UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIOLOGY, SOCIAL POLICY AND CRIMINOLOGY

DIGITAL TRACES, SOCIOLOGY, AND TWITTER

Between false promises and real potential

by
Olivier R. Philippe

Thesis for the degree of Ph.D in Web Science

June 2016

Abstract

We are told that society changes. It evolves toward a more fluid, active and horizontal form of socialisation (Bauman, 2000; Castells, 2009; Sheller & Urry, 2006; Urry, 2000; Wittel, 2001). As much as a change in the social, it is also a change in the conception sociology gives to the sociality and society as large. There is a shift in social theory with recent raises of post-demographic perspective, and post-modernism (Latour, 2005, 2011, 2013; Ruppert, Law, & Savage, 2013).

Along these, *post* changes, another revolution takes place, in the use of new technology and mainly the Web. These new uses are associated with an explosion of the quantity of digital traces, often labelled as *Big Data.*

This *Big Data* paradigm brings radical changes in research that are perfectly suitable for computational researchers and other data scientists, and they are taking full advantage of it (Hey, Tansley, & Tolle, 2009). This new landscape, in society and in research, puts pressure on sociology, and other social sciences fields, to find adequate answers to these societal, theoretical and methodological challenges. The best proxy of these challenges and the tensions between scientific fields and the new form of social interactions are the *Social Network Sites (SNSs)*. They represent the extreme case of horizontal and fluid interactions while producing an incredible amount of accessible *digital traces*.

These digital traces are the essential bricks for all the research using SNSs. But despite this importance, few researches actually investigate and explain how these digital traces are produced and what is the impact of their context of access, collection and aggregation. This thesis focuses on these digital traces and the gap left in the literature. This empty space is however conceived as a central point of tension between sociological positions on how to define new social interactions, and methodological principles imposed by the logic of Big Data.

The work is articulated around one specific social network, *Twitter*. The reason for this choice lays in its openness, the easy use of its APIs, and, in consequence, by the fact it is the most extensively studied SNS for now.

I begin the work on the definition of the new form of sociality using the network concept as the key concept around which several notions, such as social, cultural and

technological can be articulated. I conclude that none of these evolutions are independent and need to be seen as co-integrated. In consequence, the change in the social interaction needs to be seen as much as a factual change, than a change in our way to interpret it. From this conception of the *network* and the importance in our understanding of social interactions, I retrace the evolution of the notion of Big Data, specifically with the example of Tesco and their ClubCard. This is the first step to locate the technological changes into a more comprehensive methodological framework. This framework, the *transactional perspective*, is decomposed to understand the consequence of such position applied on SNSs and specifically on Twitter. This is the first explanation of why the research almost entirely focuses on the Tweets and what are the consequences on our understanding of the interaction on the social web service. Then I use this first iteration of the definition of a digital trace to build a new definition of what a Social Network Site is and centre this definition around the concept of activity and context.

I operationalise these concepts on Twitter to develop a new method to capture social interaction and digital traces that are often put aside due to the difficulty of their access. This method takes into account the limits imposed by the Twitter APIs and describes the consequences they have on the generation of a dataset.

The method is based on a constant screening of sampled profiles over time. This method allows people/us/you to reconstruct the missing information in the profile (the trace of the changes in the friends' and followers' lists). This information creates the measure of context (the user's network) and activity (tweeting and adding or removing links) defined earlier.

The obtained dataset provides an opportunity to see the importance of the aggregation process and the flexibility offered by the digital traces.

Then following this, I developed three analyses with different levels of aggregation for different purposes. The first analysis was to test the hypothesis of the influence on users' context activity on their own activity over time. The second analysis, did not use the time as a measure of aggregation but tested the same hypothesis on an individual level. And finally, the information about the activity itself is analysed in order to see to which extent the digital traces obtainable contain the sufficient information about the change in activity itself.

DECLARATION OF AUTHORSHIP

I, Olivier R. Philippe, declare that this thesis and the work presented in it are my own and have been generated by me as the result of my own original research.

DIGITAL TRACES, SOCIOLOGY, AND TWITTER: Between false promises and real potential

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
1. [Delete as appropriate] None of this work has been published before submission [or] Parts of this work have been published as: [please list references below]:

Signed: ....

Date: June 30 2016

Acknowledgements

To Susan Halford, I want to say my most honest, and deeply grateful thanks! Thank you for your abnegation and persistence in your help, despite my stubborn attitude and my erratic communication, and for the long meetings spent in deciphering my English.

Thank you Les Carr for you insightful supervision and assistance in helping me to finish my PhD without having to live under a bridge, which is infinitely more comfortable.

I also want to thanks Jeff Vass for having accepted to jump on the (chaotic) bandwagon. I really enjoyed all our discussions, related or not to the PhD, and your advice has been invaluable.

Also thank you to Wendy Hall for having given me the opportunity on the first place to do this PhD.

I want to also thank Susan Stock who gratefully accepted the job of proof reading several sections of this work, when my writing was at its worst. Thank you to Jennifer for her help during the early stages of this work. I also want to thank Claire for her last minute help (and therefore essential).

Je ne pourrais faire de remerciements complets sans mentionner mes parents. Merci pour votre soutien, tant financier que moral, durant toutes ces années. Promis, quoiqu'il arrive, c'est la dernière!

I finalment, gracies Ari per la teva paciencia i el teu suport. No m'hauria sigut possible aquest putu treball sense la teva ajuda. Soc molt afortunat de tenir-te al meu costat.

Table of Contents

Illustration Index

Figure Index

Index of Tables

Introduction

This work is based on premises drawn from the ideas presented in three articles: *the Coming crisis in sociology* (Savage & Burrows, 2007). Latour's conference talk on *Networks, Societies, Spheres: Reflections of an Actor-Network Theorist* (Latour, 2011) and *Measuring user influence in twitter: The million follower fallacy* (Cha, Haddadi, Benevenuto, & Gummadi, 2010). These three articles represented, at the start of this PhD, the formerly accepted understanding in sociology, the new methodological development and an optimistic future for the field, but they can also define the sociological aspect of Web Science. The first article urges sociologists to take advantage of transactional data produced by marketing sciences and to incorporate their methods in sociology to maintain the field's practice of using state-of-the-art, new methods and insights, as was the case in the previous century when qualitative methods and statistics were used.

Latour's talk is about the importance of traces in a descriptive sociology where contextualized activity is central and offers a radically new and different perspective for sociology.

The article on Twitter is one of the first that uses Big Data (the entire set of tweets published from Twitter's inception in 2006, up until 2009) to analyse and answer questions about social interactions and, in this case, to measure influence.

From these readings, the purpose of this study is to integrate these three different perspectives, where each is focused on the emergence of a new kind of data, generated from interactions on Social Networking Sites, and each has a potentially different impact on sociology. Indeed, this paper is positioned at the intersection of three contemporary phenomena that are rapidly evolving and impacting sociological research: the opportunities presented by the availability of a large amount of digital data; the recognition that the existence of this new kind of data might have generated a new perspective in sociology; and the tradition of social network analysis where human behaviours are viewed as being expressed through this data and thus analysed as such.

However, there has not yet been a comprehensive analysis, from a sociological perspective, of the impact of this new kind of data on sociological research, nor of its advantages and limitations. The main argument of the present work is that such an integration requires three distinct developments in order to effectively bring together marketing methodology and sociological theory on a computer science terrain.

Such an interdisciplinary perspective, as with any hybridization, is not a clear and direct interaction between these disciplines but rather represents several levels of integrations, as illustrated in each chapter of this paper and as explained in greater detail below.

A first question to establish is therefore: *to what extent can Big Data be used in sociological theories?* Hence, the first chapter develops the idea of a fluid and networked society, as conceptualised in network and information theory as well as the mobile paradigm. This specific position raises concerns about how to capture relevant information. This is the first link with practices on Social Network Sites; how a profile can present the opportunity to understand the principle of a network society/sociality and its associate principle of a dynamic sociology in opposition to a static understanding of our social interactions.

Secondly, Savage & Burrows' call for the use of transactional data implied a shift from the use of traditional data to something radically new for sociology. This requires a deeper examination into the nature of this data and how it is produced. Since it is not an established methodology that is well integrated in the field, there is a need to understand its potential and its limits before applying it. With this in mind, a deconstruction of marketing's generative processes of transactional data is needed. This is an essential step towards understanding where the advantages and disadvantages of such information lies.

However, the decomposition of transactional data into a process, rather than a product, is different from the idea put forward by Savage & Burrows. The data, or Big Data, is the result of a controlled process and the question is not: *how to integrate this result within a sociological perspective?* But *how to integrate the process itself?*

Therefore, shifting from a result (transactional data) to a process (transactional perspective) demands the understanding of the application of this process within another field. We can no longer assume that Big Data from Social Network Sites is similar to those generated within marketing perspectives; we need to develop a specific perspective and understanding for the social field.

This is why the deconstruction (or reconstruction) of the process specifically applied to Social Networks is essential. It helps us to understand which factors influence the process and, in return, the data used in the research, and points to the growing importance of the essential components of the Big Data; the *digital trace*.

Finally, this interest in digital traces needs to be integrated with theoretical positions developed for the first question. The study of such traces needs an emphasis on the activity and the context as the essential foundation of a more dynamic sociology based on network logic and mobile sociality. Methodology is also essential as the information about network logic and mobile sociality is contained in the digital traces (as they are a direct measure of a behaviour produced in a specific context).

In short, the research questions are as follows:

- Which recent sociological theories or conceptions can best integrate the potential of Social Network Site data.

- How is transactional data produced and to what extent is this specific form of data production applicable to social interactions that occur on Social Network Sites.

- How to integrate digital traces within a specific methodology, taking into account sociological theories and the advantages of this new source of information

All these developments, even if they integrate marketing and computer science (or, at least, an understanding of the way data creation works) are dominated by a sociological perspective.

However, to frame these general questions in a practical framework, I apply these questions to Twitter and, more specifically, to the relationship between influence and the context. This specific application will not only try to examine the extent to which the theoretical positions developed above can help us gain other insights on data but, also, from a more computer science perspective, allow the development of a practical tool.

The red string, transversal to this work is the *digital trace.* The aim is to see how this trace evolves at each step of its creation and how the comprehension of this process is essential to know what computer science and sociology study when their object is online behaviour.

Therefore, this application can be seen, from a sociological point of view, as a case study of the theory developed; for a computer scientist as a tool to gather and analyse data differently; and for a Web Scientist, as the integration of the processes of data-creation with a specific focus on the context and activity on Twitter.

1 The Concept of Network

1.1 Introduction

The concept of *network* is central throughout this work. (Un)fortunately, this word, *network*, is used for several concepts in different disciplines. Sometimes it represents a type of social organisation; and sometimes a method of enquiry. It can also refer to a methodological perspective or a type of website. Additional variations of use can be found in different theories with different purposes, which are sometimes incompatible.

Within this messy context, I will also use the word *network* in different contexts, from different fields, and to answer different questions. This heterogeneous use of the word is not due to any strange linguistic obsession for it. It is more due to a global tendency to see, understand, and reduce (or extend) everything to a node, a link, a web…

This could be a manifestation of the actual evolution of society into a more networked structure. We are supposedly living in a more organic, connected society, therefore, the notion of network becomes more natural as we evolve in such a world. Another explanation could be found in the spectacular penetration into society of the internet and the technological realisation of the concept of network itself. These two naive propositions represent two main concepts that are often linked together, *social* and *technological* evolution.

This section addresses the question of the interaction of those two main concepts, and the evolution of our society towards a *network society*. To answer this interlocking question, I use Castells to retrace the origin of this evolution and the idea of a network as a *social structure*. With social structure as a basis, I link Castells' approach to a network society, with Lash's considerations on the *information society*, I provide some examples of the concepts developed on the evolution of a network society which extends the concepts to a more recent situation. This approach is helping to bridge the emergence of a network society (Castells), to the consequences of it (Lash) and to see how it is deployed within the web in general.

Nevertheless, Castells still has an important impact today and can explain, with accuracy, some recent developments of our societies, even if those developments occurred after the initial publication of his book (Castells, 2011). However, even if Castells forecasts numerous current social phenomena, his analysis can be seen as being outdated in some perspectives. This critique is less about his ability to understand our social structure than the construction of his rhetoric. What makes Castells' argument clear and powerful is the description between a *before* and an *after*, a *not (so) networked society* and a *network society*. This type of argument, extensively used by sociologists, helps to present ideas in a clear perspective, but as soon as the prophetical *after* becomes the new *now*, the argument loses its power of explanation.

To overcome this issue, I link that theoretical development with Lash's conception of *information society*. His argument will be seen as the consecration of a network society. Even if his argument is also built on a distinction between a *before* and an *after*, the objective is to describe an information society (which to some extent relies on a network society), and it is possible to understand his *now* as the *after* developed by Castells.

However, this development of theories in parallel does not consider the disadvantage of the outdated issues, since Lash's book was written in 2001. I therefore extend these theoretical developments with examples from the Web. These extensions represent what currently happens in this *network/information society,* and how the Web is the last creation/manifestation of this logic. The consequences of this technological and social evolution ultimately results in a new scientific environment with a specific data production which, I will explain, is different from the knowledge production that is specific to the information age. That development will be discussed in the next chapter.

1.2 Network as a structure

For Castells, the network society is the result of a more complex and co-evolutive process between social structures and technological opportunities, economical logic and informational flux. The network is, therefore, not the way to deploy the change, but a mix between a reason and a result. It is why the use of words such as *structure* and *logic* adjoin the word *network* in different moments in his book; the former referring to the result of the network society, while the latter is more about the possible condition of its realisation.

However, before going further, it is essential to go back to the definition of *network* used by Castells.

I use his latter definition, logic network, (the result of the evolution of our society) as the starting point of this section.

> I shall first define[1] the concept of network, since it plays such a central role in my characterization of society in the Information Age. A network is a set of interconnected nodes. A node is the point at which a curve intersects itself. What a node is, concretely speaking, depends on the kind of concrete networks of which we speak.

(Castells, 2011, p. 501)

This definition is not really different from what we can expect from a classical definition of a network. However, he insists on the notion of this network deployed within the social organisation, as a structure:

> Networks are open structures, able to expand without limits, integrating new nodes as long as they are able to communicate within the network, namely as long as they share the same communication codes (for example, values or performance goals). A network-based social structure is a highly dynamic, open system, susceptible to innovating without threatening its balance.

(Castells, 2011, p. 502)

This type of structure can be characterised by its *openness*, *fluidity* and *resilience*. This structure is the result of an evolution from the *Industrial age* to an *Informational Age* made possible by technological evolution. However, the technology itself is not enough to explain that evolution, and Castells speaks more about a *technological paradigm*, than a techno-determinism. This paradigm is going to help to see how and why we are living in a different social organisation than before, and to see what the consequences are of such an organisation, on different aspects of the society.

[1] Ironically defined in the conclusion

1.2.1. New technology paradigm

Initially, the concept of technological paradigm, developed by Dosi, tries to answer the question of the causal relationship between economic growth and technological change (Dosi, 1982, 1988). The idea was to overcome the overly simple approaches which were supposed to explain a unidirectional and linear relationship between the economic and technological processes: *demand-pull* and *technological push*[2] (Von Tunzelmann, Malerba, Nightingale, & Metcalfe, 2008).

Dosi is directly inspired by the scientific paradigm developed by Khun and his development about scientific research and revolution (Kuhn, 1970). This inspiration is clearly stated in the very definition of the *technological paradigm*:

> In broad analogy with the Kuhnian definition as a ”scientific paradigm”, we shall define a ”technological paradigm” as ”model” and a ”pattern” of solution on selected material technologies.
>
> (Dosi, 1982, p. 152)

This conception offers the same advantages of a Khunian perspective with a technological paradigm defined by its own resources, trajectory and objectives. This is the starting point for Castells to render a shift in the society, a new technological paradigm based on the information. This new paradigm, even if it is mainly explained in terms of technological and economical process (Dosi’s heritage), will impact the entire society, form the economic logic of production to the symbolic interactions and mechanisms of power (or more precisely of inclusion/exclusion). A technological paradigm is characterized by its selected technologies and this informational paradigm does not infringe on this specificity. However, this type of list has shortcomings and does not necessarily help us to understand what the processes are that are in place. However, the five characteristics of the informational paradigm are useful to understand the evolution of our society and the associated condition of its realisation as well as its consequences. This is why these characteristics form the structure of the following sections. As stated above, the web is understood as the last manifestation of this evolution but, to exaggerate the point of Castells, and to supersede the importance of his developments (work) in

[2] The first explanation insists on the influence of the market leading the innovation, while the second gives the technological change more autonomy.

understanding the web. The world is numeric but that characteristic has little to do with the web: the web is only its last manifestation.

Technologies act on information

The first feature of this paradigm is *technologies are action on information*. This means that information is no longer a by-product of the production but its final goal. Not only are the *Information Communication Technologies* (ICT) developed for communication, but the idea of information itself is the final product, transforming a means to an end. As stated by Lash:

> In information-capitalism labour power operates with not practical, but discursive knowledge; operates with no classical, but information machines; and works on not raw material, but on raw or semi-finished information in order to produce informational goods.
>
> (Lash, 2002, p. 142)

The manifestation of this prevalence of information over other products, or modes of production, can be found, for Lash, in the importance of *Intellectual Property* (IP) and the idea that prototypes are more important than the ownership of production lines.

Using the example of the CD, he said that owning the model of the CD, the copyright on the songs, or the software which can generate millions of copies is far more important than the production line. It is not the product which is important but the *artefact* (Lash, 2002, p. 22). This artefact represents the process of innovation. It is the result of a network of departments working on knowledge, on product innovation and, to use Thrift terminology, *creativity*.

This creativity becomes *a value itself* (Thrift, 2005, p. 133), but to assess this value, the result of a knowledge production, some difficulties need to be overcome as the knowledge itself is difficult to measure. Thrift lists five reasons for this difficulty:

> The non-convexities; the absence of a complete set of market; the lack of homogeneity; a strong asymmetry and a degree of which knowledge represent public good
>
> (Thrift, 2005, p. 22)

Not only is the value difficult to assess, the transformation of this value into capital is also a problematic task in a world of flow, characterised by a form of capitalism based on the knowledge. The main point of information is circulation (goods, knowledge, money, people…) so the question is about the control of circulation while it is monetized.

The idea is to valorise the asymmetry of knowledge. But this asymmetry (in the case of the CD) is lost as soon as (sometimes even before) the first CD is released. From that moment, everyone can easily reproduce the content, the information contained on it. That is the inherent characteristic of information: it is easily accessed and shared.

These characteristics are antinomic with the control needed to monetize the information. It is why the IP is so important. It is a tool that controls the sharing of information and allows the owners to valorise their work and their rights (which may be unrelated to each other). Lash is following Deleuze and his concept of solidification of object, Lash speaks about "congealed social relations" (Lash, 2002, p. 193). The Intellectual Property[3] is this solidification, a crystallisation of the knowledge into a space of exclusion/inclusion owned by someone or some company. It is the "right to *exclude* others from valorising that object" (Lash, 2002, p. 196). It is in a sense, an extension of the temporary asymmetry of knowledge within a legal form.

The point I am trying to make is, as highlighted by Lash about the Power of exclusion and inclusion in a network society (or as we will see later, by Castells, and the importance of the place in the structure), is that the IP is more of a logic from the industrial world, applied to a knowledge society. It is a paradoxical condition (crystallisation of a flow into a temporary stable form) which allows a power over the network by allowing a decision on the inclusion/exclusion of people inside this innovation network. Therefore, this temporary crystallisation is an artefact, an interconnected place, where people are confronting each other.

But this crystallisation, the IP, does not follow the principle of a network society or an information age, not the concept of solidification itself, but applying this solidification to the control over information. The solidification is based on the idea

[3] In its three distinct forms; "*Patent*: intellectual labour inscribed in goods; *Copyright*: intellectual labour inscribed in meanings; *Trademarks*: affix meaning to an *array* of goods." (Clive WJ J Granger, 1969).

of capital build on knowledge but treated as any other goods. However, the information is important only when it is shared and transmitted and as soon it is transmitted, there is no more control over it. As soon as the information is released, it is possible for others to have access to it and to diffuse it too. That is why there is such tension between new uses of technology, such as the peer-to-peer and copyright holders who try to preserve their rights. But when it makes sense for Lash to compare it with the production line, nowadays, the Intellectual Property is compared with other forms of economic power such as *prosumer* (I will describe this in detail later) or the importance of the database and information about users themselves (this point is developed in the second chapter). The idea of IP is a transition between two modes of production but certainly does not represent the age of information, like the television does not represent the informational age. However, the idea of adapting the *crystallisation*, the Deleuze's concept of *solidification* makes sense in a conception of economic power within a networked environment. This idea will be used later but in the conception of Social Network Sites and not adapted to describe the necessity for companies to control their knowledge capital, but to show how the interaction changed and can still be found solidified in one way or another. For him, it is not just that the information is becoming an end in the commercial and production world, it is also a shift in the foundations of our society. The evolution of means and ends differed from one society to another, and he retraces the evolution over time.

The traditional society was specified by the idea of *good life* as the end, the goal to reach. The transformation of nature was used to reach such goal, representing the means. Later, in an industrial order, the transformation of nature, or the production of goods (*manufacturing* became the end of this society, while the capital (of production but also of ideology to ensure the order) was the means to reach this goal. Lately, the informational order is using information (ideology and knowledge capital) as an end, as the final goal (Lash, 2002). Moreover, the prevalence of information is that it becomes, *"referent, end and meaning"* at the same time (Lash, 2002, p. 80).

Pervasiveness of effects of new technology

The second feature is the *pervasiveness of effects of new technology*. This feature does not only impact the economic sphere but the entire set of human activities:

> Because information is an integral part of all human activity, all processes of our individual and collective existence are directly shaped (although certainly not determined) by the new technological medium.
>
> (Castells, 2011, p. 70)

Even if this pervasiveness were a certainty at the time the book was written, we can see that it holds more importance today. We can see this pervasiveness through the expansion of the Internet or the explosion of smartphones[4]. These technologies have not only integrated our everyday life with new devices, they also modify our integration of virtual interactions and the types of uses we have with and within it. We can use, as an example, the evolution of debates and questions about the impact of the web on our social interactions. When the Internet started to spread, at the same time as Castells' writing, the debates were about the impact of anonymity on the web, with the use of pseudonyms and nicknames instead of real names. It was also the debate about the impoverishment, or enhancing, of virtual interactions in comparison to face-to-face relations (Kraut et al., 1998).

Nowadays, the question of anonymity (such as the impact of anonymity on the aggressive behaviours (Moore, Nakano, Enomoto, & Suda, 2012) or the comparison of virtual and real relations, are still important but, we can observe the emergence of new questions showing how the virtual and the real world are seen as interconnected, or at least, how they interact with each other where there are less boundaries between the real and the virtual, but also that the web does impact every other aspect of our life. Scientific literature can highlight these new concerns arising from the pervasiveness of the effects of new technology, and newspapers are also full of examples, more or less dramatic.

These new questions, such as those raised in the literature about online bullying of teenagers and identity theft, are a manifestation of this integration of real-virtual, on another extend than only conceptual. This evolution does not mean that the previous issues about anonymity do not exist, but now the question is how is technology integrated in our lives and how does it impact all aspects of our lives, with a tendency to be more and more pervasive as the technology reaches, by its

[4] The number of mobile broadband subscriptions was 268 million in 2007 and is expected to be 2.09 billion in 2013 while the number of wired-broadband subscriptions was 284 million in 2007 and is expected to be 696 million in 2013 (Union, 2013).

miniaturisation, its reduction of costs and the change of our habits, every single part of our environment.

Networking logic

The third characteristic is *Networking logic*. The technology makes the implementation of the network structure possible (Castells, 2011, p. 71). The point that is being made is not that the networking logic is new and did not exist previously, but that the technology gives the material condition to extend this logic to the whole social structure, making the morphology more important than the social action itself. The consequence of it is that the presence, or not, in the network is more important than the action itself. Someone, in some areas/cities outside the network will have no power. It is their position and the position of their network which matters[5] (Castells, 2011, p. 500). This idea is seeing power as a matter of position and morphology and the capability to be in the network. The influence of structure and position is shifting the previous debate about relationships between classes towards a more inclusion/exclusion relationship between people. This is now a matter of inclusion/exclusion more than class relationship (Lash, 2002). As seen, with IP, that not only the position, but the right to include/exclude is also important. This gives a less structural version of this logic.

Flexibility

The fourth characteristic is the *flexibility*. This is the ability for the structure to change without destroying itself. Organisations, as well as society, are changing more and are more fluid. Tools provided by the technology allow this flexibility:

> Turning the rules upside down without destroying the organization has become a possibility because the material basis of the organization can be reprogrammed and retooled.
>
> (Castells, 2011, p. 71)

Instead of calling it *flexibility in organisation*, Lash, uses the disorganisation terminology to insist on the decline of a traditional organisation defined as: *hierarchical systems of normative rules* (Lash, 2002, p. 27). It is the shift from social

[5] For instance, in their study, Vitali and al. show, through a network analysis, how the position in the network of ownership in the financial market matters. A company in the core of this network has 50 % of chance to be considered as a top holder in comparison to the 6% for companies in the in-section (Castells, 2004, p. 8).

norms to the importance of cultural values as disorganisation is about value (Lash, 2002, p. 41). But the consequence is also an increased fluidity and movement:

> Disorganisation for their parts are fluid and mobile. They form, they de-form, they break up and come together again in different places. Their existence is one of being on the move.
>
> (Lash, 2002, p. 41)

Then, from the evolution of technology, which provides tools to sustain the fluidity in organisation, we have a definition as disorganisation, which is not only a form of more mobile organisation, but a shift from norms to values. This is a radical change in the component of organisation itself.

Convergence of specific technologies into a highly integrated system

The fifth and last characteristic is the *convergence of specific technologies into a highly integrated system*.

This interest follows the idea that it is our culture which modifies our perception of reality and therefore our experiences and our social interactions. Castells, following Postman's argument, down this logic.

Firstly, the communication modifies the culture, as all media, transmitting culture, are in fact communication. Secondly the culture modifies our interpretation system and, finally, our reality is an interpretation and not living *sui generis*. Therefore we are experiencing the reality differently in the function of our interpretation system and it is the same as experiencing our social interactions (Castells, 2011). From this perspective, we need to understand how the culture is impacted by the media and the consequence of it before stating its influence on our social interaction. To understand the impact of technology and media on culture, we are still following Castells and his effort to depict the evolution of media as:

> A system of feedback between distorting mirrors: the media are the expression of our culture, and our culture works primarily through the materials provided by the media
>
> (Castells, 2011, p. 365)

We then have an interrelated system whereby the media and the culture modify our perception and our way of developing interactions and, therefore, following Castells, the nature of our interactions themselves. Now it is possible to see the main impact, the following link between the technological paradigm and the social interactions.

1.2.2. Evolution of media and evolution of our social interactions

Castells built his network logic concept with the evolution of media. The media evolution is, beside the technological evolution which made it possible, an evolution of the message and its channel. Message and media represent different aspects and their forms evolve with the media, the political and the economical context. It is evolving from the Gutenberg galaxy to McLuhan Galaxy and from this Galaxy to the interactive multimedia (Internet) which represents the outcome of this evolution and imposes the network logic through the co-evolution of three concepts, *integration*, *diversity* and *interactivity*.

McLuhan stated the famous *the medium is the message*. He was talking about the supremacy of television. This medium became the main media and imposed on all other forms of expression (*i.e. newspapers, radio,…*) its own logic by integrating all messages in one channel.

The point is not saying that the other media disappeared but they were integrated in this media:

> The real power of television, as Eco and Postman have also argued, is that it sets the stage for all processes that intend to be communicated to society at large, from politics to business, including sports and art
>
> (Castells, 2011)

This normalization of all channels, all variety of communication through a unique media is the reason for the transformation of the culture. If any message needs to be integrated in one media, and as culture is communication, the media is the culture. This is the idea of mass-media and the homogeneity of culture or the McLuhan Galaxy.

But the idea of mass-media is problematic for Castells. He mentions the fact that media has to be interpreted by individuals, regarding/in relation to their contexts, aptitudes and sensibilities. Moreover, television represents the idea of mass-media but, when states lost control over its diffusion and private companies invested in the market, the very idea of a mass-media was flawed by the diversification of public and channels. Not only limited to one or two television stations, the market starts to exploit every possible existing niche (or creates new ones) to be competitive. Then the message became the media, as the diversification and the specificity of media raised.

Nevertheless, the two characteristics, the integration of all communication in one media and the diversification of this media, are not sufficient to create an informational society as there is still a lack of interactivity. The interactivity first appears with the Minitel, in France, but was really developed by the introduction of the Internet and, again, it is a co-evolution of the technology and the social context.

This late concept, the *interactivity*, associated with the integration and diversification imposes the network logic. Everything that is to be distributed needs to adapt to this logic:

> […] the price to pay for inclusion in the system is to adapt to its logic, to its language, to its points of entry, to its encoding and decoding. This is why it is so critical for different kinds of social effects that there should be the development of a multi nodal, horizontal network of communication, of Internet type, instead of a centrally dispatched multimedia system, as in the video-on-demand configuration.
>
> (Castells, 2011, p. 405)

Therefore, the culture is modified by the need for the message to adapt to the existing network logic. Before, it was the television logic with one single channel and the idea of mass-media. Now, the interactivity is breaking, by the fundamental logic of it, into small inter-connected, horizontal channels, or media. This has a profound impact on our socialisation processes:

> […] because they are the symbolic fabric of our life, the media tend to work on consciousness and behavior as real experience works on dreams, providing the raw material out of which our brain works.
>
> (Castells, 2011, p. 365)

This perspective is the idea that there is no difference between reality and virtuality, both are representations understood through signs. The result of this change is called the *real virtuality*:

> A system in which reality itself (that is, people’s material/symbolic existence) is entirely captured, fully immersed in a virtual image setting, in the world of make believe, in which appearances are not just on the screen through which experience is communicated, but they become the experience.
>
> (Castells, 2011, p. 404)

This is the idea of new socialisation, accompanied with a timeless time and a non-linearity of event and narrative developed through these relations (Castells, 2011, p. 491).

Then the technology, but also the culture, makes our social relationships horizontal, networked and flexible. It is not only because the media evolved but also because of our relationships which are shaped by them and adopt the same characteristics.

The web service, Youtube is the best example to show the evolution of the media since the evolution of the television. It can be considered as a new form of television, with the principle of maximum interactivity proned by Castells.

1.2.3. Youtube Galaxy or the media theory

When Castells speaks in terms of *network logic*, the need for horizontal interactivity, it is still the notion of delivering content from one side to more active consumers on the other side. It is still a limited version of the networking logic as there is a distinction between the two roles.

At the time the book was written, the full deployment of the horizontal and fully interactive version of the web was only an ideal aiming to offer the possibility to everyone to express their own ideas, to develop and share whatever they wanted, regardless of any physical borders or external control. This idea can be found in the “declaration of independence of cyberspace (John, 1996). Later, with the appearance of the web, the same ideas were represented but it is finally with the so-called *web 2.0* that we have seen the realisation of this ideal.

Not that the technologies did not exist before, the Usenet board, the weblogs and the forums already permitted everyone to share, comment and participate, but it is only with the conjunction of a higher, material accessibility to the web and a simplification of use of technology that the *pure interactivity* reached a mature level. In this perspective, I need to re-contextualise the idea of networking logic into a more contemporary example.

I have developed with Lash and the idea of Intellectual Property that in an informational society, knowledge is more important than the production line. I stated that vision was problematic in the sense that power is gained through exclusion rather than inclusion. As well as the IP, another form of economic power exists alongside production lines that is more representative of an information society raised at the same time as the *pure interactivity* came to a reality. Lash stated that there are no longer receivers and audiences but primarily users (Lash, 2002, p. 76). This is exactly what Youtube shows.

Youtube offers a common space which brings together the roles of consumers and producers. Moreover, the interactivity and the status of users is so extended that there is no longer initial separation between the two roles even if the activities of users can differ.

When users register for the Google service, they can comment and/or upload videos, to create their own channel, create links with other users (to some extent, create an audience), subscribe to other channels and, of course, watch video. The only difference between a producer and a consumer is the degree of participation and the wish of each user to publish or not. It is, in some sense, a full extension of the principle of the horizontal, where roles are blurred and cannot be considered as different. Everyone has the opportunity to interact with other users but also, everyone has the same rights and opportunity to publish and share content. Every consumer is a potential producer. It is the fundamental reason for Youtube being as stated by its slogan: *Broadcast Yourself*[6]. Let's invent another Galaxy to represent this idea, the *Youtube Galaxy*, to paraphrase Castells[7].

This idea, the disappearance of any distinction between consumers and producers has found a recent interest under the term *prosumer*. This concept can be defined as: "a

[6] Before/Recently this slogan was removed due to some graphical changes on the webpage.

[7] Or McLuhan.

process involving the creation of meanings on the part of the consumer, who re-appropriates spaces that were dominated by institutionalized production, and this extends to the exploitation of consumer creativity on the production side" (Paltrinieri & Esposti, 2013). We can see from this definition that a debate exists on the dynamics of such a new form of consumption. Some sociologists of consumption see the *prosumerism* as a new form of creativity, creating new subjective interpretation of some mass media products (a similar idea in Castells who denied the concept of mass-media but it is with a more active perspective here, while Castells spoke more about personal context and different sensibilities). Others are pointing out the alienation process due to the *work for free* and the creation of content re-appropriated by industry and governments (Fuchs, 2010; Rey, 2012; Ritzer & Jurgenson, 2010). Beside this Marxist analysis, Paltrinieri et al. explain that this form of consumption was raised in the past few years alongside the rise of Social Networking Sites[8]. The reasons why SNS are so suitable for prosuming are:

- *The production and sharing content on Web 2.0;*
- *The abundance of produced and published content;*
- *The unpaid work of those who produce the content; and*
- *The online spread of culture of free content.* (Paltrinieri & Esposti, 2013)

This analysis is still an economic perspective on power and production, but I think the *prosumerism* can represent a more cultural re-appropriation of the media, going further than the simple necessity of an adaptation of the network logic to survive. Lash offers a useful highlight on this extended consequence while he is talking about information itself.

1.3 Social Network Sites and the last manifestation of informational society

I want to go even further than the concept of prosumerism by questioning the fundamental characteristics of Social Networking Sites in comparison with the other form of participation on the web, and linking this reflection in a cultural sense, as developed by Lash.

[8] The terms was coined the first time by Toffler in (1981).

Nowadays, the interaction and user-content generation have seen a different development in the rise of many social network sites and a new mode of communication. Instead of being centred around one person (weblog) or one community or topic (forum), these SNS are centred around the person, the user. A weblog brings people who are interested in the same subject to one person (a media producer).. The forums are a more interactive version[9]. On a forum, sections are separated into subjects, or topics, and people join them and comment, share ideas or argue in a discussion.

But the main point remains a common interest, where the principle of fluidity is external to the interest itself. People are brought together on this sole point of interest. It is a space where the global and nomad cultures can be tribal again, as Lash's interpretation of the *Global Village*. The interest for the topic is such as the church in the middle of a village, the place where travellers stop by and exchange in a traditional perspective but with the characteristics of an impossible narrative on an individual level. Here, adding the distance and the technology, people don't only meet (yes they still do) in a physical place but in a forum where other people can share information and interact with them. But, the narrative is contained in the subject, within the topic, the board itself: it is external to the users. It is an externalised culture built on temporary re-appropriation of each subject by users, creating a community.

With SNS, the logic is different, the point of interest is the person itself, the discussion is around this person, within people who are part of this network. It is not only a change in our way of discussing problematics but a radical and qualitative change in the way we create our environment, our own media. It is obvious that the possibility of using SNS is still limited by what the market is offering as possibility. Interacting on *Facebook* is not the same as interacting on *Instagram* or *Pinterest*, but they share the same ground: you are creating your media by your own activity and the interactions with the social links you are building on it.

The consequence, with SNS, is that it is no longer the media which is pervading our interactions, or our culture but, as this pervasion has already occurred, it is through our interactions that we create the culture. A narration is rebuilt around the person but it is made by temporary and weak interactions. Therefore, we have a more

[9] Even though they existed before the weblog under the version of Usenet.

fragmented, horizontal media based on communication itself but also based on the individual. In a sense, it is no longer possible to *be famous for 15 minutes* but we are all famous to our 150 friends for 15 seconds. The result is that not only is everyone a producer, as described for Youtube, or a prosumer, but that everyone is in an interaction, creating their own media, content and, with the interactivity, their own public. *The interaction is the media*.

I have seen the evolution of society with a macro perspective, essentially focused on the evolution of technology and the predominance of information in our society. I have argued, following the authors, that this evolution impacts on our culture and our socialisation. This has an impact, not only on how we interact, but also on how we conceive the world: *connected*, *self-centered*, *networked*, …

1.4 Network as a practice

Wittel's object of study is the network as a practice (Wittel, 2001, p. 52). His conception of network is derived from the structural and technological features developed by Castells. However, he conceived it as a new practice deployed by people in their everyday social interactions. Wittel's idea is to translate the macro-sociology of a network society into a micro-sociology of the information age (Wittel, 2001, p. 52). To do so, Wittel develops the concept of network sociality which describes how to create and maintain networks.

For him, the network sociality is a consequence of the individualization in our society. We have lost our community sense for the profit of our own story: *Network sociality is not based on a shared history or a shared narrative. Instead, it is defined by a multitude of experiences and biographies* (Wittel, 2001, p. 65). This idea of individualisation finds a direct echo in Castells' work, as the loss of legitimizing identity (Castells, 2004, p. 355)[10].

This network sociality is deployed through ephemeral and intense relations and is considered as a new form of organisation in work situations. His ethnography study (about people working in media) showed how people are focused and how they work on one project to which they are heavily committed, for a short period of time. As soon as the project is finished, they shift to another project with other collaborators

[10] The legitimizing identity is the identity imposed by dominant institutions to extend and rationalize their domination vis à vis social actors (Boltanski & Thévenot, 2006).

but with the same intensity and short term perspective both in terms of project and relationships.

A direct consequence of the individualisation and the ephemeral relations is that the sociality shifts from narrative to information. The disappearance of narrative is solely explained by the individualization, but the ephemeral relations impose another constraint. As people experience the relationships with more weak ties for a shorter time period, they need to speed up the transmission of information between them, reducing the amount of detail but increasing the efficiency in a timeless space of flows. This follows the same importance of information itself as developed by Lash and the informational development of the society where the relationships become presentational instead of representational. We are not developing long relationships, we need an instantaneous presentation of ourself (Lash, 2002, p. 76).

In conclusion, the network sociality is:

> […] a sociality based on individualization and deeply embedded in technology; it is informational, ephemeral but intense, and it is characterized by an assimilation of work and play. Furthermore, I suggest that certain features of the practice of networking might be ’new’: it is widespread practice in urban post-industrial spaces […]
>
> (Wittel, 2001, p. 72)

Witell’s work is a translation of macro-perspective into a micro focus. It is bringing the terminology developed by Castells, the network as a structure, to transform it as a practice and echoes Lash’s perspective on the development of a new form of socialisation where the process is displaced by informationality:

> The principle of society becomes displaced by the principle of information. An order in which sociality becomes displaced by the principle of informationality. Sociality (long-lasting and proximal) is replaced by informationality (short duration and at a distance)
>
> (Lash, 2002)

For him, the primary qualities of information are flow, disembeddedness, spatial compression, temporal compression, real-time relations (Lash, 2002, p. 2). Not surprisingly, the features of such processes remain essentially similar to the characteristics of network, developed in Information age. However it remains a different concept of network, not only as a logic and a structure but as an already shared practice.

The network practice as developed by Wittel represents the idea of endless reproduction in Lash, where we start new projects over and over, and new relations without link, without a narrative between them. The Social Networking Sites have the same characteristics of weak ties but present the advantage (the technological materiality of flexibility) to help to deal with these numerous and ephemeral relations.

However, I have seen that SNS can represent a re-territorialisation of the users and their lost narrative, as they are building their own channel, their own interaction and public, and create a central point where it is their own story which becomes the temporal/temporary focus point of others while they are interacting.

1.5 Network as a constraint

I have spoken about *horizontality*, *flow*, *fluidity*, *long distance relations* …All these concepts have a more general concept: *mobility*.

As Lash and Castells said, flows are not only a flow of information, they are a flow of people, a flow of goods, and a flow of money. In other words, there is communication about everything. In this mobile and communicative world, sociology needs to grasp people while they are on the move. This is why Sheller and Urry developed a new method and approach to catch this movement.

For them, the sociology was focused on statism for too long, ignoring the movement of all aspects of social life, from family to work and from leisure to political protests (Sheller & Urry, 2006, p. 208). The idea is really a united need in the methods of inquiry and the evolution of society. It is not saying that such a movement did not exist in the past, or that the fluidity and movement are more present today, but it is more about a profound lack of sociological understanding of the notion of mobility.

This conception of the reality relies on the concept of flexibility and network logic. They want to understand the *quickening of liquidity*. They want to understand the immobility and the creation of zones of connectivity and its opposite, the zones of exclusion (Sheller & Urry, 2006, p. 210). Again, the fundamental ideas are similar to the previous ones. The main point is the connectivity of people or places. And the correlation of this connectivity is necessary to address questions such as fleeting (or ephemeral), distributed (or networked), multiple (or diverse) and complex connections, ties, and relationships, by dealing with the sensory and the time-space compression.

The difference here is that their perspective is methodological. The main goal is to observe people's interactions by taking into account the mobility and the fluid interdependence. They develop specific methods and tools to reach it, such as walking with the street moving, or the 'time-space diaries',…But these techniques deserve a specific idea, following the bodies and its internalities in its *sensescape*: Bodies are not empirically fixed and given but evolve performances to fold notions of movement, nature, taste and desire, into and through the body. Bodies sense and make sense of the world as they move bodily in and through it, creating discursively mediated sensescapes that signify social taste and distinction, ideology and meaning (Büscher & Urry, 2009, p. 102). Therefore they conceive their object of study as experiencing actors and the need to gain access to this experience. Their methodological approach answers a unique conception of the reality. The new configuration of a moving society is not only a new process by which interaction differs, but also a constraint to access this sensescape. To assert this idea, they follow Simmel's arguments. Not only on the evolution of society in a big city, the impact on how we interact, and the time perspective with the constant move, but also on eye-interaction. The interaction is considered as the most important type of relationship we can have:

> People cannot avoid taking through the eye without at the same time giving. The eye produces "the most complete reciprocity; of person to person, face to face" (page 112). The expressive meaning of the face provides a special kind of knowing.
>
> (Sheller & Urry, 2006, p. 217)

Considering the body as a sensitive object in a symbolic world, the access to this pure relation is essential to understand the interaction it has on the move, experiencing different places, different time and creating ephemeral interactions.

Then the constraint of network is not only because there are more movements than before, or because sociology neglected this facet of our lives, but also because the consideration of our self, as an interpretative system within a context, changes the perspective we need to have to understand it. One direct consequence is this ethnography on the move, following the subject/body/system while it's interacting in different situations.

1.6 Methodological consequences

The emphasis of information and media leads him to conclude that it is not the network logic which invades the economical and the cultural but the media itself: *Media theory is only possible in an age in which social and cultural life has been pervaded by the media.* (Lash, 2002, p. 66). Therefore, when such a thing happens, sociology starts to be a *"mediology" (Lash, 2002, p. 206)*.

The media theory, or mediology is the extreme version of the impact of multimedia and electronic communication developed in Castells, the fifth idea of Castells about the convergence of technology and the network logic

So far, we have developed the idea that society evolved into a more networked world (network as a structure), in which people deploy a network sociality (network as a practice). But to study this practice, there is a need to develop specific tools to deal with the effect of more fluidity and movement (network as a constraint). This last development pointed out that the notion of network is a social conception but also a point of view which implies different perspectives on the object. Therefore we have to deal with the characteristics of it: its structure, its flexibility, its horizontalilty and its mobility. But, how do we apply this idea where traces are vanishing or too big to follow? How do we deal with a partial perspective of the world, where only users' activities appear and where the body which experiences them is not accessible? With electronic data it is not possible to access these experiences or feelings. It is impossible to access the eye as "*the most direct and purest interaction that exists (Simmel, 1997, 111)*" (Sheller & Urry, 2006, p. 217) and it is impossible to study the physico-symbolic architecture of a city to understand the impact on identity (M. Castells, 2011). In other words, how do we add the field constraints and limits to the already complex imbrication of society reality, sociological perspectives and methodological implications.

In fact, what is needed, is to go back to the original point of this chapter. I briefly mentioned the scientific paradigm of Khun as the foundation of Castells' argument through Dosi and the notion of *technological paradigm*. The breakdown of technological change, seen as a puzzle-solving trajectory, developing its own logic, tools, questions and answers, leads to the unveiling of this social phenomenon. Now the consequence of that is a profusion of information, an everyday-life, creation of information. This situation, where the information is the first element of response, is the crystallisation of interaction within the Social Network. This is the final manifestation of the network sociality, following the technological paradigm as developed by Castells.

For Castells the evolution of the technological paradigm brings a new form of information society, by modifying the forms of production, and, in a neo-marxist perspective, by modifying the forms of socialisation.

Wittel shows this network structure in practice, modifying the way, the duration and the purpose of our social relations.

For Lash, this change in society, the social relations and the production mode is also a change in our culture, from a representational culture, where object and culture are outside the world, into a change in a technological culture where culture is pervasive and within the world. The emphasis of information and media leads him to conclude that it is not the network logic which invades the economical and the cultural but the media itself: "Media theory is only possible in an age in which social and cultural life has been pervaded by the media." (Lash, 2002, p. 66). Therefore, when such a thing happens, sociology starts to be a *"mediology"* (Lash, 2002, p. 206).

