Effectiveness of User-Curated Filtering as Coping Strategy for Information Overload on Microblogging Services

by

Simon de la Rouviere

Thesis presented in partial fulfilment of the requirements for the degree of Master of Arts in Socio-Informatics in the Faculty of Arts at Stellenbosch University

Department of Information Science,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Supervisor: Mr. K. Ehlers

April 2014

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: .....

Abstract

Effectiveness of User-Curated Filtering as Coping Strategy for Information Overload on Microblogging Services

We are living in an increasingly global and connected society with information creation increasing at exponential rates. The research sets out to help solve the problem of mitigating the effects of information overload in order to increase the novelty of our interactions in the digital age. Online social-networks and microblogging services allow people across the world to take part in a public conversation. These tools have inherent constraints on how much communication can feasibly occur. Become too connected and a user will receive too much information to reasonably process. On Twitter (a microblogging service), lists are a tool for users to create separate feeds. The research determines whether lists are an effective tool for coping with information overload (abundance of updates). Using models of sustainable online discourse and information overload on computer-mediated communication tools, the research found that lists are an effective tool to cope with information overload on microblogging services. Quantitatively, individuals who make use of lists follow more users and when they start using lists they increase the amount of information resources (following other users) at a greater rate than those who do not use lists. Qualitatively, the research also provides insight into the reasons why people use lists. The research adds new academic relevance to 'information overload' and 'online sustainability' models previously not used in the context of feed-based online CMC tools, and deepens the understanding and importance of user-curated filtering as a way to reap the benefits from the increasing abundance of information in the digital age.

Uittreksel

Die Doeltreffendheid van Gebruiker-saamgestelde Filtrering as 'n Strategie vir die Hantering van Inligtingoorlading op Mikroblog-dienste

("Effectiveness of User-Curated Filtering as Coping Strategy for Information Overload on Microblogging Services")

Ons leef in 'n toenemend globale en gekonnekteerde samelewing waarin inligtingskepping toeneem teen 'n eksponensiële koers. Hierdie navorsing het ten doel om die newe-effekte van die oorvloed van inligting te verlig sodat daar meer waarde uit ons interaksies in die digitale era kan geput kan word. Aanlyn sosiale-netwerke en mikroblog-dienste laat mense wêreldwyd toe om deel te neem in 'n openbare gesprek. Hierdie aanlyn gereedskap het egter inherente beperkinge op hoeveel kommunikasie prakties moontlik is. Wanneer gebruikers té gekonnekteer raak, word daar te veel ingligting ontvang om redelikerwys verwerk te kan word. Op Twitter ('n mikroblog-diens) is lyste 'n hulpmiddel waarmee gebruikers afsonderlike strome van inligting kan skep. Deur die gebruik van modelle van 'volhoubare aanlyn diskoers' en 'inligtingoorlading', bewys hierdie navorsing dat lyste 'n doeltreffende hulpmiddel is om die oorvloed van inligting te verlig op mikroblog-dienste. Kwantitatief volg gebruikers wat lyste gebruik meer gebruikers vergeleke met die wat nie lyste gebruik nie. Wanner hul lyste begin gebruik, volg hulle gebruikers teen 'n hoër koers as dié wat nie lyste gebruik nie. Kwalitatief bied die navorsing ook insig oor die redes vir die gebruik van lyste. Die navorsing onderstreep die akademiese relevansie van 'inligtingoorlading' en 'aanlyn volhoubaarheid' modelle wat nie voorheen gebruik is in die konteks van stroom-gebaseerde aanlyn gereedskap nie, en verdiep die begrip en belangrikheid van gebruiker-saamgestelde filtrering as 'n manier om die voordele te trek uit die toenemende oorvloed van inligting in die digitale era.

Acknowledgements

Thanks to Kobus for being a fantastic supervisor. Thanks to all the people at the MIH Media Lab: I will miss all the intellectual discussions around the coffee machine. Thanks to my parents for the support and allowing me to move back home to save money.

Dedications

This thesis is dedicated to all the people downstream from my feeds who at some point had to deal with irrelevant updates.

Contents

Declaration i

Abstract ii

Uittreksel iii

Acknowledgements iv

Dedications v

Contents vi

List of Figures viii

List of Tables ix

Table of Abbreviations x

1 Introduction 1
- 1.1 Information Overload . . . 3
- 1.2 Online social-networks . . . 4
- 1.3 Problem Statement . . . 6
- 1.4 Research Question and Design . . . 7
- 1.5 Structure . . . 7

2 Literature Review 9
- 2.1 Limited Capacity of Social Dynamics . . . 9
- 2.2 Information Overload . . . 14
- 2.3 Microblogging . . . 27
- 2.4 Information Overload and Twitter . . . 31
- 2.5 Filtering Methods . . . 31
- 2.6 Lists on Twitter . . . 33
- 2.7 Conclusion . . . 35

3 Models 37
- 3.1 Choosing a model . . . 37

3.2 Use Cases of Models . . . 38
3.3 Physical Constraints to Information Processing . . . 42
3.4 Group Social Dynamics . . . 43
3.5 Information Overload Online . . . 43
3.6 Agent Model of Gonçalves *et al.* (2011) . . . 45
3.7 Model Usage . . . 48
3.8 Conclusion . . . 48

4 Research Design **50**
4.1 Quantitative . . . 50
4.2 Qualitative . . . 56
4.3 Statistical Analysis . . . 58
4.4 Technical Implementation . . . 59

5 Results and Analysis **60**
5.1 Experiment 1: Change in following rate . . . 60
5.2 Experiment 2: Effectiveness . . . 65
5.3 Survey . . . 67
5.4 Conclusion . . . 71

6 Discussion and Future Work **72**
6.1 Discussion . . . 72
6.2 Research Focus . . . 75
6.3 Future Work and Recommendations . . . 77

7 Conclusion **80**

Appendices **83**

A Technical Details **84**
A.1 General . . . 84
A.2 Experiment 1 . . . 84
A.3 Experiment 2 . . . 86
A.4 Survey . . . 86
A.5 Data Analysis . . . 86
A.6 Conclusion . . . 86

B Twitter's Terms of Service **88**

C Initial reasons for using lists **90**

List of References **92**

List of Figures

2.1 Diversity and Bandwidth Trade-off (Aral and Van Alstyne, 2011) . 10
2.2 Resource-based Model of Sustainable Social Structures (Butler, 2001) 13
2.3 Conceptual framework for research on information overload (Eppler and Mengis, 2004) 19
2.4 Model of Information Overload in Online Interaction Spaces (Jones *et al.*, 2004) 21
2.5 Framework of information overload on online social-networks (Koroleva *et al.*, 2010) 25

3.1 Social strength as connections increase (Gonçalves *et al.*, 2011) . . . 46
3.2 Simulated Agent Model from Gonçalves *et al.* (2011) 47

5.1 List Usage over time against Friends Count (log scale) 61
5.2 Friends Count change of Twitter Users over time (per day). 62
5.3 Average per run across 2 weeks for non-list users. 64
5.4 Users collected per run. 65
5.5 Distribution of friends count split between lists and no lists. 66
5.6 Percentage of users who use lists above a certain threshold. 67
5.7 Distribution of size of lists. 68

List of Tables

4.1 Questions about Information Overload 57
4.2 Reasons for List Usage 57

5.1 Statistics for List usage over time. 62
5.2 Reasons why people use lists 68
5.3 Scores 69
5.4 Survey Correlations 71

Table of Abbreviations

Abbreviations

API = Application Programming Interface
CMC = Computer Mediated-Communication
IRC = Internet Relay Chat
OSN = Online Social-Network

Chapter 1

Introduction

Humans are social beings. Across the ages we have developed from nomadic tribes, settling down and forming agrarian societies, migrating to massive urbanised cities and connecting across time and space with the radio, telephone, television and the Internet. With new tools we find new ways to form the social systems we are a part of. When we figured out how to farm and tame animals, we settled down, and benefited from new economies of scale. When we discovered proper sanitation, ways to transport resources, trade and other forms of industrialization, the social systems we had grew even larger into cities. One of our defining strengths as a species is the way in which we organise ourselves. We benefit when we work together: the sum is often greater than the parts. With the tools we develop we redefine how we reap the benefits from the network effects of our society. As we come together and reorganise, we have to deal with new problems such as keeping a city sanitised, keeping people fed and figuring out how to govern increasingly larger social systems.

The economy of the 21st century is moving to one where the use of knowledge leads to the most innovation (Drucker, 1992). The rules and practices that determined success in industrial and labour-intensive economies need rewriting in a knowledge intensive world where we are highly interconnected and becoming increasingly globalized. The challenges that new technology and increasing global connectedness brings are myriad.

At the turn of the 20th century technology quickly accelerated and presented a slew of unprecedented problems. With the advent of the personal computer

and Internet, we are reorganising and participating in new and fascinating social systems. We build open source software, such as Linux, for the benefit of everyone, build free, digital encyclopaedias such as Wikipedia, read news of earthquakes faster than the earthquake itself on Twitter, share culture across nations and explore virtual worlds such as World of Warcraft. We are digitising: putting systems once analogue and ephemeral (such as oral histories) to bits and bytes and extracting value from it. There is so much novel information waiting for us each day to use, and it is not slowing down.

Information creation is increasing exponentially (Eppler and Mengis, 2004), and with it new problems are introduced. How do we continue to benefit from being able to be instantly connected to 2,4 billion people (Nielsen, 2013)? All these people are contributing to this problem every day by posting updates, reviewing places, checking-in, etc. Dealing with all the information is an unprecedented problem. In order to contribute to the knowledge economy finding the right information (and people) and using it properly is becoming increasingly difficult due to this. A society that can effectively filter the deluge of information, make it novel and useful, is one that will have a higher chance to succeed in the current age.

There are 2 broad parts to this puzzle: the social system and information overload. As humanity moves ahead to functioning in increasingly larger systems, it is important to understand the limitations of it. Information does not exist in a vacuum: it is dependent on an entity intending to encode a message and sending it to another entity (Shannon and Weaver, 1948). In the knowledge economy intellectual capital in the form of people and what they can do is what contributes the most to growth (Drucker, 1992). By first understanding how individuals function in social systems through the use of these new tools, it can be determined why information overload occurs and how to mitigate it. For example, Butler (2001) looked at online social structures to determine the interplay between the size of the community, the amount of communication and ultimate sustainability of the system. The social systems we design in part contribute to what information can be shared (Jones *et al.*, 2004). If an individual has to contend with information coming from a larger social system (being connected to a lot of people), it inevitably increases the amount of in-

formation they have to deal with, whether it is the actual information being passed or understanding the context of all the actors in the system. It results in information overload.

1.1 Information Overload

When an individual's information processing requirements exceeds their information processing capacity, information overload occurs (Eppler and Mengis, 2004). This can cause stress, confusion, pressure, anxiety and low motivation to individuals (Haksever and Fisher, 1996). Information overload has been studied in various fields: from accounting (Schick *et al.*, 1990), organisation science (Tushman and Nadler, 1978), marketing (Jacoby, 1984) and on the internet by researchers Jones *et al.* (2004) and Butler (2001).

When information overload occurs, it changes subsequent interaction dynamics (Jones *et al.*, 2004). Individuals adjust their behaviours, whether it be through various personal changes such as improvements of information management (Bawden, 2001), or through more drastic actions such as removing themselves from the cause (Koroleva *et al.*, 2010). An example of research into information overload was done by Jones *et al.* (2004) that looked specifically at online communication and how the design of the tool affects our space in it and the information that has to be dealt with.

The model assumes that '...(1) an individual must invest more cognitive resources to process large, complex group computer mediated communication (CMC) than small-scale group CMC; (2) decisions made by individuals to employ various information-overload coping strategies will affect the dynamic of virtual public discourse; and (3) the nature of cognitive resources required to process group CMC relates to the CMC technology in question.' Jones *et al.* (2004) tested this model with usenet forums and online chatrooms (Jones *et al.*, 2008). In the USENET study, at the maximum average communication load, the complexity of messages decreased as the amount of communication increased; simpler messages generated more responses than complex messages; and users ended active participation (Jones *et al.*, 2004). In the online chatroom (IRC) study, it was found that as more users joined, the messages posted

per individual decreased; and that the number of posts in a chat channel will increase up to a certain point, after which it remained constant (Jones *et al.*, 2008). Different CMC tools enable certain kinds of communication, while it might restrict others.

This increasingly globalised world is being connected through online social-networking tools. It is here where through data analysis tools, we can study how people interact in increasingly large and highly connected systems. It is an ideal way to study how to mitigate information overload and what tools contribute to solving it. The CMC tool that is going to be looked at is thus online social-networks (or more specifically microblogging services).

1.2 Online social-networks

Online social-networks provide individuals with a means to stay in touch across the world. Richter and Koch (2008) describes six basic functionalities that online social-networks provide. They are identity management, expert finding, context awareness, contact management, network awareness and exchange. These functionalities in turn provide processes such as editing of profile data, exchanging of views, finding of other users, discovering common context and being able to cross link with others. Boyd and Ellison (2007) define *online social-networks* (OSN) as services that allows individuals to a construct a profile within a bounded system, articulate a list of users with whom they share a connection with and view their lists and other's lists in this system. An integral part of an OSN is its capability to read updates from the people a person is connected to (Richter and Koch, 2008). This is called a stream or a feed. Various OSNs do this in different ways, depending on how the connections are established. An OSN allows for interaction to occur in various places. On Facebook, for example, interaction can occur by posting on the walls of individuals, interacting in groups, pages and chat for example. The place where interaction occurs the most on Facebook is the stream (Sandberg, 2009). As connections increase, the amount of information present in this stream increases. At a certain point, the amount of updates an individual will see, will be more than they can handle: the information processing requirements will exceed the information processing capacity (Koroleva *et al.*, 2010).

This is cause for concern. Online social-networks have become an integral part of the new global society, with Facebook alone having more than 1 billion users (Facebook, 2012). Since it rose to prominence, users have been connecting to people from their past, but also to people they are currently meeting. Going forward, new people are constantly being added to existing social graphs[1]. If a user moves to a new city, they will meet new people and subsequently would want to add them as friends on Facebook for example. An individual's online social graph thus expands as they connect with new people, resulting in more updates: an increase in the information processing requirements on that online social-network. Koroleva *et al.* (2010) studied information overload on Facebook and derived a conceptual model based on a framework for information overload by Eppler and Mengis (2004).

Microblogging is a subset of online social-networks that focuses on posting small updates ('micro blog posts'). Its feature set is not as broad as other online social-networks. Twitter is a prominent example of a microblogging service. Its role as an information conduit facilitates rapid diffusion of information. It contributes, for example, to giving voices to people in the Arab spring (Stepanova, 2011), and it allows people from across the world to take part in a global and public conversation (Dorsey, 2012). Twitter is an important part of the increasingly global society. Its default public nature and open API (to access the data), makes it a perfect tool to study how information overload affects users.

Some research do not explicitly study information overload on Twitter, but have studies that indicate indirectly that it is occurring. For example: Borgs *et al.* (2010) takes a game-theory approach to discover what utility users gain from following various celebrities. They discover that as a user tweets more over a period of time, the rate of unfollowing increases. Gonçalves *et al.* (2011) studied how connections and interactions get affected as a user becomes more connected on Twitter. They found that interactions on Twitter (through @-mentions) per connected user increase as users become more connected, but

[1]In this context, a social graph is a graph that depicts the personal relations between people.

decreases when it reaches a certain point (150-200 connections). Grabowicz *et al.* (2012) found that in smaller 'clusters' on Twitter (less than 150 people), users interact more (through @-mentions) than retweets. When the groups gets too big, replies and retweets are equally common. Comarela *et al.* (2012) found the same results when they measured what factors affect response rates on Twitter. Retweets are easier to do and require less information requirements than replying, and is thus more abundant. Comarela *et al.* (2012) proves information overload by arguing that users reply to tweets that are far back in the stream (10% of retweets occur 800 positions back in the timeline). Novelty and interesting updates are a dominant factor in whether users feel information overload (Koroleva *et al.*, 2010). Some users on Twitter feel that only 36% of tweets are worth reading (André *et al.*, 2012). No overarching study has however been done to examine whether information overload exists on Twitter. A more detailed literature review will canvas these ideas laster.

1.3 Problem Statement

Because of the increasing ubiquity of the internet and its continued impact on society, it is important to determine how to mitigate and deal with information overload. Online social-networks are at the forefront of this. More and more users are becoming connected through them: increasing the amount of updates each user will have to deal with. For example, what was once 120 average friends in 2009 on Facebook (Marlow, 2009) was 190 in November 2011 (Ugander *et al.*, 2011). If the information processing capacity of users do not increase, information overload will happen because of an increase in the information processing requirements (Jones *et al.*, 2004).

Online social-networks provide ways to deal with information overload: either through algorithmic filtering or through user-curated filtering (Grineva and Grinev, 2012). Facebook, for example, uses EdgeRank; an algorithmic machine-based filtering method to show relevant updates to users. Instead of showing every action a user takes to all connected users, Facebook determines what is relevant for a user. Their approach is to not have information overload happen. Twitter on the other hand, does not filter any updates. It is the onus of the user to filter their updates. Users can do this through the use of lists;

filtering users into various topics (Yamaguchi *et al.*, 2011).

Information management (through network control) is a coping strategy for information overload (Koroleva *et al.*, 2010; Eppler and Mengis, 2004). By using Twitter lists, users are filtering their stream into topics. By doing this, users are increasing the value of the information to that what they want to read, reducing the information processing requirements (fewer irrelevant updates to read). Context is an important indicator of whether information overload will occur (Eppler and Mengis, 2004). Compared to algorithmic filtering, user-curated filtering requires effort on the part of the user. The act of filtering a stream into topics, increases the value of each stream, because it creates a better context in which that information is processed (Butler, 2001). By understanding how user-curated filtering helps to mitigate the effects of information overload, the research aims to further the understanding of how to solve the problems that is faced in the knowledge economy. Systems and tools can then be better designed so that individuals and society as a whole can benefit from all the novel information being generated.

1.4 Research Question and Design

The research focuses on measuring the effectiveness of user-curated filtering (through Twitter lists) as a coping/countermeasure strategy for information overload on micro-blogging services. Detailed design and methodology is explained in a later chapter.

To measure the effectiveness, thorough research is done to look at users who use lists and those who do not and whether their usage significantly differs. This includes a quantitative and qualitative aspect.

1.5 Structure

In order to understand how these groups of users differ in their interaction on Twitter (and their usage of lists), an established model has to be used. The second chapter covers a broad spectrum of literature focusing on social systems and information overload; moving from its occurrence in various fields, to cases

where it has been studied online. Based on this literature, chapter 3 details the model that is used to determine information overload on Twitter. The current studies on information overload on Twitter is then compared to this model. Chapter 4 consists of explaining the research design and methodology: taking the model explained in chapter 3 and creating a version that can be used to specifically test the information overload threshold on Twitter between users who use lists and those that do not. Chapter 5 consists of detailing the results and analysis of it: figuring out what it means. Chapter 6 consists of the discussion of the results (whether the research questions have been answered) and recommend further work in the field. The final chapter concludes the research and ends off the narrative and the importance of the research.

Chapter 2

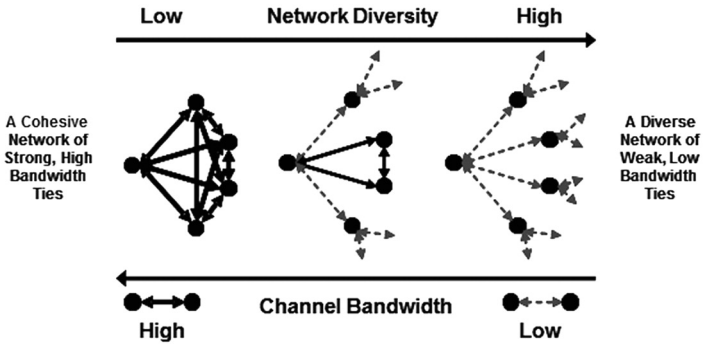
Literature Review

Two core sections of literature are reviewed to understand and eventually mitigate information overload. Understanding the social systems where it occurs is important, as well as understanding what happens when too much information is present that can not be processed by individuals. Several fields define information in different ways. In Shannon and Weaver (1948), information is technical and means something that reduces uncertainty or entropy in the receiver of the information. In systems theory it is any type of pattern that influences the formation or transformation of other patterns (Casagrande, 1999). What it ultimately comes down to is that information does not exist in a vacuum. There is the intent of encoding something so that another entity can decode and (hopefully) understand it. Due to communication having inherent entropy (Shannon and Weaver, 1948), it correlates to our ability to organise in groups (Miritello *et al.*, 2013*a*; Grabowicz *et al.*, 2012).

2.1 Limited Capacity of Social Dynamics

The dynamic push-and-pull of social systems can provide insight into how CMC tools can be designed to maximise novel information consumption. The studying of social systems are broad and can range from research into evolutionary anthropology (Dunbar, 1998) to models of online social structures (Butler, 2001). As mentioned in the introduction, we are using technologies and tools like the internet to connect on an unprecedented scale, and in order to benefit from this new global society, the limits of our capacity to benefit needs to be understood. This can be very broad, thus literature will be

Figure 2.1: Diversity and Bandwidth Trade-off (Aral and Van Alstyne, 2011)



reviewed that give insight into how social dynamics can limit novel communication (and information consumption), how we interact online and how this could be used to create tools to potentially mitigate information overload for individuals and groups.

