspective.

Our Perspective

We have identified the problems with the computer perspective, with its focus on complexity reduction. These methods offer extensive quantitative descriptions and strong predictions but also often lack a depth of understanding of the behaviours and what they represent, which processes are behind them, and why they are modified under different circumstances. As mentioned before, it is because the complexity reduction is only one side of the Big Data issue. We also need to understand how to reduce social complexity.

First of all, qualitative studies and traditional quantitative surveys exist to understand the impact of ONS on people, from the usual interests of young people, to the factor and risk of disclosure information to sub-cultures within ONS (Nosko, Wood, & Molema, 2010; Robards & Bennett, 2011; Shin & Hall, 2011). But it is not our interest now to see how sociologists are trying to understand these relatively new practices.

Sociology has a history of powerful methods to control the quality of the information, testing hypotheses,

establishing causality and knowing the limits of the generalisation of their conclusions- all aspects that do not seem to be primordial within the current studies on Big Data. We know that data and methods will shape the theory, but we also know that the theory has to be adapted to the field. In any scientific process, the data, the methods, and the theory are intrinsically linked. Therefore, our theoretical background is chosen for its pertinence regarding the possibility offered by the datasets obtained from ONS, such as having a trace of all interactions, but also the limits of possible analysis that can be applied on it. We are going to use the three "Vs" (*volume*, *variety*, *velocity*), but this time under the perspective of a sociologist and not of a computer scientist. We think the sociological perspective reflects on the content whereas the computer science perspective places an emphasis on structure.

1. Variety – Post Demographics

People are posting about their preferences, tastes, profiles, attachments, networks and so on. They are building networks with people who matter to them. This information differs from our standard social science instrument. It is not driven by pre-defined categories – class, gender, race, ethnicity or other supposedly stable traits that are often used in sociology to categorize people, but information about, what people actually say on ONS. It is different from data gathered by surveys or interviews as it is built from the data and not decided in advance which categories makes sense and which do not. This point was made by Rogers who coined the term *post-demographics data* to represent this idea of shift in theoretical paradigm: "*It also marks a theoretical shift from how demographics have been used 'bio-politically' (to govern bodies) to how post-demographics are employed 'info-politically', to steer or recommend certain information to certain people.*" (Rogers, 2009). We see from this explanation, the affiliation to the marketing world, where recommendations and predictions are central and need to be individualized.

Developing a marketing approach to building social-categories seems to be more efficient than staying on the concept of demography and nation states as usually envisaged with regards to information we can obtain. Pointing out the lack of demographic information in Big Data issue is necessary, (Boyd & Crawford, 2011) but we can put this importance into perspective with the current pertinence of social-categories in a global and almost generic space (Barber, 2003; Lash, 2002).

As we live more and more in a global space where we are all inter-connected without national limits, on Twitter and on Facebook, it is not necessarily the physical place which is important, but the topic discussed, the place within the network, the imagined communities mentioned above; based not on myths and traditions, but based on profiles, "*like*" or other *"Here more people you might know*" recommendations constantly asked by any ONS, The identity is atomised and externalised within the groups in which we belong by the technology (Lafontaine, 2003). It is not based on our economical status or a stable environment but as a fluid identity, changing and intense (Wittel, 2001).

And even if the initial place is important, it is possible to re-construct it from what people are saying, and to whom, like the infograph published by Facebook which sketches the continents using their users' interactions

(Butter, 2011). Even more generally, we can reassemble traditional social classes with behaviour on the web by analysing which websites users visit and how often people are browsing them (Sharad Goel, Jake M. Hofman, & M. Irmak Sirer, 2012). These ideas of non-pertinence of traditional social-categories is not saying we need to completely remove them, or declare they are not pertinent; it is replacing them with a more naïve approach based on what people say and do and with whom.

. With information, profiling users we can see where and what people are doing, not if they are in one category or another. The idea here is to use this clustering not as an end, but as a means for dynamic construction of social categories. It is ultimately studying their activity rather than putting them in pre-built boxes, having the advantage of being *data-driven* with flows of digital traces.