The media theory, or mediology is the extreme version of the impact of multimedia and electronic communication developed in Castells, the fifth idea of Castells about the convergence of technology and the network logic.

In other words, *technology*, *production*, *socialisation*, and *culture*, every aspect of our lives, are modified by this network/technological/informational perspective, as a result, as a cause, as a product, everything at the same time. This represents one aspect modifying the other aspects. As all the authors I have mentioned claimed, it is not a linear relationship but a co-evolutive process where interactions between different spheres of activity are creating a new reality.

But in science, do we shift from different concepts of reality because this reality changed or because we unveil a new form of socialisation? Then, the question is not which technological invention or economical reality creates new social practices but to what extent our perception is modified, and what are the consequences for our comprehension. In sciences, the paradigm revolution is a concept used by Khun to explain a radical change in sciences when previous models failed to explain some phenomena. After a period of crisis, new models (and/or new instruments) bring a different concept of the problems studied and the way to understand them. These shifts from a previous paradigm to a new one lead to a different concept of the world. The world itself could still be the same but the perception has completely changed:

> Led by a new paradigm, scientists adopt new instruments and look in new places. Even more important, during revolutions scientists see new and different things when looking with familiar instruments in places they have looked before.[…]Nevertheless, paradigm changes do cause scientists to see the world of their research-engagement differently. In so far as their only recourse to that world is through what they see and do, we may want to say that after a revolution scientists are responding to a different world.
>
> (Kuhn, 1970, p. 111)

An adequate methodology needs to be deployed in order, not to understand the difference between a less networked situation and a more flexible sociality, or a new versus old form of interaction, but to understand what is composed and how such sociality is deployed:

> New rules of sociological method are necessitated by the apparently declining powers of national societies since it is they that have historically provided the intellectual and organizational context for sociology
>
> (Urry, 2000)

Instead, as quoted by Sheller and Urry, we are following the idea of a change in science itself to understand phenomena. Rifkin notes that contemporary ’science’ no longer sees anything as *static*, *fixed* and *given* (Sheller & Urry, 2006, p. 193); *rather, apparent hard and fast entities always comprise rapid movement and there is no structure separate from process* (Sheller & Urry, 2006, p. 212). It is a co-evolutive process, on one side, the evolution of society as depicted by Castells and Wittel, and on the other side an evolution of our tools to grasp this evolution. This change will have an effect by giving a pre-conceptualised shape of the studied phenomenon.

> We need to see this shift not as simply a matter of scientific progress or the advance of knowledge, but also as related to the remaking of cultural hierarchies, and the redrawing of class, gender, and national boundaries, crystallized through a mode of knowing the modern rational nation in the image of social scientific methodology which mobilized sociological aggregates.
>
> (Savage, 2010, p. 237)

And yes, the network society produced a different type of knowledge, following the same idea as the technological paradigm. While for the network society, technology allowed the resilience of the network and the network logic to match our world representation within a similar perspective, the databases and the Big Data development sustain another type of available data, while the evolution of privacy, and showing our lives on the web, allow another type of information to become available. These processes force sociologists to rethink the way they create knowledge about society or they are condemned to disappear. This may not be a new scientific paradigm, but it is at least a big change for our comprehension of ourselves. In the past, when all media adapted to the logic of the television, it looked as though all sciences needed to adapt to the logic of computer science. However, this may not be necessary. But to see if an alternative is possible, we need to understand the consequences of the explosion of data and the associated methods.

1.7 Conclusion

This chapter, by introducing some core concepts of the i*nformation age* attempted to locate the latest developments of the web within a larger societal and theoretical evolution. The society evolves as shown, but the theoretical tools also change our perspectives on the issues at hand. Both leading to a more fluid and horizontal perspective that previous perspectives on class and static vision could not grasp.

From Castells, the main concept used here is the word *network*. Not only did I use it in the sense of Castells definition of a network as a social morphology, but I also highlighted the evolution of the concept within other perspectives, the local practices, as developed by Wittel in the continuity of Castells's work on an ethnographic level, and methodological implications, as seen by a brief incursion in Khun's definition of paradigm.

Overall it helped to see how Social Network Sites can be relocated in a media theory to understand the consequences they have on the everyday practices and social interactions and how we can locate them in the larger evolution of the media in general.

However the last section, the methodological conception of the network, tempered the idea of novelty presented by the fluidity and argued that it may be the evolution of science itself that shifts our attention from a class dividing image of the society to a constant changing horizontal society. Beside this inherent reason to limit the aspect of the novelty of fluidity, some limits of media theory can also be voiced.

These previous sections did not bring forth any critiques on the Castells' work and the media theory in general. Mainly because the chapters' purpose was to introduce the concept of network, fluidity and the novelty shown by Social Network Sites as a continuity of the information society. The critiques never really question the idea of an information society as described by Castells and the evolution of mode of production. However some limits of this vision have been highlighted since the publication of Castells' work on the extent of this society and the methodology of his analysis.

First, the idea that we are living in a fluid and changing society is not necessarily something new. The very idea of a revolution in our form of production and socialisation is not something that was not already present (Garnham, 1998; Webster, 1997). The flow and flexibility are more a long term evolution of capitalism, as finance has already shown, than a creation of the information technology (Garnham, 1998, p. 118).

In the same idea, this importance of traditional capitalism *versus* the revolution advocated by Castells can also be tempered. The idea that information takes over traditional economy and mode of production may be a distorted version of the reality. For instance, Fuchs points out the fact that in 2008, even if the information technology industries account for 4.59% of the total assets of the world's largest corporations, the more traditional oil and gas industry account for 6.21%. A proof that the fossil industry still has more power than any other information society, even if for Castells, communication is power (Fuchs, 2009, p. 106).

The world may be more globalised, but millions of people are still outside the loop and not impacted by this new form of socialisation. The consideration of them being excluded in terms of position in the network and not having access to the information age does not bring any theoretical tools to understand their situation (Webster, 2004). On the same idea, it is seems unlikely that the factory worker or the soldier on the ground would agree to the idea of a horizontal society when they can easily decide the course of their action on the field (Fuller & Webster, 2005).

All these arguments, the previous existence of flow, the persistence of the traditional form of production and the limited impact of information society overall on the population, moderate the argument of a revolution brought about by information technology. These arguments may be seen as different interpretations of the phenomenon without really implicating that the Castells' argument is wrong overall but they need to be taken into account when the term information society is used.

Alongside these previous critiques, the methodology is not spared either but may raise more fundamental concerns. Castells is often praised for his sociological work on a macro-scale. He builds his argument on the largest changes in economy and demography and uses numerous sources to prove his theory. But his use of the table and data provided does not apply without problems.

> […] the relative richness of their empirical detail and paucity of language often make it difficult to tell the difference between what is presented as fact, interpretation or explanation – or, for that matter, justification or criticism.
>
> (Fuller & Webster, 2005)

The critique is that Castells may be partisan when he develops his argument because he has spoken to a specific audience, political decision makers rather than academics (Fuller & Hond, 1999). That is linked to the poverty of his language. By heavily relying on the empirical data tables and figures to assess his argument and using the technical language of the already globalised elite, it is hard to see when his argument is either a justification or a criticism (Fuller & Webster, 2005). This argument echoes the problem of defining the word, network, itself:

> Castells's reliance on the overused concept of networks hardly helps to clarify his vision. He stresses informationalism's tendency to reduce social norms to a recognition that several parties may realize their goals by temporarily acting in a concerted fashion, which in turn defines a specific network (1:171 ff). But, stripped of tables and jargon, this sounds like the definition of normativity put forward by the Austrian school of economics, which provides the intellectual foundation of contemporary neoliberalism.
>
> (Fuller & Hond, 1999)

The apparent collusion between the sociologists and the elite he studies, creating a normative argument is not the only aspect of his methodology that raises concerns. His heavy use of tables and data on a macro scale is also criticized due to some loss in the methodology of aggregate data and the lack of direct statistical demonstration of his hypothesis (Fuller & Hond, 1999).

This last critique about the pitfall of the macro-perspective adopted by Castells will be answered in the following chapter with the dissection of what is Big Data, where it comes from and its component, the digital trace. To be able to build a more adapted perspective on the issues raised by Social Network Sites, an introduction of social constructivism and the actor-network theory will serve to develop more adapted tools to the analysis of the social relations in-situ.

2 Transactional Data and Actor-network theory

2.1 Introduction

In the previous chapter, the development of the argument was focused on the evolution of society as more fluid and networked. One main reason for this is definitively technological, as pointed out by Castells.

I carefully tried to avoid any technological determinism, by focusing the analysis on the development of social practices issued from the technological paradigm and the network concept[11].

Here, the perspective is slightly different. It is based on the technical evolution prior to any social evolution. It is not excluding one from the other or stating any prevalence. The separation is simply for the purpose of explanation.

One of the consequences of this technical evolution (or a cause: it is never clear which it is in co-evolutive processes), is the explosion of information. Not only is everything information and communication in a cultural sense (Lash, 2002) but the information itself grows exponentially. This growth has several practical consequences, such as the emergence of new fields or techniques using, and producing, this information. I am going to focus on one specific use of this explosion of digital information with the transactional data as an economic potential for any company to better understand their own working and that of their customers or services users. This development of transactional data is important as its premises and characteristics lead, to some extent, to the vast majority of Big Data analysis, even if the data ceases to be about transaction.. This is why I will describe Big Data based on the practitioner's, rather than theoretical definition. This definition will be used as a yardstick to highlight the differences between data generated in the context of Social Network Sites (SNSs) before completely abandoning the name Big Data, in the following chapter.

The capacity of the world's technology to store information is estimated to be at around $2.9\text{x}10^{20}$ optimally compressed bytes (Hilbert & López, 2011). For digital

[11] However, as pointed in the conclusion of the previous chapter, Castells is not free of some forms of techno-determinism.

information, some researchers estimate the production to be around 13.2 zeta bytes per year (Makarenko, 2011). These numbers provide only an estimate, but this clearly indicates that we are producing far more information.

This increase in production is possible and it is needed by technology itself. It is impossible to do anything online without leaving a trace of the activity. Every time we go onto a website, the website shares information (with cookies) on our computer, interacts with our browser and any request we make, and stores this information. This is, in a sense, the passive version of information increase; a by-product of the technique. However, this is not only the sole explanation.

We can see an economic perspective, where the production of information is a part of a strategy in which there is a need to enhance the knowledge companies have about their environment, their customers and their competitors. Social practice also participates in this explosion of information when people produce more information about themselves on the web by writing blogs, updating Facebook statuses, tweeting about a political question, or uploading a picture of the not-so-interesting meal we had in a restaurant, onto Instagram.

It is obvious that these three reasons, the *technological*, the *economic* and the *social* are artificial distinctions and it is possible to redefine them differently or find other sources of data production[12]. But I will stick to this arbitrary separation in order to develop the idea of the evolution of data, produced for different purposes, and within different contexts, in order to show how different they can be. I am going to start with the transactional data and the production, storage and analysis of data in a controlled environment before understanding this same production in a less structured context, the Social Network Sites.

Do not be misled, I know that SNS are themselves created in order to generate transactional data to ensure specific business models[13]. But we are going to see they are different as the purpose of data production itself differs, as well as the context where it is produced and more important the control over this production. These

[12] Such as science, for instance the LHC creates a petabyte of data per second by taking 40 million snapshots of protons collision.

[13] For instance, Facebook introduced on April 2013 a new structured method to update status which allowed users to select different emotions, from a list, associated with the status, and link activities to objects (such as watching movies or reading books). In other words, they extended the now traditional "*like*" to more qualitative options. This is not using a pen and paper, but the idea is similar to the Tesco model, extending the information they can analyse and treat to bring a more complex picture of their customers (Raylene, 2013).

differences lead to another reality which doesn't necessarily occur in transactional data: *the social complexity*.

2.2 Data complexity and the 3V definition

This use of transactional data has grown enormously, recently, with the web and the facility to collect data on every customers' actions (E. Apeh & Gabrys, 2011; E. T. Apeh, Gabrys, & Schierz, 2011; Büchner & Mulvenna, 1998; Srivastava, Cooley, Deshpande, & Tan, 2000). From the already large number of 10 million Tesco Clubcards (Rowley, 2007), we are now in a situation where a company like Amazon can produce an individually customised webpage for every one of their users (Linden, Smith, & York, 2003). This phenomenon, also known as *Big Data*, is not necessarily new. In the late 80s, the whole US census was stored in 100GB and was considered to be a Big Data 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 is characterised by a huge amount of information where issues such as storing, computing and analysing data cannot be answered by traditional dataset tools or traditional statistical software. Therefore, the definition is not bounded by a typical size limit but it 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 issue from a computer science perspective and uses the 3Vs: volume, variety and velocity (Stapleton, 2011).

2.2.1. Data Volume

This unprecedented quantity of data requires new approaches to deal with it, because it is impossible to treat the data with traditional algorithms. Even the methods for storing the data has changed e.g. the big social web-services such as Facebook and Twitter tend to more often use NoSQL database types allowing a more scalable solution, a more efficient approach towards the exponential growth of information. This represents one kind of integration of computer science and marketing, and is the first step in the creation of Big Data as we currently understand it, simultaneously providing the first difficulty. Collecting and storing massive amounts of information is not the same problem as recalling small amounts of information. In fact, finding what to study within Big Data, and how to study it, are two sides of the same coin, and the answer is found in the same place where the issue was created: the computer.

With computers, one statistical technique has become feasible and has been used broadly, from gene analysis to marketing studies: the clustering methods (Kaufman & Rousseeuw, 1990). The main purpose of this technique is to group similar items together and answer the question of volume in Big Data. Then, the remaining question is to develop more efficient algorithms in order to treat more and more data (Ester, Kriegel, Sander, & Xu, 1996; R. Xu & Donald Wunsch, 2005). But clustering is only the first step in the analysis when dealing with transactional data. Before it is even applied, several effects of the volume of data need to be addressed, such as noisy data (Ertöz, Steinbach, & Kumar, 2003), long tail effects with power law distributions (Hsu, Chung, & Huang, 2004; Park & Tuzhilin, 2008) or missing data (Honda & Ichihashi, 2004). However, the data also comes in a less practical format and several steps are needed before we are able to compute it and extract useful information - data reduction. This is because the datasets are usually too large to be analysed and require the removal of some useless information. This is known as data cleansing, which is the way to deal with missing and inconsistent data. We also need to perform data integration, which means combining several datasets and, finally, data transformation, which refers to normalizing the data to produce a more efficient analysis (E. Apeh & Gabrys, 2011).

2.2.2. Data Velocity

The constant renewal and flow of this data also creates a great difficulty in storing and treating it in real time, or at least in a short enough period of time to still be useful. The main consequence of the velocity aspect of digital information is that it adds non-negligible difficulties. The algorithms need to be scalable and very efficient to process massive data.

Another issue relates to analysis: to decide, within this fast flow, what information will be treated and when to ensure optimisation, not only of the algorithms but also of the selection of relevant data. This point is developed later. The algorithms need to analyse vast quantities of data but without the traditional delay. The importance of the speed of execution covers two different requirements. Firstly, people (especially in business) need to have data as quickly as possible. Secondly, and more importantly, the stream of data often lacks memory, meaning that the information which is not processed in real time is lost.

2.2.3. Data Variety

Data can take heterogeneous forms. It can be numbers, words, sentences, pictures, addresses, and so on. It can also come in different formats, which is more or less easy to store and analyse. However, ultimately, from a Big Data perspective, there is the need to reduce and fit any information into rows and columns. This format is needed for the Database, not only for the database structure but also because of the need to analyse quantifiable measures. And it is this translation which is problematic, yet crucial, as it has a later impact for the translation of human behaviour into a manageable form. 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* that is the difficulty in recalling and analysing vast and different amounts of information in real time.

2.3 Transactional Data

2.3.1. Introduction

Before going further, we need to define what *transactional data* means. *Transactional data* recovers the information a company or an administration holds on all customer transactions. At the beginning it was produced by banks to give them information and tracking all the transactions. It was a necessity for the bank system to work. Later, administration also used it to keep a track of everything going on, and finally commercial companies developed their own use of transactional data through the Business Intelligence paradigm (Luhn, 1958). Marketing was the last stage of the evolution which resulted with the development of the *Customer Relation Management* (CRM) (Rygielski, Wang, & Yen, 2002). On the technical side, the development of relational databases was a big step forward in this system, allowing the storage and manipulation of data in a more efficient way (Codd, 1970). Collecting all this information is effortless as databases are designed to collect and store information in a transactional fashion (Jacobs, 2009).

Moreover, with the *democratisation* of the storage technologies and computers in general, companies collect and store vast quantities of information about anything and everything.

The leitmotiv seems to be the more, the merrier, even if they don’t know how to use the data stored (Callan & Teasdale, 1999). On the web, the tendency is even more extreme, when customers view their webpages in order (or not) to buy items, several types of information can be displayed: the customers are able to know how many items are available and how long it will take to receive them, their prices, which other similar items other people viewed, the evolution of the price over time, and the cost of the delivery based on their geolocation information. All this information is based on the knowledge the company has about itself, its logistics and comparisons of customers' needs. If the customers buy an item, the website has to display which items they are actually buying, in what quantity, which price and from where it has to be dispatched. But all this information is only the visible tip of the iceberg. More information is recorded than needed for the service itself. We can think about which pages the clients are visiting, how long they are staying on them, the items they are looking at and, when they made their last visit, and more.

We can see these two technological particularities, designs of databases and democratisation of computer prices as the first motors of Big Data emergence. However, technology is nothing without use and application. That is the domain of marketing and Business Intelligence who really capitalised on this technological particularism into a new form of data use. Since then[14], Big Data stopped being a side-product of logistics and technology to become a goal in itself.

To illustrate this development from side-product to an ongoing process in itself. I will describe the example of the *Clubcard* and *Tesco* as one of the earliest and successful applications of that logic.

2.3.2. A pro-active approach: Tesco and the loyalty card

It seems obvious now that companies such as Facebook and Google base their entire business on the knowledge they can extract from their customers, but it wasn't so obvious in 1995 when Tesco launched its ClubCard and started to take advantage of all the transactional data.

This production and use of transactional data was not the first application of transactional data, nor was it the first time a loyalty scheme was installed with the

[14] The earlier article talking about Business Intelligence founded for this work (this is not to say it is the earliest one) is from 1958. The argument developed here states that the relational database helped to discover the importance of Business Intelligence and the use of data. Obviously, there is a chronological issue in the argument. However, as for any application of technology, the appearance of the database, and the wide spread use of it, are two different things. Moreover, the logic inherent in business intelligence is still the same. The evolution of technological possibilities and practices are not part of a linear process and can evolve at a different speed and level.

aim of rewarding customers[15]. But this example is one of the most representative of the application of the loyalty card and can show us two main characteristics: the production of data into a controlled environment and the arbitrary definition of relevant data. Even if this example is not directly related to the production of data on the web it still has the same properties in terms of the analysis of data complexity, how it is generated and how it is solved.

Humby et al. gives a good account of the deployment of this ClubCard (as they are part of the project) (Humby et al., 2008). From this historical description I am more interested in the steps leading to the collection, creation and analysis of the data, as the objective of this section is to see, through an example, the pro-active creation of transactional data and the consequence of such production into this specific economic environment.

Initially (and it is still the case), the ClubCard aimed to retain customers by rewarding them with three distinct mechanisms:

- Pure loyalty: *"Strengthening the existing bond between the customer and the retailer, so the retailer can find out what the customer wants, and give that customer more of it"* (Humby et al., 2008, p. 10).
- Pull loyalty: *"Attracting customers by augmenting a retail offer, so customers will find that buying one product means they get an offer on another, linked product"* (Humby et al., 2008, p. 11).
- Push loyalty: *"creating a scheme to encourage us to use a way of shopping that we would not have done before - pushing customers through new channels, or trying to create new types of behaviour"* (Humby et al., 2008, p. 11).

However, it rapidly appears that the main advantage of this card was the amount of information it provides and making a better and more adapted relationship possible with the customers, achieving the advantage over competitors really to the point that some researchers said the loyalty rewards are themselves useless in comparison of the advantage of the knowledge (Rowley, 2007; Stone et al., 2004).

[15] The Co-op organisation did give a dividend back to its customers, decades ago in UK. Also loyalty schemes were common in US.

To be able to generate data, they had to install new systems to collect information. One called the EPOS (Electronic Point of Sale), a magnetic reader to collect all the transactions from every customer (where, when, which price). But another important type of information was missing; the name and the address of these customers (Humby et al., 2008).

To collect (generate) this information, the pen and paper technology was first used in asking for the addresses and names of people when they asked for the ClubCard. Then, this information, collected through an old-fashion method, was stored in the magnetic card, and the EPOS was able to generate new combined information between the *"where, when, which price"* and the *"who, from where"*.

Therefore, an accurate collect of information was possible and generated in the first three months, 50 million shopping trips (containing 2 billion purchased items from 5 million ClubCards)[16].This vast source of information, generated by customers had the potential to transform the relationships the retailer could have with its customers:

> [It is] equivalent of a continuous customers survey based on a panel of 5.5 millions actives customers
>
> (Humby et al., 2008, p. 74)

However, this first translation of each cardholder into *"a complete record [where] each shopping trip will be updated"* (Humby et al., 2008, p. 98) is only potential knowledge until a proper analysis can make sense of the data. In 1995, storing huge amount of data was possible, but resulted in too great a volume to be analysed in time[17].

However, they managed to analyse a small section of data initially. First they sorted customers into small samples. They used 1% of their customers during a year and then shifted to 10 % of the total customers. They also reduced the number of items and took into account only the items which were responsible for 90 % of the sales (reducing the number from 45000 items to 8500 items).

From this first reduction of data, they created 80 buckets defined as *"of products that appeared from the make-up of a customer's regular shopping baskets"* (Rowley,

[16] Later, there were 10 million cardholders, creating 1.600 million new data items each month ('Galaxy Zoo', n.d.).

[17] Time is one of the most important problems in data warehouse alongside the format, the scale, the quality, the cost, the culture and the corporate ego (2012).

2007, p. 371). Then, they wanted to apply this 80 dimensions clusters on 10 million customers, but the problem of time processing (or software and hardware upgrade costs) did not make it possible. Therefore they applied the cluster analysis on the first sample and re-applied it to the full dataset.

But, with this creation of clusters, they missed the reason *why* these customers can be grouped together. From their perspective it was important to understand the reason why people bought specific items together to personify the relation with these groups.

Therefore, they created *Tesco Lifestyle* segmentation, a 27 clusters for which they added the explanation needed:

> As a retailer Tesco found it a lot easier to picture the individuals who made up groupings like High Spending Superstore Families, and go from there to think of ways to encourage them to consolidate more of their spend at Tesco. It's a grouping that has meaning. You can picture them. It helps managers in the business think about helping real people, with real-life needs.

(Humby et al., 2008, p. 136)

Later, they improved their division of their customers to answer an issue they were facing with the bucket methods (some overlapping and indistinct clusters) but also because they finally had the technological capacity to record every customer's transaction and analyse it.

This last evolution is the *Rolling Ball* method. This improvement followed the combining of products. They added descriptive information about the customers to understand the link between the *what* and the *why*, offering a better understanding and a detailed picture of the customer segments and their buying habits (Erickson, Ph, & Rothberg, 2005).

Not only did they improve their methods of collection and analysis, they also improved their methods for measuring their marketing strategy. It was essential to have a suitable "*why*" so they changed the way they assessed the comprehension of their population with new ways of aggregating and making sense of the segmentation itself.

Up to the lifestyle segmentation method, they used the RFV analysis to assess their method. This acronym represents the *Recency*, *Frequency*, and *Value* criteria which are defined as follows:

- *Recency*: A simple measurement of when the customer last shopped. It is a good measure for discovering customer defection, but this measurement is not good enough alone. (Humby et al., 2008, p. 95).

- *Frequency*: How often the customer shops. It is *"How robust the relationship is between the customer and the brand"* (Humby et al., 2008, p. 96).

- *Value*: The value of a shopping trip.

Later they modified the RVF analysis to develop their own model to assess the success of the loyalty scheme: *the loyalty cube*. This measurement is essential for adding more granularity and the possibility of understanding customers' behaviour. It is a 3-way, interconnected, dimensional space with:

- *Contribution*: Indicates the degree of profitability that the customer represents for the company. It is not a direct measure of loyalty, but a measure of the present relationship between the customer and the company.

- *Commitment*: Indicates the future value of a customer and is contained in two elements: The likelihood of a customer remaining a Tesco customer; the customer's future loyalty, and the concept of *"headroom"*. This represents the potential for a customer to be more valuable in the future. If the value is small, it means they already buy the maximum. If the value is big, it means the shop can try to sell them more items.

- *Championing*: Customers can talk about Tesco to others and have an impact on the company. They can talk about the Loyalty scheme and have an impact on the diffusion of information by mentoring their friends and family connections and enrolling them (Humby et al., 2008, p. 129). It is also a measure of potential: customers with a low value, for instance, can have a long-term impact through this behaviour.

Finally they can build the *"why"* information that is needed but missing from the different clusters. One of the best examples of this implementation is when Wal-Mart tried to implement their shops in the United Kingdom.

Immediately afterwards, Tesco analysed their customers behaviours to find out which products were the cheapest that they purchased; subsequently they targeted Wal-Mart consumers by decreasing the price on almost 300 items (Rohweder, 2006).

This successful use of data into an actionable marketing practice is only the tip of the iceberg. Since the introduction of the Tesco card (1994), the situation completely changed and the production of data but more importantly, the analysis of these data has reshaped almost in every field, not only in Business.

The culminated version of these transformations of science by the profusion of data can be found in the 4th Paradigm; the idea of a scientific revolution that allows new discoveries and the emergence of new fields. I will briefly outline what the 4th paradigm is to be able to contextualise the impact of data creation into science, before explaining why, the very existence of a 4th paradigm, and the hopes associated with it, calls for a specific approach in social sciences.

2.3.3.4th Paradigm, positivism and Actor-network theory

The 4th paradigm, a term coined in 2007 by Jim Gray (Bell, Hey, & Szalay, 2009; Hey et al., 2009), or the data-intensive science, is the idea that it is possible to develop a new epistemological approach with the quantity of data produced and the associated technological developments.

The term highlights the difference between the three previous science paradigms, *experimental*, *theoretical*, and *simulation*:

> Originally, there was just experimental science and then there was theoretical science, with Kepler's Laws, Newton's Laws of Motion, Maxwell's equations and so on. [...] the theoretical models grew too complicated to solve analytically and people had to start simulating. [...]. At this point, these simulations are generating a whole lot of data, along with a huge increase in data from the experimental sciences.

> The new model is for the data to be captured by instruments or generated by simulations before being processed by software and for the resulting information or knowledge to be stored in computers. Scientists only get to look at their data fairly late in this pipeline. The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science as a new, fourth paradigm for scientific exploration.
>
> (Hey et al., 2009)

The data-intensive science is seen as the last episode of an increasing complexity in science, made possible by the advance of knowledge and technologies. One previous state creating the conditions for the following state to emerge. At the end, the scientists look only at the data generated by a complex set of powerful tools, and do not, either experiment directly, theorise or simulate, but fish into the dataset generated. They are at the end of the complex pipeline and never have to touch, see or manipulate their object of inquiry anymore.

Hey et al. use the example of the telescope in their introduction to exemplify the process. The observations are made through a complex network of powerful telescopes, the data gathered are then stored in a grid of super computers and the astronomers manipulate the data behind their computers, in their office or biologists only using computer to study molecular biology (Bell et al., 2009; MacMullen & Denn, 2005). In fact a lot of massive projects in natural sciences are *de facto* data intensive sciences, such as the LHC, or the current monitoring of the global climate change requiring a real-time monitoring from a network of data. This network and all the associated satellites, sensors, double the quantity of information every two years (H.-D. Guo, Zhang, & Zhu, 2015).

Generating a vast quantity of numbers is not new in Business with the transactional perspective, neither in natural sciences. However in social sciences (excluding the economy) it is rather unusual. The surveys, case studies, interviews and all methodologies deployed in the field cannot compete with any data-intensive natural science project.

Although, the web opens the possibility, not only to create and access data on people's economic transactions but also on people's interactions, movements, feelings, identities, through the data collection on Social Network Sites (Chen et al., 2014). For instance, Twitter, generates (or generated in 2014) 500 million messages per day (Twitter, 2014). These messages are, and this is the real novelty, an interaction between people. This is where social science, and sociology in particular, can see a massive interest. Not necessarily the idea of having access to a vast quantity of information, but that information is about social interactions.

Indeed, social science embraces this *revolution* with the same enthusiasm, with the same hopes as in the past: numbers will make it possible to understand, categorise, predict any social interactions with much more precision, accuracy than before.

Some commentators have exaggerated this idea of the 4th paradigm and imagined that the experimentation, formulation of hypothesis, or any input from the scientists would be irrelevant. The only thing that a scientist would have to do it is to point to the dataset and let the *algorithms* create information and knowledge (Anderson, 2008; Prensky, 2009; Steadman, 2013).

2.3.4. The idealistic approach of the 4th paradigm and the social constructivism

However, aside these voluntary excessive views, the more common accepted idea in natural science is not a radical perspective as such but more a combination of a traditional scientific work (Liew et al., 2016; Shah, Cappella, & Neuman, 2015) and a constant shift between paradigms (Kitchin, 2014). The human is even put back in the loop to identify and lead the process, even with the latest machine learning process used in the Integrating Advanced LIGO Detector. In their description, the authors show how it is primordial to associate what they call *citizen science*[18], when massive datasets are used. It helps to overcome the limit of any computable and automated process (Zevin et al., 2016). Under this conception, the fourth paradigm is another tool added to the toolkit of science. But even with these clarifications about

[18] Citizen science can be defined as: is a volunteer who collects and/or processes data as part of a scientific enquiry (Thomas, Grier, Song, & Paxson, 2011). This inclusion of citizen science is often use in combination of massive dataset, such as the ones created in astronomy and astrophysics. One of the precursor and most successful application of that idea is the Galaxy Zoo, launched in 2007 (Daries et al., 2014; Heffetz & Ligett, 2014; Sweeney, 2002). This project aims to recruit people from the public to classify images captured by the telescopes, several peer reviewed publications came from that project, as well as the discovery of new type of astronomical object (Sayes, 2014).

the feasibility of the automated processes using Big data, or the inclusion of previous paradigms as the integrated part of the current workflows of research, the use of Big Data still participates to a form of *positivism*.

Here, it is not the version thinking that science is fundamentally good and the progress that it brings will inevitably enlighten humanity. This positivism is a more persistent one and has been something we can find especially in natural sciences and even more in technological and business areas which advocate one vision of science and technology. Several definitions of positivism can be found in the literature with more or less refinements and complexity, but here is a clear and simple definition:

> In the positivism paradigm, the object of study is independent of researchers; knowledge is discovered and verified through direct observations or measurements of phenomena; facts are established by taking apart a phenomenon to examine its component parts.
>
> (Krauss & Putra, 2005, p. 759)

The idea is that the *universe*, the *nature*, the *phenomena,* has an accessible version of it. It exists as a reality out-there, a truth, that science will eventually find out by applying logic and rigorous methods. Or the same argument turned in the other way, whatever science finds and says, is true (under the assumption it followed the appropriate methodology).

This positivism also exists in sociology and is linked to the inclusion of statistics and quantitative methods in the field. When Quetelet associates the random and fixed traits of an *average man* and developed his *physique sociale*, he hoped to be able to predict all human variety and being able to predict its future based on the use of probability. Durkheim and his work introducing statistics to study the suicide is another famous example of a positivism perspective in sociology. Not only, did he use the statistics and promote an idea of quantitative methods in sociology, but he also introduced the notion of *fait social* and established this notion as a fundamental rule in his sociological methodology*: La première règle et la plus fondamentale est de considérer les faits sociaux comme des choses* (Emile Durkheim, 1894, p. 21).

The consequence of such perspective is to consider the social as a natural object that sociology can only study by applying a rigorous quantitative approach by aggregating all individuals' actions into a *fait social.*

> Since the law of causality has been verified in the other domains of nature and has progressively extended its authority from the physical and chemical world to the biological world, and from the latter to the psychological world, one may justifiably grant that it is likewise true of the social world.
>
> (Émile Durkheim, Lukes, & Halls, 1982, p. 159)

Durkheim claims that the social world is no different than the physical, the natural world. It is possible to apply causality and for him, research using quantitative studies tends to prove that. That idea perpetrated through the 20th century with the extensive use of surveys and to some extent, proved him right. These quantitative methodologies did revolutionise the field and allowed to bring sociology up to an accepted level of recognition. Simultaneously, sociology brought methodologies and new tools to the statistics field, carried by the extensive and developed use of survey methods, by researchers but also by governments and private companies (Clogg, 1992).

However, since these changes in sociology, several critiques towards the use of statistics and the positivism in general have been voiced. Aside the inherent critiques against the statistics used in the case of suicide studies (Pescosolido & Mendelsohn, 1986), more general concerns about the liability and the power of the statistic in sociology also raised some concerns. It is not so much about the usefulness of it, but more about its proclaimed capacity to answers all the questions that society raises and the idea that scientific method as developed in natural science is the only adequate method to study human interactions. The best illustration of this importance of quantitative methods in sociology comes from Fisher, a biologist who founded the modern statistics:

> Statistical methods are essential to social studies, and it is principally by the aid of such methods that these studies may be raised to the rank of sciences.
>
> (Fisher, 1958, p. 2)

The idea was that, only numbers and the scientific methods could works, and that science, in all its positivism, will bring the answers and it is the only acceptable method. As we can see, the link between quantitative methods in sociology and the realm of science is not new. It did not wait for the 4th paradigm to develop the exact same discourse. The Big Data and the 4th paradigm are more of a *bis repetita,* more data, and more powerful analysis as the golden path for sociology to overcome the disappointment of the statistics, which could not hold the exact same promises.

But the new status of positivism with Big Data, does not only come from the enthusiasm of the Big Data participant but from the very nature of the Big Data itself and the 3V outlined earlier. The complexity generated by the 3V imposes restrictions on what it is possible to do and what is not. Therefore, methodologically, the constraints favour some approaches over others:

> As with earlier critiques of quantitative and positivist social sciences, computational social sciences are taken to task by post-positivists as being mechanistic, atomizing, and parochial, reducing diverse individuals and complex, multidimensional social structures to mere data points […]
>
> (Kitchin, 2014, p. 8)

From Kitchin, there is also a repetition of history, but he centred his critique towards the tension between the absence of data complexity and the complexity of the human interactions.

Reducing the complexity to a *mere data point* is one cost of the 4th paradigm applied in social sciences. We have to be aware of this discrepancy between a technologist approach to human interactions, because it involves a massive dataset, and more sociological approaches on these same human interactions to try to render the complexity of it.

The solution is to take a step back from the idea that the data represents some inherent and unfiltered quality of the world and the analysis applied on them will reveal the true face and the only possible vision of the phenomena studied. In other words, avoiding this post-positivism approach mentioned by Kitchin.

To this effect, sociology did not only debate about social interactions, but also how knowledge is generated, how science creates truth and how some objects can be transformed into social phenomenon. This questioning of how science generates knowledge has been developed extensively with the science studies about other subjects than social sciences.

Sociology has developed strong fields over the last century to de-construct the idea of positivism, *sociology of knowledge*, *science and technology studies*, *actor-network* *theory*, these fields are devoted (entirely or partially) to the study of how knowledge is generated and incidentally, how science creates its own knowledge.

One fundamental concept that underlies these approaches is the idea that knowledge in general, and science in particularly, even if it is about natural events, is *socially constructed*. This conception, without being opposite to a positivism perspective, brings some nuances that can help in the present context but need to be introduced to avoid some confusion.

The *social construction, or social constructionism* is a concept established by Berger and Luckmann's in their book *The Social Construction of Reality*[19] *(Berger & Luckmann, 1991)*. Their work, primarily concentrated on the sociology of knowledge, provides a theory on how social institutions are formed but also maintained. For them, the reality exists both as objective and subjective:

> As an objective reality, society presupposes habitual and meaningful actions which have become typified, and thus institutionalized. An institutional world is experienced as both legitimate, and having an objective reality that originates from "the typification of one's own and others' performances" (Berger and Luckmann 1966: 89). The institutional world's existence is however but a small part of a common knowledge of sedimented and more or less constant meanings that are transmitted and maintained from one generation to the following ones.
>
> (Segre, 2016, p. 96)

Then, people can experiment with this reality through the others, with socialisation and language. This reality is then internalised, and seems obvious for the individuals, with no need of questioning their existence. They applied their approach on social institutions only, but as pointed out by Lynch, the ontological perspective of their concept inherently called for an application to science too:

> Using their approach thus requires us to make a judgment about which phenomena are socially constructed and which are not, a judgment that is particularly difficult to make when faced with phenomena that are authoritatively presented in our own society as "natural".
>
> (Lynch, 2016)

[19] The first publication was in 1966.

As soon as the box of questioning phenomena that seems *natural* is open, there is no real reason why it should be limited to society only or to *bad science.* This is what happened later. Before, sociology of knowledge made a distinction between what was true and what was false, between beliefs and Nature, or *Social* and *Nature*. The underlined assumption was that a natural truth exists and therefore it was possible to study with science. As the methods in science were devoted to unveil this truth, when it was carried out under the right methods and practices, it was beyond questioning. Only what was possibly seen as wrong from an external observer could be considered as social constructed.

The use of the *social constructivism* within the study of natural sciences came with the *strong program* started by Bloor (1977). With that program, the sociology of knowledge applied the *symmetry principle* and that which can explain errors can also explain the production of truth.

Since the development of this *symmetry* and the application of social constructivism to the natural sciences, a lot of confusion (and even a so-called *science war*[20]) have been raised. Stating that science can be constructed through social relationships rather than being an understanding of the reality may seem off limits for some people. However, the social constructivism is less controversial than what can be expected.

The idea is *reality is socially defined but this reality refers to the subjective experience of everyday life, how the world is understood rather than to the objective reality of the natural world (Andrews, 2012, p. 40)*. This is not saying that there is no natural world, but that we construct the concept of it rather than discovering it.

This construction, as based on a reality, a locality, is different than a relativist claim. It does not remove the physical world for a fully socially constructed science.

> Social constructionism in science and technology studies is neither relativist, reductionist, nor naively deterministic. The focus on humans as quintessentially social does not eliminate the idea of humans as biological or thermodynamic systems. Stressing the social does not erase the various physical and natural

[20] An interesting episode for the critical studies and sociology in general. A useful website (accessible on the 22/05/2017) links to different resources about that event http://www.math.tohoku.ac.jp/~kuroki/Sokal/.

> substrata on which the social interactively nests. Genes and neurons are clearly part of the larger picture. It may be that the difficulties these ideas pose for us are rooted in a wrongheaded system of categories and classifications that separate mind, body, brain, and social order.
>
> (Restivo & Croissant, 2008, p. 225)

If we accept the idea that the science, or the comprehension of the Nature that we have, is socially constructed, and that construction is about the concept, the representation we have of it, it is possible to question, or to understand how knowledge is produced in a more complex way than assuming it is only unveiling an underlying, and always there, truth.

This idea then has been developed, not only in *sociology of knowledge* but was also a core concept for the *social construction of technology (SCOT)* introduced by Bijker and Pinch (1987) as well as the *sociology of sciences and technology.* In these disciplines, sociologists, rather than concentrate their attention to what scientists and engineers say, are interested on what they are actually doing. These fields have been prolific in the last few decades. One piece of work, among all that has been produced during the last few decades, that applies social constructivism is the Latour and Woolgar's ethnographic work with the *Laboratory life: The construction of scientific facts* (Latour & Woolgar, 1986). In this work, the authors depict the work of the scientists as if it was any tribe or any culture that an ethnologist will study.

2.3.5. The generalised symmetry as evolution of social constructivism

They investigated the work in the laboratory and how the creation of fact is far from the idea of a clean and perfect research but rather closer to a messy work. The clean and perfect vision of science conveyed by paper, publications and conference is often far from the work itself where data and experiences are often inconclusive and how much of the scientific work is about choosing and making the data *speak* in order to publish. This work, even if it does not mention the word *actor-network theory (ANT)* introduces many of the concepts that are later used in what can be seen as one evolution of sociology of science and technology.

Actor-network theory is taking its origin in the science studies:

> ANT has been developed by students of science and technology, and its claim is that it is utterly impossible to understand what holds society together without reinjecting in its fabric the facts manufactured by natural and social sciences and the artefacts designed by engineers. As a second approximation, ANT is thus the claim that the only way to achieve this reinjection of things into our understanding of social fabrics is through a network-like and social ontology theory.
>
> (Latour, 1996, p. 370)

This idea of relational materiality is similar to the social constructivism. And Latour and Woolgar's work on laboratory life was, in some way, a work of constructivists. However, while at first glance it seems that actor-network theory is a refinement of social constructivism, both approaches are different and the evolution of ANT, since Laborary life, puts the authors on a more critical approach.

> When we say that a fact is constructed, we simply mean that we account for the solid objective reality by mobilizing various entities whose assemblage could fail; 'social constructivism' means, on the other hand, that we *replace* what this reality is made of with some *other stuff*, the social in which it is 'really' built.
>
> (Latour, 2005, p. 91)

What poses a problem for Latour is the idea that explaining Nature with the Social through *social constructivism* is only another way to reproduce error that characterises modernity. It is either the *Nature* that explains the truth (in case of positivism) or the *Society* that explains truth and falsehood (in case of social constructivism). This is an impossible position for Latour, both Society and Nature need to be explained and neither of them can take a predominant role in the type of explanation given.

> But Society, as we know, is no less constructed than Nature, since it is the dual result of one single *stabilization process.* For each state of Nature there exists a corresponding state of society. If we are to be realist in the one case, we have to be realist in the other; if we are constructivist in one instance, then we have to be constructivist for both.

(Latour, 1993, pp. 94–95)

For Latour, the problem with the conception of sociology of knowledge and the idea of constructivism applied to *Nature* or to *Society*, to *belief* or to *science*, is the manifestation of an asymmetrical position of the sociologist/anthropologist. This asymmetry has been problematic since modernity.

The asymmetrical position conceives either one of the pole (Nature or Society) as the source of explanation for both of them. But for ANT, both aspects should be understood at the same time and not one being an anchor to explain the second. We have to use the same explanation for any of them and not having a prior judgment of what can be defined as Natural or Social before even studying it.

> To talk of hybrids entails more than a simple mixing of the two opposite poles, for these come at the end of the production process; they are forged out of the heterogeneous materials used in both manufacture and represent Nature and Society.

(Murdoch, 1997, p. 744)

The *quasi-objects* are at the same time human and nonhuman (e.g. artefacts, organisations, structures,…) and traditional perspectives such as social constructivism will fail to explain them because, rather than unveil their interconnections and their association, make a work of *purification.* This work of purification either suppresses the object, stating it counts for nothing (sociology) or that the object is everything (social constructivism) (Latour, 1993, p. 53).

This idea of refusing the dualism between human/non human, has been conceptualised under the *principle of generalized symmetry*. The concept has been firstly developed[21] by Callon:

> It is similar to D. Bloor's principle of symmetry but is considerably extended. The goal is not only to explain conflicting viewpoints and arguments in a scientific or technological controversy in the same terms. We know that the ingredients of controversies are a mixture of considerations concerning both Society and Nature. For this reason we require the observer to use a single

[21] There are not the first one to include nonhuman. As reported by Sayes about the manufacture device in Marx theory (2012) or the already existing critic of Blondel toward the *personalism* reported by Bencherki (Jahr, 2016).

> repertoire when they are described. The vocabulary chosen for these descriptions and explanations can be left to the discretion of the observer. [...] given the principle of generalized symmetry, the rule which we must respect is not to change registers when we move from the technical to the social aspects of the problem studied.

(Callon, 1986)

This *generalized symmetry* helps to locate the sociologist/anthropologist on the median point between the two poles of Nature and Society in a more fruitful way than social constructivism:

> He is not allowed to use external reality to explain society, or to use power games to account for what shapes external reality. In the same way, he is of course forbidden to alternate natural realism and sociological realism by using 'not only' Nature 'but also' Society, in order to keep the two original asymmetries even while concealing the weaknesses of the one under those of the other (Latour, 1987).

(Latour, 1993, p. 96)

By including human and nonhuman on a same ontological level, Latour and Callon, with the generalised symmetry, take a third approach and give them agency and consider them on the same level as human. This is one of the most controversial ANT positions but it does not stop Latour making provocative claims:

> I suddenly understood that the nonhuman characters had their own adventures that we could track, so long as we abandoned the illusion that they were ontologically different from the human characters. The only thing that counted was their agency, their power to act, and the diverse figurations they were given.

(Latour, 2013, p. 291)

Behind this assertion that seems to cut any possible compromise between different theories, the conception of agency is more complex and probably more chaotic, even for proclaimed supporters of ANT (Sayes, 2014). The agency is not necessarily what expected, and if the previous quote is not taken too literally it is possible to conceive it within ANT with more nuance.

First, agency should not be understood as its full extend. Not only ANT refuses the dichotomy distinguishing human to object, it also recuses the dichotomy between agency and structure. To these dichotomy it offers a different conceptualisation of the action. An object will not take an action by itself, or have any wish or desire to do something or something else. The action is never taken alone, disconnected from others or other object but they participate to any action in the world (Bencherki, 2012). The participation of object in the world of human is heavily linked to this specific notion of *action* within ANT:

> Action is *overtaken* or, as one Swedish friend transcribed this dangerous Hegelian expression, action is *other-taken*. So it is taken up by others and shared with the masses. It is mysteriously carried out and at the same time distributed to others. We are not alone in the world. 'We', like 'I', is a wasp's net; as the poet Rimbaud wrote:'*Je est un autre'*.

(Latour, 2005, p. 45)

The actions are rarely taken by either human or nonhuman but often by both, it is almost never a human-to-human connection or object-to-object but often a process that passes through a mix of both (Latour, 2005, p. 75).

