PRESENTING TIERED RECOMMENDATIONS IN SOCIAL ACTIVITY STREAMS

A Thesis Submitted to the College of

Graduate Studies and Research

In Partial Fulfillment of the Requirements

For the Degree of Master of Science

In the Department of Computer Science

University of Saskatchewan

Saskatoon

By

WESLEY WALDNER

PERMISSION TO USE

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building
110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5C9

ABSTRACT

Modern social networking sites offer node-centralized streams that display recent updates from the other nodes in one's network. While such social activity streams are convenient features that help alleviate information overload, they can often become overwhelming themselves, especially high-throughput streams like Twitter’s home timelines. In these cases, recommender systems can help guide users toward the content they will find most important or interesting. However, current efforts to manipulate social activity streams involve hiding updates predicted to be less engaging or reordering them to place new or more engaging content first. These modifications can lead to decreased trust in the system and an inability to consume each update in its chronological context. Instead, I propose a three-tiered approach to displaying recommendations in social activity streams that hides nothing and preserves original context by highlighting updates predicted to be most important and de-emphasizing updates predicted to be least important. This presentation design allows users easily to consume different levels of recommended items chronologically, is able to persuade users to agree with its positive recommendations more than 25% more often than the baseline, and shows no significant loss of perceived accuracy or trust when compared with a filtered stream, possibly even performing better when extreme recommendation errors are intentionally introduced. Numerous directions for future research follow from this work that can shed light on how users react to different recommendation presentation designs and explain how study of an emphasis-based approach might help improve the state of the art.

ACKNOWLEDGMENTS

I first must acknowledge the irreplaceable contribution of my graduate supervisor, Dr. Julita Vassileva. With her guidance and through our regular discussions about the field of social computing in general, I was able to search for unturned stones while exploring my interests in this area of research. I would not have made it to this point without the freedom to study something that was meaningful to me. I would also like to thank her for seeing something in me while I was still an undergraduate student and for giving me this opportunity that I would not have otherwise considered.

Thanks go out to the other members of my committee—Dr. Winfried Grassmann, Dr. Regan Mandryk, and Dr. Chris Zhang—for their valuable input, corrections, and particularly for making sure that I kept the big picture in mind as I completed my studies and wrote this manuscript.

Finally, I need to acknowledge the love and support of my family and of my wife, Jennie, in particular. Thank you for keeping me focused on the goal, for your encouragement, and for always knowing that I would make it to the end.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

Recent advances in technology have created opportunity and demand for the development of innovative approaches to connecting people from around the world using online social networks. Typically these innovations have started as novel adaptations of existing communities. Facebook, for example, sought to reproduce the social experience of college online [Lynley & Edwards 2014]. Avatar-based chat platform IMVU initially envisioned users importing their existing contacts from other chat services to interact in 3D before eventually learning that their users, gifted with the choice of anonymity and new, fanciful physical identities, preferred to meet new people [Ries 2011]. Similarly, Twitter was born out of the idea of sending text messages to small groups of people [Sagolla 2009], presumably whom the sender already knows.

However, many of the new ways in which we communicate with others in these social networks grew organically out of affordances of the data-driven online platform. In the online version of social interaction, no longer did we need to physically sort and store letters received from distant friends and family members, remember or write down the content of spoken conversations, or even ask as many probing questions such as “What’s on your mind?” or “What’s happening?”—the platform would ask them for us.[1] All of our personal updates and communications with other members of our network are stored as raw data to be sorted, categorized, and manipulated in any way the service sees fit. While the connectedness of the Internet did create some possibilities that never before existed, such as sending a picture instantly to someone on the other side of the world, online social networks specifically innovated in the area of how communications are presented to users. With all of these data so readily available for

[1] At the time of writing, these were placeholders for status update inputs on Facebook and Twitter, respectively.

consumption, it was a simple step technologically to present in bulk, as a one-stop shop for interesting information about friends and family and anyone else one might want to connect with online. Such collections of social interaction data were both a natural next step and an essential innovation for social networking sites.

1.1 Social Activity Streams

Facebook's News Feed, which displays status updates and activity from members of a user's friend network, was not introduced until more than two and a half years after the site's initial launch [Sanghvi 2006]. Despite being met with immediate opposition [Zuckerberg 2006], it has long been the focal point of the landing page for Facebook users, and did not take long to become one of the network's most well-known and heavily-used features. Twitter's home timeline is a similar feature for a different type of network, one that might better be called an "information network" [Lapowsky 2013]. The home timeline collects Tweets—the fundamental 140-character text (and media) building blocks of Twitter—from all of the users in one's network and displays them strictly in reverse chronological order, with some exceptions made in some client applications for conversational Tweets displayed in-context. Such features are commonly called "social activity streams" and can be found in the most popular online social networks today, including Google+, LinkedIn, Instagram, Tumblr, Flickr, MySpace, Hi5, QZone, VKontakte, and Sina Weibo.

A user's social activity stream comprises the actions of a particular set of users. However, it may be more useful to distinguish between three primary ways in which content comes to appear in one's stream: curation, personalization, and evolution. These sources of change to the content in one's social activity stream and their primary actors are summarized in Table 1.1.

Curation. Conceptually similar to the more common term "customization," curation refers to any user-initiated action intended to control the content of one's stream. This includes changing the members of one’s network, keyword or other specific content-based filtering, or listening for certain topics either from within or outside of one’s network. Facebook’s friend request and Twitter’s follow features are examples of activity stream curation.

Personalization. Personalization, on the other hand, refers to any system-initiated action that affects what a user sees in her stream. An online social network may attempt to show users more relevant content by rearranging the stream so that interesting content is shown first and uninteresting content is shown later or completely hidden. How content is deemed interesting or uninteresting depends on the system, but the process would likely involve machine learning techniques that take into account attributes of the content one most often interacts with.

Evolution. Finally, evolution refers to changes to one’s social activity stream that originate from within the network itself. Examples of stream evolution include users changing posting frequency, general style or topic, or privacy settings so that the activity stream content deviates from expectations or from assumptions made during the curation process.

Table 1.1: Sources of change to social activity stream content

Source	Actor(s)
Curation	The user
Personalization	The system
Evolution	Other users is the user’s network

Curation and evolution are both entirely user-controlled, albeit by different users, at least to the extent that the system provides the necessary functionality. Designing systems that enable and encourage users to engage socially is an interesting problem, but designing intelligent systems that are helpful while being minimally intrusive is perhaps a more difficult one. Black-box personalization features such as Facebook’s rearrangement of News Feed content are finetuned to increase measures of user engagement rather than user satisfaction[2], and tweaks to the system remain outside of user knowledge and control. One can debate which of these measures would be more virtuous to optimize and even how much difference there really is between them, but the benefits of flexibility and potential harms of information overload make personalization efforts essential.

As more and more information about our world is disseminated via these streams, and as a greater number of important sources of information emerge, this is a problem that is only increasing in importance. Today the top online social networks are massively popular, and their creators tremendously successful. Facebook is the largest and most valuable social media company in the world and ranks 10th among all companies [Kilgore 2015]. The company earned $12.47 billion in revenue during 2014, 58% higher than their total in 2013 [Facebook 2015]. During 2014 they boasted an average of 890 million daily active users and 1.39 billion monthly active users. As of March 24, 2015 their market cap was $238.78 billion [Yahoo Finance 2015]. In December 2014, Instagram, which was bought by Facebook in 2012 for $1 billion [Upbin 2012], was valued at $35 billion and had 300 million monthly active users [Alba 2014], while WhatsApp, which Facebook paid $19 billion to acquire in 2014 [Facebook 2014], hit 700 million monthly active users in January 2015 and processes over 30 billion messages every day [Sherr 2015]. Twitter, on the other hand, earned $1.403 billion in revenue in 2014, which was a 111% increase over 2013 [Twitter 2015]. The company, which was reported to have a market cap of $23 billion in December 2014, reported 182 billion timeline views in 2014's fourth quarter among its 288 million average monthly active users. In addition, Facebook users spend an estimated 39 minutes per day on the site on average [eMarketer 2014], and Ipsos has reported

[2] Facebook News Feed Team member, personal communication, October 10, 2014

that the average online social network user spent 3.6 hours per day socializing online [Ipsos 2013] and that millennials in particular spend 5 hours daily consuming user-generated content [Ipsos 2014]. Clearly this is an important topic at this point in history, and while some might feel that it would be a worthy goal to decrease these numbers, there is at least some value in optimizing that time.

If people spent over 211 billion hours on Facebook in 2014, how much of that time was spent contributing and consuming interesting content, and how much was spent simply searching for it? What about Twitter, where the average user follows about 100 others [Beevolve n.d.] and over 500 million total updates are posted every day [Twitter n.d.]? In an effort to make their services “simpler and more relevant” [O’Brien 2014] or to increase user engagement, several online social networks have abandoned a strictly chronological presentation of social activity streams. After all, the only thing any user is guaranteed to see is what is presented at the start of his session, so it makes sense to showcase the most engaging content first. This is the case, as mentioned above, with Facebook’s News Feed. Recommender systems can be and have been used to personalize streams and pick out the most and least interesting content for a particular user [Wang et al. 2010]. However, filtering and rearranging streams using recommender systems have a number of inherent and/or common disadvantages:

- Recommenders can never be perfectly accurate all of the time, so some interesting content will be missed and some uninteresting content included.
- Recommenders tend to filter away anything that cannot be predicted as interesting based on the users’ past behaviour, and this can cause the user to miss potentially very important information (the “filter bubble” problem [Pariser 2011]).

- Often no indication is given that any content was hidden, and there is no way to view and correct specific instances of content filtering.
- If the hidden content *is* retrievable, it is often either difficult to perceive or not indicated where the content belongs in the stream chronologically. This logically leads to an inaccurate perception of stream activity and relative quality of content.

Filttr and TweetDeck are Twitter applications that allow users to filter their streams by certain keywords, phrases, or authors, and SocialFixer is a similar service created for Facebook. However, these services all require users explicitly to define filtering criteria, and while Filttr can show where posts have been filtered in the stream, each one must be expanded individually in order to be read. In addition, showing a filtered, uniform stream offers no flexibility for different reading situations: the user may want to spend more or less time reading than it would take to consume all unread items in her stream. As a result, it is unlikely that the *n* content items the user consumes are the *n* most interesting ones, even if the filtering recommender is perfect. Besides the direct drawbacks listed above, stream filtering can ultimately result in decreased trust [Nagulendra & Vassileva 2013], which can make it less likely for users to return to the system for continued use [Pu & Chen 2006].

Information streams are not unique to social media. In all sorts of domains, order of presentation is a powerful dimension to utilize in order to convey information. In many information streams, it is useful to display items in chronological sequence to preserve context. This is especially important when the items combine to tell a story, such as in conversations, commentary, or real-time accounts of events. Thus it would seem beneficial to be able to recommend an item within a stream without uprooting it from its original context. While stream filtering certainly can help with information overload, the side effects are both undesirable and avoidable, and stream efficiency still has much room for improvement.

1.2 Research Questions

This research introduces a novel method of presenting recommendations in place in information streams in general and Twitter timelines in particular that addresses the information overload problem without filtering out any content. Along the way we will seek to answer the following questions:

1. Is the method immediately intuitive? Can it be used as intended to improve reader efficiency and satisfaction, and how easy is it to use?
2. Do users gain additional insight by exploring a two-dimensional stream visualization that highlights interesting content at a glance?
3. When evaluating recommendation presentation, we typically want to consider the effectiveness of the design separate from a recommender algorithm. How persuasive is the design itself when removed completely from a recommender system?
4. How does this method compare with a current standard on the completely subjective measures of perceived accuracy and trust?
5. In this presentation method and in social activity streams in general, is it worse to recommend uninteresting content or to hide something interesting from the user?

I will look at each one of these questions separately over the course of several user studies across Chapters 3 through 5. But first I will discuss the relevant literature that gives insight into what recommender systems do and why they are necessary, how users experience them, what properties this new presentation design should adopt, and previous work by other researchers and designers that may serve as a foundation.

CHAPTER 2
LITERATURE REVIEW

In order to develop a way to recommend items in-stream we must understand the human factors of recommender systems, including what affects users' trust and perception of recommender accuracy. Knowing about the types, applications, and evaluation metrics of recommender systems is also essential before setting out to build one. Additionally, much research has been done on social information visualization and recommendation presentation that can serve as a starting point for an appropriate presentation method. We want to know whether a suitable method has been discovered in a different field that can be adapted to social activity streams. Failing that, we want to bring together knowledge from the various related fields to develop a method that should be effective according to the research that has already been completed. This chapter summarizes the relevant literature on information overload, recommender systems, information visualization, and recommendations within social networks in particular.

2.1 Information Overload

For the vast majority of human history, relationships have been limited to those people who have face-to-face interaction or, in some cases, communication by letter-writing. It is only recently that technology has allowed this restriction to be lifted to permit virtually real-time communication by any two people anywhere in the world with access to an Internet-connected device. While a person could now conceivably befriend a thousand others, what is their actual ability to maintain a relationship with this number of people? According to Robin Dunbar, the number of people with which we can sustain meaningful relationship has historically been about 150 and is constrained by the number of neurons in our neocortex [Dunbar 1992]. Further, that number has not changed as a result of technological advancements [Dunbar 2011]. However, while we may not consider all of our online social contacts to be “friends”, we are still influenced by them. This is just one of many sources of information overload that we encounter online.

There are references to information overload brought on by the World Wide Web at least as early as 1994 [Maes 1994], and this was an especially big problem in the pre-Google days [Berghel 1997]. But while the web has become increasingly more organized and navigable, it also contains more and more user-generated content every day that continues to raise the risk of overload. Our information processing limitations have been well-noted since at least 1956 when George Miller wrote about the “magical” number seven, plus or minus two [Miller 1956], referring to the number of chunks of data we can hold in working memory at any given time. Since we have limited capacity for information processing [Thorson et al. 1985], the demands can often exceed our capacity [Rogers & Agarwala-Rogers 1975; Schick et al. 1990] or time constraints [Grisé & Gallupe 1999]. The negative effects of information overload are also well-documented. Potential consequences include “information fatigue” [Oppenheim 1997], cognitive strain [Malhotra et al. 1982], decreased decision-making accuracy [O’Reilly 1980], and numerous related anxieties [Bawden & Robinson 2009].

How best to cope with information overload depends on the domain, the type of information, and the data-processing requirements. Losee proposed ranking electronic messages by expected importance or economic worth in order to minimize overload [Losee 1989]. This strategy involves reordering items and may not be feasible in all applications, but it is a reasonable solution in others. Simon proposed the concept of “satisficing” as an alternative to “maximizing” or “optimizing” in order to deal with volumes of information for which a thorough consumption and decision-making process would be unrealistic [Simon 1956]. However, according to Schwartz et al., some people are susceptible to decreased satisfaction when employing such strategies [Schwartz et al. 2002], a phenomenon that became popularly known as “the paradox of choice”. A potential solution would be somehow to identify which items might be relevant to the user and present only those, hiding the irrelevant items to reduce information overload. This is the primary function of recommender systems.

2.2 Recommender Systems

2.2.1 Types of Recommendations

Product recommendations. Many different types of systems have made use of recommenders in order to increase user satisfaction and to address the information overload problem. Some of the earliest such systems were independent entities that recommended products that people needed to go elsewhere to try or buy [Burke 2002]. Recommender systems existed before the rise of the World Wide Web in the 1990s, but the increasing popularity of e-commerce websites during this period brought increased interest in putting the wealth of new user data to use in new ways. In addition, as the number of available items grew, so did the need to provide assistance to users to help them find products they would like.

Subscription-based recommendations. The need to assist these users grew over the next decade as recommenders for entertainment services began to gain popularity. With services such as Netflix, new sources of entertainment are being suggested rather than new products, and this is an important distinction.[3] With subscription services, the goal is not to guide users toward a product to purchase as with product recommendations, but to keep customers who are already subscribed satisfied and interested so that they will not only recommend the service to others by

[3] Recommending a subscription service to new prospective customers is considered a product recommendation, while recommending specific products or services only available to subscribed members is considered a subscription-based recommendation.

word of mouth, but also so that they will not cancel their own subscriptions to the service. As a result, subjective measures such as user satisfaction would become increasingly more important in evaluating recommender systems. These subscription-based systems are closely related to product recommenders, and, together they would drive the majority of research in the area through the 1990s into the early 2000s.

Social and information recommendations. Another form of recommender system that was present in the beginning but has begun to emerge much more recently is related to information. Recommendations within social activity streams is a recent trend, but systems operating on similar concepts were being developed in the early 1990s as well, such as Tapestry (email lists) [Goldberg et al. 1992], GroupLens (Usenet articles) [Resnick et al. 1994], and DailyLearner (news) [Billsus & Pazzani 2000]. Today, social networking sites use recommender systems to filter activity streams and news feeds and to recommend potentially interesting sources of information (e.g., other nodes in the network). This type of recommendation, particularly when dealing with streams of information that are always changing and are targeted to perhaps a select few with special knowledge, must be examined differently than product recommendations. These types of social recommenders will be examined in Section 2.6.

2.2.2 Types of Recommenders

Nearly all of the research on recommender systems as well as existing implemented systems can be categorized into one of three approaches: content-based, collaborative, and hybrid.

Content-based recommenders. Content-based recommender systems use explicitly-stated item features and user preferences as a basis for making recommendations. These systems commonly compare items to other items based on their features, either from textual analysis or gathered from metadata in order to determine similarity between items. Thus if a user likes item

A, the system can then confidently recommend an item B that has many features in common with item A. Stated user preferences can also be matched with item features as an additional source of data even before any previous ratings have been made. However, if items have not been given any metadata and do not lend themselves naturally to feature-based analysis (e.g., videos), content-based systems will not be able to recommend them. Another limitation of content-based recommenders is their inability to discover new areas of interest for a user [Lü et al. 2012]: they will only recommend what matches their existing preferences or items they have liked in the past.

Collaborative recommenders. Whereas content-based methods consider similarity between items, collaborative filtering methods find statistical similarity between users based on their ratings of items. This method is widely used because it avoids many of the problems inherent to content-based systems. Collaborative methods are able to recommend any type of item, even without explicitly-stated features or textual content, and they can recommend items from completely new genres or categories that a user may not have been previously exposed to [Adomavicius & Tuzhilin 2005]. This may result in more serendipitous discoveries, but users may also require more explanation about the recommendations or they will be unwilling to try them [Indratmo 2010]. Because collaborative systems rely on user ratings to produce recommendations, they will invariably face what has been called the cold-start problem [Lü et al. 2012]. In these cases, the system will not have enough information to provide recommendations to this particular user. Some recommender systems also use a measure of trust instead of solely relying on user similarity to weight ratings for collaborative methods [Lü et al. 2012].

Hybrid recommenders. Due to the limitations of both content-based and collaborative recommendation methods, most current recommender systems employ a hybrid approach whenever possible. Normally the system will be primarily based on the collaborative filtering approach but will be supported by content-based methods to overcome the cold-start problem [Burke 2002]. Since these systems are still based on content-based and collaborative techniques, the primary research into hybrid systems deals with finding the most effective ways to apply the different methods into a combined score to produce accurate recommendations. Burke [Burke 2002] summarizes the different approaches commonly used in hybrid systems to calculate final predictions.

2.3 Evaluating Recommender Systems

2.3.1 Objective Metrics

Recommender systems have traditionally been evaluated objectively by measuring how accurately they can predict user ratings of new items. The two most fundamental metrics are *precision* and *recall* [Herlocker et al. 2004], which come from the field of information retrieval. If each item in a superset can be classified as either *relevant* or *irrelevant*, then precision and recall can be conceptualized with the Venn diagram in **Figure 2.1**.