2. Velocity – Activity

Focusing on the activity, actual behaviour as an available and relevant information in Social Stream is only possible through a certain theoretical lens which emphasises the dynamic nature of behaviour instead of static states. Theses on theoretical tools already exist to analyse the dynamism of social interactions instead of measuring their stability. Since the emergence of the information age (Castells, 1996) different theories use concepts such as fluidity (Wittel, 2001), liquidity (Bauman, 2000) and mobility (Sheller & Urry, 2006), but all have different outcomes or conception about ICT.

Generally we can see a higher interest of movement than for stability and fixed characteristics. This more recent object in sociology could be for two independent reasons; the societal (Lash, 2002) or methodological (Latour, 2005). We think that combining some aspects of these two perspectives on activity will help us to understand the reason of studying activity.

Activity - *velocity* represents the change in a more rapid and networked society. Everything happens so fast in the informational age, the change from any one state is so dramatically quick that time is compressed. In result, the activity itself and the meaning of it are combined the *sense* and the *practice* are united. Following the idea of Garfinkel, his empiricism phenomenology, and his definition of reflexivity as "*[...] no longer separate but 'incarnate' in activities*" (Lash, 2002, p. 17), we can see it as the *velocity* of ONS. For instance it is almost pointless on Twitter to publish an old tweet, the main characteristic is the instantaneous and the flow of tweets otherwise the information is considered as outdated.

But, as pointed out by Lash, a signal only becomes information when some meaning is attached (Lahs, 2002), otherwise it is only noise. In the chaos of our information age, only meaning is important and it is only realised by the sense-maker and his environment. What is important here is the perspective of a change in society which leads to an epistemological difficulty: the activity is fast and directly contains the information, but we need to understand to whom the activity makes sense and it is why we think that the interest for activity in actor-network theory can complete our understanding of activity.

The actor-network theory expresses the importance of such perspective by stating that only activity and traces left by people can be investigated as a manifestation of social interaction. From this perspective, the

social: "*[…] designated two entirely different phenomena: it's at once a substance, a kind of stuff, and also a movement between non-social elements. In both cases, the social vanishes*" (Latour, 2005). We also have this idea of activity, of movements which need to be studied.

This is the reason why the concept of post-demographics coupled with activity will allow us to have another view on social interaction. It is not only building categories on new sorts of information, it is also building information on what is actually happening.

But these theories pose some limits in order to fully understand activity such as context for actor-network theory (Latour, 2005). Following the importance of the description and context highlighted by this theory, there is virtually no limit to where we can extend or stop the description of any phenomenon, we only need to follow the traces where they lead. Consequently, one limit Latour suggest is the limit of the writing, the size of the book or the article (Latour, 2005). We think that ONS Data with the characteristic of *volume* offers (imposes) alternatives.

3. **Volume – Whole Population**

In order to draw categories based on post-demographic data generated by activity, and given the *volume* of data, we need to apply specific analysis to retrieve relevant information, a social network analysis or to apply a cluster analysis. These methods consider data as a complete set and not aiming, at any point, a statistical representation of a larger population, even if they are de facto a sample of a larger dataset. It is here precisely where the interaction between data complexity (computer science perspective) and social complexity (social science theoretical tools) give us the opportunity to pose further questions that sociological debates raise at the beginning. First, as the datasets are considered as a whole population, we can make a sample within it to test more local phenomenon with the only need to verify if the sample is representative of the dataset, not the distribution of the whole of Twitter or the whole of the ONS users or the whole population. This possibility is because we are turning the limit we found as fragmented world into an advantage. Instead of trying to know the distribution of the whole of the ONS users, which will present problems such power law distribution or limiting our extension of description to the limit of our thesis space, we are deciding that the dataset is the whole population. Therefore, we have the statistical advantage to be able to narrow the distribution of our sample from the whole population, as well as not having the issue about the absence of information about the sampling methods of the API. However, the problem of the possibility to extend our conclusions and comparing it to other studies remains.

Paradoxically here we think, that it is qualitative methodology which could help us exceed the limits of whole dataset on specific population. If we are considering each dataset about any ONS, any period, as a specific, but complete case study, we could maybe understand how comparisons are possible if we understand processes but about different types of populations. It would have to be done without any a-priori and could also answer one issue scientists are facing about the access of ONS and the lack of control over it. One dataset is one case study and there is no reason all conditions will be similar later. We can still do

replications over the same dataset, impossible in the case of traditional qualitative study such as field observations or participative observation.