An example is the study by Miritello *et al.* (2013*a*) on limited communication capacity. They studied an anonymized mobile phone operator network comprising of 20 million users making 700 million connections over a period of 19 months. The first relevant finding from the research was that when individuals formed ties to disparate parts of a network, at the cost of reducing the amount of communication events, they had disadvantaged access to novel information.

This is the diversity to bandwidth trade-off that exists in social networks. This was also studied by Aral and Van Alstyne (2011) who investigated e-mail networks and discovered that having a more diverse network results in lower access to novel information. It is based on the idea that the same level of communication cannot be maintained with many ties. A small network will have larger bandwidth between each node. Although a person with a disparate network has potential access to more novel information than a small, highly connected network, that novel information is not passed on, because of lower communication activity. Figure 2.1 shows this visually.

In terms of network structure, there is thus a position where information consumption of novel information is at its highest: not too dense, and not to sparse/disparate. Over the time period Miritello *et al.* (2013*a*) also found

that there is a conservation principle in communication. Any deactivated ties are replaced with new ties over a 7-month time period. Their communication capacity remains constant. There was also a correlation found between the number of connections being created against the communication capacity. Individual communication is characterized in terms of communication capacity and communication activity (Miritello *et al.*, 2013*a*). In other words, individuals exhibit different thresholds for communication as well as at what rate they activate and deactivate new ties.

Although this research does not delve to deeply into areas such as cognitive psychology, it is worth mentioning key research in the area that fits explanations for limited capacity of social dynamics. Of note, is research done by Dunbar (1998). He is an anthropologist and and evolutionary psychologist, who studied the correlation between the size of the neocortex (in the brain) of primates and the size of the groups in which the primates organise. He found that due to our brain's ability to process information (the neocortex), humans are limited in the amount of stable social relationships they can handle. This is because an individual needs to know who each person is and how each person relates to the other people in the group. The number that came up, was 150, now known as Dunbar's Number. A validation of Dunbar (1998) has been studied on online tools as well (Gonçalves *et al.*, 2011).

Some technologies such as e-mail and the telephone allow people to connect with each other, but it relies on established networks that as Miritello *et al.* (2013*a*) and Aral and Van Alstyne (2011) has shown, are subject to natural tie activation and deactivation. People meet new people and grow apart from others. However, newer CMC tools such as online social-networks rely on a network to deliver information (Sandberg, 2009). Natural tie activation and deactivation does not exist in the same form. Online communities are thus highly dependent on the design of the tool.

2.1.1 Resource-based model of online social structures by Butler (2001)

Butler (2001) studied how membership size and communication activity af-

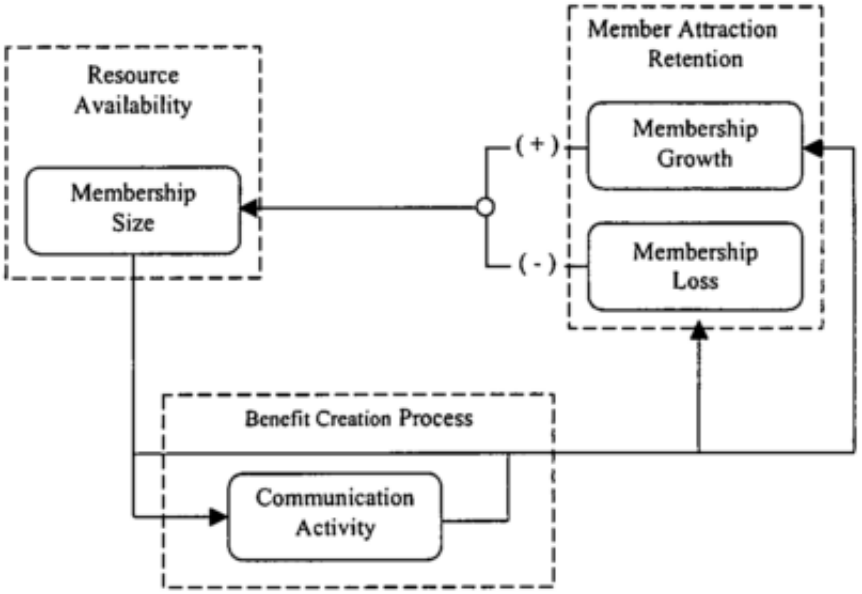
fects the sustainability of online social structures. He presents a resource-based model that is affected by resource availability, benefit provision and the social structure's ability to attract and retain members. In order for a social structure to be sustainable, it must maintain access to a pool of resources that can be turned into valued benefits to its participants: the benefits should outweigh the cost of membership. If the primary resource are members, increasing membership size, increases the available resources. However, size can have an adverse effect on converting these available resources into valued benefits for the participants. As an example in traditional social structures (face-to-face conversation), individuals end up having less time to talk and participate. Another factor of large memberships is that members lose track of who in the group can provide the resources they want. The result of the difficulty of processing the available resources into valued benefits to a social structure's members, means that a social structure has difficulty attracting and retaining members. The intra-group complexity that increases with membership size can hopefully be reduced when using computer-mediated communication tools.

Communication is essential for a social structure to turn the available resources ('other people') into benefits for individuals (Butler, 2001). Without communication activity, the resources will remain dormant. It is expected (as with membership size) that when communication activity increase, more resources are turned into benefits for its members. However, as an example, communication activity that is deemed beneficial to one member may be seen as noise to the other.

When users partake in such a social structure, they have to expend time and energy. Higher communication activity requires more time and energy to extract the benefits. In other words: higher communication is only beneficial to an individual if the benefits provided by it outweigh the costs of being exposed to it.

As can be seen in Figure 2.2, it explains the interplay between resource availability, benefit provision and member attraction and retention. An increase in membership increase resource availability, which coupled with communication activity creates benefits to the members. Both lead to member attraction and

Figure 2.2: Resource-based Model of Sustainable Social Structures (Butler, 2001)



retention. Social structures are thus faced with the fundamental problem of balancing the positive and the negative consequences of size and communication activity (Butler, 2001).

Butler (2001) took the resource based sustainability model and studied an e-mail list community. The interplay between membership size, communication activity and member attraction and retention was as expected. Size is positively associated with an increase in communication and variation which negatively impacts sustainability. 'The negative impact of more varied communication activity on the retention of members is significantly greater than the positive impact on member attraction' (Butler, 2001). The increasing amount of communication as well as the increasing variedness of it, leads to the members of the group being overloaded. It is simply too much to deal with.

The literature show there are natural limits to how people communicate (offline and online). Miritello *et al.* (2013*a*) showed that the concept of communication capacity exists. In Butler (2001), too much communication from too many people stifle the processing of information. And as Dunbar (1998) mentions, our capacity to interact in social systems are dependent on our ability to process

information. An underlying theme amongst the limitations of social dynamics is the capability of individuals to process information regarding larger groups. The more people a person is connected to, the more social relationships they have to take into consideration (Dunbar, 1998; Romero *et al.*, 2011) as well as simply deal with increase in information (more communication) (Butler, 2001; Miritello *et al.*, 2013*a*).

There are two parts to this: simply maintaining information about the social ties (few people can manage more than a certain amount), as well as trying to maintain increased amounts of communication. If the group size was stable, but communication increased, information overload will eventually occur. There are limits to the amount of information the brain can reasonably process. Information overload occurs when the information processing requirements exceed the information processing capacity (Tushman and Nadler, 1978) (Galbraith, 1974).

2.2 Information Overload

Information overload has been studied in various disciplines: from organisational context (Eppler and Mengis, 2004), to cognitive psychology (Gonçalves *et al.*, 2011) and online (Jones *et al.*, 2004).

An important paper is 'The Magic Number Seven, Plus or Minus Two: Limits on Our Capacity for Processing Information' (Miller, 1956). Miller tested whether an increase in input information changed an individual's capacity to make absolute judgements. In other words, whether the increase of information result in erroneous judgements. He reviewed various studies relating absolute judgements of unidimensional stimuli. An example is study by Pollack (1953) on individuals' capability to identify tones. With only 2 or 3 tones, listeners never confused the tones. At 4 it was rare, but 5 or more tones confusions became more frequent. For multi-dimensional stimuli on the other hand, additional independent variables increase our capability to make absolute judgements, but at a decreasing rate for each variable (Miller, 1956). To quote Miller (1956): 'The point seems to be that, as we add more variables to the display, we increase the total capacity, but we decrease the accuracy of any

particular variable. In other words, we can make relatively crude judgements of several things simultaneously.'

Driver and Streufert (1969) expanded upon this with their own research on information processing, looking more specifically at what impact the connectedness of the input information (integrative complexity) have on the output. The more complex the input, the complexity observable in perception and decision increases to an optimal point, and starts to decline afterwards.

In short, the research of Miller (1956) and Driver and Streufert (1969) show that an increase in information leads to more optimal actions. The novelty of any additional information to making optimal decisions has diminishing returns. Thus at a certain point, any additional information will not lead to more optimal actions, but actually worsen the situation.

Miller (1956) and Driver and Streufert (1969) formed the basis of subsequent research into information overload in various disciplines. Various aspects of information overload have since been studied: from the causes, to the symptoms and coping strategies in various fields.

2.2.1 Information Overload in Management

Eppler and Mengis (2004) reviewed literature from a broad variety of disciplines including organization science, accounting, marketing, MIS and related disciplines to analyse the definitions, situations, causes, effects and countermeasures of information overload. Eppler and Mengis (2004) provides an extensive look into the various aspects of information overload found within an organisation.

In the realm of management, information overload is more concerned with how the individual's performance (adequate decision making) varies with an increase in the amount of information. It was found that performance increases with the increase of information up to a certain point, after which performance declines (Chewning *et al.*, 1990). In marketing as well, information overload occurs when the volume of information exceeds the individual's processing capacity to make adequate product buying decisions. To simplify, Tushman and

Nadler (1978) and Galbraith (1974), defines information overload as the point where information processing requirements are greater than the individual's information processing capacity. Other definitions include more components, including time and the characteristics of the information (such as novelty, ambiguity, uncertainty, intensity or complexity) (Schneider, 1987). Beyond these objective conceptualizations, there are definitions that define information overload as subjective: feelings of stress, confusion, pressure, anxiety and low motivation signal information overload when dealing with surplus information as used by Haksever and Fisher (1996) and O'Reilly III (1980). Information overload research is grouped into the following 3 clusters: causes, symptoms and countermeasures (Eppler and Mengis, 2004).

2.2.2 Causes

The causes of information overload studied is grouped into 5 categories by Eppler and Mengis (2004):

- personal factors
- information characteristics
- task and process parameters
- organizational design
- information technology

Information overload emerges as as a mix of these factors. They ultimately influence either the information processing requirements or the information processing capacity (Eppler and Mengis, 2004).

In organizational design, for example, a move to a more decentralized structure can lead to greater information processing requirements, because it creates the need for more intensive communication and coordination (Tushman and Nadler, 1978). The characteristics (level of uncertainty, level of ambiguity, novelty, complexity and intensity) of the information is itself an important factor (Schneider, 1987). In other words, improving the quality of the information can have a positive effect on an individual to process the information (Simpson and Prusak, 1995). Personal factors also contribute to information

overload such as: personal skills (Owen, 1992), level of experience (Swain and Haka, 2000) and motivation (Muller, 1984). Complex tasks or processes increase the information processing requirements (Eppler and Mengis, 2004). If it is not based on reoccurring routines, it adds to the information processing requirements (Schick *et al.*, 1990). Information technology makes it easier in certain ways to process information, such as using e-mail as an asynchronous form of communication. However this also increases the possibility that useless information might be used in these new channels (Edmunds and Morris, 2000). In other words, information can increase information processing capability, but also increase the information processing requirements (Eppler and Mengis, 2004).

2.2.3 Symptoms

Eppler and Mengis (2004) groups the symptoms of information overload into the following categories:

- limited information search and retrieval strategies
- arbitrary information analysis and organization
- suboptimal decisions
- strenuous personal situation

When information processing requirements exceed the information processing capabilities of the individual, they become less effective at identifying relevant information (Jacoby, 1984), selective about the information they process (Herbig and Kramer, 1994), less capable of identifying relationships between details and overall perspective (Schneider, 1987) and requires more time reach a decision (Jacoby, 1984) which might en up being suboptimal (Malhotra, 1982).

2.2.4 Countermeasures

In terms of an organization, Eppler and Mengis (2004) groups countermeasures into the same categories that cause information overload:

- personal factors

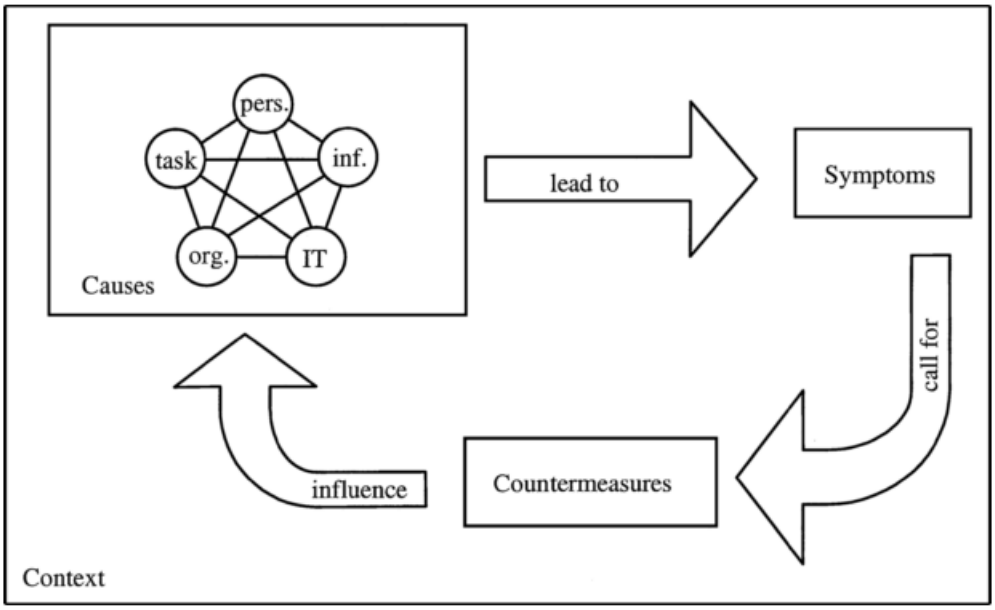
- information characteristics
- task and process parameters
- organizational design
- information technology and application

For information itself, its value is increased if delivered in the most convenient way and format (Simpson and Prusak, 1995). Countermeasures for personal factors involves training users to augment their information literacy (Schick *et al.*, 1990) and working on personal time management skills and techniques (Edmunds and Morris, 2000). Galbraith (1974) proposes various ways that organisations can change their design in order to decrease information overload: such as the creation of lateral relationships, coordination by goal setting, hierarchy and rules (creates less uncertainty), creation of self-contained tasks and creating slack resources. However, studies have also shown that collaborative and interdisciplinary work can result in increasing information overload rather than decreasing it (Wilson, 1996). Standardization of task and processes is an effective countermeasure against information overload (Schneider, 1987). Information technology systems that make it easier to prioritise information (Schick *et al.*, 1990), provide quality filters (Edmunds and Morris, 2000) and decision support systems (Cook, 1993) are ways to countermeasure information overload.

Based on Eppler and Mengis (2004), a framework was proposed for structuring research on information overload within an organization. In short: causes lead to symptoms of information overload which calls for countermeasures that subsequently influences the original causes. The framework can be seen in figure 2.3.

This study proposes three testable models related to an employee in an organization that can be used to test the causes, symptoms and countermeasures of information overload. By using the different categories, employees can be surveyed against each category. The independent variables in for symptoms and countermeasures are those that already occur, while the dependent variable is the occurrence of information overload for those individuals.

Figure 2.3: Conceptual framework for research on information overload (Eppler and Mengis, 2004)



In the reviewof literature of information overload in the business domain by Eppler and Mengis (2004), it was found that a lot of the research are based on experiments, with some studies using surveys, qualitative interviews, document analysis and formal modelling methods. They suggest that research methods (other than experiments) be used in order to triangulate the findings: such as ethnographies, action research, case studies and longitudinal studies. This will result in more informed hypotheses and refined experiments.

2.2.5 Information Overload on the World Wide Web

The literature review of Eppler and Mengis (2004) gives a great background on information overload that have been found in organizations and related disciplines, however it does not touch on information overload occurring on the web. It was written in the early 2000's and only published in 2004. It touches on e-mail, but looks at it in context of an organization. A lot has changed in information technology since 2004 (the popularisation of online social-networks and smartphones as an example) that contribute to new problems in information overload.

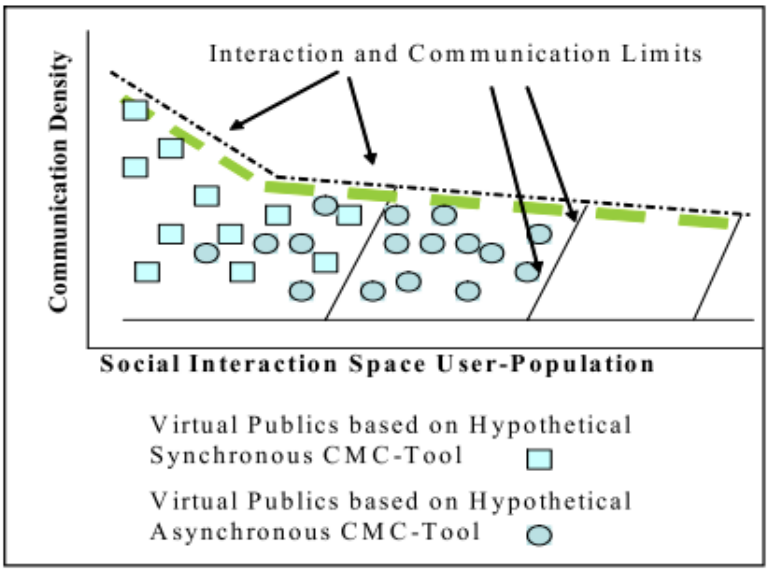
2.2.6 A model of Information Overload in Online Interaction Spaces (Jones *et al.*, 2004)

Jones *et al.* (2004) did a study on information overload and the message dynamics of online interaction spaces. They propose a model and explored that model empirically. The model focuses on analysing the impact that individual-focused information overload coping strategies have on public, open, online group discourse. They use the term, 'virtual publics', to define 'delineated, computer-mediated spaces, that enable a potentially wide range of individuals to attend and contribute to a shared set of computer-mediated interpersonal interactions'. Each computer-mediated communication (CMC) tool enables a certain range of social interactions. Using appropriate measures of these interactions, it can shed light to what extent these virtual publics enable and constrain user interactions. For example, some constraints only become apparent when they reach critical mass or when information overload starts to occur.

Jones *et al.* (2004) posits that sustainable interaction dynamics in virtual publics are constrained by information overload. Existing patterns of communication will change once information overload starts to occur. Countermeasures to information overload are grouped into two options, according to Jones *et al.* (2004). The user can either end participation or change their behaviour and enact countermeasures to deal with the information overload. The countermeasures include: increasing effort, learning new information management techniques, failing to respond or attend to some messages, producing simpler responses and storing inputs to respond when time permits. Users have to make more effort to reply to a chain of messages as opposed to replying to a single message (Lewis and Knowles, 1997). Analysis of individual information overload and its effect on virtual public message dynamics can be achieved by studying how behaviour changes once it occurs (Jones *et al.*, 2004). If membership increases, it is likely related to an increase in communication. However at a certain point, information overload might occur, which will impact subsequent dynamics, as individuals try to cope.

Jones *et al.* (2004) theorise that 'an individual must invest more cognitive resources to process large, complex group CMC than small-scale group CMC;

Figure 2.4: Model of Information Overload in Online Interaction Spaces (Jones *et al.*, 2004)



(2) decisions made by individuals to employ various information-overload coping strategies will affect the dynamics of virtual public discourse; and (3) the nature of the cognitive resources required to process group CMC relates to the CMC in question'. The model that Jones *et al.* (2004) proposes can be seen in figure 2.4.

In short: As the population of an online interaction space increases, so does the average maximum communication load given the variable it is measured against. More asynchronous tools/spaces allow for a higher load because users do not have to be co-present to process the information (Jones *et al.*, 2004).

2.2.6.1 Application: Information Overload on Usenet

This model has been used to explore information overload in online interaction spaces, including USENET discussion forums and chat channels. From this theory, Jones *et al.* (2004) proposed 3 research questions related to communication through online USENET discussion forums: 'How does the volume of interactive group communication relate to the complexity of message content?', 'How does the complexity of the initial postings relate to the chances of gaining a response?' and 'How does the the volume of interactive group communication relate to user participation patterns?'. They posit then that

at the point of average maximum communication load: 'as the number of interactive communications increase, there will be a decrease in the complexity of response messages', '...simple group CMC messages will be more likely to generate response than complex messages' and '...as group CMC complexity increases, there will be an increased tendency for individuals to end or reduce active participation.'

The first hypothesis was supported. Jones *et al.* (2004) found that as the newsgroup got more active, the messages that were posted got smaller on average. The second hypothesis was also supported. Simpler messages generated more responses. Finally, the third hypothesis was also supported: as communication complexity increase, individuals tend to end or reduce active participation.