The theory of action in ANT is viewed as the process of interconnected *mediators*. The objects, the nonhumans, transport the action, as well as the humans and this transportation through the association is called *mediators.* A succinct definition, or effect they have is that they transform the meaning of the element they carry (Latour, 2005, p. 39), they are not passive. However, they are connected to the action, bringing meaning to it and transforming it. The mediators are often opposed to the *intermediaries,* their counterpart. An *intermediary* will not change the action or the meaning, it is a black box that will not impact anything and will not give any useful information. ANT considers the *social* or the *Nature* in sociology as *intermediaries* and therefore not giving any explanation, just replacing what is happening by a black box. But if you open the black box, then a full network of mediators can be unveiled. Rather than explaining a phenomenon with Society, Power, etc, it needs to unveil how the action is deployed by the human and the non human.

The mediators in ANT are the backbones of the approach. They are conceived as the real interest for ANT and what is make our world complicated rather than complex.

But how the action is carried and the meaning transformed by objects does not mean that they are *causing* or *doing* but they are creating more mediators and make mediators *to do* thing (Latour, 2005, p. 217).

> This is why it is only when we forget all the costly work that is necessary for the production and sustenance of even the most meager nonhuman agent that we could even pretend that nonhumans inherently have agency. Simply put, nonhumans do not have agency by themselves, if only because they are never by themselves.
>
> (Sayes, 2014, p. 144)

The agency in ANT is therefore different than a conscience of willingness but it is only the case because the theory of action and the introduction of mediators transform the ontological definition of what social is.

Second way to understand the full extend of the inclusion of nonhuman (Law, 1999, p. 3) and why ANT refuses the traditional dichotomies such as agency/structure, human/non-human, far/close relationship (Latour, 1996) or micro/macro scales (Latour, 2005), is to integrate the notion of mediators within the specific concept of *actor-network*[22]. It has been already outlined by the idea of an action *other-taken*, and the mediators *transporting* the action, that connectedness is central in ANT. But the conception of network is not connecting people in the technological sense, neither in organisational sense as Castells uses it. The concept takes origin from the Diderot's uses of *réseau* to *describe bodies and matter, rather than surface* (Latour, 1999, p. 370, 2005, p. 130). ANT uses the word *network*, as a concept, a tool rather than a description:

> So, network is an expression to check how much energy, movement, and specificity our own reports are able to capture. Network is a concept, not a thing out there. It is a tool to help describe something, not what is being described.
>
> (Latour, 2005, p. 131)

But the use of network alone cannot represent the ontological and methodological perspective of ANT. It is only with the conjunction (and not the opposition) of the

[22] Which, as for the term theory, is a relation of love-hate in ANT. They consider the choice of words actor and network as wrong and at the same time the best advocate of their ideas, despite some attempts to replace them with more accurate ones such as *wordnet* (Darch, 2014), *assemblage* (Latour, 2005), or *actant-rhysomes* (Vitali, Glattfelder, & Battiston, 2011).

word *actor*, that ANT deploys its own regard on the social and the society. The network, not seen as a structure, is a set of traceable actions, a *trace* left by an agent, and actor, an actant. The actor is not outside the network but makes the network, while it is deployed and made consistent through it:

> […] there is not a net and an actor laying down the net, but there is an actor whose definition of the world outlines, traces, delineates, describes, files, lists, records, marks or tags a trajectory that is called a network. No net exists independently of the very act of tracing it, an no tracing is done by an actor exterior to the net.

(Latour, 1996, p. 378)

This vision of actor-network finds its origin on the specific application of semiotics while ANT was still a sociology of sciences and technology. Applying semiotic to scientific discourse forced the sociologists to deploy other methods than relying on social explanations of phenomenon unveil to them:

> ANT does not assert that all the other domains of social science are fine and that only science and technology require a special strategy [...] It claims that since social accounts have failed on science so pitifully, it must have failed everywhere, science being special only in the sense that its practitioners did not let sociologists pass through their turf and destroy their objects with 'social explanations' without voicing their dissent loud and clear.

(Latour, 2005, p. 101)

From this critic toward traditional methodologies, ANT proposes a methodology inspired by ethnography and attached to the description of the social object rather than finding underlying causes of the events described. This is the third important aspect of ANT to understand why nonhuman having agency is appropriate and how ANT can duck some critics by claiming a methodological approach (but not without some limitations, see (Sayes, 2014).

It is well-known that despite its name, Actor-network Theory is not a theory but a research methodology centred on the descriptive aspect of reality (Latour, 1996, 1999, 2005; Law, 2009), or a position (Mol, 2010). It is interested in *what holds* *society together,* not the *why* but the *how* it is enacted and constructed (Latour, 1999; Law, 1999).

This approach promoted by ANT, blurs the traditional work of sociologists. As said, it is to do sociology without the use of Society. It is understanding social relationship without bringing power into the equation, and by studying human with non-human. The use of the word *society* is abandoned in favours of associations, mediators and network.

> There is no society, no social realm, and no social ties, *but there exist translations between mediators that may generate traceable associations.*
>
> (Latour, 2005, p. 108)

The traces are the manifestation of the social activity. Without them, the only possible assumption is that the social does not exists. This idea of trace is deeply rooted in the descriptive approach advocated by Actor-Network Theory. Under ANT, the descriptive methodology is used to unveil these traces. It is the work of the sociologist to discover them and to retrace them into the conception of a network of action, to render them visible. That essence of ANT into the type of work required by sociologist, retrace the social, rather than explaining will be deployed in the next chapter.

2.4 Conclusion

In the first chapter, I outlined the change in society that led to a new form of socialisation. Associated with a more fluid and organic social life, technology and raise of digital life, reinforce this aspect. Therefore, sociology has a need to find new approaches and new source of data to grasp this constant moving world.

This chapter, was solely focused on the consequences of the digital world, the increase of digital data and the apparition of new form of research and potential analysis with the profusion of data. I introduced the example of the ClubCard with Tesco to exemplify these news uses of data that were previously only considered as by-product. That helped to introduce the culminating use of this side-product with the Big Data and the 4th paradigm, notion initially developed for the natural science but becoming pertinent at the same time as we are more and more in a *Youtube Galaxy*.

However, this profusion of data is associated with a resurgence of a positivism, where numbers are seen as the ultimate answer to all the questions. To this pitfall, I briefly introduced the notion of social constructivism and how the construction of science, even natural sciences, may be dependent on other aspects than the method to reveal an underlying truth. That led to the actor-network theory.

In order to highlight the necessity to understand how these data are produced and how some inherent limits of their production and analysis need to be deconstructed, I described some useful concepts from Actor-network theory. The main outcome of this section is the necessity to unveil how data are produce, not only on a technical or economical perspective, but by including the nonhumans and by taking a descriptive approach. This is, in my opinion, crucial if we want to deploy a more sociological approach to a rather mechanistic phenomenon.

While not denying the potential of the Big Data for the social sciences, it is important to understand how it is constructed. Understanding the way it is collected, aggregated and analysed will help to understand why some conclusions are made while other conclusions are avoid or not even mentioned. This is, in some way, the practical version of the lampost and the drunk man searching his keys under the light. I am not saying the drunk man is stupid to search his keys under the light. On the contrary, he is smart. Why would he tries to find something where there is no light to help him. That metaphor describes this work. Rather than pointing out the lack of light in some area of the research, it will try to understand how research is build upon digital traces and what are the link in term of outcomes, basically describing the lampost rather than the drunkman. It is carrying a work of investigation of Big Data and the associated methods. But the task is not done by only studying the scientists in-situ, as Latour and Woolgar did with Laboratory of Life, but being one of this practitioner, deploying the understanding of how data is created, generated, and which limit, actors and actants are deployed within its manifestation.

The difference between actor-network theory and this work, is that they study external phenomena and reflect on it. The work they carry is a sociological work. Here, the position is not sociological, neither it is a computer scientist. Nevertheless, it uses the tools developed in both fields to develop a particular and interdisciplinary research on Social Network Sites. This is the *construction* of the research process, while trying to see what it is currently underlined.

The following chapters will be devoted to this task. The next one will be an introduction of what is a digital trace and how it can be conceived within the ANT framework. The help of the Tesco card will be at use before finally turning the eyes toward Twitter and specifically the API as a obligatory passage point, to use ANT vocabulary.

3 Digital traces

3.1 Introduction

Briefly introduced in the previous chapter, ANT relies on the existence of trace to assess the social. Rather than supposing the existence of any super structure or organisation, they are attached to describe on detail how the social is deployed, by unveiling these traces left by the *social in formation*. When action is taken, this action translates through mediators (human and nonhuman) and this translation leaves *traces* of the modification of the social ties.

The traces are left, *after*, the association has been made. This is due to its fluid and ephemeral nature.

> The adjective 'social' designates two entirely different phenomena: it's once a *substance*, a kind of stuff, and also a *movement* between non-social elements. In both cases, the social vanishes. When it is taken as solid, it loses its ability to associate; when it's taken as a fluid, the social again disappears because it flashes only briefly, just at the fleeting moment when new associations are sticking the collective together. […] It is traceable only when it's being modified.

(Latour, 2005, p. 159)

Rendering them visible is a complicated task as they disappear as soon as the *social* is formed. The task of the sociologist should consists, for ANT, to make these traces visible through meticulous descriptions, and letting aside any temptation to *explain* the social, *after* its formation. To render these traces visible, it is important to adopt the *right locus* (Latour, 2005, 164). What Latour means by the right locus is a methodology to render the *social flat*.

Usually, there exists a difference between the micro and macro level. To trace the different social in formation, the description needs to overcome the idea of individual versus aggregation. Rendering the social flat is to remove this two levels of *zoom* by investigating the localities of the association rather than taking an higher view of it.

The solution offered by ANT is to make a compromise between the two scales and rather than zooming in and out, to keep both the local and the global side by side.

The researcher should stick to the s*cale of the actor* versus the *scale of the measure* (Latour, 2005). The user creates its own frame of reference:

> It is not the sociologist's job to decide in the actor's head what groups are making up the world and which agencies are making them act. The job is to build the artificial experiment – a report, a story, a narrative, an account – where this diversity might be deployed to the full.

(Latour, 2005)

The actor can mobilise a local concept and than *zoom in/out* to justify their action based on a broader and more general idea such as *culture, society, economy,...* This constant shift cannot be properly grasped by the researchers if they operate an *a priori* zoom out, and use general concepts to describe the reason of acting of the actors. By doing so, the researchers miss the constitution and the dynamic of zooming in and out from the actor[23]. To be more precise, the idea of micro and macro still exists in the ANT methodology but it is deployed through flat connection and flat description:

> [...] there are two different ways of envisaging the macro-micro relationship: the first one builds a series of Russian Matryoshka dolls – the small is being enclosed, the big is enclosing; and the second deploys connections – the small is being unconnected, the big one is to be attached.

(Latour, 2005)

This connectedness between different levels needs careful descriptions and is and emphasis, one more, on the traces left by the action.

The task is then way harder, not only the researchers need to capture the social in formation during its brief moment of deployment, and before it is solidified, but they also needs to flatten the social and to be able to follow the traces wherever the local contingencies lead them.

[23] Boltanski and Thévenot use this idea of constant shift from local to global in their work to explain the mechanism of justification mobilised by actors in different situations. To anyone interested in the subject and how a post-Bourdieu sociology is possible by decomposing the justification and bringing them into a classified system (Humby et al., 2008, p. 130).

To answer these difficulties, technology helps to make the social more persistent and the follow of the traces more manageable:

> The more science and technology extend, the more they render social ties physically *traceable.*

(Latour, 2005, p. 119)

So not only the technology helps to retrace the social in some context such as science or finance, but also in more everyday interactions with the extend of profiles and the web in general. Aware of this potential, Latour used the potential of web profiles, not only to deploy a flat perspective of social interactions, but also to develop his theory about what it is called the *1-LS* and *2-LS* and the notion of *aggregation* rather than *collective/global*.

Briefly defined, the 2-LS means *two level standing point* and represents the theories that make distinction between micro and macro level.

> In 2-LS social theory, the most current approach to handling the distinction between macro-structures and micro-interactions consists in establishing a first level of individual entities, then adding to them a few rules of interaction, in order to observe whether the dynamics of interaction lead to a second level, that of aggregation which has generated enough new properties to deserve to be called a 'structure'.

(Latour, Jensen, Venturini, Grauwin, & Boullier, 2012)

They use it in contrast with their own theory, the 1-LS, *one level standing point* that is focus on the individual only but deployed in the network (Latour et al., 2012).

The reason why it is possible to deploy 1-LS is with digital traces, and profiles in particular, relies in the type of navigation. The hyperlinks offers a technical possibility to navigate through them on the same level. The description of a user, an actor, will be a list of attributes. This list of attribute on the web will be a list of hyperlinks to other profiles or other websites. These hyperlinks constitute a network of connection between pages and information. This is the definition of the actor in ANT, *a network of different elements that stabilises*. If one of these element is a university where the actor works (to follow the example given by the authors), then it is possible to navigate to the school and obtain a list of attributes (which could be a list of employees) that compose the school. That is the 1-LS:

> [...] circulating in such a way from actor to the network and back, we are not changing levels but simply stopping momentary at a point.
>
> (Latour et al., 2012)

The way they deal with this issues, and overcome the dichotomy is using visualisation tools to explore and to describe the dataset. For them, as it is easier to collect data about individual, we shouldn't reduce individual to atoms nor developing restricted vision of the social that "*[...] assumes that there first exist simple individual agents, then interactions, then complex structures, [...]"* (Latour et al., 2012).

Along this lines, they refuse to aggregate individuals but prefer to only associate them. In the Chapter 7, section 7.2 Levels of aggregation , when I will need to apply this idea, I will take another path. Nevertheless, their idea that digital traces brings something that allows to test new form of methodology is not necessarily new, as seen in the 4th paradigm. Here, the difference is the accent is on the ontological definition of actor, aggregate and network that primes the vision of collecting data. This type of approach is probably more complicated to develop, but at the same time give other opportunities to see different aspects of the social ties online.

One critic though about the Latour's article using digital traces, is the lack of proper definition of what a digital trace is. They resume it only as a trace in a database, but even if it is a starting point, much can say about the digital trace and how it can lead to a specific situation, shaping the *method assemblage* by being an *obligatory passage point*. Maybe the reason why they did not develop further the description of the digital trace is the lack of specificity and detail about *where* they collected their data.

The next sections will redefine what a digital trace is. First, it will reuse the ClubCard, to introduce some concepts useful concepts. Then the attention will be fully directed toward Twitter. It is only by describing how Twitter works, how it gives access to data for researchers, which information is collected, and how the need of number that we can understand how the current research is shape.

3.2 Digital trace in the world of dataset

To come back to the example of Tesco's loyalty card deployment can help to highlight two characteristics inherent in transactional data; *data complexity* and *control*[24]. It is possible to understand data complexity as similar to the broader issue of Big Data; while control over the environment will be more associated with a specific vision of data generation and the consequences on a methodological level. This data complexity is simultaneously generated and solved by technologies, the data analysts', and the marketing managers', needs.

It was generated because the technology was available to store vast amounts of information but even that vast amount of information was not enough. They wanted to extend the information to meet the marketing management's needs by extending the *where, when, which price* with the *who, from here*. To do so, a technological upgrade was implemented to collect the needed information. Then, as soon as the complexities of storage collection, were overcome, they faced the complexity of analysis.

At that time, specific restrictions and imaginative solutions to the situation created several versions of data analysis. Each version was increasingly complex and contained its own limits, but also a new potential, until the technological capacity allowed them to do a complete analysis on the full data set. But that was still, however, with their original version developed on a sample. In a sense, at each step, they dealt with data complexity on several levels, step by step, and each time they reached a new level, new opportunities (technological or marketing) led them to generate more data complexity and again having to face it.

3.3 ClubCard as an apparatus

This is what Ruppert et al. called "*an apparatus*", a similar idea to Foucault's dispositif. They defined it as a combination of use of analytical procedure, infrastructure and personal (Ruppert et al., 2013). The idea here is the same. As they

[24] It is important to note that these concepts are not two more characteristics of the Big Data, as the 3Vs were described in section *2.2 Data complexity and the 3V definition*. These concepts are the result of the process of generating relevant and valuable data, rather than the properties of the dataset itself. While the 3Vs are great to pinpoint advantages and disadvantage of collecting and storing Big Data, more in a pure computer scientific perspective, the data complexity and data control are more about the consequences of these Big Data and how to use it.

state, they are purposeful assemblages (Ruppert et al., 2013, p. 30) with more or less control over it.

I want to focus on the very reason why this works on a methodological level: the control over all processes of data generation. The argument here, beside the recognition of the inventiveness employed to extend every limit of their field, is that control over their data allows the reduction of data complexity.

The data is created in a controlled environment: a shop (or many shops). The data is about a specific behaviour; buying items, from which act specific information is extracted: when, where, which items, for how much, and is associated with other information collected through other ways; who, and from where. All these collection and analysis activities are aimed at one specific goal: enhancing the relationship with the customers by providing targeted vouchers. This goal is aimed at a specific population, the active ClubCard owners. But in order to be able to perform new analyses and develop more adapted solutions for the CRM, not only were more effective data collection and better algorithms needed, but also better measurements and a better definition of their customer population. This is the reason behind the transition from RFV to the loyalty cube. They created more and more analytical complexity to grasp an extended version of their population and to create new understanding about the same data as they had before.

This evolution of measurement is a manifestation of this control. To summarise, we have a simplified version of human behaviour, in a digital format, upon which specific clustering analyses, aiming at one specific goal, are applied. The direct conclusion of this control is that people using this data can conclude that: *you are not where you lived* (Humby et al., 2008, p. 95), as stated in the time of geo-demographic data, but *you are what you eat*.

And yes, it is true, but only because this assertion is made from data generated by a retailer about active customers with the goal to sell more. The inscription of human behaviours into a database is the first process of translation, the individual becomes the record, the trace within row and columns, and nothing more, and all information which is not retained in the table is lost. Depending on the capacity of the database to store more or less information, being accurate or not, the impact can be destructive and lead to the disappearance of all human complexity.

Poster, in his interpretation of the database's impact, is going in this direction and even further. For him it is the grammar, the language of the database itself which leads to the disappearance of human complexity:

> I contend that the database imposes a new language on top of those already existing and that it is an impoverished, limited language, one that uses the norm to constitute individual and define deviant

(Poster, 2013, p. 95)

This is common position for post-structuralists to fear the destructive power of database.

> According to our view, the power to constitute consumer identity is tied to linguistic power and in the mode of information, that power is located within the database.

(Zwick & Dholakia, 2004, p. 40)

It is not only that the database imposes a new language, but it also creates apparent relationships that for the post-structuralists, do not exist outside the database.

Nevertheless, the digital record can make visible, some relationships that were not visible in the first place, and that do not exist.

> […] the structure or grammar of the database creates relationships among pieces of information that do not exist in those relationships outside of the database. In this sense databases constitute individuals by manipulating relationships between bits of information.

(Poster, 2013, p. 96)

I think a more accurate vision is that the database and the digital traces re-contextualize the relationship through the analysis and allow us to see what is *invisible* otherwise. There is a major difference between the actions to *render visible* and to *create un-existing* relationships. So, the following citation from Ruppert et al. when they cite Strathern et al. is on the same line:

> […] rather than being decontextualized, the digital actualizes relations and connections that are otherwise beyond perception and thus inherent to the very imagining of social relations

(Ruppert et al., 2013)

But these relationships that are made visible are, and Poster is right, within the database. Stating that the database renders some relationships visible doesn't solve the problem of de-contextualisation of the database language itself.

The database makes this contextualisation less accessible by reducing the complexity into a column and row of digital records. But, this is to forget that the database is not an entity out there in isolation. The database is included in a complete process of knowledge or, at least, data production. At the beginning, when transactional data was stored as a by-product into the database, they were not used for any analysis. This data existed because the electronic system needed them in order to work properly. As soon as an interest in this massive amount of data developed, an entire bank of information was built around the data. New methods of data collections, reports, better algorithms, departments, and so on….

Then the database was used to make assumptions. But these assumptions, this knowledge, is an entire apparatus where the database plays a central role but only because it is integrated into a network of inscription devices. Therefore, it is only by trying to understand the link between the database and the overall environment that it is possible to re-create this contextual information. The database doesn't possess a grammar, or a power *per se*, other than the binary language. The use of it by specific people for a specific goal within specific resources is where the power and the grammar is located (if we want to talk about grammar).

Here, it is the digitalization of the behaviour which creates this re-contextualization, but only if we integrate the artificial environment generated by the data production in the same perspective as a science apparatus creates a human subject through the survey, qualitative methods and all the devices and relations deployed during the research.

Then this artificial environment, instead of being a limit, is an advantage and creates a context around the measurement and the conclusion that the Tesco marketing teams reach but, more importantly, it can properly deal with data complexity.

It addresses the full complexity of the data but it has to deal with a reduced, or at least a controlled, social complexity. The behaviours are the unambiguous buying of items. It is they who added more complexity around it by clustering, and creating a why, in order to have a sense of it in an aggregate version.

In a sense, this transformation of human behaviour into a single entry in a database is not losing any information because the database itself is an environment: the Tesco warehouse storing the database containing specific information in a specific location, Tesco Store, for a specific population, the active ClubCard holders. It is possible to see that data generated in a controlled environment explains the reason for data complexity but it also helps to deal with it. The loss of information during the translation of human behaviour into a database can be found again with these generative processes. That is what we are going to call a transactional approach to summarize these integrated approaches, a specific apparatus characterised, not for its specific environment but with its specific control over the data production.

Therefore, the transactional perspective has three fundamental aspects: control, *translation* and *context*. The danger is when these three aspects are not met and the conclusions of any use of digital traces are made outside that box. That is the fear of Poster and his fellowships: that the government's and companies' uses have added the loss of control over the new human created within the database. I still argue that the language of the database[25] that they use, and the conclusions of such new insights and realities, is made outside their context of production, their apparatus. However, even if the consequences are real, they are beyond the scope of the production of the knowledge itself, and the realities unveiled are not less real or less true than any others made with other methods of inquiry. They are just limited to one aspect, as Humby summarized in a negative image:

> We are all individuals, but when it comes to our shopping habits we have a lot more in common with many other customers than we have significant differences.
>
> (Humby et al., 2008, p. 149)

The next section develops this transactional perspective applied to the SNS. Later, I will explain why this is one of the biggest problems, the occlusion of the apparatus, that currently faces the research in SNS. The lack of control over the context of

[25] Which is for me an overly complex concept to describe a reduction of complexity into a manageable and computable form, surveys, administration forms, and all tools developed to collect information reduce the complexity. We are not creating a new grammar or a new language for each of them.

production of the behaviour (it is no longer limited to consumers), the quality of data (through limited access) and the translation of small behaviour as representative of human complexity to finally define what a digital trace is in the context of Twitter and associated research.

3.4 Transactional perspective within SNS

I analysed the process of transactional data research using Tesco as an example case, analysing marketing research with direct applications within a controlled environment. In this section, I will do the same to highlight some key features of the current research within Social Network Sites such as the general epistemological perspective, the theoretical framework and their respective technical aspects. After describing the research, the goal is to focus on mapping the constraints emerging from each key point while trying to see why the research is taking this transactional perspective and why Tesco states *We are what we eat*, and research on Twitter can state *Our influence is our Retweet*. Understanding the process of data generation and the specific goal within a controlled environment gives another perspective on how such statements make sense but also where their limits are.

The idea here reveals the process of producing the realities of Twitter as an object of enquiry for research. Therefore, the research itself and what it mobilises, the epistemological perspective, the theories and the technical aspects, are seen as many inscription devices in the sense of their original definitions from Latour and Woolgar:

> an inscription device is any item of apparatus or particular configuration of such items which can transform a material substance into a figure or a diagram which is directly usable by one of the members of the office space.
>
> (Latour & Woolgar, 1986)

Here, instead of a physical substance, the researchers transform digital traces into a figure or a diagram and render a complex reality intelligible and predictable, where multiple interactions occur. This is what Tesco data analysts did, and it is what researchers using Social Network Sites do. The idea is to try to unveil this apparatus through the different inscription devices used. But instead of a microscope or a telescope, the data accessed through API, then analysed through the lens of a specific theory which in turn will shape the computation of social interactions and transformation of digital traces into new visualisations or results. From there, it will be possible to develop another approach and understand the specificity of the problems raised by this new method, the transactional perspective, over a social object, Social Network Sites or, at least, over social interactions.

Previously I briefly outlined the process of data production within transactional data in the case of Tesco. The aim was to describe a transactional perspective, rather than a research on transactional data, and to outline the necessity of control over the environment that renders a specific vision of Tesco's customers possible. The process is identical to any research or description of the reality and it needs to render invisible some aspect of the reality through the method, or the method assemblage:

> If a statement in endocrinology (or medical sociology) corresponds to a reality out-there, if it simply seems to describe it, then this is because most of the assemblage within which it is located has been rendered invisible, Othered.
>
> (Law, 2004, p. 88)

The final result can be understood by reversing the process, making the invisible, visible. The categories built around the consumers are the result of this assemblage as they are the enactment of one reality through different inscription devices such as the till machine or the database, they produce a specific reality. It is not that this reality doesn't exist, but the whole process of translation enacts it while rendering the others invisible:

> method is productive of realities rather than merely reflecting them. And that parts of the out-there are made visible while other parts, though necessary, are pushed into invisibility.
>
> (Law, 2004, p. 89)

This is why, talking about transactional perspective is more accurate than being focused on the transactional data solely. It is an entire method, aiming to reduce reality (as every method does) but with specific tools and specific methods. The data will imply specific tools, which themselves are embedded in particular methods, reflecting theories. The data itself and its nature do not translate the complexity of the method assemblage, and changing data without changing the method will also produce the same effect. It is why now I will try to understand this perspective on another source of vast information: Social Network Sites.

Derived from weblog into a more interactional setting, these Social Network Sites, along with older forms of participative communication on the web (weblogs, message boards, Usenet), form this new source of data. Zhao et al. coined the term *social text streams* to talk about them (Q. Zhao & Mitra, 2007). This term is reduced to a social stream by (Sayyadi, Hurst, & Maykov, 2009). From the first weblog, sixdegree.com in 1997, to the famous Facebook or Twitter, these websites have seen their customer pool increasing dramatically. For instance, Facebook claims 1.23 billion users monthly users for its financial report in January 2014 (Facebook, 2014); Twitter claims 241 million monthly users, also in its financial report for the year 2013 (Twitter, 2014), and Youtube, one billion monthly unique users ('YouTube - Statistics', 2014)[26]. Beside this penetration among everyday life of internet users, they are not only creating data on their everyday activities, but they are characterised by the active production of their users on a (more or less) voluntary basis.

As well as in the marketing field and shop data analysis, these users produce digital traces, in mass. In fact it is not simply more data. To make data manageable, Big Data always associates to the quantity an antinomy in the form of a small amount of data about lot of people, as Tesco did, or a lot of data about a small amount of people, as not many people do.

But it is not only this double aspect, big and small data, which makes this Social Stream so attractive, it is also the conjunction of different factors of the data they generate, for instance, the nature of the data production. It is in looking at behaviour, whereby people leave a trace when they actually do something. This quality is needed to be seen in comparison to the inferred information collected through surveys. This almost direct link between the behaviour and its record offers a new opportunity for social science to study human interaction without interfering with it. It is in a specific context of social network sites but this context, or this interaction, can be extended much further.

The interest can be in the link between SNS and external events like predicting political elections (Skoric, Poor, Achananuparp, Lim, & Jiang, 2012; Tumasjan, Sprenger, Sandner, & Welpe, 2010), risk of demonstrations and riots in some areas (Baker, 2012; Bollen, Mao, & Zeng, 2011)or even flu epidemics (Culotta, 2010).

[26] They don't communicate the period of their statistics but the site was accessed in February 2014.

Others studies examine the activity over SNS to understand information diffusion through SNS and its difference from newspapers (W. X. Zhao et al., 2011).

They can also emphasize an understanding of users through their profiles, activities or networks (Hernandez-Orallo, 2013) and can show their political orientation, for example whether a democrat or republican in America (Kosinski, Stillwell, & Graepel, 2013), or the personality profile of a Facebook user (Back et al., 2010). Focusing on Facebook and the user's network instead of their own activity, other studies have shown that it is possible to know their sexual orientation by gathering information on the sexual orientation of their friends (Jernigan & Mistree, 2009)[27]. Some other studies extend the research to various platforms, for example predicting interactions on Youtube from an algorithm developed on Twitter.

This list is obviously not exhaustive. Nevertheless, the point here is to show how research on Big Data, Social Network Sites and the transactional perspective is prolific and varied. It is prolific because the availability of data is increasing over time as more and more people use SNS in daily life and for different activities. It is varied because many different SNS exist, developing their own characteristics in this competitive market, but not that much because researchers develop different analysis.

All these studies emphasise the importance of having a simple digital trace in order to manage the complexity of the interactions. The essential building block of them remains the same through all of them: the digital trace of a simple behaviour, aggregated in large volume to find insights about more complex phenomena. A like for Facebook (Kosinski et al., 2013), a tweet rate in Twitter (Bild, Liu, Dick, Mao, & Wallach, 2015), a subscription on Youtube (Wattenhofer, Wattenhofer, & Zhu, 2012), and so on.

These behaviours appear simple but are becoming more complex in the research process. For instance, in the article on the popularity of Youtube, the views are not of a single dimension. The authors broke down the measures in three separate datasets[28]:

- The User subscriptions of users,
- commenting activities (reduced to the knowledge of comments or not)

[27] The purpose of this section is not reviewing the entire Facebook research. A good review was done by Wilson, Gosling, & Graham (Wiener, 1956).

[28] They had access to the full dataset.

- user-uploaded content (Wattenhofer et al., 2012)

All these metrics were quantitative only, reducing the entire context to a one-dimensional measure for use in a network analysis. For instance, under the assumption that a comment is about content and a subscription is about a social link, they were able to distinguish different measures of popularity. For them:

> […] there is a dichotomy between the dynamic of content (user's content) and the dynamic of social (user's users), where the content consumption and social influence are similarity active, but largely separate components of the same system.
>
> (Wattenhofer et al., 2012)

A complex behaviour, here a comment, is transformed into a single measure, and then it is aggregated into a dataset. This is the dataset, composed of one type of digital trace, which contains the contextual information and allows researchers to create distinct versions of the same idea, the popularity in either a content or social perspective. This reduction of complex behaviours into single measures is a manifestation of the trade-off for contextual information to volume, where the reconstruction of the dataset contains the information.

This research shows that in order to deal with the number imposed by the 4th paradigm, the research reduces the information about the user, taking only one simple digital trace and then creating knowledge from it by applying analysis upon this simple behaviour. This is specifically the same idea as in the transactional perspective but, instead of reducing the number of items collected, research limits the social interaction to one aspect only.

The best example of the application of the transactional perspective applied to Social Network Sites is the study of Twitter and the measure of influence. This research possesses all the characteristics of an interdisciplinary research and is, at the same time, the archetype of a transactional perspective applied to digital traces about interaction in a more open context than a shop.

3.4.1. Network theory, influence and Retweet

The retweet (RT) appeared in 2007. It is a practice on Twitter that consists of republishing a tweet (a 140 character message) from someone else to one's own followers at the click of a button and which is often used to understand users' behaviour and predict activity. It became current practice among Twitter users before being incorporated as an official mechanism by the Twitter team (Kooti, Yang, Cha, Gummadi, & Mason, 2012). Since Cha's study in 2009, where they disprove the importance of the number of followers/friends as being a good measure of influence, and said that the retweet was more important (Cha et al., 2010), a lot of studies put the retweet rate, or behaviour, in a central position. In their study they used three different metrics to assess the influence of a Twitter user. The *in-degree* which is based on the number of followers a user has and it represents their audience. The *retweet* influence is based on the number of retweets containing the user's name, and represents the ability of the user to generate content. And finally, the *mention influence*, based on the number of tweets that contain the user's name and represent their capacity to engage others in a conversation.

As for the research on Youtube, the idea here is that several metrics, based on the same digital traces, represent different concepts of influence, the audience, the content generation, and the conversation. They discovered that the audience doesn't have a significant impact on the *retweet* influence and the *mention* influence. However, an essential difference in the collection data is that the *retweet* and *mention* are collected over time, whereas the audience is collected as a snapshot. This is a same issue in the study on Youtube's popularity (Wattenhofer et al., 2012), I will come back to this specific point later. Sometimes, the retweet is used to measure different phenomena, such as homophily (J. Weng, Lim, Jiang, & He, 2010) or a transitivity phenomenon (Golder & Yardi, 2010). On other occasions it is used as a variable measuring attention (L. Weng, Flammini, Vespignani, & Menczer, 2012) with the dual role of the hashtag (Yang, Sun, Zhang, & Mei, 2012).

However, one of the shared characteristics is the use of the influence from a network analysis perspective. The nature of the SNS, connecting profiles of people into a network more or less open, is the perfect situation to apply network analysis on a real-life field.

In fact, their model, and the theory behind it, is a direct application of the definition of influence as described in network analysis. When they explain the information cascade (Cha et al., 2010), diffusion tree (Bakshy, Hofman, Mason, & Watts, 2011) or the neighbourhood effect (Z. Xu & Yang, 2012) with the retweet, it is because the behaviour is a good direct manifestation of these concepts. But the retweet could also be the manifestation of other types of concepts. However, the network analysis has an advantage over other methods. It is not (only) suitable to represent this behaviour because the interactions are more networked nowadays than they were previously. It takes advantage of an intrinsic evolution of the method itself and the evolution of technology, such as clustering, which finally allowed Tesco to take advantage of their data.

The notion of influence, within network analysis, is a powerful tool to understand interactions between people, where the possibility offered by the evolution of a mathematical model and a computable model is essential. The network analysis uses the graph theory but, to perform computer analysis, it transforms the data into a suitable representation, the matrices:

> Matrices are an alternative way to represent and summarize network data. A matrix contains exactly the same information as a graph, but is more useful for computation and computer analysis. Matrix operations are widely used for definition and calculation in social network analysis, and are the primary representation for most computer analysis packages (GRADAP, UCINET, STRUCTURE, SNAPS, NEGOPY).
>
> (Wasserman & Faust, 1994)

So it does not seem that the only explosion of data that brings the network analysis as the main method on Social Network Sites, it is also because this method has the advantage of fitting perfectly with the nature of data generated from these web services. It uses the same format, the adjacent matrices or adjacent lists, to make the link between the data, the field and the theory, with the advantage of being easily computable. The network analysis is also adapted to answer specific questions such as the social influence or the spreading of messages. Specifically, regarding the question of social influence, one of its major aspects is dynamic evolution. Data needs to be timed and able to represent an evolution over time (Mason, Conrey, & Smith, 2007). This is what the studies, mentioned above, are doing (Guille, Hacid, & Favre, 2013), on Twitter but also on other Social Network Sites (Nettleton, 2013; Union, 2013). However, to do this, the information of time needs to be found somewhere. In the case of Twitter, it is found in the essential brick of the web service, the tweet. This location of the essential information will have repercussions on the type of results, which I am going to explain. Before my explanation, I present two comments. First, the study on Social Network Sites takes advantage of the Social Streams created by users on these websites. The idea is to get as much data as possible (epistemological perspective) to develop new models by transforming the complex digital traces into a manageable form. Secondly, this manageable form has to be associated with existing methods of analysis (methodological implications). Tesco developed different clustering algorithms to build their customer cubes, but still had to reduce the information to 80 products only. Research on Twitter uses network analysis as a tool to transform the complexity of interaction into adjacent lists, or a matrix, and reduces the interaction to node and edges (theoretical perspective). However, this conversion of the digital trace needs appropriate tools and technological possibilities to collect and store the data.

3.4.2. The technical possibilities - The APIs

All SNSs are, at least marginally, sometime drastically, different to each other. They offer different mechanisms of interaction, expect the users to publish all sort of personal data and give different granularity in term of privacy settings. Besides all these differences, that directly impact on the users, the access to the data is also different between them and can also change. The impact on the users is less than on the people that develop services to interact with the SNSs, or in this case, researchers who aim to study the social deployed on these web spaces.

No SNS gives full access to their data. The reasons are endless but the result is that only a limited and heavily controlled access is given to the researchers. This control can discourage some researchers investigating the phenomenon, even if it is more likely that the technical knowledge has been a bigger reason (D. Boyd & Crawford, 2011). But ultimately, it definitively shapes the direction any researches will take.

The hypothesis here is that the control on access, associated with the trend of the 4th paradigm reduces the scope for researches on only one aspect, the data which is at the same time accessible and in quantity.

The API

All Social Network Sites give different access to different type of data but they all tend to user what is called an *Application Programming Interface (API)*. An API is a computer concept which is used to describe an interface between two programs in order to communicate information.

They often come in a form of a functional library that includes different ways of interacting between programs. It is composed of modules to make the programs work together. On the web it is slightly different. It often works in the form of specific request messages with a structured response, in a specific format ('Application programming interface - Wikipedia, the free encyclopedia', 2014).

The API gives an easier access to the data for researchers (among everyone who is interested in the data), and makes the control over which data is available easier for the company who owns and store it. This is the technical entry point to the data and it is where the fate of many researches is being decided. Because the access can be radically different in regards to the resources accessed, it is what ultimately leads to the decision of which data are going to be collected and studied.

Specifically on Twitter, the API is divided in two main accesses, the Stream API and the REST API[29]. They work differently, have different types of information and different limits.

By using the REST API, the client (the researcher, using a piece of software that can talk to the API), sends a request to the API to access some specific information. The Twitter API responds to the request by sending back the relevant information. As soon as the information is received, the connection between the client and the API is closed.

The Stream API works differently than the REST API. The client still establishes a connection to the Twitter servers, but this time the connection remains active, open, and the data is sent on continuously to the client. This change is important as it allows information to be collected in almost real time. As soon as information is available (let's say someone's tweet), the client receives it.

The type of information, the second main difference, that is accessible through the Stream or the REST API is also different.

[29] This is not true, more APIs exist, such as the Ads API, but they are irrelevant for this context. Also, in the future some may be added, as Twitter is the only entity that has control over it. The developer documentation is available on the following webpage: https://dev.twitter.com/overview/documentation

There is a lot of different ways to access the information. For the REST API, the number of different addresses that give different information is 97[30] at the time of this work. The information can change over time but it is possible to say that the REST API gives access to almost everything, while the Stream API can only give access to the tweets. But by the nature of connection, the REST API gives the possibility to download historical data (to a certain limit), while the Stream API (being an ongoing connection) give only real-time data, it is impossible to have past tweets with the Stream API.

The third difference is the type of limit that Twitter imposes on the API[31]. As seen, the REST API works on a mode of requests-response. Therefore, Twitter imposes a limit on the number of requests it is possible to do for a specific time window which is 15 minutes. That means that after the client reaches the limit of the number of requests it can possibly make, Twitter stops sending information back. Also, not all information requested is sent at once, that means that several requests are required to build the entire dataset asked for.

For instance, asking for a list of followers gives 5000 user objects per requests, and the number of requests is limited to 15 every 15 minutes. If the users have 5702 followers, two requests are needed to have the entire lists. If the request concerns the information about users' profile, the limit is 100 users per request (and still 15 requests every 15 minutes). After 15 minutes all the limits are reset and Twitter API sends back the information on new requests. These limits are really constraining if the goal is to collect a huge amount of data.

It is not a problem with the Stream API as the information is in real time so there is no windows time and limit on calls. However, others limits are imposed. As well as for the REST API, different endpoints exist to connect to the API. Three types of Stream API exists, the Public Streams[32], the User Streams[33] and the Site streams[34].

[30] The details and all available information are on this webpage: https://dev.twitter.com/rest/public. It is important to note that not all of them are used to give information, some are to publish information, such as tweets or profile information. The API is used by any developer that wants to develop a program that interact with Twitter.

[31] Both API face another limit that is inherent in Twitter. In contrast to other SNS, such as Google+ or Facebook, where the privacy settings are numerous, on Twitter, only two settings are possible, public or private. These settings have an impact on the information that can be obtained through the API. The first setting allows anyone to see the profile and the tweets, as well as following the user. The latter allows more privacy because the tweets are only seen by people who are following the user and this is only possible after the user has accepted the request.

[32] https://dev.twitter.com/streaming/public

[33] https://dev.twitter.com/streaming/userstreams

[34] https://dev.twitter.com/streaming/sitestreams

The endpoint most used in research is the Public Stream. This endpoint has three ways to filter the tweets received: The statuses/filter[35], the statuses/sample[36] and the statuses/firehose[37]. The Public Stream is the one mainly used for research. It has different options. It is possible to obtain a sample of all public statuses of around 1% of the tweets. Another option is to filter tweets to obtain a more specific selection. However, this selection is also limited to 1%. If the collection of tweets is filtered on a limited number of keywords or users[38], it is possible to have an entire dataset for this specific subsection of tweets or users. But if the number of tweets collected through the filters exceeds this limit, the API will only return a number of tweets corresponding to the cap of the sample Public Stream.

The main purpose of this description is to show that the APIs have their own logic. The importance is that this logic is going to shape any data collection. Not only the SNSs have specific interaction and information given by the users, but also the type of API they give access to is different. I suggest that this last constraint is the most important one as it controls which data can be studied, and ultimately how the field progresses in that area.

On Twitter specifically, it can be seen that only one type of API fulfils the requirements imposed by the easy access and the huge volume needed under the *Big Data* idea, the Stream API. As this API only gives access to the tweets, it is natural that the vast majority of the research is focused on this aspect only[39].

But not only the API has an importance on the choice on which interaction is studied, the object retrieved itself is important, here the *tweet object*.

[35] https://dev.twitter.com/streaming/reference/post/statuses/filter

[36] https://dev.twitter.com/streaming/reference/get/statuses/sample

[37] https://dev.twitter.com/streaming/reference/get/statuses/firehose

[38] For now, this limit is 400 keywords, 5000 users or 25 0.1-360 degree location boxes (‘REST API v1.1 Resources’, 2013).

[39] It is also possible for some researchers to access other datasets and obtain a bigger sample. For instance at the University of Southampton, the web-observatory has access to 20% of the public tweets, providing more research opportunities. Other companies that have partnerships with Twitter, also offer the full access to data such as gnip or datashift but these methods are not the norm and more often it is the Public streams that are used for a research perspective. It is the easiest method to collect tweets but it can be combined with the REST API to combine advantage of both API. However, it is difficult to know which API is used in studies on Twitter. It is quite rare that the aforementioned studies specify the method of access to the data. Often, they only state the number of tweets or the number of users their data contains.

A tweet object, retrieved by any method, provides a range of information, such as the author of the tweets, their time of creation, the text, any hashtag, whether it is a retweet, and if it provides information about the original author, and so on.

For a good visual description of what a tweet is like when downloaded with the APIs, see the Appendix Visualisation of the tweet object.This description is outdated (2010). The key and values changed since the publication and are probably going to change more in the future. The only way to keep an updated version of the tweet structure is to go directly on the Twitter website[40]. The relevant structure of the tweet at the time of this work is detailed in the methodology chapter. However, even if the visualisation is outdated, it is still represents a good idea of how a tweet is represented through the APIs and what is a tweet in research.

Now, I will describe the type of information that it is possible to obtain, the tweet. Usually, this description is not given because people tend to know what a *tweet* is and what a *mention* means. However I think it is essential to understand what information is collected and what is not, as that is fundamental to any analysis.

The tweet – As an obligatory passage point

Access to the API can produce two types of information, encapsulated in a JSON[41] file. It can contain a tweet or the profile information, or even both. There are the only two types of information that can be obtained, but they are more complex and complete than it seems at first glance. This is the trace that almost exclusively used in the research, the tweet. On a more abstract level, it represents a communication from a user to their audience (their followers), or between a user to another user(s) (in the case of a *mention*). It can also be a shared message (in the case of a Retweet).

To differentiate between these different abstract levels we need to understand the structure of the tweet itself. When Wattenhofer et al. created two datasets representing two behaviours (Wattenhofer et al., 2012), it was based on two different metrics from the beginning, the *comments* and the *subscriptions*. Within Twitter it is different, all different concepts used in all research are based on the same containers, the tweet, which contains the different information needed. For instance, in Cha's

[40] For a complete review of the information contained in the tweet object: (Latour, 1999).

[41] The JSON format is a data interchange format widely adopted by web services for its lightness and for the good balance between machine and human reading. It is a text format, independent on the (computer) language used. It is a dictionary which has key-value pairs, where the key is the representation of the data and the value, the data itself (Humby, Hunt, & Phillips, 2008, p. 100).

study, they created 2 different datasets. One was based on the number of the retweets and the another one on the *mentions* (Cha et al., 2010). Apart from the fact that the Twitter API gives this information in a handier way[42], all information is contained in the message itself[43]. Moreover, the tweet can contain links that are sometimes analysed themselves to enrich the tweet message (Celik, Abel, & Houben, 2011). This method extends the digital trace to external links and websites. The digital trace is multidimensional and offers different possibilities for research. But the 4th paradigm imposes a logic of number (instead of representativeness as for the traditional approach) and within this logic, the issues raised by Big Data. Therefore, studies need to narrow down the complexity by focusing on only one dimension of the tweet. One type of information is translated into a metric where only one aspect is relevant, often the retweet in the case of information cascade and influence, hashtags in the case of community study. Other information can be used to filter the dataset collected, such as the language or the geographic area, often used when studies try to predict elections (Gayo-Avello, 2012).

But one other important aspect in the tweet information is the time stamp information localised in the key '*created_at*'. This information cannot be found anywhere else. This information is essential to study evolution of network as pointed out earlier and only the tweet embed it.

This unique presence of the key *created_at* transforms the *tweet trace* from another manifestation of behaviour, as well as the low cost of data collection through the API allow us to see the *tweet* as an *obligatory passage point* for most of the research on Twitter.