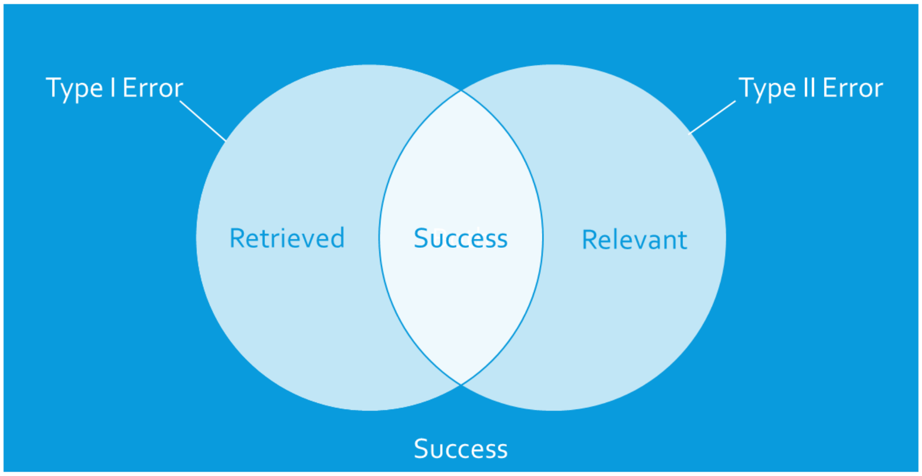


Figure 2.1: Venn diagram of cases in information retrieval

In Figure 2.1, the circle on the left represents the set of items predicted to be relevant by the recommender, while the circle on the right represents the set of items the user would actually consider relevant. Retrieval can be considered successful if a relevant item is retrieved or if an irrelevant item is not retrieved. Then, if A is the set of retrieved items and B is the set of relevant items, we have the following formal definitions of precision and recall:

$$precision = \frac{|A \cap B|}{|A|} \tag{2.1}$$

$$recall = \frac{|A \cap B|}{|B|} \tag{2.2}$$

Essentially the recommender system can only control the size and contents of set A and should try to match set B as closely as possible to maximize these two metrics. Precision and recall can be measured by obtaining a set of ratings from the user and dividing it into a training subset, from which the system learns the tendencies and preferences of the user, and a test subset, against which predicted ratings are evaluated for accuracy. After training, the system will select a subset of items from the test set predicted to be relevant and compare them to the items in that set that the user actually rated highly. Precision and recall can then be used, if desired, to calculate single accuracy metrics such as F1 [Sarwar et al. 2000] and Mean Average Precision [Herlocker et al. 2004]. Precision may be more important than recall in recommendations involving items that require a large commitment of time of resources [Gunawardana & Shani 2009; Tyler & Zhang 2008], but this may not be the case in contexts where irrelevant retrievals are more easily ignored or forgiven. These methods are intuitively effective at measuring a recommender's accuracy, but human factors are important to consider as well when evaluating the overall value of such a system, as is discussed in Section 2.3.2.

2.3.2 Explanations

The straightforward approach outlined in Section 2.3.1 for objectively evaluating recommender systems is oversimplified. In real-world situations, the items being recommended are new to users, and they haven't yet formed opinions on them. They can be influenced by the way they believe the recommender to be working, by how much they trust it, and by how these recommendations are presented. Recently, designers have begun to realize the importance of *explanations* in recommender systems. Explanations are any ways in which the recommender communicates to the user about how or why certain recommendations are given, or by giving a level of confidence in a recommendation. Tintarev and Masthoff [Tintarev & Masthoff 2007] identified the following seven aims of explanations in recommender systems:

- *Transparency*: The user's level of understanding about why or how the system does what it does
- *Scrutability*: The ability for users to correct the system
- *Trust*: The user's confidence that the system will do what it is supposed to do
- *Persuasiveness*: The ability for the system to change the user's mind about an item somehow
- *Effectiveness*: The extent to which the system helps users make good decisions
- *Efficiency*: The extent to which the system helps users make decisions quickly
- *Satisfaction*: The extent to which the user enjoys using the system

These seven aims are all dimensions on which recommender systems can be evaluated and are all distinct from the objective measures of precision and recall. Any recommender system benefits from attaining a higher level of any of these qualities.

2.4 Trust and Persuasiveness

Much of the research on users' trust in recommender systems links it with transparency [Tintarev & Masthoff 2010; Sinha & Swearingen 2002] and focuses on using explanations to drive trust [Pu & Chen 2006; Tintarev 2007; Friedrich & Zanker 2011; Wang et al. 2010]. Users have been shown to have greater distrust in a system that hides certain information from them if they do not understand or cannot see why [Nagulendra & Vassileva 2013]. However, users can trust a recommender without it offering any features to explicitly enhance transparency and without offering explanations for its recommendations. Trust may be directly influenced simply by how accurate users perceive the recommender to be [McNee et al. 2003], and this perception can be driven by other unrelated factors such as persuasiveness. Trust levels can also vary depending on the users' knowledge of the underlying recommendation mechanism. For example, users lose trust as a result of inaccurate recommendations faster when the recommendations are personalized [Harman et al. 2014].

Since users are more likely to return to recommenders they trust [Pu & Chen 2006] and persuasiveness can result in increased acceptance of recommended items [Cremonesi et al. 2012], it would be beneficial to build recommenders that are persuasive without betraying the users' trust. Though users have been shown to detect systems that manipulate predictions, they will also tend to rate items more similarly to predictions shown to them by a recommender, regardless of its accuracy [Cosley et al. 2003]. Adomavicius et al. found a similar result in a consumer study, in this case determining the phenomenon to be caused by anchoring effects [Adomavicius et al. 2013]. Objective measures of recommender accuracy are also important, but perhaps not as important as a user's perception, which is likely to be affected more greatly by factors related to user experience [Cremonesi et al. 2012]. Indeed, more persuasive recommenders are likely to have higher ratings of user satisfaction [Nanou et al. 2010].

2.5 Information Visualization

Information visualization involves mapping data attributes to visual attributes [Deller et al. 2007]. Depending on the application, information visualization can make trends or complex relationships immediately clear, separate effects of multiple variables, draw attention to particularly relevant data, or use metaphors to give meaning to the story behind a set of data. Cleveland and McGill [Cleveland & McGill 1984] identified an inexhaustive list of ten "elementary perceptual tasks" (shown in Figure 2.2), each of which represents a visual attribute one might use to visually code a single data attribute.

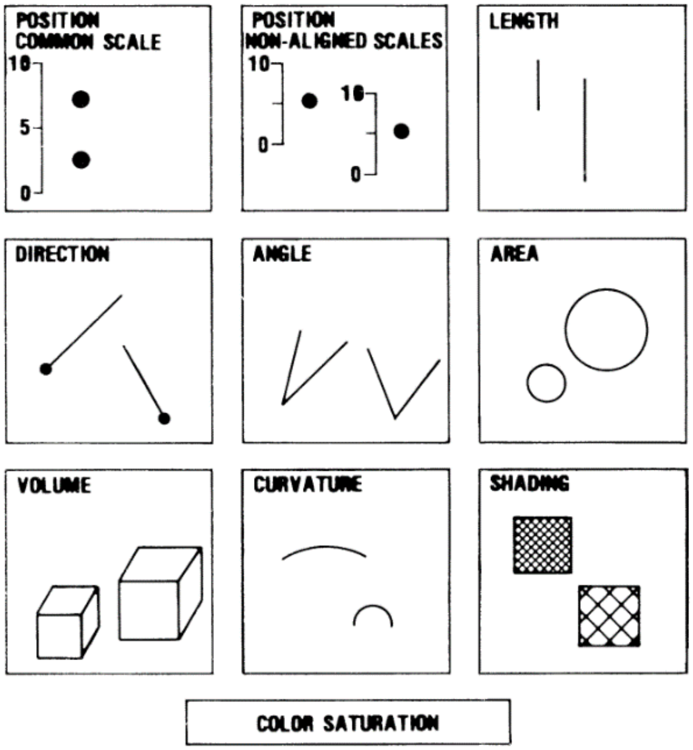


Figure 2.2: Ten Elementary Perceptual Tasks
(reproduced with permission from [Cleveland and McGill 1984])

More specifically relevant to this research is the concept of visual search. We want to display many single items and make it especially easy for users to find items that are recommended to them. Parallel search [Treisman & Gelade 1980], where an object immediately stands out due to one or more basic visual features, is much faster than serial search, where each item must be examined one at a time. In order to speed visual search, visual features should be chosen that can be detected by preattentive vision, which include size, colour, and intensity, among other attributes [Deller et al. 2007]. As an example of this technique, Suh et al. used enlarged, highlighted text to draw attention to certain keywords in web pages [Suh et al. 2002] (shown in Figure 2.3).

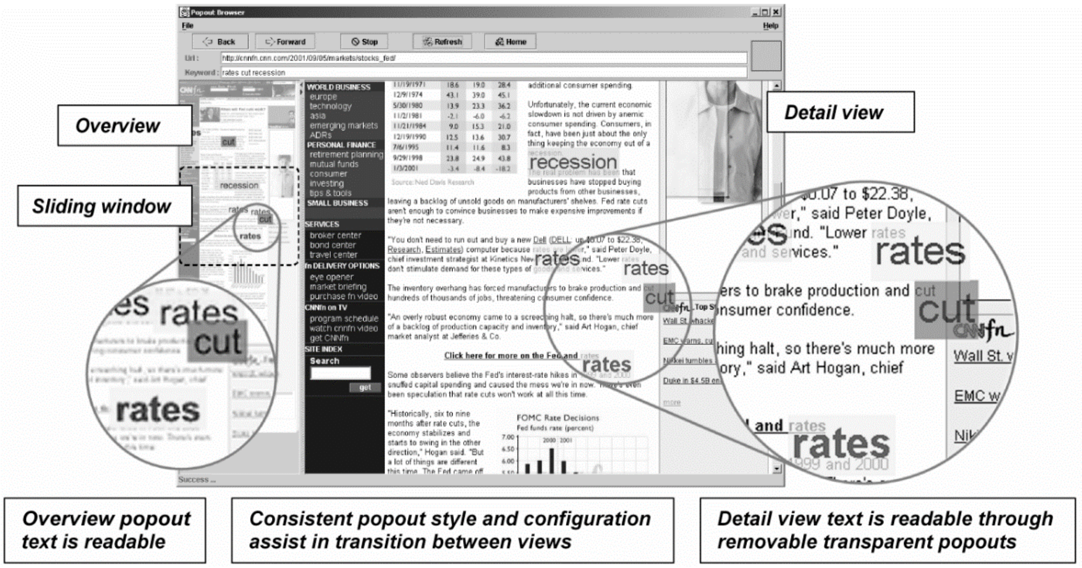

Popout Browser
Back
Forward
Stop
Refresh
Home
Keyword: rates cut recession
Overview
Sliding window
Detail view
rates
cut
recession
Overview popout text is readable
Consistent popout style and configuration assist in transition between views
Detail view text is readable through removable transparent popouts

Figure 2.3: Popout Prism Screenshot
(reproduced with permission from [Suh et al. 2002])

Additional examples of information visualization serving enhancement of social awareness in online communities and as explanation mechanisms for social recommender systems can be found in Section 2.6.

2.6 Social Recommendation

The issue of recommendations within Twitter timelines in particular from a filtering approach has been tackled by Sriram [Sriram 2010], among others. In addition to a naïve Bayes classifier, C4.5 decision tree and sequential minimal optimization algorithms were used to classify Tweets into categories such as “news”, “opinion”, “deals”, and “events”. Support was also added for user-defined classes. Wang et al. [Wang et al. 2010] also studied recommendations of updates across both Twitter and Facebook, focusing only on recommendation effectiveness without suggesting filtering as a solution to the information overload problem. They studied the value of textual and non-textual features in accurately predicting whether an update will be liked, disliked, or neutral. Machine learning algorithms such as decision trees, support vector machines, Bayesian networks, and radial basis functions were compared for performance. Chorley et al. studied users’ consumption decision-making in Twitter by presenting different metadata without revealing the content of the Tweets. They found that qualitative metadata, particularly pertaining to the author’s identity, drove decision-making more than quantitative metadata such as retweet counts, which was effective in comparison to other quantitative metadata [Chorley et al. 2015]. These studies give a good idea of what features to use in Twitter recommenders based on a machine-learning approach.

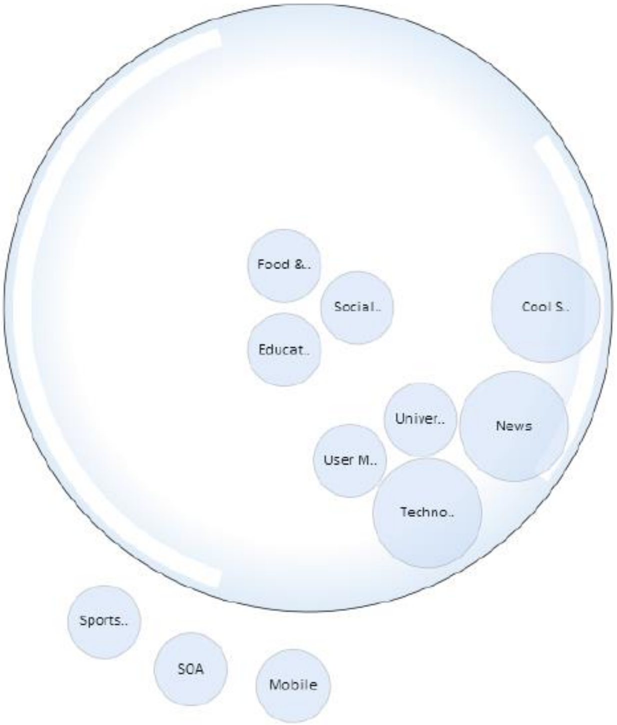


Figure 2.4: Madmica’s Filter Bubble Visualization
(reproduced with permission from [Nagulendra & Vassileva 2013])

Some of the drawbacks of information filtering in social streams have been identified and addressed by Nagulendra and Vassileva, such as decreased user trust [Nagulendra & Vassileva 2013]. The “Filter Bubble” visualization in social networking site Madmica[4], shown in Figure 2.4, allows users to view which updates have been hidden, and it also gives control to show or hide posts on certain topics from certain users. However, it remains difficult to get a sense of where posts belong in the content of the social activity stream without restoring them to a visible status. This is likely not as important in Madmica as it is in Twitter, where updates may quickly become less relevant as they age.

[4] http://madmica.usask.ca/

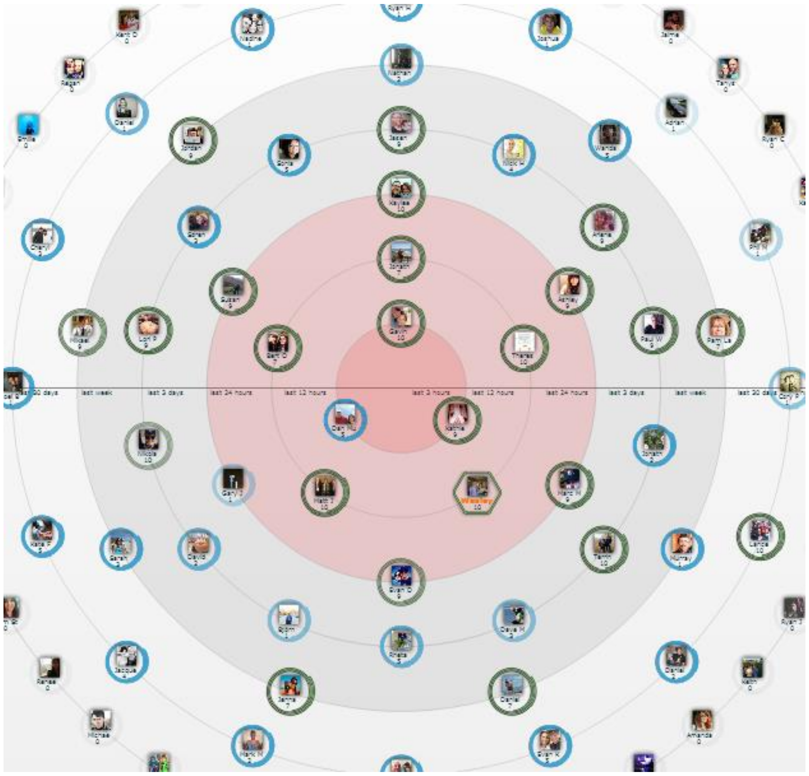


Figure 2.5: Rings Screenshot

Rings[5] [Shi et al. 2014] (shown in Figure 2.5) is a visualization system for Facebook friend networks that codes recency, quantity of recent posts, and average social impact of those posts. The system successfully increases user awareness of lurkers and the most active recent contributors within one's own network. Rather than emphasizing which individual posts were most impactful, Rings focuses on the users and their relative activity levels within the friend network. The information that the visualization provided was demonstrated to be interesting for users and was not easily discoverable through Facebook's own default interface. The main drawback of the design is that it was not necessarily useful for popular Facebook functions such as everyday social activity stream consumption.

[5] http://rings.usask.ca/

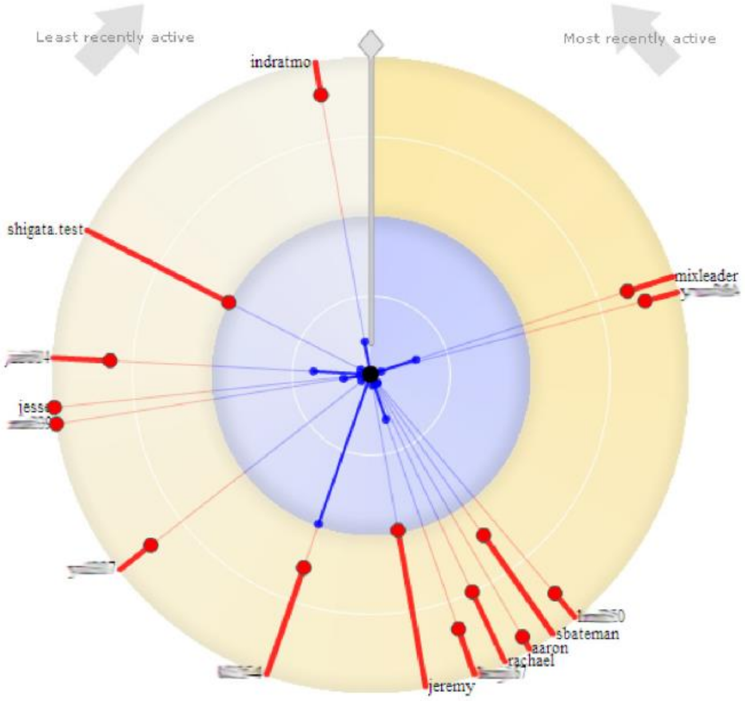
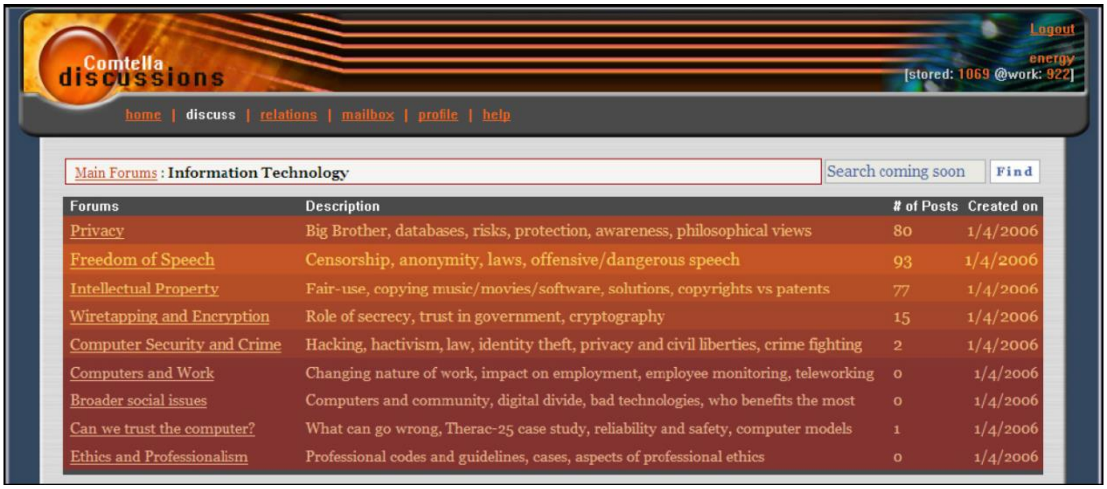


Figure 2.6: KeepUP Recommender System
(reproduced with permission from [Webster & Vassileva 2007])

KeepUP [Webster & Vassileva 2007] (shown in Figure 2.6) visualizes a user's network of influence in a RSS recommender system that allows for user interaction. While it does primarily model the network rather than the content, it also tracks topics that each user has commonly liked or disliked. The transparency provided and affordance of user control over others' influence on recommendations allows users to shape their own filter bubbles.