2.2.6.2 Information Overload on Chat Channels

Jones *et al.* (2008) also used the information overload model introduced in Jones *et al.* (2004) to analyse information overload that happen in online chatrooms. Certain types of CMC tools allow for different types of communication. For asynchronous CMC tools such as USENET, user population can be larger before communication limits become a problem. Subsequently, synchronous tools allow for greater communication (eg receiving instant responses) because the users have to be co-present. This co-presence also means that communication occurs fairly quickly (Jones *et al.*, 2008).

The main focus was to determine what the individual information processing limit is that constrains the interaction dynamics of chat channels. The hypotheses put forward were as follows: 'As the size of the active chat channel user community grows the maximum average number of chat channel messages posted per individual (message density) will decline to minimal asymptotic level.' and 'The number of active participants of a chat channel (posters) will grow to an observable asymptotic level. Beyond that size, the number of posters co-present in an IRC channel will remain constant.'

It was found that message density declines as the number of users increase. In other words, the maximum average number of chat channel messages posted

per individual decreases. Above 30, it starts to decline substantially, hitting a limit of 220. For posters, the same trend occurs. Message density decreases as the number of posters increase up to 39.

As predicted by information processing capacity (IPC) model of Jones *et al.* (2004), the bounds on the rate of posting to synchronous CMC tools (such as IRC channels) are much lower as compared to asynchronous CMC tools such as USENET (Jones *et al.*, 2008). Up to a channel size of 14 users, the ratio of users to posters is linear. Beyond that, the ratio begins to shrink until it reaches a ratio of 20%. This occurs at the limit of 150 users. After this point, the data becomes less predictable.

This paper gives empirical evidence to the IPC model proposed by Jones *et al.* (2004). It shows how individuals dealing with information overload affect the interaction dynamics of the entire CMC tool. The research by Jones *et al.* (2004) is a departure into a more empirical approach as emphasised by Eppler and Mengis (2004). Future research directions at the end of Jones *et al.* (2004) are more large-scale analysis of virtual public discourse, and more thorough empirical examination of the impact of the various ways in which individuals respond to information overload on virtual publics.

The tools discussed so far have been based on public CMC tools. In other words, each participant partakes in the same space. Each user on a USENET forum will have to deal with the same amount of information. The same goes with a chat channel. Each user in that channel will have to deal with the information. Online social-networks, on the other hand, are CMC tools, where each user is subject to their own experience. The user's experience is largely determined by what connections the user chooses to make.

2.2.7 Information Overload on social media and online social networks

Richter and Koch (2008) describes six basic functionalities that online social-networks provide. They are identity management, expert finding, context awareness, contact management, network awareness and exchange. These

functionalities in turn provide processes such as editing of profile data, exchanging of views, finding of other users, discovering common context and being able to cross link with others.

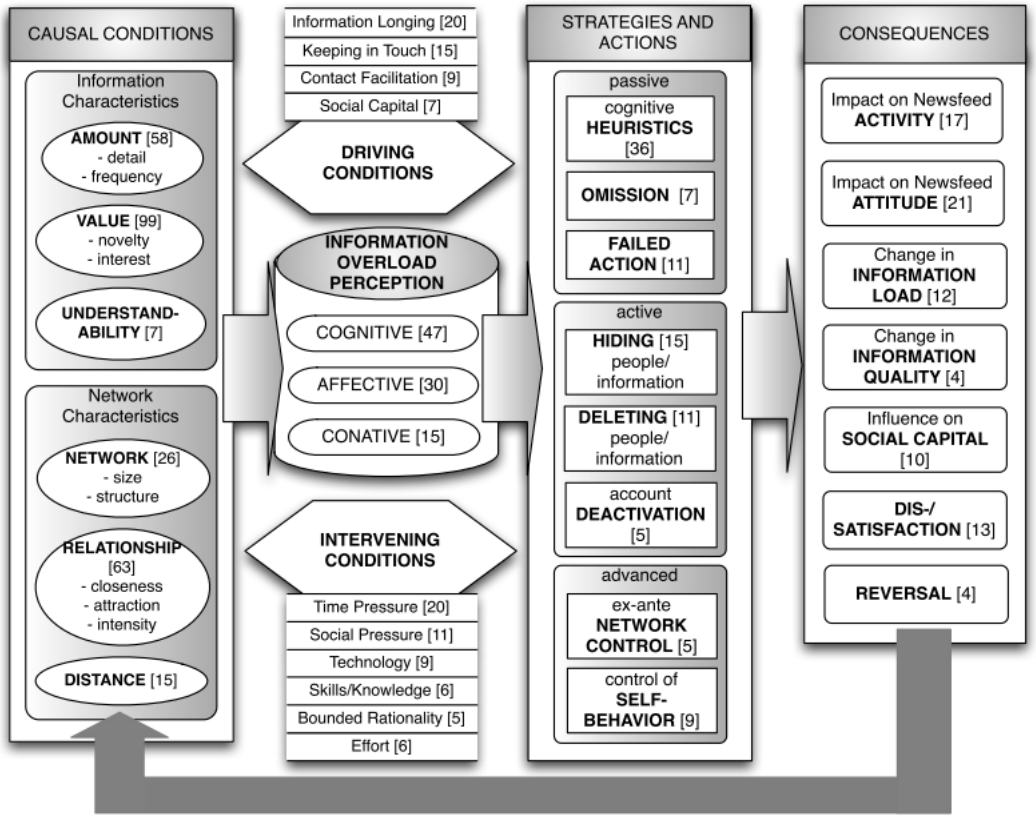
Boyd and Ellison (2007) define *online social-networks* (OSN) as services that allows individuals to a construct a profile within a bounded system, articulate a list of users with whom they share a connection with and view their lists and other's lists in this system.

An integral part of an OSN is its capability to read updates from the people a person is connected to (Richter and Koch, 2008). This is called a 'stream' or a 'feed'. Various OSNs do this in different ways, depending on how the connections are established. An OSN allows for interaction to occur in various places. On Facebook, for example, interaction can occur by posting on the walls of individuals, interacting in groups, pages and chat. The place where interaction occurs the most is the stream (Sandberg, 2009). As connections increase, the amount of updates will also increase, which increases the information processing requirements that can lead to information overload.

Koroleva *et al.* (2010) explored information overload on Facebook; finding out the main sources of information overload, how users cope with it and what the consequences are to a site like Facebook. Facebook is the world's largest online social-network with over 1 billion users (Facebook, 2012). They pioneered new social features such as the tagging of people in photos and stream communication (known as the 'news feed'). Based on the 12 in-depth interviews, Koroleva *et al.* (2010) created a conceptual model of information overload on OSN based on the information overload framework by Eppler and Mengis (2004) (see figure 2.5). The number in the brackets indicate the amount of times it was mentioned in the interviews, indicating the relative importance of each variable.

Koroleva *et al.* (2010) puts the causal conditions of information overload into two groups: information and network characteristics. Information characteristics are grouped into 3 further categories: amount, value and understandability. Too much detail, and frequency are related to the amount. Novelty

Figure 2.5: Framework of information overload on online social-networks (Koroleva *et al.*, 2010)



and interest are major determinants of the value of the information (Eppler and Mengis, 2004). Understandability is the final characteristic of information that determines information overload. If a user has to take time to understand it, and the context, it increases the information processing requirements for the users (Koroleva *et al.*, 2010).

As mentioned earlier, an online social-network creates a unique experience for each user based on whom they choose to connect to. In other words, their network also determines whether they experience information overload (Koroleva *et al.*, 2010). The characteristics are broken down into network, relationship and geographical distance. The network's size is important, but so is the friends' network size. The structure of the network is also important. Expanding networks might increase the amount of connections that a user is not fully interested in, which can increase the perception of information overload, because users aren't receiving valuable and relevant information. The

relationship characteristics are broken down into level of closeness, level of attraction and communication intensity. Finally, geographical distance has an effect, because local events or happenings aren't always relevant to a wider, global audience.

Following the causal conditions, there are also intervening conditions such as time pressure ('On a hectic day I wouldn't follow the xyz I'm not really interested in... But when I have my holidays I just go and look at people.'), social pressure, bounded rationality, effort, skills/knowledge and technology (Koroleva *et al.*, 2010). 'Driving conditions' are conditions that drive the user to 'weather' through a lot of information due to certain desires. Conditions such as information longing, keeping in touch, contact facilitation and social capital drive people to endure conditions that cause information overload.

When confronted with information overload, users apply different coping strategies and actions to overcome it. Based on their interviews they grouped the responses into the three categories: active, passive and advanced. The passive strategies are: cognitive heuristics, omission and failed action. Cognitive heuristics is a passive strategy that relies on simple persuasive cues to identify relevant information and is usually employed in situations with low motivation and limited ability to process information (Sicilia and Ruiz, 2010). These heuristics are different for each individual. Koroleva *et al.* (2010) found that the following heuristics were used: 'friend-based', 'distance-based', 'information-based', 'explicit' and 'self-centered'. The other two passive strategies are 'omission' and 'failed action'. Active strategies that are used are the hiding of information; deleting people/information; and as a final resort deactivating the account. Advanced actions and strategies involve using tools to control the network and exercising control of self-behaviour (in other words, not posting a lot, because you realise it might not be relevant to everyone in your network).

These strategies and actions lead to consequences that change the dynamics of interaction on Facebook. Newsfeed activity decreases (to lessen information overload); attitude towards the newsfeed might change (when it repeatedly does not provide relevant updates); information load might decrease; social capital might change if there are adverse effects after a user deleted someone;

and finally dissatisfaction (Koroleva *et al.*, 2010).

Users desire tools that help them deal with information with the least effort possible (Ariely, 2000). The study by Koroleva *et al.* (2010) shows that the most relevant information for users come from: close friends at different geographical locations, wider circle of friends with matching interests, and any friends who share new and important information.

This study provides a different insight into information overload, especially with regards to online social-networks. Instead of empirically measuring interactions as with Jones *et al.* (2004) and Jones *et al.* (2008), Koroleva *et al.* (2010) took a grounded theory approach to determine how information overload occurs, and what the strategies and consequences are.

2.3 Microblogging

Microblogging shares some functionality with online social-networking tools. The main difference is that it is usually first seen as a tool for information sharing, rather than a purpose purely for connecting socially (Kwak *et al.*, 2010). Microblogging services 'allow users to exchange small elements of content such as short sentences, individual images, or video links' (Kaplan and Haenlein, 2011). The name originates from the term of posting micro (small) blog (web log) posts: a condensed journal of sorts. Microblogs are also found in online social-networks (such as Facebook (Koroleva *et al.*, 2010)). There are several microblogging services such as Twitter, Sina Weibo, Tencent Weibo, App.net, Yammer (more enterprise focused), FriendFeed, Jaiku and Identica. The largest ones are Twitter, Sina Weibo and Tencent Weibo. Weibo is the Chinese word for microblogging and both weibos are popular mostly in China (Rapoza, 2011). Twitter was one of the first (if not the first) microblogging sites, and has been subject to a lot of research (as opposed to Sina Weibo and Tencent Weibo). It has also played important part in world events such as the Arab spring, as is thus an important part of the global society. Another factor for using Twitter is the API, which allows any developer to gather public content for research. The research thus focuses on Twitter as the microblog of choice.

2.3.1 Twitter

To better understand the context and space in which users communicate, a brief overview of Twitter will be given, as well as some research on how users interact on it. Twitter was founded in March 2006, and has 140 million active users as of March 2012 (Twitter, 2012). Users create updates (no longer than 140 characters). Users can then follow any user's stream that are open to the public (default). It is asynchronous, so users don't have to mutually accept a relationship to see each other's updates. In other words, a celebrity can have millions of followers, but doesn't have to follow anyone. A 'friend' on Twitter is someone whom a person follows. In other words, if you follow 300 people, you have 300 friends. It is important to remember this, as a friend is not necessarily a mutual relationship on Twitter (such as on Facebook).

In an update (called a 'tweet'), users can mention other users, eg '@simondlr', and that users will receive a notification of it. These notifications can be viewed in another feed purely based out of these mentions. If a user wants to rebroadcast someone's update, it is called a retweet. For example, when a user posts a funny cat picture, and the user that follows them finds it funny and want to reshare it with their followers, a retweet is the appropriate action. Users can also send direct messages to each other if they both follow each other. A direct message is not public, and acts as a private way to talk to people on Twitter. Communication on Twitter is restricted to 140 characters. Besides viewing updates from all the users a person follows, a user can decide to add users into lists: the name of which is defined by whatever the user deems appropriate.

One of the first and important studies about Twitter was conducted by Java *et al.* (2007). This was almost a year after Twitter started. They found the following insights. There is a correlation between the in and out degrees. Users followed by many users, also follow a lot of users. The largest portion of tweets are about daily chatter (daily routine and what people are doing) and 13% of all posts contained links.

Huberman *et al.* (2008) also did extensive research into understanding Twitter. Of tweets, 25,4% were directed at someone (through the use @-mentions). An

interesting finding related to information overload and limited capacity is that users post more as they get more followers, it eventually saturates (posts as a function of the number of followers). Posting more does not result in more followers after a certain stage.

Another interesting piece of research came from Kwak *et al.* (2010), who studied whether Twitter's users behave the same as in online social-networks. Kwak *et al.* (2010) got the similar results to Huberman *et al.* (2008). Initially the number of tweets are correlated to an increase in followers until about 100, where it remains relatively flat until 1000. After 5000 followers, the number of tweets increases by an order of magnitude. The same trend occurs when correlating numbers of tweets to friends ('people you follow'). There is thus a correlation between activity and the expansion of the network a user is connected to. That network grows, but then slows down after a point, where more activity does not grow the network as before.

Kwak *et al.* (2010) also ranked users on Twitter based on the number of followers, their PageRank (Google's manner of ranking their search engine) and by number of retweets. Ranking by followers is very similar by ranking by PageRank, but very different to the retweet network. This provides an interesting insight into influence on Twitter, as was also extensively researched by Cha *et al.* (2010). They confirmed that having a lot of followers does not equate to a larger amount of retweets. What's more important is the content. This gives again more validation to Twitter acting as an information sharing network (Kwak *et al.*, 2010).

In June 2009, Cheng and Evans (2009) did an extensive study on Twitter. They found that 85% people of Twitter tweet less than once a day. A surprising 5% of users account for 75% of all activity. The larger percentage of these are bot accounts posting content (such as posting links to news items automatically). 0.29% of people follow more than 2000 people. 92% of users follow less than a 100 people and 94% of people have less than 100 followers. They found that friends and follows remains relatively equal until 150 (for both). After that they do not accrue as many followers as the amount of people they follow. It is well worth noting, that this dataset was retrieved in 2009, when

Twitter experienced it first massive growth spurt, so some of the data might be a bit biased due to this. The majority of the users were thus in their initial stages of their usage of Twitter.

Gonçalves *et al.* (2011) studied Twitter in an interesting way. They tried to validate whether Dunbar's number applies to online discourse. The idea was to test if online discourse enables us to function in larger groups, or whether we still are limited by our brains capacity to digest and process social information. Gonçalves *et al.* (2011) looked at how how strong relationships decline over time as we become more connected on Twitter. They defined the strength as the weight of each connection to other users over all the interactions that user has. An interaction (as they define it) occurs when a user @-mentions another person. If they @-mention a person often, the weight between the two people increases (it's a stronger relationship). So, for example, an individual with strong social strength will talk a lot to a few people. Initially, the strength increase as users become more connected (they connect and talk more), but saturates between 100 and 200, where afterwards it starts decline again. This is similar to the finding by Miritello *et al.* (2013*a*) and Aral and Van Alstyne (2011). Having too many connections reduces the possible 'bandwidth' between existing nodes (or people). Gonçalves *et al.* (2011) further sets out a model to explain this, and will be looked more in depth in chapter 3, where it will be used to explain what models will be used to base the research design on.

Grabowicz *et al.* (2012) clustered users together on Twitter based on the interactions (@-mentions) between them. In groups smaller than 150 users, the abundance of mentions are higher compared to retweets. In groups larger than that, mentions occur with the same frequency than retweets. Because mentions require more effort, this suggests people converse more in smaller groups, reminiscent of studies by Gonçalves *et al.* (2011), Aral and Van Alstyne (2011), Dunbar (1998) and Miritello *et al.* (2013*a*). In smaller groups, processing communications and interactions can more easily be done due to having to keep a smaller set of relationships into account.

2.4 Information Overload and Twitter

Limited empirical research on information overload on Twitter exist. The research either presupposes that it happens (Bernstein *et al.*, 2010), develop their own definitions (Comarela *et al.*, 2012) or unintentionally reveal data that point to information overload (Grabowicz *et al.*, 2012). There has been research done that looked into the effects abundant communication on Twitter, but it has not been explicitly compared to information overload theories and models.

2.5 Filtering Methods

On Twitter, users create their feeds in two ways. The main way is simply through following other users. The other way is to put users in lists. They are not mutually exclusive. You do not need to follow someone to put them in a list. The difference with lists, however, is that additional context can be specified by the users. Examples are by 'news account' or by 'musicians'. The user takes the onus to filter the information through an additional context. It thus serves as a way to process the information in a way that fits the user.

Another way of filtering content in feeds is through algorithmic filtering. These are ways in which the site takes the onus on figuring out what is relevant to show to the user. On Facebook, for example, there are a lot more actions happening (confirming for events, liking posts, etc), and showing all of them will simply be too much for the user. They use EdgeRank to filter out content that would be relevant to each user (Applum, 2013). Each interaction on Facebook is given a score based affinity, weight and time decay. Affinity is how close a relationship is with another person (interact more and the affinity increases). Weighting is determined by what actions Facebook deem to be more important. Photos for example have a larger weight attributed to it versus a simple status update, due to it being more engaging to users. Finally, the timing is important. The longer the time after the interaction, the more it loses value.

User-curated filtering and algorithmic filtering solve the problem of processing novel information in different ways. The one gives that control to the user,

and the other does it through algorithms. It has been shown however that not many people want to bother to curate their own feeds. Lists on Twitter for example are used by only about 24% of the people (Yamaguchi *et al.*, 2011), and Mark Zuckerberg (creator of Facebook) himself said that when they introduced lists on Facebook, no one used it (Siegler, 2010). Since then, the CEO of Facebook said, they've introduced 'smart lists': list that are automatically populated with people (such as family, close friends and acquaintances (Ross, 2011)). With algorithmic filtering, a site can thus with their best effort, make the stream a lot more manageable. Instead of users leaving (Butler, 2001), they stay on the site.

Algorithmic filtering might, however, not be the best way to solve the problem of getting the most novel information. It might make information processing easier (by stripping away information), but it might not be the most novel, because the filtering algorithm needs to infer (based on behaviour) what is relevant to each user. This problem can also be exacerbated when users start living more and more inside an echo chamber, receiving information based on a positive feedback loop. For example, a user will like a photo of another users. Their 'affinity' has thus increased in the ranking algorithm. This means that the user will see more content from the other user, increasing the chances of seeing content that is relevant. Through this process, other content from other users are not shown which might have been relevant and novel. Research on Facebook by Bakshy *et al.* (2012) showed that the chance of sharing information increases drastically when it is from a weak tie. In other words, weak ties have novel information, and Facebook uses this as an example to show that Facebook is not an echo chamber. However, due to Facebook's EdgeRank algorithm, weak ties will increasingly be relegated lower scores as users interact on Facebook.

This problem does not just exist on feed-based CMC tools, but on search engines as well. It is called the 'filter bubble'. When searching on Google, the objective information is filtered based on certain features ('what you've searched in the past' or 'where you are searching from') to give more relevant results. However, as is the case again, it is still a best guess, and the control is not given for the user to curate the way they might want to. Some users

have different information processing thresholds (Jones *et al.*, 2004) and thus will trade extra time and resources for non-filtered results.

In terms of a global society, there are differing opinions and thus there also exists an ethical question to filtering. If a user chooses to be friends on Facebook with differing political affiliations, is it ethical to filter out conflicting opinions to what a user has stated? User-curated filtering thus has its place in the 21st century, as it focuses on giving control to users to objectively filter the way they want their feed to be.

2.5.1 Benefits of Human Actors

Another facet to allowing users the control to filter information is that these outliers that want to do the effort are often a powerful resource that benefits the rest of society. Yamaguchi *et al.* (2011), for example, looked at Twitter's list usage. Although only 24% of users use lists on Twitter, they put 80% of users into lists. Their desire to curate means they are adding contextual benefit to the rest of Twitter. More research on lists on Twitter will be looked at later in this chapter.

Another example is Facebook's photos feature. There are few users that tag the rest in photos. 12% of the users tag 35% of the users on Facebook (Hampton *et al.*, 2012). Their desire to curate information gives benefits to the rest. Another fantastic example is Wikipedia. In May 2013, only about 10 500 people make more than 100 edits a month on Wikipedia. This is out of the total of 1 691 208 users on Wikipedia who have made more than 10 edits over their whole lifetime (Wikimedia, 2013).

2.6 Lists on Twitter

Lists is the tool that is the focus of this research. There has not been a lot of research been done on Twitter lists. Most of the research examines ways in which lists can be used to facilitate discovery on Twitter.

On Twitter, users can consume information in 3 ways: following and through

lists (creating their own ones or choosing to subscribe to lists created by other users). In order to consume content from a user in a list, they do not have to follow them. Lists can either be private or public. Velichety and Ram (2013) studied the relationship between these 3 forms of information consumption, namely: 'following', 'membership' (own lists) and 'subscription' (other lists). Their research shows that these curators prefer more to either follow or list/subscribe users: not both. They found the following:

- Users on average follow 43% of the people they list.
- Users list 28% of people they follow.
- Users follow 11% of the members in the lists they subscribe to.
- 9% of the people a user follows are also members in the lists they subscribe to.