The term, introduced by Callon (1986) can represents an object, an human (Law, 2009), or even states (Bueger & Bethke, 2014). It can be define as an object that forces the others actors in the network to converge through it and shape their action and their content. The tweets, and its extension, the Stream API are the central point to where the researchers will collect the digital traces provided by Twitter. That specific *door* to the social also provides the timestamp, essential information to see the social in action. As it is an ephemeral trace and link, having only a snapshot (as the profile information gives) will not help to see how it has been constructed. The

[42] The key *'retweet'* and the key *'mentions'* in the *'entities'* array.

[43] Sometime, only using these methods can be tricky and result in incomplete qualification of tweet as retweet because different practices co-exist (Azman, Millard, & Weal, 2012).

necessity to capture the change live is inherent to the notion of social in formation within ANT. Therefore, either it is possible to observe it when it is actually deployed, or collect traces of it. In the latter case, the only possibility is to have a timestamp to be able to locate in time when the actor-network is formed.

What is important is not so much the presence in the tweet trace than its absence anywhere else. The Stream API is adapted to collect dynamic data on a large scale, fulfilling the logic of the 4h Paradigm. This data embedded in the tweet, can be easily reduced to a single dimension in order to deal with the size of the dataset. This is a perfect assemblage but rendering only one aspect of the reality, not a wrong one, but one aspect enacted by this combination of advantages and limitations on every aspect. The tweet is the ClubCard from Tesco. While the ClubCard and the record of the customers basket helped to redefines the user, it had to be comprised as the entire manifestation of the Tesco's strategy to enhance their knowledge, technical and technological knowledge and advancement, and the control over the environment, the shopping habits over their own shops.

3.5 Conclusion

Along this chapter, the notion of trace has been developed. The trace in general is considered as an essential brick of the ANT methodology. The conception of the social as ephemeral phenomena brings the need of a necessity to capture its formation rather than its established result. Following this idea, the trace becomes the only way to see this transformation.

The notion of trace is rather a methodological perspective on how to study the social rather than a specific element of it. It can be historical accounts, documents and notes from a laboratory, and reports from actants. The diversity and the qualitative nature of this data are the leverage for a full account of the *social in progress*. In the digital age, this brick can take the form of a record in a database and bringing a more malleable and quantitative aspect to the trace of the social. The nature of this record opens several new opportunities. Latour fully takes advantage of them while testing his 1-LS theory and the possibility to navigate and to visualise different actor-networks. His interest was only in a practical demonstration of his theory and he did not investigated the consequences of the transformation of a trace into a digital row.

To illustrate the impact this transformation on research, I used the ClubCard example and to study the transactional perspective which leverages such particularities. Digital traces and databases are centrals for the creation of knowledge and the resulting increase in power as demonstrated by Thrift(2005).However, this creation of knowledge is the result of an apparatus where the company has control over the data. This control can be seen as the context surrounding the digital trace. The data collection, the data storage, the type of analysis and the purpose of the analysis gives the missing information that has been lost in translation. By transforming complicated behaviour into manageable row in a database, the control over data creates actionable knowledge that would have not be possible otherwise.

The process of research using digital trace is not different from the transactional perspective in business as it uses the same approach on data. Consequently, the control over the data is an essential element to understand what the digital trace represents. It is only logical that if the control over the digital trace is intrinsically linked to its use a description of how and where the data is collected, is essential. This step is often concealed, even if it can bring back the context surrounding the digital trace.

About Twitter and the associated researches, the main technical aspect (similar to the tilt machine for Tesco) is the Twitter API. The description of this API has the objective of contextualising the technical context in which research can be deployed. In this specific case, the control over data is displaced from an omnipotent control to a limited access to data provided by the APIs. This limited access brings different types of constrains and possibilities on the research.

Another restriction (or opportunity) plays a major role in shaping research. As seen, the urge of numbers amplified using digital traces will have an impact on which type of API and which digital trace is often used. By understanding the interplay between the state of the research in one hand; and the specific technical limitation from Twitter on the other hand, it is possible to see the importance given to one resource over another one.

As a result of this research specificity and Twitter technical possibilities, method assemblages for studying Twitter often use the Stream API and the tweet to *plug* the research. In return, the analysis are often based on Social Network Analysis as they fit perfectly with the nature of data connected.

Claiming that the entire research landscape only uses Social Network Analysis and the Stream API would be an obvious mistake. Some researches adopt a more digital ethnography methodology. This could be done through case studies and mixed approaches with offline and online content (Postill & Pink, 2012), textual analysis of the content of one hashtag during the Ferguson's protests (Bonilla & Rosa, 2015) or specific context such as education (Chretien, Tuck, Simon, Singh, & Kind, 2015; Veletsianos, 2017). Another type of researches gather data from different digital sources. For instance, Fu & Shumate combine data from SNSs with data contained in hyperlinks (from news media) to study the impact of SNSs over the news media (Fu & Shumate, 2017).

These examples add information and context to the digital trace by adding trace that can be found outside the Twitter API. While these approaches are interesting, their methodology falls out of the scope of this study. The accent here is to see what it is possible to study through Twitter only, considering the only accessible digital traces.

Having reduced the notion of digital trace to what is accessible through Twitter only, there is still a variety of behaviours and interactions that is possible to collect. Within that idea, some researches combine different metrics from Twitter, reviewed by Riquelme & Gonzáles-Cantergiani (2017). In their review, they list all different studies that investigate the influence over Twitter and which metrics and algorithms have been used. The three metrics used are the *follow-up, the retweet, the mention, the favourite and the like.* Beside the *follow-up*, all these metrics are found in the tweet. Moreover, the studies measured the influence in a social network analysis context. Some studies were using other types of profile information such as the study on the impact of the displayed name on the number of followers (Mueller & Stumme, 2017), or the longitudinal study on follower count on a dataset of 507 users (Hutto, Yardi, & Gilbert, 2013). But in the former example, they only used one snapshot of number the of followers, while in the second, the focus was only on the tweet itself. Additionally, they were using the followers mainly as proof of popularity and not as an object of interest itself. Despite using other data than the tweets the studies outlined above were still conceiving the profile and the network solely in a quantitative perspective. The different aggregations made were to create a measure that can validate or not the algorithms or the hypothesis. The aggregations were never used as a measure of context, which is what this work will develop. These examples reinforce the idea developed on the importance of the Stream API and ultimately, the Tweet as *an obligatory passage point*.

In this work, the focus is on the conception of the social activity within a context and within a specific method assemblage. The transactional perspective developed in this chapter helped to describe how this assemblage works. However, the core idea of this work lays into an ANT conception of the network and this conception proscribes the aggregation method. It advocates for an extended flat descriptions under the principle that aggregation reduces the complexity of the social captured by the digital trace. The position here is more nuanced and claims that the control over the data, transformed into a method assemblage, has a role to play in the contextualisation of the digital trace. Describing this control is essential to see how the technical context framed this work and more generally, to see how the digital traces shape the research on Twitter.

The description showed how complex it is and how much more it is than a row in a database. Opening it up, something that Latour hasn't done unveils what it is possible or not. Consequently, any definition off digital traces should be extended to its method assemblage.

In summary, the definition of the *digital trace,* as used in research on Twitter and for this work specifically, has to include the notion of *trace* as methodological manifestation of the social, the control (or absence of it) over the data collection through the API as a *context* and the current state of the 4th paradigm as a *pressure*. The API is interpreted as the *obligatory passage point* between the research and the row in the database controlled by Twitter. The former is pressurised to access digital trace in quantity while the latter has to impose control on the type of data and the quantity that can be accessed by third-parties. However, this conception misses the most important element of the equation: the actant performing the social activity on Twitter. This actant is not anymore the person or the bot behind the computer but an aggregate of translated traces of its activity through the API, resulting in the negotiation between two different apparatus, the researcher and Twitter.

This definition of the digital trace is mostly descriptive but the idea of a translation of a trace through the API as the connection point between two apparatus can raise some concerns. The next chapter is devoted to these questions and starts with some existing critiques stated by social scientists and humanists about the use of data itself.

After outlining some limits of the use of data, the last section will give a definition of Social Network Site focusing on a simple description that can be translated into a methodological object and taking into account this limitations described.

4 Activity and context

4.1 Introduction

The rise of social media and the associated amount of produced data does not only bring enthusiasm among social scientists. Several critics and warnings have also been voiced and continue to be. These critics cover all aspects of research and concern the collection, the data analysis and the interpretation of digital trace, as well as the ethical issues about it.

> issues of **representativeness** and **uniformity** […] challenges of **interpreting** these sorts of social data […], extend to the **ethics** of research practice and personal **privacy**, the **value** of **theory** and **reasoning** in relation to prediction and engineering, and, of course, the application of **appropriate** and **rigorous** modes of **inference** […] also because of the **acquisition**, **archiving**, and **analysis** of these types of data [...]
>
> (Hutto et al., 2013)

For Shah, Cappella & Neuman, this list of issues that research faces, starts to be solved in the current research by numerous and various methods. This enthusiasm can be shared but criticism still occurs alongside great achievement in the Big Data era.

For instance, one of the earliest Big Data triumphs using digital traces from social activities on the web is the flu prediction by Google (Shah et al., 2015). In 2008, the year that Google launched the service(‘Google Flu Trends’, 2015), Google claimed they could predict the flu spread from one to two weeks prior to the *Centers for Disease Control and Prevention (CDT) Influenza Division*, which uses lagged data (‘Tracking flu trends’, 2008). Google enhances its service until the shutdown in 2014 (Ginsberg et al., 2009).

But, as Lazer reports in his article, in 2013, this apparent success still failed comparatively to traditional methods used by the CDT and it over represents some correlations. To explain this failure[44], two elements are offered by Lazer.

[44] It is important to mention that the author recognises the contribution of that research. However, even Google seems to have to realise the limit of the tool as the service is discontinued (or the lack of economical advantage, or any reason that is not mentioned because it is a private company).

The first argument is about what the author called the *Big Data Hubris*:

> […] often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis
>
> (‘Google Flu Trends’, 2015)

That idea that Big Data will overcome any other methods was already debated in the chapter 2.3.3 and the 4th Paradigm. A position against this evangelism, has already been taken in this work. However a second argument about how this data is generated is also stated by Lazer. This second argument is about the algorithm dynamic:

> Algorithm dynamics are the changes made by engineers to improve the commercial service and by consumers in using that service. Several changes in Google's search algorithm and user behavior likely affected GFT's tracking.
>
> (Lazer, Kennedy, King, & Vespignani, 2014, p. 1203)

Here, this issue reaches a problem deeper than the transparency issue. Google needs to tweak its algorithm of its search engine to build a better service. The only goal they try to achieve is their service and the quality of it. To enhance the service, the inputs are the users' behaviours, they use this activity to modify and to test in real time different algorithms and designs. The method to improve their service may follow scientific methodology but the later impact is not to improve knowledge, but to improve user experience and profit. This situation damages any research on this generated data. This is not only an issue about the access to some data while some others are invisible for the researchers, it is also the pro-active generation of such data that questions any further validity of any research. When transparency seems to tackle only the issue of accessing the data, the whole generation of it should be also under scrutiny. What is the value of saying that Google flu search can predict better than the CDT when the search engine itself is not consistent but more important, its changes are not visible and accessible? Short answer: not very much[45].

However, what Google do is what Lin calls *better engineering*:

> [..] the goal is to improve a particular metric, the only thing that matters is improving that metric, […] real-world user behavior is the ultimate validation
>
> (Lazer et al., 2014, p. 1204)

[45] This is maybe the reason why Google shut down the service after a while.

This perspective, not focused on understanding but on building a good predictive model, not being worried about the knowledge itself but on the practical outcome is used in opposition of the *better science.* This latter idea is more in line with a traditional goal in social science:

> […] many social scientists are using big data to understand the complexities of human behavior, such as how individuals form and maintain social ties and the dynamics of influence and power.
>
> (Lin, 2015, pp. 39, 40)

This distinction has the benefit to eliminate several issues that social scientists mistakenly associate with one type of research (*best engineering*) while it only suits the second one (*better science*). It helps to re-frame the current status of the research in social sciences and overcome a simple *bad research* definition.

However, his critics about the misconception of social scientists on Big Data takes into account the asymmetry in data availability between academics and industry. But this asymmetry for him is nothing new. He compares it with the access to the clay tablets under Sumerian time. For him, the current asymmetry is something that has been always experienced and we should trust the industry to sort out real issues… (Lin, 2015, p. 41)

Even in the improbable situation where it is safe to blindly trust the industrialists the lack of transparency has an indirect impact on the research currently done. As seen in the previous chapter, not everything is available and research is often focused on one type of data (the retweet for instance). This lack of transparency mentioned often *forces* the researchers to use the most public social media (often Twitter) and the most prominent trace (the tweet). These *opportunistic methodologies* of doing research will poison both *best engineering* and *best science* by collecting data that is available instead of using data that suits the needs.

Therefore, to do a *better science* an in depth study on theoretical aspects is needed prior to work with the data itself. This is what Resnick, Adar and Lampe advocate in their paper ((Lin, 2015, p. 44). The need to associate a theoretical framework to avoid the misuse of data by pure facility of access by articulating the theoretical constructs that can explain the best the communication on Social media with the fragmented data that it is possible to obtain.

The article lists 4 types of theoretical constructs that will be later linked to the available data and the appropriate methods to study them. These theoretical constructs represent two aspects, the *human* (*individual* or *social*) and the *technology* (*single site* or *across sites*). Depending on which level you are, the theory in social sciences will have different concepts to study.

	Individual	Social
Human-Based		
Static	Gender; age; extraversion; loneliness	Trust; tie strength
Dynamic	Increased loneliness	Increased tie strength
	Single site	Across sites
Technology-based		
Static	Presence and use of particular features	Communication multiplexity
Dynamic	Entry; exit; participation trajectories	Stability of membership

Table 1: Reproduction of Table 1: Taxonomy of constructs […] (Resnick et al., 2015, p. 195)

Added to what Resnick et al. developed an integrated approach between the theory and the data. They established the best way to deal with the fragmented traces that it is possible to obtain from social media.

They list 5 challenges to face with digital traces; *variability* of behaviour between and within subgroups; *change* in behaviours and product over time; *unreliable* self-report; difficulty of *inferring causality* from observational data (Resnick, Adar, & Lampe, 2015). To deal with all of these challenges they come up with a solution and are currently developing an ideal platform for data collection and experimentation called *Mtogether.* The key point they advocate is that the method needs to be *person-centric* to avoid the problem of data transparency. This person-centric method relies on the subject installing software in order to collect information about their practice on a longitudinal perspective.

This method is a mix between traditional approach and use of digital trace. But then, they face the traditional issues of such methods like the problem of diversity in their sample, as well as the natural observation it is possible to obtain without interfering with the studied population, But as they said:

> The research community will benefit from continued reflection on what we can hope to find under different lampposts
>
> (Resnick et al., 2015, p. 197)

This work takes another perspective but, as well as Resnick, tries to be consistent between the type of data obtained and the answers it tries to find and develop the theoretical background from two angles, unified with one problem. The aforementioned authors talked about it but it is not always explicit in the literature. This problem is the central point, or at least should be, of any study on digital traces produced by Social Network Sites: *people*. It seems common sense to state that *people* should be studied when *social* networks are the object of inquiry. But since computer scientists have decided that number is what is powerful and at the same time, sociologists slowly lose grasp on that object, that central point, *people contained in the social*, is often eclipsed by the number of traces collected. At the end, the manifestation of these people, is at the very last chain of the research's process. As seen, it is possible to categorise this approach as *best engineering basket* and not necessarily see it as an entirely negative idea. At the same time it is possible to re-focus on what sociology does, *best science*.

As soon as the goal of the research is set, the people, a course of action with a clear objective, meaning a clear definition of the target, is needed. This is what these two following sections will do.

To define what *people* means under the specific realm of digital traces, a negative definition can be taken. Instead of focusing on the lamppost, the focus will be where the light is off[46]. By following the remarks raised by the above authors (and many others before and after them), it is known that all information is not accessible. This lack of information did not inhibit sociologists from inventing innovative ways to access it, to go to it (Resnick et al., 2015, p. 204). The problem here is that the lack of information is specifically about socio-demographic information, one of the most important pieces of information when it is about people, at least as soon as a sociological approach is taken. Without this information it is impossible to tackle issues about representativeness, knowing who are the *people* studied and how the inherent differences about people impacts the studied phenomenon. However, this problem is inherent in specific aspects of knowledge production. The section will

[46] As with the drunk, there is more chance to find the keys in the dark than under the spot of light, but everyone has already searched under the lamppost.

show that, not only the issues are more an issue of specific theoretical position, the importance of socio-demographic information is also a missed factor in the importance of context in the research.

To say it differently, take the scenario of a sociologist wanting to study workers in a specific factory to show the impact of the management system over the power diffusion among workers. If that study shows that in fact it is the maintenance team that hold the power, more than the director, despite a centralised system (Savage, 2010), no one really cares about the demographic information here. This type of theory uses the individual rather than the system to explain the game of influence and power and finally ends up with the explanation of structure. The system is a human production, a social construct that is self-limited in the social relation between its members and cannot make sense outside this network (Friedberg & Crozier, 1977).

This work does not follow the French organisational sociology[47], it is an example of how theories in sociology do not necessarily go exclusively for the demographic information and they can find other ways to create valuable information. It is true that sociology has been the first champion in the use of statistical methods and the need demographic information (Friedberg & Crozier, 1977, p. 50), but it is not the only angle taken. Sometimes that type of information is irrelevant or incompatible with the angle developed. But to *replace* the demographic information by something that make sense, a comprehensive approach is needed. The context where the social is created becomes as important as the social itself.

That it is why the second point is developed, not about the *people*, but about the *place, the space* where they interact. The *Site* in *Social Network Site.*

Resnick et al's approach is to take into account the site-specific definition to study web-based interactions, the argument here, is not starting with this difference but rather trying to find the core definition of every Social Network Site before developing the specifics. The difference in approaches between their work and this one can be explained by the fact that they are trying to see what are the flaws in the research within digital traces, whilst here the accent is on the flaw of research in digital traces using inadequate perspectives.

[47] The authors refuse this label and prefer to mention their work as the “sociologie de l'action organisée” – “sociology of the organized action of people”

In Social Network Sites, and if the focus is on people, it is possible to state that the most important aspect of this websites is the profile. Unfortunately, this importance is inversely proportional to the extent of it being studied. A clear definition of the profile, on both its technical aspect and abstract definition will give the missing *context* to study the *people*. Not why they use the Social Network Site, not who are the people that use the website, but *who are these people within the website*.

From these two approaches, it is possible to bring them together and define what is possible to study, in terms of available data, what is studied, in terms of object, and where this data is produced, in terms of context.

From there, a specific methodology will be developed in order to put in practice the ideas developed in the last point. This *ideal* situation will be tempered by the reality of Twitter and the limits imposed by the API and other factors, making every decision a movement toward the construction of the reality out-there.

4.2 Lack of information: The socio-demographic issue

Often, the data is incomplete. For instance, a criticism often raised by social scientists is the lack of socio-demographic information and the associated issue about the lack of representativity. It is possible to argue that this issue is a misunderstanding from sociologists about the strength of digital trace and a temptation to use old techniques with new data without developing a deep understanding of its specificity at first. The first reason is a clash between different epistemological perspectives. The statistics and the 4th paradigm's logic are different and not necessarily compatible. The second reason is about the nature of digital trace itself as a measure of activity and the different approach that socio-demographic categories represent. But before that, the understanding of the problem of representativity, specifically to Twitter, is needed.

The data often lacks information about characteristics such as the gender, the income or the geolocation. Moreover, when this information is provided, there is no direct method to ensure the validity of it. Even when the information is given, it is difficult to know if the user tells the truth or not. Besides that, it remains the problem of abbreviation and pseudonyms, making any extraction of information more difficult. Either there is no information accessible (for ethics reasons or because it is not provided), or this information is not reliable. This type of information is essential in a context of using Twitter in relation to external events. But, SNS are worldwide and the problem of socio-demographic information in this case is not only due to the lack of it, but also because people are talking about worldwide events even when they are not relevant to the external event itself, in term of geo-localisation for instance.

But another problem within SNS is, to use the famous Peter Steiner's quote: "*On the Internet, nobody knows you're a dog*". Then, using data generated by a dog will be clearly an issue when the goal is to predict, for instance, who is going to win the next elections in The Netherlands, since dogs cannot vote in this country. Therefore, to remove dogs from the dataset, or more seriously, everyone (people who cannot vote for various reasons) and everything (companies, bots, …), socio-demographic information is needed.

Gayo did a meta-analysis to see if the studies using Twitter data were efficient in their prediction or if it was only due to random information (Burrows & Savage, 2014). He discovered that the power of the model to predict election results did not predict better than random predictions. One main issue is the lack of socio-demographic information. Without knowing who tweets, it is hard to know if the data is relevant or not for the specific election. Therefore, the cleansing of data based on socio-demographic information is essential to ensure that the study uses appropriate data.

He did mention some studies, which have applied data cleansing instead of just taking the entire dataset they could collect. The cleansing was based on the language used (when it is feasible, cleaning on the English language is obviously impossible as it is spoken worldwide, while filtering on Dutch seems more reasonable) and the location of the tweet or the user's profile who posted the tweet.

Other studies exist which, instead of surveying the population of specific SNS, try to predict with other methods, such as language, to identify the age (Gayo-Avello, 2012) or the gender from the name (Nguyen, Gravel, Trieschnigg, & Meder, 2013). Others developed a classifier to extract the ethnicity, gender, geographic location, language and race by using the name, the first name and the location when it is provided (Liu, Al Zamal, & Ruths, 2012; Liu & Ruths, 2013), or even the occupation ((Bergsma, Dredze, Van Durme, Wilson, & Yarowsky, 2013).

But, even if it is possible to deal with socio-demographics information, it remains a double issue from the representation of any data collected through Twitter and consequently a bigger issue when it comes to predict external events.

An inherent problem is the over-representation of some categories of population in Twitter population. Not only is there a lack of socio-demographic information but some categories of population are also over represented in the users of SNS, for example in terms of race and ethnicity (Sloan & Morgan, 2015; Sloan, Morgan, Burnap, & Williams, 2015). The Pew Research centre surveys the proportion of Twitter users among American Internet users with variables such as age, education, gender, race and urbanity. The most typical Twitter user is a Hispanic woman aged between 18-29 who went at least to college and lives in an urban area (Mislove, Lehmann, Ahn, Onnela, & Rosenquist, 2011).

This issue of over-representation gives more visibility to some categories of people in comparison of their weight in the *real world*. An important issue when the matter is to predict social phenomena based on the age and the location such as elections.

Tesco faced a similar problem when they realized that one of the cluster was loosely defined in comparison of the others. They found out that this cluster contained two different socio-economic categories of people and couldn't discriminate them on their habits. Later, they were able to discriminate them based on the time of their shopping habits rather than on the items they bought; they could also distinguish divorced and single people.

The lack of this specific information makes several assumptions impossible about who the people are and what to expect from them. It is also more difficult to connect this information to other previous knowledge gathered from different methods such as census and surveys. The census methods collect this information to categorise the population and make sense of the finding, being sure to know who is the respondent, to which category they belongs and finally being able to compare their behaviour/attitude/... collected in other categories. But the difference with SNS such as Twitter is the link between the two types of knowledge. When people extract information online, the idea is to extract every useful piece of information from active users. When they compare that to external events such as elections, the information is based on sampling through survey or census and this sample is itself based on the previous knowledge about the population. The main problem about linking online and offline activities is that the two methods work differently. While surveys are mainly based on sample and the central concept of representative sample (the way to handle complexity is to reduce the number of entry), Big Data works under the assumption that volume is better because such knowledge of the composition of the dataset is unknown before the analysis. The inherent problem that the researchers face when they try using Twitter to predict behaviours within the virtual world is the collapse of two different approaches based on different logic. In conclusion, the Big Data has an advantage but, depending on the resources to date cannot, be used to answer everything. At least, if Big Number is the main quality researched.

But what is a profile and how to define it? The socio-demographic information is not present and it does not seem feasible to to expect such information about the users. In fact, a redefinition of the profile is needed to be able to understand what valid information is available and what are they supposed to represent. The next section will try to give another approach to the profile than the usual equation *profile=person*.

4.3 Definition of the profile on Social Network Site

Twitter can be summarise to the type of action it allows to its users. First, the users create a profile with some basic information that are going to be shared publicly. Then from the profile, the users can post/remove tweets to themselves. The other possibility is to add another user to their friends lists. It will allows them to see the tweets that the other user publishes. And that’s it! Some variations around these two main actions exist, such as retweeting, favouriting a tweet, blocking another user, adding it to a list…[48] I am not saying that these different actions are not important, they are even essential, but they are possible because one of the two initial actions has been taken by the users themselves or by another user. There is no Retweet without tweet and there are no followers without linking profiles. It is possible to consider every Twitter profile as a composition of these two fundamental basic units; *tweets* and *links*. However, the apparent simplicity of mechanism of exchanges, restricted to message and friends lists, becomes a complex network as soon as a closer look is taken at them. This complexity emerges from the imbrication of these

[48] Worth noting that the exchanges are more complex than that, as Twitter adds or removes some interactions, like the lists of users, the ability to favourite a tweet. These are important too but did not seems to be central at the time of this work. The situation may have evolved since.

two bricks rather than on their composition. A profile is,visually different than a tweet, have different purposes, the first one is there to represent user, the second to transport a message from the user. However, at several levels the profile and the tweet are identical.

The first similarity is technical and comes from the translation from the API. I have shown during the chapter 2 the structure of the tweet as a JSON file. The profile is exactly the same, except that the information contained in the JSON is about the profile rather than a tweet[49].

The second similarity is more on the consequences of their interlinks. The Twitter profile is itself nothing more than a trace of these communications. Adding a temporal perspective, a Twitter profile[50] is a crystallized version of previous social activities. It then becomes the reference for any further activity and create a context that describes the user. Considering the Twitter profile as a context created from past digital traces implies an evolution over time. This dynamic construction is done with the two basics units and therefore follows two different mechanisms.

The part of the context constituted with links evolves with the adding-removing actions. The links can be either followers or friends. The followers are the Twitter profiles that follow the account. This list is outside the control of the user but represent the potential audience of any next tweets. A contrario, the list of friends is created with an active action from the profile's owner and represents the potential exposition to new tweets. Of course these two contexts can (and do) overlap, when users are following under the reciprocity principle, especially if the users belong to the same group of interest. On this matter, the geographical distance and the homophily are strong predictors (Duggan & Brenner, 2013; Sloan et al., 2015). Gallos et al. decomposed the reciprocity with variables such as gender and age and found different patterns (Hopcroft, Lou, & Tang, 2011). Overall this reciprocity is a strong predictor of following behaviour, explaining why followers and friends networks are overlapping.

[49] In fact, when a profile is retrieve through the API, it is possible to retrieve the last statuses of the user, which is the full tweet in JSON. And when a tweet is retrieved, the information about the profile is possible to get too. This makes the trace itself identical, rather than similar, between the tweet and the profile. They both contains the other.

[50] Except the name, picture profile and other information such as location and description

The part of the context containing the tweets is an aggregation of small contexts themselves. As seen earlier in the chapter, the tweet contains more information than only the text or the mentions and can be a link to other profile or hashtag.

A profile, here taken as a tangible and finite unit is only another network, with endless links to other account which are themselves linked to other accounts. There is no global and local on Twitter, it is impossible to define a profile as a finite entity as it is only a temporary collection of other profiles and tweets.

This vision of an Ouroboros has a technical justification for its existence but makes it difficult to define where a Twitter profile starts and ends and where and when an action is effectively the social activity or a part of the context where the social activity takes place. In fact the openness and the endless link to other part of network is, in a sense, the perfect example of what Latour calls an open network.

To summarise: The contexts are the traces of past social interactions aggregated in a same place, while the actions are the social interactions being performed from this context.

This apparent confusion results from the flattening of the digital trace and following the social in formation (or the tweet creating the link between users). A profile is a tweet and a tweet is a profile. Not because they contain the same information (they do not), but because they are technically identical, a single point of connection that contains its own meaning, its own context. This context is folded into a link to other profiles and tweets but gives the possibility to track it down, to follow and map it.

This situation, created on some aspects by the technical requirements, leads to a new realm where the by-product of the sociality becomes the central point of attention. It is the consumers basket for Tesco, the big data, and so on. These artefacts are a manifestation of their use, not a translation of what they are in a sociological perspective. In this idea, the profile is a direct manifestation of the idea of connectedness that Latour oppose to the idea of zoom (Gallos, Rybski, Liljeros, Havlin, & Makse, 2012). Profiles can link to another point and form a network of contexts. These contexts can be bigger than the profile that is linked to it because it is a collection of profiles, or a single profile with lot of followers or friends. This context is somehow more important, bigger in the sense of the number of links that it contains and in turn the number of other contexts it is linked too, same idea for a tweet containing a popular hashtag. That hashtag will be a link to a bigger community/audience/…

The size of a community of a hashtag is reached by opening the link, by unfolding that context and following every mention or user, but there is no zoom in the sense of getting a higher level, a macro perspective versus a micro perspective. All links and contexts are flat and sometimes folded in one link that it is possible to explore by expanding it, by tracking back the list of followers or the history of a Retweet chain (Latour, 2005, p. 187). So not only is this the perfect idea of a connectedness and oligopticon materialised for direct interaction between profile users, but unfortunately this illusion of a perfect translation of these concepts does not hold the reality of the APIs (which is detailed in the the Methodology Chapter).

Furthermore, even if science and technology make traces more visible and lets everyone see the lines that were barely visible before, […] (Tinati, Carr, Hall, & Bentwood, 2012), some actions are still invisible. Some conscious activities are beyond the scope of digital method, such as reading, browsing other tweets or users' profile. These activities do not leave a trace, or at least in a tangible (as long as the numeric can be tangible) way that it is possible to collect. The problem is that even if they don't leave direct traces, these actions have an impact on the network and on the other users on others aspects. On Twitter, the consequences can be either on the tweet or on the link to other profiles.

On the tweet itself, the external influence on Twitter is manifested by people publishing links from other websites. In that case, it is traceable by searching for URL mentions. Myers Zhu and Leskovec showed that 29% of URL mentions come from external sources (rather than retweet) (Latour, 2005, p. 181). While this external influence is important to measure when the interest is about the network itself, it is less successful as an indicator when the focus is on the individual or a subset of it.

Another degradation of the data is the missing information by the users. People do not necessarily see all tweets, their exposition is only potential. The impact of any tweet will be minored by this absence of some receivers when the tweet is emitted. Here it is only a problem of scale of impact.

The last activity that is not leaving a direct trace concerns the exogenous activity when users decide to add or remove links. The reasons can be the use of external services (such the ones that give recommendation of other users to follow), read external websites with the users mention or the use of a search engine (internal to Twitter or Google when they have agreements for indexing the tweets). All these reasons will be missed and the reason why people follow each other will not be unveiled completely. However, as showed by Antoniades and Drovolis, the endogenous reasons to follow someone[51] make up for 42 % (but only 18% is explained when it is about unfollowing) (Myers, Zhu, & Leskovec, 2012).

These cases show that the omniscience brought by digital traces has some boundaries. The solution could be to trace the information through mixed methods but the present works redefine the object, taking into account these limits, rather than trying to collect all available information. The lens used will always limit the field, here the lens is the definition of the activity and its materialisation in a reachable form. An activity is any recordable voluntary action that takes place within a social context and leaves a digital trace that is visible for the researcher.

However, even if different mechanisms of message diffusion co-exist (the context is within the tweet, or the context is within the profile), only the mechanism of tweet containing information is so far extensively studied. The unique use of Stream API gives great information about how a message evolves within its own context and allows us to collect an extensive amount of information. But it is impossible to know about the evolution of the user's context and their activity with that method.

4.4 Definition of Social Network Site

Social Networks Sites have evolved since the first introduction. They have different audiences, goals, interfaces, features, philosophy and so on. With the success of leaders such as Facebook, Social Network Site drifted from being a specific type of website to a concept applied to any online (and offline) activity. If you play, run, work, or even weigh yourself, chances are that there is an option to share your activity with friends, to create a list or to browse other people who use the same device/service/tools, and so on.

Social media involves, to different degrees, the disclosure of information. This disclosure can take several forms. It can be fully disclosed, completely private, or be a mixture of both. But at some point the data is publicly available and accessible to a

[51] They define an endogenous reason as the exposition to a retweet followed by the action to follow the original author of the tweet.

wider audience than traditional chats or email exchanges. This disclosure can affect the message sent or shared with the network or affect the profile information. Depending on which aspect is easily disclosed, the status of public-private will vary.

Echoing the diversity in Social Network Sites, several definition can be found in the literature. In their article, Kietzmann et al. describe a social network as complex honeycomb :

> a honeycomb of seven functional building blocks: identity, conversations, sharing, presence, relationships, reputation, and groups. Each block allows us to unpack and examine (1) a specific facet of social media user experience, and (2) its implications for firms. These building blocks are neither mutually exclusive, nor do they all have to be present in a social media activity. They are constructs that allow us to make sense of how different levels of social media functionality can be configured.

(Antoniades & Dovrolis, 2015)

However, this definition presuppose some functionality and are less a description of what a social network site is than an a priori construction of the object itself. On a better approach, Bucher definition has the advantage to put an accent on the human-nonhuman component of social network in her article:

> In the Facebook universe, users need not be humans connecting with other humans (see, for example, Kendall & Zhou, 2010). Users can be business pages, songs, or newspaper articles. Being social simply means creating connections within the boundaries of the system. Every click, share, like, and post creates a connection, initiates a relation. The network dynamically grows, evolves, becomes. The network networks. The social in social media is not a fact but a doing. The social is constantly performed and enacted by humans and non-humans alike.

(Kietzmann, Hermkens, Mccarthy, & Silvestre, 2011, p. 243)

This definition is only about the social aspect of the Social Network Site. Moreover, it does not help to build a methodological definition of the SNS.

Another example is the definition of Social Network Site:

> We suggest that social media platforms are better seen as posttransactional spaces that compute and trade the expressive and communicative social fabric they engineer
>
> (Bucher, 2015, p. 2)

This definition offers the advantage to highlight the expression as the result of the engineered communication's options offered but the Social Network Sites. But the accent on the trading aspect has to do more about the type of social platform they were studying (a platform where users tags and mentions which items and brands they want to buy) rather than an intrinsic description of all SNSs.

However, these definitions are more about the comprehension of the Social Network Site after they interpret it within their framework, or before their analysis, in order to shape the research within a specific boundaries.

This work will do the same but with an attempt to have a more descriptive approach on what a Social Network Site is from a user perspective. To achieve that goal, I think the best method is to start from the earliest definition it can be found and rebuild a more generic definition from it.

To do so, going back in 2007, an eternity in web time, it is possible to give a more simple definition, although a riskily outdated one. Boyd and Ellison gave a good starting point:

> What makes social network sites unique is not that they allow individuals to meet strangers, but rather that they enable users to articulate and make visible their social networks. […] While SNSs have implemented a wide variety of technical features, their backbone consists of visible profiles that display and articulated lists of Friends who are also users of the system.
>
> (Alaimo & Kallinikos, 2017, p. 177)

This definition is based on profile information and network connection, both under the principle of visibility. Without necessarily stating that the previous definition is erroneous, I think that the main characteristic of a SNS shouldn't be a separation between profile and network, but profile and message. Several reasons to shift the distinction can be highlighted.

This distinction will help to be more generic toward different types of SNS, it helps to make a distinction between activity and context and lastly can explain the methodological consequences of different types of information disclosures.

The user's network is information disclosed on their profile. It represents their audience, in case of an asymmetric SNS, or a friendship in the case of a symmetric network. But is not the only way to communicate with people. If someone publishes a post on Facebook, it reaches the audience composed by friends (and usually friends of friends in the case of an interaction and depending on the privacy settings). If it is a tweet, it reaches the audience composed by the followers. The bigger the network of followers, the biggest the chance to reach large amount of people. But when the message reaches the others' timeline, it is impossible for the user to know who actually sees and reads it.

A study done on the Facebook network showed that more people read, or are at least exposed to the post than the user actually realises (D. M. Boyd & Ellison, 2007, p. 211). On Twitter, no such study is available but researchers, by using a website where they asked Twitter users to rate their friends' tweets show that only 28% of them are worth reading. It is not giving any information on the reading rate but gives the value of the message itself on Twitter and that it is not because lots of people read the message that it is something interesting. Is it important to know if the user is read or not, or has an impact and how to measure it? Both are important questions (not necessarily in sociology but at least in other fields).

But on Twitter, as shown in the section The tweet – As an obligatory passage point, the tweet itself is a more complex digital trace than it can appear at first glance. One aspect of this complexity is raised by different opportunities to communicate. The hashtag, can be used to tag the content of a message such as the tag's use in other SNS like Flickr, but more importantly, it can also be used to reach a community (Bernstein, Bakshy, Burke, & Karrer, 2013). In the latter case, the users try to reach another audience than their network

The hashtag offers the possibility for anyone to see the tweet within a specific timeline composed of all tweets containing this hashtag. This timeline, is a new network, more ephemeral (the time people are using the hashtag), but it is more active and more interactive within users than on other websites where the hashtag is used as tag content (Yang et al., 2012).

What is important here is that the network of Followers/friends stops to become relevant for the information diffusion of a message. The audience reached by the message is not limited to the network anymore. In fact, the message itself contains its own context, its own audience through the hashtag.

The information diffusion of a message then follows two mechanisms, one is through the network of Followers/friends and the second with the hashtag, or words, included in the tweets. In the first case, the context where the message is spread is within a potential audience known in advance by the user. In the latter, the audience is more dynamic and less known in advance than in the case of the network of Followers/friends.

Of course, when a tweet is transmitted, both mechanisms of diffusion are active, within both networks. The network information is therefore only relevant as profile information about the user who emits the tweet, more in a line of popularity or social capital.

The network information is therefore only relevant as profile information about the user who emits the tweet, more in a line of popularity or social capital.

The profile is composed both by the network of Followers/friends and other peripheral information, which is more static (such as the name or the unique id), but also with previous messages or any other recorded activity.

> A Social Network Site is a network of **actors** jointly engaged in **activities** with other users. These activities leave **digital traces** that are crystallized in a profile. This profile is the user's **context** of any further activity.

This personal definition has major changes from Boyd's definition (Z. Xu & Yang, 2012), it is more suitable to be adapted to a larger variety of Social Network Sites and is based on less preconceptions.

The first element to be removed is the idea of the *list of friends*. This idea of *friends* does not necessarily represent the asymmetry in SNS such as Twitter. Moreover, the list of friends is only one type of link possible. As developed before, the list of friends is only one possible link. The tweet also contains a set of connections within the mentions and the hashtags. Both represent a network as valid as the list of friends, but are not contained within any profile. Users that have no list of friends still have a profile and still can access and participate on the SNS, as well as being members of discussion around hashtags or mentioning other users.

The second element removed from Boyd's definition is the idea of visibility. Again this visibility is not necessarily a required feature to be a member and actively participating in Twitter.

Instead of these elements, the personal definition is centred around four notions, *actor, activity, digital traces* and *context.*

The notion of *actor*, besides being closer to the ANT vocabulary, reflects the variety of possible *individuals* used in Boyd's definition. Twitter is not only composed by individuals. Organisations, that communicate with their public, personalities, who have a personal assistant using and managing the account, bots, who spam erotic pictures and scams on popular hashtags, dogs, which are trying to vote in Netherlands… The list is endless and can be composed of humans and non-humans, algorithms or groups of people. The notion of *actor* avoids any pre-made distinction and can reflect the complexity of what an actor can be. This is also important to avoid biases on the ontological definition of what a profile is, or at least what is not: a profile is not equal to an individual!

This first assertion is supported by the notion of *activity.* It does not assume the reason why the SNS and why people are using it, like the *meeting strangers* does. Actors can use SNS for various reasons and behaviours, the activity can be sharing information to strangers or spying on acquaintances but they are the only information available within the SNS, the action of the users. This is why the occupation, gender or social classes need to be predicted with algorithms.

And all of them, regardless of their goal, leave *digital traces*. These digital traces are the only proof of presence and past activities. They form the profile, not the list of friends but the entire digital traces the SNS allows to collect and to concentrate in one place, the profile. This is the context which represents the actor. The trace being by definition in the past, any further activity starts from this context and allows other users, or researchers, to localise the new production of social interactions within that context.

This is how the profile needs to be conceptualised in research. Not a translation of an individual (or an organisation, or anything else), but the collection, crystallisation of the past activities created and shared on the system.

That personal definition represents more precisely the principles underlying any SNS without giving erroneous definitions based on a preconception of who are the users and what are their intentions.

It also redefines the profile and what is happening on the SNS as digital traces, which redefines the object of any study on SNS. That has an important consequence. Without refuting the possible association between a person and a profile, it stipulates the limit of the available information. There is only activity from an actor, this is the valuable information present the past activity and the context of where it comes and what it is. Any additional information, such as who is the actor, needs to be accessed with other methods or it falls into the irony of prediction[52].

The next chapter is the methodology. I will base the development of the idea of context and activity to Twitter specifically (and avoid a sterile debate on a too big picture). This task will be the opportunity to describe potential solutions (and often new issues). It will be used as a demonstration of a possible response to the actual crisis we are facing in front of all the new opportunities. This response takes into account the development of theoretical perspectives on the importance of context and the need of a dynamic methodology in the data collection. This *ideal* situation will be tempered by the reality of Twitter and the limits imposed by the API and other factors, making every decision a movement towards the construction of the reality out-there. Computer scientists have decided that number will be what is powerful. I think these numbers are important, but in more ANT perspective, I think the process on *how* this number is generated is equally important. Unveiling the decisions toward the data collection and data creation is what can make a change and getting back the control over the process of data creation rather than leaving the number speaking in isolation.

[52] The social classes and socio-demographic information was used to predict the behaviours, now we use the behaviour to predict these categories.

5 Methodology

5.1 Introduction

The previous chapters highlight how Twitter research is a method assemblage and how the deployment of any tools or study on the object only casts a partial light on the social media. As a method assemblage, the constraints imposed by Twitter through the obligatory passage point to the API, the need for research to sustain numbers in order to stick with the 4th paradigm, and the specificity of methodology employed, often lead the research to a specific direction.

I described these processes and ended up with a personal definition of Social Network Sites that gives importance to two aspects, the activity and the context. However, as any method assemblage, a definition alone will have no connection to the empirical ground.

> […] we as researchers are part of a world that is constitutively multiple, relational and emergent. So, to engage with our 'empirical object' is to enact them, but to engage is also to be enacted. As researchers we are, to some degree or other, also performed through our method assemblage as it interacts with other assemblages such as those inhabited by the 'object of study'.

(D. M. Boyd & Ellison, 2007)

This assemblage of the study object is represented by the Twitter API, which type of data it allows to collect and with which limit of frequencies, the JSON file, which type of data it contains. Of course, these assemblages are made with other assemblages, such as the business model that also ultimately decide which information is possible to have, or the technical aspect and cost of maintaining the API (which lead to a reduction of the access over years). However, the method assemblage of this work is the translation of the activity and context into the *actual* activity and context. The definition of SNS given earlier was an enactment of this research through the empirical object. Now, the description of the API and JSON object will rearrange this definition and create a different assemblage than initially.

This work is going to use these objects to enact this thesis. In that way, not only the API and JSON object will deploy agency, but the initial goal of this work too, trying to *plug* into that already formed assemblage. Through that plug, both object, the perspective of this research (definition of activity and context) and the digital trace left by the API and the user will create a new assemblage. A new object, or quasi object, that will deploy its agency through the whole research process.

Concretely, this section redefines what context and activity means in term of profile representation and what type of information it is possible to get (the different count of the followers/friends/tweet lists). That will, at the same time, create rules to define different level of users. These levels are only existing for the purpose of the methodology.

These practical definition of the object of inquiry will then be translated into an actual method of data collection. The digital traces does not exist for this method assemblage until they are connected to the research. And finally, by taking into account the own agency of the quasi-object *research*, the ethical aspect of the data collection will also be considered at the end as it is itself a consequence of the solution founded in the methodology.

5.2 Context and activity

The theoretical foundation of this methodology is the evolution of social interactions within the web based societies and the theories as shown in chapter 1. We are evolving, to a more accelerated, fluid world where instability is the main concept and the ephemerality of our relations is the norm. This is at least, a conception given by theories developed earlier (Michael, 2016, p. 134). However, these theoretical assertions are hard to test empirically, mainly because they are constructed to highlight the difference between a *before* and an *after*, artificially created for the purpose of the demonstration.

At the same time, methods of inquiry, such as advocated by ANT, give the activity of human interactions a central role to any studies. Human interaction is not what is studied directly but it is rather the traces left by these interactions, showing the social being in construction and in formation.

From this conception of the social object, it is possible to create a new methodological picture of the social questions raised in chapter one, such as, is the more active and changing environment having an impact on people who are living in it. In other words; *if an actor is living in a rapidly changing world, will that actor will also change accordingly?*

The idea of activity and traces, extensively inspired by ANT, can find a practical and already everyday application in Social Network Sites. The advantage of SNS is inherent to their nature: social information is recorded and accessible.

As defined in chapter 4.4 Definition of Social Network Site. These profiles are composed of traces left by previous activities and, in turn, are the context of any further user's interactions. Social Network Sites give all information needed to test the hypotheses of a "*more accelerating world*, *more changing relation*", not necessarily because they are also fast changing, but because they provide an access to this *social in formation*.

It is why SNSs provides a good opportunity to bring together these two sociological developments; a theoretical perspective of the evolution of the context by using a methodology based on trace of activity.

However, in term of research all SNSs do not necessarily offer the same opportunities for answering the above questions. They are different in terms of *complexity*, *access*, *population* and *public availability*.