Comtella
discussions
Logout
energy
[stored: 1069 @work: 922]
home | discuss | relations | mailbox | profile | help
Search coming soon
Find

Main Forums : Information Technology

Forums	Description	# of Posts	Created on
Privacy	Big Brother, databases, risks, protection, awareness, philosophical views	80	1/4/2006
Freedom of Speech	Censorship, anonymity, laws, offensive/dangerous speech	93	1/4/2006
Intellectual Property	Fair-use, copying music/movies/software, solutions, copyrights vs patents	77	1/4/2006
Wiretapping and Encryption	Role of secrecy, trust in government, cryptography	15	1/4/2006
Computer Security and Crime	Hacking, hactivism, law, identity theft, privacy and civil liberties, crime fighting	2	1/4/2006
Computers and Work	Changing nature of work, impact on employment, employee monitoring, teleworking	0	1/4/2006
Broader social issues	Computers and community, digital divide, bad technologies, who benefits the most	0	1/4/2006
Can we trust the computer?	What can go wrong, Therac-25 case study, reliability and safety, computer models	1	1/4/2006
Ethics and Professionalism	Professional codes and guidelines, cases, aspects of professional ethics	0	1/4/2006

Figure 2.7: Comtella-D Discussion Forum Screenshot
(reproduced with permission from [Webster & Vassileva 2006])

Webster and Vassileva's work in the Comtella-D online discussion forum [Webster & Vassileva 2006] (shown in Figure 2.7) presents recommendations using emphasis rather than filtering. In their system, recommendations are made collaboratively by and for other members of the community. The most recommended posts are shown in a brighter colour and with larger text in order to be visually attractive and more noticeable. The chosen colours in this case fit with an "energy" metaphor, with the more recommended posts displaying more life while the least recommended posts have a dull and lifeless appearance. This method of collaborative recommendations works well in a closed community settings, but it is not replicable in the vast, open world of Twitter.

2.7 Summary

Though content-based recommenders have some drawbacks, such as an inability to recommend certain types of content and difficulty recommending serendipitous content, these problems are non-issues for this line of research. Besides certain types of Tweets with media content, social activity stream content in Twitter comes with sufficient metadata and lends itself well to content-based analysis. In addition, we are only trying to recommend items that are already flowing out of a user's network. The filter bubble is still a concern, but serendipity is not the primary goal here. Using a recommendation presentation method other than filtering will also go a long way towards eliminating that problem.

Since this research deals with recommendation presentation methods, objective evaluation metrics are less important to us than subjective metrics that can actually be impacted by the way content is shown to users. There is no reason that the presentation would not be more or at least equally effective with a more objectively accurate recommender under the hood. As such, the applicable metrics for this recommendation context are trust, persuasiveness, and satisfaction. Trust and persuasiveness are more tangible goals to strive for, and both positively influence user satisfaction as well.

Research in the field of information visualization and social recommendation provide inspiration for potential presentation methods. Interactive social recommenders such as Rings and Comtella-D show how recent and relevant user activity and content, respectively, can be emphasized using visualization. In order to be successful, any possible solution must be able to support standard activity stream consumption both in terms of features and scalability.

The next chapter describes the proposed presentation method in detail and presents the methodology and results of an experiment designed to evaluate its usability and user satisfaction.

CHAPTER 3
TIERS OF EMPHASIS IN TWITTER TIMELINES

This chapter describes the proposed approach to presenting recommendations within social activity streams and presents the results of a pilot user study designed to evaluate the feasibility of the design as well as levels of user satisfaction and perceived recommender accuracy.

3.1 Proposed Approach

The basis of this research is an approach to presenting recommendations in social activity streams without removing items from their original context, preserving chronology and facilitating selective consumption without filtering. I have called this visual recommendation presentation design "ViTA" (Visualizing Twitter Activity) and incorporated it into a web-based application that was developed to test this approach in a real-world social activity stream—namely, Twitter's home timeline. I chose Twitter because of the large number of public updates available and the ease with which one can follow many other users to build a rich content stream.

Consider the example social activity stream represented in Figure 3.1, where updates are presented as vertically-stacked rectangles. This user will not know which updates are interesting until he reads through them. As a result, much of this time that was intended to be spent reading interesting updates will effectively be wasted.

3.1.1 Design

The idea behind ViTA is simple: show everything, but make content likely to be more valuable to the user more visually prominent. Three distinct visual tiers are used, as shown in Figure 3.2, which is a conceptual depiction of the same stream from Figure 3.1 with ViTA's recommendation presentation applied. The most highly recommended items are given the largest rectangles, largest font, and noticeable colour. Yellow was chosen as the colour since the human eye is most sensitive to light between 540 and 570nm (green-yellow) [Thomson & Wright 1947]. The least highly recommended items (those which are likely to be of less value to the user) are shown in smaller rectangles, using a smaller font, and without any colour. The second tier occupies the middle ground in all respects.

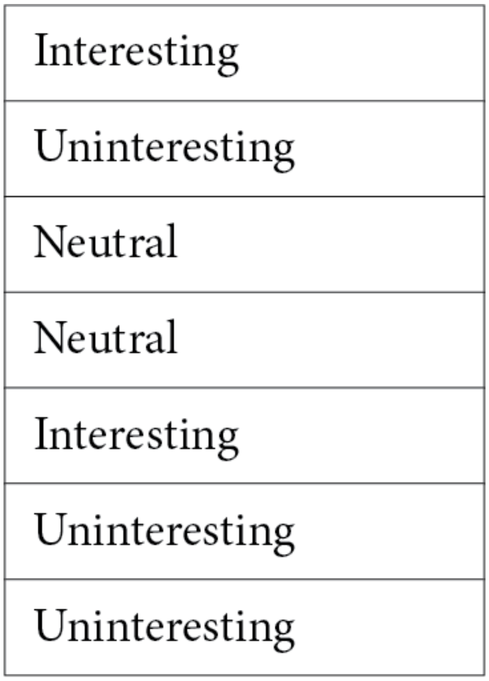


Figure 3.1: Representation of an untreated Twitter timeline

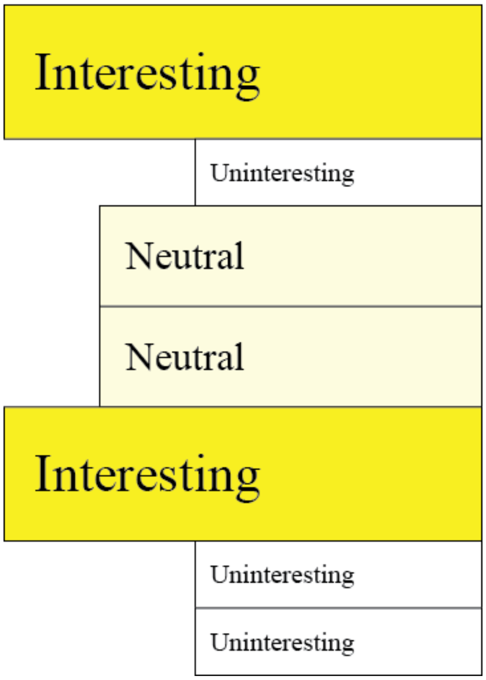


Figure 3.2: Representation of a Twitter timeline in ViTA

I chose to utilize three tiers to keep the system very simple but to allow for more than an oversimplified dichotomous differentiation. Having more tiers is advantageous because the user can move on from the top tier without needing to consume the most uninteresting content. Assuming equal distribution among tiers, offering a greater number of tiers provides the user with a greater degree of control over the quality of content she consumes. However, having too many tiers would prevent each tier from standing out in relation to the others, making it more difficult for the user to restrict his reading to a single tier.

Depicting the three distinct tiers can be thought of as a problem in visualization ordinal data. Mackinlay [Mackinlay 1986] provides a ranking of visual properties that are effective for such a task, and two of the top three have been chosen to enhance presentation of the three tiers: position and colour saturation. The colours and sizes of the objects remove any ambiguity about which tier should draw the user's focus. Besides offering a clear separation between tiers, the horizontal offsets make it easier for users to read through a particular tier while ignoring the less interesting ones. The horizontal alignment of Tweets within a tier facilitates chronological consumption of a particular tier, and the user's eye can return to a single reference point and scan up or down to find other Tweets within the same tier.

The design itself is system-agnostic: it could be applied to any collaborative or content-based recommender. It is also simple enough to be adapted to a number of different purposes. For example, the tiers need not represent a general recommendation score; they may indicate any type of categorization on a scale that could be mapped to three ordinal values. Further, it is not necessary that social activity streams be the platform: this design could be applied to any information set where chronology and differentiation between items on at least an ordinal scale both have value. However, social activity streams are a very natural application and the one we will explore exclusively in our experiments with the system.

3.2 One-Dimensional Stream Pilot Study

3.2.1 Goals

The primary goal of this evaluation is to determine the suitability of the design for presenting recommendations in a social activity stream. Most importantly, we want to know whether it is easy to ignore de-emphasized content, because this design otherwise may not be a reasonable alternative to stream filtering. Of secondary interest was the accuracy of the recommender system. The current implementation was not expected to perform exceptionally well, but it should perform well enough to gain the users' trust. If users do not trust the recommender at all, they will not gain anything by reading only the recommended Tweets.

3.2.2 Recommender System

All recommender systems need some way to make inferences about user preferences. In Twitter, the value a user finds in a particular Tweet can be explicitly indicated by retweeting—sharing with one's followers—or favoriting, which is a public expression of interest or appreciation as well as an action that catalogues Tweets that can be easily viewed later. The potential downside to these actions is that they are completely public, which might be an incentive or deterrent to performing those actions in certain situations. For the purposes of training a recommender, it would be preferable to have private actions so the user feels free to be completely honest. Twitter also does not provide a way to indicate disinterest in a Tweet.

To address these shortcomings, I added private "like" and "dislike" features, used exclusively to train the recommender and to provide visual feedback by assigning "liked" and "disliked" updates automatically to the emphasized and de-emphasized tiers, respectively. These two actions are denoted by familiar "thumbs-up" and "thumbs-down" icons, which are displayed on the side of every Tweet on hover.

Users are given an opportunity upon first loading the application to rate individual Tweets from their home timelines as interesting using the retweet, favorite, and like icons, or as uninteresting using the dislike icon. After the user supplies 30 ratings, the recommender becomes active. Previous research on recommender systems using content-based methods suggests that ten ratings may be enough [Wang 2010], while Movielens.org[6] requires at least 15. However, in Twitter, such a small number of small updates from a single point in time may not be

[6] http://movielens.org/join (accessed May 6, 2014)

representative enough to attain an acceptable level of accuracy. Giving ratings on items as quickly consumable as Tweets requires little effort, so a slightly higher target was used for this application just to be sure.

The recommender uses a naïve Bayes classifier to predict whether unrated Tweets would be interesting to the logged-in user. Every ten minutes, the classifier is retrained using features from the rated Tweets stored in the database. Then all unrated Tweets are classified as interesting or uninteresting. The recommender uses Bayesian probability to classify Tweets using the relative frequencies of feature occurrence in the two classification groups, and it predicts how likely the Tweet is to fit into each classification. For the purposes of generating a single score, it assumes that all Tweets must be either interesting or uninteresting, and calculates a final value as a function of the posterior probabilities. The recommendation score is given by the following formula, where *T* is the Tweet being classified, *I* is the set of interesting Tweets, and *U* is the set of uninteresting Tweets:

$$score = \frac{P(T \in I)}{P(T \in I) + P(T \in U)} \tag{3.1}$$

For display to the end-user, Tweets are assigned to one of the three categories based on this score: Tweets with scores of less than 1/3 are classified as “uninteresting”, Tweets with scores greater than 2/3 are classified as “interesting”, and the rest are classified as “neutral”. The following features are included in the classification procedure:

- Content author and retweeter (if applicable)
- All hashtags and user mentions
- Tweet type(s): photo, link, retweet, reply, quote, manual retweet, and comment
- Popularity (number of retweets and favorites)

- Length of text
- Number of numeric digits

These features are all used to try to classify different types of Tweets. For example, a user may be partial to a particular user's updates when they make long posts with many numbers that have been retweeted many times and contain no hyperlinks.

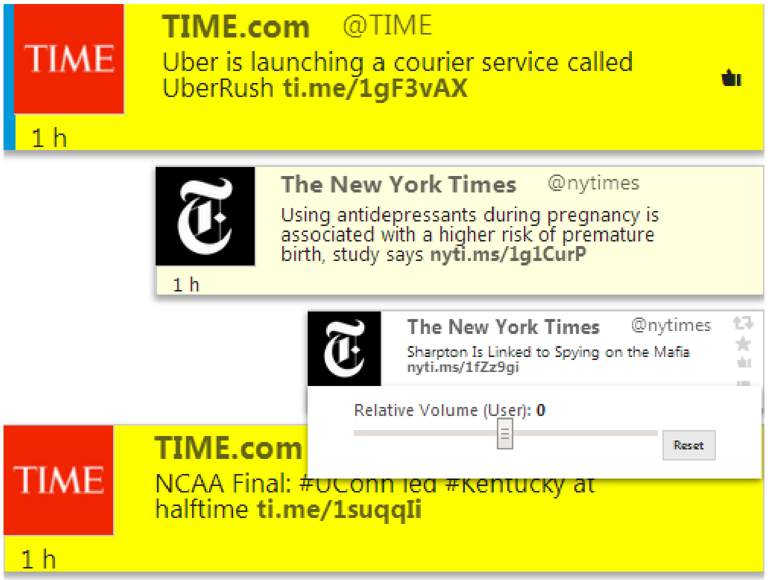
3.2.3 Implementation

The ViTA web application uses the Twitter API along with custom server and client software to present a custom social activity stream layout in the web browser. The software implementation of this application consists of three basic components: a client, server, and database. The server connects directly to the Twitter API and to the database and just sends the necessary updates to the client. A full JavaScript technology stack was used, consisting of Node.js, Express.js, MongoDB, Socket.IO, and AngularJS.

The application's view consists of a linear vertical display of Tweets in chronological order, to remain consistent with the official Twitter client and thus familiar to users. Each Tweet is shown in its own rectangular box (see Figure 3.3). Embedded images can be expanded by clicking but are displayed as thumbnails by default, as showing the full images would add too much size to the Tweet, artificially increasing its visual prominence in the stream.[7] Each Tweet appears differently depending on its recommendation tier. In addition to the content-based recommender system, users can increase or decrease the relative influence of any of the users they follow by means of a "User Volume" control. The control can be accessed by hovering the cursor over a user's avatar and moving a slider to the left or to the right. Since there are three recommendation levels, which can be thought of as –1, 0, and +1, influence can be changed from

[7] At the time that this application was being developed, Tweets were limited to only one image each.

a minimum value of –2 to a maximum of +2. For example, if a level of –1 is chosen then all Tweets from the selected user, unless explicitly rated, will appear as one level lower than the rating they were assigned by the recommender, to a minimum level of –1. Tweets that have been retweeted, favorite, or liked are displayed in the “interesting” tier regardless of their actual recommendation score. However, to distinguish these Tweets from ones that were merely predicted to be interesting, a small vertical bar is displayed at the left-hand side of the rectangular Tweet element. The same is also done for disliked Tweets.



TIME.com @TIME
Uber is launching a courier service called UberRush ti.me/1gF3vAX
1 h

The New York Times @nytimes
Using antidepressants during pregnancy is associated with a higher risk of premature birth, study says nyti.ms/1g1CurP
1 h

The New York Times @nytimes
Sharpton Is Linked to Spying on the Mafia nyti.ms/1fZz9gi

Relative Volume (User): 0

TIME.com
NCAA Final: #UConn led #Kentucky at halftime ti.me/1suqqIi
1 h

Figure 3.3: Partial screenshot of a ViTA timeline

3.2.4 Procedure

Twelve participants were recruited for this pilot study using word of mouth and through extended Facebook and Twitter friend networks. Participants ranged from casual to expert Twitter users. Five of them only read their Twitter timelines a few times a month, while seven do so several times per day, and the other about once a day. Follow network sizes were distributed evenly in size between 10 and 500 friends.

Initial instruction informed users of the purpose of the study and that they would be testing a prototype of a system that presents recommendations in a Twitter timeline. Estimated time to complete the study, including the final questionnaire, was 20–30 minutes. Upon logging in, users received further instruction. After reading about the purpose of the application and a brief explanation of the recommender and the different rating tiers, users were asked to complete the following tasks:

1. *Rate 30 Tweets*: Required to train the recommender.
2. *Tiered Reading*: Once the recommender is active, read through the timeline one tier at a time.
3. *Normal Reading*: Read through everything chronologically.
4. *Survey*: A link to an external survey is provided in the top bar of the application after the recommender is activated.

3.2.5 Results

The survey consisted of questions broken down into the following categories: Twitter usage, recommendation presentation, recommender performance, and design feedback. The results for categories 2 and 3, which were on a six-point Likert scale, are outlined in the following sections.

Recommendation presentation. Users were asked how easy it was to accomplish three objectives in the tiered and normal reading tasks. More than half of the participants said that it was “very easy” to read only the most emphasized Tweets, and all but one said it was at least “easy”. Ignoring only the most de-emphasized Tweets was more difficult, while consuming the entire stream chronologically was rated as the most difficult task, though all tasks had generally positive responses. The difficulty of the “Read All” objective is least relevant because users were able to disable the visualization styles altogether in the application settings. However, in some different use cases it may be valuable to use the tiers to convey information to users where it is also important that they consume all content. In general, these results suggest that this method of recommendation presentation could be a viable alternative to stream filtering, given an accurate recommender. These results are shown in Figure 3.4.

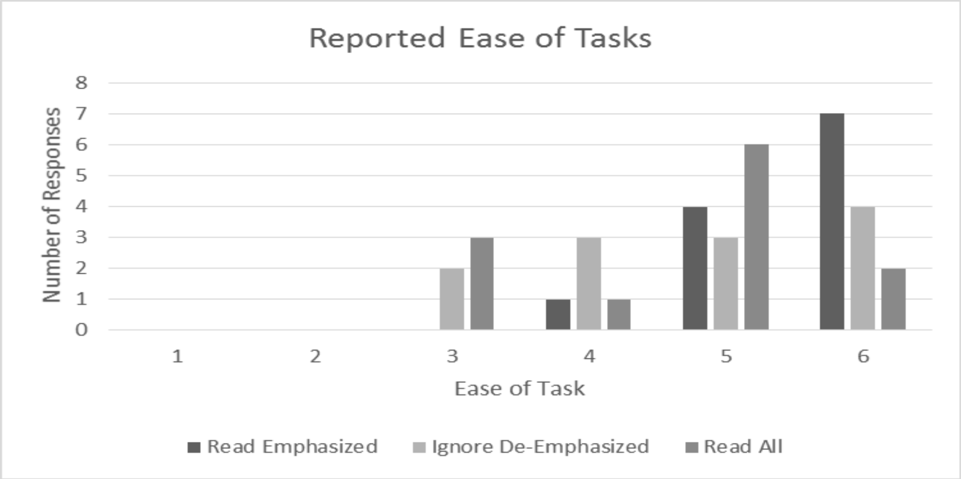


Figure 3.4: Reported ease of reading tasks

Recommender accuracy. Users were asked to rate the recommender's accuracy in placing unrated Tweets into the highest and lowest tiers. This very simple subjective evaluation showed slightly better results than expected, shown in Figure 3.5. It is very possible that an unbiased test of accuracy using pre-determined ratings in training and testing sets and cross-validation would tell a different story and that users are more forgiving of recommendations that are slightly off or just better than the alternative. At the same time, user perception of recommendation accuracy reveals more about trust than objective measures of precision and recall. User preferences would be quite easily inferred by the naïve Bayes classifier if the user follows others who Tweet only about a narrow range of topics. To get a more reliable indication of recommender system performance using this subjective method of testing, a larger sample size is needed.