Yamaguchi *et al.* (2011) looked into tag-based user topic discovery using lists. They found 24% of people created a list. 80% of people, in full, were put in a list by these 24% of Twitter users. Because users can be followed and put in lists, Yamaguchi *et al.* (2011) determined if there is a correlation, and there was. Users who are followed a by a lot of users, are also in a lots of lists. The more interesting part that Yamaguchi *et al.* (2011) studied was the descriptions users used for naming their lists. These lists give insight into why people use lists. The top 10 tags are in order: list, bot, news, music, friends, friend, shop, info, famous, media. These are roughly grouped into four categories: topic tags (sports, music), property tags (famous, politician), personal tags (friends, conversation) and nonsense tags (tags such as list and bot that do not add any semantic meaning). Kim *et al.* (2010) also looked into using Twitter's list feature as a tool discover latent characteristics of users. They qualitatively determined whether these tags represent the users by questioning a sample of Twitter users about it, and this was found to be the case.

Another example of using lists as discovery is research by Wu *et al.* (2011), who also categorised users on Twitter through the use of lists. They looked at the top users on Twitter and determined in what way they are categorised through lists. Using those tags as seeds, they used snowball sampling and crawled through lists with those tags. They then split this sample into 'elite'

users and 'ordinary' users. They found that about 0.05% of Twitter garner about 50% of the attention on Twitter. They also found that almost half of the information that originates from media accounts (the category that posts the most URLs) passes to the masses on Twitter through an intermediary layer: through another set of users. These users are not of the elite users, but are more connected (follow more accounts) than their followers. This gives validation that a very small percentage of users on Twitter creates the most content, while there are another subset of information curators that create benefit for the rest of Twitter that act as information gatekeepers and disseminating information through the network (Wu *et al.*, 2011).

Greene *et al.* (2012) studied the usage of lists in news curation for topical events. Lists are employed to put in users to tweet about a certain event. However, this has to be done manually. If these feeds can be automatically generated when news events occur, it can create hyper relevant feeds to certain topical events.

2.7 Conclusion

This literature review looked into the limited capacity of social dynamics, information overload (offline and online), microblogging, Twitter and filtering methods. As Eppler and Mengis (2004) emphasises: information overload can come from a myriad of sources. What's important for the scope of this research, is looking at how to effectively design CMC tools to increase beneficial communication. As Jones *et al.* (2004) has shown with chat rooms: that chat room itself has a limit where information overload starts to occur based on the design of the CMC. The processing requirements for individuals in that chat room then exceed their capacity to process the information. It only means that they are experiencing information overload in relation to the CMC in question. As Butler (2001) emphasises: different online social structures can maintain different levels of communication.

A better understanding has been gathered to better approach the problem of determining what benefit user-curated filtering has on microblogging services. There are several models and theories that seem apt to use. However,

none of them have been used in the context of Twitter. Before setting off with the research design, it is important to figure out if these models can well be used to study the effectiveness of lists as a coping strategy for information overload. If they can measure what is intended to measure, then the design can go ahead. Because these models have not been used in the context of Twitter, this research will hopefully add academic relevance to these models.

Chapter 3

Models

There currently exists no framework purposefully designed to measure effectiveness of coping strategies on microblogging services, however there are models and theories fit to use with online CMC tools, such as Jones *et al.* (2004) and Butler (2001). Because these models are broadly defined, it begs the question whether these models can be used to measure what the research intends to determine. This is the intention of this chapter.

3.1 Choosing a model

In the literature there are various frameworks, models and theories discussed. The ones that stood out that can potentially used for the research is the research by Miller (1956), Driver and Streufert (1969), Dunbar (1998), Aral and Van Alstyne (2011), Butler (2001), Eppler and Mengis (2004), Jones *et al.* (2004) and Gonçalves *et al.* (2011).

These models can be grouped as follows:

- Physical Constraints to Information Processing: Miller (1956), Driver and Streufert (1969), Dunbar (1998) and Gonçalves *et al.* (2011).
- Group Social Dynamics: Aral and Van Alstyne (2011), Butler (2001) and Jones *et al.* (2004).
- Information Overload Online: Jones *et al.* (2004), Butler (2001) and Gonçalves *et al.* (2011).

The study by Eppler and Mengis (2004) is an overview of information overload literature, but provides a framework for qualitatively measuring the causes, symptoms and countermeasures of information overload. The different pieces are not mutually exclusive. Physical constraints to information processing lead to problems with maintaining social group dynamics. Online models of maintaining social group dynamics differ slightly to real life to various other factors, such as the dependence on the design of the CMC Jones *et al.* (2004) and lack of natural tie deactivation Miritello *et al.* (2013*a*).

3.2 Use Cases of Models

Before delving further into choosing models fit for answering the research question, it will be useful to see in what context these models have been used in further research.

3.2.1 Miller (1956)

His work has been used in various research from understanding memory (Craik and Lockhart, 1972), how the prefrontal cortex functions (Miller and Cohen, 2001), language learning (Brown and Yian, 2000), multimedia learning (Mayer, 2002), perception in chess (Chase and Simon, 1973), innovation management (Van de Ven, 1986), to testing new end-user information systems (Davis Jr, 1986).

The research in most cases were used to explain limitations of our cognitive capacity to deal and process information effectively. It's used to provide evidence and support to larger models and theories that look into inefficiencies of information processing by human brains.

3.2.2 Driver and Streufert (1969)

Their work has been used in research ranging from measuring group effectiveness (Gladstein, 1984), selecting decision making strategies (Huber and Power, 1985), understanding the brands people choose (Jacoby *et al.*, 1974), to managing product innovation (Brockman and Morgan, 2003).

In the above research, it is also used in a similar fashion to Miller (1956) to explain limitations, such as in the case of Jacoby *et al.* (1974), when an abundance in brands as options, affect consumers' decisions.

3.2.3 Dunbar (1998)

His work has been used in various research from understanding human empathy (Decety and Jackson, 2004), how humans acquire new languages (Kuhl, 2004), to studying virtual guilds in World of Warcraft and the limitations of designing online communication mediums (Gonçalves *et al.*, 2011).

The research is used to gain a deeper understanding into how (and why) we organise as social groups, as is the case with Williams *et al.* (2006), which showed that subgroups began to form when guilds became too large, supplanting some of the social elements to smaller groups.

3.2.4 Gonçalves *et al.* (2011)

This research is quite new and looks at physical limits of information processing that also exists online. The model at the end of the research explains why an increase in information on feed-based CMC inevitably leads to results that explain limits to social information processing (more in depth look at it later in the chapter).

Since then it's been used mostly to look at social features of online networks (Grabowicz *et al.*, 2012), studying collective attention on Twitter (Lehmann *et al.*, 2012), communication strategy in mobile phone networks (Miritello *et al.*, 2013*b*) and even linking it back further to Dunbar (1998): looking into the limits of social cognition in online spaces (Dunbar, 2012).

The research has mostly been used since to contrast similar results with offline interactions and explaining behaviour on Twitter.

3.2.5 Aral and Van Alstyne (2011)

The research is also new and has not been used as extensively as research by Miller (1956) or Dunbar (1998). It's been used in research studying the

dynamics of organizational networks (Ahuja *et al.*, 2012), instigating social contagion (Aral and Walker, 2011), strength of intermediary ties on Twitter (Grabowicz *et al.*, 2012), understanding collaborative user-generated content (Ransbotham *et al.*, 2012) and communication strategies when faced with limited communication capacity (Miritello *et al.*, 2013*a*).

It's primarily used to understand how groups of people function, especially in relation to the use of weak and strong ties. For example, in Aral and Walker (2011), strong ties are activated when people share products and services (as opposed to weak ties).

3.2.6 Butler (2001)

To summarise from the literature: Butler (2001) developed a model that looked into online social spaces and the relation between the amount of people and the communication activity. It was mainly focused on the trade-off that exists between the amount of people in the community and the increasing communication activity that occurs as a result of it.

Butler (2001) has been used in understanding online communities (Adamic *et al.*, 2008), social capital and knowledge contribution in electronic networks (Wasko and Faraj, 2005), linking theories of group attachment to design of online communities (Ren *et al.*, 2007), and looking at a resource-based view of information systems research (Wade and Hulland, 2004).

It has been mainly used to understand limitations of the size of communities and the reasons why people form and contribute to them.

3.2.7 Jones *et al.* (2004)

To summarise from the literature review, their model focuses on analysing large-scale changes in user and group behaviour when individuals start experiencing information overload on group CMC tools. To quote:

> '(1) an individual must invest more cognitive resources to process large, complex group CMC than small-scale group CMC; (2) decisions made by individuals to employ various information-overload

> coping strategies will affect the dynamics of virtual public discourse; and (3) the nature of the cognitive resources required to process group CMC relates to the CMC technology in question.' The model is used to understand the group-level usability of CMC tools through use of analysing on a large-scale the naturally occurring patterns of sustained interactive online communication. It also looks specifically at what coping strategies does to the discourse dynamics of the CMC, which is relevant to this research (looking into lists on Twitter).

Jones *et al.* (2004) has been used in predicting continued participation in newsgroups (Joyce and Kraut, 2006), studying information cascades (Duan *et al.*, 2009), limits to organizing in CMC like Wikipedia (Butler *et al.*, 2008), sustainability of open source projects (Oh and Jeon, 2007; Toral *et al.*, 2010), studies into the interplay between social and structural dynamics on online spaces (Ridings, 2010), social bookmarking websites (Benbunan-Fich and Koufaris, 2009), instant messaging in organizations (To *et al.*, 2008) and investor communities (Gu *et al.*, 2007).

To *et al.* (2008), for example, studied the factors affecting the adoption of instant messaging as a tool within organizations. One of the hypothesis was that critical mass would have a positive impact on the attitude of IM use within an organization. This was due to the assumption that if more people use it, the more useful it becomes. However, it was found that critical mass has a negative effect on the attitude towards it and To *et al.* (2008) attributes this result to Jones *et al.* (2004): the fact that when critical mass is reached, information overload can start to occur and thus negatively affect discourse (and in the case of To *et al.* (2008), negatively affect attitude).

The research is mostly used to give evidence of information overload on online spaces. The framework is also used in studies where message length (from posters) can be used to correlate effects of too much information/communication on the CMC.

3.2.8 Eppler and Mengis (2004)

Since Eppler and Mengis (2004) is a review of literature, it's often used as basis for other research into information overload, such as Bawden and Robinson (2009) into the effects of information overload or Stokols *et al.* (2009) into the effects of rapid change on humanity.

As mentioned, these models can be grouped roughly into 3 categories: physical constrains to information processing, group social dynamics and information overload online. These models are not mutually exclusive, as can be seen how research has been used that incorporates some elements of the other. It seems that physical constraints to information processing along with structural dynamics affect how we cope in groups. Then, group social dynamics along with the design problems related to online communication gives insight into information overload in online spaces and how to cope with it.

3.3 Physical Constraints to Information Processing

Miller (1956) and Driver and Streufert (1969) looked at what happens to our abilities when we increase the amount of information we have to deal with. The models share some similarities and points to an important aspect of information overload: with the increase of information, actions taken improve, but not linearly. In other words, at a certain point, any increase in information results in worse decision-making: not further augmenting it any way. Miller (1956) explains this when looking at what effect the increase of independent variables have on judgements. Miller (1956) and Driver and Streufert (1969) both compare an increase in information against 'transmitted information' and 'decision quality' respectively. Dunbar (1998) goes a bit deeper and asks why this is the case, and discovers the size of our neocortex is a potential reason why this happens.

While these models delve deeper into what happens when we process information, it is a bit broad. Using just these models to design research might leave out intricacies better explained with models that fit group social dynamics and

information overload online.

3.4 Group Social Dynamics

These models look more into what happens what lack of being able to processing information in due time does to social dynamics. Although Aral and Van Alstyne (2011) and Butler (2001) looked at tools used through the internet, their models are applicable to the offline world as well. This is evident as shown by the research of Miritello *et al.* (2013*a*), that looked at a telephone network and found similar results to the theory on diversity-bandwidth tradeoff by Aral and Van Alstyne (2011). Butler (2001) also mentions that 'the use of CMC infrastructures is not likely to fundamentally change the problems underlying the development of sustainable social structures.'

3.5 Information Overload Online

The models focusing on information and online discourse are those by Jones *et al.* (2004), Butler (2001) and Gonçalves *et al.* (2011). Of these, Jones *et al.* (2004) goes the furthest into information overload online, because it focuses on the dependence of the design of the CMC as an important factor.

Other research have used the model proposed by Jones *et al.* (2004) in more depth to not only study group sustainability in terms of structural dynamics, but also in terms of social dynamics (Ridings, 2010). While Jones *et al.* (2004) quantitatively measured changes in discourse, they did not account for the social dynamics of the CMC in question. Ridings (2010) found for example that a CMC designed for information sharing, can turn into a social one as users develop relationships, which in turn affect discourse dynamics. New users can come in, only see social discussions and not the information they seek and subsequently not join.

While research by Butler (2001) on online group sustainability is not specifically focused on information overload, it is relevant to the research and has been used in conjunction with Jones *et al.* (2004). As mentioned in the literature review by Koroleva *et al.* (2010) and Eppler and Mengis (2004), the

novelty of information is a large contributor to stop information overload from happening. The resource-based model Butler (2001) relies on communication activity to turn resources into benefits to users. However, too much diverse communication might turn into noise as what is useful to one user may not be deemed by the other. This is relevant, because user-curated filtering aims to filter the information into more relevant channels, increasing the novelty of the information. For example, using lists on Twitter, a user can filter people who usually talk about coffee into a list. When reading that list, the user has an expectation of content and is not inundated with other varied communication that might not be relevant to that topic.

The models proposed by Jones *et al.* (2004) and Butler (2001) both looked at online spaces where users interact in the same space. In other words, for most cases, a forum and chat room will look the same for the individuals involved in the participation. If a user posts a message in a chat room, all the other people in the chat room will see the message. There might be interactions that can happen out of the usual context, such as private messaging, but for the most part, interaction happens where most users can see it. On the other hand, online social-networks are different. Users interact largely in their own 'spaces' dependent on who they are connected to in their social graph. If a user from India (for example) is not connected to a user from Brazil, they won't interact in the same space, considering that feed-based interaction is where most interactions occur (such as the case with Facebook (Sandberg, 2009)).

On online social-networks, the community a user interacts with is largely formed by who the user chooses to connect to. Both Jones *et al.* (2004) and Butler (2001) only looked at group behaviour in online spaces where all the individuals interact in the same space. Before continuing the research, it must first be determined if these models can be used to look at information overload on online social networks. In other words, do the models that are based on groups function the same in groups that don't share the same online space?

The theory by Jones *et al.* (2004) has 3 assumptions, the important one for this case, is that an individual invests more cognitive resources to process large,

complex group CMC than small-scale CMC. This was indeed the case with forums (Jones *et al.*, 2004) and chat rooms (Jones *et al.*, 2008). When users interact in the same space, it was proven by looking at how users' behaviour change once they cannot process all the information. On Twitter with messages restricted to only 140 characters, it can, however, not be measured this way.

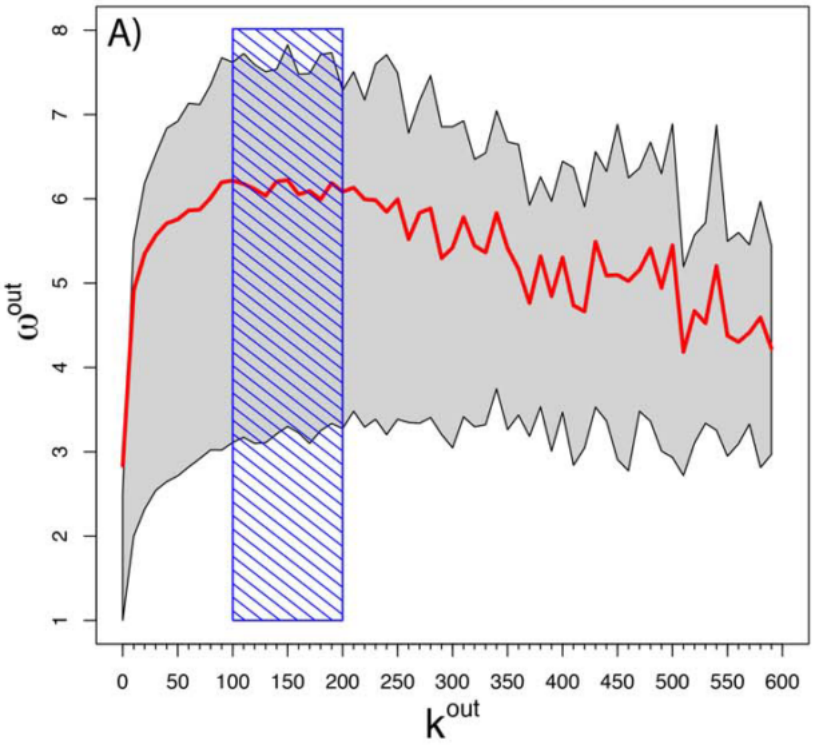
The information a user has to process on Twitter are their information feeds (main and lists), their replies (@mentions) and their direct messages. If they follow more people, this will increase. If more people follow them, it might increase as well if the user wants to reply to their mentions. An example of this could be a celebrity who follows only 50 users, but has 2 million fans that constantly reply to them. As previously mentioned, Gonçalves *et al.* (2011) studied how social interactions on Twitter change as they become connected to more users. In order to explain the behaviour of the decline of strength of social relationships as connections increase, they created a model based on information processing on Twitter.

3.6 Agent Model of Gonçalves *et al.* (2011)

Their model is as follows: Assume an agent, i, has to process replies (@mentions) when other agents reply to them. Once another agent sends them a reply, it is added to an internal queue (for the person to process). Due to physical constraints of processing information (as shown by Miller (1956) and Driver and Streufert (1969)) and the speed with which users can reply back (if they so choose), means there is a limit to how large this queue can get in a certain timeframe. When resources become finite, users have to prioritize what messages are the most important to reply. Thus for each time step, the agent chooses a random message to reply to that is in the queue proportional to the priority of the agent that sent them a reply. Popular and socially active agents are more likely to be replied to. Any messages replied to are deleted from the queue, and any incoming messages that exceed the queue are simply discarded.

This simple model is then run for a network of $N = 10^5$ nodes for timesteps $(T) = 2 * 10^4$ over 1000 runs. The network used is a directed, heavy-tailed

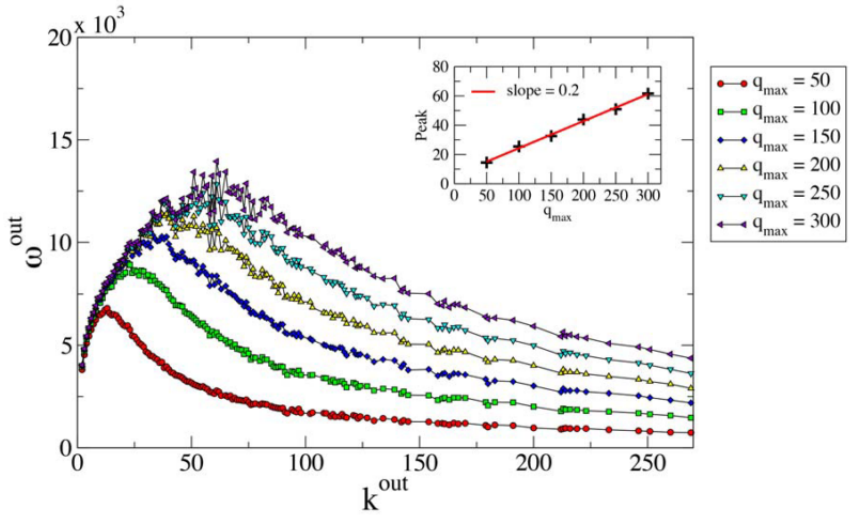
Figure 3.1: Social strength as connections increase (Gonçalves *et al.*, 2011)



network with a power-law distribution similar to what is found in Twitter. The simulated graph plots reflect what was originally measured by comparing the amount of agents a specific agent replies to (on the x-axis) against the average social strength of the agent's connections. To reiterate the average strength is defined as the average weight of all connections divided by all the amount of connections. The strength of the weight is equal to how much interaction exists between the agents. The model is consistent with what was found. In other words, as expected, stronger relationships decline. Agents reply to other agents, but it is spread out over other agents as the amount of connections increase (see this in figure 3.1 and figure 3.2). Figure 3.1 shows the actual results from Twitter. As a user replies (interacts) with more users, the average weight of connections over all connections decrease. Figure 3.2 shows the same results with the simulated model.

This simple model that looks at information consumption on feed-based online social network shows what was expected. If it is taken that a person can only process certain amounts of information over a certain time, increasing the

Figure 3.2: Simulated Agent Model from Gonçalves *et al.* (2011)



source of information flow leads to users having to prioritise their interactions. In online spaces where everyone interacts in the same space (such as a chat room), the amount of information is proportional to everyone posting in it. In online spaces such as a feed-based online social network (such as Twitter), the amount of information a user is subjected is mostly reliant on how many people they are connected to (in and out) (Sandberg, 2009). The higher the degree of a node (the more connections a node has), the more information they will receive, and the tougher it will be to eventually process all the information.