All these aspects play a role on the decision of which SNS to study. In the present case, it is Twitter that has been chosen. One reason is the relative easy access to the data by using the API. This SNS also has the advantage to have public profiles rather than private or limited ones. Another reason, probably the most important, it is the simplification of the interactions on Twitter. When other SNSs offer a variety of mechanisms to interact: building their network by adding or removing other Twitter's profiles and communicating with tweets to these profiles (within or outside their network). This reduction in the number of types of interaction allows the research to be more focused on the consequences of the type of data collected rather than on the complexity of the exchanges.

The method presented in this work will track the evolution over time of a selected set of users, following who they are adding in their network, to who they are tweeting and who adds/removes and tweets to them. The measures collected represent this evolution of contexts and changing relations.

5.3 Definition of activity

In Chapter 3, the profile has been given a conceptual definition. This is the place where the activity is recorded and it is through the profile that it is possible to access the past and future activities, being a context itself and containing links to other contexts. The flattening of the profile links and tweets into open networks has consequences on how to define activity and context. The two are only differentiated with a time perspective. The context is the past, while the activity is the social in formation. From there, it was pointed out that the digital traces do not represent the entire possible activities and therefore impose specific limits to what it is possible to collect, and by extension, what it is possible to study. As a consequence, only the tweets and adding or removing someone is considered as activity.

This section is about the actual digital representation of the activity and the associated contexts. It will operate a zoom on the actual digital traces that constitutes the activity and the profile by going through the composition of the profile itself.

5.3.1. Profile representation

This section describes the required steps to build a *basic unit of analysis* based on the information available in the profile and the workaround to build a *trajectory of activity* for the profiles. They are both linked as the basic unit of activity is created by calculating the difference between two instants and this succession of instants is what creates the trajectory itself (see Table).

From count to activity

The first type of trace collected is the digital traces from the profile. The profile can be summarised as a list of *key-values,* exactly like the tweet[53]. The interest in these is limited to the following ones:

- *screen_name*: The name of the profile. That is the same value as the one appearing in the tweets and on the profile page. The collection of this information is for checking only. It is easier to visually check the data collection using the *screen_name*. Later this *screen_name* is dropped.

[53] Every information is stored on a key-value format, A key is the type of information stored, and the value, the information itself.

- id_str[54]: A number that is the unique identifier of the profile. More interesting than the *screen_name* because it does not change even if the user changes their profile name. This is used in the data collection as the unique identifier too. Later it is replaced by a random number to make identification impossible.

- *followers_count*: This is the number of followers the users has at the time of the access to the profile. This information is stored to measure activity

- *friends_count*: Same as for the previous key but about the *friends.*

- *statuses_count*; Needless to say that it is the exact same as the previous one but for the *statuses*, meaning the number of tweets the user has posted. If a user removes a tweets, this number decreases.

There are lot of different keys that may be interesting but these 4 keys are the only ones that are going to be stored for this work.

From list to context

Beside these quantitative (and static) values, it is also possible to access the list of links and tweets the user has. For the lists of followers and the list of friends, the information obtained is a list of identifiers (id_str) of every user who is following, or being followed by, the user. No other information is given but the id_str gives a way to access to the user profile information.

Again, it is a snapshot, therefore no change is accessible. The only way to infer that change is by taking the difference between two snapshots. The interest in these lists is qualitative and will be used later in the sampling technique.

The twitter list is more complete than just a list of identifiers. It is a list of full information for each element of that list. So there is no need for the followers and friends, to use the identifier to access more information. As described in The tweet – As an obligatory passage point in Chapter 3, the tweet is a complex JSON file, giving a list of key-values, similar to the profile information (but adapted to information about the tweet). This re-enforces the idea that a tweet or a profile is the same conceptual context and open network. They both contain information about

[54] This key has the same value as the *id,* except that the format is a string for the former and in an signed 64 integer for the latter.

themselves and link to anything mentioned in it. As for the profile, only a few key-values are interesting here:

- *id_str*: As well as for the Profile, this is an unique identifier for the tweet.
- *retweet_status*: If the tweet has been retweeted or not
- *mentions*: Any other twitter user account mentioned in the tweet
- *hashtag*: Is a list of any hashtag inserted in the tweet.

The more important information is then a link to another context. A link to another user's profile (the mention) or a link to an aggregation of tweets around a topic (the hashtag).

However, the three metrics only represent the state of the users at the time of the data collection. It is nothing like an activity measurement. In fact it has been used in many other ways. Hofer and Aubert even link these numbers with the perceived social capital. The *followers_count* to the bonding social capital[55], while the *friends_count* is associated to the bridging social capital (Castells, 2011; Wittel, 2001). Which mean that these numbers can even reflect a wider social capital outside the SNS. But they failed to represent the social in formation. That is why it is necessary to collect the data from the same profile several times, to capture the evolution of these numbers. It is then possible to build a measure of activity, a trajectory of it (see Table 2: Trajectory of activity).

id_str	type	t_0	t_1	$t_{\dots}$	t_{n-1}	t_n
123456	Followers	134	135	…	241	143
123456	Friends	734	666	…	934	1400
123456	Status	1300	1310	…	1356	1372

Table 2: Trajectory of activity

To compute a measure of activity with this type of table, it is simply taking the difference between two snapshots of the {followers, friends, statuses}_count. To take into account the different time period between two snapshots, the difference needs to

[55] They are using Putnam distinction about the two different type of social capital, the bridging one is inclusive and associated to weak ties, while the bonding social capital is more exclusives and associated to stronger ties (such as friends and family.

be divided by the difference of the gap between the two snapshots, to have a weighted measure in function of the time passed between them.

At the end of the data collection, each user has these associated vectors of these created activity measures. For the *followers_count*, it is not an activity from the users themselves, but from another user that decided to follow them.

From context to sample

It is impossible to get a representative sample of Twitter's users as the information is not available on Twitter (2013). And even if the information is available, it is not necessarily trustworthy. However, this problem is not directly a concern here, as there is no need to assess hypotheses based on socio-demographic information but to analyse digital traces within Twitter. This epistemological decision to exclude any other form of data that is not produced within Twitter gives in return more freedom on how to use the data generated. However, the limit of the API does not allow you to go on a full *Big Data perspective*, but can give the opportunity to use some advantages associated with digital traces and SNSs. First advantage, as the representativeness is not needed anymore, the sample does not have to be fixed in advance in order to respect this essential characteristic and to represent a larger population. The sample can change over time according to other rules of inclusion/exclusion, dynamically generated during the data collection. When research is based on tweets and hashtags, the sampling is dynamic in nature, the data collected is in function of the tweet published, there is no way to know in advance what the dataset is going to contain, tweet or users. But if the perspective is the users instead of a hashtag, the dynamic is different and puts in play what it is possible to call *contingent connectivity*. If the perspective is a tweet collection, the method is equivalent to a type of *fire&forget*: the tweet is collected as soon as it is published, it is stored and then the next tweet is collected, and so on.

With an approach centralised on users, the *contingent connectivity* is slightly more complicated. Users add but also remove other users but may add them again later, it is not a linear process. Therefore, in order to track these changes, it is necessary to capture this flow of activity that is composed of these two actions of adding/removing but that can apply several times for the same user. The hypothesis lying under this non-linear conception of the *contingent connectivity* is that these actions of inclusion/exclusion are the sole accessible information of attention from the users. This assumption makes an inference from the digital traces available, and the intention/state of mind of a user. It is trying to access the *eyes of the users* in Simmel's terms. Therefore, this *contingent connectivity* is a function of these activities and is supposed to represent a temporal attention. The method of sampling is capturing this concept of attention by collecting more data around it when it happens.

Adding data dynamically extends the possibilities to analyse the context surrounding the action of adding/removing. However, there is a moment where this focus on a user needs to stop otherwise it will be the same as trying to collect the entire Twitter's network. This is where the concept or *contingent connectivity* reaches its limit. Or said in more accurate way; there is no way, using the digital traces, to know when a user shifts their attention from the users they added or removed, if they ever does. In regards of this inherent limit of digital traces, the decision to stop collecting data about the context surrounding an added or removed user, is purely arbitrary, and the focus is stop after two days but start again if there is a new interaction.

These *temporarily of pertinence* and the resulting *temporarily focus* of collection of data according to the interactions creates a methodological need of three independent lists of users with different purpose and rules of sampling.

These ideas result in the creation of three different levels of users that are tracked down, in function of the interest and the interaction. They are sampled but with different time periods and with different granularity of details.

5.3.2. Type of user collected

Main user

This set of users contains the first users to be collected. They are the population of interest for this research. The way to collect them and the reasons to choose them is detailed in Chapter 6.1. As they are central to the study, this set of users doesn't change over time except for two reasons. If they set up their account on protected, it is impossible to access information, except for those who are following them[56]. They can also remove their account or suspend it. In these two last cases, the result is the same, the account is unavailable for data collection.

[56] On this setting they have to accept the request, more like the others SNSs.

For the users on that level, all the information described above is collected: *profile information*, *list of followers* and *friends* as well as their *tweets*. The collected information gives the activity metrics but also the context of where it occurs. This information creates the rules to dynamically sample the users on the other level. It will reflect their activity and interaction with other profiles and results in two types of samples.

The contextual users

This set is created by each user collected in the followers and friends lists that are recorded. They are representing the social context. Under the hypothesis the activity of the context influences the activity of the users, the profile information is collected, then an activity vector is created for each of these users. The list of their followers and friends and their tweets are not collected. This set of users changes over time and only contains the users that are currently followed or being followed at the time of the snapshot. They are removed from the set and no more information is collected when they are not following or being followed by any other user in the dataset. Therefore, as this set is the function of the activity of the other users in the dataset, it is dynamically created over time and changes between each snapshot.

These changes, adding/removing are the rules that lead to the third type of set, the activity users.

The activity users

This list depends on who is added or removed from the followers and friends lists from the main users. Under the principle of temporarily of pertinence, these users are considered as temporarily more important for the main users, at least for one of them. Therefore the data collection reflects this importance and collects more information about these users. Beside the profile information, as for the contextual users, the followers and friends list and the tweets are also collected (as for the main users).

5.3.3. Example

The logic of the different users' sets is described below with a fake example. It assumes that the data is collected four times.

Let the set of main users Main:

- Main_{t0} = {User_1, User_2, User_3}

First snapshot:

The first snapshot t_0 is going to download the information about these users (the profile information, the followers list, the friends lists and the tweets)[57].

After the data collection, it is possible to see that each of these users have a set of followers. For example:

- U_1 has the set Followers1_{t1} : {User_2, User_4, User_5}
- U_2 has the set Followers2_{t1} : {User_5, User_6, User_7}
- U_3 has the set Followers3_{t1} : {User_4, User_8}

From these sets, the set of contextual users is built by taking the combination of all the Followers sets minus the users that are in the Main set:

- Context_{t1}: (Followers1_{t1} ∪ Followers2_{t1} ∪ Followers3_{t1}) - Main_{t1} = {User_4, User_5, User_6, User_7 , User_8}

Second snapshot:

For this snapshot, t1, the set of Main remains the same, but the set *Context* is added to the data collection. The profile information is collected and stored for further analysis. The sets *Followers* for the set of *Main users* now are:

- U_1 has the set Followers1_{t2}: {User_2, User_4}
- U_2 has the set Followers2_{t2}: {User_5, User_6, User_7, User_9}
- U_3 has the set Followers3_{t2}: {User_4, User_8}

The User_1 has one less follower (User_5), while the User_2 has a new followers (User_9) since the last snapshot. Therefore the set *Change* for every users is created from this difference:

- U_1 has the set Change1_{t2}*:* Followers1_{t1} Δ Followers1_{t2}: {User_5}
- U_2 has the set Change2_{t2}*:* Followers2_{t1} Δ Followers2_{t2}: {User_9}
- U_3 has the set Change3_{t2}*:* Followers3_{t1} Δ Followers3_{t2}: {Ø}

[57] Only the followers is used for the example, but it is exactly the same for the friends.

Then the set *Activity* is built by taking the combination of all the Change set minus the users that are in the Main set:

- Activity_{t2}: (Change1_{t1} ∪ Change2_{t1} ∪ Change3_{t1}) - Main = {User_5, User_9}

The *Context* set has a slight modification, just to be sure the users that are collected for activity are not included:

- Context_{t2}: (Followers1_{t2} ∪ Followers2_{t2} ∪ Followers3_{t2}) – (Main_{t2} ∪ Activity_{t2}) = {User_4, User_6, User_7 , User_8}

Then the third snapshot works for the second ones but adds the activity set to collect their profiles and the followers-friends lists, as well as the tweets.

To summarize, the different Snapshots T = {t_0, t_1, …, t_n) have the different sets collected:

- Main User_{tn} = {User_1, User_2, … User_m)
- Activity_{tn} = (Followers_1_{tn} Δ $\text{Followers\_1}_{tn+1}$) ∪ ($\text{Followers\_m}_{tn}$ Δ $\text{Followers\_m}_{tn+1}$) – Main User_{tm}
- Context_{tn} = Followers1_{tn} ∪ Followers2_{tn} ∪ Followers3_{tn}) – (Main_{tn} ∪ Activity_{tn})

A flow chart of this process can be found in the Illustration 1: Flow Chart. Now that the data collection has been clarified, the way to collect this information needs to be explained too. Beside the necessity to understand the underlying mechanics, this exercise renders visible some choices that have been made and the reasons why they have been done. Collecting the digital traces to build the different sets described above raises a number of difficulties, mainly in terms of access to the information and the limit imposed on the service by Twitter.

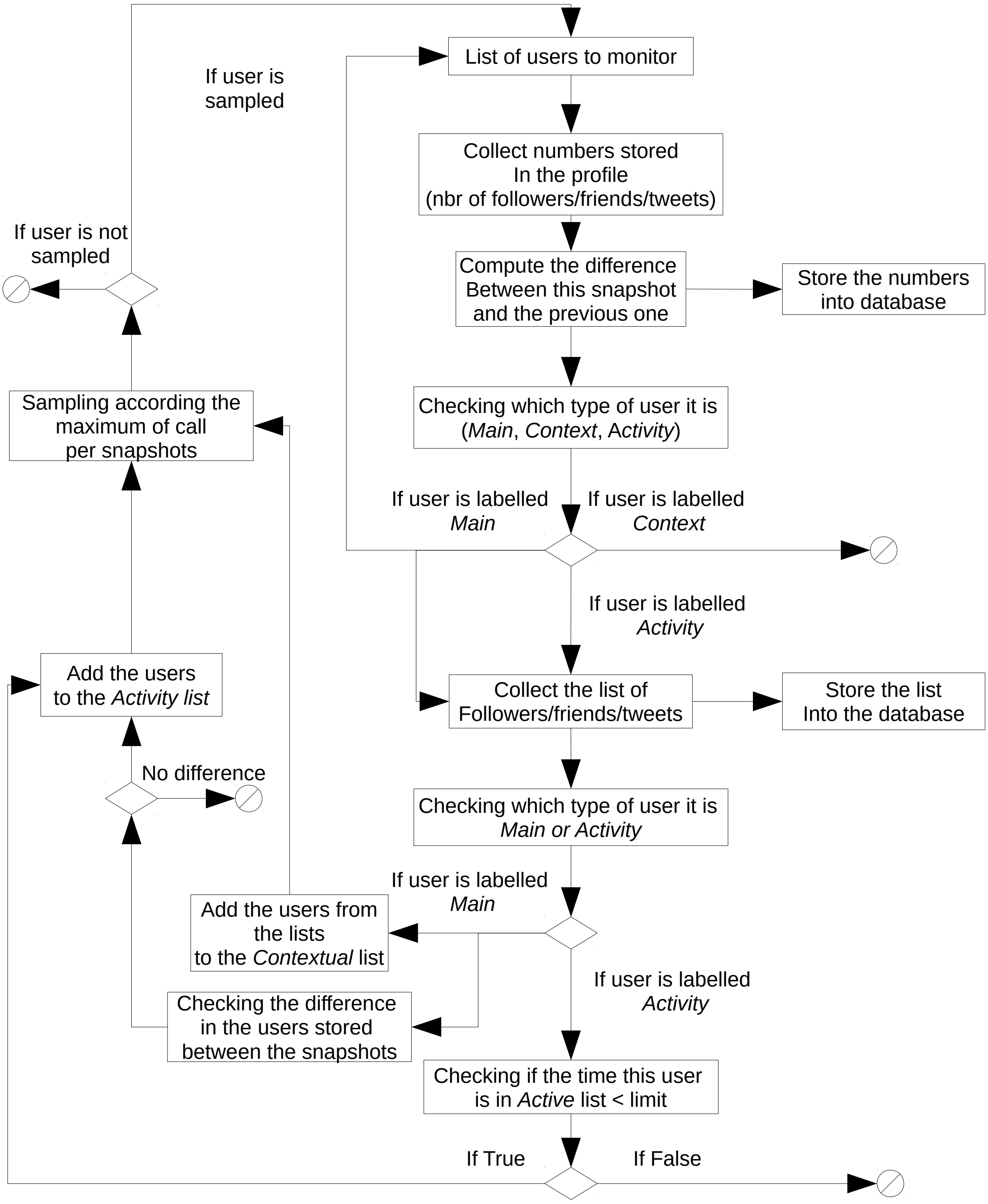

List of users to monitor
If user is sampled
Collect numbers stored In the profile (nbr of followers/friends/tweets)
If user is not sampled
Compute the difference Between this snapshot and the previous one
Store the numbers into database
Sampling according the maximum of call per snapshots
Checking which type of user it is (*Main*, *Context*, *Activity*)
If user is labelled *Main*
If user is labelled *Context*
If user is labelled *Activity*
Add the users to the *Activity list*
Collect the list of Followers/friends/tweets
Store the list Into the database
No difference
Checking which type of user it is *Main or Activity*
If user is labelled *Main*
Add the users from the lists to the *Contextual* list
Checking the difference in the users stored between the snapshots
If user is labelled *Activity*
Checking if the time this user is in *Active* list < limit
If True
If False

Illustration 1: Flow Chart

5.4 Use of API

The three types of information, *profile information*, *link lists*, *tweet lists* require specific API access to the REST API. Beside the fact they have their own limits, they also works slightly differently.

As seen in the Chapter 3, section 3.4.2 Twitter owns the right to modify or stop, basically, do whatever they want to, with the access to the data through their API call (and they modify it extensively). Therefore the following description is only accurate at the time of the data collection (December 2015) and not necessarily after.

Overall, the API works with calls and a window frame of 15 minutes. Within that time, the calls are counted and if a limit is reached the API just sends a *Pause message* rather than the information requested. After 15 minutes, the limits are reset and it is possible to get the information again.

Two types of identification exist, the *app authentication* and the *user authentication.* They have different purposes but also different limits[58]. For this work specifically, the App authentication has the advantage to offer a higher limit for the tweets and the links calls, while the User authentication gives more calls for the profile information. Even if there are different addresses to access different resources they share the same overall limit number of 450 calls per 15 minutes, no matter which information is requested. To obtain access to the APIs, it needs to get keys from Twitter by registering an app from their website, associated to a Twitter account. These keys are unique and make it impossible to call the API from several applications using the same keys (or from the same IP).

For this work, a mix of authentication is used to have a balance the number of profile and the number of links it is possible to requests.

5.4.1. The different APIs

User profile

To collect user' profile information, that is required for the all users set, the API resource used is the users/look_up call[59]. This resource allows us to obtain the details of 100 users per call using their id_str or their screen_name. The limit is 180 calls

[58] A chart is showing the differences in the limits between these two mode of authentications. Source: https://dev.twitter.com/rest/public/rate-limits

[59] Source: https://dev.twitter.com/rest/reference/get/users/lookup

every 15 minutes. Each call returns a fully hydrated users object which contains all information about a user profile as well as its last tweet.

The protected accounts are still accessible with this API endpoint. However, it has been decided to discard them and not store any data if the users decided to be on protected.

Followers/list and Friends/list

To obtain the followers list of a user, several types of calls are possible. For the purpose of the research the *follower/list call*[60] and the *friend/list call*[61]. These return a cursor object of up to 5000 *id_str*. The information returned is therefore a list giving the identifier to be used in association with the *user/lookup call* to get information about them. The number of call is drastically limited to 15 calls every 15 minutes (and 30 if the *app authentication* is used). So the maximum of followers (or friends, as it is two different calls) is equal to 15*5000, 75000 every 15 minutes. This number which can seem big is the real bottleneck of the method employed here, because it limits the maximum of users to 15 per 15 minutes (and could even been reduced to less than 8 for another reason as explained below).

Using the REST API forces to limit the size of the sample collected. Given the way it works it is only possible to fetch a limited amount of information. The bottleneck is the *friend/list call* (and the *follower/list* call) which limits to a maximum of 15 calls (if the users have less than 5000 links) every 15 minutes. To this low limit, another one put this number even lower.

If a user has more than 5000 links, it will take more than one call to fetch their list, reducing the total users it is possible to monitor. To be able to have a reconstruction of the profile, several snapshots are needed to trace the evolution of the network. One solution is simply to wait until all users' network profiles are being fetched before starting a new cycle. But doing this makes it impossible to control the time period between two snapshots, as the time will expand for every new user added. Therefore, to increase the number of users, one solution it is to extend the time between two snapshots.

[60] Source: https://dev.twitter.com/rest/reference/get/followers/list

[61] Source: https://dev.twitter.com/rest/reference/get/friends/list

A second limit occurs when a user has more than 5000 followers. A call retrieves 5000 *id_str* but if the users has 5003 followers, it costs 2 calls to have the entire set of followers (and 3 if the user has 10004 followers, …). It had been decided to limit the collection up to 5000 maximum. Not that the users that have more than 5000 users are discarded, but only their 5000 more recent links are collected.

However, even if only 5000 users are targeted, there is still a need to collect more than 5000 in order to do a comparison of the evolution of the network list. When Twitter returns a collection of users, the list is on a reversed chronological order. The first *id_str* given is the last one added. Therefore, it can show the lasts user's interactions. But, the problem with this order and the fact that it is impossible to fetch the entire list of users, is when the user monitored removes a user beyond the first 5000. If the user removes, let's say the 14025th link, but adds at the same time two users, the list fetched will have one element different than before but will not represent the proper change. Therefore, if a user has more than 5000, a second call is made to collect the second list of 5000 (or less). Then, it is possible to catch any difference in the first list of followers-friends but only for this 5000.

There is one remaining limit, if the user removes or adds more than 5000 users at once, it is still impossible to know which ones of them the user has removed and which ones the user did not. In this case, the record of change is discarded and only the new list is recorded. In short, more data are collected in order to do the analysis, but overall it is still a fragmented vision.

It remains a last problem with the snapshot's method. It is impossible to notice a hit change occurring between the interval of two snapshots. If a user adds or removes another user within that time, it will not appear on the record. Therefore, extending the time between snapshot shouldn't be done too much otherwise it loses its granularity and its ability to capture any change.

Even if these limits exist, this method allows us to capture the activity of the user but also its context. The limits are similar to constraints in other types of method of social research, interviews or survey, they do not stop us doing the research but it needs to be kept in mind when conclusions are drawn.

Also, they present the need to be flexible. It is possible to tweak them by playing with the number of clients, or the time between two snapshots or the number of links collected. Therefore it is possible to adapt them according to the priorities of the research.

Tweet lists

The tweets can be collected with the Stream API and the REST API. It has been decided to use the REST API with the statuses/lookup[62]. This API endpoint allows us to retrieve up to 100 tweets from one user per request. The number of requests for the user authentication is 180 requests per 15 minutes, while the app authentication can have 450 calls per 15 minutes. These limits are enough for this work as the main bottleneck comes from the links APIs.

5.5 Building a script

Now that the different types of rules to create set of users, and the different API calls have been detailed, it is time to get deeper into the deployment. Some important decisions have been translated here but basically all the constraints and the choices made have been explained in the previous section.

A program has been developed for this work, developed with Python[63] to interact with the Twitter API, and using MongoDB[64] to store the information into a database.

The structure of the program is schematised in the Illustration 2: Infrastructure of data collection. On a general level, a server controls every aspect implying, comparing previous records with new information collected, calculating the limit of the sample that it is possible to collect, storing the information in a database and pushing new data to be collected. The clients, on the other side, are querying a database to know which information to request to Twitter and are dealing with the temporisation when the limit of request is reached.

The server control is the most complex piece of that software. The actions taken are different if it is a profile information (and if it is main, activity or contextual user), a link or a tweet.

[62] Source: https://dev.twitter.com/rest/reference/get/statuses/lookup

[63] Python is a programming language. The version used for this work is the Python 3.4. Source: https://www.python.org/

[64] MongoDB is a NoSQL database, no particular reason has presided to its choice except the necessity to have a database that can handle several Gigabytes of data and millions of records. The version used for this work is the 3.2. Source: https://www.mongodb.com/

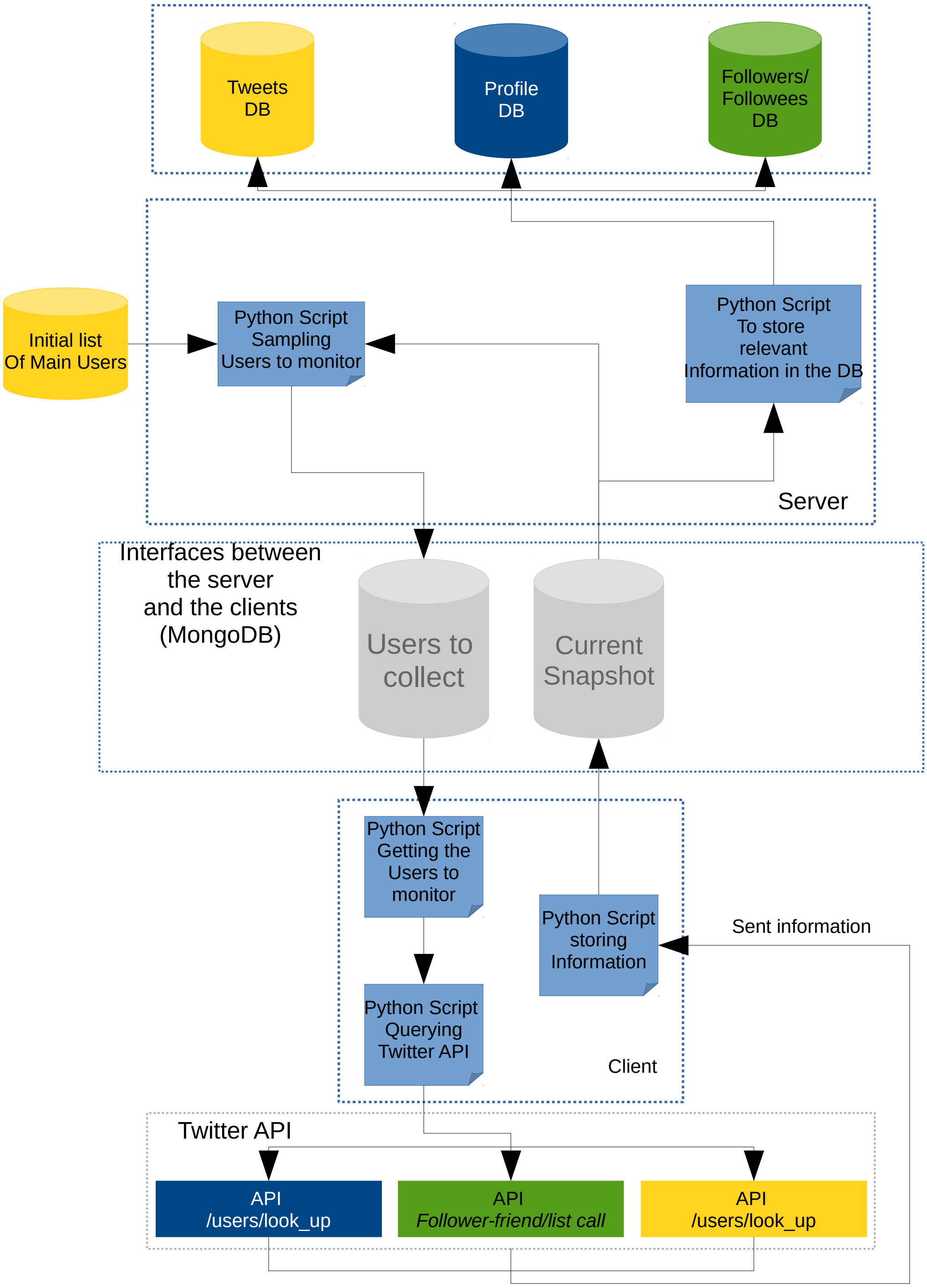

Tweets DB
Profile DB
Followers/ Followees DB
Initial list Of Main Users
Python Script Sampling Users to monitor
Python Script To store relevant Information in the DB
Server
Interfaces between the server and the clients (MongoDB)
Users to collect
Current Snapshot
Python Script Getting the Users to monitor
Python Script storing Information
Sent information
Python Script Querying Twitter API
Client
Twitter API
API /users/look_up
API Follower-friend/list call
API /users/look_up

Illustration 2: Infrastructure of data collection

For the tweet, the operation is quite simple. It queries the temporary collection that contains the new tweets and stores[the key-values needed (see section From list to context) into the permanent database, than deletes the records from the current snapshot database. For the users it requires more decisions and processes.

5.5.1. Creating the set of users

If it is a profile record, the first operation is the same as for tweets, selecting the right key-values, storing then into the database and deleting the record from the current snapshot database[65].

After that, the program checks if the profile is a user from the set of *contextual*, *main* or *activity* users. If it is a *main user*, it records the *id_str* for the next snapshot by mentioning the fact that the profile information, the followers, the friends and the tweets needs to be collected. If it is another labelled user (*activity* or *contextual*) it just deletes the records from the current snapshot database and stores the information into the long term database.

For the links, what the server receives is a list of *id_str* associated with the owner of that list. This lists represents the *contextual users* for the next snapshot. Therefore, it creates a record from all these *id_str*, labels them as *contextual users.* This label tells the clients that the profile information needs to be collected. The second step is to check if there is a difference between this new list and the list stored from the previous snapshot. If a difference is found (one or more *id_str* that are present in one list but not in the other one), it creates a record with these *id_str* and records it in the next snapshot database. This record is labelled as *activity user.* This label tells the clients to collect the profile, tweets and links information[66].

These new labelled users are not necessarily checked during the next snapshot (except for the *main users*) due to the API limits. These limits are stable and easy to forecast, but it is impossible to know in advance the number of activity or contextual users between two snapshots. A dynamic sampling can answer this problematic, by creating a set of rules to sample the users between the snapshots.

[65] The different operations are done at the same time, the server works while the clients collect the data. As soon as a record is added in the current snapshot database, the server picks up the record and process the storage, labelling and selecting the difference between lists when it is applicable.

[66] To avoid potential duplicates (*i.e. a user being labelled main user and contextual users*), a hierarchical order ensure that it is the highest level that is privileged (*main user > activity user > contextual user*).

5.5.2. Dynamic sampling

Two types of users have limits on the number of users that can be part of the next snapshot, the *contextual users* and the *activity users*. For the *contextual users* the limit is the one imposed by the *user/look_up*, the activity users is the limits imposed by the *followers/list* and *friends/list*.

It was mentioned earlier that it is possible to extend the limit by increasing the time between two snapshots (give more time windows). Another way to do it is to increase the number of connections to the API by obtaining more keys. It is a mix of these two solutions used here.

The following formulas are used to calculate between each snapshot the size of users that it is possible to collect for the next one.

$$total_{call} = time * number\ of\ client$$

$$maximum_{link} = total_{call} * 20$$

$$maximum_{activity} = maximum_{link} - total_{main}$$

$$maximum_{contextual} = ((total_{call} * 100) - total_{link})$$

This limit changes when more clients are collecting data, when the time between the snapshots is increased, and when one of the set of users have less members in it. However, when the limit is hit, a sampling is still done. The program randomly selects a record from the sample database. It does it for the activity and the contextual users, separately. The users that are not sampled, are labelled as *extra* and are still input for the next snapshot. However they are only collected if the clients have finished all the others users and are waiting for the next sample to be computed by the server.

Also, for the activity users, an extra step is done but prior to that. Users have different size of lists, some have more followers/friends. That means they will have higher activity and then a bigger number of users to be included in the activity set. A risk is that one user will hit the limit alone by having a high activity. This high activity can be seen as high in comparison with the limits but not relatively to their own links. For every user, a first sampling is done based on the size of their links and here what it is achieved by the code:

```
If the number of link is lower than 1000, than 100%
of the activity links is stored in the sample pool.
If the number of link is between 1000 and 2000, than
70% of the activity links is stored in the sample
pool.
If the number of link is between 2000 and 5000, than
50% of the activity links is stored in the sample
pool.
If the number of link is between 5000 and 10000, than
20% of the activity links is stored in the sample
pool.
If the number of link is higher than 10000, than 10%
of the activity links is stored in the sample pool.
```

With that, a proportion of activity is recorded, limiting the risk of having a user taking the entire limit of the APIs.

Now that the way to collect information and the sampling rules have been explained, the ethical issues remain. This dynamic sampling method and the associated collect of information raise specific questions and problems alongside the traditional issues brings by digital traces. The point of the next section is to highlights these problems and how to deal with them without jeopardizing the research or the users' privacy.

5.6 The (im)possibility of an ethical attitude

Two fundamental principles in ethics are particularly challenged by the digital traces and the SNSs, the *informed consent* and the *respect of privacy/anonymity*. The global reason is the quantity of data, how and why they are generated. They are more public, or at least more publicly available, than before. The very nature of the SNSs makes them more open to the others and it is technically easier to collect them on huge quantity. Also the nature of real-time and, sometimes, the impossibility to get historical data make it more complex to ask for prior consent. The respect of privacy suffers from the potential discrepancy between the nature of data as public and the conception of the users of what is public and who is the public, alongside the greater difficulty to anonymise the data themselves.

5.6.1. Informed consent

This notion comes directly from the Nuremberg trials in 1946 when the allies judged the medical experiments conducted by the Nazis (Gayo-Avello, 2012). The Nuremberg code is directly created after these trials and have 10 principles to respect and among them the informed consent[67] (Mitscherlich & Mielke, 1949, p. 180).

This definition evolved over time and become eventually the declaration of Helsinki (Weindling, 2001) but the fundamental principle remains the same: asking for prior and informed consent.

A recent new conception of the informed consent defined by Friedman et al. (Association, 2001) and summarized by Gomer et al. (2005), has 6 components:

> **Disclosure** (providing adequate information), **Comprehension** (the individual having sufficient understanding of the provided information), **Voluntariness** (the ability for the individual to reasonably resist participation), **Competence** (the individual possessing the requisite mental, emotional and physical capabilities), **Agreement** (a reasonably clear opportunity to accept or decline participation) and Minimal **Distraction** (the consent process itself not being so overwhelming as to cause the individual to disengage from the process).

With digital trace, this fundamental precept is more difficult to obtain and sometimes is not obtained at all. The most extreme example known is the Facebook experience of *emotional contagion*[68]. During two weeks, in 2012, Facebook manipulated the timeline of two sets of users (N= 689003). For one set of users they reduced the number of negative posts while for the other group they reduced the number of positive posts. Later, Kramer et al. used this data to write a paper on emotional contagion and found out an effect of the manipulation (2014). Right after a massive

[67] To obtain the list of the 10 principles: http://www.hhs.gov/ohrp/archive/nurcode.html (ASH, 2016).

[68] The idea that emotional contagion is possible via text-only communication. After a user makes a status with emotion content, the friends exposed to it are more likely to choose the same words later (Silvertown & Al., 2009).

outcry hit the authors of the papers and numerous articles and newspaper articles were written about the apparent lack of ethics from the researchers (Kramer, Guillory, & Hancock, 2014).

This example is extreme because it involves an experiment on emotion without any informed consent. However, to defend themselves or the researchers, two interesting arguments were made. First, the users gave their agreement to be part of this type of research when they subscribed to the service and secondly, the Cornell University stated that, as long as the data collection was not made by the researchers, but internally to Facebook, the paper did not fail any ethical requirement (Meyer, 2014; Schroeder, 2014).

For the first point, every user has to accept the User Licence Agreement. In that agreement they accept the possibility to be part of experience of having their data and behaviour analysed. To go back on Twitter, they explicitly state the same agreement. For instance about the Log Data (but it is also the same with other information):

> “We may also receive Log Data when you click on, view or interact with links on our Services [...] **Twitter uses Log Data** to provide, understand, and improve our Services, **to make inferences**, like what topics you may be interested in, and to customize the content we show you, including ads. If not already done earlier, for example, as provided below for Widget Data, we will either delete Log Data or remove any common account identifiers, such as your username, full IP address, or email address, after a maximum of 18 months.”

https://twitter.com/privacy?lang=en
(Carberry, 2014)

It is slightly different than having the timeline manipulated[69], but the fact is that they can do it as the user agreed on it. That raises another debate on what the user really agrees to. It seems that more often, the users do not necessarily understand these agreements and the notion of *consent* is correct in a legal way but less in an ethical perspective (‘Privacy Policy’, 2016).

[69] But this manipulation is also a part of the agreement. Every timeline on Facebook is composed by algorithm and since February 2016 it is the same on Twitter (C. W. Granger, 1980).

Moreover, this agreement covers Twitter (or Facebook in the case of the Kramer et al. research) but does not protect any external researchers. As soon as the researchers collect the information themselves (as here), they are not covered at all. There are some rules that guide the use of the API Twitter publishes them for good practice but these guidelines are for developers and it is less for the researchers than the service providers that are aimed at.

If the researchers want to fulfil the requirements to be considered as ethical, they cannot rely on any facilities offered by SNSs, neither thinking of being covered by their agreements, they are not driven by the same ethics standards as research and do not offer any safeguard for it. But if the informed consent remains the standard that the researchers want to have when they are doing digital trace research, it is going to be impossible; often it is impossible to ask for prior consent, or even consent at all.

For instance, it is impossible to collect tweets from more than two weeks in the past. If a discussion occurs around a hashtag, it is impossible to know in advance who is going to participate to it and asking them their consent in advance. In case of profile information, as seen in Chapter 2, historical data does not even exist. Any time a consent is asked, the data vanished. Sometimes, the informed consent is voluntary omitted.

Suicide and sociology is not a surprising combination, it is therefore quite natural that several studies on suicide are using digital traces and SNSs. Some of them are relatively traditional in their approach (Marreiros, Gomer, Vlassopoulos, Tonin, & schraefel, 2015), but others are taking full advantages of the specificity of digital traces[70] (Sueki, 2015; Won et al., 2013). In these two latter articles, they collect data about users they suspect to have suicidal thoughts. The Colombo et al. study used newspapers to find which of the users collected have or have tried to, commit suicide. The O'Dea et al. article builds a method to build an automated system to detect and identify such users.

These two studies share another common point. They both add a significant gap between the data collection and the analysis. For the first study, the reason is methodological (they had to wait for the user to decease in order to identify him). For the second study, this gap is artificial.

[70] Traditional in theory and level of analysis: the article from Won et al. uses a traditional study of macro economic variables and celebrity suicide rates on the suicide but on population level. Traditional in method: The article from Sueki et al. uses a traditional survey.

They had to wait 3 months between the data collection and the analysis to respect the requirement formulated by their ethical committee[71]. No reasons are given for this artificial gap of three months, but what is possible to suppose is that during that gap suicidal users that expressed such thought may have actually done it and are therefore already dead during the analysis. It is possible to also think that this gap could avoid psychological distress for the researchers. Then, as soon as the period of uncertainty is over, the science could be peacefully and ethically be deployed. The study of "*the levels of concern among suicide-related tweets*" (Colombo, Burnap, Hodorog, & Scourfield, 2016), could be done without the mentioned *concern*.

Cynicism aside, this method is interesting because the time between the data collection and the analysis creates a distinction between the two. Therefore, a different regime of ethical principles can be applied. Exactly like for the Facebook research and the justification given by Cornell University. With digital traces, the issues seems to come principally from the mode of data collection, more than from the analysis of the collected data. The reason seems that for data analysis, it is easier to respect the traditional standard of anonymity and confidentiality[72]. But even if there is a will to ask for consent, not necessarily prior to the data collection, there are others issues raised, again, by the size of the data.

If million tweets and information about associated profiles are collected, as it is often the case in Twitter studies, it is almost impossible to contact each individual users. The real issue is not that there is no way to contact them, but to fight the spam problem, Twitter imposes drastic limitation on direct message. To be able to ask for prior consent, either the direct message or a mention message are possible. If the direct message is the method, the users need to follow the other users to be able to send them a message. The first limit is then on the number of new followers per day it is possible to add 1000 new users per day. It seems a lot but it is nothing compared to the number of users it is possible to collect in one hour using the Stream API. Even if 5000 followers can seem a lot, you then need to add another limit. After following 5000 users, Twitter imposes a limit based on ratio. It is not possible to add more people if that ratio followers/friends is not reached. Obviously, the ratio is not known even if some website and forum do some test to discover it. And even if the

[71] Beside that they decided to not asking for any consent under the principle of precaution and to not invade the privacy of their users.

[72] But the size of data can jeopardizes the confidentiality. Some example can be found in the social network analysis (Lash, 2002, p. 196).

number of users is reached, it is impossible to send more than 1000 direct messages per day[73].

An account does not necessarily imply a real person behind. It could be an organisation, a group of people or a robot and in this context, asking for prior consent can lead to remove some participants that will not pose any ethical issues while they will offer valuable insight for the study. For instance, an organisation will have different behavioural patterns than an individual, and it is expected that the study will show these differences.

Another difficulty is the possibility that users will not see the message in time. Some accounts are not active or very active, therefore the user associated to the account can miss the message before the collection of data. By sending several messages with a sufficient time lap between them, it is possible to collect the data (automatically), and ensure the maximum visibility of it. By waiting for a prior consent, the relevance of the information is not possible. For instance, it is impossible to collect the tweets older than a week. If the user doesn't reply in this interval, the information about the context is lost.

To summarize, it seems unlikely that research involving digital traces will be able to apply an ethical principle that was originally designed after the medical atrocities in Nazis' camps. Consents in research is often abandoned for several reasons, velocity of data, absence of historical data, volume of users/subjects to reach, nature of these users. These issues are concentred during the data collection, while the data analysis seems to be less problematic.

5.7 Ethic process

The methodology shows that the process of data selection-collection and analysis is automated. This is an advantage, in regards of ethical principles. One problem is invading the private life. The main argument used by companies such as Google is that if it is automated, it is not a real privacy invasion. This argument was used when some outcries reached the press about the automatic processing of emails stored in Gmail (O'Dea et al., 2015, p. 188). That content is among the private conversation, while the information collected on Twitter is already public.

[73] All these limits can be found here: https://support.twitter.com/articles/15364

The collection involves several analysis steps to identify which user's profile to collect. However, these analysis, as well as the data connection does not need any human intervention. The entire process is done by the scripts developed for this work. Some random manual checks were done to ensure the quality of the process, but theses checks did not need any specific user identification, neither to go back to Twitter to assess the quality of the data collected.

5.7.1.Prior knowledge and data collection

In the specific case of a protected account, Twitter still allows to have access to the profile information but not the links (Friends and Followers). In that case, the information is discarded as soon as it is automatically identified as protected. That means that it is not even stored in the database.

The data collection will take place without participants knowledge. However the final data analysis will not be done if some ask to be removed from the database.

The reason is not a need of covert observations but that the selection of the sample is based on activity of users. Therefore it is impossible to ask prior consent for observation.

The second reason is the nature of Twitter itself. An account does not necessarily imply a real person behind. It could be a organisation, a group of people or a robot. Some account are not active or the user can not connect on its Twitter account during the period of the research and missing the message.

It is important to note that this option was not offered for users in the activity network, neither the users from the friends links nor the followers links are aware that data is collected about them. The reason is that it is impossible to contact all of them. There are too many of them and Twitter will consider as spam if any attempt is done to contact all of them. However, even if the API gives full information about their profile, only a limited amount of information is collected (see earlier in this chapter).

For all these reasons, it is not possible to ask prior consent from the users, however it is possible to ask the participant if they want to opt-out.

During the data collection, I specially created twitter account for the research to contact each individual to ask if they have any objection about the analysis of these data.

User were contacted via the Direct Message system of Twitter to give a link to a website which provided more information about the study, the harvesting of the data, the process of anonymization and the contact information for any enquiry.

At the bottom of the web-page, an opt-out form can give them the opportunity to be removed from the dataset if they wish to.

I considered that if the participant did not express their wish to be removed from the dataset, that I can use the data for the analysis purpose.

This twitter account can be found here: https://twitter.com/op1e10.

5.7.2. Data storage and anonimization

On Twitter the identity is ensured by a unique identifier (represented by the *id_str*) which is the most sensible information stored (the username is not kept as it does not bring any added value to the identifier). With this identifier, it is possible to go back to the profile (or the tweet). This information is needed to communicate with the API, to access the right profile, and during the analysis. The list of followers and friends, as well as the list of mentions in the tweets stored only contains these identifier. Therefore they are essential to the analysis and during the data collection. However, prior to any public access (by anyone who requests access to the dataset), every identity will be matched in a separate database with a random number. Later, this number will be used to conduct analysis. However, some information about profile location will still be stored.

The destruction of the database containing the concordance between random number and Twitter id will be done, only at the very end of the research. It is to allow me to be able to remove people if they are asking for it, even after I started to analyse the data (a possibility is to keep this database as long as the dataset is available, if this latter option is better, then the database will be encrypted and stored on a different server than the server hosting the dataset).

Any identifiable information is published in this work or on Github[74].

The data were stored on a virtual machine on University server. The only person who had access to it was me. The table containing the link between the Twitter Id and the random number was stored on the personal computer at university. If someone obtained access to the server, they would still not be able to identify the users and accessing the personal computer would not grant access to the server. The ethic application process can be found in Appendix 9.3 Ethic Application.