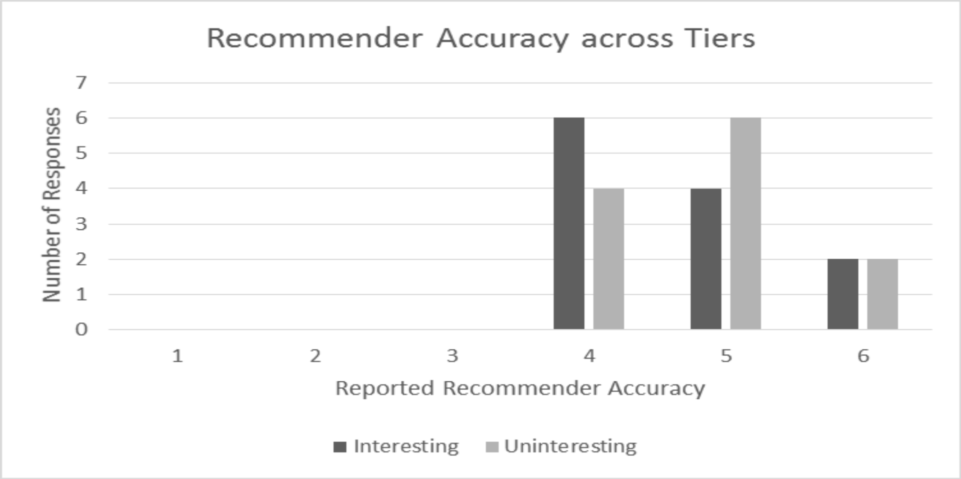


Figure 3.5: Reported accuracy by recommendation tier

Design feedback. Users were finally asked to rate the helpfulness of the User Volume feature and to give open-ended feedback about the user interface and application features. The helpfulness ratings of the User Volume feature are shown in Figure 3.6.

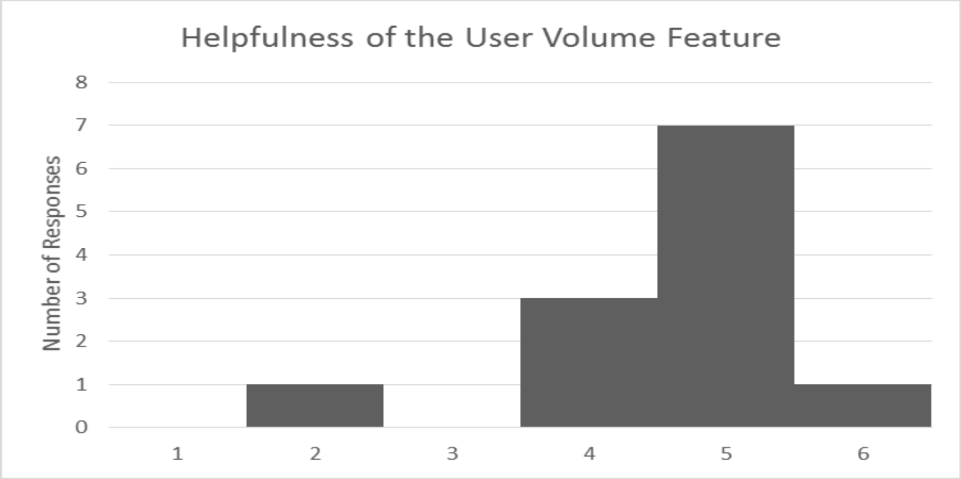


Figure 3.6: Reported helpfulness of the User Volume feature

3.2.6 Discussion

Results from the evaluation suggest that this presentation method is usable and that there would be value in doing future experiments. Based on user feedback, the recommender is sufficiently accurate to be used in those future studies as well. To repeat this study with a larger group of users would likely not add much benefit; in order to answer our research questions it would be better to compare the emphasized timeline with a filtered timeline to see which one users prefer. The primary takeaway from this pilot study is that this is a suitable design to match up against a stream filtering approach.

The greatest drawback to this pilot study was the limited sample size. Of course a pilot study using even a small number of participants is more helpful than none at all, since it forces the designer to consider the implications of releasing a system to the public further in advance. While the feedback was helpful for informing decisions about future enhancements to the system, it is unwise to make any firm conclusions about the results gathered from the Likert-scale questions. As well, most of the questions asked in the questionnaire and all of those used in the analysis were subjective and may have been positively biased. Future experiments will be designed to try to minimize this effect.

3.3 Two-Dimensional Stream Pilot Study

A unique challenge in active social streams is that it can be very difficult to catch up after some time away without skipping over many interesting updates. A simple solution would be to use filtering to provide the highlights of the stream that appeared during the time of absence.[8] However, this solution would still present Tweets taken out of their natural context. This is not necessarily a problem for all content, and it might be better than nothing, but it may result in major parts of the story being missed and it does not allow the user to make decisions in finding interesting content from the rich activity stream.

[8] Several months after this visualization was presented at the IntRS'14 workshop at RecSys'14 (early October 2014), Twitter released its own "While You Were Away" feature in its official apps.

A stream visualization that follows Shneiderman's visualization seeking mantra—"Overview first, zoom and filter, details on demand" [Shneiderman, 1996]—and simultaneously depicts all updates from within a specific time range while differentiating between the most popular and most recommended ones is a potentially useful alternative to stream filtering. It should allow users to explore more or less deeply depending on the amount of time they have available. By using a two-dimensional visualization that recommends and emphasizes the most important and interesting updates for a particular user at a particular time, users will have increased awareness of the most impactful updates in their networks and will be able to consume time-relevant updates more effectively and efficiently without needing to filter their streams. Such a design was implemented in a pilot study alongside the one-dimensional ViTA visualization described in Section 3.2.

3.3.1 Design

Overview first, zoom and filter. The design was heavily inspired by the interactive Rings visualization for Facebook [Shi et al. 2011], except to represent updates rather than users. A primary goal of the two-dimensional visualization was to provide an overview of stream activity while still facilitating common stream-consumption use cases of Twitter. As with Rings, the background for the stream visualization comprises a number of concentric circles around a central point (see Figure 3.7). The central point can be thought of as the immediate present. Each background circle, in increasing distance from this central point, represents a point in time further in the past. The distance between time circles remains close to constant, but the time represented increases at greater distances from the centre to allow more room at the present where there is less angular spread and where users a likely to focus their attention in order to read the latest updates. Each update is represented by a solid circle placed on the background. Thus the amount of time since an update was posted is coded in the visualization as the distance between the update circle to the background centre (the “now” point). Because larger objects are naturally more prominent visually, the recommendation score is coded with the circle radius. With this combination of visual mappings, Tweets that are more recent and more relevant to the user will occupy more space close to the central region of the visualization. Colour opacity was chosen as the mapping for Tweet popularity, which is calculated as a normalized sum of the number of retweets and number of favorites. Appropriate default minimum and maximum values are in place to prevent unreadable results, and users are able to personalize the appearance so that it works best for the throughput level of their own streams. The complete list of visual mappings is shown in Table 3.1.

Table 3.1: Variable-attribute mappings for ViTA's two-dimensional timeline visualization

Variable	Visual Attribute
Recency	Distance from center
Recommendation Score	Size
Popularity	Colour opacity
Unread/read	Shape (circle/horizontal line)

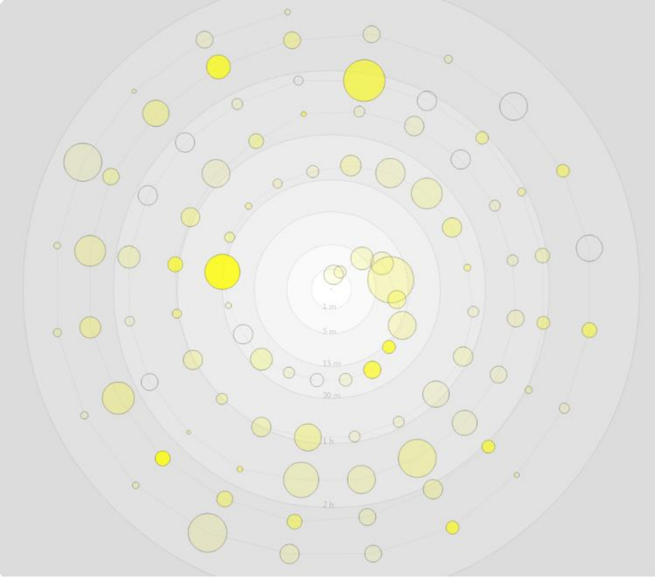


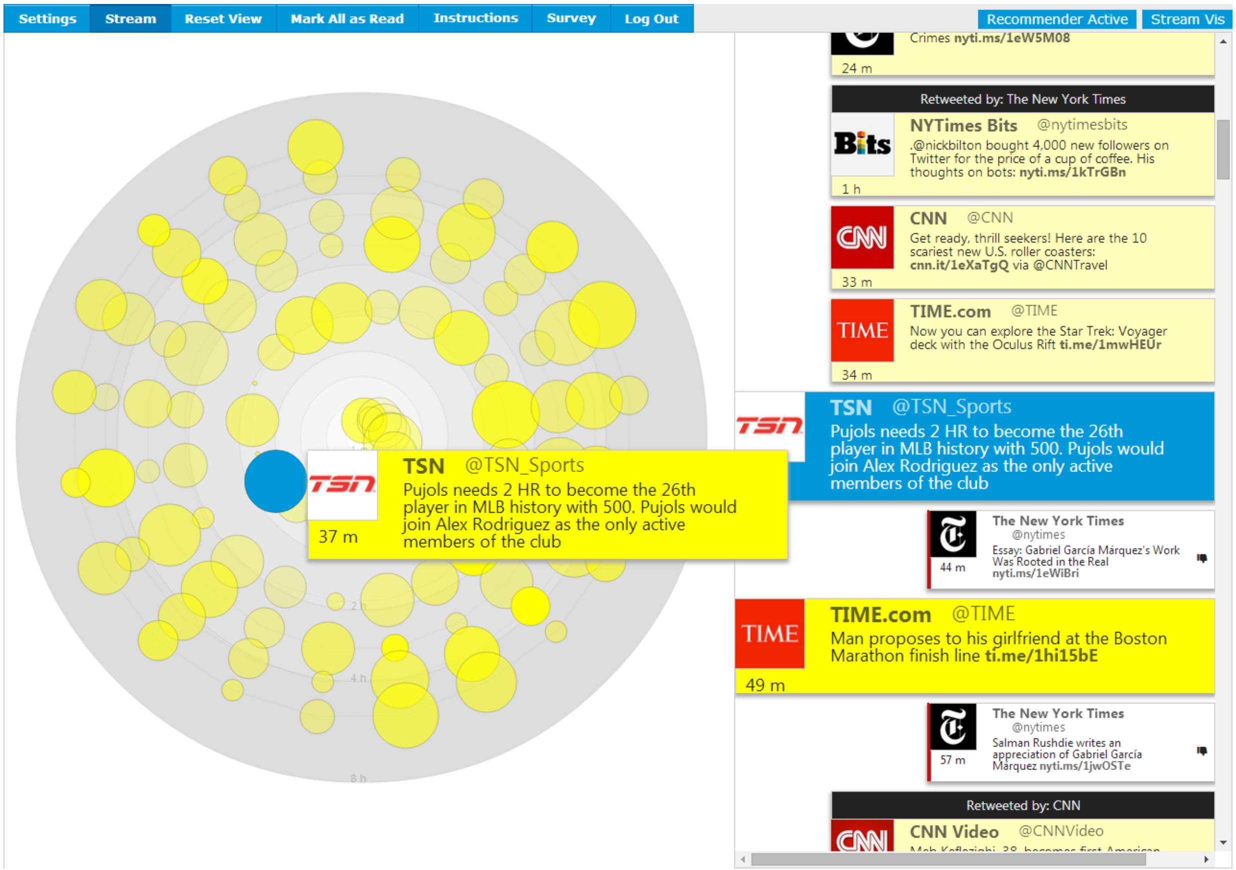
Figure 3.7: Partial screenshot of ViTA's two-dimensional timeline visualization

Details on demand. Showing hundreds of complete Tweets onscreen at one time would of course cause overcrowding and would overwhelm the users; this is why circles are being used as placeholders. The actual content of the Tweets is hidden until the user's cursor hovers over one of the circles. On hover, a small card-like element appears next to the cursor that displays the Tweet's content, including thumbnails of embedded images, and the Tweet author's user name and avatar. Additionally, there is a linear stream panel that can be docked along the right side of the window, which contains the one-dimensional ViTA visualization described in the previous pilot study in Section 3.2. When the user interacts with a Tweet in either view, the corresponding Tweet in the other view (including the circle representation in the visualization) will be highlighted to help draw a connection between the two stream representations. This may be helpful for a user who is reading the linear stream and wants to see the impact of a particular Tweet in relation to others around it. It also makes it easier to switch back and forth between views at any given time.

A filtering feature was also added in order to see how trust is affected when users, rather than the system, have full control over filtering. Users can move two sliders, one labelled "Min" and the other labelled "Max", to select a range of recommendation scores to allow through the filter. Setting the minimum value higher will exclude Tweets with low scores, while setting the maximum value lower will exclude Tweets with high scores.

3.3.2 Implementation

HTML5's canvas technology was considered for rendering the visualization, but elements and event handlers would be easier to manage if each component was a Document Object Model (DOM) node. Instead, Scalable Vector Graphics (SVG) technology was used to allow for creation of vector images, which can scale to arbitrary sizes without losing detail. SVG elements are defined using XML and can be used within HTML markup just like regular DOM elements. Because all of the graphics are scalable, I added a feature that allows the user to zoom in and out to the position of the mouse cursor by scrolling the mouse wheel. Panning in the visualization is also allowed by clicking on an open area and dragging the cursor in any direction.



TSN @TSN_Sports
Pujols needs 2 HR to become the 26th player in MLB history with 500. Pujols would join Alex Rodriguez as the only active members of the club
37 m

Crimes **nyti.ms/1eW5M08**
24 m

Retweeted by: The New York Times
NYTimes Bits @nytimesbits
.@nickbilton bought 4,000 new followers on Twitter for the price of a cup of coffee. His thoughts on bots: **nyti.ms/1kTrGBn**
1 h

CNN @CNN
Get ready, thrill seekers! Here are the 10 scariest new U.S. roller coasters: **cnn.it/1eXaTgQ** via @CNNTravel
33 m

TIME.com @TIME
Now you can explore the Star Trek: Voyager deck with the Oculus Rift **ti.me/1mwHEUr**
34 m

TSN @TSN_Sports
Pujols needs 2 HR to become the 26th player in MLB history with 500. Pujols would join Alex Rodriguez as the only active members of the club

The New York Times @nytimes
Essay: Gabriel García Márquez's Work Was Rooted in the Real **nyti.ms/1eWiBri**
44 m

TIME.com @TIME
Man proposes to his girlfriend at the Boston Marathon finish line **ti.me/1hi15bE**
49 m

The New York Times @nytimes
Salman Rushdie writes an appreciation of Gabriel García Márquez **nyti.ms/1jwOSTe**
57 m

Retweeted by: CNN
CNN Video @CNNVideo
Meb Keflezighi, 38, becomes first American

Figure 3.8: Screenshot showing both ViTA timeline visualizations side-by-side

3.3.3 Goals

In order to determine suitability for a large-scale quantitative study using this visualization tool, a smaller pilot study was necessary to identify pain points, streamline the experimental process, and determine the best way to collect the necessary data. The pilot study tested the usability of the system and the appropriateness of the variable mapping arrangement. Feedback was gained from the users on the following qualities of the system:

- Usefulness of the visual-emphasis approach to presenting recommendations
- Usefulness of the user-controlled filtering feature
- Ability to navigate the two-dimensional interactive visualization
- Usability in general
- Sources of particular difficulty

3.3.4 Procedure

Though more users were recruited via Facebook, only two participants ended up completing the study. They were required to complete, in order, all of the tasks listed in this section.

Explore and rate Tweets. Participants were required to rate Tweets to train the recommender. To do this, they were instructed to read through either the one- or two-dimensional visualization timeline in chronological order, rating especially interesting and uninteresting Tweets along the way. Thirty ratings were sufficient to produce what users deemed to be accurate recommendations in the previous ViTA pilot study, so the recommender was activated after 30 ratings. At this point users were to make any necessary adjustments to the default settings now that the size of the Tweets had changed to reflect recommendation scores.

Timeline reading. Participants were instructed to traverse their timelines chronologically, reading *only* the emphasized Tweets, first using the one-dimensional timeline, and then using the two-dimensional timeline.

User volume. In order to evaluate the usefulness of the User Volume feature in this setting, participants were instructed to identify some users they wanted to see more or less of in their timeline and then to use the User Volume slider to make that user's updates more or less visually prominent.

Filtered timeline reading. Finally, participants adjusted the filter settings to test the recommender and visualization’s joint effectiveness in another way. First they increased the minimum filter amount to show only the most highly-recommended Tweets, and then they reset and decreased the maximum filter amount to show only the least highly-recommended Tweets. After each adjustment they read their timelines and evaluated the effectiveness of the change.

Survey. A link to the questionnaire appeared after the recommender became active. Participants completed this survey as the final step in the study.

3.3.5 Results

The questions in the 20-part questionnaire were broken down into the following categories:

1. Twitter usage
2. Recommendation presentation
3. Recommender accuracy
4. Design feedback

The results for categories 2–4 are outlined in the following sections. Responses for categories 2 and 3 were on a six-point Likert scale.

Recommendation presentation. Participants were asked the following three questions for both the one- and two-dimensional timeline visualizations:

1. How easy was it to read only the most emphasized Tweets in your timeline?
2. How easy was it to ignore the de-emphasized Tweets in your timeline?
3. How easy was it to read through all Tweets in the timeline together in chronological order while the recommender was active?

Responses to these questions are shown in Table 3.2 and Table 3.3. Generally, the response to the one-dimensional visualization was very positive, while response to the two-dimensional visualization was mixed, but mostly negative. Both users found it at least as difficult to read the entire stream chronologically in both cases as it was to read only the emphasized Tweets or ignore the de-emphasized Tweets. This can be considered a positive result because it suggests that recommendation emphasis may be a viable alternative to filtering for stream consumption.

Table 3.2: Reported ease of tasks using the one-dimensional timeline visualization

	1	2	3	4	5	6
Emphasized	-	-	-	-	1	1
De-Emphasized	-	-	-	-	1	1
Combined	-	-	-	1	1	-

Table 3.3: Reported ease of tasks using the two-dimensional timeline visualization

	1	2	3	4	5	6
Emphasized	1	-	-	1	-	-
De-Emphasized	-	-	-	1	1	-
Combined	1	-	1	-	-	-

Recommender accuracy. With regard to recommender accuracy, the following questions were asked:

1. How accurate was the recommender in emphasizing interesting Tweets?
2. How accurate was the recommender in de-emphasizing uninteresting Tweets?
3. How strongly do you agree with the following statement? “As you increased the minimum filter value, the application showed a generally more interesting timeline.”
4. How strongly do you agree with the following statement? “As you decreased the maximum filter value, the application showed a generally less interesting timeline.”

Responses to these questions are shown in Table 3.4. Subjective evaluations of recommender accuracy do not necessarily tell the whole story, but they are a very important component, especially in social activity stream recommendation. Participants may be especially forgiving in this setting because reading an uninteresting Tweet causes little harm. Overall these results are not surprising because of the results found in the previous ViTA pilot study, which used the same recommender. However, it shows that this small set of users did not punish the recommender regardless of their opinion of the visualization method.

Table 3.4: Responses concerning recommender accuracy and manual timeline filtering

	1	2	3	4	5	6
Interesting	-	-	-	-	2	-
Uninteresting	-	-	-	1	1	-
More Interesting	-	-	-	-	2	-
Less Interesting	-	-	-	-	2	-

Design feedback. With regard to the user interface and feature design, the following questions were asked:

1. How useful was the “User Volume” feature?
2. Which timeline presentation style would you most prefer for regular use?
3. What did you like most about the user interface?
4. What did you like least about the user interface?
5. Which application feature did you like most?
6. Which application feature did you like least?
7. Do you have any other comments or suggestions?

The first question had responses on a four-point Likert scale, while the second question asked the users to choose between three options. The others were all text fields that allowed for open-ended responses.

Because Tweet interest can fluctuate greatly even within the set of Tweets from a given user, it was unclear how helpful the User Volume feature would be. However, both participants reported that the feature was useful. When asked which timeline presentation style they would most prefer for regular use, participants were given the choice between showing everything equally, showing everything with varying levels of emphasis, and filtering out the most uninteresting Tweets. Neither participant said that they would prefer everything to be shown equally, while each of the other two options was selected once. Without more participants these responses are not very useful, but it does suggest an appetite for users to have some processing done on the content in their timeline, as not all updates are created equal.