The first hypothesis by Jones *et al.* (2004) then holds true on Twitter: 'an individual must invest more cognitive resources to process large, complex group CMC than small-scale CMC'. As Gonçalves *et al.* (2011) shows, the larger the network becomes from which the agents receives information, more cognitive resources are needed to process it.

The two models that will be used for this research are that of Jones *et al.* (2004) and Butler (2001). The model on information overload by Jones *et al.* (2004) is adequate to be used for this study. It has been used in various other research since, and is more focused on information overload on online group discourse, as opposed to the effects of individual information overload and its effect on decision-making (such as Miller (1956)). As mentioned: to measure the effect of user-curated filtering as a coping strategy, not only individual

effects will be gauged, because these individual coping strategies subsequently affect group discourse. Jones *et al.* (2004) does this by measuring on a large scale the interactions that occur in the specific medium. Considering that Twitter is a largely open online social-network with a public API to easily access this data, the model is ideal.

Butler (2001) is also adequate and adds important distinctions to how user-curated filtering might be effective against information overload. As a group adds more members, resources increase, which when turned into benefits by communication provides in the case of Twitter, novel information. However, as Butler (2001) states a larger group has more diverse communication. By filtering the information as the user sees fit, it constrains some of the varied communication into benefits that the specific individual would want.

3.7 Model Usage

The important parts of the models that will be used for the research are the emphasis of Jones *et al.* (2004) on how coping strategies affect discourse and dynamics, and the emphasis of Butler (2001) on the effects of membership size ('resources') on communication activity. The research design will explain more in depth, but the basic hypothesis from the theory is that users start using lists when they follow more people on average (when they start experiencing information overload) in an attempt to continue receiving benefit from the CMC. When they start using lists there will be a difference in usage behaviour compared to users who do not use lists that indicate it's effectiveness for coping with information overload. The expected result is that they will increase their resource base of new novel information at a greater rate, because they can now filter out the varied communication into different channels with additional context. Over time, users who use lists will be able to follow and process a larger amount of users' Twitter updates.

3.8 Conclusion

Since there are no models fit for the study of information overload on feed-based CMC tools, it's important to emphasise at this stage that developing

a new model for the research is not part of the scope. Academically, there are models that have been used to study information overload on CMC tools. They have however not been used in the context of feed-based online social structures. Before setting off to design a new model, it's useful to first determine if the current models are applicable to feed-based CMC tools. By doing this, this research also adds new academic rigour to these models as they are being used in a new context.

Of the various models, some are more apt to be used for the research. Given the above explanations, the models that the experiments will be derived from and tested against will be that of Jones *et al.* (2004) and Butler (2001). They are fit to be used in the context of feed-based CMC tools. For qualitative research, the frameworks proposed by Eppler and Mengis (2004) will be used.

Chapter 4

Research Design

The purpose of the research is to test the effectiveness of user-curated filtering as a coping strategy for information overload on microblogging services. In order to do this, the usage behaviour of the people who use lists needs to be compared with users who do not use lists. The research design consists of several parts, comprising of *quantitative* and *qualitative* research. Both will be explained in more detail in the following sections.

4.1 Quantitative

The mode of Jones *et al.* (2004) is used primarily by doing large-scale data analysis. This section of the research will include gathering large amounts of data to determine whether lists are being employed as a coping strategy and whether they are effective. Before delving further into the design of the experiments, first an explanation of where the data will be coming from.

4.1.1 Twitter's API: Accessing Data

Online public CMC often have Application Programming Interfaces (API) that make it easy to access the interactions that take place on these platforms. Twitter has such a public API (Twitter, 2013*a*), in which data can be gathered and used for research. By nature of it being public, researchers can use it if they conform to the terms of service. The relevant parts of the Terms of Service can be found in the addendum B.

In short, anything a user creates is still owned by that user. However Twitter retains the right to 'use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute such Content in any and all media or distribution methods (now known or later developed).' This 'Content' can then be syndicated, broadcasted, distributed and published by Twitter or third party partners.

In order to use this 'Content', it has to be gathered through the use of the API: 'you have to use the Twitter API if you want to reproduce, modify, create derivative works, distribute, sell, transfer, publicly display, publicly perform, transmit, or otherwise use the Content or Services.' This API has further rules and restrictions on what is allowed. These are called the 'rules of the road'. 'You may use the Twitter API and Twitter Content in connection with the products or services you provide (your "Service") to search, display, analyse, retrieve, view, and submit information to or on Twitter.'

A developer is allowed free reign to content on Twitter, through the previously mentioned licensing agreement, except in the following circumstances: selling, renting, leasing, sublicensing, redistributing or syndicating content. The only content that is not allowed to be aggregated or cached is the geographical information contained in tweets.

Coming back to the terms of service: 'You are responsible for your use of the Services, for any Content you post to the Services, and for any consequences thereof. The Content you submit, post, or display will be able to be viewed by other users of the Services and through third party services and websites (go to the account settings page to control who sees your Content). You should only provide Content that you are comfortable sharing with others under these Terms.'

The publicly available data is thus free to be mined and used for analysis and research.

4.1.1.1 Rate Limiting

By using the API to access the content, applications and users are subject to rate-limits. In other words, there are restrictions to how many calls an application can make for an authenticated user for a certain time period. This is not ideal when doing large-scale data mining of Twitter's data. For example, fetching user profile information, an application is restricted to 180 calls per 15 minutes for each authenticated user. A developer is not allowed, due to the terms of service, to register multiple applications to mine the data due to the following clause restriction: You 'may not use multiple application API keys for the same use case.' Due to the way, rate-limits are applied, there are legitimate work-arounds. This bucket of 180 calls per 15 minutes only applies to an authenticated user for that specific application.

If multiple users give access, then there is a larger 'bucket' of calls that can be made to the API. Twitter gives this as an example strategy to work with rate-limiting in their FAQ: 'Scale your use of the API with the number of users you have. When using OAuth to authenticate requests with the API, the rate limit applied is specific to that user-token. This means, every user who authorizes your application to act on their behalf, has their own bucket of API requests for you to use.' (Twitter, 2013*b*)

Multiple users will thus have to authenticate with an application in order to get enough buckets to use to data-mine the relevant data from Twitter. Because the research will include a qualitative survey as well, these users will be asked to authenticate the application when completing the survey.

4.1.2 Design

When users start experiencing information overload, they use information management tools to ease the load (Eppler and Mengis, 2004). People employ these strategies at various levels, depending on how well they cope (Jones *et al.*, 2004). Once users reach these levels, involvement patterns change (Jones *et al.*, 2004). To determine whether lists are being used as a coping strategy, two things need to determined. At what point do lists start being used? At that point (and subsequent behaviour) are there any visible changes in involve-

ment patterns of Twitter users? According to Butler (2001), if the negative effects of too much communication activity is mitigated, it will bring about an increase in membership size (following more users).

If it has been determined that there is indeed a change in dynamics, then it needs to be determined in what way it affects the individual and discourse dynamics. In other words: in what way do lists help? If it is effective, it would allow users to deal with more communication activity and be able to follow more users.

4.1.3 Experiment 1: Change in Following Rate

Experiment one will gather data on at what point users start using lists. To do this, the public stream will be crawled to find a random sample of users. The public stream is a subset of all the existing tweets occurring as they happen. The users that will be gathered are thus users who are actively tweeting when the data is being collected. With each of these users, a further API call will be done to retrieve what lists they have created or have subscribed to. Using this data, users who have created only one public list (not subscribed) in the previous 2 weeks, will be stored. In other words, the first list that a user creates. If these conditions are met, the users' anonymised information will be stored along with the time at which the data was collected.

At various points during the 2 weeks, it can be determined what the average of the friend count is. Just to reiterate, a friend is classified as someone a person follows (it is not necessarily a mutual relationship). This way, the average following rate can be determined. Because the users are randomly sampled, this can be extrapolated, instead of 'watching' a select few users over a period of time (bypassing technical limitations). This way, much more data can be collected.

This data will be compared to another dataset, which consists of users who do not use lists. If the point is determined at which point users start using lists (say 600), then for a period of 2 weeks users will be monitored who do not use lists and see how their following rate changes. If Jones *et al.* (2004) holds true and lists are being used as a coping strategy, then a change in involvement

patterns will be seen between the two groups. At least 10 000 users for each group will be used. In the very least, the expected change will be that when users start using lists, they will accrue more resources to digest at a faster rate than normal, because they suddenly have tools available to split the varied communication into additional contexts. Because users can choose to list a person and not follow them, membership counts will also be taken into account.

For the second group, technical limitations are not as strict, because list information will not be needed. These users will be randomly sampled from Twitter (when it has been determined at what point users start using lists) from the public stream. The 'get users/show' endpoint allows 180 requests per 15 minutes.

4.1.4 Experiment 2: Effectiveness

Lists allows users to filter and compartmentalise the varied communication activity that occur on Twitter. According to Butler (2001), mitigating it, will result in increased benefit provision, as the communication that is happening is delivered with additional context.

4.1.4.1 'Membership Size'

Following more users invariably result in being subject to more information (Butler, 2001). If it is the case that lists help in processing information, users who use lists on average will be able to handle a higher information load than those who do not. In other words: users who use lists should on average follow more users (including having members in lists).

This is based on the assumption that if a user does not yet experience information overload, they will more easily accrue more users (increase in membership size) (Butler, 2001). Those who experience information overload do not easily follow other users due to being overwhelmed. This heuristic is evident in Butler (2001): member churn happens when communication load is too high. It's also evident on Facebook (a similar feed-based online social-network) (Koroleva *et al.*, 2010). Users will continue accruing users if they can manage the information load. The novelty of new information has diminishing returns

(Butler, 2001).

There are outliers such as users following a proportionally large amount of users that don't use information management tools (such as lists). They employ a heuristic as a coping strategy, to wilfully not read all the updates in a feed (Koroleva *et al.*, 2010), but instead just read snapshots of the feed even though they might miss out on potentially novel information. As expected these outliers account for a very small percentage and shouldn't impact the study. 81% of Twitter users follow less than 100 people (Bourne, 2010; Beevolve, 2012). Outliers included, social systems still display the expected behaviour as shown by the research of Butler (2001) and Jones *et al.* (2004).

The Twitter API endpoints that was used, was the public stream and the 'get lists/list' endpoints. The public stream contains a snapshot of tweets that happen on Twitter (in near real-time) which contain tweet data as well as the data of the user. It does not however contain information on lists usage. A separate endpoint ('lists/list') must be used to get that data. The public stream has no rate-limit on it. You can process as much as is possible. However, the 'lists/list' endpoint is rate-limited to 15 calls every 15 min per authenticated user. To be able to process larger amounts of data, more authenticated users (per app) are needed. The qualitative section, explained later, asked users to authenticate with an application so that enough tokens could be gathered to ease the processing of the quantitative data and to stay within the limits of the terms of service of Twitter's API.

At least 30 000 unique users were gathered from the public stream. Then it was determined whether they use lists or not. If they use lists in any form (created their own, or subscribed to another user's lists), it is regarded that this person is using lists. Along with list usage, the amount of people a user follows and the amount of members are in the lists they created or subscribed to were also stored. As previously stated, the idea is to determine whether there is a difference in the average of the amount of users a user can handle between those who use lists and those who do not.

Because users were sampled randomly from the public stream, the average

amounts are expected to be higher than the whole user-base of Twitter. Active users (those who tweet at least once a month) follow more people on average (Basch, 2012). This bias does not impact the study as the difference between the two groups will be measured, not whether the average amounts have significance on their own.

4.2 Qualitative

In order to attain enough user tokens to collect enough data in sufficient time, Twitter users have to voluntarily authenticate a Twitter application. This is a great opportunity to get further qualitative data from users to confirm why they use lists and whether they are experiencing information overload. Eppler and Mengis (2004) proposed a framework to question users whether they are experiencing symptoms related to information overload. This part of the research will be more exploratory as no framework currently exists to specifically test symptoms of information overload on online spaces. If an increase in users equates to eventual information overload there should hopefully be a correlation between the qualitative questions and how many users a person follows.

4.2.1 Survey

The survey have two main sections: The first part are questions asking whether the person is experiencing, or has experienced symptoms related to information overload on Twitter. Koroleva *et al.* (2010) derived a model for testing information overload on Facebook (also a feed-based social network) based information overload literature such as Eppler and Mengis (2004). As mentioned in the literature review, Koroleva *et al.* (2010) states it occurs due the characteristics of the information and the network the users finds themselves in. Along with Butler (2001) and Jones *et al.* (2004), the following questions are asked to determine whether the users are experiencing information overload. It uses a 5-point likert scale ranging from 'Strongly Disagree' to 'Strongly Agree'. The questions can be seen in Table 4.1

Table 4.1: Questions about Information Overload

I have felt in the past that I cannot keep up with all the tweets.
I often see irrelevant updates.
The content on Twitter is sometimes boring.
There are more non-interesting content on Twitter than interesting content.
Sometimes I don't understand what people are tweeting about.

Table 4.2: Reasons for List Usage

I use lists to put in my friends whose updates I want to see.
I use lists to filter my home feed into topics (news, sports, etc)
I use lists because my home stream became unmanageable
I use lists because it was difficult finding relevant information in my home feed.
I use lists to put in people who I'm not sure I want to follow yet.
I use lists to filter people by location.

The two main concepts it tests, is whether there are too many tweets occurring (due to a large network), and whether the information being tweeted is relevant. As Butler (2001) shows: the larger the group (source of information), the more varied the communication becomes and the more it lends itself to the user receiving irrelevant updates. After this, a screening question is shown asking whether the user, uses Twitter's lists feature. A few selection boxes will be shown consisting of use-cases of Twitter lists. The user ticks those that are applicable. The options are in Table 4.2

If there are other reasons, there is a box for the person to fill in other reasons that are not on the list. The network characteristics mentioned in Koroleva *et al.* (2010) was data-mined in the quantitative portion of the study. There won't be as many people doing the survey as was data-mined, but it will add reputability to the data if the users who took the survey behave consistently with the expected results: which is, at a certain point information overload occurs. Users then employ lists as a coping strategy to filter their stream into a more manageable manner.

As short preliminary evidence: from the @twimemachine Twitter account (36 000+ followers) [1], the responses were gained on why people use lists on Twitter.

[1] This account was used because it has a lot of followers and is owned by the researcher.

This was used, along with the theory, as basis for the options in the survey. The answers can be found in the addendum C.

4.3 Statistical Analysis

For both quantitative experiments, the difference in the average (mean) between the two groups (lists vs no-lists) was measured by using a t-test with a statistical significance of $p = 0.05$.

The qualitative section is a bit more expansive. When designing surveys, not only is testing for significance important but also testing for reliability and validity. The first few questions tests if people experience symptoms related to information overload on Twitter. A construct, such as testing for information overload, usually comprises a few questions which together try to measure that underlying construct. When testing for reliability, it determines whether those questions measure the same thing. This statistical test is called the Cronbach Alpha. When testing for validity, you determine whether the questions actually represent what is supposed to be measured. This is a lot more difficult, as expansive measures and correlations have to be drawn between the questions and expected results. In the literature there is no model for validity on questions to determine whether a user experiences information overload.

Eppler and Mengis (2004) gives an overview of symptoms that people experience when they are overloaded with information and proposes a test based on their framework. However, the questions are very general. As mentioned, those questions have been combined with research from Koroleva *et al.* (2010), Jones *et al.* (2004) and Butler (2001) to best try and determine whether the first 5 questions measures 'symptoms' related to information overload on Twitter. However, as this hasn't been done before, the validity of the construct is questionable and falls into the area of 'exploratory research' for the scope of this research. It aims to measure symptoms related to information overload, not whether users are definitively experiencing information overload.

As the qualitative section is more exploratory, interesting correlations (spearman correlation) could potentially be drawn between separate questions and

the amount of users a person follows. This will hopefully provide a basis for future work into this area.

4.4 Technical Implementation

The experiments (both quantitative and qualitative) were coded primarily in Python (with the Flask web framework), running from an EC2 AWS instance and saved to a MongoDB database. Results were analysed using Python and Matlab. The addendum contains a more thorough explanation of the technical implementation.

Chapter 5

Results and Analysis

The results gained from the quantitative and qualitative experiments were developed and gathered over a period of 4 months (from March 2013 to June 2013).

5.1 Experiment 1: Change in following rate

The first set of data that was gathered for this experiment was to determine at what point (number of *friends*) users start using lists. The users were gathered from the public stream and the criteria were as follows (in increasing granularity): 1) they had to use lists, 2) use only one public list and 3) that (one public) list was created over the past 2 weeks (from the point of data collection). This means we get users who have recently created their first list. Twitter's public stream is a randomized subset of users who have recently tweeted.

5.1.1 Main Findings

Over the course of a month, after some duplicate users were removed, 14 231 users' data were gathered.

Figure 5.1 plots this data, through the use of Matlab. The x-axis shows the time in seconds over a span of 2 weeks. The y-axis is the number of people a user followed (*friends*), as well as the number of people in the lists at the time of data collection. An example of such a data-point is as follows: the public

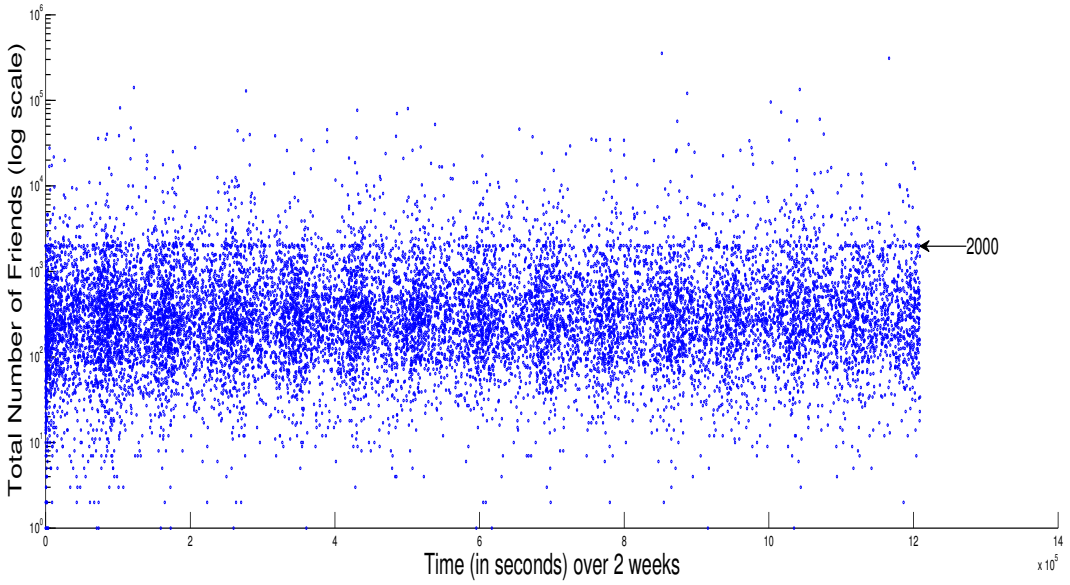


Figure 5.1: List Usage over time against Friends Count (log scale)

stream is crawled, finding a user that created a list (is the only list) and that the list's age is less than 2 weeks. At that point, the number of people that a user follows (along with number of people they have in their newly created list) is stored alongside the time difference between when the data was gathered and when that list was created. This gives us an idea how users' friends count changes over time once they start using lists.

Looking at the graph, it seems that most users start using lists above 100 friends: as is evident that after 1 day, users on average follow 622 users. Over time, the changes can bee seen in table 5.1. It is important to note that is already far higher than the number of friends the majority of Twitter users have (81% follow less than 100 (Bourne, 2010)). At the end of the 2 weeks (the whole dataset), users on average follow 882 users. This is expected as users in the period of 2 weeks start following more users (from the point that they created a list). At 2000, there is a small anomaly that is Twitter's protection mechanism against spam accounts. It is capped at that level, so users do not follow more than 2000 people, unless they themselves have close to 2000 followers. The arrow on 5.1 shows where this line is.

A linear regression was fit over the data on a per second scale. Theoretically, a non-linear regression could work, due to the fact that users, once discov-

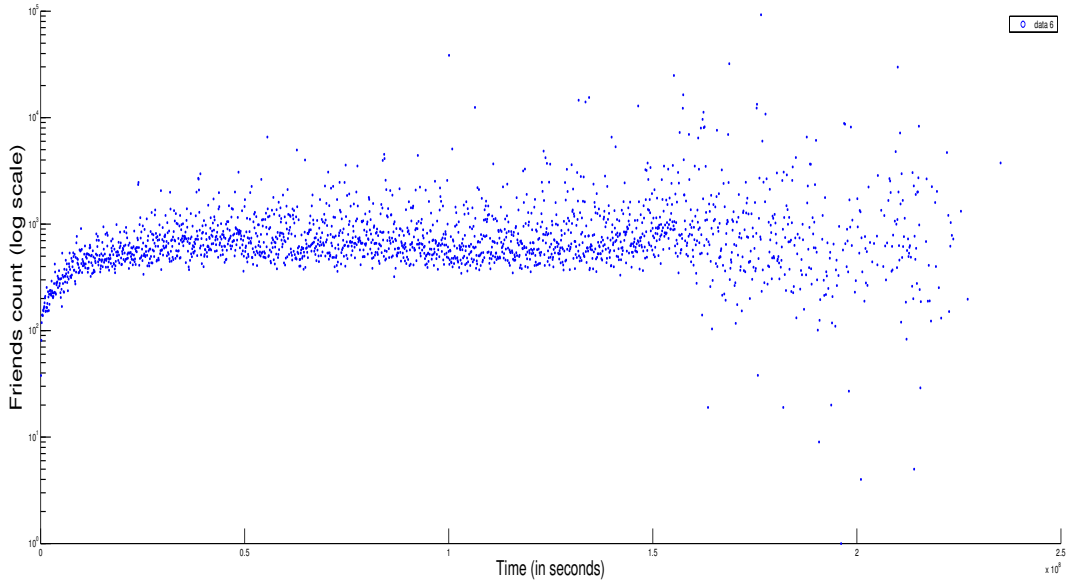


Figure 5.2: Friends Count change of Twitter Users over time (per day).