[74] The code developed for the analysis and the data collection can be found here: https://github.com/Oliph/PhD-WebScience.git

6 Data Collection

6.1 The entry point

One main question remains, about the population to study. As it is pretty clear now, a selection needs to be done prior to the collect of data, as it has a limited amount of users it is possible to screen. I have not talked about the population until now because the main focus of this work is about the use of digital trace within a specific environment, Twitter, and the unveiling of the specificity of such data. Being focused on a specific group, or a question about one particular event such as elections or news spreading would have raised other questions that are inherent to the studied population and I would have needed to account for these, instead of being entirely dedicated to the methodological aspect. This being said, there is an inherent necessity to apply the developed concepts to at least one specific group to support the development of the thesis.

To understand the conceptualisation of the sampled group in this methodology, a short detour to St. Brieuc Bay can be helpful.

In an ANT based study, Callon was interested in the population of scallops and fishermen in St. Brieuc Bay, or at least that was the title of the article "Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc Bay" (Callon, 1986). The novelty of that paper is in the focus on the social interaction between the actors, the fishermen, the researchers and the science community. The main point was to demonstrate how science and nature but also the social shape all of them in their attempt to define themselves and the others. It is that symmetry between social and science/nature the author highlights, as new methodological tools to give a precise idea of this development. This interaction between the two, the nature and the society is called the translation and is defined as:

> […] the mechanism by which the social and natural worlds progressively take form. The result is a situation in which certain entities control others.
>
> (Rushe, 2013)

With this methodological posture, *Pesten maximus*' population and its mode of reproduction becomes a *device of interessement* for all actors to build identity, alliances and negotiate the construction of the Nature-society. This is the first translation of the scallops for the actors studied by Callon.

But if another perspective is taken and the article itself and Callon himself who are taken as the point of reference, the scallops become less important. The defined object, the fall of scallops population became an excuse to apply the demonstration of the sociology of translation. At the end, neither the scallops, the fisherman, nor the researchers are the main interest of the researcher. On the contrary, as stated in introduction of the article, the main goal is:

> […] outlines a new approach to the study of power, that of the sociology of translation.
>
> (Callon, 1986, p. 8)

Actors that are described in their process of translation become, as a whole, or as a blackbox a *device of interessement* for Callon, losing their ability to be interesting on their own but they are important to show the methodology needed for a sociology of translation. The fact that scallops are studied is incidental.

Here it is the position is shared; the group studied will not necessarily have an intrinsic interest. It is more a *device of interessement* to apply a specific methodological posture. It is an excuse to show the implications and possibilities of research using digital traces and the interaction between technical limits imposed by the API, the data generated by the actors and the sociological questions that is possible to raise within such environment.

Being interested in profile limits the ideal *type* of account of phenomenon, it is possible to study. First, the trending topics or the most famous Twitter's accounts are the worst place to start. They are the best shot when the focus is on the tweets themselves because it is easier to collect them and the huge generation of data helps to apply several models that need such amount of information. But using the REST API complicates the task to obtain a sufficient granularity on the data collected. Therefore one quality of the group studied should be on a smaller scale. This scale is defined by the number of calls is possible per hours and the granularity wanted.

Another type of users, or at least conversations that is often studied on Twitter are the political debates (Callon, 1986, p. 1). Here again, this is not appropriate for the present research. These types of discussion are often heated, dependent on the news and often polarized with confrontation with an opposite group. One aspect that will make the group easier to study is a more stable group or discussion, or at least not necessarily confrontational.

Third characteristic that is wanted is a relative presence on the network otherwise it is impossible to study. Not everything is possible to study, a statement that seems to be a truism but not often understood.

One type of group of people that match these constraining limits is be people with a hobby. A group of people sharing a common interest can be more stable and less polarized than political debates. They are often present on the Internet (but not necessarily on Twitter). It is also possible to limit the group studied on other characteristics than the hobby itself. It also presents the advantage of not being necessarily focused on socio-economical characteristics. Not that they cannot play a role on the population itself, but it is possible to see it as less critical.

One group that matches all these characteristics is the *amateur astronomers*. They are relatively stable (considering that becoming an amateur astronomer implies a minimum in learning and equipments), are potentially well versed in technology and are present on the web (they are well known to use several tools of crowd-sourcing for instance).

To identify a plausible group of them, there is, in United Kingdom, an association that brings these sky observers together, the *British Astronomical Association*. To paraphrase their website[75], this organisation, created in 1890, supports and promotes the amateurs observers in United-Kingdom. They have a variety of activities, including the publication of a journal, a handbook and other email listing. They

[75] https://britastro.org/

encourage mutual help among observers and even distribute prizes and medals to their most honourable members.

This association has a Twitter account under the screen name @BritAstro[76].

Already, by translating the concept of amateur astronomers to the British Astronomical Association and then to the associated account, it is not astronomers anymore that are studied but profiles related to the profile @BritAstro. In comparison to Callon's research, we can say that BritAstro is the St Brieuc Bay of this work, a location to limit the area of the research on a web based sociality.

It could have been possible to obtain a list of members of the association, contacting these members, trying to get their twitter account, the permission to follow them and being sure that the person was an amateur astronomer. It would have been interesting if the research was about them but the point is to show the advantages and limits of digital traces. In this optic, the population of interest becomes the profiles associated with BritAstro[77] with a high chance that it is composed of amateur astronomers.

The data about the twitter accounts are summarised in the following section. It helps to give a first idea of the network of interest. Rather than being a hashtag, it is a more stable network based on a profile.

The idea is not to collect the information about that account in particular, but collecting the information about its network. At the end, BritAstro is merely the interest but helps to define the network of interest as the: *network of actors that are following BritAstro.* Nothing more can be said at this point but the following section give some information about it.

6.2 Information about the account

This account, created in August 2009, has at the time of writing, 8239 followers and 80 friends as well as 1681 posts with an average of almost 5 tweets per week.

[76] https://twitter.com/BritAstro/

[77] On a side note, even if some could argue that the link between following an account associated to amateur astronomers and the fact that it is linked to actual amateur astronomers is weak and not sufficient enough, two answers can be given. First, is that a population of interest is always redefined through the methodology. Here the limit is the digital traces and therefore only what is accessible through it. Second, as seen in chapter 3, a profile is not that different than an hashtag. They are both temporary constructed network bringing in the same *place* tweets and profiles. The only main difference concerns the fact that a profile is supposed to be centred around an actor, while the hashtag is more around a topic, and that the profile can be often seen as a more stable network, even if here it is an *a priori* rather than a proved point.

The year 2015 was not a prolific year for the account with only 135 tweets compared to the 354 the year before. However, taking a look at the number of retweet per tweet on average, the numbers remains the same, an average of 30.69 for 2014 and 28.68 for 2015. See Error: Reference source not found and Figure 1: BritAstro Followers over the years.

Year	Number of Tweets per Year	Number of Retweet	Average of Retweet per Tweet
2010	175	580	2.708772
2011	285	772	2.708772
2012	263	1314	4.996198
2013	402	2738	6.810945
2014	354	24261	68.5339
2015[78]	177	7986	45.11864

Table 3: Tweet per years for the account @BritAstro

The apparent average of 1 tweet per week is misleading. If a closer look is taken for the last two months of 2015 it is possible to see that the activity is more intermittent. By bursting activity it is more one or more tweet per day separated by several days without activity. It is hard to categorise this account as a really active account.

[78] The year 2016 is not represented in the table as it is only in January at the time of writing.

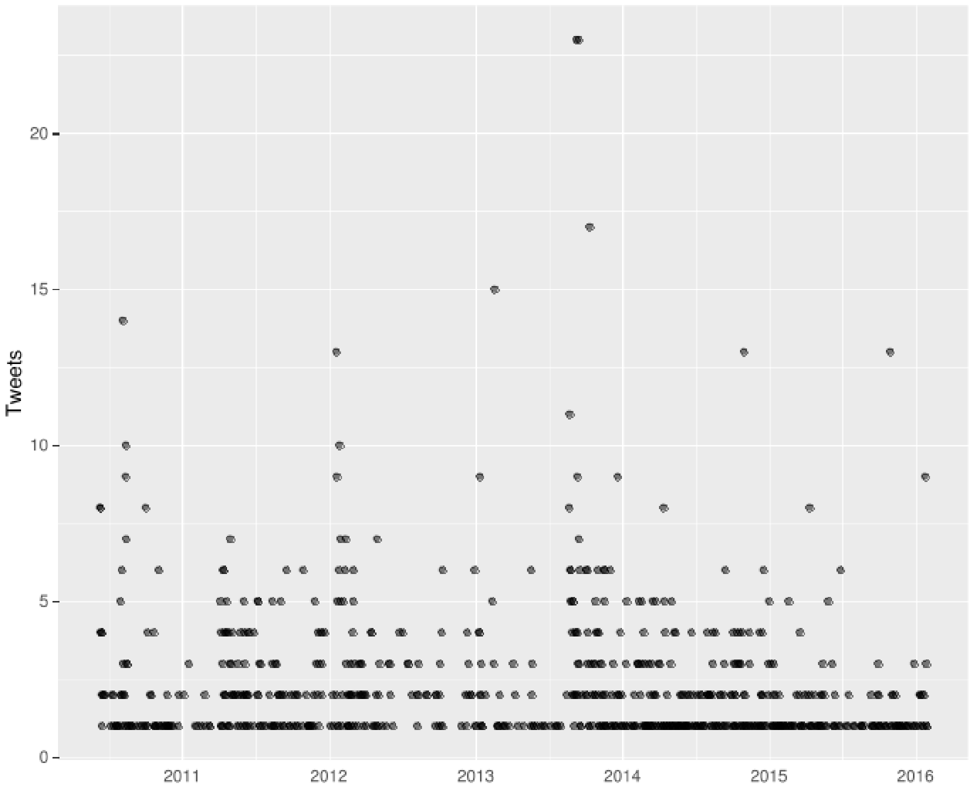


Figure 1: BritAstro Followers over the years

Now, the interest is on the list of followers and how to sample it. The following section is about them and which steps have been conducted to reduce the size to a core of interest.

6.3 Containing the object @BritAstro

It has been defined in the Chapter 5.1 Introduction that encountering the empirical object will redefine the research itself by reconfiguring the method assemblage.

The limits of the API access were one of the main translation of the definition of the SNS into a practical version, creating de facto a new heterogeneous object. These limits, not only redefine the SNS, but also redefine the object by limiting what it is possible to access in term of quantity and type of digital traces. To recreate the empty information, the time stamp in this case, of the users' activity, a more important stress is imposed on the quantity of information that is possible to collect; the REST API being more restrictive than the STREAM API. To bear with this restriction, not only this work has to study one object of interessement (@BritAstro account), but also has to limit the extent of its ramifications.

Latour stated the only limits of the description of a network is the size of the book (Poell & Van Dijck, 2015; Pond, 2016). Later, talking about digital traces, he also admitted that the power of visualisation, even if bring new horizons with digital traces, will still face practical issues, giving only a limited light on the object, despite the 1LS approach. Here, the interest is still to adapt quantitative methods on the empirical object. In ANT, the object being endless (the very reason they are defined as *actor-networ*), some restrictions need to be put in place. But, instead of applying a sampling method toward the full population, as a random sampling will do, the constraint will be based on the past digital traces of the list of profiles following @BritAstro.

This section will highlight some choices that had to be made in order to keep the volume of data and the number of call to the API in the right limits.

The subset will be the entire followers list of the account. This list of 8239 is too big to be screened in detail with the actual limitations. However, before randomly sampling that list, several cleaning are needed. These filtering steps are needed to ensure that privacy is respected, inactive accounts are not screened, and some limitation from the API are met.

The very first step done is to ensure that only unprotected users are going to be monitored. There is still some information available that is possible to collect but to respect some ethical standards, these users are going to be removed. This is done by looking if the field *protected* is set on *true* in the *JSON file* representing the *profile*.

Protected	FALSE	TRUE
	7401	838

Table 4: Number of BritAstro Followers with a protected account

These 838 followers are removed from any further analysis.

The second step is to remove inactive users. It will be useless to collect data about people that don't have any activity. This step is the first decision to take. Defining who is active and who is not is arbitrary but in this case it is also reflects the constraints imposed by the API.

The decision has been taken that users who did not tweet the last two weeks. Using this sole limit is inadequate in the sense that Twitter's users can have different patterns of publication and could have sporadic activities that spread on a time span longer than two weeks.

To avoid that side effect only the users with less than 3 tweets per week on average are removed. And to be sure to not remove too recent accounts that cannot satisfy these conditions, only users that have created their account later than the last two months are removed.

The scale of interval, per month, has been decided accordingly to the windows of data collection, two weeks initially. Therefore, this definition of *inactivity* is shaped by the method and not by the theory. This last transformation of the object of inquiry into a research artefact is rooted on the limit of this research but it could be changed in the future.

	TRUE	FALSE
Last tweet less than 2 weeks	3750	3651
At least mean of 3 tweets per weeks	5572	1829
Account older than 2 months	7401	0

Table 5: Active versus Inactive BritAstro' s Followers

As the last condition is fulfilled by all followers' account, only 1030 users are removed from the initial dataset. At the end of this first steps, the dataset comprise 6371 users.

From this filtered followers ids that were *unprotected* and *active* recently, all their followers and friends were collected. This collected a total of 8854779 links, the entire links of this dataset. The last filter to apply is the one to select only the one that belong to a community.

As said, the people who follow an account can be seen as potentially expressing an interest in that account, the same as the people who tweet with a hashtag. However, the goal here is not only to screen the people who follow BritAstro but try to study an online community too.

Several ways are possible to build this *group*. Here the solution that was kept is to subset only the people that are following the BritAstro's followers. Instead of using a fixed number, a threshold of at least 10 % of links being BritAstro's followers is necessarily to be included in the final subset[79]. The limit of 10% is arbitrary. After

[79] A big problem on social network is the proliferation of bots. Not that they are not interesting themselves but more because they generate a lot of noise, especially on small dataset like here. A common way to remove them is to check after a period of time which account has been suspended and remove them under the assumption they were bots (Law, 1999). However, such method is not

testing several thresholds it is the one that seems to be the best balance between detecting active people tweeting about astronomy and leaving out other followers that will not interact with the other followers.

On another side note, it could have been 20% or 100% that would have not necessarily change the outcome of the research. More data would have been collected but some more drastic subsetting rules would have to be applied in order to deal with the API limits.

The final subset obtained contained 737 unique users that are going to be used as the subset of interest.

This subset can be therefore be defined as:

> Twitter profiles that are following BritAstro, and are following each other for at least 10% of their total following, creating, to some extent, an online group of same interest.

The definition clearly outlines the limits of the subset and is not making any other assumption of its existence than the rules that lead to its creation. This online group is the starting point of the data collection that span over 60 days. This limit of 60 days is again, a technical constrain, this time not within Twitter, but on the side of the researcher, me. I could only get 500 Gigabytes of space in a virtual machine to collect data. The amount of information stored during the data collection, even if it is on a small subset, rapidly growth as all profile lists are stored. As a result, 60 days was the usual time period it was possible to monitor before attaining this limit.

The result is a dataset containing the information about 30405721 users, 72683786 tweets, 70993 lists of followers (containing up to 10 000 user ids), 73555 lists of friends. In terms of activity, the data collection recorded 59530042 changes in profiles between two snapshots and the details on 33999 changes in the followers and friends lists.

The next chapter is going to describe the analysis on this dataset. It will also refine the different theories and hypothesis by following the constraints of aggregations and

needed here, as the main sample is filtered *a priori*. The last filter, selecting only the users who share 10 % of links with the others, should be enough to avoid them. As a measure of precaution, a manual check of the entire sample has been done and no bots were detected (which does not mean there were no bot in the subset).

statistical analysis. The reasons and consequences of these decisions are outlined before presenting the results.

7 Analysis: Influence of context over the activity

7.1 Introduction

The idea from the beginning is to see if the activity of the context has an impact on any user who evolves within that environment. The activity and the context, as seen in chapter 1, are the new *socio-demographic categories.* In a fast changing world, sociology has to adapt and the predetermined categories are not relevant anymore. This is the reason why the vocabulary of *network* supplants the vocabulary of social classes, and may be replaced by activity and context. This change is accompanied by an increase of data, bringing new methods and new paradigm. However, as seen in the Chapter 2, this new paradigm will not necessarily answer the problematic and may often seen as a post-positivism. One way to avoid this pitfall is the methodological and ontological approach developed by ANT (Chapter 2.3.4). With their tools in hand, it was possible to see the deployment of digital trace within method assemblage in the ClubCard first, and then with Twitter (Chapter 3.3). The deployment of the method assemblage was then described further with the definition of Social Network Sites and ultimately its practical translation through the Twitter *API representations* (CHAPTER 3.4.2). As stated by Marres et al.:

> The context of implementation is at least as strong as determinant of 'method' as the implemented measures.
>
> (Latour, 2005)

As any research, the quasi-object composed by the access to the data, the definition of the *device of interessment* (Chapter 6.1) has to be plugged into the network of the research itself. Here more limit was imposed on the data collection due to the choice of using the REST API and some technical limitations on data storage. Rather than seeing at it only in term of limits, these descriptions helped to redefine the heterogeneous object created through the process of the research itself, and not being convinced we are studying *real behaviour*.

The last step remains, the analysis itself. But before doing it, the analysis itself requires specific *data transformation.* This last step is crucial and has consequences on the potential conclusions that we can draw. A good understanding of all the limits and potentials of the collected digital traces helps to avoid massive pitfalls in the interpretations (Marres & Gerlitz, 2016, p. 30).

This step is intrinsically linked to the type of analysis that is going to be used. Therefore the decision on the transformation needs to be framed within the specificity of the analysis rather than the full potential of the dataset. Data transformation is about losing some information, or at the very least, framing the perspective into a more narrowed angle.

But knowing that the data needed to be transformed and re-framed before analysis, also means that the dataset has the potential to be adapted for several approaches and perspectives. This flexibility offers a way to explore different types of aggregation and different angles of analysis on the same population and on the same processes. This is something new offered by the digital traces. On more traditional quantitative methods, or experiences, the collected data already fits the analysis, the collection of data, through questionnaires or devices to record the experience is carefully designed with the analysis in mind.

With digital traces, the data collection has to face the technical and API limitations first. Some could argue that surveys face limitation too. While it stands true, the difference with digital trace is that as soon as the API limitations are overcome, the quantity and variety of data is bigger than what is traditionally obtained with surveys. As consequence, the by-product is often bigger than what is collected through simple format of questionnaires.

Obviously the data collection is itself engraved into the process of the analysis, but the data collection can sometimes be easier to process, and numerous information is collected that was not necessarily needed for the principle analysis in mind at that time. This gives a more versatile dataset that can be explored, by adapting the data transformation to different statistical tools or algorithms, often through different types of aggregation or use of different metrics[80].

What remains is the question of the transformation itself. The word *aggregation* has been mentioned without context. In fact this word is what brings together a sociological perspective, defended by ANT about how to study social phenomena,

[80] If a mixed approach of qualitative and quantitative methods is taken, there are even more possibility. But for this work, only quantitative methods are used.

and the practical approach of dealing with digital trace. The last tension between numbers and sociology, adopted in this work (before the next one).

7.2 Levels of aggregation

As developed in the Chapter 3 on the digital traces, Latour explicitly condemns the distinction of micro and macro perspective. It takes the advantage of the information provided by the web on individual to deploy his conception of the 1.5LS perspective; with the possibility to aggregate data about individual and navigate through the different profiles, it is possible to develop a picture of the social in formation that it is a the same time local and global. His conception of the use of digital trace relies heavily on visualising tools; ANT advocating for a descriptive approach.

Even if the idea of Latour and the ANT is appealing, and looks promising for the use of digital traces, this work takes another perspective. For Latour, the aggregation is associated to the 2LS method, while it's 1.5L deploys a *monadological principle* and for him it is superior to retrace the social in formation.

The idea of aggregation is clearly understood as a limited comprehension and grasp on the complication of social phenomena. The aggregation is obtained when enough individuals are collected and put together under the same umbrella such as structure or society.

But digital traces and their associated dataset have the advantage to contain a more complete picture. The cost of entry (API restrictions, development of tools to interact with the API, storage,…), as soon as being paid, is easier and it is possible to collect on a bigger extent. That means that the collected data already overcomes the traditional *micro-macro* distinction. A conclusion that Latour et al. already made. However, they use this advantage to fully explore the connection and the travel into the data that visualisation tools can offer. But if the goal is to analyse the data in another way than visualisation, the methodology faces again the harsh reality of the distinction, but this time with the advantage of the digital dataset. As said, the dataset itself does not suffer from a distinction between micro-macro scale, more information is available without being forced to have the limit of a *2-LS theory.*

It is possible to conceptualise different levels of aggregation and use these different levels with different types of analysis. In consequence, the exploration is displaced from the visualisation, in descriptive methodology, to the exploration of different levels of aggregations, following different rules.

Probably that pure ANT will condemns this approach. However, as mentioned by Latour, every instrument will develop a highly focused but limited view of the whole (Diesner, 2015) and the visualisation on all actors/profile may not be appropriate. And developing a method of assemblage will inevitably pose some limits; the extend of description limited by the length of the book (Latour et al., 2012), the number of profiles it is possible to follow with the *monadological principle* (Latour, 2005), or as explained in the chapter 6.1, the limit imposed by the API itself.

Therefore, the approach taken here will be an hybrid of the ANT perspective. First, there is no presupposition of any structure. The structure is construed by the accessibility of the traces. In the present studies, it is the profile snapshot. This profile give access to different lists, the friend lists, the followers lists and the tweet lists. These will compose the different contexts. But rather than only having an aggregation on these lists, the aggregation will follow the practices deployed by Twitter's users that create these lists.

I think that it is an interesting interdisciplinary perspective that have the merit to bring together a better understanding of the limit of the statistical tools used, and still using robust and trusted methods. Not only it is possible to take advantage of the methods, but it is also allows us to better use some aspect of the dataset that is impossible to use entirely. A median position that cut us off from an ANT perspective but tries to find a middle ground that should characterise a webscience approach.

To summarise, one dataset is composed by digital traces, collected through one methodology, to understand one social interaction about one population. This dataset presents enough flexibility to build different angle of analysis and different rules of aggregation to study the same population - social interaction - digital traces with different tools.

Three different aggregations are used here, network, individual, activity. But the decomposition is methodological rather than ontological. The theory is developed on the integration of all findings and not by connecting a *higher level to a lower level*.

For the first level, the *network perspective,* the aggregation is done in two steps. First, the users are aggregated into groups. The *main users* is the set of users/individuals, as described in the methodology section about main users, the *followers users* includes any users who follow at least one of the main users and are not a member of the main users, and the *friends groups* which contains the users that main users are following. Each group is then aggregated based on the snapshot created during the data collection. This creates time series to analyse with the Granger Causality.

The second analysis is using the same groups as for the network perspective, but this time it is not aggregated with the snapshot, but on individuals, hence the *individual perspective*.

The last analysis, the *activity perspective,* does not rely on the previous grouping but use the entire dataset but this time it is the activity itself that is the central point of analysis, by allowing a proxy on how much information Twitter, and the digital traces, can give us when a new link is formed or destroyed between the main users and the context link.

The following sections will detail the analysis on these three levels, respectively, the *network*, the *individual* and the *activity* level.

7.3 Network scope

The data have been collected over time, the only way to create a measure of activity. The obtained data points, can be transformed into time series. To build the time series, every data points needs to be aggregated, by snapshot, into one measure. It is possible to do it separately for every network, and by type of activity (tweets and links). The result is a time series for each network and for each of their activities. The downside of this method is the loss of individual information. The mean (or any measure of centrality) looses some variability and information about the range of the data. Nevertheless, this aggregation gives the possibility to find better interactions between activity and context such as finding the direction of a link between two time series.

7.3.1.Data creation

To aggregate, it is important to take into account the dynamic nature of the dataset. The *main network* has associated *context networks* composed for one by the main network's *followers* and for the second one by the main network's *friends.* These networks evolved over time and the composition of them was changing between the snapshots. The members of one network are not always the same, some are included or removed. Consequently, the set of links are in function of the links the main network possesses at the time of the snapshot. Added to that, as shown in the figure, the number of recorded profiles is different than the number in the users list. That can be due to the fact that the user itself has removed their account, or the was not included in the sampling methods. Aggregating the numbers of followers and friends for all users in the network will give erroneous results. If one user is removed, their number will be removed from the total and the change will represent this disappearance rather than an activity change. We are not interested in the total of links, but the change in the context, its activity. The Figure 2: Number of profiles per snapshot and per type of network[81] shows the global picture of the data collected and the size of the sample that was supposed to be collected. The difference are due to the limit of the collection and one hiccup in one day (a short lost of the internet connection). As seen the number of followers grows over time while the number of friends remains stable over time.

The *main network* can also evolve over time. Some users may change their protected settings, going from *not protected* to *protected.* They also can have their account suspended or they can leave the service.

Normally, the aggregation of activity should not suffer from the change over time. The measure itself is the difference between two snapshots and is not the absolute count of followers, friends, or statuses.

However, the size of the friends context is much bigger than the size of the main users, making comparison less easy. A solution is to calculate the influence by using the average of absolute change[82] per group and per time. As the groups have different

[81] The scale of the y-axis are free so each graphs need to be read independently.

[82] Rather than the signed value, adding or removing is considered as an activity it is not the net change in the network that is interesting here.

sizes, to make the comparison easier, a pre-processing is needed. The score is divided by the size of the group before the mean per snapshot is calculated[83].

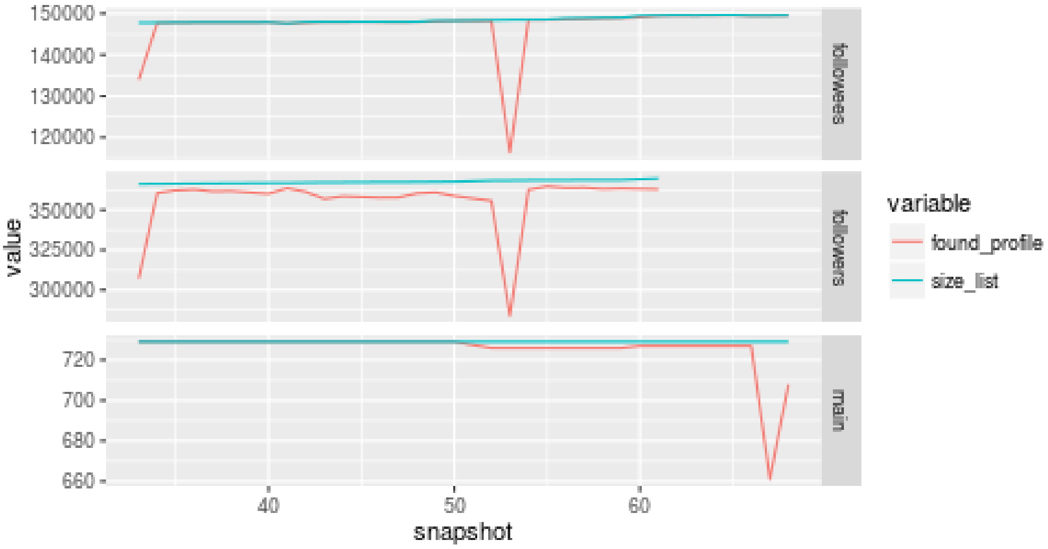


Figure 2: Number of profiles per snapshot and per type of network

Another important aspect for the aggregation is the overlapping of the network. The sample method uses a measure of overlapping between the networks to indicate a measure of cohesion between the BritAstro followers (see chapter 6.3 Containing the object @BritAstro). This overlapping is problematic when we need to aggregate different networks. These aggregations are no longer independent because the same users will be found in different time series. Any subsequent analysis will therefore be biased by them and it will be impossible to draw any conclusion. However, theoretically, the idea is to measure the influence of one network over another one. But not the influence between friends and followers. In that sense it is not mandatory to check for independency between the followers and friends networks. The pre-processing only needs to aggregate a measure per snapshot based on unique users, also remove any presence of the main network in the followers and friends network. That is the principal rule to respect the independency of all networks.

To summarize the steps explained here. For every snapshot, the set of followers (the exact same method applies for the friends) of each main user is extracted and this set is specific at the time of the snapshot. Then, all the set from all main users for that snapshot are added together. As soon as all the set are added, the present main users are removed, as well as all duplicated entries (due to overlapping in the different

[83] The followers activity lines are shorter because I stopped collecting any new data toward the end. The reason is that I reach the maximum of the capacity of the server. This network is not essential for the latter analysis, except for the activity scope analysis. Therefore the impact is not important.

networks). The result, for each snapshot, is a set of all unique followers for the main users, minus any user presents in the main users.

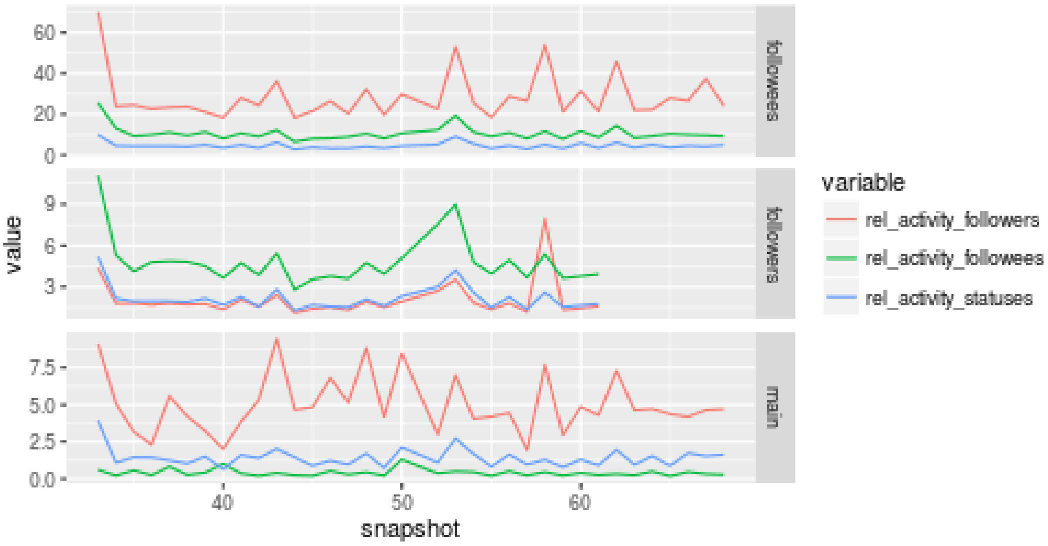


Figure 3: Activity, relative to the size, per snapshot and per network

From this set, every measure of activity, the link activity (adding-removing links) and tweet activity (adding removing tweets) is averaged, giving a single measure for each snapshot. The operation is repeated and the gives for every network (main, followers, friends) 3 time series (see Figure 3: Activity, relative to the size, per snapshot and per network).

A last point is to take into account is the outliers. In Big data, or any methods that uses a *whole population* dataset is not confronted to this issues, the size of the data often smooth the extremes. But the sampling method here is more traditional on the aspect of the size. The impact of the outliers could be more important. However, after analysis with and without the outliers no differences were found in the network scope's analysis. The same verifications are needed for the individual levels and activity level.

Now that the dataset are cleaned and pre-processed, it is possible to apply analysis on them.

The idea from the beginning is to see if the activity of an environment has an impact on someone else. As stated in the previous section, the activity is defined with the change in friend network and the change in the statuses. However, the influence may not be direct and it can have a lag between the action and the influence.

7.3.2. Analysis: Granger causality

Time series are stochastic processes (random variables) and it is a possible to apply regression and see the link between them. However, applying static regression is problematic. In that case, the effect of one variable on another one has to be immediate. For instance, a static regression will suppose that the change in activity from the friends' network or any links will have a direct and immediate effect on the main network.

Also, it is mandatory to use an auto-regressive model with time series analysis to avoid spurious correlations. What happens is that a link found between two time series could be due to the auto-correlation of one or both time series with themselves. In that case it is not a correlation between the two time series but the evolution of the time series themselves, called *unit-root*.

Not only it is possible to avoid this type of spurious correlation, but it is also possible to perform causal analysis by taking advantage of the natural order of time. The order implies that an event can impact another event in the future but cannot influence an event in the past. Several models exist to test causality with time series. For this work, Granger Causality model is used. This model has been developed by Wiener (Latour et al., 2012) and later developed by Granger (Wiener, 1956). This test is widely used across disciplines, less in sociology and on social media but has been used on Twitter and social protests (C. W. Granger, 1980; Clive WJ J Granger, 1969).

In that article the authors studied the potential link between social media participation and onsite protests. They collected posts on Twitter and Facebook about three on-going protests at that time, the *Indignados*, the *Occupy* and V*inegar* protests. They found that communication on Twitter and Facebook forecasted the onsite protests (with a bi-directional link in the *Occupy* movement). That means that on some point the online communication caused the onsite protests.

The notion of causality is different than we may think. In the case of the Granger causality, a specific definition is used. To understand this limitation, we need to understand what the Granger causality is in the first place. The best summary is from Guo et al.:

> The basic idea of Granger causality can be traced back to Wiener[84] who conceived the notion that, if the prediction of one time series is improved by incorporating the knowledge of a second time series, then the latter is said to have a causal influence on the first. Granger[85] later formalized Wiener's idea in the context of linear regression models. Specifically, two auto-regressive models are fitted to the first time series– with and without including the second time series – and the improvement of the prediction is measured by the ratio of the variance of the error terms. A ratio larger than one signifies an improvement, hence a causal connection. At worst, the ratio is 1 and signifies causal independence from the second time series to the first.
>
> (Bastos, Mercea, & Charpentier, 2015)

The first thing to note in that definition is the sense of causality. This is not a definition in the usual counter factual definition (if A did not happen, B cannot happen either), nor in a deterministic one (if A happens, B always happens). Rather, it has to be seen in term of knowledge, and the knowledge of the past to be more precise. The causality here is that A is considered to *granger-cause* B if the knowledge of the past of A improves the prediction of B better than if only the knowledge of the past of B is used. This is a limited definition of causality but this allows one to have the direction of the relationship between A and B (A granger causes B but not the inverse, or B causes A but not the inverse, or the relation is bi-directional). It is also important to note that the granger causality supposes that no other variable causes the relation. However, in social sciences and moreover with sociology, it is impossible to not violate this assumption. The *third variable* influence is common when the study is on complex phenomena (and why ANT claims it is impossible to study them like that). This issue is a constant limit of a method of investigation focused on digital traces solely, but this limitation is overcome by the benefit of applying a specific methodology on that type of data. Also, as noted by Bastos et al., even in the presence of unknown causes, the Granger causality is useful as it still can give the influence of a variable A on B, even if an unstudied mediating variable plays a role (S. Guo, Ladroue, & Feng, 2010). Added to that, even if the Granger causality is a data-driven method, it is possible to build theoretical

[84] (Kramer, 2012)

[85] Investigating Causal Relations by Econometric Models and Cross-spectral Methods ('JSON', 2014) and Testing for causality: a personal viewpoint ('Tweets | Twitter Developers', 2013)

hypothesis to lead the search of granger-causality. And in that sense, the theoretical foundation of the hypothesis limits the problem of unknown variables.

From a statistical formulation, to perform a granger causality, some other prerequisites have to been met. The most important one is the stationary nature of the data series (Bastos et al., 2015). A time series will be considered as stationary if it has the following statistical properties:

1. The mean μ (mean) is constant
2. The variance σ^2 is constant
3. The autocovariance function between X_{t1} and X_{t2} only depends on the interval t_1 and t_2

If one or two of the time series are not stationary, it is impossible to test for Granger causality or it will lead to spurious regression (Hyndman & Athanasopoulos, 2014). The reason behind the possible spurious regression is that non stationary time series have tendency to *wander*[86]. Then this random walk may lead to regression between time series when they are completely independent.

Checking for the level of integration of the time series is primordial when we want to apply the direct Granger method. For this purpose, several tests are available to check that the time series is stationary. In this work, the Augmented Dickey–Fuller test and the KPSS test are both used, as a cross check is preferable in this situation (Clive W J Granger & Newbold, 1974; He & Maekawa, 2001).

If the time series is found to have a unit root, not being stationary, it is still possible to differentiate it by removing the past value and apply a direct granger causality on the differentiated time series.

Formally, a granger causality test, check if the following model gives better prediction than a random one. Considering two stationary time series, the auto-regressive uni-variate model for y and x as following:

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_m y_{t-m} + e_t$$

Formula 1: Time series of Y

[86] For a nice, funny and lenient explanation of this issue, see (Murray, 1994).

$$x_t = a_0 + a_1 x_{t-1} + a_2 x_{t-2} + ... + a_m x_{t-m} + e_t$$

Formula 2: Time series of X

This model is then augmented with the value from x giving the following estimated model:

$$y_t = a_0 + a_1 x_{t-1} + a_2 x_{t-2} + ... + a_m x_{t-m} + b_p x_{t-p} ... + b_q x_{t-q} + e_t$$

Formula 3: Granger model

Testing the null hypothesis is that the time series x does not granger-cause the time series y:

$$H_0 : b_1 = b_2 = ... b_q = 0 \; with \; q = nbr \; of \; lags$$

Formula 4: Null hypothesis

If the null hypothesis is rejected, it is possible to consider that x granger cause y with q=lags as restriction. Also, by inverting x and y, it is possible to test for a bidirectional causality.

Now that the model to test is specified, we need to clarify which interactions are going to be studied.

7.3.3. Formulation of hypothesis

Several hypothesis can be drawn from the dataset but we will stick to the ones that are aligned with the definition of activity explained above. Two types of activity are possible for the user: posting/removing tweets and changing who the user follows.

On the context side, two types of context are possible, the friends and the followers. The former is more important as it is from these links that the main user can see the posts. The influence is the activity of the context network.

To reformulate in others words, all users that have been sampled from @BritAstro are aggregated to define one network, *the main network.* From this network, their friends are collected and aggregated together under this rule, to define a second network, the *context network*. This network is considered, as the structure of connection on Twitter suggests, to be the immediate context of the individuals from the *main network*. The individuals from the *main network* see their tweets and their interactions and choose to follow or unfollow them.

In the hypothesis that a more active and changing world brings more activity and changes, this is a chance to see if the global activity of a context has an impact on another context. Not to check on any representative of the change, but trying to see the deployment of such impact on the social network. Also this is a chance to takes a closer looks on not a smaller dataset. Rather than being a limitation, it is a opportunity to see if some findings from big dataset containing the *whole population,* still stands true on a smaller scale.

The hypotheses reflect the theoretical definition of activity as well as the methodological constraints, and can be formalised as follow:

1. H1: The increase or decrease of the *tweet activity*, defined as posting and removing tweets, in the *Context network,* will cause an increase or decrease in the *tweet activity*, defined as posting and removing tweets, in the *Main network.*

2. H2: The increase or decrease of *link activity*, defined as adding and removing friends in the *Context network,* will cause an increase or decrease in the *tweet activity*, defined as posting and removing tweets, in the *Main network.*

3. H3: The increase or decrease of *tweet activity*, defined as adding and removing tweets *in the Context network,* will cause an increase or decrease in the *link activity,* defined as posting and removing friends, *in the Main network.*

4. H4: The increase or decrease of *link activity*, defined as adding and removing friends in the *Context network,* will cause an increase or decrease in the *tweet activity,* defined as posting and removing tweets, *in the Main network.*

To test these 4 hypothesis, 4 different times series are then needed. The tweet activity and link activity for the Main network and the context network (see Figure 3: Activity, relative to the size, per snapshot and per network).

7.3.4. Checking that time series are stationary

As explained earlier, I first need to be sure that the time series are stationary. The 4 times series are double tested here, using the KPSS test and the ADF test, implemented in R. For the statuses for the main network, the time series has a unit root, according to the ADF and KPSS test. This is the same for all time series. The full results are available in Table 6: Stationary – Results)

		Augmented Dickey–Fuller test *p < 0.01		KPSS Test * p > 0.01	
		Level form	P value	Level form	P-value
Main Network	Statuses	-4.3629*	0.01	0.15727*	0.1
	friends	-4.3146*	0.01	0.25872*	0.1
friend Network	Statuses	-3.9255*	0.0239	0.11512*	0.1
	friends	-4.7659*	0.01	0.19072*	0.1

Table 6: Stationary – Results

Every time series is considered as stationary with no unit root. That means it is possible to apply a direct granger causality test. The Augmented Dickey-Fuller test, rejects the null hypothesis that the time series is auto-correlated. While the KPSS test keeps the null hypothesis that the time series is stationary.

It is also possible to apply a Vector Auto Regressive model on the time series to find the best lag to model in the granger causality. That means that the test check if the value of the time series y is granger caused by the value from the time series x at the time t – number of lags. Here, that means that the n lags is the number of past snapshots that will influence the value of y. If n is 3, that means that the influence of the context over the main users will be due to the activity of the context from one snapshot before, two snapshots before and three snapshots before. However I decided to test for the granger causality with a VAR model of order 1. The time period of a snapshot is almost a day and Twitter is a relative fast social network, I pose the hypothesis that more than one day is great a lapse between the context activity and its potential influence and one day being the maximum. Also this decision makes the interpretation of the results easier and avoids confusion of from when the influence comes.

Results

The first of the granger causality tests was performed to determine if the activity in statuses from the context network provides significant information for forecasting the same type of activity in the Main network. This is the first hypothesis. The hypothesis H1 only supposes a unidirectional granger causality from the friend network to the Main network. However the inverse relation has also been tested to see if the influence can be bi-directional or in the inverse direction. The same type of tests have been carried out for all the potential relations between the two networks as exposed in the hypothesis section. The results are shown in the Table 7: Granger-causality relationships. This table provides a breakdown of the Granger-causality tests and displays results with *p* value and the *F*-statistic.

	From	To	*F*-statistic	*p* value
H_1	Context Tweet	Main Tweet	1.6795	0.2049
	Main Tweet	Context Tweet	0.352	0.5575
H_2	Context Link	Main Tweet	0.2275	0.6368
	Main Tweet	Context Link	2.0139	0.1662
H_3	Context Tweet	Main Link	5.2233	0.02952*
	Main Link	Context Tweet	1.3349	0.2571
H_4	Context Link	Main Link	0.1265	0.7246
	Main Link	Context Link	1.8545	0.1834

Table 7: Granger-causality relationships

One relation is significant, the relation between the activity in tweet in the context and the activity is the links in the main user network ($p < 0.05$). This relation is unidirectional as the reverse granger causality is not statistically significant ($p > 0.2$). Therefore the hypothesis H3 was confirmed. The increase of *tweet activity* granger causes the increase of the *tweet activity* in the Main network and the inverse is not true.

The other results are not statistically significant and I have to reject the 3 others hypothesis. This also means that the double checking of hypothesis that the main network can granger cause activities in the context network are not confirmed either. This is still a good finding as these relations will not be explainable under the theory of the influence of the context developed here.

7.4 Regression with individuals

The previous analysis puts an accent on the network scope by using Granger causality. This method came with a cost, the necessity to reduce all the users from a network to one measure per snapshot. In consequence, all the information about individual variability is lost in the data transformation process.

But the information about individual users exists and this is why, still under the goal of measuring the influence of context over activity, now the accent is on the the individual and not the evolution over time.

Even if the data collected is time stamped, this is only a side product of some methodological necessity (an activity can only be a change, therefore a before and an after is needed). The goal is to see if the activity of the users is influenced by their context. The definition of activity and context is similar to the definition above, but the data creation changes on several key points.

7.4.1. Data creation

The measures of activities are created by taking the average of the activity from all the snapshots, for each individual. These measures are straight forward:

$$Link\ Activity = \frac{\sum Link\ activity_i}{Nbr\ of\ snapshots}$$

$$Tweet\ Activity = \frac{\sum Tweet\ activity_i}{Nbr\ of\ snapshots}$$

To create the measure of context, it is also important to take into account the dynamic nature of the links as during the network scope. To deal with this characteristic, a script was developed to parse the database. For each main user, to collect the links at the time of the snapshot. Then, when this list is retrieved, to collect the activity of the user at that time and sum up with the other from the list. The result is a value of activity of the context at that time of the snapshot. This sum of activity is then averaged by the number of links collected for that snapshot. This transformation is needed to control the effect of the size of the list. At the end, the same transformation of the measure as for the main users activity is applied, the average on the number of snapshots[87]. In this way, the indicator of activity represents the context over time, even if the time changes.

7.4.2. Formulation of hypothesis

I pose the hypothesis that both types of activity will be influenced by the link activity and/or by the tweet activity from the friends network. The time is not a variable anymore, a simple regression analysis is done to assert the causal effect of each measure independently. As the time variable is not in play anymore, the formulation of the hypothesis is slightly different. This is the overall level of the user activity that is tested, rather than the evolution over time.

Understood in line of the theories developed in the Chapter 1, I expect that if people evolves and want to stay connected in a fluid world, they also have to express a high mobility. To test this hypothesis, it is possible to apply correlation between the users' activity and their context's activity. However, as developed in the chapter 1, the idea of more flexibility of the context should increase the flexibility of the people within that environment. The reason is not mechanical but more a consequence of this logic. A flexible and active person will seek an active network otherwise this person will not be exposed to new tweets and new interactions. If the activity of the context is low, like no one is tweeting, or no new interactions are made, an active user will probably leave that context to seek a more pro-active one, therefore, their own activity could be predicted by the activity of their own context.

While a more static and less active user, by remaining static will finish to be disconnected and left from the active network and will finish by remaining in a static network.

To test this hypothesis, the tweet and link activity are used as proxy, they are the only conscious activity. Therefore, three simple linear regression models will be tested to see if these proxy support the hypothesis.