The open-ended responses revealed some useful suggestions for future improvement. Participants found the two-dimensional visualization relatively difficult to use and understand, suggesting that the presentation and interface could be more intuitive. The greatest source of trouble was lag due to a feature inherent to the AngularJS framework that makes it a bad fit for this type of dynamic visualization. AngularJS's two-way data-binding feature uses "dirty checking", which continually checks watched variables to see whether their values have changed. Because of the large number and large size of variables used in this visualization, as well as the processing needed to translate that raw data into SVG attributes, the script executed very slowly. Significant performance enhancements are possible by using a different application framework.

In designing the visualization, I attempted to mitigate any potential performance problems by allowing users to limit the number of Tweets displayed on the page at one time, and in testing this seemed to work well. It is unclear whether the users missed reading about this feature in the instructions or if it did not have the same positive effect in their environments. It may also not be as practical in higher-throughput streams to limit the number of Tweets shown too much.

3.3.6 Discussion

In a direct comparison, the two-dimensional ViTA timeline visualization does not seem to be as usable or effective as its one-dimensional counterpart. Though no firm conclusions can or should be made from the limited results of this pilot study, participants were split on the general concept of emphasis versus filtering and indicated that significant enhancements would need to be made to this particular design for them to consider using it regularly.

The greatest drawback to this pilot study, even more so than the previous one, was the extremely limited sample size. I gained useful feedback on specific design elements, but it is unclear whether the opinions of these few are rare or would be shared by many others in the population. The procedure also could have been better designed to have more structured tasks rather than an open-ended, exploratory approach. Because the two-dimensional visualization is so different from the familiar Twitter stream and shows such a great volume of information, participants are already at risk to be confused and disoriented by it. Providing more structure could ease them in more slowly and help them understand what each of the visual dimensions represents. For this reason it could also be argued that the design itself also does not deal well with the problem of information overload.

It was likely difficult to read all Tweets chronologically due to the spiralling arrangement. Because they are aligned at certain angles, the farther the circles are from the centre, the greater the distance to the next Tweet chronologically, and several other Tweets that share that same angle may actually be much closer. At these distances it may be difficult to follow the arc naturally to find the next Tweet and to differentiate it from other Tweets that may have been posted several minutes earlier or later. This problem may have been mitigated by utilizing a greater number of angular positions along the arc with increased distance. In general, it may be a drawback of the design that no variable was being mapped to the angular dimension other than to follow chronological sequence. This decision was made in order to make the timeline readable in the spiral. Using distance from the centre alone would make it too difficult to compare Tweets to determine any kind of order and context within the stream. However, some other ideas were considered to utilize angular position, including representing different topics or to cluster Tweets by authors' network similarity.

3.3.7 Conclusions

A larger sample size is desirable before writing off the two-dimensional visualization as a completely useless tool for stream consumption, but it would likely benefit from some design changes. It is possible that the visualization is better served as a complementary view to provide social activity awareness and a general view to support a primary one-dimensional, textual stream. Some possible reasons users preferred the one-dimensional stream are that it supports a more passive browsing style, shows a deeper level of information at one time, is more familiar, is much simpler to understand, and contains larger targets for mouse interaction. More information needs to be gathered about the particular weaknesses of the existing system, and more usability testing is needed to improve the system before carrying out a large-scale user study. Ultimately my line of research will no longer pursue the two-dimensional timeline visualization approach, but it could be a topic for future work.

CHAPTER 4
THE PERSUASIVENESS OF PURE EMPHASIS

In the first ViTA pilot study it was unclear what degree of the users' favourable overall response was due to the persuasiveness of the presentation. Perhaps the system was better at making its users believe it was doing a good job than it was at actually doing that job. Before moving on to a larger study using the ViTA system, it is important to establish the persuasive power of this approach to recommendation presentation in order to interpret future results properly. Since the ViTA pilot study attempted to evaluate several features at once (recommender accuracy, readability of the stream, appropriateness of visualization, and usefulness of user-level volume control) by only subjective means, there were many potential confounding factors for any one measurement. A strict evaluation of persuasiveness could only be effective by stripping out all unnecessary features while trying to make the experience as uniform as possible across the set of participants.

Recommender accuracy can vary across users for a number of reasons. First, users may have vastly different feedback behaviour. For example, one might give a much higher quantity and variety of information in the form of favorites and retweets than another, better training the recommender and resulting in more helpful recommendations. Second, users' follow networks will be differently distributed. If one user follows many users who all Tweet similarly to one another, the results will be different than for a user with a very diverse network. In the first case, the recommender will need to be more discriminating and might find more subtle differences between interesting and uninteresting Tweets. Meanwhile, in the second case, because of the great differences between Tweets, the recommender may be more likely to interpret irrelevant details as having predictive power in a limited sample size. Finally, two different users are likely to have different stream consumption habits. This can depend as much on the user's stream content as the user's personality. One user (A) might passively skim Tweets to find something of interest, while another (B) might focus more deeply and follow conversations. They may find different things valuable in their timelines, so a recommender perceived as accurate for User A may not be perceived as accurate for User B.

4.1 Design

For this study, I developed a new web application based on the ViTA presentation principles, code-named ViTA Tempest. Tempest leverages ViTA's recommendation presentation design and was intended to appear as a simple online survey. For each part of the survey, users were presented with a set of six Tweets, and the part consisted of two tasks:

1. Select the Tweet that you find to be *most* interesting.
2. Select the Tweet that you find to be *least* interesting.

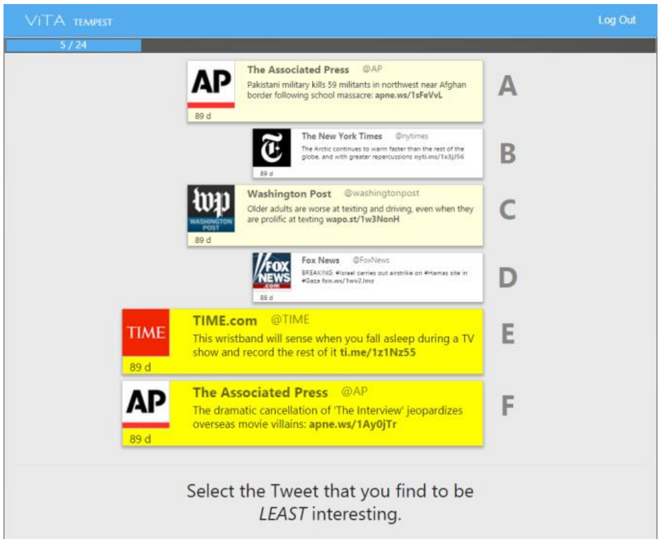
The goal of this study is to remove all of the sources of variation between users listed above and isolate how much the emphasis and de-emphasis of Tweets influence users' subjective evaluation of recommender accuracy. In other words, we want to see whether users were more likely to choose emphasized Tweets as most interesting and de-emphasized Tweets as least interesting. However, in order to evaluate the persuasiveness of the presentation method completely independently of the accuracy of the recommender itself, it was necessary to remove the recommender system from the equation entirely. In Tempest, the content-based recommender is replaced with an algorithm that determines at random which Tweets will be emphasized, which will be de-emphasized, and which will be neutral. The reason for having six Tweets in each set was to have the minimum possible number of Tweets while having at least two Tweets at each level of emphasis. Indeed, the algorithm was designed to choose *exactly* two of the six for each of the three levels by producing random permutations of the array $\{-1, -1, 0, 0, 1, 1\}$.

In order to accomplish our goal of an evaluation of pure persuasiveness, it was necessary to address the three sources of recommender accuracy variation identified above, which are all problems that need to be considered in our design:

1. Feedback behaviour
2. Follow network distribution
3. Stream consumption habits

Replacing the recommender with a random one partially addresses the side effects of the second problem. In addition to this effort, Tempest presented the exact same set of Tweets to all users. Only the order of Tweets within each set was randomized so as to remove any potential confounding effects of sequencing, such as the serial position effect, which is a tool than can be used to make recommenders more persuasive [Felfernig et al. 2007] but would confound what I wanted to evaluate in this study. The order in which the sets are presented is the same for all users. Showing the same Tweets to all users and standardizing the mechanism for feedback by requiring users to choose exactly one most interesting and one least interesting Tweet from each fixed-size set addresses the first and third problems.

Requiring the same sets of Tweets for all users presents an interesting problem: How should these Tweets be selected for a set of users, none of whom will have exactly the same interests? The natural inclination is to select Tweets that are each equally likely to be found interesting across the entire population. However, the problem of selecting Tweets to match this criterion is beyond the scope of this research and to solve it would be a tremendous undertaking. Therefore, a simpler approach was taken: select recent, popular Tweets from popular accounts that are likely to be relevant to a large portion of the population. Sixty general-interest Tweets were selected from popular media accounts focused on news, sports, science, technology, business, finance, and health.[9] To make this simpler approach work, it is necessary to randomize the emphasis levels between users, not just between Tweets, so that each user sees a different distribution of emphasis levels within each set of six Tweets. If each user saw the exact same distribution then the inherent, general interest level of a Tweet would confound the results. For example, a highly interesting Tweet emphasized for all users would skew the results to indicate a more persuasive system than actually exists. Since an equal level of interest between Tweets could not be guaranteed, I had to ensure that (1) each Tweet was equally likely to be assigned each emphasis level for a given participant, and (2) each Tweet was placed into each emphasis level roughly an equal number of times across all participants. No effort was taken to have a representative sample of Tweet topics in each set because I suspected users would be more likely just to select Tweets from the topic that is most interesting to them. The more users are able to develop conscious rules to govern their selections, the less the subconscious factor of emphasis, which we are hoping to detect, is likely to play into their decisions.



The Associated Press @AP
Pakistani military kills 59 militants in northwest near Afghan border following school massacre: **apne.ws/1sFeVvL**
89 d

The New York Times @nytimes
The Arctic continues to warm faster than the rest of the globe, and with greater repercussions nyti.ms/1x3jJ56
89 d

Washington Post @washingtonpost
Older adults are worse at texting and driving, even when they are prolific at texting **wapo.st/1w3NonH**
89 d

Fox News @FoxNews
BREAKING: #Israel carries out airstrike on #Hamas site in #Gaza fxn.ws/1wv2Jmz
88 d

TIME.com @TIME
This wristband will sense when you fall asleep during a TV show and record the rest of it **ti.me/1z1Nz55**
89 d

The Associated Press @AP
The dramatic cancellation of 'The Interview' jeopardizes overseas movie villains: **apne.ws/1Ay0jTr**
89 d

Select the Tweet that you find to be *LEAST* interesting.

Figure 4.1: Screenshot of the ViTA Tempest survey

[9] For the sake of simplicity, Tempest used only English-language Tweets and targeted English-speaking users.

4.2 Treatments and Hypotheses

For the results to be valid, users needed to believe that the emphasis levels actually meant something. If they were told from the beginning that the emphasis was determined randomly, we would be unlikely to measure any effect, regardless of the design's actual persuasive potential. To investigate the effect of users' perception more deeply, participants were split randomly into two equally-sized groups, A and B. The first user to log in for the first time was placed into Group A, the second into Group B, the third into Group A, and so on. Depending on which group they were placed in, participants were shown one of the following messages upon logging in:

Group A: ViTA emphasizes Tweets in your timeline based on their popularity.

Group B: ViTA emphasizes Tweets in your timeline by learning from your Twitter history to determine your preferences.

Group A's message was left open-ended, but it suggests that the Tweets with more retweets and favorites were emphasized more greatly. As for Group B, since every participant in the study was required to sign in with their Twitter accounts, there was no reason to believe that this statement was untrue.

Hypotheses. As part of the analysis we have two null hypotheses to explore:

1. The task (select the *most* or *least* interesting Tweet) will have no effect on the level of emphasis of the Tweets users choose in response.
2. The recommendation mechanism explanation treatment groups will have no effect on the level of emphasis of the Tweets users choose for either task.

4.3 Pilot Study

The ViTA Tempest pilot study involved six users recruited from Facebook and from among computer science graduate students of the MADMUC lab at University of Saskatchewan. The purpose of the study was not to analyze any of the data, but to try to eliminate any bugs or flow problems before launching the study to a wider audience. During testing, I did discover two problems.

First, some users were unable to get past the login page. Though I could not reproduce the error myself, I eventually narrowed it down and guessed that it was due to an error on the server in parsing the session cookie. I replaced my simple method, which used regular expressions, with Express.js's cookie-parser middleware, which is built to handle cookie parsing robustly. This change solved the problem for those users who were experiencing difficulty.

Second, though the server was recording timestamps for every individual answer submission from the participants, it was not recording the time they started. Because of this, there was no elapsed time measurement for the first task. I added a timestamp to log the time they began the first task in order to solve this problem. These timestamps were vital because they could identify users who rushed through the tasks and did not take time to read the Tweets to make an informed decision. Participants who were deemed to rush would have their data removed from the analysis.

4.4 CrowdFlower Study

Crowd-sourcing platform CrowdFlower was used to recruit a much larger set of participants than for the pilot study. When launching a survey job in CrowdFlower, one can choose how many times the survey should be completed. To start, I launched ViTA Tempest for 20 users, again so that any problems I might have encountered could be caught before releasing it to a larger group. These 20 responses came back in about an hour, and since everything seemed to go well, I immediately launched it for another 100 users. It is not possible to add users to a job once it has been launched, so it was necessary to copy the job while keeping IP restrictions in place from the original job so that no one could complete the survey twice. Unfortunately, these measures did not successfully prevent duplicate submissions from the original launch, but I was unaware of this until after launching a third time for 200 more users. However, I had already prepared to identify duplicates manually, so these submissions would not be counted twice in my analysis. In total, 320 English-speaking users were recruited to participate in the study. There were no requirements regarding Twitter experience, but all CrowdFlower contributors targeted needed to have a valid Twitter account in order to log in and complete the survey, and were at least level 2 on CrowdFlower's three-level quality scale, which is determined by their history of contributions.

The CrowdFlower job consisted of an external link to the survey and a verification code input field. Users were to complete the survey, after which they would be presented with a verification code to copy and paste into the field on the CrowdFlower page. The verification code was constructed from the Twitter user's ID using a hash function so that it was possible to easily identify duplicate or false entries. CrowdFlower provides the capability to validate fields by regular expressions, but this was unknown to me at the time. This could have been used to eliminate users who submitted responses without even looking at the survey, since all verification codes were 32-digit hexadecimal numbers. Those users were ultimately eliminated, but it took manual work.

To eliminate duplicate and false entries, I first removed any responses with a verification code that was not a 32-digit hexadecimal number. I then eliminated any results with a verification code that had already been submitted in the past, which indicated that the same Twitter user completed the survey more than once. The test tasks were designed to have five unquestionably uninteresting Tweets (such as "boring" or "i'm bored") and one Tweet that was an interesting fact. I also investigated whether any users did not take enough time to read the six Tweets in any given set, but none of the remaining responses seemed to be rushed. After eliminating all of the invalid responses I had 170 remaining participants, and after removing test tasks, the result set consisted of 20 responses for each user: 10 selections for which Tweet was most interesting and 10 selections for which Tweet was least interesting.

4.5 Results

The responses for each of the 170 remaining users and each set of Tweets were paired together to give 1700 rows of data, each with two data points. The first value was the emphasis level of the Tweet the user selected as most interesting, and the second was the emphasis level of the Tweet the user selected as least interesting. Since the emphasis level of each Tweet was determined randomly, these two sets of values should only differ significantly if the visual presentation of recommendations is affecting participants' judgment of the Tweets' interestingness. In other words, if the two sets are significantly different, then it suggests that the recommendations are persuasive. The Wilcoxon Matched-Pairs Signed-Rank Test is more appropriate than a *t*-test in this case to compare the two groups, since the data is ordinal rather than interval and since the two values for each question are not independent—this should be obvious because the Tweet chosen for the first task cannot also be chosen for the second.

Hypothesis (1): Null hypothesis rejected. Pairing together the responses for each of the two tasks by Tweet set for each participant, the Modified Wilcoxon Signed-Rank Test [Oyeka & Ebuh 2012] gave a *z*-score of 6.84. Since the sample size was large ($n = 1700$), a *t*-test should also give a similar result. As a sanity check, a *t* statistic of 6.90 was also calculated. These results show a very strong significance, with a *p*-value of far less than 0.0001. In fact, given this result and assuming hypothesis (1) is true, the probability of obtaining a sample of this size with a difference at least this extreme is less than 10^{-11}. This leads us to reject hypothesis (1) and accept that users are more likely to choose higher-emphasis Tweets to be most interesting than to be least interesting. This isn't necessarily a surprising result, but its decisiveness is in large part due to a massive sample size and does not answer the question of how great the effect actually is.

Table 4.1: Cross-group selection rate by task and emphasis level

Task	Emphasis Level	Selection Rate (%)
Select Most Interesting	Emphasized	41.9
	Neutral	33.1
	De-Emphasized	24.9
Select Least Interesting	Emphasized	33.4
	Neutral	31.0
	De-Emphasized	35.6

If participants were truly choosing the most and least interesting Tweets without being influenced or persuaded by the recommendations, we would expect values approaching 33.3% across the board. Instead, as shown in Table 4.1 and illustrated in Figure 4.2, an emphasized Tweet was selected as most interesting 41.9% of the time—25.8% higher than the 33.3% we would expect by random chance. Meanwhile, de-emphasized Tweets were selected as least interesting 35.6% of the time, which is 6.9% more than expected and 43.0% more often than they were selected as most interesting. Both de-emphasized *and* emphasized Tweets were chosen more often than neutral Tweets as least interesting. This all can be explained by the fact that the emphasized Tweets are more visually prominent and likely attract user attention earlier and more strongly. Thus, regardless of the task and regardless of their content, there is an inherent bias toward these Tweets because of their visual presentation.

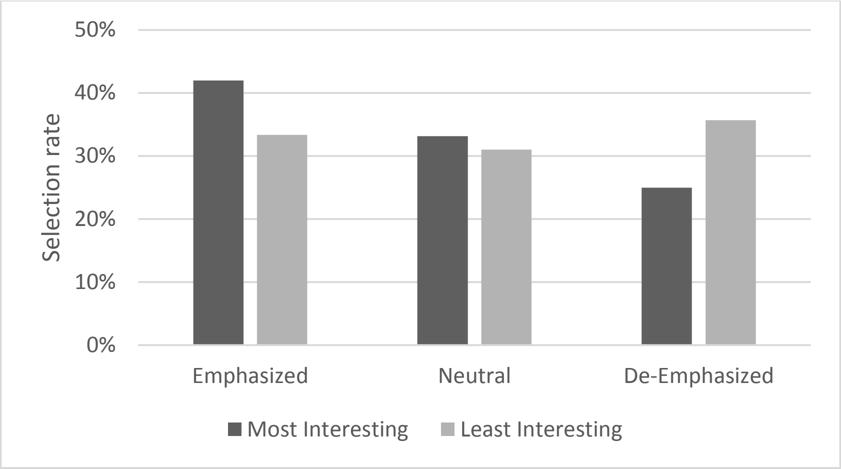


Figure 4.2: Selection rate by emphasis level and task

Hypothesis (2): Null hypothesis accepted. In order to test hypothesis (2), it is necessary to compare the two groups: A, which was primed with the statement that the highlighted updates are the most popular; and B, whose members were told that updates are personally recommended to them based on their Twitter history. In this case I used the Mann-Whitney test and looked at the two tasks individually. Each data point in the sample was a selection for one of the survey tasks (choose the most interesting or choose the least interesting update), and for each task I investigated whether there is a significant difference between the two groups. The results of this analysis are shown in Table 4.2.

Table 4.2: Mann-Whitney test statistics comparing groups A and B for each task

	U	μ	σ	z
Most Interesting	351575	360800	9457.2	0.385
Least Interesting	350817	360800	9457.2	1.056

Neither of these *z*-scores suggest abnormal samples given the assumptions of hypothesis (2), so we accept that there is no statistically significant difference between treatment groups A and B for either task. The insignificant differences between the two groups by task at each level of emphasis are depicted in Figures 4.3 and 4.4.