Table 5.1: Statistics for List usage over time.

	Average	Median
1 day	622	213
1 week	802	258
2 weeks	882	267

ering they have a tool to process more information, immediately follow more users, and then at a slower rate accrue new friends. However, for higher order polynomial regressions, the difference in the norm of the residuals is negligible (difference of 100 between linear and 10th degree).

The R-squared is 0.000517, which indicates substantial variance. To determine how it compares with Twitter's whole user-base (across all time-periods), 162 121 users were taken randomly from the public stream, and the time difference between when their account was created to the time of data collection was plotted in figure 5.2 (for easier display purposes, it was averaged to a day, instead of the per second scale). Overall, this variance for a linear regression, on a per second scale, is also big (R-squared of: 0.00195), and thus while the data looks random, this is what the distribution of Twitter's user graph looks like. Compared to theory of Jones *et al.* (2004), it also makes sense as users have different thresholds.

The linear regression slope for the users who start using lists is: 0.00034174. This regression is statistically significant from 0 with a p-value of 0.0066853. Although the correlation over time is really small, it is important to note that the time series is based on a second, resulting in an absolute, nominal change of 413 friends over 2 weeks (when mapped to the regression of 0.00034174 over 1209600 seconds). This is very substantial considering that predominantly, users, on average, follow less than 100 people on Twitter (Cheng and Evans, 2009).

The second part of this experiment compared the rate of change when users use lists, to users who do not use lists across the same section. A randomly sampled group of 20 000 users were gathered from the public stream across 2 days that do not use lists, and follow between 600 and 1800 people. This is around the average of the previous group.

Over the course of 2 weeks, their following count was measured about roughly every hour. 3 821 780 datapoints were gathered over the 2 weeks. Each datapoint is the following count of one of the users at a certain time period. Duplicate users were removed from the dataset to result in 18 543 unique users, with a total of 3 625 378 datapoints being used. For display purposes, instead of showing all the datapoints, each run was averaged. There was enough granularity to show a clear trend over time and can be seen in figure 5.3.

Similarly to the first part, the linear regression is very small: 3.1654e-05, which results in a much smaller absolute change over 2 weeks, namely: 38. This regression is also statistically significant different from 0 with a p-value of 0. Comparatively, users did not change their behaviour as much as they did when they use lists. It meant they kept following new users when they could over a 2 week period at a lower rate. The linear regression coefficient for users who start using lists (0.00034174) was higher than the averaged trendline for users who did not start using lists (3.1654e-05). A t-test was done to determine the significant difference of the regression against a confidence of 95% and was found to be significant. Thus over a 2 week period, users of lists ended up adding more information resources than those who did not. There was thus a marked difference in the change of following rate (overall and more granular).

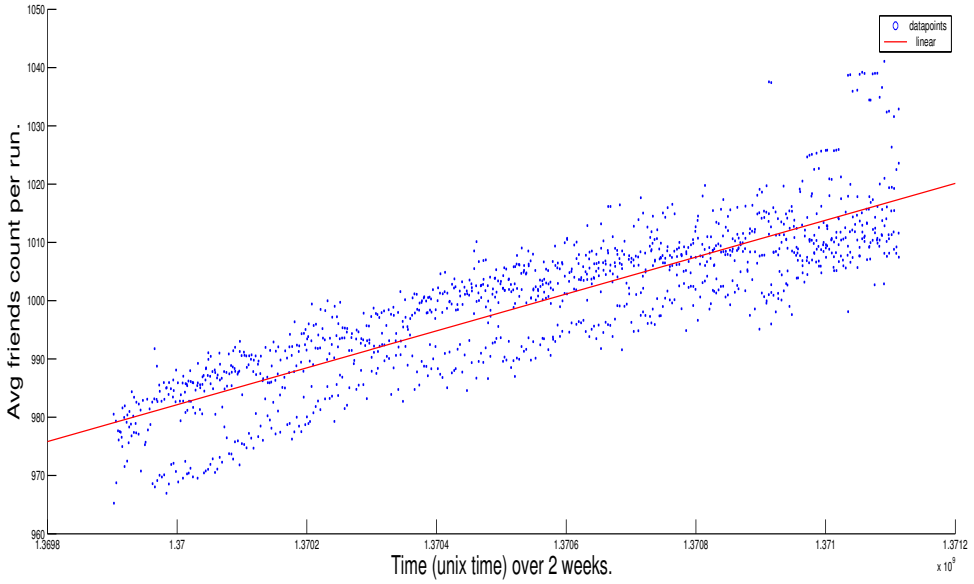


Figure 5.3: Average per run across 2 weeks for non-list users.

5.1.2 Other Findings

An unexpected and interesting result from the data was the gated effect over time. In periods of 24 hours there are more people creating lists. This is a daily trend over the 2 weeks (there are 14 gates). This gated effect had no effect on what was intended to be measured (a trend over 2 weeks is what is wanted). The assumption is that the gates relate to how the data was collected. If it is assumed that when users are active on Twitter (tweeting) they create their lists and they use Twitter generally in 24 hour intervals (daily), then this gated effect will appear. Users that are tweeting now are likely active in 24 hour periods: roughly, the same time each day. In other words, users created lists when they were on Twitter (actively tweeting).

Twitter's API rate limits and non-consistent uptime sometimes bring about errors (tokens failing, or Twitter not accepting requests) in data collection. Some users' data might not be collected every hour. However, this does not affect the result we want, because if there is data at the beginning of the 2 week period and data at the end with none in between, then you will have information on how many new friends the users accrued (and the rate of change). However, having more granularity might provide extra insight. For example, the change in *friends* between 600 and 1800 could follow a trend that it is not

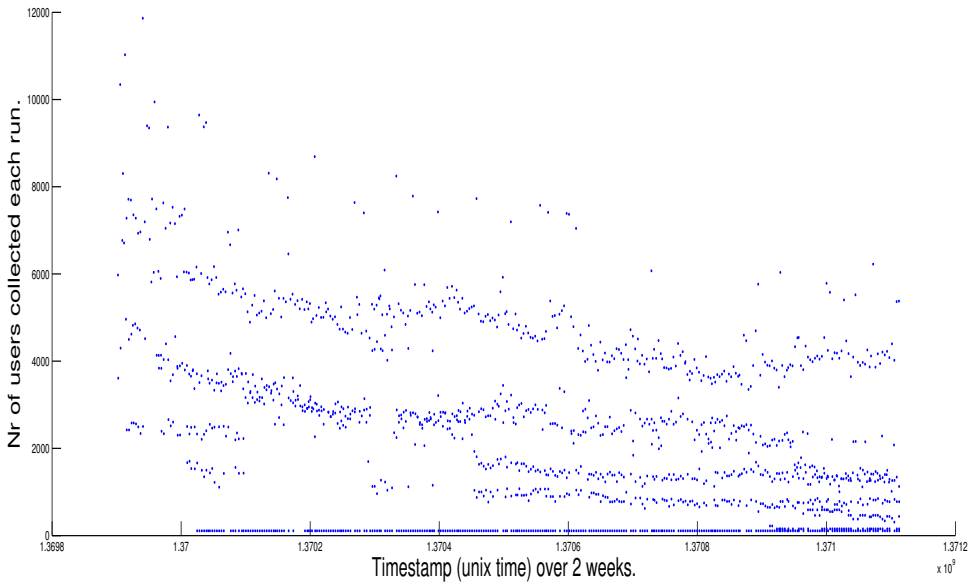


Figure 5.4: Users collected per run.

expected.

As expected, Twitter's API did not behave consistently. Over time tokens failed and in some cases Twitter's API went down. There were thus runs where few users' data were gathered. The distribution over time can be seen in figure 5.4. Over time, the amount per run decreased. The runs that could only gather less than 200 users were removed from the dataset. Only more successful runs (with more users) were used for this study.

5.2 Experiment 2: Effectiveness

The next experiment determined if there was a visible difference in the amount of 'resources' (friends) users who use lists have vs those who do not use lists. A total of 31 684 users were randomly sampled from the public timeline, along with information on whether they use public lists or not. Of the whole population, 16,67% uses lists. This is consistent with other literature studying list usage (Yamaguchi *et al.*, 2011). The users who use lists, on average, follow 962 users, while the the users who do not use lists, on average follow 388 users. As expected this is above the overall average for Twitter, due to a bias to more

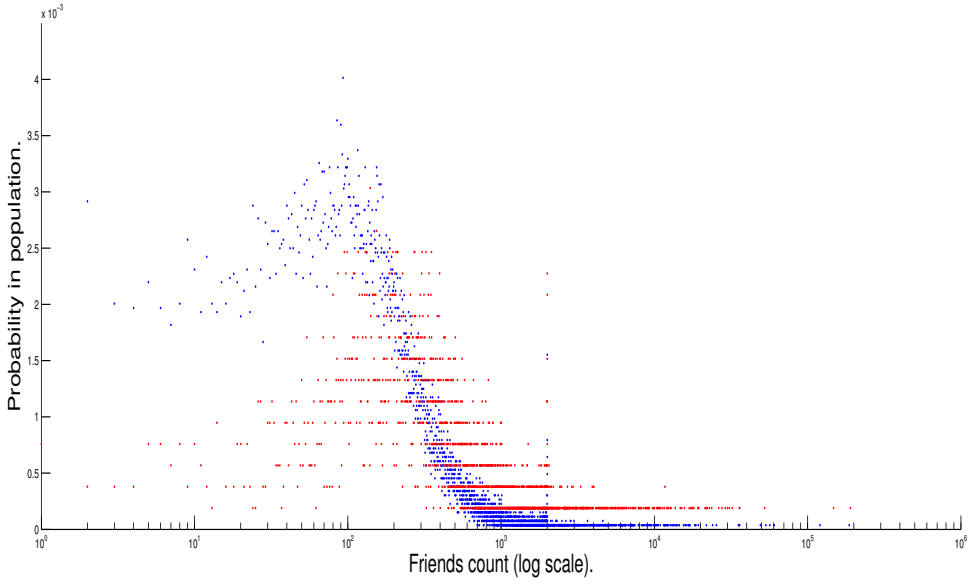


Figure 5.5: Distribution of friends count split between lists and no lists.

active users (those who tweet). Active users (those who tweet) follow more people (Basch, 2012; Beevolve, 2012). The mode for people who use lists are 140, and the mode for people who do not use lists are respectively 106 and 93. Figure 5.5 displays the (normalised) distribution.

The y-axis is the probability of a user (in the sample) following a certain amount of users (x-axis). The x-axis is a logarithmic scale. The graph is split between those use who use lists (red) and those who do not (blue). The anomaly at 2000 friends is again due to how Twitter protects against spam.

A two-tailed t-test to test the significant difference in means between the two populations (lists vs no-lists) was statistically significant (against significance p of 0.05). The results thus indicate that there is a significant difference in how many users a person follows split between those who use lists and those who do not.

It seems that users of lists are outliers. 81% of Twitter users follow less than a 100 people (Bourne, 2010). On average 16,67% use lists, however this increases over time. In figure 5.6 this can be seen. The x-axis is the percentage of users above a certain amount of friends, who uses lists. So for example, of all the users above 2000, 45% use lists. It increases rapidly, and the stabilises around

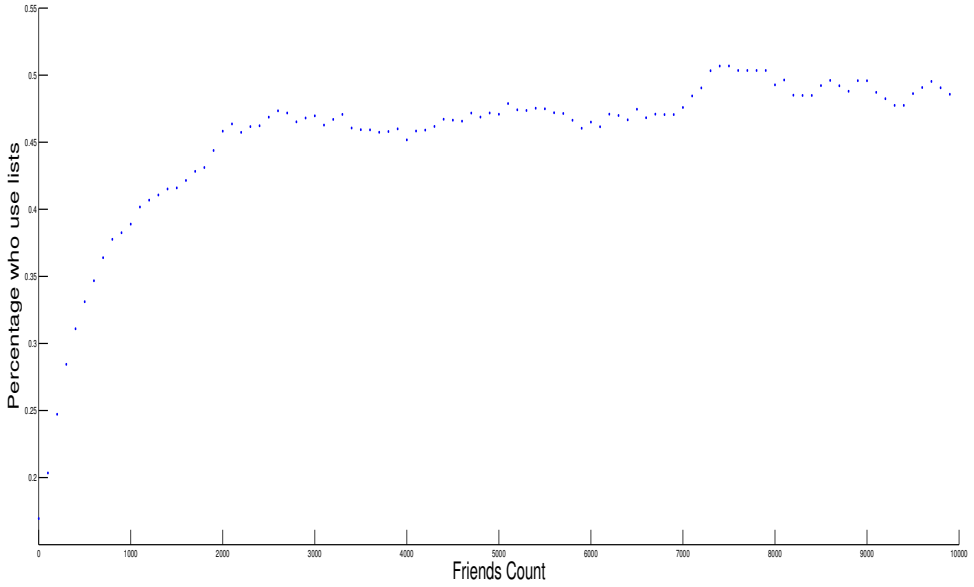


Figure 5.6: Percentage of users who use lists above a certain threshold.

post-2000, where it shuffled between 45% and 50%. This means those users who actively take on large number of followers, also actively use lists a lot more.

The distribution of the size (member count) of lists are in graph 5.7. As expected, the majority of lists are less than a 100. During the course of the research, the cap of lists were 500. At the end you can see it trending upwards to 500 again. The average is 82 and the median is 21. If member count is combined with following count, then the average is 1303 (vs 962).

5.3 Survey

The survey was done to get user API tokens for the quantitative section, to get an understanding on why users use lists and to find a correlation between users experiencing symptoms related to information overload and the friends count. The correlation between symptoms and following counts is exploratory as no models or theory exist that correlate survey questions to definitive results about information overload.

The questions were solicited from various social media accounts. 125 respon-

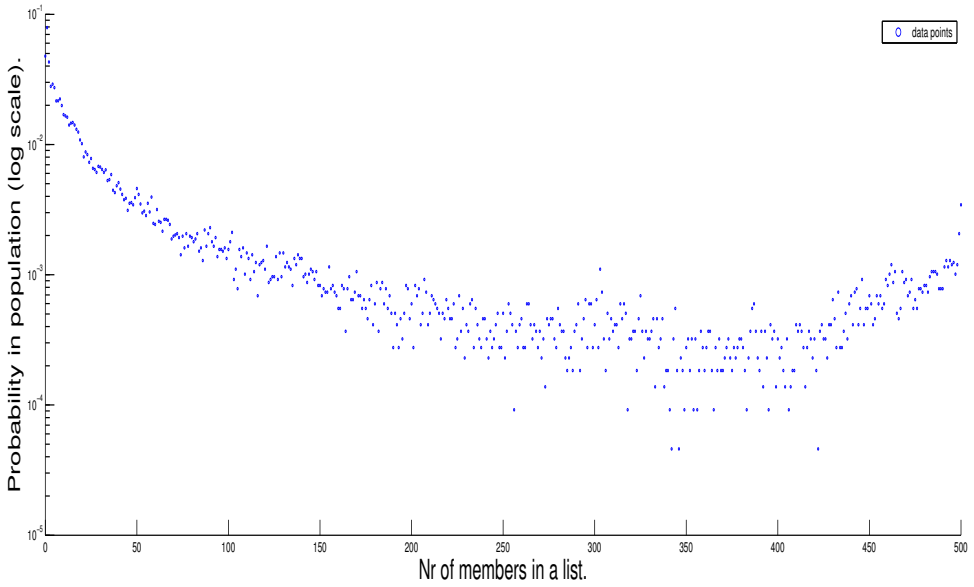


Figure 5.7: Distribution of size of lists.

dents answered the survey. Of those, 20,7% said they used lists. The reasons stated in the survey are tabulated in table 5.2.

The 'other' use cases people used were (posted verbatim):

- 'I follow some curated and automated lists but don't check them often'
- 'I use lists to monitor tweets from people whose updates around a topic I don't want to always see, but would like easy access to.'
- 'Use it to filter infrequent, but important, posters into a lower traffic area so I can see all they're updates.'
- 'by industry'

Table 5.2: Reasons why people use lists

Reason	Amount
I use lists to put in my friends whose updates I want to see.	19
I use lists to filter my home feed into topics (news, sports, etc).	18
I use lists because my home stream became unmanageable.	13
I use lists because it was difficult finding relevant information in my home feed.	8
I use lists to put in people who I'm not sure I want to follow yet.	6
I use lists to filter people by location.	5

Table 5.3: Scores

Question Nr.	Average	Median
I have felt in the past that I cannot keep up with all the tweets.	3.78	4
I often see irrelevant updates.	3.77	4
The content on Twitter is sometimes boring.	3.5	4
There are more non-interesting content on Twitter than interesting content.	3.04	3
Sometimes I don't understand what people are tweeting about.	2.82	2

- 'I keep a private list of people I stalk.'
- 'Lists+Accounts+Flipboard=Control'
- 'I use lists to keep a casual eye on people I used to follow'
- 'friends from university so I can find and read relevant tweets about my major'
- 'people who I need to follow, vs people it doesn't matter if I miss stuff.'
- 'Find tweets to retweet on appropriate accounts - I have multiple.'
- 'I use a list for having all people active on one of the communities I engage in bundled without them occupying my stream.'

A theme that wasn't covered in the main points, is that users use lists purposefully as a 'lower traffic area' so they can stay more easily up to date with a specific set of users. The average scores for each question can be seen in Table 5.3.

5.3.1 Statistical Analysis

The first step was to determine whether the questions measure the same construct. As mentioned this is done with Cronbach Alpha. The result was 0,696: which is acceptable (especially for exploratory research). The questions thus measure the same thing: which is questions about experiencing symptoms related to information overload derived from the framework of Eppler and Mengis (2004).

Of those who experience symptoms related to information overload (greater than 3 on the likert scale), follow on average 253 users and of them only 20,73% use lists. Those who do not experience symptoms related to information overload (less than or equal to 3), follow on average 268 users and of them only 46,5% use lists. Although users who do not experience symptoms related to information overload follow more users, the difference in means between the populations is not statistically significant. Using a t-test to determine the results, ended in a p-value of 0.36. This means that for the observed results there is a 36% chance that it arises from chance. More on this in chapter 6.

The next step was to determine if there are any correlations between the symptoms of information overload with the average number of followers split between those who use lists and those who do not. If they do not use lists, does information overload symptoms increase the more people they follow?

5.3.1.1 Survey Correlations

The first correlation measured was whether an increase in people you follow equate to a rise in symptoms of information overload. The result was not statistically significant (p=0.59). There is thus no correlation. The next test was to check if there is a difference in correlation between those who use lists and those who do not. There was no statistically significant correlation for both those who use lists (p=0.29) and those who do not (p=0.5). There was, however, a stronger correlation between the amount of people you follow and information overload symptoms for people who use lists (than people who do not use lists).

With the questions grouped together there weren't any significant correlations. Each question separately could provide more insight to whether there is any correlation with the amount of people you follow to information overload. This will also be split between those who use and those who do not. As mentioned in the previous chapter, the statistical test that will be used is the Spearman correlation. This measures if there is any monotonic trend (versus Pearson's correlation that measures only a linear trend).

As can be seen there are variances in the strength of correlations for each

Table 5.4: Survey Correlations

Questions	Total	No Lists	Lists
Q1:	p = 0.12. r = 0.10.	p = 0.34, r = 0.04.	p = 0.025 r = 0.325
Q2:	p = 0.30 r = 0.05.	p = 0.14 r = 0.12	p = 0.64 r = -0.06
Q3:	p = 0.16 r = 0.09.	p = 0.17 r = 0.1	p = 0.21 r = 0.14
Q4:	p = 0.7 r = -0.04.	p = 0.68 r = -0.05	p = 0.35 r = 0.07
Q5:	p = 0.97 r = -0.17.	p = 0.92 r = -0.15	p = 0.71 r = -0.09

question. The most significant correlation was for question 1: 'I have felt in the past that I cannot keep up with all the tweets'. Further discussion of the results occur in the next chapter.

5.4 Conclusion

The main findings were positive and confirms the expected behaviour when users start using lists. The next chapter contains further discussion into the meaning of the results.

Chapter 6

Discussion and Future Work

In this chapter, the results are discussed and then compared to the theoretical base to determine whether the research questions were answered. Future work and recommendations are given as well to further research in this area.

6.1 Discussion

6.1.1 Experiment 1: Change in following rate

The varied communication that exists when following too many users, decreases the rate at which users increase their 'resource base' (Butler, 2001). When there is opportunity for further benefit provision, and an increase in novelty, then users on average increase the resource base (Butler, 2001; Jones *et al.*, 2004; Koroleva *et al.*, 2010). A coping strategy (or countermeasure) helps address issues related to information overload. By mitigating the effect of too much communication (too many tweets), users are capable of increasing their novelty by following more users.