1. *Model 1:* The main users' tweet activity will be predicted by their context's tweet activity: $User_{tweet} = \mathrm{B}_0 + Context_{tweet} * \mathrm{X} + \epsilon$
2. *Model 2*: The main users' link activity will be predicted by their context's tweet activity: $User_{links} = \mathrm{B}_0 + Context_{links} * \mathrm{X} + \epsilon$

[87] The same analysis have been done without taking the average per snapshot and the results were identical.

3. *Model 3*: The main users' tweet activity will be predicted by their context's link activity: $User_{tweets} = \mathrm{B}_0 + Context_{links} * \mathrm{X} + \epsilon$

4. *Model 4*: The main users' link activity will be predicted by their context's tweet activity: $User_{link} = B_0 + Context_{tweet} * X + \varepsilon$

One issue that can be highlighted at this level is the nature of the proxies created. As seen, the measure is the absolute change which includes adding and removing users. There is no distinction between the two types of behaviour. However, this problem would be important if the metric used was about the followers. If a user chooses to remove friends, that is a choice and that still shows a conscious active behaviour. In case of a follower, it is a consequence of their own actions. But this consequence is not controlled and cannot be seen as conscious activity. Therefore, this issue is not that important as long as it is not used to measure the followers' activity.

7.4.3. Result

But it is possible to focus on the activity itself. The network studied, even if it remains relatively stable over time, still has tweet posting and more interestingly, adding and removing users quite frequently. That is the object of the last analysis, the activity of removing or adding someone but not with the perspective to explain the action itself, but to see the information contained in the dataset about the digital trace of the activity.

The first model (Model 1: Statuses vs Statuses) is significant ($p < 0.5$), therefore the rate of tweet activity in the context network can predict the rate of user's tweet activity. However, the R^2 is extremely low: 0.006334 (see Figure 4: Regression model 1 - Statuses versus Statuses).The residual falls along the line and therefore seems normally distributed (see Figure 8: QQ Plot of residuals - Statuses versus Statuses in the Appendix 9.2).

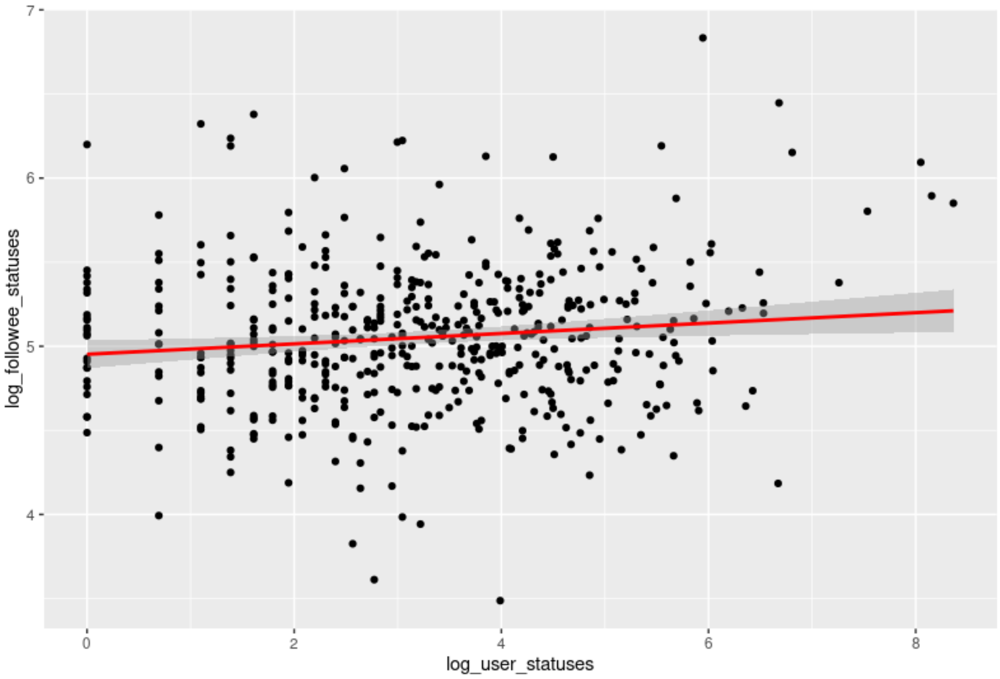


Figure 4: Regression model 1 - Statuses versus Statuses

The second model (Model 2: Links vs Links) is also significant ($p < 0.001$) with a higher R^2: 0.1688 (see Figure 5: Regression model 2 - Links versus links).The residual falls along the line and therefore seems normally distributed (see Figure 9: QQ Plot of residuals - Links versus Links in the Appendix 9.2).

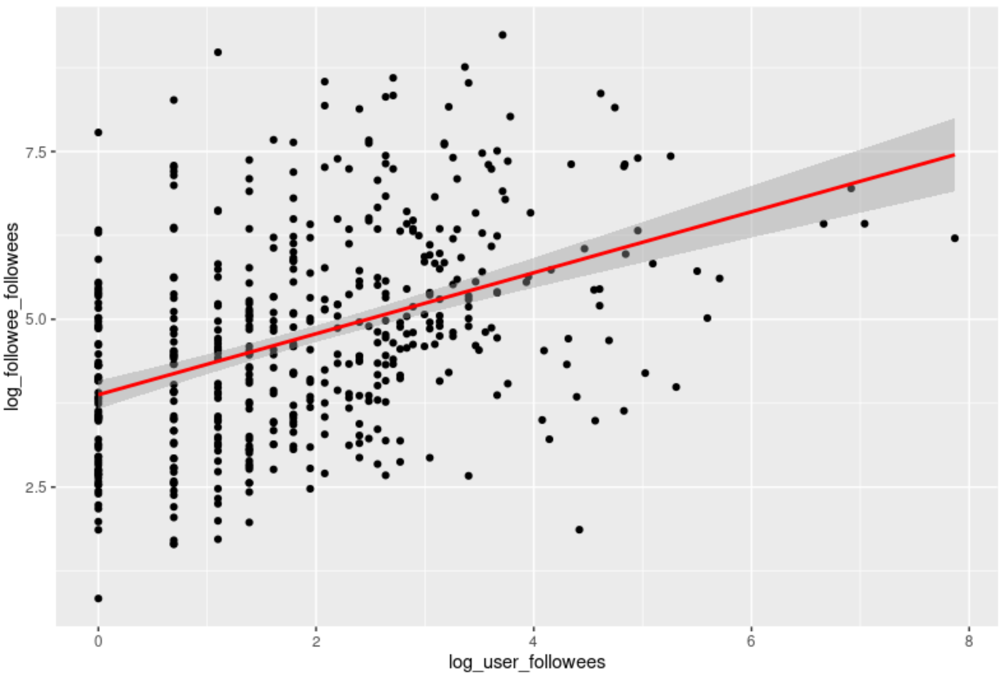


Figure 5: Regression model 2 - Links versus links

The third model (Model 3: Links vs Statuses) is also significant ($p < 0.05$) but does have a low R^2 : 0.006334 (see Figure 6: Regression model 3 - Statuses versus Links).The residual falls along the line and therefore seems normally distributed (see Figure 10: QQ Plot of residuals - Statuses versus Links in the Appendix 9.2).

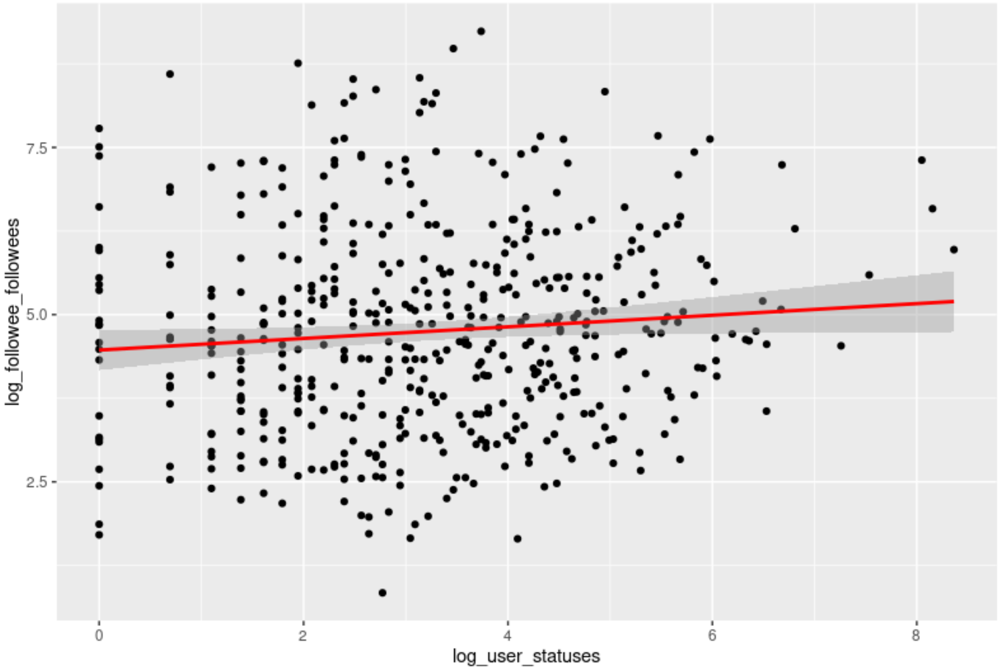


Figure 6: Regression model 3 - Statuses versus Links

The fourth model (Model 4: Links vs Statuses) is also significant ($p < 0.001$) but does have a low R^2 : 0.06502 (see Figure 7: Regression model 4 - Links versus Statuses).The residual falls along the line and therefore seems normally distributed (see Figure 11: QQ Plot of residuals - Links versus Statuses in the Appendix 9.2).

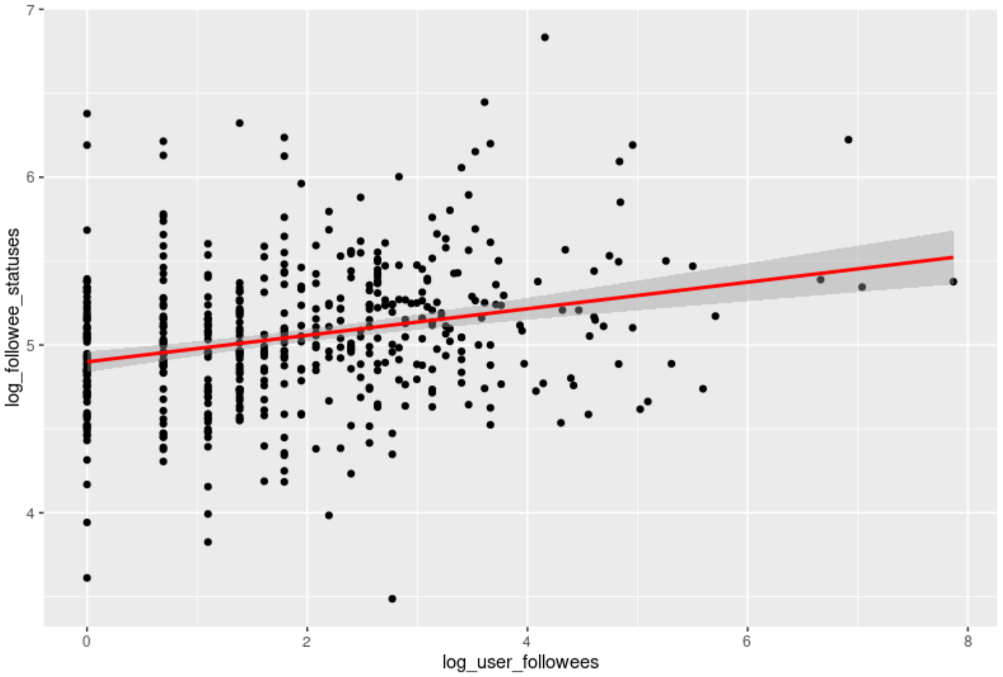


Figure 7: Regression model 4 - Links versus Statuses

7.5 Activity, attention and limit on the method

This last analysis applies another level of aggregation. Earlier, the aggregation was on the variable of time, forcing the analysis to study the activity on a network scale. Then when the time variable has been removed, it was possible to study the individuals themselves rather than their aggregation. Now, the level of aggregation is done on the type of activity. However, even if the individual and their associated links are not used anymore in the aggregation, they are still useful. They are going to be used to generate the context of each of these activities.

I has been mentioned several times that the tweet and the link activity are the only conscious activity on Twitter that leave traces. As the action is conscious, it needs the attention of the user (*a contrario* to the followers metrics). But this attention needs a context to emerge from.

This context, as defined earlier, is created within the tweet and the profile. They both contain an embedded network of links to other tweets or other profiles (as seen in Chapter 3). Therefore, the information has to be found in both of these contexts. In the profile, the information collected is the links of the users that are responsible for the activity and the links of the users who are added or removed. From the tweet the information collected is any mentions and any hashtags. They both represent a link to another network, another context within Twitter. The presence of a URL is discarded. It also represents a link to a context, but this context is external to the Social Network Site.

The time, without being an element of integration, is also essential to know which information and at what time it has to be retrieved. The information about the snapshot is important to relate the activity in the profile with the activity in the tweet (the tweet contains a time stamp that is matched with the time stamp of the snapshot).

This is obviously not the first time that the concept of attention has been studied on Twitter (Guy, 2013). However these articles develop a methodology that includes all the dataset and splits the attention among the studied elements. They obviously include only the collected digital traces. But that has an important consequence, the absence of the digital trace is omitted. Here the novelty is to include this absence in the analysis. This is possible because the aggregation is on the trace of activity itself (as for the referenced studies) but it considers the contextual information around as the entire available information. This contextual information cannot be found outside the tweet and the profile.

For instance, if a user A adds a user B, that means they had to have seen the user B's profile name (and link) somewhere. If one of the user A's friends, let's say user F, has the User B in their list, and retweets one of the user B's tweets, this tweet becomes visible for user A.

In that case, the visibility of user B for user A is through the retweet. This retweet could be the reason why user A added user B (give their attention to the tweet) but this is because they were exposed to the tweet and the user B mention in the tweet. This example shows that the cascade of information makes the action of adding a user possible. However it is not the only case, any link in the tweet or in the profile that links to user B is also valid, that can be a hashtag or a friend of a friend. The key point is that the links need to be carried to user A. This example also shows how it is possible to track the origin of the information. If we know that user A follows user F and user F retweets user B, it is possible to draw some hypothesis on the influence of the retweet on user A's activity. If we extend that to the other possibility, hashtags and mentions, it is easy to see how and why different research is possible on the notion of influence and to track down that origin.

Then if we list all the possibility of links that each context possesses, and despite that, no link is found between user A and user B prior to the activity, it is impossible to know where the influence comes from and how the user A knew of the user B's existence. It is easy to guess different methods that user A can use to get into user B contact. They can use the search engine within twitter, meeting them in person first, clicking by luck on their profile or having their internal program randomise the search and add of new links

In all these scenarios, the links between the two users is impossible to find in any of the visible context, moreover on Twitter and the limits of the API. It is a *black hole* in the omniscience of the digital traces. This is the goal of this last analysis, trying to explore this *black hole* by using the notion of context and activity as proxy to see its presence (but as a black hole, it is impossible to see it).

In consequence, it is possible to split the activity of adding[88] a user into two categories, the *endogenous* and the *exogenous.* The *endogenous* is all the activities that have at least one digital trace that can link both users prior to the activity. The *exogenous* is the category which contains all the activity that does not have at least one digital trace of potential link between the two users.

Therefore, with the assumption that a conscious activity needs an attention first. Considering that attention can only emerge from a context and that context is only visible through digital trace, in absence of those traces, we can suppose that the reason of the activity is external to Twitter.

7.6 Data creation

To build the data for this analysis, a script identify all the activity collected. It retrieved the identity of the person responsible for it (adding or not) and the identity of the user added or removed. Then, by going back in time with the help of the time stamp on the snapshot, it reconstructed the context of both users. First it got the links that both users had at that time and checked for the users that are present in both profiles. That is the profile context. The second context, the tweet is also reconstructed. It get in the past, up to 5 days, collects any mentions present in the tweet as well as hashtags. Initially the option to check for the presence of common hashtag or presence of the users mentioned in at least one of the users' link lists was

[88] This exploration makes only sense in the case of the adding activity. In case of removing a user, the presence of a digital trace is pretty obvious.

used. However, as I did not collect information about the users mentioned, neither about the people participating to the hashtag, a more relaxed approach I decided to include any mentions and any hashtag as a presence of a digital trace of endogenous context.

The dataset obtained is then visualised in a simple categorical analysis to see the proportion of endogenous digital traces and the proportion of exogenous.

7.7 Results

The data is composed of 17197 observation of activity. After removing any activity that have one of the user sets up as protected (in that case it is impossible to access to their context), the number is 16993 with 6049 *removed* type of activity and 10944 *added* type of activity (only the latter is included in the analysis). The Table 8: Presence of digital traces – Results contains all the information about which type of digital trace is present in the dataset.

Presence of a digital trace				
	Hashtag	Shared users	Mentions	Any
FALSE	8554	5902	7867	3749
TRUE	2390	5042	3077	7195
Percentage	21.83 %	46 %	28.1%	65.7%

Table 8: Presence of digital traces – Results

Even if I was really general with the definition of what constitutes a trace of a potential link, the results show that only 65.7% of all the interaction have at least one potential explanation within Twitter. This result is on the entire dataset (minus the protected accounts) and shows that an important part of behaviour within Twitter would not have a direct digital trace of potential connection between users. The part of interaction that have a common links seems higher than the rest of the type of connection. There is no possibility from this simple analysis to withdraw any conclusion about the importance of the links versus the importance of the retweet but show that, even if the million followers fallacy is still considered as valid, the rate of information about common links is significantly higher.

This number is quite high, considering the number of studies on Twitter implying the influence of retweets or the mentions or the hashtag.

8 Conclusions

To conclude, the work of the thesis will be broken into several sections. The first, more general, section returns to the entire work as a global argument around the specificity of the digital traces, and places the findings within the context of actor network theory. The second section returns to the results obtained with the methodology adopted during the work and presents an analysis of the findings. The last section, is devoted to the interdisciplinary aspect of this work and the differences of this work compared to the current field in both computational social science and sociology, placing an emphasis on the absence of socio-demographic information and the representativity.

8.1 On the Actor-Network Theory Perspective

Throughout the thesis, work has been has been affiliated to actor-network theory (ANT), even if ultimately I diverged from the core idea of the ANT methodological principles a number of times when answering specific hypotheses. Alongside these divergence with the ANT, this work also diverge on the way the ANT is applied. The reasons can be found in one critique over ANT, but also on a critique over numerous papers about the impact of digital traces from sociologists.

From past, studies of sciences and technologies, ANT has been used to study researchers in action. It is not limiting its object to that and also develops empirical tools to apply ethnography and description principles on the deployment of the social or on controversies. However the core development of ANT was to deploy tools on science studies to build a critique on sociology at large. Here, instead of studying the deployment of a method assemblage on some social network site (SNS), I developed the argument about the limit and potential of digital traces and SNS *while* studying it.

This echoes a critique on several social sciences studies that describes the impact of digital traces within a larger scope without necessarily explaining the issues of them directly on the field. In the introduction of this work I outlined a problem: not only that sociologists and computer scientists had misunderstandings between each other, but they were actually working on the same object, hoping that the other-world would answer their own questions. Savage feared that sociology loses grasp on its own field when applied to online interaction, while Cha was over-reaching conclusion on the influence of people over other people is measured by the retweet.

It is not that their conclusions or fears were not valid, but it is how they reached these conclusions that has raised some questions. Each was interpreting the other field from a distance, taking a global picture of the other's achievements and pitfalls, and based their conclusions on this vague panorama (from Latour's definition[89]). Cha applied the large notion of influence over the massive data he has in hand. Savage understood the potential of the computing field and the rise of the social network under the scrutiny of the critical sociological were everything is suspicious. More recently, Savage and Burrows bring precision on their initial position (Burrows & Savage, 2014), and other sociologists have developed in-depth argumentation on how social sciences are and need to be deployed into the apparatus of digital devices (Ruppert et al., 2013), but there were few attempts to ground the theoretical developments of these articles into the real world of digital traces and identify in detail where the pitfalls and opportunities are.

However it is only possible to highlight the consequences of digital traces by describing how they are conceived, integrated and analysed in-situ. This approached helped me to deploy ANT on different occasions throughout this work and helped me to create a methodology with which to answer the initial hypothesis. That methodology offers some novelty and was not solely based on an opportunistic perspective. However, that mixed approach had some drawbacks such as not being able to deploy analysis on massive datasets, and arbitrary limiting the population I studied.

The first occasion ANT was deployed was the critique of the 4th Paradigm (or Big Data) as a new form of positivism. Despite the pertinence of such a paradigm with the profusion of data created by SNS and the importance of such online networks within our current form of sociality there is a risk to see the social over the web as only possible under the angle of Big Data and that without considering the impact of such paradigm on the analysis itself and therefore on the potential conclusions drawn on top of it.

[89] "[…] panoramas, as etymology suggests, see *everything*. But they also see *nothing* since they simply *show* an image painted (or projected) on the tiny wall of a room fully *closed* to the outside." (Bruno Latour, 2005, p. 187).

However, the critique against Big Data is not sufficient to refuse using the digital trace itself. In ANT, the emphasis being on the *trace* left by the *social information* also benefits from this digital incarnation. Latour also uses this opportunity to develop a specific methodology to study profiles on the web. However these digital traces are often considered as a black-box, either because the researchers are not interested in what they are (except the impact of a record within a database) or because what produces the digital trace (ultimately the company that owns the SNS), acts as a black-box (L. Weng et al., 2012; Zhu & Lerman, 2016)

Research on SNS are enabled through a contact point - a plug - APIs that allow researchers to collect data. In the case of Twitter, it is the STREAM API that is typically preferred due its ability to allow a high quantity of tweets to be collected in real time (Cook, Conrad, Fowlkes, & Mohebbi, 2011; Lazer et al., 2014).

The tweet is therefore considered as the basic unit for most researches, but the tweet itself - or more precisely, its JSON representation - is a set of hyperlinks (and other less *connectable* information) that can be seen as a technical translation of a pure actor-network. It can be unfolded, flattened and each link can be followed. This link can also be seen as their own full actor-network that can be unfolded and followed too. This is exactly the reason why ANT also developed an interest on hyperlink profiles on the web (Riquelme & González-Cantergiani, 2016).

An equally important aspect of the tweet is the fact it is the only digital trace that it is possible to collect with a timestamp[90]. Timestamps, associated with the massive data collection permitted by the STREAM API, transforms the tweet into an *obligatory passage point*. It becomes the essential plug that holds together the method assemblage composed of the research that need to access digital traces, and Twitter that provides them. Therefore the digital trace may be redfined as: *an interface between the research and the social information, provided by Twitter, being the result of a choice of analysis and technical limitation.*

It is not a broad perspective, but a grounded description based on the technical understanding of the nature of the digital traces. It is not only a trace contained in a database - and the mode of storage is not what has more consequences of our understanding on the social in formation - but the method assemblage itself composed of the need to access huge quantity of data and what is really accessible

[90] The profiles have a creation date but lack the granularity of the tweet.

through the API: the actor-network of the researches and actor-network Twitter plugged together.

Whilst this definition offers a description of what the use of digital trace in social sciences is, it also helps to understand not only the 4th paradigm under the light of ANT, but also the distance between what is studied and what it is supposed to represent. That description is the first outcome of the ANT approach applied to this research. It does not take any position on the good or bad of Big Data, but simply reintegrate the different aspects in play when Big Data is used.

However, the description of the digital trace alone misses the importance of the profile as seen by the users and cannot help to see how activity and context can be redefined within Twitter. Therefore the following definition of an SNS is used:

> A Social Network Site is a network of **actors** jointly engaged in **activities** with other users. These activities leave **digital traces** that are crystallized in a profile. This profile is the user's **context** of any further activity.

That is the first infringement of the principles advocated by ANT. For ANT the description should prevail on any temptation from the researcher to frame the object prior to the research itself. However, I will argue that this definition was an attempt to reword what a digital trace is within the perspective of the profile and the SNS itself. It does not add any preconception on which type of practices are deployed or the reason why people use it. The emphasis placed on the activity and context is a rewording of the social in formation within the SNS.

The second advantage of this definition, is the possibility to use the profile itself as an empirical object by focusing on the aspect of context, actor and activity without first being polluted by complex definitions (Latour et al., 2012).

That transformation into an empirical object was the last step into transforming the whole work into a method assemblage. The first requirement was to decide which device of interessement I will use. That choice was arbitrary and I decided to use the BritAstro account. However, that account was only the seed, a starting point to select a population of users that I monitored to answer the hypothesis. The decision about studying the profiles and monitoring them over time faced several technical limitations. The first one, that ultimately shaped the majority of this research was the necessity to use the REST API rather than the STREAM API. The drastic limits on the number of calls on the API lead to the use of a smaller population and a reduction on the time frame of data collection. The second limit, due to the data collection, was the level of travel it was possible to do. Deploying a full description of the activity as advocated by the 1LS methodology would require a visualisation approach and a more flattened data collection.

This is the most divergent methodological choice in this thesis compared to ANT. This is not only different, it is against ANT. Instead of sticking to a descriptive approach I chose to *aggregate* the links. That aggregation allowed me to use quantitative methods to study the dynamic between the context and the activity, both being defined as the same action (a creation or destruction of a link to a profile or a tweet) but being separated on the aggregation level I applied on them.

Despite this hybrid approach between an ANT perspective and a more traditional methodology, this created a specific empirical object where the creation of the data that is labelled as important to study is dynamically constructed with the activity of the population I studied.

This dynamism in the data collection, not only allows to bypass some of the limitations imposed by the REST API, it is also a new way to conceive a *sample*. There is no prior and imposed category that decides which users or person has importance, but it is the trace they left which decides. Added to this dynamic method of data collection, the type of aggregation also took different angles to answer different questions. The data collection allowed to aggregate the data under principles that were not predetermined, but followed a different angles. By casting several lights on the same dataset, this mitigated the critique about developing aggregation and not being able to see the whole dataset, and instead navigate within it as visualisation tools would have allowed.

These aggregations also give the opportunity to not only answer the questions about the impact of the activity over context, but also to two more issues highlighted in the thesis. The problem about the predefined socio-demographic categories, and the problem of representativity. Only by refusing the opportunistic approach and by building in detail the description of the translation of the digital traces into an empirical object, was it possible to develop such hybrid objects that at the same time render the complexity of digital traces as a distant manifestation of the social.

The next section will detail the conclusions it is possible to draw by applying the method outlined here. The last section will describe in more detail the issue of socio-demographic categories, which is more a sociological issue, and the issue of representativity in the paradigm of whole population, which is more of an issue in computing applied on a social object.

8.2 On the Results

The analysis gave various results that did not necessarily support all the hypotheses, but nevertheless provided interesting results. The first type of analysis, the Engel Granger model, was used to see if the activity of different networks influences each other over time and, if so, in which direction. Over the four hypotheses, only one was supported by the analysis. The model showed that the activity of tweets from the friends network *granger-causes* the link activity in the main network. This relation was unidirectional, there was no support of a *granger cause* from the main network to the friend network with the same activities.

Embedding these findings back into the theory, I interpret this as confirmation of Hypothesis 3 as partial evidence of influence of the context over the activity. The friends network is the direct influence (but not the only one) for any twitter user. It is from this network that the user's timeline is populated with tweets. Therefore if they see more tweets, they have a greater chance to see more tweets, and so they will be exposed to more possible new links, hence the possibility to create more links. The interesting point from this finding is not necessarily in confirming an influence in the activity rate of one context over another one, but a possible confirmation that this influence is short term. The influence was calculated on one lag only. That means that when the context network increases the number of tweets published or removed, this activity creates more activity in the creation (or destruction) of the links for the *Main network* the following day.

Also, the influence is from the tweets to the link activity. Throughout this research, the profile has been defined conceptually and technically as equal to a tweet. That means that both represent a network. However a tweet has a more ephemeral existence as a network than a link that remains more stable over time. The fact that these two types of activity have a relationship can support the idea that high activity, even about ephemeral phenomenon, can quickly impact longer term network formation.

However, the creation of the network, even if it was based on a degree of connection between its members, was mainly a methodological construction. The reason for the creation of this network was to create the possibility to analyse the activity over time. This is why the second analysis was designed to study the individual users rather than aggregate their data into an artificial network.

Tracking the activity over time of each user allows the creation of a measure representing their average rate of activity during the data collection period. The more active the user, the higher the score. The same metrics were built for their respective network while respecting the dynamic nature of their connections. The idea for this analysis was to see if the more active users also have a higher active network. These hypotheses were directly inspired by the idea that a more fluid and changing society requires an increase of fluidity from its members or they would be excluded. Applied on Twitter, it will mean that if a user is not highly active, an active context would exclude him. Then over time the more active users from their context would leave and only the less fluid and less changing ones will remain. On the contrary, an active user would seek other active users and these would be reflected on the overall rate of activity. Again, the tweets and links activity were used as a proxy to build models to test the hypothesis of inclusion/exclusion.

The three linear regression models were statistically significant, however the R-squared were rather low, except for the model 2. This model, the relation between the activity of adding/removing friends in both the users and their context, had a higher R-squared and fo this reason it is possible to see that the homophily may have created an overlap between their respective friends' networks and this relation reflects the activity of shared links. However the fact that all the models were significant is a good indication of the possible influence of the activity context over users' context and that it is possible that the low explicative power of the models is due to the dataset analysed and the nature of interaction on Twitter. The quantity of data may be one of them. The users collected in the dataset are rather quiet and there is no one account that is extensively followed and unfollowed. This is not the reason, but to find a potential effect of the "poor get poorer" and the "richer get richer" effect as found in (Antoniades & Dovrolis, 2015), a bigger dataset may have been needed to observe this long-term effect on social media and to have a better discriminative power to distinguish low activity users and high activity users.

The second reason could be found in the nature of the interactions on Twitter. Myers and Leskovec found that the interactions on Twitter work more on a burst mode (2014). Some actions create a burst in the network that is then propagated and subsequently impacts the network, with the network being shaped through these burst dynamics.

Here again, the reason for not finding such events could be linked to the size of the dataset rather than the absence of relation. But then, the question remains that if something is found in *Big Data*, does that mean it happens everywhere and in more specific situations? To what extent, are the findings relating to massive dataset, where everything can be found, still applicable on a smaller scale, like here, or for more specific users? That question asked, without any straight answer at this point, it still remains that this burst activity cannot be found in the dataset. As shown, the network has a relative constant activity, but it remains stable over the data collection time - the users screened are not really highly active users.

Two issues can therefore be found: the type of users collected (their behaviour is not necessarily really active), and how social media evolves according to research (by burst). To address these issues, it will be necessary to replicate the data collection on another population and use this method first to identify which users are more active to pinpoint the right entry point.

Finally the analysis aims to see what proportion is present of at least one digital trace surrounding the activity of adding someone in the network. The goal was to see to what extent the digital traces of the behaviour within Twitter can be explained with the visible trace of a past link between users, or if the behaviour would have external reasons, outside the APIs. The exploration into the presence of digital traces, prior to the activity of adding a user, shows that only 65.7% of these interactions have at least one of these potential digital traces. That means a huge proportion of these interactions do not have any digital trace that is possible to analyse, a *black hole* of 34.3%. The analysis raises important concerns on the research on SNS.

In almost the entire body of research regarding Twitter, the focus is on one aspect at a time; the retweet, the presence of the hashtag, or the number of followers and so on. The algorithms in these aforementioned research perspectives are extremely developed, but their approach will never question this simple fact: *the representativeness within the Social Network Site*. The problem is inherent to their methodology and the logic of the *whole population.* If the dataset collected all the information, and the study is only on one aspect, there is no need to worry about the representativeness of the phenomena within the dataset, this is just one step of selecting the appropriate data to apply the algorithm. The activity that is not a hashtag or a retweet, is not considered interesting because they do not have this feature under scrutiny. It is rarely questioned to what extent the *absence* of this phenomenon in the other interactions is problematic. This shows again some potential risks associated with the new paradigm of *whole population.*

8.3 Final words about socio-demographic categories and the representativity

This work is not representative of the actual work in computer research (but never tried to be). The methods used in this work involved the use of a more traditional approach, such as sampling and regression analysis. This is distant from the new era of machine learning and the neural network analysis that are currently disrupting the AI field and already our everyday life. One could argue that this work, even before being finished, was outdated.

There are two reasons why this is not the case, even considering the advancement in machine learning or any deep learning and all associated computing development. First, the use of small data, rather than full dataset, is going to be a reality for all of us. This reality will probably not change and even if it changes, it will only move the barrier towards more data being produced, but still not all of it available. The practical reason is obvious; the owners of the API gain considerable power from their data, and the access to the API is often led by strategic and economic reasons, rather than a profound interest in the research. That entire debate about the control over API is not the debate here, but the consequences are visible: social scientists will always have limited access to data and various APIs.

The second reason is more ontological. Due to the nature of the data, from its object, production and inherent logic, sociology will always be needed. The glorification of Big Data comes with the associated illusion of omniscience. It is possible to see everything because everything is recorded. This is the specific assumption that the last analysis tried to tackle. It illustrated the potential existence of a *black hole* in the available information despite the idea of *whole population*. However, if the perspective is uniquely based on the number collected and on the size of the sample, the absence of information is directly in the missing information about the existing one. The main idea is that information is there in the data, but not accessed. But I have not been concerned about that type of missing information but the absolute absence of it.

To unveil this absence, a shift in methodological perspective was needed. It is only by being aware of the importance of the context and the digital traces they leave as the only way to study *social in formation* (Latour, 2005). By recognising that, if information cannot be found in the context, any attempt to understand the reason and the logic of the actor behind the profile is impossible. On this work, the idea of a black hole is a consequence of this perspective, and were specifically understood as the missing information about the activity of removing and adding a user in the friends links. Obviously the situation is different in other SNS and the specific conclusion about the formation of links are only applicable on Twitter. However, this exists on any system producing digital traces. There is no point to defining where the information and which information is missing, since the situation also evolves over time. For example, Tesco approached a new way of collecting information because they were missing one type of information to refine their algorithms (Humby et al., 2008). All SNSs are building different strategies to gather as much information as they can, but some information will always be missed.

Since some information will always be missed, the algorithm will always try to fill that gap by creating a new way to predict this gap. However even the most modern neural network does not give you any insight into the question of *why*. But this is not the goal anyway; the purpose is categorisation, prediction, and modelling, and not understanding the complexity of the underlying social activity.

I showed that the digital trace contains its own complexity. The very process of accessing, collecting, storing, aggregating and analysing creates, at each step, a new layer of translation. The digital traces are imbricated into a network of technological, methodological and practical necessities, opportunities and logic. This specific complexity inherent to the digital trace leads to one conclusion: the process of translation ends up with a digital trace that merely represents the people or the social behaviour.

This conclusion is essential to understand the debate about the post-demographic posture adopted and defended by several sociologists (Rogers, 2009; Ruppert et al., 2013), as well as the associated debate about the over representativity on SNSs and the lack of socio-demographic information.

The use of the socio-demographic categories makes sense when it is directly about a question on what creates their very definition, such as racism or sexism to take the most obvious examples, or in the case of predicting specific event where representativeness is essential, such as elections. However in other cases, the use of these categories is more a collusion of two worlds created with incompatible rules. First, the social categories are built through surveys, reinforced through media and interviews, and used to categorise complexity into smaller and more manageable boxes.

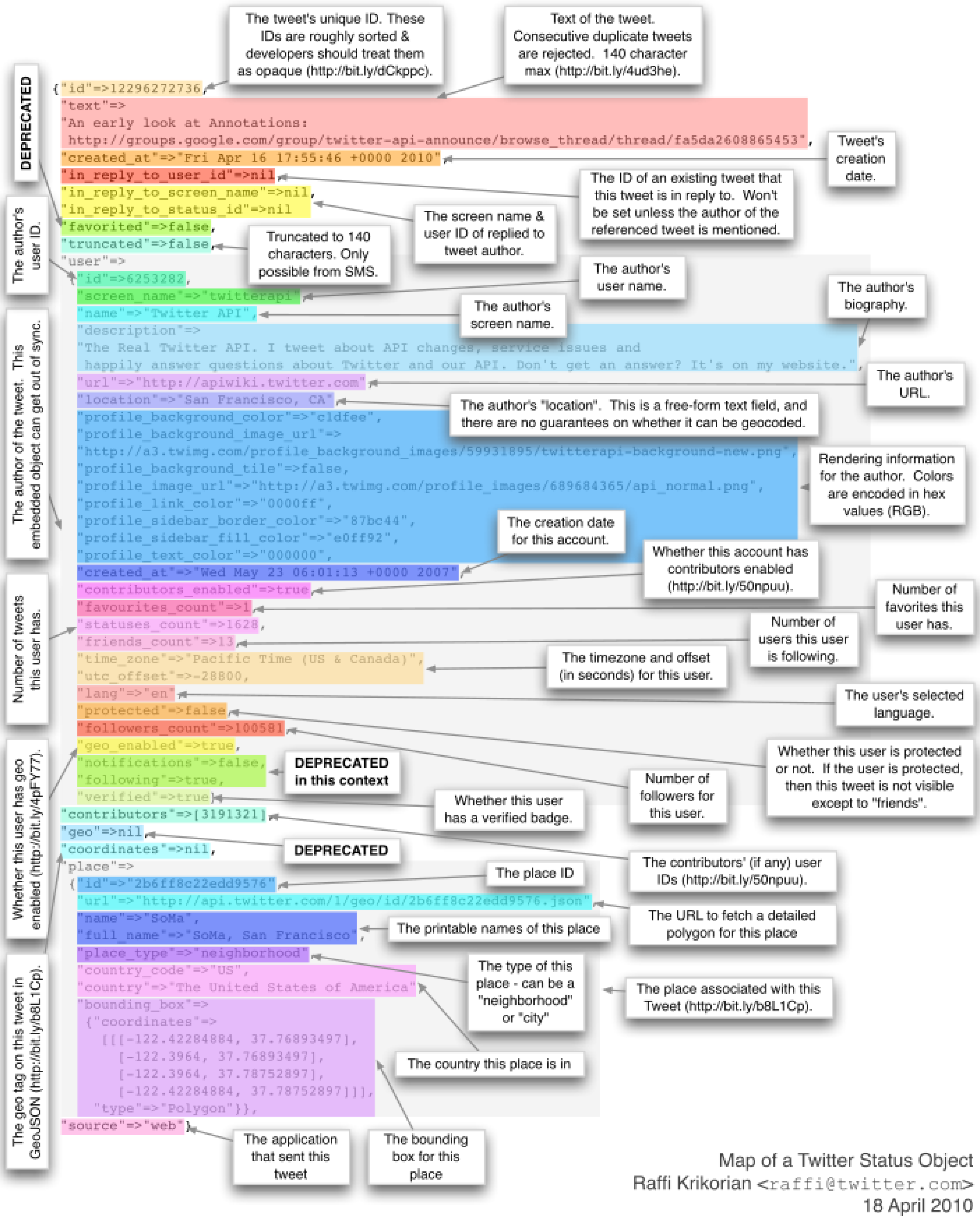
This is exactly the same purpose of the *transactional perspective*, but instead of being built upon the limit of the statistic, it is based on digital information about current activity. They have both their advantages and limits but bringing them together to explain behaviour on the web is more about adding more distance to the studied object rather than affording any solution. The main reason has to be found in where they join together: they are studying people.

However, in the first place, social categories are already not people anymore, they are the result of surveys, theories and *macro* perspectives imposed on the phenomenon studied. More importantly, the digital traces are not equal to people either. It is only by understanding this distinction that sociology can make a difference. The empty space of information left by any digital trace, is where sociological critics are needed, first to unveil it, and second to understand it with the criticality that has made sociology such a valid field, because even if the digital trace are not people, they are definitively the trace of the social in formation.

This is the illusion that this work tried to break without refuting the full potential of the digital traces and the incredible change they bring into the social sciences. However we need to acknowledge that we have to uncover their false promises, hidden in the details, and this work unveils some of them as well as showing what it is possible to do.

9 Appendices

9.1 Visualisation of the tweet object



```
{"id"=>12296272736,
"text"=>
"An early look at Annotations:
 http://groups.google.com/group/twitter-api-announce/browse_thread/thread/fa5da2608865453",
"created_at"=>"Fri Apr 16 17:55:46 +0000 2010",
"in_reply_to_user_id"=>nil,
"in_reply_to_screen_name"=>nil,
"in_reply_to_status_id"=>nil
"favorited"=>false,
"truncated"=>false,
"user"=>
  {"id"=>6253282,
  "screen_name"=>"twitterapi",
  "name"=>"Twitter API",
  "description"=>
  "The Real Twitter API. I tweet about API changes, service issues and
   happily answer questions about Twitter and our API. Don't get an answer? It's on my website.",
  "url"=>"http://apiwiki.twitter.com",
  "location"=>"San Francisco, CA",
  "profile_background_color"=>"c1dfee",
  "profile_background_image_url"=>
  "http://a3.twimg.com/profile_background_images/59931895/twitterapi-background-new.png",
  "profile_background_tile"=>false,
  "profile_image_url"=>"http://a3.twimg.com/profile_images/689684365/api_normal.png",
  "profile_link_color"=>"0000ff",
  "profile_sidebar_border_color"=>"87bc44",
  "profile_sidebar_fill_color"=>"e0ff92",
  "profile_text_color"=>"000000",
  "created_at"=>"Wed May 23 06:01:13 +0000 2007",
  "contributors_enabled"=>true,
  "favourites_count"=>1,
  "statuses_count"=>1628,
  "friends_count"=>13,
  "time_zone"=>"Pacific Time (US & Canada)",
  "utc_offset"=>-28800,
  "lang"=>"en",
  "protected"=>false,
  "followers_count"=>100581,
  "geo_enabled"=>true,
  "notifications"=>false,
  "following"=>true,
  "verified"=>true},
"contributors"=>[3191321],
"geo"=>nil,
"coordinates"=>nil,
"place"=>
 {"id"=>"2b6ff8c22edd9576",
  "url"=>"http://api.twitter.com/1/geo/id/2b6ff8c22edd9576.json",
  "name"=>"SoMa",
  "full_name"=>"SoMa, San Francisco",
  "place_type"=>"neighborhood",
  "country_code"=>"US",
  "country"=>"The United States of America",
  "bounding_box"=>
   {"coordinates"=>
     [[[-122.42284884, 37.76893497],
       [-122.3964, 37.76893497],
       [-122.3964, 37.78752897],
       [-122.42284884, 37.78752897]]],
    "type"=>"Polygon"}},
"source"=>"web"}
```

DEPRECATED

The author's user ID.

The author of the tweet. This embedded object can get out of sync.

Number of tweets this user has.

Whether this user has geo enabled (http://bit.ly/4pFY77).

The geo tag on this tweet in GeoJSON (http://bit.ly/b8L1Cp).

The tweet's unique ID. These IDs are roughly sorted & developers should treat them as opaque (http://bit.ly/dCkppc).

Text of the tweet. Consecutive duplicate tweets are rejected. 140 character max (http://bit.ly/4ud3he).

Tweet's creation date.

The ID of an existing tweet that this tweet is in reply to. Won't be set unless the author of the referenced tweet is mentioned.

The screen name & user ID of replied to tweet author.

Truncated to 140 characters. Only possible from SMS.

The author's user name.

The author's biography.

The author's screen name.

The author's URL.

The author's "location". This is a free-form text field, and there are no guarantees on whether it can be geocoded.

Rendering information for the author. Colors are encoded in hex values (RGB).

The creation date for this account.

Whether this account has contributors enabled (http://bit.ly/50npuu).

Number of favorites this user has.

Number of users this user is following.

The timezone and offset (in seconds) for this user.

The user's selected language.

Whether this user is protected or not. If the user is protected, then this tweet is not visible except to "friends".

Number of followers for this user.

DEPRECATED in this context

Whether this user has a verified badge.

DEPRECATED

The contributors' (if any) user IDs (http://bit.ly/50npuu).

The place ID

The URL to fetch a detailed polygon for this place

The printable names of this place

The type of this place - can be a "neighborhood" or "city"

The place associated with this Tweet (http://bit.ly/b8L1Cp).