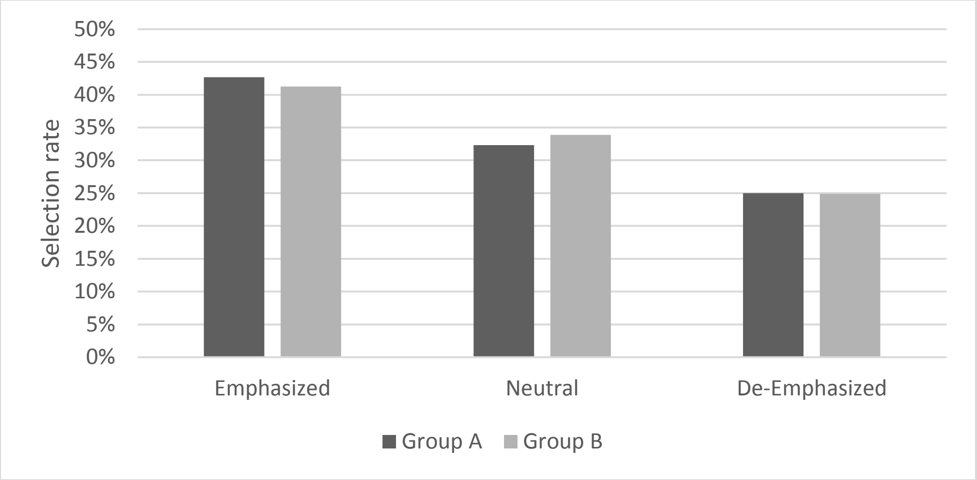


Figure 4.3: Most interesting Tweet selection rate by emphasis level and group

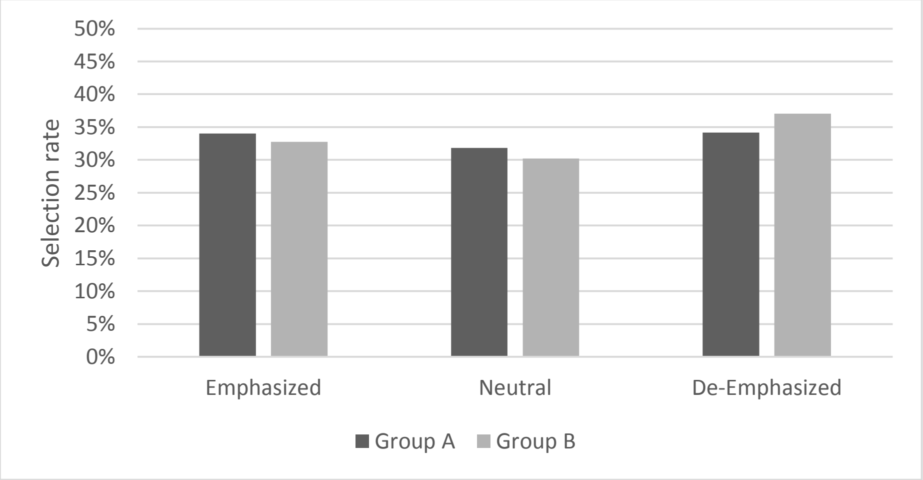


Figure 4.4: Least interesting Tweet selection rate by emphasis level and group

Effect of Tweet Order. If we look at selection rates by option letter (which reflected the order in which the individual 6 Tweets were listed in each of the 10 sets), we can clearly see that it was a good experimental design choice to randomize the order of Tweets within each set for each user. Participants were more likely to choose Tweets toward the beginning of the list as most interesting and Tweets toward the end of the list as least interesting. Overall, 56.6% of response pairs had the most interesting Tweet coming before the least interesting Tweet in list order. Figure 4.5 shows that the distribution of Tweet choices by option letter for the two tasks are almost mirror images of each other.

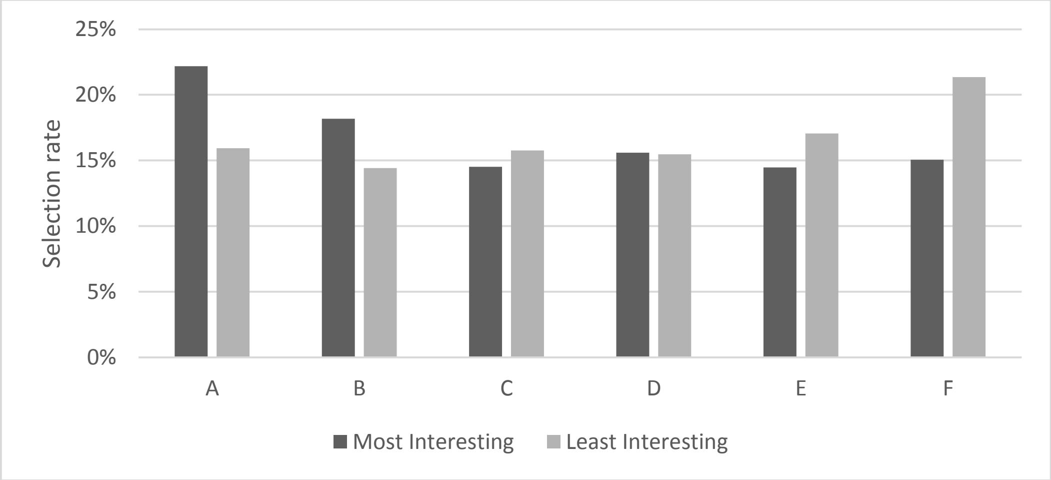


Figure 4.5: Cross-question option letter selection rate by task

This finding suggests a likely effect of the order in which the two tasks are presented. Unless there is another reason for the phenomenon, this effect would disappear if this order were also randomized. In the experimental design, the order of tasks was kept consistent to make the survey less confusing, to try to minimize the chance that a participant completed the wrong task by mistake, for example by picking the most interesting Tweet first in a set because that was what they did first for the previous set. However, finding a way to vary the order in which tasks were presented within each set while making it more clear which task was to be performed each time would have led to a better design. Meanwhile, this also tells us that there are other factors besides emphasis that are impacting users' decisions. Users generally chose Tweets closer to the top as most interesting and later Tweets as less interesting irrespective of emphasis level, as shown in Figures 4.6 and 4.7. Since all Tweets have an equal chance of being at each level of emphasis, this suggests a systematic strategy. And since *everything* within a given set is randomized, any dishonest strategy the user adopts (consciously or otherwise) that does not take emphasis level into account is only going to cause the results to tend toward the expected averages and away from the significant result that ultimately was found. In other words, the relationship between selected Tweets and emphasis levels may actually be stronger than demonstrated in this study. It is possible that this effect only reflects the behaviour of a particular subset of users within the sample rather than the population as a whole.

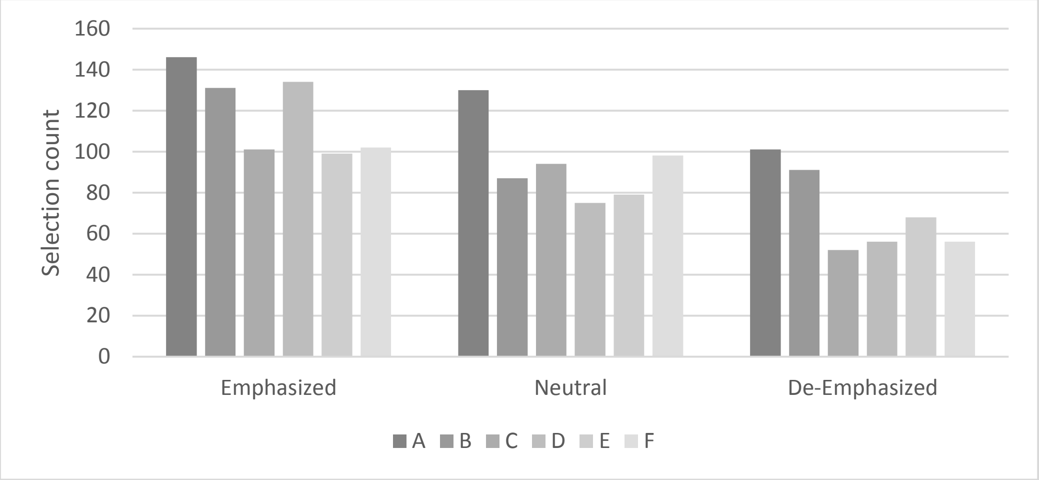


Figure 4.6: Emphasis levels of most interesting Tweets by option letter

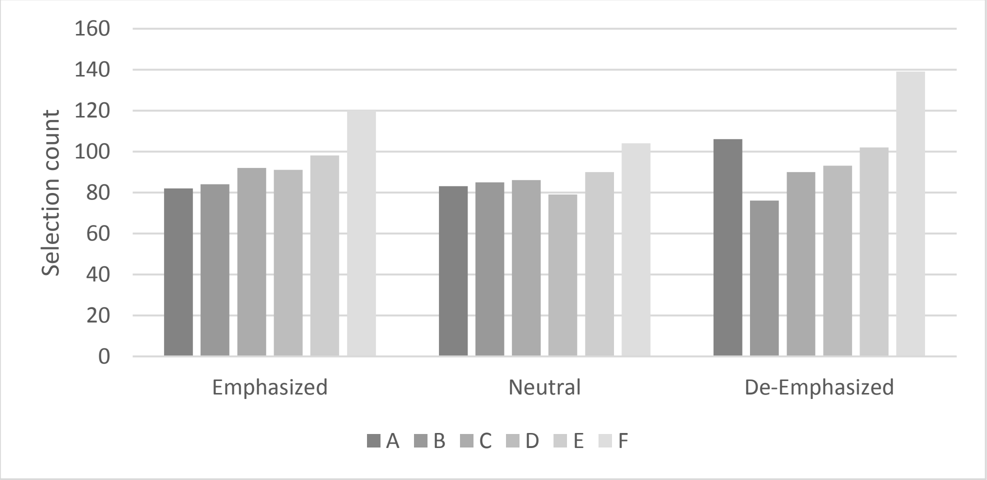


Figure 4.7: Emphasis levels of least interesting Tweets by option letter

4.6 Discussion

The results clearly show that recommendations in this study were persuasive, but the underlying cause of this persuasiveness is not obvious. It seems reasonable to conclude that the visual prominence of the highly-emphasized Tweets caused them to be chosen more often in general (37.7% of selections across all tasks were highly-emphasized Tweets). However, this does not explain why more de-emphasized Tweets were more often chosen as least interesting or why the disparity between numbers of emphasized and de-emphasized Tweet responses was so great for the question of which was most interesting. The most natural explanations are that the participants simply wanted to agree with the recommendations or that they actually *did*—albeit subconsciously—agree with the recommendations. Perhaps it is a more subtle variation of that second possibility: if the highly-emphasized Tweets are more colourful, easier to read, and generally more pleasant to look at, participants may be conflating their visual experience of the Tweet with its actual content. This generally positive perception would be more likely to cause the user to select it as the most interesting of the set even if they would have chosen a different one if the comparison was by content only. Even more simply, content of the larger Tweets may have stuck more in users' minds because of their visual prominence or because they tended to read them first. A variation on this study emphasizing Tweets using less appealing visual characteristics such as less legible fonts or colours associated with negative emotions would be an interesting direction for additional experiments.

Another possible factor carrying an even simpler explanation is that some participants may have used emphasis level to break ties. This is especially likely if they did not spend much time reading and deciding. However, this factor alone is not enough to explain the great difference in persuasive power of the presentation between the two tasks: emphasized Tweets were chosen as most interesting more than 25% more often than expected, while de-emphasized Tweets were chosen as least interesting less than 7% more often than expected.

As mentioned in the analysis of the results, a better design would have randomized the order of the tasks within each set for each participant rather than asking for “most” before “least” in every set. We saw that users’ choice for most interesting Tweet was likely to be earlier in the list than their choice for least interesting. Considering that users were overall much more likely to choose an emphasized Tweet as most interesting, those cases would leave them with only one emphasized Tweet among the remaining five to choose from as least interesting. Choosing emphasized Tweets because they are more visually prominent could thus explain part of the effect of least interesting Tweets being chosen among the less emphasized ones simply because of the order of the tasks.

The explanation of the recommender’s decisions (popularity-based or personalized) had no significant effect on the emphasis levels of Tweets chosen as either most or least interesting. This is not necessarily a surprising result, but neither is it trivial. One might expect participants to agree more with “crowd-sourced” recommendations (based on numbers of retweets and favorites) due to social influence and conformity [Laporte et al. 2010] and the anchoring effect that has been detected in e-commerce websites [Adomavicius et al. 2013]. Similarly it would seem natural for users to fight against personalized recommendations that are inaccurate. Harman et al. found this to be the case when comparing personalized and non-personalized recommenders and that participants trusted an inaccurate personalized recommender less than an inaccurate non-personalized one [Harman et al. 2014]. In fact what I found on average is that Group B (personalized) agreed with the random recommendations more often than Group A (popularity) only for de-emphasized Tweets (i.e. the not-recommended items, or those that would be filtered away in a standard filtering recommender system), though the difference was not statistically significant.

Persuasiveness may vary depending on the individual participant. The results of this study do not imply for any given user the level to which they are susceptible to being persuaded by this design, but rather an average level across the population. Additionally, there may still be some variation in how users consume the Tweet set. For example, one user may read the emphasized Tweets first rather than reading from top to bottom, stopping when they have found something particularly interesting. While I hoped that the pretense that users' feedback is being used to evaluate and improve a recommender system would convince participants to read all Tweets before responding, it is quite unlikely that all participants did.

The possibility has occurred to me that some users may have just picked their favourite from the pair of emphasized Tweets and their least favourite from the pair of de-emphasized Tweets. We could eliminate results that strictly follow this pattern, but such participants may have still been acting honestly. To remedy this, we could have included a test task at the beginning of each set of tasks with only one correct answer that required users to read through all Tweets. It might be necessary to enforce a timeout period before allowing users to answer so that there is no incentive just to guess until they get it right. This step would add time to the study and inconvenience for the participants, but may give more accurate results. In any case, there is no reason to expect the difference to be so great that the acceptance of either hypothesis would change. Using eye-tracking is unfortunately not feasible in crowdsourced studies, but a smaller study in lab conditions may be used to explore the relative frequencies of patterns of reading the tweets.

This study does not compare persuasiveness of this design to other methods of presenting recommendations. The goal of this study was to measure the absolute persuasiveness of this particular design rather than to suggest it is the most persuasive design possible. However, isolating and examining the effects of the individual elements of emphasis (rectangle size, text size, colour, horizontal offset) versus a baseline, plain textual indication of recommendation tier (“Recommended”, “Neutral”, or “Not Recommended”) is well worth a future study.

4.7 Conclusions

The work described in this chapter demonstrates how recommendations can be agreeable even when they make no attempt to be accurate, and it shows that the ViTA presentation design is significantly persuasive. While this design can easily be applied to domains such as social activity streams, the effects may vary in different contexts. The next step is to evaluate this system within users’ real-world Twitter timelines and directly compare the three-tiered emphasis approach with one that filters out the updates predicted to be least interesting. Future research may also probe more deeply into the factors that influence persuasiveness, including comparisons with alternative designs and isolation of the various visual elements that compose the ViTA design.

CHAPTER 5
EMPHASIS, FILTERING, AND ERROR INSERTION

So far none of this work, aside from a minor qualitative evaluation as part of the two-dimensional visualization pilot study, has directly compared a ViTA stream to a filtered stream. We know that presenting certain information about filtered items can increase users' trust in a system [Nagulendra & Vassileva 2013], but I want to investigate whether this same positive effect can be demonstrated in an all-inclusive stream that uses different levels of emphasis instead of filtering. However, even if this design is successful, the inherent fallibility of personalized recommender systems implies that there may be sources of error that promote distrust. Therefore, I want to test which type of error—wrong recommendations or missing interesting information—has a greater negative effect on trust in the context of social activity stream recommendations.

5.1 Goals and Tool Design

The goals of this study are as follows:

1. To understand how two competing presentation methods affect subjective measures of trust and accuracy in users' real-world social activity streams
2. To investigate the effects of intentional errors within those streams

I sought to achieve these two goals respectively by (1) gathering trust and accuracy feedback from participants who viewed timelines with two different presentation methods and (2) artificially emphasizing lowly-recommended Tweets or de-emphasizing highly-recommended Tweets at different times within those timelines.

To accomplish this, I needed to modify the original ViTA application in order to experiment with different timeline presentations. Based on user feedback from the earlier pilot studies, there are many changes that could be made to the application to make it more user-friendly or to meet different use cases. Some suggestions included:

- Custom colour themes
- Optimize speed for two-dimensional visualization
- Provide more direct user control over recommender (like the User Volume slider)
- View Tweets from a particular user
- Post new Tweets without leaving the application

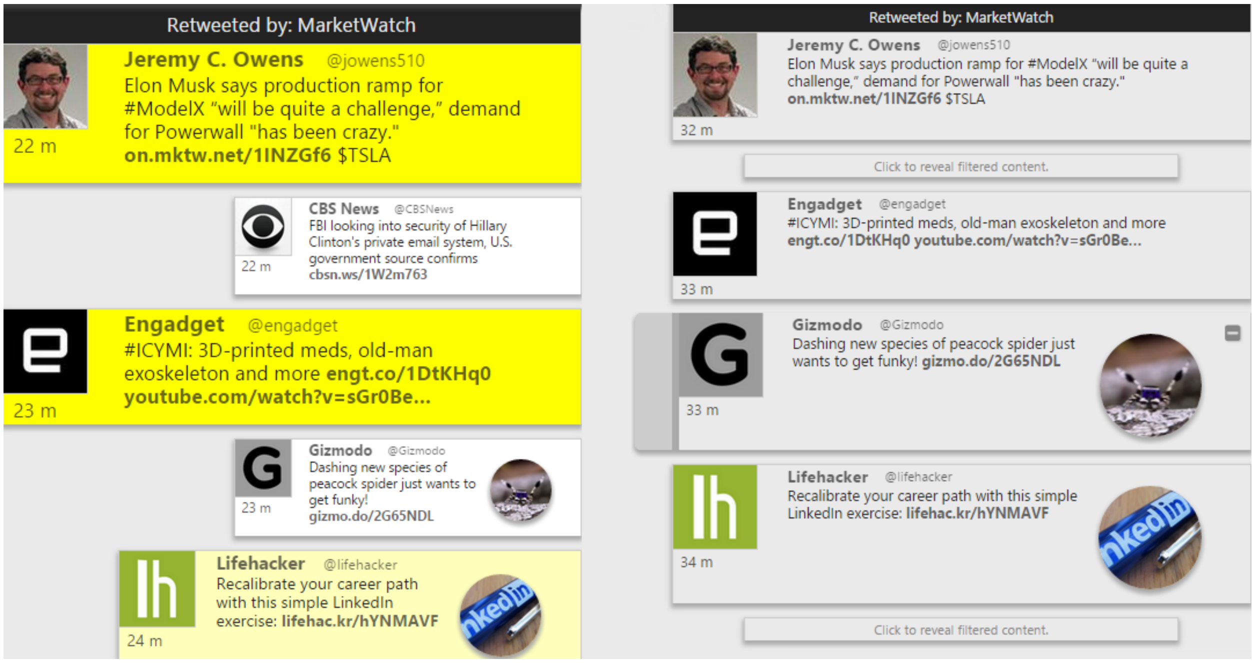
These are all excellent suggestions, and effort would be taken to implement them for a general-purpose Twitter application. However, any new features that would add to the complexity of analysis without adding value or that would distract from the purposes of the study were left unimplemented for this experiment. Instead, I only made changes that exclusively make the application easier to use or understand or that make it easier to measure those things that I intended to measure. The following is a complete list of changes for ViTA 2.0:

- Removal of the User Volume slider
- Removal of the two-dimensional visualization
- Removal of retweet and favorite features
- Addition of a filtered timeline mode where the existence of hidden Tweets is indicated and those Tweets are expandable
- Ability to scroll through the stream using the mouse wheel while the cursor is positioned anywhere on the screen
- A small number of sub-modes in which the recommender exhibits dishonest behaviour for the purposes of the study

- Fewer ratings required to activate the recommender based on research that claims the difference in perceived accuracy gained by gathering more than 10 ratings may not be worth the negative user experience of added effort [Pu et al. 2012]

5.2 Experimental Design

The two primary things I wanted to measure with this study were users' trust in the recommender and perceived recommender accuracy. These measures were evaluated across two groups and three timeline treatments for a total of six treatment combinations. First, the users were divided into two groups in the same way that was done for the ViTA Tempest study. Group A saw timelines with *emphasis*, while Group B saw timelines with *filtering*. The two different timeline presentations are shown in Figure 5.1.