This was found to be the case. Lists start being used when users follow on average around 620 users. In the 2 weeks after list creation, this increased to around 900. At the point of list creation, users continue following new users. Over the 2 weeks there is a statistically significant increase (correlation). Across this same spectrum (time and friends) for users who do not use lists, the rate of change is less. Thus when users start using lists, they start following more users than if they would have if they did not start using lists.

6.1.2 Experiment 2: Effectiveness

It is clear from the first experiment that when users start using lists, they increase the amount of information they can process, more so than users who do not use lists. The next experiment determined to what extent lists help overall with coping with information overload. If lists help, then the result should be that users can handle a larger 'resource base' (more friends) than those who users who do not. Over time, there should be a statistically significant difference between the two groups of users. This is what was found. Those who use lists on average follow more users than those who do not. Lists are thus an effective way to cope with information overload. It increases the novelty and information that can be processed.

An interesting statistic from this experiment is how the usage of lists as a percentage, change based on how many friends a user has. The more friends a users has, the higher the chances are that will use lists. Lists are not used a lot (only by about 16% of the users, but when users follow more than 2000 people, it increases to almost 50%). This gives more validation to the idea that when faced with a large set of resources to try and benefit from, users employ information management tools (Jones *et al.*, 2004; Butler, 2001). When following a smaller number of users, information management tools are not used as they are not as needed.

6.1.3 Survey

The survey explored some of the qualitative aspects around list usage on Twitter. From the results, a conclusive statistical result was that of the users who do not experience symptoms related to information overload, 46,5% of them use lists. Much more so than 20,73% of those who do not. Only one of the tests to determine correlations between symptoms of information overload and how many friends a user has were statistically conclusive. It was for Q1: 'I have felt in the past that I cannot keep up with all the tweets.' This is consistent with what was found in the previous experiments. Combined with the reasons why people use lists and when they start using lists, this makes sense. Of the group of people who have felt they cannot keep up with the tweets, it would be users who use lists. Overall, Q1 had the best results. In the future work

section, further explanation will be given in order to design a better framework for hopefully measuring information overload qualitatively as well as quantitatively on micro-blogging services.

The reasons people use lists provide great insight into why it is employed by users and further emphasises its use as an information management tool by, for example, using it to create a purposefully lower traffic area and using it to filter by topic.

Although the questions as a construct were reliable (they measured the same thing), it did not correlate to the amount of users a person follows. There could be several reasons for this.

Jones *et al.* (2004) states that users have their own thresholds. They might be experiencing symptoms related to information overload at various levels. The granularity of the Likert scale could also affect the options users choose. An average of 1 to 5 may not be granular enough to compensate for users' different thresholds over the 1 to 20000 people they might follow.

The first three questions were not statistically significant, but correlations were generally positive (not by much). If there are not enough datapoints, statistical significance testing fails easier, because it is more to difficult to prove that the results arise out of chance. Increasing the datapoints can perhaps improve the dataset. By increasing the amount of people being questioned to 300+ will already improve the results (at least the statistical significance).

Another reason why there were statistically non-significant correlations is that the questions used simply do not relate to following count. Increasing the 'resource base' increases the amount of information, however, most of the questions were subjective (based on symptoms of having to process too much information).

Finally, it could also be the case that the questions were the correct questions, and they simply do not correlate to the amount of *friends* a user has as a proxy for information overload.

6.1.4 Threshold

Combining the 2 quantitative experiments provides interesting insight. Users start using lists close to the overall average. It suggests that users who already follow a lot of users eventually switch to using lists. Following a lot of users means a higher information load. These users seem to be able to handle a lot more information in their streams before they even consider using lists (than none-list users). This is consistent with literature as Jones *et al.* (2004) explained: users have different thresholds at which they can deal with information. However, when following about 800 users, even as an active user, those users are already outliers. As mentioned previously, 81% of Twitter users follow less than a 100 people (Bourne, 2010). This is also emphasised given how few people use lists (less than 20%). It is also known that a very small percentage of people on Twitter produce the most tweets on Twitter (Wu *et al.*, 2011). These users are the active users, consuming, producing and curating the data. At the beginning of June 2013 (after the data collection), Twitter announced changes to the lists features that reflect the need to cater to these outliers. Users can now create 1000 lists (up from 20) with each list having a maximum of 5000 accounts (up from 500) (Blagdon, 2013).

An interesting case study on lists was the recent Boston marathon bombings. A user (Danny Sullivan) created a list that included reporters and other users that were in the Watertown region where the hunt for the suspect took place. Danny follows over 3700 people (an outlier) and has over 15 public lists. These people are the information producers that helps curate information for the majority of Twitter users that only consume information. By expanding on the usage of lists in this fashion, they are empowering the advanced users to help filter information overload so that novelty can be increase not only for the individual, but for others as well.

6.2 Research Focus

To reiterate the research purpose: It '...focuses on measuring the effectiveness of user-curated filtering (through Twitter lists) as a coping/countermeasure strategy against information overload on micro-blogging services.

To measure the effectiveness, thorough research is done too look at users who use lists and those who do not and whether their usage significantly differs. It is done quantitatively and qualitatively.'

Effectiveness is defined as being 'successful in producing a desired or intended result.' By measuring the difference of usage between users of lists, and those who do not use lists, the results showed that user-curated filtering is an effective coping strategy for information overload on microblogging services. When users opt to start using lists, they follow people at a higher rate than when they do not use lists, indicating that users could take on more information (as the theory suggests (Butler, 2001)).

Similarly, over time, the gap becomes larger, showing that users of lists eventually follow a lot more people. At the new level, users can function (and possibly experience information overload) at this new higher level (Jones *et al.*, 2004), group effectiveness still exists, as these outliers work to filter and curate their streams, not only for their benefit, but for others as well (Butler, 2001; Wu *et al.*, 2011).

In terms of the research question, the qualitative results give subjective credence to why users use lists, especially the top-cited reason: 'I use lists to put in my friends whose updates I want to see.' Users use lists to create a purposefully lower traffic area so that they don't have to contend with extra information from users that they don't regard as equally important. In terms of Butler (2001), this means users are using tools to not only to split up varied communication, but also to weight the resources (other users) to their liking.

The models of Jones *et al.* (2004) and Butler (2001) gave insight into why usage patterns differ between users who use lists and those who do not. It helped not only to understand what to expect when people start using these tools, but also why outliers exist (Jones *et al.*, 2004).

6.3 Future Work and Recommendations

For the quantitative experiments there are several possible improvements and expansions. For experiment one, the dip in the friends count is interesting. The suspected reason for this is due to users taking people out of their 'following' stream and putting them into them into lists, specifically for a lower traffic area, so they can be up to date. This was not proven, and not within the scope of the research. Knowing who the person is following and whom they put into lists would provide further insight into this. The timespan for the rate of change experiments were done over 2 weeks. Very few longitudinal studies have been done on Twitter. When Jones *et al.* (2004) studied USENET behaviour they chose the time-period to study (a week) based on their judgement of the CMC in question. Since little literature exists that suggest timespans for changes in user behaviour on Twitter (especially with lists), an exploratory 2 weeks were chosen. There could be other patterns for a longer period that could provide further insight to the how usage changes over time.

The other bottleneck was Twitter's API. In the past Twitter responded to research requests where large amounts of data could be anonymized and used for research. However their growth and recent focus on developing Twitter as a business means they are not providing it any more. Certain API accounts could also be whitelisted, allowing much more requests per hour (up to 20 000). This is also no longer the case as Twitter eventually migrated in May 2013 to a new version of their API (v1.1). All API requests have new buckets per endpoint and all requests have to be authorised. This makes it increasingly difficult to gather enough research data. There are service providers such as Gnip that provide access to all of the tweets for a fee (they work with Twitter). However, after requesting the data for the research, they responded that they only store tweets and not any other longitudinal data. There are rising micro-blogging communities that although they have a lower volume, have much better APIs. An example is App.net, a recent micro-blogging community that is a paid-for model.

The survey served a few purposes: gathering tokens for the quantitative experiment, finding out why users use lists, and trying to determine if the questions

derived from the literature equate to following counts. The results found in the survey can be used to derive better questions (use those who had the highest correlations and rephrase those that did not). More research into this can definitely be done to create a framework for qualitatively measuring information overload on feed-based CMC tools.

Overall, the research provides insight into the effectiveness of user-curated filtering. Due to the power of outliers contribution to information creation, it would be interesting to determine how much lists help in the diffusion of information on Twitter. A large percentage of the users who consume information on Twitter, use lists. However, there is not a clear correlation yet, if they contribute to information diffusion. There are clues that makes this seem to be the case (as shown by Twitter's recent changes to lists), and that the outliers anyway produce the most content. Statistically it might be the case.

6.3.1 Possible connection between prefrontal cortex and information threshold

There is the possibility for very interesting research that involves scopes from neuro-science and cognitive psychology. The research of Jones *et al.* (2004) shows that outliers will exist: people have their own levels of information overload. However, on average, and for the most part, users experience information overload at the same levels, which is why we see results such as shown by Dunbar (1998). There exist outliers online that far exceed most people's capability to manage and process information. An example is the blogger, Robert Scoble. To quote what he said: 'Something really changed on my feed about two weeks ago. Facebook started showing me a lot more sponsored posts from brands I follow. What did that change for me? I unliked more than 800 brands like Shell, Wells Fargo, etc. Things that I 'like' but that I rarely wanted to hear from.' Another one: 'Oh, some day I'll tell you about why I wrote more than 1,500 Gmail filters. They throw away more than 300 emails every day. Every day. It's the best thing I ever did for my productivity.' It's an incredible amount of effort, and he grasps the potential return from curating information. If we assume the social brain hypothesis of Dunbar (1998): that larger brains survived due to being able to process and benefit from increas-

ingly larger social circles, then being able to process and wade through the vast amount of information in the digital age, will bring about greater chance of evolutionary success. There exists thus an interesting possibility and research to determine whether outliers in terms of information and social management exhibit different prefrontal cortex behaviour. Are there clues there to why these outliers can manage such an abundance of information? And what does this mean for creating tools to effectively curate information?

Chapter 7

Conclusion

The research set out to determine the effectiveness of user-curated filtering (through Twitter Lists) as a coping/countermeasure strategy against information overload on micro-blogging services (Twitter). In 1970, Alvin Toffler, popularised the term 'information overload' in his book 'Future Shock' (Toffler, 1970), three decades before the world wide web became a part of society. Since then, the amount of information the world produces has exploded (King, 2011). It is affecting us in unprecedented ways, and research needs to be done to determine how we can extract the most novelty from the deluge of information.

Information overload is occurring in various places each day, from the organization (Eppler and Mengis, 2004) to especially the world wide web (Jones *et al.*, 2004). Like most systems in the world, you can not take humans out of the equation. On online social-networks, the main sources of information are other users. Depending on the design of the website in question, it inevitably reaches a threshold where communication activity is hampered: any additional activity becomes more and more infeasible, not creating any additional novel information (Jones *et al.*, 2004; Butler, 2001).

What is also evident is that users will try to increase their novelty in social systems up to the point where they will have to start coping. In real life social networks, people have always needed to cope. People make time for different people in their lives. The oft-used heuristic is to simply 'leave': contacts between good friends can lead to them becoming acquaintances and eventu-

ally strangers again (Miritello *et al.*, 2013*a*). The results of this heuristic is evident in research by Dunbar (1998): people not easily forming large groups with strong social relationships.

With the advent of online communication, this heuristic is also used a lot as is shown by Jones *et al.* (2004) and this research (very few people opt to use lists). Because online communication tools are restricted by their design, natural tie activation and deactivation are hampered, and thus information management tools, such as lists are needed. These help to constrain the communication activity and make it more manageable so that further benefits can be provided to the individual (Butler, 2001).

The research confirmed and extended the models of Butler (2001), Jones *et al.* (2004) and Eppler and Mengis (2004). These models have not been used in the context of micro-blogging services such as Twitter before. Their assumptions and hypotheses about behaviour on Twitter still holds true. Increasingly varied communication do hamper the amount of novel and resourceful information that can be processed. The design of the system also has an effect on how much information can be processed. If lists are not used, the behaviour looks different than when lists are used. User-curated filtering (through lists) is a coping strategy as it changes the dynamics of the interactions of individuals. It is also an effective coping strategy because it mitigates varied communication found when resources increases. This ultimately leads to an increase in the size of the resources users can handle (Butler, 2001).

Although user-curated filtering is predominantly used by outliers of information consumption, it is still valuable in the context of the 21st century. While algorithmic filtering can help everyone (as no input is needed to intentionally filter; catering for the masses), user-curated filtering tools help in two ways. As explained earlier (in the literature review), on large CMC tools such as Wikipedia, a small percentage of active gatekeepers provide a surprisingly large benefit to the community. On Twitter as well, a small percentage of people produce the most content. There is value in the small outliers who actively find, consume and produce information. As was shown, the users who follow more than 2000 people, consistently about 45% of them use lists. In

terms of benefits for individuals, user-curated filtering also has merit over algorithmic filtering. As one of Twitter's co-founder said: Twitter is a tool for public conversation (Dorsey, 2012). Algorithmically filtering the public conversation has a problem of potentially silencing voices that could be important.

This research can also be applicable to other fields, beside microblogging services (and online social-networks). It's not just about providing information management tools to users, but it is realising that some users are adept at it. These outliers that avidly organise and curate information can provide benefit for a lot of people. In a company, for example, there could be people that can do information management better than others. It's a trade-off that has to be considered by the company, but there are cases where rather making a tool that fits a small subset of users, but gives them immense power in curation and organization, will benefit the rest much better than making a tool that has to be understood and used by a larger population. The normal users won't contribute much, and the experts will not be empowered enough.

To conclude: as humanity is currently reorganising around a global society and its reliance on an abundance of information, it is important to understand how to deal with connecting on unprecedented levels and how to gain the most benefit from this. This research on the effectiveness of user-curated filtering on micro-blogging services as a coping strategy for information overload hopefully adds to the understanding of this, and takes it a step further so that we (as individuals and society) can benefit from one of the most amazing and exciting times in recent history.

Appendices

Appendix A

Technical Details

Crawling large quantities of data from can be done in several ways, using different languages and technologies. This addendum serves to give an overview how the data was collected for the research. Other researchers can learn from this in order to do something similar in the future. One of the biggest hurdles was working around the limitations of the API.

A.1 General

The language that was used, was Python. Python has a large community of people and libraries, as well as having syntax that is easy to write and read. Since Twitter's API gives their data in JSON, a document-oriented database tool was used to store it, namely: MongoDB. The scripts were run mainly from a micro instance from Amazon's Elastic Compute cloud service. The specifications of the micro instance was enough to sit within the free tier they provide. The only thing that was upgraded was the disk space, as 8gb for both the data and the operating system of the instance was not enough. Fabric, a python library, was used to quickly deploy code changes to the instances. Matlab and python was used for data manipulation, presentation and statistical analysis.

A.2 Experiment 1

This experiment was responsible for collecting two data sets. The first was determining at what point users starting using lists. The bottleneck here was using the 'get lists/list' API endpoint, as it is limited to 15 calls per 15 min

per user token. About 130 users authenticated the application, so the maximum amount of users to check was about 7 800 users an hour. If each user didn't qualify the requirements: created first public list in the past 2 weeks, then it was discarded. This was a slow process, and took about 2 - 3 weeks to collect the data. This experiment wasn't time sensitive, so it could afford to fail. This did happen occasionally in order get a clearer understanding of why errors occur for the more time critical experiment. The cases where it failed the most often was due to Twitter's API timing out. It is accepted that not all requests will succeed. When it was a critical part in the code (such as fetching a new batch of users to check from the public stream), the code would fail.

API requests can be fulfilled faster than the rate limit, and thus the code was forcefully set up so that it could only digest the correct amount of users for every 15 minutes. After the data was collected, the script waited the allotted time for the API buckets to be full again (for all the tokens) before the script would loop again.

The second data set was to compare this data to the set of users that do not use lists over that time period. This was a more difficult task as it was time dependent. So with any sort of failure, the script should've just continued. With any token failure, or erroneous API calls, the script should not stop. This would make the script not collect how many users a person follows at that specific moment (over 2 weeks). A lot more users' data could be gathered due having to only use 'get users/show' endpoint which allows 180 requests every 15 minutes. However, as was shown in the results chapter, this wasn't always the case. Tokens failed, and a lot of errors occurred.

If tokens failed, they were discarded for the current round (every 15 minutes). So if there was an API error after getting 100 (out 180) users, then the token was discarded until the next run. Ideally, for the future, you would want to be able to perfectly cycle tokens and use all the available hits. However, in order to do this, you would probably have to keep track yourself how much hits has been done (per token). The remaining rate limits for each endpoint can be retrieved, but this adds overhead in terms of how fast you could collect data, because you have to do additional requests. More granularity will mean

more overhead. If it is a lot of users (with a lot of tokens), you could easily not retrieve the potential amount of user data within the time limit.

A.3 Experiment 2

This experiment is similar to the first data set of experiment 1 without the 2 weeks and first list limitations. Users were sampled from the public stream, and checked whether they used lists or not. This was done much faster due to not having the limitations. Other than, it was exactly similar.

A.4 Survey

The survey was also coded in Python, using the Flask web framework. Heroku (a cloud hosting service) was used to host it. MongoDB also served as the document-store, with Twitter's Bootstrap serving as the design base. The Twython twitter library was used to work with Twitter's content.

A.5 Data Analysis

The tools that were used to do the data analysis was Python and Matlab. Python was used to do basic analysis (such as averages, medians, etc) and to also manipulate the data from the database to store it in text files so that Matlab can easily import it. Matlab was used to draw the graphs and figures as well as calculating the statistical correlations (t-test, Cronbach alpha and Spearman's correlation).

Matlab struggled when it had to work with 4 million datapoints for the second data set of the first experiment. It took almost a minute to redraw any of the graphs in different ways. Due to that many datapoints, the whole graph could not be displayed in the thesis. It easily reached 80mb+ of vectorised data.

A.6 Conclusion

There is little guidance on the web that shows how to effectively data mine Twitter for research. The time used to develop and do research on Twitter can

hopefully be reduced in the future by the addition of this addendum.

Appendix B

Twitter's Terms of Service

The 'Terms of Service' can be found at: https://twitter.com/tos. The 'Developer Rules of the Road' can be found at: https://dev.twitter.com/terms/api-terms.

Twitter allows users to extract data from Twitter. They are however bounded by a terms of service. These are the relevant parts extracted from it. Twitter Content (or Twitter data) is defined as: 'any information, text, graphics, photos or other materials uploaded, downloaded or appearing on the Services' (Twitter, 2013*c*). The rights contained is explained in this chapter from the Terms of Service:

'You retain your rights to any Content you submit, post or display on or through the Services. By submitting, posting or displaying Content on or through the Services, you grant us a worldwide, non-exclusive, royalty-free license (with the right to sublicense) to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute such Content in any and all media or distribution methods (now known or later developed).

You agree that this license includes the right for Twitter to provide, promote, and improve the Services and to make Content submitted to or through the Services available to other companies, organizations or individuals who partner with Twitter for the syndication, broadcast, distribution or publication of such Content on other media and services, subject to our terms and conditions for such Content use.

Such additional uses by Twitter, or other companies, organizations or individuals who partner with Twitter, may be made with no compensation paid to you with respect to the Content that you submit, post, transmit or otherwise make available through the Services.

We may modify or adapt your Content in order to transmit, display or distribute it over computer networks and in various media and/or make changes to your Content as are necessary to conform and adapt that Content to any requirements or limitations of any networks, devices, services or media.

You are responsible for your use of the Services, for any Content you provide, and for any consequences thereof, including the use of your Content by other users and our third party partners. You understand that your Content may be syndicated, broadcast, distributed, or published by our partners and if you do not have the right to submit Content for such use, it may subject you to liability. Twitter will not be responsible or liable for any use of your Content by Twitter in accordance with these Terms. You represent and warrant that you have all the rights, power and authority necessary to grant the rights granted herein to any Content that you submit.'

Appendix C

Initial reasons for using lists

The literature gave clues as to why users used lists (mostly from Eppler and Mengis (2004)). In order to stream-line the answering of why users use lists, an initial set of responses were gathered, and grouped together with the literature to create options for users to choose. If it wasn't amongst these options (in the survey), then there was a field to add additional uses. Some tweets has since the data collection (November 2012) been deleted, or is part of a private account. They thus can't be referenced.

Responses:

- I just started + I love them It organizes my interests + keeps my homepage clean for hobbies I dont want to read about all day (_indelibe, 2012)
- yes: one for news, one for friends, one for sports.
- when I followed back I used them to sort out who I preferred hearing from. Eventually I unfollowed who I preferred not to.
- I follow an array of different people on twitter, so I group my friends into private lists, so I don't miss their tweets (aprilskye_, 2012)
- Yes, to split up who I follow, eg real people, companies, software news, work stuff etc.(AdamDempsey, 2012)
- I don't, but I have heard people put users they don't actually follow in lists to keep them out of the main timeline. (rinbrand, 2012)

- yeah, to see all the people I actually know irl (fueledbyrobert, 2012)
- I use twitter lists to store lists of people who I normally wouldn't want to read on my feed but read occasionally. (nobuyukinyuu, 2012)
- I can make a widget of the lists and put it on my blog on the causes section. If my blog is approved.
- Yes. Accounts that are some sort of interesting but too much "chaty". It prevent them to take over my TL

List of References

AdamDempsey (2012 November). Yes, to split up who i follow, eg real people, companies, software news, work stuff etc.
Available at: `http://twitter.com/AdamDempsey/status/281026720746844163`

Adamic, L.A., Zhang, J., Bakshy, E. and Ackerman, M.S. (2008). Knowledge sharing and yahoo answers: everyone knows something. In: *Proceedings of the 17th international conference on World Wide Web*, pp. 665–674. ACM.