Conclusion

This article was an attempt to show how an interdisciplinary perspective is essential to deal with Big Data. Not only by using computer science to resolve traditional question in sociology but by redefining the possible questions we can ask with the nature of the data we can obtain. By doing this reflection prior to any work on Social Stream data, we ensure that we avoid questions not necessarily relevant to the field (lack of information) and we have more opportunities to discovers new interesting phenomenon. This is not a solution to issues raised by other debates in sociology about the pertinence of Big Data use in social-science but more an attempt to integrate both views into a single approach and putting them into the perspective of already developed theories which are enthusiastic about the idea of digital traces without necessarily understanding the data from a computer-science perspective.

References

Anderson, B. (1991). Imagined communities: Reflections on the origin and spread of nationalism.

Apeh, E., & Gabrys, B. (2011). Change Mining of Customer Profiles Based on Transactional Data. *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on* (pp. 560–567). IEEE.

Barber, B. R. (2003). *Jihad vs. mcworld*. Random House.

Bauman, Z. (2000). *Liquid modernity*. Cambridge: Polity.

Bitly blog - You just shared a link. How long will people pay attention? (2011, September 6). Retrieved September 7, 2011, from http://blog.bitly.com/post/9887686919/you-just-shared-a-link-how-long-will-people-pay

Bizer, C., Boncz, P., Brodie, M. L., & Erling, O. (2012). The meaningful use of big data: four perspectives--four challenges. *ACM SIGMOD Record*, *40*(4), 56–60.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*.

Boyd, D., & Crawford, K. (2011). Six Provocations for Big Data.

Büscher, M., & Urry, J. (2009). Mobile methods and the empirical. *European Journal of Social Theory*, *12*(1), 99–116.

Butter, P. (2011). Visualizing Friendships. Retrieved January 12, 2011, from http://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919

Callebaut, W. (2012). Scientific perspectivism: A philosopher of science's response to the challenge of big data biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*.

Castells, M. (1996). *Rise of the network society: The information age: economy, society and culture*. Blackwell Publishers, Inc.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2008). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014.

Golder, S. A., & Yardi, S. (2010). Structural Predictors of Tie Formation in Twitter: Transitivity

and Mutuality. *Proceedings of the Second IEEE International Conference on Social Computing. August 20-22, 2010. Minneapolis, MN*.

Hargittai, E., & Hsieh, Y. P. (2010). Predictors and Consequences of Differentiated Practices on Social Network Sites.'. *Information, Communication & Society*, *13*(4), 515–536.

Hong, L., Convertino, G., & Chi, E. H. (2011). Language matters in Twitter: A large scale study. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.

Jacobs, A. (2009). The pathologies of big data. *Communications of the ACM*, *52*(8), 36–44.

Kirmse, A., Udeshi, T., Shuma, J., & Bellver, P. (2011). Extracting Patterns from Location History.

Lafontaine, C. (2003). Nouvelles technologies et subjectivité: Les frontières renversées de l'intimité. *Sociologie et sociétés*, *35*(2), 203–212.

Lash, S. (2002). *Critique of information*. Sage Publications Ltd.

Latour, B. (2005). *Reassembling the social*. Oxford University Press Oxford.

Latour, B. (2011). Networks, Societies, Spheres: Reflections of an Actor-Network Theorist. *International Journal of Communication*, *5*, 796–810.

Law, J. (2008). Actor network theory and material semiotics. *The new Blackwell companion to social theory*, 141–158.

Lehmann, J., Gonçalves, B., Ramasco, J. J., & Cattuto, C. (2011). Dynamical Classes of Collective Attention in Twitter. *Arxiv preprint arXiv:1111.1896*.

Lerman, K., & Ghosh, R. (2010). Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*.

Macskassy, S. (2012). A Systematic Investigation of Blocking Strategies for Real-Time Classification of Social Media Content into Events. *International AAAI Conference on Weblogs and Social Media; Sixth International AAAI Conference on Weblogs and Social Media*.