Retweeted by: MarketWatch

Jeremy C. Owens @jowens510
Elon Musk says production ramp for #ModelX "will be quite a challenge," demand for Powerwall "has been crazy." on.mktw.net/1INZGf6 $TSLA
22 m

CBS News @CBSNews
FBI looking into security of Hillary Clinton's private email system, U.S. government source confirms cbsn.ws/1W2m763
22 m

Engadget @engadget
#ICYMI: 3D-printed meds, old-man exoskeleton and more engt.co/1DtKHq0 youtube.com/watch?v=sGr0Be...
23 m

Gizmodo @Gizmodo
Dashing new species of peacock spider just wants to get funky! gizmo.do/2G65NDL
23 m

Lifehacker @lifehacker
Recalibrate your career path with this simple LinkedIn exercise: lifehac.kr/hYNMAVF
24 m

Retweeted by: MarketWatch

Jeremy C. Owens @jowens510
Elon Musk says production ramp for #ModelX "will be quite a challenge," demand for Powerwall "has been crazy." on.mktw.net/1INZGf6 $TSLA
32 m

Click to reveal filtered content.

Engadget @engadget
#ICYMI: 3D-printed meds, old-man exoskeleton and more engt.co/1DtKHq0 youtube.com/watch?v=sGr0Be...
33 m

Gizmodo @Gizmodo
Dashing new species of peacock spider just wants to get funky! gizmo.do/2G65NDL
33 m

Lifehacker @lifehacker
Recalibrate your career path with this simple LinkedIn exercise: lifehac.kr/hYNMAVF
34 m

Click to reveal filtered content.

Figure 5.1: Emphasis timeline shown to Group A (left) and filter timeline shown to Group B (right)

Emphasis timelines. The timelines that were shown to the members of Group A were generally the same as the timelines used in the first ViTA pilot study, with some minor cosmetic changes. Specifically, image thumbnails were changed from square to circular, and support was added for Tweet quoting and multiple images, two features new to the Twitter API.

Filtered timelines. Group B's timelines were more similar visually to a regular Twitter timeline that the users would be familiar with. The unfiltered Tweets all had uniform appearance, while the filtered Tweets were shown as default by small horizontal bars bearing the text "Click to reveal filtered content." Once clicked, a bar expanded to full size, appearing as a regular Tweet. However, to distinguish it from the unfiltered Tweets, a grey tab was added to the left-hand side. Users were given the ability to re-collapse these Tweets by clicking a small icon on the right-hand side.

For members of both groups, after logging in with their Twitter accounts, the experiment consisted of five distinct steps:

1. Recommender training
2. Timeline treatment 1
3. Timeline treatment 2
4. Timeline treatment 3
5. Timeline ranking

Recommender training. The first step was identical for both groups. Participants were shown a recent segment of their timeline consisting of 100 Tweets. From this segment they were instructed to select ten favourite and ten least-favourite Tweets to train the recommender system. Recall that the retweet and favorite features were removed from this version of the application, so only the like and dislike features were available. A tally at the top of the screen kept track of how many likes and dislikes the participant had made. The system would accept no more and no less than ten of each type of rating. It was necessary to require the same number of likes and dislikes so that the recommender was better able to predict both positive and negative ratings more equally. Since the timelines were going to be manipulated in later steps, it was important to have a good balance of emphasized and de-emphasized Tweets for those in Group A, and of filtered and unfiltered Tweets for those in Group B. Once twenty total ratings were given, the participant was permitted to continue to the next step.

Timeline treatments. Each participant stepped through the following three timeline treatments, though not all in the same order:

1. *Honest recommendations*: The recommender did its best to personalize the timeline based on the user's preferences.
2. *False positives*: Starting with honest recommendations, the 10% of Tweets with the lowest predicted rating were artificially emphasized or unfiltered, depending on the group.
3. *False negatives*: Starting with honest recommendations, the 10% of Tweets with the highest predicted rating were artificially de-emphasized or filtered, depending on the group.

Naturally, the *false positives* treatment resulted in a louder (Group A) or busier (Group B) timeline, while the opposite was true for the *false negatives* treatment. In each treatment step, each user would see the same Tweets in the same order, so that a direct comparison between treatments would be possible. Participants from both groups were assigned one of three sequences in which to step through the three timeline treatments. The three sequences were created by a Latin Square design and distributed evenly based on the order in which the participants first logged in. After reading through each step completely, taking at least sixty seconds per step, participants were asked to rate the timeline on a six-point Likert scale for accuracy and trust by answering the following two questions:

1. [Group A:] How accurate was the recommender in emphasizing interesting Tweets and de-emphasizing uninteresting Tweets?

   [Group B:] How accurate was the recommender in filtering out uninteresting Tweets?
2. How much would you trust this recommender to show you the most important information in day-to-day use?

Timeline ranking. After all three timeline treatment steps were completed, participants were required to choose the best and worst treatment by number (first, second, or third) on each measure (accuracy and trust). The reason for this seemingly redundant task is twofold: first, it requires participants to break ties in cases where they assigned the same rating to two or more treatments, and second, it serves as a sort of test question, to identify where responses that should agree with each other do not; participants who give the lowest ranking to a treatment they end up ranking first in the end likely have not answered truthfully or carefully throughout.

Participants were recruited using CrowdFlower. I conducted an initial run of 20 users to ensure everything was working as expected before opening it up on a larger scale. In total, 333 users completed the study. For this study I added a few more safety measures to ensure a higher standard of data quality. The verification code was generated in the same way as for the ViTA Tempest study, but this time I added regular expression validation to the form on CrowdFlower so that users were far less likely to submit a response without completing the survey. Before displaying anything to the user, the application itself also checked whether the logged-in Twitter user had already submitted a response to the survey. This was done to prevent the duplicate submissions that appeared in the previous study.

5.3 Hypotheses

As part of the analysis we have the following hypotheses to explore:

H1. Accuracy and trust ratings will be higher for Group A (with emphasis) than for Group B (with filtering).

H2. Accuracy and trust ratings will be highest for the *honest recommendations* timeline treatment.

H3. Accuracy and trust ratings will be lowest for the *false negatives* timeline treatment.

H4. Accuracy and trust ratings will be strongly and positively correlated ($r \geq 0.7$).

The reason I expect *false positives* to do better than *false negatives* is that I expect participants to punish more greatly recommenders that hide potentially interesting information than those that highlight uninteresting information erroneously, since the inconvenience caused to the reader by missing interesting information is potentially more severe. If the participants were unaware of the interesting posts that were hidden, then this may not be the case. However, the instructions say to read through the timeline completely, including hidden or de-emphasized Tweets.

5.4 Results

We start with a top-level overview of the differences between the groups, then further break down the findings by timeline treatments and, finally, by sequence configuration to investigate any potential ordering effects that may have survived. Looking at the ratings given by both groups for accuracy and trust, it is most natural to compare the honest timelines only, since we are ignoring timeline manipulation for now. Summaries of the relevant data for accuracy and trust are shown in Figures 5.2 and 5.3.

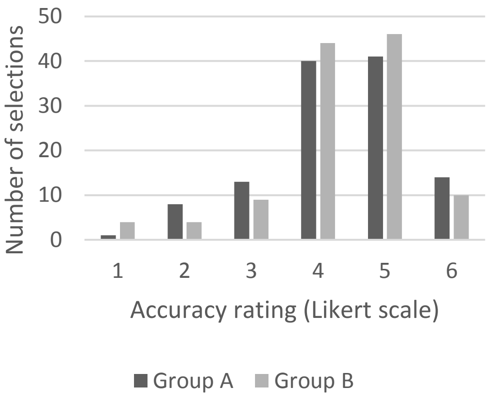


Figure 5.2: Accuracy ratings for honest timeline treatment

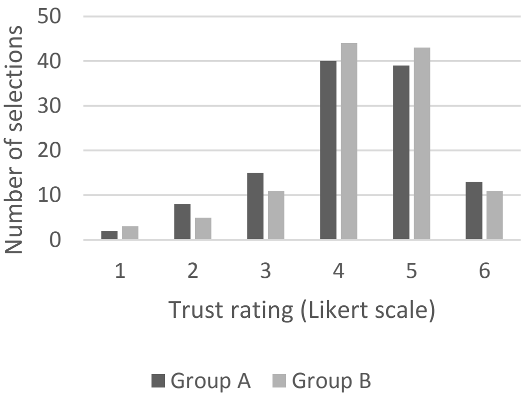


Figure 5.3: Trust ratings for honest timeline treatment

As shown in Table 5.1, the Mann-Whitney test produces *z*-scores of −0.11588 for accuracy ratings and −0.40559 for trust ratings. These correspond to *p*-values of 0.90773 and 0.68504 respectively, meaning that these results are well within the expected range if we assume no differences between the groups. Therefore, we accept that there is no significant statistical difference in our subjective measures of accuracy and trust between the two different timeline presentation methods, and we need to reject H1.

Table 5.1: Mann-Whitney test statistics comparing ratings by groups A and B

	Treatment(s)	***U***	***μ***	***σ***	***z***
Accuracy	Honest	6784.5	6844.5	517.76	−0.12
	All Treatments	63490.0	61600.5	2686.55	0.70
Trust	Honest	6634.5	6844.5	517.76	−0.41
	All Treatments	61007.0	61600.5	2686.55	0.22

From the graphs in Figures 5.2 and 5.3 it seems that Group B gave slightly higher ratings overall in this sample than Group A on both variables measured. While Group A gave low scores (2–3) more often, it also gave the highest score (6) more often for both accuracy and trust than did Group B. However, on the whole, this finding defies the expectation that Group A would score higher overall on both metrics. Unsurprisingly, across the entire data set, accuracy ratings correlated strongly to trust ratings, with a correlation coefficient of 0.85699, which satisfies hypothesis H4. Participants gave identical ratings for accuracy and trust in 72% of all survey responses.

Hypotheses H2 and H3 deal with the differences between timeline treatments. So far we have only examined the cross-group effects of the *honest recommendations* timeline treatment, but here we want to compare the differences in accuracy and trust ratings between all three treatments and determine whether those differences are consistent between the two groups as well. We can establish whether the difference between two treatments is statistically significant on either accuracy or trust by performing Mann-Whitney tests on each pairwise combination of treatments. However, using ratings as we did to compare between groups will make comparisons more difficult; an evaluation using rankings in this case will actually give more information. This is because, as shown in Table 5.2, 32.5% of all participants gave the exact same accuracy rating to all treatments, and 39.3% gave the same trust rating to all treatments. Another 41.5% and 39.7% gave ratings within a range of just one step, for accuracy and trust respectively. So clearly for many participants there was little to choose between the three timeline treatments. As mentioned earlier, the tiebreaker ranking was introduced in anticipation of this problem, so that we could look at which timelines users would choose if they needed to pick just one.

Table 5.2: Occurrence rates for differences between participants' highest and lowest ratings—a rating range of 0 represents users that gave the same rating to all timeline treatments

Rating range	**Rate of occurrence (Accuracy)**	**Rate of occurrence (Trust)**
0	32.5%	39.3%
1	41.5%	39.7%
2	16.7%	14.5%
>2	9.4%	6.4%

When using ratings to calculate test statistics, as shown in Table 5.3, the only statistically significant difference is between the *honest recommendations* and *false positives* timeline treatments when measuring accuracy for the filter group (B), and, being a two-tailed test, even this difference is only significant at the $p < 0.10$ level.

Table 5.3: Mann-Whitney test statistics comparing accuracy and trust ratings and rankings by timeline treatments

Variable	Group	Treatment Comparison	Rating ***z***	Rank ***z***
Accuracy	A Emphasis	Honest vs. False Neg.	0.37	0.49
		Honest vs. False Pos.	0.70	−0.26
		False Neg. vs. False Pos.	0.44	−0.64
	B Filter	Honest vs. False Neg.	1.19	1.60
		Honest vs. False Pos.	1.71*	2.01**
		False Neg. vs. False Pos.	0.29	0.47
Trust	A Emphasis	Honest vs. False Neg.	0.70	1.84**
		Honest vs. False Pos.	0.47	0.19
		False Neg. vs. False Pos.	−0.21	−1.32
	B Filter	Honest vs. False Neg.	1.21	1.50
		Honest vs. False Pos.	0.90	1.33
		False Neg. vs. False Pos.	−0.34	0.02

However, when we use rankings, this result is significant at the $p < 0.05$ level, and there is also a significant difference at the $p < 0.05$ level between *honest recommendations* and *false negatives* when measuring trust for the emphasis group (A). Though the trend is that *honest recommendations* did outperform the other timeline treatments on both measures across both groups within this sample, the difference is not statistically significant overall. *False positives*, on the other hand, only outperformed *false negatives* on measures of trust, and it did so to a statistically insignificant degree. For these reasons we ultimately must reject H2 and H3. However, there are some other interesting observations we can make about the data obtained from this particular sample.

In Table 5.3, positive *z*-scores correspond to an overall higher rank of ratings for the first treatment listed in the middle column. So, for example, when looking at the *Rating z* column, for all comparisons except *Negatives vs. Positives* when measuring trust (which have negative *z* scores for both groups), the first listed treatment scored higher. While the statistical tests indicate that we should not make conclusions about the population in general, this tells us that, in our sample, true to expectations, the *honest recommendations* treatment seemed to outperform both the *false negatives* and *false positives* treatments on both measures in both groups, though *false positives* did better on accuracy for the emphasis group when only rank is considered. Participants in our sample also rated the *false negatives* timeline as slightly more accurate, while rating it slightly less trustworthy, particularly among members of the emphasis group.

While these observations hold true across both groups, there are some noteworthy differences between the groups as well. These differences are calculated by comparing ratings along the Likert scale, and the test statistics are summarized in Table 5.4. Rankings could not be used in this comparison because participants had no way to compare to a timeline they had not seen. The perceived accuracy of the *honest recommendations* treatment was relatively higher than the other two for the filter group compared to the emphasis group. We also see a similar, though slightly lesser effect when looking at trustworthiness. However, as alluded to above, the *false negatives* treatment performed relatively better for the emphasis group than for the filtering group, meaning that it wasn’t as problematic when important content was de-emphasized as it was when important content was hidden from the filter group. While the filter group gave higher ratings for the *honest recommendations* timeline, comparing timeline treatments across groups, we see the emphasis group rating the other timeline treatments as more accurate, while trustworthiness scores from both groups were essentially equivalent.

Table 5.4: Mann-Whitney test statistics comparing ratings by groups A and B for different timeline treatments

Variable	Treatment	***U***	***z***
Accuracy	False Negatives	7216.0	0.72
	False Positives	7182.5	0.65
Trust	False Negatives	6893.5	0.09
	False Positives	6822.5	−0.04

A closer look at the data reveals that the only advantage the emphasis timeline had over the filter timeline in this study was in the cases of deliberately inaccurate recommendations. Interestingly, though, analyzing scores in the context of their sequence (i.e. looking at treatments as *first*, *second*, or *third* instead of as *honest recommendations*, *false negatives*, or *false positives*) reveals that the emphasis group A gave higher ratings on average for accuracy at all steps, and for trust at steps 2 and 3. We can further divide each group into those participants who were assigned each of the three sequences of timeline treatments:

1. {*false negatives*, *honest*, *false positives*}
2. {*false positives*, *false negatives*, *honest*}
3. {*honest*, *false positives*, *false negatives*}

The ratings given by participants separated by group and sequence are summarized in Figure 5.4, while the final rankings are shown in Figure 5.5. Looking at the set of participants broken down in this way, two discoveries stand out above the rest:

1. The *false positives* timeline treatment was extremely polarizing on both measures for members of the emphasis group (A): while it was chosen almost equally as lowest and highest rank, only 12.2% of these participants implicitly placed it in the middle.
2. Among those participants who saw treatment sequence (1), those who saw the emphasis timeline (A) gave generally high ratings across the board, while those who saw the filtered timeline (B) gave much lower scores to the *honest* and *false positives* treatments.

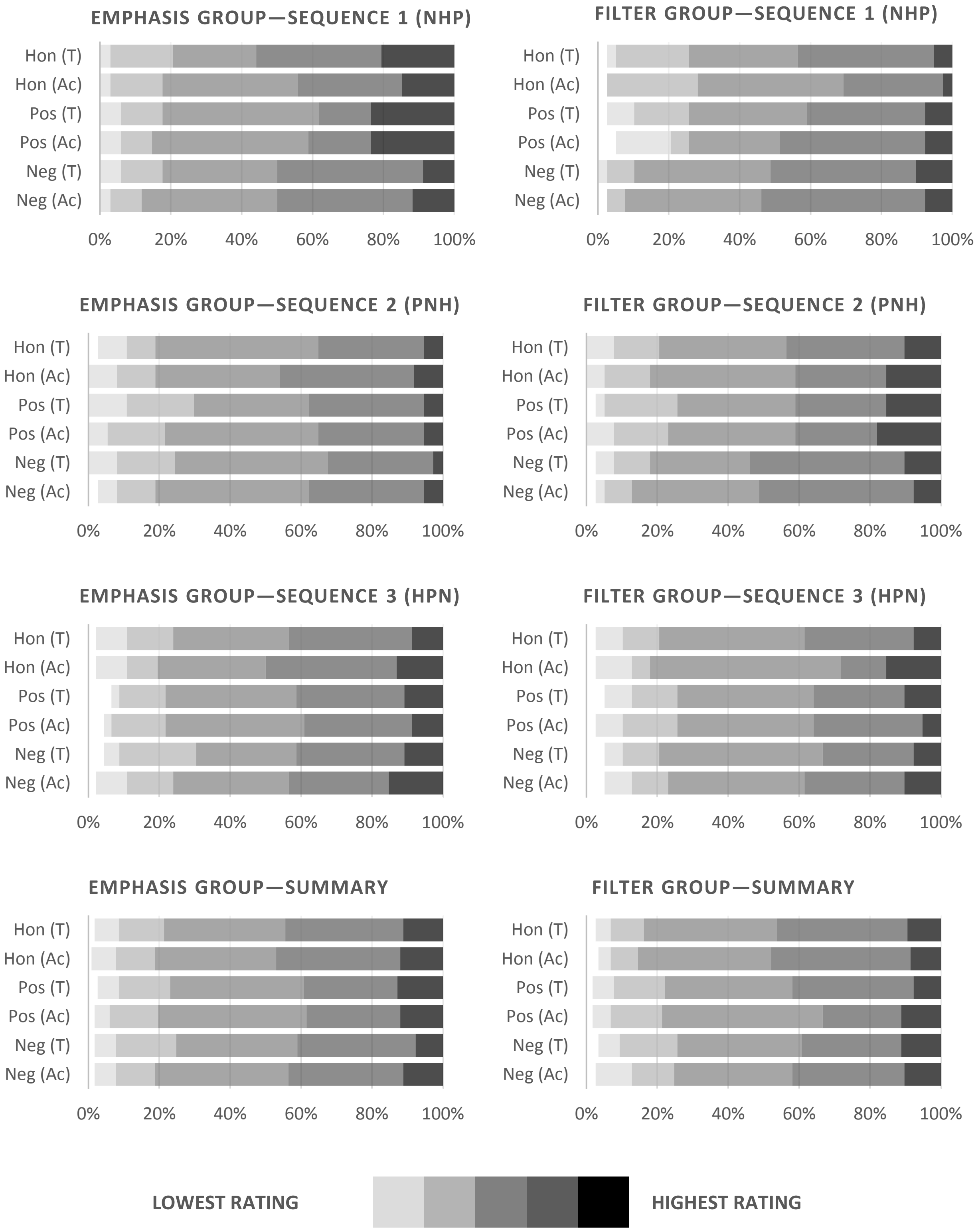


Figure 5.4: Accuracy (Ac) and trust (T) ratings for each treatment, by group and sequence assignment