The country this place is in

The application that sent this tweet

The bounding box for this place

Map of a Twitter Status Object
Raffi Krikorian <raffi@twitter.com>
18 April 2010

9.2 QQ-Plot of residuals

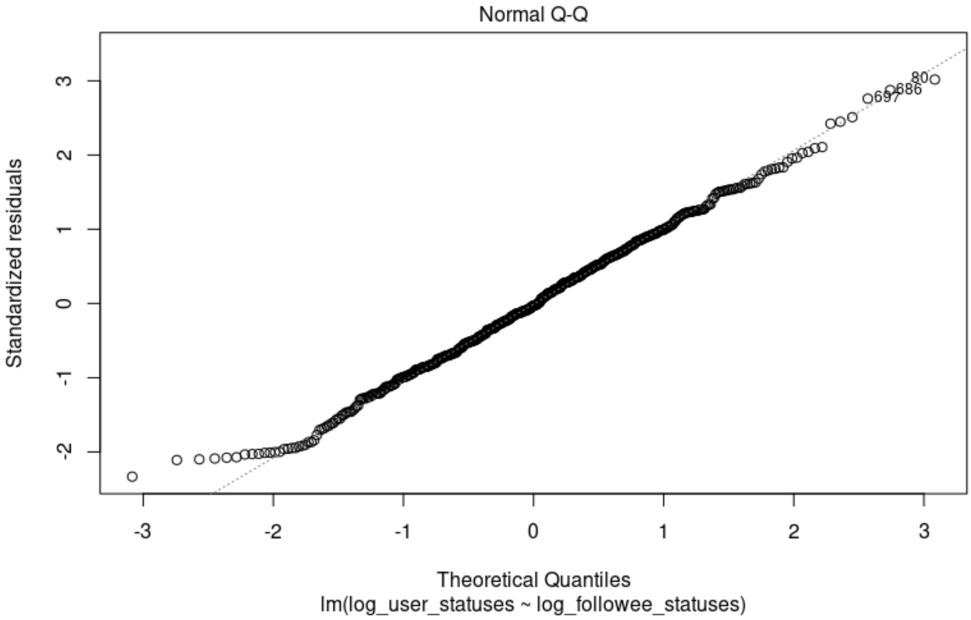


Figure 8: QQ Plot of residuals - Statuses versus Statuses

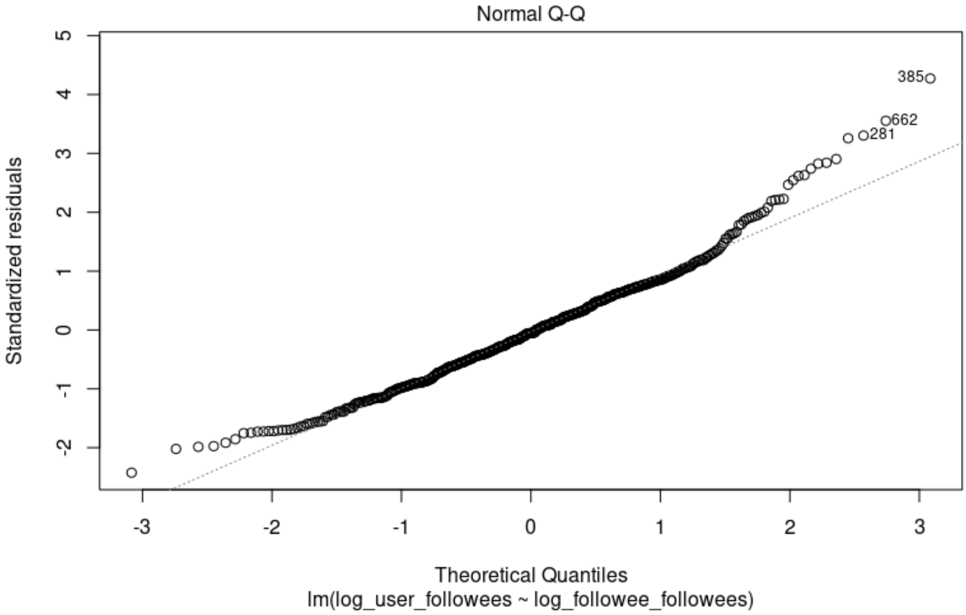


Figure 9: QQ Plot of residuals - Links versus Links

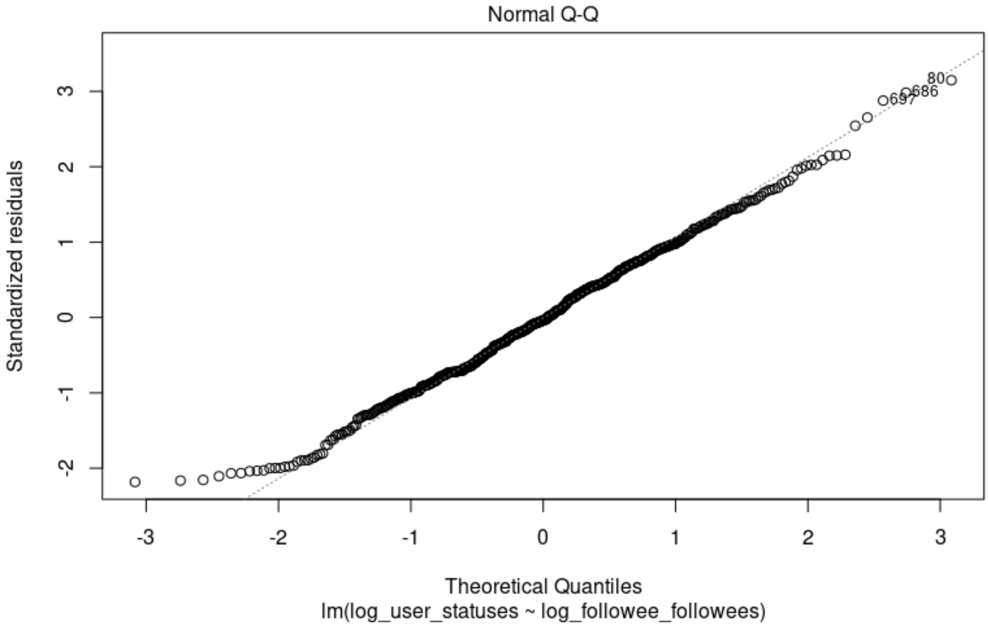


Figure 10: QQ Plot of residuals - Statuses versus Links

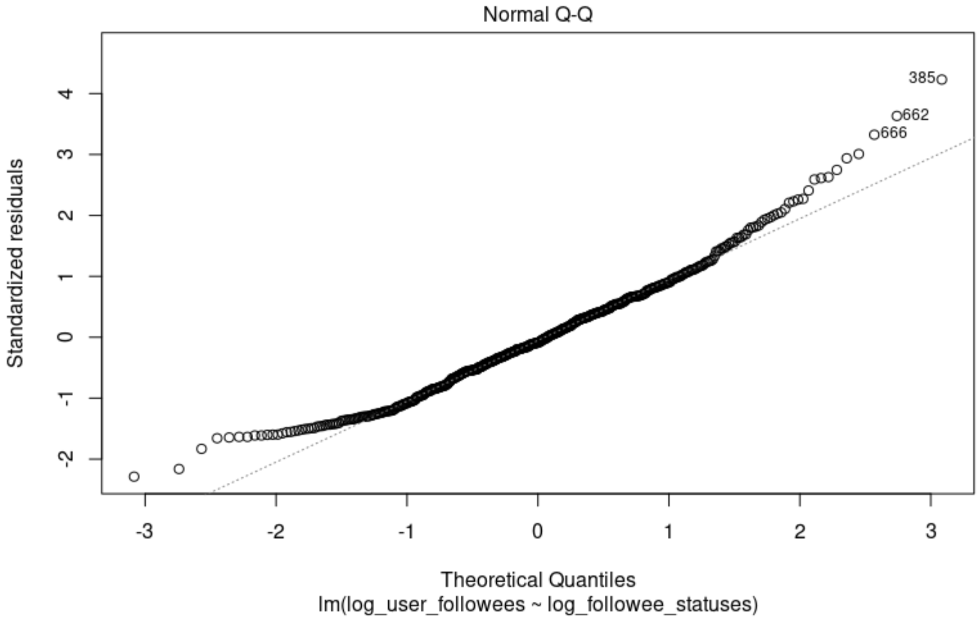


Figure 11: QQ Plot of residuals - Links versus Statuses

9.3 Ethic Application

1. Name(s): Olivier PHILIPPE

2. Current Position: PhD Student

3. Contact Details:

Division/School Sociology and Social Policy, Social Sciences

Email op1e10@ecs.soton.ac.uk

Phone **********

4. Is your study being conducted as part of an education qualification?

YES

5. If Yes, please give the name of your supervisor

Susan Halford, Jeff Vass, Les Carr, Wendy Hall

6. Title of your project:

Implications of Big Data and digital traces in WebScience

7. i) What are the start and completion/hand-in dates of your study?

1St October 2011 – 30th October 2014

ii) When are you planning to start and finish the fieldwork part of your study?

1st February 2014 – 30th April 2014

8. Describe the rationale, study aims and the relevant research questions of your study

This study aims to develop a new methodological approach to conduct research within social media, specifically the Twitter micro-blogging service. Whilst there is considerable research interest in the 'influence' on social media studies of Twitter have – to date – only considered the extend and circulation of 'Retweets' (where users pass on an original post to their own followers).

The aim of this research is to extend the measure of influence and providing contextual information about followers and users' network and explore the relationship between tweets, re-tweets and the network of followers that surround a user.

Research on influence on Twitter is based on the use of Retweet measure only. The aim of this research is to collect information about the list of followers and friends to see test the hypothesis of using different metrics instead of the sole Retweet measure.

By collecting information about the activity of users' link, I will also be able to see if the activity of their environment will impact their own activity.

The global perspective of this research is inter-disciplinary. It aims to bring together sociological considerations about research methods and how they impact the corresponding theories, and computer science, what are the opportunities but also the technical constraints imposed by the system employed for the information collection. In this perspective, the research is mainly focused on the process of the data collection, from the conceptualisation of what constitute an important information among the available data, to the implementation of the collection per se.

As a consequence, the interest is not on participants themselves but on the process of collection and analysis within a digital resource. To achieve this objective, I will collect information about user activity on Twitter only and how they dynamically built their network.

9. Describe the design of your study

The goal of the study is to collect publicly available data on Twitter and analysing the behaviour of user on the Retweet measure.

Access to the data

This data are publicly available via two Twitter Application Programming Interfaces (API).

The Stream API allows researchers to access to the tweets and the REST API provide an access to the users' profile information.

The terms of use for these APIs are set up by Twitter company and the research comply with this rules. This rules can be found at the following address: https://dev.twitter.com/terms/api-terms.

Selection of the population

The research will explore the activity of 200 initial Twitter users within 3 different groups in order to explore the impact of network on the influence.

1. Online group with a clearly existing offline community.

2. Community formed around an Hashtag (a word following the symbol # to allow user to participate to the same discussion) to study the specificity of a community developed around a specific and temporal interest.

3. A randomly chosen users to understand the participation on a more individual perspective.

Information collected

In each case, the study will collect Profile information, Network information and Twitter activity.

1. Profile information.

The profile information will contain the screen_name (the name displayed on the account user), the location (if it is set up by the user), the language (if it is set up by the user) and its id_str (an unique identifier created for each account on Twitter to identify them even if they change their screen_name).

The screen_name is only used to perform some manual check on the data to ensure the script is working properly and will be drop as soon as the collection is done. The language and the location are used jointly to give some information about the sharing characteristics between the participants. The id_str is essential to collect information and to be sure it is the same user is tracked.

2. Network Information.

Two type of links can be found on Twitter, the Followers and the Friends. The Followers is composed by the people who are following the user. The Friends are the people the user is following.

These network information are collected at the same time as the Profile information. But these are static, there is no possibility offered by Twitter to track and build the dynamic of these networks. In order to re-create this information and to obtain a dynamic records of the networks, these two lists are collected recursively for the first and the second list of participants (see point 10). Every time the script screens an user, the list of its current Followers and Friends is updated as well as any change in comparison to the previous list. The list itself contain only id_str. However, this list of id_str is used to access the API and collect Profile Information too, depending on the position of the user in the different list (see point 10).

3. Twitter Activity.

The twitter activity is the tweet posted by the users on their public Timeline.

I access to this information by the Stream API as well as the REST API, depending at which moment the collection of the tweet occurs.

The information collected about the tweet is the text itself, the URL (if it presents), the mentions of other users (if presents), and the time when the tweet was posted.

The text and the URL are used later for the analysis while the presence of mention is directly used to build the sample.

If an user tweet to someone else or being tweet by someone (if this user is in the first list), then it is considered as activity and it is used to track this person and collect the profile information about them.

Stream and REST API.

The Stream API is used to collect the tweet in real time while the REST API is used to collected past tweets. The use of both API is to ensure an efficient approach of the API but it is also needed to be sure all the information is collected.

If an user for them first or the second list as a network activity with a new user, the REST API is used in order to collect its last 3500 tweets. Then it is added to the list of users screened by the Stream API.

Then, from the Stream API, if a user tweet to someone or is being tweeted, the user is added to the list and being screened too.

Storing data

The collection doesn't involve any analysis, neither human intervention (except some checking to ensure the data are properly collected) as the process is automated through a script developed for this purpose.

During the data collection, I will use a specially created twitter account for the research to contact each individual to ask if they have any objection about the analysis of these data.

User will be contacted via the Direct Message system of Twitter to give a link to a website (document attached) which will provide more information about the study, the harvesting of the data, the process of anonymization and the contact information for any enquiry.

At the bottom of the web-page, an opt-out form can give them the opportunity to be removed from the dataset (for the reason of an opt-out system, see the point 13 and 18) if they wish to.

It is planned to send a first message to the people added to the first level of the list (see point 10) before the collection. Then I send a DM right after the collection of the data, then a second message a week after and finally a third message one more week later. After 4 weeks, I will consider that if the participant did not express her wish to be removed from the dataset, that I can use the data for the analysis purpose.

This twitter account can be found here: https://twitter.com/op1e10

Data analysis

The purpose of the thesis is to develop a method adapted to a digital situation. Therefore, the interest is more located on the development of the method and its theoretical implications rather than information about users themselves.

However, to test to test the hypothesis about the influence of context over influence and network users, different metrics are used.

The evolution of number of Followers and Friends

The link shared within the tweets

The number of mentions and Retweet an user do and received.

To conduct the analysis, the dataset will be completely anonymized, removing any information which could lead to the identification of the user (see point 20 and 21).

The analysis will use the information collected about the evolution of the network and see if there is any

The data will be kept at least until the completion of the PhD.

10. Who are the research participants?

There is three levels of participant. The primary one, the second one and the contextual one.

1. The first list of participants: the main users

This people are selected with the method presented above and represent three distinct dataset.

The consent is asked prior to the data collection, as removing them later is more damageable for the research purpose than when they are on the second level.

The information collected about this participants is the Profile Information – Network Information and Twitter Activity.

2. The second list of participants: the activity users

The second list of participant is dynamically created. It depends with who the participants from the first list interact with. If they mention, Retweet, add, remove, are being mentioned, Retweeted, added, removed, the other user is added to a second level list.

The information collected about them is similar as for the first list.

The first difference is that if there is no further interaction after a week, the user is dropped from the list and no further information is collected about her, unless an interaction is again detected with an user from the primary list.

The second difference is that if they interact with other users, this interaction is not used to gather more people, it is just an information kept to know their activity, while when an user from the first list has an interaction with an user, this interaction is used to know which user is needed to be tracked.

3. The third list of participants: the contextual users

The third list is created by every users in the following and friends list from the primary and the secondary list. The amount of information is only limited to the id_str, the number of status published, the number of friends and followers. No more information is collected.

No consent will be asked for this list of user as they are representing a social context rather than an interest on individual level. It is only use to be able to draw a network graph and see the overlapping interaction between the user from the first and the second list.

11. If you are going to analyse secondary data, from where are you obtaining it?

N/A

12. If you are collecting primary data, how will you identify and approach the participants to recruit them to your study?

Please attach a copy of the information sheet if you are using one – or if you are not using one please explain why.

See above, section 10

13. Will participants be taking part in your study without their knowledge and consent at the time (e.g. covert observation of people)? If yes, please explain why this is necessary.

The data collection will take place without participants knowledge. However the data will not be analysed before consent has been given.

The reason is not a need of covert observations but that the selection of the sample is based on activity of users. Therefore it is impossible to ask prior consent for observation.

The second reason is the nature of Twitter itself. An account does not necessarily imply a real person behind. It could be a organisation, a group of people or a robot.

Some account are not active or the user can not connect on its Twitter account during the period of the research and missing the message.

For all these reasons, it is not possible to ask prior consent for user on the second list.

14. If you answered 'no' to question 13, how will you obtain the consent of participants?

N/A

15. Is there any reason to believe participants may not be able to give full informed consent? If yes, what steps do you propose to take to safeguard their interests?

N/A

16. If participants are under the responsibility or care of others (such as parents/carers, teachers or medical staff) what plans do you have to obtain permission to approach the participants to take part in the study?

N/A

17. Describe what participation in your study will involve for study participants. Please attach copies of any questionnaires and/or interview schedules and/or observation topic list to be used

Only observation, no interaction or question to the participants

18. How will you make it clear to participants that they may withdraw consent to participate at any point during the research without penalty?

During the collection is not possible to have the consent, but for the analysis, a private message will be sent with an Twitter account created for this purpose. This account will give detail about the research and contact details.

The message will give a link to a web page describing the purpose of the research, the respect of the anonymity and the possibility to remove the data from the dataset.

19. Detail any possible distress, discomfort, inconvenience or other adverse effects the participants may experience, including after the study, and you will deal with this.

N/A

20. How will you maintain participant anonymity and confidentiality in collecting, analysing and writing up your data?

The collected data will not be anonymized at the first stage of the collection as the identity (represented by the id_str as an unique identifier used by Twitter) is important to ensure the quality of the dataset.

However, prior to any analysis, every identity will be match in a separate database with a random number. Later, this number will be used to conduct analysis. However, some information about profile location will still be stored.

The destruction of the database containing the concordance between random number and Twitter id will be done, only at the very end of the research. It is to allow me to be able to remove people if they are asking for it, even after I started to analyse the data (a possibility is to keep this database as long as the dataset is available, if this latter option is better, then the database will be encrypted and stored on a different server than the server hosting the dataset).

21. How will you store your data securely during and after the study?

The University of Southampton has a Research Data Management Policy, including for data retention. The Policy can be consulted at HYPERLINK "http://www.calendar.soton.ac.uk/sectionIV/research-data-management.html"http://www.calendar.soton.ac.uk/sectionIV/research-data-management.html

The data will be stored on a virtual machine on University server. The only person who has access to it is me. The table having the correspondence between the Twitter Id and the random number will be stored on the personal computer at university. If someone has access to the server, he/she will still not be able to identify the users and an access to the personal computer will not grant access to the server.

22. Describe any plans you have for feeding back the findings of the study to participants.

The feedback about the finding will use the same methods as for contacting the users before and during the analysis, described in 18.

23. What are the main ethical issues raised by your research and how do you intend to manage these?

The collection of data should not be a problem as it is only done for publicly available information. However, people are not necessarily aware about the possibility to analyse their profile. It is why the anonymity is ensured for the analysis and not for the collection.

The opt-out approach is chosen due to the very nature of Twitter. It is impossible to know if the user is a human, a bot, or a company. Therefore, it is only if the user actively express a desire to not be include in the analysis that all data about him/her/its, will be removed.

24. Please outline any other information you feel may be relevant to this submission.

N/A

10 Bibliography

Alaimo, C., & Kallinikos, J. (2017). Computing the everyday: Social media as data platforms. *The Information Society*, *33*(4), 175–191.

Anderson, B. C. (2008). The End of Theory : The Data Deluge Makes the Scientific Method Obsolete. Retrieved 15 May 2016, from https://www.wired.com/2008/06/pb-theory/

Andrews, T. (2012). What is Social Constructionism? *The Grounded Theory Review*, *11*(1).

Antoniades, D., & Dovrolis, C. (2015). Co-evolutionary dynamics in social networks: a case study of Twitter. *Computational Social Networks*, *2*(1).

Apeh, E., & Gabrys, B. (2011). Change Mining of Customer Profiles Based on Transactional Data (pp. 560–567). IEEE.

Apeh, E. T., Gabrys, B., & Schierz, A. (2011). Customer profile classification using transactional data (pp. 37–43). IEEE.

Application programming interface - Wikipedia, the free encyclopedia. (2014, February 17).

ASH. (2016, April 10). The Nuremberg Code.

Association, W. M. (2001). World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects. *Bulletin of the World Health Organization*, *79*(4), 373.

Azman, N., Millard, D. E., & Weal, M. J. (2012). Dark retweets: investigating non-conventional retweeting patterns. In *Social Informatics CN - 0000* (pp. 489–502). Springer.

Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., & Gosling, S. D. (2010). Facebook profiles reflect actual personality, not self-idealization. *Psychological Science*, *21*(3), 372.

Baker, S. A. (2012). From the criminal crowd to the mediated crowd: the impact of social media on the 2011 English riots. *Safer Communities*, *11*(1), 40–49.

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an influencer: Quantifying influence on Twitter (pp. 65–74). ACM.

Bastos, M. T., Mercea, D., & Charpentier, A. (2015). Tents, Tweets, and Events: The Interplay Between Ongoing Protests and Social Media: Tents, Tweets, and Events. *Journal of Communication*, *65*(2), 320–350.

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

Bell, G., Hey, T., & Szalay, A. (2009). Beyond the Data Deluge. *Science*, *323*(5919), 1297–1298.

Bencherki, N. (2012). Mediators and the Material Stabilization of Society. *Communication and Critical/Cultural Studies*, *9*(1), 101–106.

Berger, P. L., & Luckmann, T. (1991). *The Social Construction of Reality: Treatise in the Sociology of Knowledge*. *New York*. London: Penguin Books.

Bergsma, S., Dredze, M., Van Durme, B., Wilson, T., & Yarowsky, D. (2013). Broadly improving user classification via communication-based name and location clustering on twitter (pp. 1010–1019).

Bernstein, M. S., Bakshy, E., Burke, M., & Karrer, B. (2013). Quantifying the invisible audience in social networks (pp. 21–30). ACM.

Bijker, W. E., Hughes, T. P., & Pinch, T. J. (1987). *The Social Construction of Technological Systems*. *The Social Construction of Technological Systems* (Vol. 1).

Bild, D. R., Liu, Y., Dick, R. P., Mao, Z. M., & Wallach, D. S. (2015). Aggregate Characterization of User Behavior in Twitter and Analysis of the Retweet Graph. *ACM Transactions on Internet Technology*, *15*(1), 1–24.

Bloor, D. (1977). *Knowledge and Social Imagery*. *The British Journal of Sociology* (Vol. 28). University of Chicago Press.

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

Boltanski, L., & Thévenot, L. (2006). *On justification: economies of worth*. *Princeton studies in cultural sociology*. Princeton: Princeton University Press.

Bonilla, Y., & Rosa, J. (2015). #Ferguson: Digital protest, hashtag ethnography, and the racial politics of social media in the United States. *American Ethnologist*, *42*(1), 4–17.

Boyd, D., & Crawford, K. (2011). Six provocations for big data. *Social Science Research Network Working Paper Series CN - 0000*.

Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer Mediated Communication CN - 0002*, *13*(1), 210–230.

Bucher, T. (2015). Networking, or What the Social Means in Social Media. *Social Media + Society*, (1), 1–2.

Büchner, A. G., & Mulvenna, M. D. (1998). Discovering internet marketing intelligence through online analytical web usage mining. *ACM Sigmod Record*, *27*(4), 54–61.

Bueger, C., & Bethke, F. (2014). Actor-networking the 'failed state' — an enquiry into the life of concepts. *Journal of International Relations and Development*, *17*, 30–60.

Burrows, R., & Savage, M. (2014). After the crisis? Big Data and the methodological challenges of empirical sociology. *Big Data & Society*, *1*(1), 2053951714540280.

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

Callan, R. J., & Teasdale, A. (1999). Hotel guest history as the foundation for database marketing: Embracing a pilot survey of UK hotels. *Journal of Vacation Marketing*, *5*(2), 140–153.

Callon, M. (1986). Some elements of a sociology of translation: domestication of the scallops and the fishermen of St. Brieuc Bay. *Power, Action, and Belief: A New Sociology of Knowledge*, *32*, 196–223.

Carberry, J. (2014, March 11). Media statement on Cornell University's role in Facebook 'emotional contagion' research.

Castells, M. (2004). *The Power of Identity: The Information Age: Economy, Society and Culture*.

Castells, M. (2009). *Communication power*. OUP Oxford.

Castells, M. (2011). *The rise of the network society: The information age: Economy, society, and culture* (Vol. 1). Wiley-Blackwell.

Celik, I., Abel, F., & Houben, G. J. (2011). Learning Semantic Relationships between Entities in Twitter. *Web Engineering*, 167–181.

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. P. (2010). Measuring user influence in twitter: The million follower fallacy.

Chen, M., Mao, S., Liu, Y., Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A Survey. *Mobile Netw Appl*, *19*(2), 171–209.

Chretien, K. C., Tuck, M. G., Simon, M., Singh, L. O., & Kind, T. (2015). A Digital Ethnography of Medical Students who Use Twitter for Professional Development. *Journal of General Internal Medicine*, *30*(11), 1673–1680.

Clogg, C. C. (1992). The Impact of Sociological Methodology on Statistical Methodology. *Statistical Science*, *7*(2), 183–196.

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM*, *13*(6), 377–387.

Colombo, G. B., Burnap, P., Hodorog, A., & Scourfield, J. (2016). Analysing the connectivity and communication of suicidal users on twitter. *Computer Communications*, *73*, 291–300.

Cook, S., Conrad, C., Fowlkes, A. L., & Mohebbi, M. H. (2011). Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. *PLoS ONE*, *6*(8), e23610.

Culotta, A. (2010). Detecting influenza outbreaks by analyzing Twitter messages. *arXiv Preprint arXiv:1007.4748*.

Darch, P. (2014). Managing the Public to Manage Data: Citizen Science and Astronomy. *International Journal of Digital Curation*, *9*(1).

Daries, J. P., Reich, J., Waldo, J., Young, E. M., Whittinghill, J., Ho, A. D., … Chuang, I. (2014). Privacy, anonymity, and big data in the social sciences. *Communications of the ACM*, *57*(9), 56–63.

Diesner, J. (2015). Small decisions with big impact on data analytics. *Big Data & Society*, *2*(2), 1–6.

Dosi, G. (1982). Technological paradigms and technological trajectories: a suggested interpretation of the determinants and directions of technical change. *Research Policy*, *11*(3), 147–162.

Dosi, G. (1988). Sources, procedures, and microeconomic effects of innovation. *Journal of Economic Literature*, 1120–1171.

Duggan, M., & Brenner, J. (2013). *The Demographics of Social Media Users, 2012*. Pew Research Center's Internet & American Life Project.

Durkheim, E. (1894). *Les règles de la méthode sociologique*. *Revue Philosophique de la France et de l'Étranger*.

Durkheim, E., Lukes, S., & Halls, W. D. (1982). *The rules of sociological method*. New York: Free Press.

Erickson, G. S., Ph, D., & Rothberg, H. N. (2005). Expanding Intelligence Capabilities : Downstream Knowledge Targets. *Journal of Competitive Intelligence and Management*, *3*(2), 8–15.

Ertöz, L., Steinbach, M., & Kumar, V. (2003). Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data (Vol. 47).

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise (Vol. 1996, pp. 226–231). AAAI Press.

Facebook. (2014). *Facebook Reports Fourth Quarter and Full Year 2013 Results*. Facebook.

Fisher, R. A. (1958). *Statistical methods for research workers*.

Friedberg, E., & Crozier, M. (1977). L'acteur et le système. *Ed Seuil, Paris*.

Friedman, B., Lin, P., & Miller, J. K. (2005). Informed consent by design. *Security and Usability*, (2001), 503–530.

Fu, J. S., & Shumate, M. (2017). News media, social media, and hyperlink networks: An examination of integrated media effects.

Fuchs, C. (2009). Some Reflections on Manuel Castells' Book 'Communication Power'. *tripleC*, *1*(7), 94–108.

Fuchs, C. (2010). Labor in Informational Capitalism and on the Internet. *The Information Society CN - 0090*, *26*(3), 179–196.

Fuller, S., & Hond, F. Den. (1999). Review Essay. *Science, Technology, & Human Values*, *24*(1), 159–166.

Fuller, S., & Webster, F. (2005). Another sense of the information age. *Information, Communication & Society*, *8*(4), 459–463.

Galaxy Zoo. (n.d.). Retrieved 16 May 2017, from https://www.galaxyzoo.org/

Gallos, L. K., Rybski, D., Liljeros, F., Havlin, S., & Makse, H. A. (2012). How people interact in evolving online affiliation networks. *Physical Review X*, *2*(3), 31014.

Garnham, N. (1998). Information Society Theory as Ideology: A Critique. *Loisir et Société / Society and Leisure*, *21*(1), 97–120.

Gayo-Avello, D. (2012). A meta-analysis of state-of-the-art electoral prediction from Twitter data. *arXiv Preprint arXiv:1206.5851*.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). 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.

Gomer, R., & Gerding, E. (2014). Consenting agents: semi-autonomous interactions for ubiquitous consent (pp. 653–658). ACM.

Google Flu Trends. (2015, November 16).

Granger, C. W. (1980). Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and Control*, *2*, 329–352.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 424–438.

Granger, C. W. J., & Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*, *2*(2), 111–120.

Guille, A., Hacid, H., & Favre, C. (2013). Predicting the Temporal Dynamics of Information Diffusion in Social Networks. *arXiv Preprint arXiv:1302.5235*.

Guo, H.-D., Zhang, L., & Zhu, L.-W. (2015). Earth observation big data for climate change research. *Advances in Climate Change Research*, *6*(6), 108–117.

Guo, S., Ladroue, C., & Feng, J. (2010). Granger causality: theory and applications. In *Frontiers in Computational and Systems Biology* (pp. 83–111). Springer.

Guy, N. (2013). A test for second-order stationarity and approximate confidence intervals for localized autocovariances for locally stationary time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *75*(5), 879–904.

He, Z., & Maekawa, K. (2001). On spurious Granger causality. *Economics Letters*, *73*(3), 307–313.

Heffetz, O., & Ligett, K. (2014). Privacy and data-based research. *The Journal of Economic Perspectives*, *28*(2), 75–98.

Hernandez-Orallo, J. (2013). A short note on estimating intelligence from user profiles in the context of universal psychometrics: prospects and caveats. *arXiv Preprint arXiv:1305.1655*.

Hey, A. J. G., Tansley, S., & Tolle, K. M. (2009). *The fourth paradigm: data-intensive scientific discovery*. Microsoft Research Redmond, WA.

Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science*, *332*(6025), 60.

Hofer, M., & Aubert, V. (2013). Perceived bridging and bonding social capital on Twitter: Differentiating between followers and followees. *Computers in Human Behavior*, *29*(6), 2134–2142.

Honda, K., & Ichihashi, H. (2004). Linear fuzzy clustering techniques with missing values and their application to local principal component analysis. *Fuzzy Systems, IEEE Transactions on*, *12*(2), 183–193.

Hopcroft, J., Lou, T., & Tang, J. (2011). Who will follow you back?: reciprocal relationship prediction (pp. 1137–1146). ACM.

Hsu, C. N., Chung, H. H., & Huang, H. S. (2004). Mining skewed and sparse transaction data for personalized shopping recommendation. *Machine Learning CN - 0026*, *57*(1), 35–59.

Humby, C., Hunt, T., & Phillips, T. (2008). *Scoring points: How Tesco continues to win customer loyalty*. Kogan Page Ltd.

Hutto, C. J., Yardi, S., & Gilbert, E. (2013). A longitudinal study of follow predictors on twitter (pp. 821–830). ACM.

Hyndman, R. J., & Athanasopoulos, G. (2014). *Forecasting: principles and practice*. OTexts.

Jacobs, A. (2009). The pathologies of big data. *Communications of the ACM CN - 0056*, *52*(8), 36–44.

Jahr, M. (2016, March 10). Never miss important Tweets from people you follow. *Twitter Blogs*.

Jernigan, C., & Mistree, B. F. T. (2009). Gaydar: Facebook friendships expose sexual orientation. *First Monday*, *14*(10).

John, B. (1996, July 18). A Declaration of the Independence of Cyberspace.

JSON. (2014). Retrieved 28 February 2014, from http://json.org/

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: an introduction to cluster analysis* (Vol. 39). Wiley Online Library.

Kietzmann, J. H., Hermkens, K., Mccarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, *54*, 241–251.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 205395171452848.

Kooti, F., Yang, H., Cha, M., Gummadi, K., & Mason, W. A. (2012). The Emergence of Conventions in Online Social Networks.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802–5805.

Kramer, A. D. I. (2012). The spread of emotion via Facebook (pp. 767–770). ACM.

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790.

Krauss, S. E., & Putra, U. (2005). Research Paradigms and Meaning Making : A Primer. *The Qualitative Report*, *10*(4), 758–770.

Kraut, R., Patterson, M., Lundmark, V., Kiesler, S., Mukophadhyay, T., & Scherlis, W. (1998). Internet paradox: A social technology that reduces social involvement and psychological well-being? *American Psychologist*, *53*(9), 1017.

Kuhn, T. S. (1970). *The Structure of Scientific Revolution* (2d ed.). Chicago.

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

Latour, B. (1993). *We have never been Modern*. Harvard University Press.

Latour, B. (1996). On actor-network theory. *Soziale Welt*, *47*(4), 369–381.

Latour, B. (1999). On Recalling Ant. *The Sociological Review*, *47*(1_suppl), 15–25.

Latour, B. (2005). *Reassembling the social: an introduction to actor-network-theory*. Oxford University Press.

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

Latour, B. (2013). *Nous n'avons jamais été modernes*. La découverte.

Latour, B., Jensen, P., Venturini, T., Grauwin, S., & Boullier, D. (2012). The whole is always smaller than its parts a digital test of Gabriel Tardes' monads. *The British Journal of Sociology*, *63*(4), 590–615.

Latour, B., & Woolgar, S. (1986). *Laboratory life: The construction of scientific facts*. Princeton University Press.

Law, J. (1999). After ANT: complexity, naming and topology. *The Sociological Review*, *47*(1), 1–14.

Law, J. (2004). *After method: Mess in social science research*. Routledge.

Law, J. (2009). Actor network theory and material semiotics. *The New Blackwell Companion to Social Theory*, *3*, 141–158.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, *343*(6176), 1203–1205.

Liew, C. S., Atkinson, M. P., Galea, M., Ang, T. F., Martin, P., & Hemert, J. I. Van. (2016). Scientific Workflows. *ACM Computing Surveys*, *49*(4), 1–39.

Lin, J. (2015). On Building Better Mousetraps and Understanding the Human Condition: Reflections on Big Data in the Social Sciences. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 33–47.

Linden, G., Smith, B., & York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *Internet Computing, IEEE*, *7*(1), 76–80.

Liu, W., Al Zamal, F., & Ruths, D. (2012). Using social media to infer gender composition of commuter populations.

Liu, W., & Ruths, D. (2013). What's in a Name? Using First Names as Features for Gender Inference in Twitter.

Luhn, H. P. (1958). A business intelligence system. *IBM Journal of Research and Development*, *2*(4), 314–319.

Lynch, M. (2016). Social Constructivism in Science and Technology Studies. *Human Studies*, *39*(1), 101–112.

MacMullen, W. J., & Denn, S. O. (2005). Information problems in molecular biology and bioinformatics. *Journal of the American Society for Information Science and Technology*, *56*(5), 447–456.

Makarenko, A. V. (2011). Phenomenological Model for Grown of Volumes Digital Data. *Arxiv Preprint arXiv:1102.5500*.

Marreiros, H., Gomer, R., Vlassopoulos, M., Tonin, M., & schraefel, m c. (2015). Exploring user perceptions of online privacy disclosures.

Marres, N., & Gerlitz, C. (2016). Interface methods: renegotiating relations between digital social research, STS and sociology. *The Sociological Review*, *64*, 21–46.

Mason, W. A., Conrey, F. R., & Smith, E. R. (2007). Situating social influence processes: Dynamic, multidirectional flows of influence within social networks. *Personality and Social Psychology Review CN - 0096*, *11*(3), 279–300.

Meyer, M. (2014). Misjudgements will drive social trials underground. *Nature*, *511*(7509), 265.

Michael, M. (2016). *Actor-network theory: trials, trails and translations*.

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. N. (2011). Understanding the Demographics of Twitter Users.

Mitscherlich, A., & Mielke, F. (1949). Doctors of infamy: the story of Nazi medical crimes.

Mol, A. (2010). Actor-Network Theory: sensitive terms and enduring tensions. *Kölner Zeitschrift Für Soziologie Und Sozialpsychologie* , *50*(1), 253–269.

Moore, M. J., Nakano, T., Enomoto, A., & Suda, T. (2012). Anonymity and roles associated with aggressive posts in an online forum. *Computers in Human Behavior*, *28*(3), 861–867.

Mueller, J., & Stumme, G. (2017). Predicting Rising Follower Counts on Twitter Using Profile Information, *1017*.

Murdoch, J. (1997). Inhuman/nonhuman/human: actor-network theory and the prospects for a nondualistic and symmetrical perspective on nature and society. *Environment and Planning D: Society and Space*, *15*, 731–756.

Murray, M. P. (1994). A drunk and her dog: an illustration of cointegration and error correction. *The American Statistician*, *48*(1), 37–39.

Myers, S. A., Zhu, C., & Leskovec, J. (2012). Information diffusion and external influence in networks (pp. 33–41). ACM.

Nettleton, D. F. (2013). Data mining of social networks represented as graphs. *Computer Science Review*.

Nguyen, D., Gravel, R., Trieschnigg, D., & Meder, T. (2013). 'How Old Do You Think I Am?': A Study of Language and Age in Twitter.

O'Dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C., & Christensen, H. (2015). Detecting suicidality on Twitter. *Internet Interventions*, *2*(2), 183–188.

Paltrinieri, R., & Esposti, P. D. D. (2013). Processes of Inclusion and Exclusion in the Sphere of Prosumerism. *Future Internet*, *5*(1), 21–33.

Park, Y. J., & Tuzhilin, A. (2008). The long tail of recommender systems and how to leverage it (pp. 11–18). ACM.

Pescosolido, B. A., & Mendelsohn, R. (1986). Social Causation or Social Construction of Suicide? An Investigation into the Social Organization of Official Rates. *Source American Sociological Review*, *51*(1), 80–100.

Poell, T., & Van Dijck, J. (2015). Social media and activist communication. *Poell, Thomas & José van Dijck (2015). Social Media and Activist Communication. In The Routledge Companion to Alternative and Community Media*, 527–537.

Pond, P. (2016). Twitter Time: A Temporal Analysis of Tweet Streams During Televised Political Debate. *Television & New Media*, *17*(2), 142–158.

Poster, M. (2013). *The Mode of Information: Poststructuralism and Social Contexts*. John Wiley & Sons.

Postill, J., & Pink, S. (2012). Social Media Ethnography: The Digital Researcher In A Messy Web, *145*(1), 123–134.

Prensky, M. (2009). H . Sapiens Digital : From Digital Immigrants and Digital Natives to Digital Wisdom Digital Wisdom. *Journal of Online Education*, *5*(3), 1–9.

Privacy Policy. (2016, March 10). *Twitter*.

Raylene, Y. (2013, August 23). Adding What You're Doing to Status Updates - Facebook Newsroom. *Facebook Newsroom*.

Resnick, P., Adar, E., & Lampe, C. (2015). What Social Media Data We Are Missing and How to Get It. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 192–206.

REST API v1.1 Resources. (2013, July 15). Retrieved from https://dev.twitter.com/docs/api/1.1

Restivo, S., & Croissant, J. (2008). Social Constructionism in Science and Technology Studies. In *Handbook of constructionist research* (p. 822). Guilford Press.

Rey, P. J. (2012). Alienation, Exploitation, and Social Media. *American Behavioral Scientist*, *56*(4), 399–420.

Riquelme, F., & González-Cantergiani, P. (2016). Measuring user influence on Twitter: A survey. *Information Processing & Management*, *52*(5), 949–975.

Ritzer, G., & Jurgenson, N. (2010). Production, Consumption, Prosumption The nature of capitalism in the age of the digital 'prosumer'. *Journal of Consumer Culture CN - 0251*, *10*(1), 13–36.

Rohweder, C. (2006, July 5). No. 1 Retailer in Britain Uses 'Clubcard' to Thwart Wal-Mart. *The Wall Street Journal CN - 0000*.

Rowley, J. (2007). Reconceptualising the strategic role of loyalty schemes. *Journal of Consumer Marketing*, *24*(6).

Ruppert, E., Law, J., & Savage, M. (2013). Reassembling Social Science Methods: the challenge of digital devices. *Theory, Culture & Society*, *30*(4), 22–46.

Rushe, D. (2013). Google: don’t expect privacy when sending to Gmail. *The Guardian. Retrieved Dec*, *19*.

Rygielski, C., Wang, J.-C., & Yen, D. C. (2002). Data mining techniques for customer relationship management. *Technology in Society*, *24*(4), 483–502.

Savage, M. (2010). *Identities and Social Change in Britain since 1940: the politics of method*. Oxford University Press Oxford.

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

Sayes, E. (2014). Actor–Network Theory and methodology: Just what does it mean to say that nonhumans have agency? *Social Studies of Science*, *44*(1), 134–149.

Sayyadi, H., Hurst, M., & Maykov, A. (2009). Event detection and tracking in social streams.

Schroeder, R. (2014). Big Data and the brave new world of social media research. *Big Data & Society*, *1*(2).

Segre, S. (2016). Social Constructionism as a Sociological Approach. *Human Studies*, *39*(1), 93–99.

Shah, D. V, Cappella, J. N., & Neuman, W. R. (2015). Big Data, Digital Media, and Computational Social Science: Possibilities and Perils. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 6–13.

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

Silvertown, J., & Al., E. (2009). A new dawn for citizen science. *Trends in Ecology & Evolution*, *24*(9), 467–71.

Skoric, M., Poor, N., Achananuparp, P., Lim, E.-P., & Jiang, J. (2012). Tweets and Votes : A Study of the 2011 Singapore General Election. In *System Science (HICSS), 2012 45th Hawaii International Conference on* (pp. 2583–2591). IEEE.

Sloan, L., & Morgan, J. (2015). Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. *PloS One*, *10*(11), e0142209.

Sloan, L., Morgan, J., Burnap, P., & Williams, M. (2015). Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PloS One*, *10*(3).

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. N. (2000). Web usage mining: Discovery and applications of usage patterns from web data. *ACM SIGKDD Explorations Newsletter*, *1*(2), 12–23.

Stapleton, L. K. (2011, March 10). Taming big data.

Steadman, I. (2013). Big data, language and the death of the theorist | WIRED UK. Retrieved 16 May 2016, from http://www.wired.co.uk/article/big-data-end-of-theory

Stone, M., Bearman, D., Butscher, S. A., Gilbert, D., Crick, P., & Moffett, T. (2004). The effect of retail customer loyalty schemes detailed measurement or transforming marketing? *Journal of Targeting, Measurement and Analysis for Marketing*, *12*(3), 305–318.

Sueki, H. (2015). The association of suicide-related Twitter use with suicidal behaviour: A cross-sectional study of young internet users in Japan. *Journal of Affective Disorders*, *170*, 155–160.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(5), 557–570.

Thomas, K., Grier, C., Song, D., & Paxson, V. (2011). Suspended accounts in retrospect: an analysis of twitter spam (pp. 243–258). ACM.

Thrift, N. J. (2005). *Knowing capitalism*. Sage Publications Ltd.

Tinati, R., Carr, L., Hall, W., & Bentwood, J. (2012). Scale Free: Twitter's Retweet Network Structure.

Toffler, A., Longul, W., & Forbes, H. (1981). *The third wave*. Bantam books New York.

Tracking flu trends. (2008, November 16). *Official Google Blog*.

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

Tweets | Twitter Developers. (2013, September 10).

Twitter. (2014). *Twitter Reports Fourth Quarter and Fiscal Year 2013 Results*. Twitter Inc.

Union, I. T. (2013). *World Telecommunication/ICT Indicators database*. Genova.

Urry, J. (2000). Mobile sociology. *The British Journal of Sociology*, *51*(1), 185–203.

Veletsianos, G. (2017). Three Cases of Hashtags Used as Learning and Professional Development Environments. *TechTrends*, *61*(3), 284–292.

Vitali, S., Glattfelder, J. B., & Battiston, S. (2011). The network of global corporate control. *PloS One*, *6*(10), e25995.

Von Tunzelmann, N., Malerba, F., Nightingale, P., & Metcalfe, S. (2008). Technological paradigms: past, present and future. *Industrial and Corporate Change*, *17*(3), 467–484.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge Univ Pr.

Wattenhofer, M., Wattenhofer, R., & Zhu, Z. (2012). The YouTube Social Network.

Webster, F. (1997). Is this the information age? *City*, *2*(8), 71–84.

Webster, F. (2004). Cultural studies and sociology at, and after, the closure of the Birmingham school. *Cultural Studies*, *18*(6), 847–862.

Weindling, P. (2001). The origins of informed consent: the international scientific commission on medical war crimes, and the Nuremberg Code. *Bulletin of the History of Medicine*, *75*(1), 37–71.

Weng, J., Lim, E. P., Jiang, J., & He, Q. (2010). Twitterrank: finding topic-sensitive influential twitterers (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*, 1–8.

Wiener, N. (1956). The theory of prediction. *Modern Mathematics for Engineers*, *1*, 125–139.

Wilson, R. E., Gosling, S. D., & Graham, L. T. (2012). A review of Facebook research in the social sciences. *Perspectives on Psychological Science*, *7*(3), 203–220.

Wittel, A. (2001). Toward a network sociality. *Theory, Culture & Society*, *18*(6), 51–76.

Won, H.-H., Myung, W., Song, G.-Y., Lee, W.-H., Kim, J.-W., Carroll, B. J., & Kim, D. K. (2013). Predicting National Suicide Numbers with Social Media Data. *PLoS ONE*, *8*(4), e61809.

Xu, R., & Donald Wunsch, I. I. (2005). Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645.

Xu, Z., & Yang, Q. (2012). Analyzing User Retweet Behavior on Twitter (pp. 46–50). IEEE.

Yang, L., Sun, T., Zhang, M., & Mei, Q. (2012). We know what@ you# tag: does the dual role affect hashtag adoption? (pp. 261–270). ACM.

YouTube - Statistics. (2014, February 15). Retrieved from http://www.youtube.com/yt/press/statistics.html

Zevin, M., Coughlin, S., Bahaadini, S., Besler, E., Rohani, N., Allen, S., … Kalogera, V. (2016). Gravity Spy: Integrating Advanced LIGO Detector Characterization, Machine Learning, and Citizen Science. *Class. Quant. Grav.*, 1–27.

Zhao, Q., & Mitra, P. (2007). Event detection and visualization for social text streams. *Proceedings of ICWSM'2007*, 26–28.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval* (pp. 338–349). Springer.

Zhu, L., & Lerman, K. (2016). Attention Inequality in Social Media. *arXiv Preprint arXiv:1601.07200*.

Zwick, D., & Dholakia, N. (2004). Whose identity is it anyway? Consumer representation in the age of database marketing. *Journal of Macromarketing*, *24*(1), 31–43.