Ahuja, G., Soda, G. and Zaheer, A. (2012). The genesis and dynamics of organizational networks. *Organization Science*, vol. 23, no. 2, pp. 434–448.

André, P., Bernstein, M. and Luther, K. (2012). Who gives a tweet?: evaluating microblog content value. In: *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pp. 471–474. ACM.

Applum (2013 June). What is edgerank?
Available at: `http://www.whatisedgerank.com/`

aprilskye_ (2012 November). I follow an array of different people on twitter, so i group my friends into private lists, so i don't miss their tweets.
Available at: `http://twitter.com/aprilskye\_/status/281026608071061504`

Aral, S. and Van Alstyne, M. (2011). The diversity-bandwidth trade-off. *American Journal of Sociology*, vol. 117, no. 1, pp. 90–171.

Aral, S. and Walker, D. (2011). Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science*, vol. 57, no. 9, pp. 1623–1639.

Ariely, D. (2000). Controlling the information flow: Effects on consumers' decision making and preferences. *Journal of Consumer Research*, vol. 27, no. 2, pp. 233–248.

Bakshy, E., Rosenn, I., Marlow, C. and Adamic, L. (2012). The role of social networks in information diffusion. In: *Proceedings of the 21st international conference on World Wide Web*, pp. 519–528. ACM.

Basch, D. (2012 July). Some fresh twitter stats (as of july 2012). Available at: `http://diegobasch.com/some-fresh-twitter-stats-as-of-july-2012`

Bawden, D. (2001). Information overload. *Library & information briefings*, , no. 92, pp. 1–15.

Bawden, D. and Robinson, L. (2009). The dark side of information: overload, anxiety and other paradoxes and pathologies. *Journal of information science*, vol. 35, no. 2, pp. 180–191.

Beevolve (2012 October). Following count distribution on twitter in 2012, by share of twitter users (in percent). Available at: `http://www.statista.com/statistics/188686/following-count-distribution-on-twitter/`

Benbunan-Fich, R. and Koufaris, M. (2009 April). An empirical examination of the sustainability of social bookmarking websites. *Information Systems and e-Business Management*, vol. 8, no. 2, pp. 131–148. ISSN 1617-9846. Available at: `http://www.springerlink.com/index/10.1007/s10257-009-0114-8`

Bernstein, M., Hong, L., Kairam, S., Chi, H. and Suh, B. (2010). A torrent of tweets: managing information overload in online social streams. In: *In Workshop on Microblogging: What and How Can We Learn From It?(CHI'10)*. Citeseer.

Blagdon, J. (2013 May). Twitter raises limits on lists, enables 1,000 groups of 5,000 accounts each. Available at: `http://www.theverge.com/2013/5/30/4381712/twitter-raises-limits-on-lists-1000-groups-5000-accounts-each`

Borgs, C., Chayes, J., Karrer, B., Meeder, B., Ravi, R., Reagans, R. and Sayedi, A. (2010). Game-theoretic models of information overload in social networks. In: *Algorithms and Models for the Web-Graph*, pp. 146–161. Springer.

Bourne, M. (2010 December). Twitter follower semi log graphs. Available at: `http://www.intmath.com/blog/twitter-follower-semi-log-graphs/5496`

Boyd, D.M. and Ellison, N. (2007 December). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, vol. 3, no. 1, pp. 19–230.
Available at: `http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2007.00393.x/full`

Brockman, B.K. and Morgan, R.M. (2003). The role of existing knowledge in new product innovativeness and performance. *Decision Sciences*, vol. 34, no. 2, pp. 385–419.

Brown, H.D. and Yian, W. (2000). Principles of language learning and teaching.

Butler, B. (2001). Membership size, communication activity, and sustainability: A resource-based model of online social structures. *Information systems research*, vol. 12, no. 4, pp. 346–362.

Butler, B., Joyce, E. and Pike, J. (2008). Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. In: *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1101–1110. ACM.

Casagrande, D.G. (1999). Information as verb: Re-conceptualizing information for cognitive and ecological models. *Journal of Ecological Anthropology*, vol. 3, pp. 4–13.

Cha, M., Haddadi, H., Benevenuto, F. and Gummadi, P.K. (2010). Measuring user influence in twitter: The million follower fallacy. *ICWSM*, vol. 10, pp. 10–17.

Chase, W.G. and Simon, H.A. (1973). Perception in chess. *Cognitive Psychology*, vol. 4, no. 1, pp. 55–81.

Cheng, A. and Evans, M. (2009 June). An in-depth look inside the twitter world. Available at: `http://www.sysomos.com/insidetwitter/`

Chewning, E., Harrell, A. *et al.* (1990). The effect of information load on decision makers' cue utilization levels and decision quality in a financial distress decision task. *Accounting, Organizations and Society*, vol. 15, no. 6, pp. 527–542.

Comarela, G., Crovella, M., Almeida, V. and Benevenuto, F. (2012). Understanding factors that affect response rates in twitter. In: *Proceedings of the 23rd ACM conference on Hypertext and social media*, pp. 123–132. ACM.

Cook, G. (1993). An empirical investigation of information search strategies with implications for decision support system design. *Decision Sciences*, vol. 24, no. 3, pp. 683–698.

Craik, F.I. and Lockhart, R.S. (1972). Levels of processing: A framework for memory research. *Journal of verbal learning and verbal behavior*, vol. 11, no. 6, pp. 671–684.

Davis Jr, F.D. (1986). *A technology acceptance model for empirically testing new end-user information systems: Theory and results*. Ph.D. thesis, Massachusetts Institute of Technology.

Decety, J. and Jackson, P.L. (2004). The functional architecture of human empathy. *Behavioral and cognitive neuroscience reviews*, vol. 3, no. 2, pp. 71–100.

Dorsey, J. (2012 January). Dld 2012 jack dorsey: Twitter has a business model that works.
Available at: `http://techcrunch.com/2012/01/22/dld-2012-jack-dorsey-twitter-has-a-business-model-that-works/`

Driver, M. and Streufert, S. (1969). Integrative complexity: An approach to individuals and groups as information-processing systems. *Administrative Science Quarterly*, pp. 272–285.

Drucker, P.F. (1992). *The age of discontinuity: Guidelines to our changing society*. Transaction Books.

Duan, W., Gu, B. and Whinston, A. (2009). Informational cascades and software adoption on the internet: an empirical investigation. *MIS quarterly*, vol. 33, no. 1, pp. 23–48.

Dunbar, R. (2012). Social cognition on the internet: testing constraints on social network size. *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 367, no. 1599, pp. 2192–2201.

Dunbar, R.I. (1998). The social brain hypothesis. *Evolutionary Anthropology*, vol. 6, pp. 178–190.

Edmunds, A. and Morris, A. (2000). The problem of information overload in business organisations: a review of the literature. *International journal of information management*, vol. 20, no. 1, pp. 17–28.

Eppler, M. and Mengis, J. (2004 November). The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines. *The Information Society*, vol. 20, no. 5, pp. 325–344. ISSN 0197-2243.
Available at: `http://tandfprod.literatumonline.com/doi/abs/10.1080/01972240490507974`

Facebook (2012 February). Facebook s-1 filing.
Available at: `http://www.sec.gov/Archives/edgar/data/1326801/000119312512034517/d287954ds1.htm`

fueledbyrobert (2012 November). yeah, to see all the people i actually know irl.
Available at: `http://twitter.com/fueledbyrobert/status/281028237231001600`

Galbraith, J.R. (1974 May). Organization Design: An Information Processing View. *Interfaces*, vol. 4, no. 3, pp. 28–36. ISSN 0092-2102.
Available at: `http://interfaces.journal.informs.org/cgi/doi/10.1287/inte.4.3.28`

Gladstein, D.L. (1984). Groups in context: A model of task group effectiveness. *Administrative science quarterly*, pp. 499–517.

Gonçalves, B., Perra, N. and Vespignani, A. (2011 August). Modeling Users' Activity on Twitter Networks: Validation of Dunbar's Number. *PLoS ONE*, vol. 6, no. 8, p. e22656. ISSN 1932-6203.
Available at: `http://dx.plos.org/10.1371/journal.pone.0022656`

Grabowicz, P.a., Ramasco, J.J., Moro, E., Pujol, J.M. and Eguiluz, V.M. (2012 January). Social Features of Online Networks: The Strength of Intermediary Ties in Online Social Media. *PLoS ONE*, vol. 7, no. 1, p. e29358. ISSN 1932-6203.
Available at: `http://dx.plos.org/10.1371/journal.pone.0029358`

Greene, D., Sheridan, G., Smyth, B. and Cunningham, P. (2012). Aggregating content and network information to curate twitter user lists. In: *Proceedings of the 4th ACM RecSys workshop on Recommender systems and the social web*, pp. 29–36. ACM.

Grineva, M. and Grinev, M. (2012). Information overload in social media streams and the approaches to solve it. In: *21st International World Wide Web Conference, Lyon, France.*

Gu, B., Konana, P., Rajagopalan, B. and Chen, H.-W.M. (2007). Competition among virtual communities and user valuation: The case of investing-related communities. *Information Systems Research*, vol. 18, no. 1, pp. 68–85.

Haksever, A. and Fisher, N. (1996). A method of measuring information overload in construction project management. In: *Proceedings CIB W89 Beijing International Conference*, pp. 310–323.

Hampton, K., Goulet, L.S., Marlow, C. and Rainie, L. (2012). Why most facebook users get more than they give. *Pew Internet & American Life Project*.

Herbig, P. and Kramer, H. (1994). The effect of information overload on the innovation choice process: Innovation overload. *Journal of Consumer Marketing*, vol. 11, no. 2, pp. 45–54.

Huber, G.P. and Power, D.J. (1985). Retrospective reports of strategic-level managers: Guidelines for increasing their accuracy. *Strategic Management Journal*, vol. 6, no. 2, pp. 171–180.

Huberman, B., Romero, D. and Wu, F. (2008). Social networks that matter: Twitter under the microscope. *Available at SSRN 1313405*.

_indelibe (2012 November). I just started + i love them it organizes my interests + keeps my homepage clean for hobbies i dont want to read about all day. Available at: `http://twitter.com/\_indelible/status/281025166786576385`

Jacoby, J. (1984). Perspectives on information overload. *The Journal of Consumer Research*, vol. 10, no. 4, pp. 432–435.
Available at: `http://www.jstor.org/stable/10.2307/2488912`

Jacoby, J., Speller, D.E. and Kohn, C.A. (1974). Brand choice behavior as a function of information load. *Journal of Marketing Research*, pp. 63–69.

Java, A., Song, X., Finin, T. and Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. In: *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pp. 56–65. ACM.

Jones, Q., Moldovan, M., Raban, D. and Butler, B. (2008). Empirical evidence of information overload constraining chat channel community interactions. In: *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pp. 323–332. ACM.

Jones, Q., Ravid, G. and Rafaeli, S. (2004). Information Overload and the Message Dynamics of Online Interaction Spaces: A Theoretical Model and Empirical Exploration. *Information Systems Research*, vol. 15, no. 2, pp. 194–210. ISSN 10477047.
Available at:
`http://isr.journal.informs.org/cgi/doi/10.1287/isre.1040.0023`

Joyce, E. and Kraut, R.E. (2006). Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication*, vol. 11, no. 3, pp. 723–747.

Kaplan, A. and Haenlein, M. (2011). The early bird catches the news: Nine things you should know about micro-blogging. *Business Horizons*, vol. 54, no. 2, pp. 105–113.

Kim, D., Jo, Y. and Moon, I.-C. (2010). Analysis of twitter lists as a potential source for discovering latent characteristics of users.

King, B. (2011 January). Too much content: A world of exponential information growth.
Available at: `http://www.huffingtonpost.com/brett-king/too-much-content-a-world-\_b\_809677.html`

Koroleva, K., Krasnova, H. and Gunther, O. (2010). 'STOP SPAMMING ME!'-Exploring Information Overload on Facebook. In: *Americas Conference on Information Systems*. Assocation for Information Systems.
Available at: `http://aisel.aisnet.org/amcis2010/447/`

Kuhl, P.K. (2004). Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, vol. 5, no. 11, pp. 831–843.

Kwak, H., Lee, C., Park, H. and Moon, S. (2010). What is twitter, a social network or a news media? In: *Proceedings of the 19th international conference on World Wide Web*, pp. 591–600. ACM.

Lehmann, J., Gonçalves, B., Ramasco, J.J. and Cattuto, C. (2012). Dynamical classes of collective attention in twitter. In: *Proceedings of the 21st international conference on World Wide Web*, pp. 251–260. ACM.

Lewis, D. and Knowles, K. (1997). Threading electronic mail: A preliminary study. *Information processing & management*, vol. 33, no. 2, pp. 209–217.

Malhotra, N. (1982). Information load and consumer decision making. *Journal of Consumer Research*, pp. 419–430.

Marlow, C. (2009 March). Maintained relationships on facebook. Available at: `https://www.facebook.com/note.php?note\_id=55257228858`

Mayer, R.E. (2002). Multimedia learning. *Psychology of Learning and Motivation*, vol. 41, pp. 85–139.

Miller, E.K. and Cohen, J.D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, vol. 24, no. 1, pp. 167–202.

Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, vol. 63, no. 2, p. 81.

Miritello, G., Lara, R., Cebrian, M. and Moro, E. (2013*a*). Limited communication capacity unveils strategies for human interaction. *Scientific reports*, vol. 3.

Miritello, G., Moro, E., Lara, R., Martínez-López, R., Belchamber, J., Roberts, S.G. and Dunbar, R.I. (2013*b*). Time as a limited resource: Communication strategy in mobile phone networks. *Social Networks*.

Muller, T. (1984). Buyer response to variations in product information load. *Journal of applied psychology*, vol. 69, no. 2, p. 300.

Nielsen (2013 June). Internet world stats. Available at: `http://www.internetworldstats.com/stats.htm`

nobuyukinyuu (2012 November). I use twitter lists to store lists of people who i normally wouldn't want to read on my feed but read occasionally. Available at: `http://twitter.com/nobuyukinyuu/status/281041048145375232`

Oh, W. and Jeon, S. (2007). Membership herding and network stability in the open source community: The ising perspective. *Management science*, vol. 53, no. 7, pp. 1086–1101.

O'Reilly III, C. (1980). Individuals and information overload in organizations: Is more necessarily better? *Academy of Management Journal*, pp. 684–696.

Owen, R. (1992). Clarifying the simple assumption of the information load paradigm. *Advances in Consumer Research*, vol. 19, no. 1, pp. 770–776.

Pollack, I. (1953). Assimilation of sequentially encoded information. *The American Journal of Psychology*, vol. 66, no. 3, pp. 421–435.

Ransbotham, S., Kane, G.C. and Lurie, N.H. (2012). Network characteristics and the value of collaborative user-generated content. *Marketing Science*, vol. 31, no. 3, pp. 387–405.

Rapoza, K. (2011 May). China's weibos vs us's twitter: And the winner is? Available at: `http://www.forbes.com/sites/kenrapoza/2011/05/17/chinas-weibos-vs-uss-twitter-and-the-winner-is/`

Ren, Y., Kraut, R. and Kiesler, S. (2007). Applying common identity and bond theory to design of online communities. *Organization studies*, vol. 28, no. 3, pp. 377–408.

Richter, A. and Koch, M. (2008). Functions of social networking services. In: *Proc. Intl. Conf. on the Design of Cooperative Systems*, pp. 87–98. Springer.

Ridings, C. (2010). Online discussion group sustainability: Investigating the interplay between structural dynamics and social dynamics over time. *Journal of the Association for Information*, vol. 11, no. 2, pp. 95–121. Available at: `http://aisel.aisnet.org/jais/vol11/iss2/1/`

rinbrand (2012 November). I don't, but i have heard people put users they don't actually follow in lists to keep them out of the main timeline. Available at: `http://twitter.com/rinbrand/status/281031604401491968`

Romero, D., Meeder, B., Barash, V. and Kleinberg, J. (2011). Maintaining ties on social media sites: The competing effects of balance, exchange, and betweenness. In: *Proc. 5th International AAAI Conference on Weblogs and Social Media.*

Ross, B. (2011 September). Improved friend lists. Available at: `https://www.facebook.com/blog/blog.php?post=10150278932602131`

Sandberg, S. (2009 April). How many friends can you have? Available at: `https://blog.facebook.com/blog.php?post=72975227130`

Schick, A., Gordon, L. and Haka, S. (1990). Information overload: A temporal approach. *Accounting, Organizations and Society*, vol. 15, no. 3, pp. 199–220.

Schneider, S. (1987). Information overload: Causes and consequences. *Human Systems Management*, vol. 7, no. 2, pp. 143–153.

Shannon, C.E. and Weaver, W. (1948). A mathematical theory of communication.

Sicilia, M. and Ruiz, S. (2010). The effects of the amount of information on cognitive responses in online purchasing tasks. *Electronic Commerce Research and Applications*, vol. 9, no. 2, pp. 183–191.

Siegler, M. (2010 August). Zuckerberg: 'guess what? nobody wants to make lists.'. Available at: `http://techcrunch.com/2010/08/26/facebook-friend-lists/`

Simpson, C. and Prusak, L. (1995). Troubles with information overload–moving from quantity to quality in information provision. *International Journal of Information Management*, vol. 15, no. 6, pp. 413–425.

Stepanova, E. (2011). The role of information communication technologies in the arab spring. *Ponars Eurasia*, , no. 15, pp. 1–6.

Stokols, D., Misra, S., Runnerstrom, M.G. and Hipp, J.A. (2009). Psychology in an age of ecological crisis: from personal angst to collective action. *American Psychologist*, vol. 64, no. 3, p. 181.

Swain, M. and Haka, S. (2000). Effects of information load on capital budgeting decisions. *Behavioral Research in Accounting*, vol. 12, pp. 171–198.

To, P.-L., Liao, C., Chiang, J.C., Shih, M.-L. and Chang, C.-Y. (2008 March). An empirical investigation of the factors affecting the adoption of Instant Messaging in organizations. *Computer Standards & Interfaces*, vol. 30, no. 3, pp. 148–156. ISSN 09205489.
Available at:
`http://linkinghub.elsevier.com/retrieve/pii/S0920548907000669`

Toffler, A. (1970). Future shock. *Amereon Ltd., New York*.

Toral, S., Martínez-Torres, M. and Barrero, F. (2010 March). Analysis of virtual communities supporting OSS projects using social network analysis. *Information and Software Technology*, vol. 52, no. 3, pp. 296–303. ISSN 09505849.
Available at:
`http://linkinghub.elsevier.com/retrieve/pii/S0950584909001888`

Tushman, M. and Nadler, D. (1978). Information processing as an integrating concept in organizational design. *Academy of Management Review*, vol. 3, no. 3, pp. 613–624.
Available at: `http://www.jstor.org/stable/10.2307/257550`

Twitter (2012 March). Twitter turns six.
Available at: `http://blog.twitter.com/2012/03/twitter-turns-six.html`

Twitter (2013 April*a*). Twitter api terms of service.
Available at: `https://dev.twitter.com/terms/api-terms`

Twitter (2013 April*b*). Twitter faq.
Available at: `https://dev.twitter.com/docs/faq`

Twitter (2013 April*c*). Twitter's terms of service.
Available at: `https://twitter.com/tos`

Ugander, J., Karrer, B., Backstrom, L. and Marlow, C. (2011). The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*.

Van de Ven, A.H. (1986). Central problems in the management of innovation. *Management science*, vol. 32, no. 5, pp. 590–607.

Velichety, S. and Ram, S. (2013). Examining lists on twitter to uncover relationships between following, membership and subscription. In: *Proceedings of the 22nd international conference on World Wide Web companion*, pp. 673–676. International World Wide Web Conferences Steering Committee.

Wade, M. and Hulland, J. (2004). Review: The resource-based view and information systems research: Review, extension, and suggestions for future research. *MIS quarterly*, vol. 28, no. 1, pp. 107–142.

Wasko, M.M. and Faraj, S. (2005). Why should i share? examining social capital and knowledge contribution in electronic networks of practice. *MIS quarterly*, pp. 35–57.

Wikimedia (2013 June). Wikipedia statistics.
Available at:
`http://stats.wikimedia.org/EN/TablesWikipediansContributors.htm`

Williams, D., Ducheneaut, N., Xiong, L., Zhang, Y., Yee, N. and Nickell, E. (2006). From tree house to barracks the social life of guilds in world of warcraft. *Games and Culture*, vol. 1, no. 4, pp. 338–361.

Wilson, P. (1996). Interdisciplinary research and information overload. *Library Trends*, vol. 45, no. 2, pp. 192–203.

Wu, S., Hofman, J., Mason, W. and Watts, D. (2011). Who says what to whom on twitter. In: *Proceedings of the 20th international conference on World wide web*, pp. 705–714. ACM.

Yamaguchi, Y., Amagasa, T. and Kitagawa, H. (2011). Tag-based user topic discovery using twitter lists. In: *2011 International Conference on Advances in Social Networks Analysis and Mining*, pp. 13–20. IEEE.