Manovich, L. (2011). Trending: The Promises and the Challenges of Big Social Data. *Debates in the Digital Humanities, ed MK Gold. The University of Minnesota Press, Minneapolis, MN. [15 July 2011]*.

Nagarajan, M., Purohit, H., & Sheth, A. (2010). A qualitative examination of topical tweet and

retweet practices. *Proceedings of Fourth International AAAI Conference on Weblogs and Social Media (ICWSM)* (pp. 295–298).

Naveed, N., Gottron, T., Kunegis, J., & Alhadi, A. C. (2011). Bad news travel fast: A content-based analysis of interestingness on Twitter.

Nosko, A., Wood, E., & Molema, S. (2010). All about me: Disclosure in online social networking profiles: The case of FACEBOOK. *Computers in Human Behavior*, *26*(3), 406–418.

Quercia, D., Capra, L., & Crowcroft, J. (2012). The Social World of Twitter: Topics, Geography, and Emotions. *Sixth International AAAI Conference on Weblogs and Social Media*.

Robards, B., & Bennett, A. (2011). MyTribe: Post-subcultural Manifestations of Belonging on Social Network Sites. *Sociology*, *45*(2), 303.

Rogers, R. (2009). *The end of the virtual: Digital methods*. Amsterdam University Press.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. *Proceedings of the 19th international conference on World wide web* (pp. 851–860). ACM.

Savage, M. (2009). Contemporary sociology and the challenge of descriptive assemblage. *European Journal of Social Theory*, *12*(1), 155.

Savage, M., & Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology*, *41*(5), 885–903.

Savage, M., & Burrows, R. (2009). Some further reflections on the coming crisis of empirical sociology. *Sociology*, *43*(4), 762.

Sharad Goel, Jake M. Hofman, & M. Irmak Sirer. (2012). Who Does What on the Web: A Large-Scale Study of Browsing Behavior. *International AAAI Conference on Weblogs and Social Media; Sixth International AAAI Conference on Weblogs and Social Media*. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4660

Sheller, M., & Urry, J. (2006). The new mobilities paradigm. *Environment and Planning-Part A*, *38*(2), 207–226.

Shin, S. I., & Hall, D. (2011). Identifying Factors Affecting SNS Users as a Temporary or Persistent User: An Empirical Study.

Skeels, M. M., & Grudin, J. (2009). When social networks cross boundaries: a case study of

workplace use of facebook and linkedin. *Proceedings of the ACM 2009 international conference on Supporting group work* (pp. 95–104). ACM.

Stapleton, L. K. (2011). Taming big data. Retrieved March 10, 2012, from http://www.ibm.com/developerworks/data/library/dmmag/DMMag_2011_Issue2/BigData/index.html?cmp=dw&cpb=dwinf&ct=dwnew&cr=dwnen&ccy=zz&csr=051211

Thelwall, M. (2006). Interpreting social science link analysis research: A theoretical framework. *Journal of the American Society for Information Science and Technology*, *57*(1), 60–68.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (pp. 178–185).

Wattenhofer, M., Wattenhofer, R., & Zhu, Z. (2012). The YouTube Social Network. *Sixth International AAAI Conference on Weblogs and Social Media*.

Weng, J., Lim, E. P., Jiang, J., & He, Q. (2010). Twitterrank: finding topic-sensitive influential twitterers. *Proceedings of the third ACM international conference on Web search and data mining* (pp. 261–270). ACM.

Weng, L., Flammini, A., Vespignani, A., & Menczer, F. (2012). Competition among memes in a world with limited attention. *Scientific Reports*, *2*.

Wittel, A. (2001). Toward a network sociality. *Theory, Culture & Society*, *18*(6), 51.

Yang, L., Sun, T., Zhang, M., & Mei, Q. (2012). We know what@ you# tag: does the dual role affect hashtag adoption? *Proceedings of the 21st international conference on World Wide Web* (pp. 261–270). ACM.

Zaman, T. R., Herbrich, R., Van Gael, J., & Stern, D. (2010). Predicting information spreading in twitter. *Workshop on Computational Social Science and the Wisdom of Crowds, NIPS*. Citeseer.