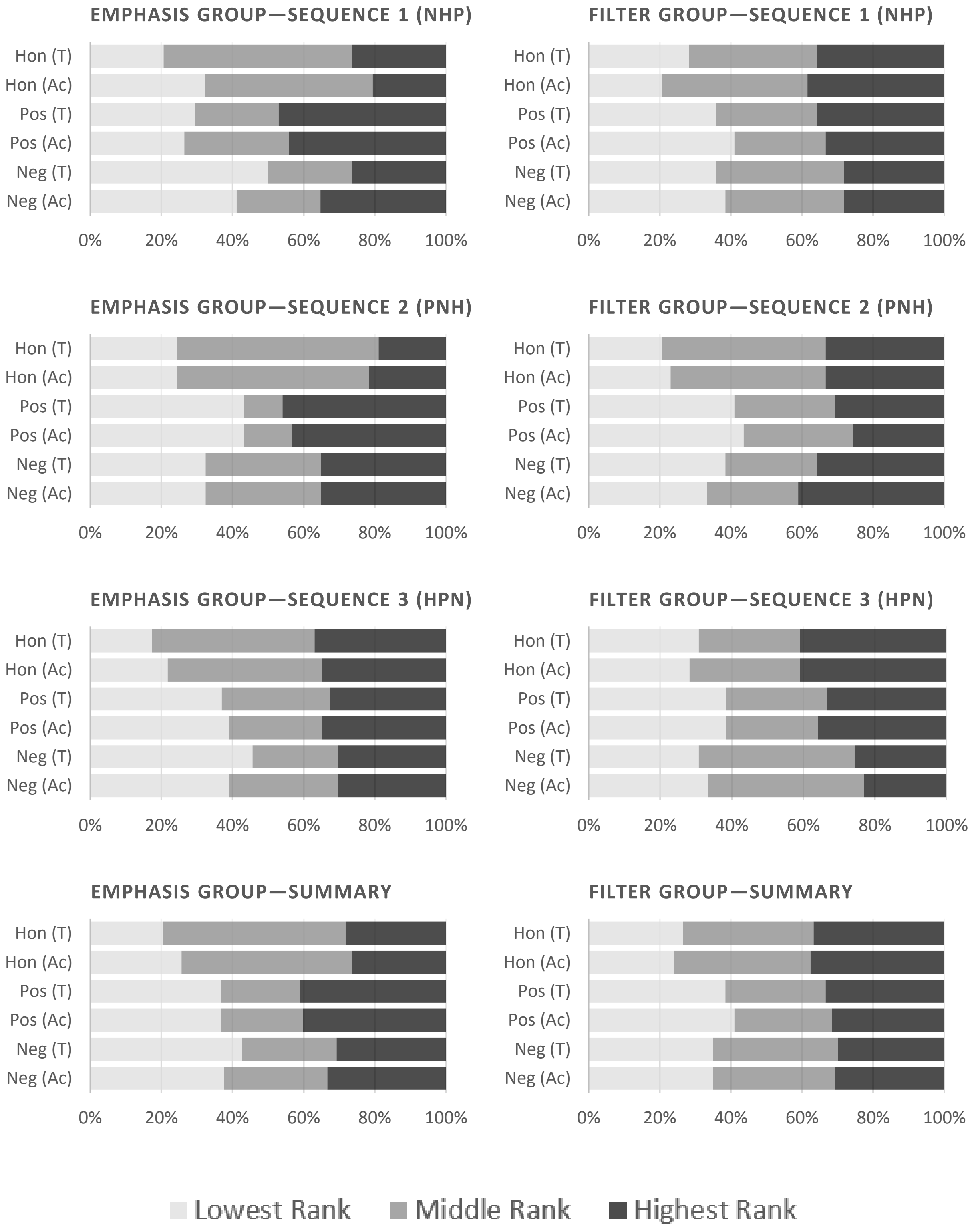


Figure 5.5: Accuracy (Ac) and trust (T) rankings for each treatment, by group and sequence assignment

5.5 Discussion

The primary goal of this study was to investigate whether a social activity stream showing recommendations styled with the ViTA presentation method would be able to outperform traditional stream filtering on measures of trust and perceived accuracy. At this size of sample, the data obtained from this study was not sufficient to conclude that either stream would be perceived as more accurate or trustworthy than the other by the general Twitter-using population. However, the fact that the ViTA method was not itself conclusively outperformed is an important finding, given that:

1. Filtering is a currently common standard—existing social activity streams such as Facebook's News Feed do currently engage in stream manipulation including reordering and filtering to reduce information overload and increase user engagement.
2. The filtered timeline used in this study offers the ability to view hidden content, which has been shown to enhance trust in such streams [Nagulendra & Vassileva 2013]. This kind of viewable filtering is not standard, but is more of an advanced baseline for comparison.
3. Presenting recommendations using emphasis does offer some unique inherent value, as shown by its persuasive potential observed in the ViTA Tempest study.
4. There is likely much room for improvement to the ViTA design, whereas strict filtering is essentially an on/off setting.

Furthermore, we may be able to learn from these results specific cases in which emphasis might have more of an advantage over filtering.

The purpose of the timeline treatments was to try to measure the effect of false negatives and false positives within social activity streams to see which one users tended to punish more, depending on the presentation method. The way that was done in this study was to take the lowest- or highest-predicted items and place them in the highest or lowest tier, respectively. However, an unintended consequence of this manipulation was that the distributions of Tweets within the tiers was not re-balanced by shifting other items up or down. As a result of this oversight, a potential confounding variable was introduced. Instead of measuring purely the effects of blatantly wrong positive recommendations versus blatantly wrong negative recommendations, the study may have been measuring the effects of seeing many large yellow Tweets (a busier looking timeline) versus seeing many small white Tweets (a quieter looking timeline) or seeing few versus many Tweets filtered out. The busy timeline may result in more cognitive overload for some users than even an untreated one, but the quiet timeline could appear dull and disappointing. While all of these effects would be interesting to study on their own, ultimately we see the results of the error effects and the timeline saturation interacting with each other, which makes it impossible to isolate either of the variables completely. It is also possible that this confounding variable made it impossible to observe a significant effect; it is conceivable, for example, that users enjoyed reading the busier timelines more, which dampened the negative effect of the inaccurate recommendations.

As we saw in Table 5.3, ratings for trust suffered across both groups when the recommender artificially de-emphasized or hid Tweets predicted to be of high value. This is important and falls in line with expectations that users would be more forgiving of briefly seeing very uninteresting content than knowing they are likely to be missing very interesting content. However, it is equally important to note that this is not just a function of the timeline containing fewer emphasized, or more hidden, Tweets. This easily can be understood by recognizing that the trust ratings were also lower when too much uninteresting content was promoted in the timeline. Meanwhile, accuracy ratings tell a slightly different story: in this case, it was worst to emphasize, or just not filter out, the uninteresting content. These errors may have been more visible because they were highlighted (emphasis group) or not hidden (filter group), so it is not surprising that they would impact a pure accuracy assessment more strongly. This also helps highlight an important distinction between accuracy and trust. When a recommender is wrong but it puts its inaccurate predictions front and centre for the user to see and evaluate, it is rightly perceived as an inaccuracy but not as untrustworthy behaviour. Conversely, when a recommender is wrong and hides something important from the user, this is likely seen as deceptive behaviour, which naturally impacts trust negatively. I suspect that the difference in accuracy impact can be explained simply by the difference in visibility of the errors, but I would have expected the effect to be more obvious with the emphasis group in comparison to the filter group than it was. More focused future study is needed to confirm these findings with statistical significance, but a potential takeaway is that minimizing or hiding content should be done carefully, with a low threshold, if maximizing trust is vital.

The *honest recommendations* timeline treatment was the happy medium for both groups on both measures, but it got higher scores from the filter group relative to the other treatments. As discussed above and shown in Tables 5.1 and 5.4, though the preferred combination overall seemed to be *filtered honest recommendations*, the other two timeline treatments did much better for the emphasis group than for the filter group, thanks to their accuracy ratings. There is a clear difference between the two timeline presentation methods that may explain this effect: the emphasis group saw Tweets placed into three tiers of predicted interest or relevance, while the filter group saw only two. Further, since the lowest tier for the filter group corresponded exactly to the lowest tier for the emphasis group, the top-rated two-thirds of Tweets for the filter group were all treated equally. This means that, out of 100 Tweets, the 1st-ranked Tweet and the 66th-ranked Tweet would have been given the same prominence in the timeline, with neither being recommended more strongly than the other.[10] If this is truly the reason, then we should expect the *honest recommendations* timeline treatment to outperform the others by a greater relative amount for the filter group; in fact, this is exactly what was found, as shown in Table 5.3.

The dearth of statistical significance in this study may be a symptom of the necessary lack of structure in the experimental procedure. Unlike the ViTA Tempest study where I was able to control most potential confounding variables, each user brought his or her own Twitter network and personal idiosyncrasies, including timeline consumption behaviours, to the table. While the data was no more subjective in this case, the paths participants took to arrive at their decisions may have varied greatly. I expected the cross-group differences might be difficult to observe, but the data suggests the within-subjects differences between timeline treatments to be more subtle than I had anticipated. I declined to make timeline presentation a within-subjects variable in order to avoid a novelty effect as well as bias to the experimental condition. Since filtering is more common and participants are likely to have encountered it in other streams, they would be likely to figure out that the emphasis timeline was the subject of study. However, seeing the degree of subtlety in the results leads me to believe that it would be worth a future study to present both timelines to all participants and try to measure both subjectively and objectively, perhaps by employing eye tracking and dwell time, which they prefer to use.

Some treatment differences seemed like they might tend toward being inconclusive (e.g. the emphasis group's *false positives* scored negatively on ratings compared to other treatments, but positively on rank), but a still-larger pool of participants is needed to tell us more about how the observed differences will generalize to the population. However, for some of the effects this experiment was designed to study, it is likely not necessary to provide such a real-world Twitter

[10] This is just an example for illustrative purposes; the recommender did not actually assign Tweets to tiers based on tertile rank, but rather on estimated likelihood of fit into one of two classifications based on user ratings during the training step ("like" or "dislike") (see Section 3.2.2 for more details about the recommender system design).

experience. While it may not be possible for comparing emphasis and filtering, future studies might better observe the differences between the effects of false positives and false negatives in a more structured experimental setting without needing a larger sample size.

5.6 Conclusions

Contrary to my goals and hypotheses, the results of this study do not allow us to conclude that a timeline using ViTA's method of presenting recommendations using three tiers of emphasis will be perceived as more accurate or trustworthy overall than a timeline using viewable filtering. We also cannot conclude whether it is more beneficial to subjective measures of accuracy or trust to maximize precision or recall, which I tried to measure by introducing intentionally inaccurate recommendations into users' timelines. However, we can see certain trends emerge within the studied sample that may be able to lead to more concrete results in future study.

One of the primary motivations for developing an emphasis-based presentation method to be used instead of standard filtering was to allow for finer discrimination between levels of recommendation. We saw some unexpected effects of this feature when the filter timeline scored slightly better than the emphasis timeline on accuracy and trustworthiness, but only when no intentionally inaccurate recommendations were included. In cases where they were introduced, the users viewing the emphasis timeline were more forgiving; they seemed less sensitive to the effects of the errors even though the false negatives were immediately visible and the false positives more prominent.

The applications of this apparent strength are not obvious; most people would not think to create a recommender for mass adoption that purposely gets things wrong. However, it does help us understand more about user acceptance of recommendations and the trust that users have in these systems. If we can find a way to get users to be more accepting of potentially erroneous recommendations, then our algorithms can swing for the fences, taking more risks and perhaps trying to offer more serendipitous recommendations to get users out of their filter bubbles.

These results also underscore how a presentation method can influence perceived accuracy and trustworthiness differently. We saw that timelines containing intentional false positive recommendations scored lower on accuracy, while those with false negatives scored lower on trust, regardless of the recommendation presentation method. In addition to what we already know about the positive relationship between transparency and trust, it also seems that recommenders in social activity streams that seek to maximize trust should tend toward optimism rather than pessimism, and should hide or minimize information conservatively.

This study may not have given the conclusive results that were expected, but in some ways this was a positive fact that drove the analysis in a more explorative direction. There are many human factors involved in addition to the different combinations of treatment conditions that add compelling layers to the results. Any one of these layers could be explored with greater focus in future study, as I will discuss in the next chapter.

CHAPTER 6
CONCLUSIONS AND FUTURE WORK

6.1 Conclusions

A piece of valuable information can be completely useless until it reaches the right people. There is a vast amount of information accessible online, but much of it will never be converted to knowledge by those who would value it. This unprecedented accessibility of information has actually led to great practical inaccessibility, as competing content crowds our limited attention, each piece threatened to be lost in the noise. As user-generated content is produced at an ever-increasing rate and its value to its consumers continues to grow, people will spend more and more time consuming information and social activity streams online. Online communities have some natural mechanisms to allow information to shine through: the mechanisms of trending topics and viral sharing organically promote content of common interest to a large group of people. However, items of great value whose target audience may be limited in scope will not rise to the top unaided. Personalized streams are therefore necessary to reduce the amount of visible content for one user, paradoxically, to increase their content consumption. This is the basic solution to the information overload problem, but we cannot stop there. Out of this solution emerge things like the filter bubble problem, and personalized streams can themselves become overwhelming as users seek to engage with a broader range of sources.

The evolution of the social activity stream has been extremely slow, despite numerous potential avenues for improvement. This research explores emphasis-based recommendations, one potential improvement that presents implicit recommendations within a chronological stream by displaying each content item with one of three visual styles to indicate its level of predicted interest or relevance to the user. The main idea of this type of approach is to make it easier for users to get the most out of their streams by allowing them easily to consume the most valuable information with context. However, through studying the impact ViTA had on real users, we also discovered some other non-trivial positive side effects. A combination of structured experimentation and subjective user studies gave us the following answers to our original research questions:

1. The ViTA recommendation presentation method is immediately intuitive. Users require no explanation of the nature of the recommendation tiers in order to understand what was being communicated. The system is also easy to use: participants reported that it was easier to read through the highest tier of recommendations than it was to read the entire timeline, and more than half of the participants in the pilot study reported the highest possible "ease of task" score. Based on limited feedback, users expressed an interest in seeing a similar presentation in their regular Twitter timeline, hinting at a user satisfaction boost.
2. Stream summary visualizations may be able to play a complementary role in stream consumption, to give context and additional insights, but the two-dimensional ViTA visualization was not preferred to the more familiar one-dimensional timeline in a limited pilot study.
3. The ViTA design is highly persuasive as demonstrated in a structured exploratory study that showed arbitrary recommendations to users under of the guise of personalized or popularity-based algorithms. Users agree with the recommendations that are shown to them much more often than chance would dictate, particularly with the highly emphasized items. This finding persists regardless of whether the users believe the recommendations are personalized or based on favorites and retweets from other users.

4. ViTA compares favorably in users' real Twitter timelines to an elastic filtered timeline that allows users immediately to view posts within context that have been hidden from them, both in terms of trustworthiness and perceived accuracy.
5. ViTA seems to be perceived as more accurate when erroneous recommendations are introduced to timelines, but slightly less accurate in the absence of intentional errors. Regardless of presentation method, users are less trusting of a system that hides interesting content than one that promotes uninteresting content. However, users perceive filter timelines with added false negatives as more accurate than timelines with added false positives.

In addition to these contributions, the tools used to conduct these experiments may also be useful in the real world to social media users or tool developers, or by researchers as part of further scientific study. The ViTA design itself could be incorporated into any sort of information or social activity stream where the order of items is important. The ViTA Tempest study, with some modifications, can be a highly structured experiment that can be used as a standard test to empirically measure the persuasiveness of a wide variety of recommendation presentation designs separately from any recommender algorithms. And in addition to these tools, this research contributes insight into the human factors of users of recommender systems in social activity streams. We know better how perceived accuracy interacts with trust and how algorithmic accuracy tells far from the whole story even when recommending items from a tiny set of social updates. But perhaps most of all, the findings of this research kindle curiosity and invite more focused investigation in a number of different directions for future study.

6.2 Future Work

There are many potential applications for emphasis-based recommendations; the Twitter timelines studied here are just one possible target, and the ViTA implementation is just one possible variation. In fact, different variations may be more effective in different types of streams, and it is unlikely that the ViTA design is ideal for Twitter timelines. This section outlines just some of the directions that this research could take in future work.

Design optimization. The design choices made during ViTA's development were not arbitrary, but they were also not empirically evaluated against any alternative designs. One or more future studies could compare different values of colour, size, horizontal offset, and number of tiers to try to isolate the effects and optimize each dimension.

Baseline persuasiveness. Further to optimizing the design, steps should be taken to establish a baseline level of persuasiveness in Tempest. We know ViTA was persuasive, but we actually don't know how persuasive it was, because simply telling users which Tweets were recommended and which were not (using plain text or arrows or asterisks, for example) would likely have been measurably persuasive as well. Once a baseline has been established, only then can the marginal persuasiveness of competing designs be known.

Stream summary visualization as complement. Though the two-dimensional version of ViTA does not seem like it would be able to stand on its own, there is likely some benefit to offering a visual overview of a social activity stream to enhance awareness and aid navigation, much like Shi's Rings visualization [Shi 2010] does for understanding activities of different friends in Facebook. Rather than trying to utilize a second dimension, which may cause users to become lost due to our more linear concept of time, perhaps a mini-map could be shown next to the full-size one-dimensional stream that indicates where highly- and lowly-recommended content can be found by scrolling through time in either direction, similar to Indratmo's iBlogVis [Indratmo 2010]. Colour might be used to indicate recommendation strength, while line length could represent popularity. Such a visualization would take nothing away from the users' ability to consume their stream, but could indicate where content fits on a timeline to show periods of high or low activity, as well as marking bursts of interesting updates.

Different applications. These studies looked at Twitter, where all updates are a maximum of 140 characters with optional photo or video attachments or inline quoted Tweets. The value of a presentation method like ViTA may be greater or lesser for a different type of social activity stream. Instagram, for example, is a stream of photos and short videos in which posts could easily be treated with similar tiers of emphasis. ViTA may work better with images than text because it is often easier to get a gist of an image at a glance than it is to read very small text. However, since users would be able to discern the value of each post more easily merely by skimming, levels of emphasis may not be required; they may even become annoying. Future study should evaluate the effectiveness of emphasis streams in various information streams such as email inboxes, to-do lists, notification centres, and rich media streams like Instagram.

Tempest improvements. In Tempest we saw the effects of asking users to choose the most interesting Tweet before the least interesting Tweet for each set. Highly emphasized Tweets may have been chosen as interesting more often simply because they were chosen *first* more often. Before using Tempest to evaluate other design variations, this issue should be addressed by varying the task order. Also, instead of displaying a set's Tweets with completely randomized emphasis, a counterbalanced or Latin Square design should be used in order to isolate the order effects better.

Isolate false positives and false negatives. A drawback of the last study was that the distribution of Tweets between emphasis tiers was not balanced in the *false negatives* and *false positives* treatment conditions. This meant that when we tried to examine the effects of erroneous recommendations we were also examining the effects of more or less busy timelines. A simple solution for the *false positives* treatment, for example, would be to move some of the less highly-recommended emphasized Tweets into the neutral tier and move some of the less-highly recommended neutral Tweets into the de-emphasis tier to keep a balance. Alternatively, intentional errors could be examined outside of a real timeline in a more structured setting similar to the Tempest experiment. For example, when showing the user what is supposed to be a set of highly-recommended Tweets, one Tweet predicted to be uninteresting could be inserted. Likewise, an interesting Tweet could be inserted into a set of uninteresting. After reading through both sets, the system could force the users to pick which one did its job more accurately. Future experiments such as this could help give more insight into whether recall or precision is more crucial to optimize when presenting recommendations in social activity streams.

Within-subjects presentation treatments. Given the subtlety of the differences between the emphasis group and the filter group in our last study, it may be worth showing both timeline types to a group of participants and asking which they prefer. Often, especially in the area of new technology, people do not know what they want until it is shown to them. Such a study may show that a significant result is possible without needing a much larger sample, but at the very least it would allow researchers to investigate what users who prefer one of the two varieties have in common.

Clearly this research only scratches the surface of possibilities for emphasis-based recommendation presentation methods. I believe that as more and more information is pushed from traditional media to consumable streams, this research will only increase in importance. Where social activity streams once were enough to deal with information overload, these streams themselves can become cognitively overwhelming, and as recommendations within these streams become more necessary and more prevalent, we need to consider more carefully how users experience them. This ought to inspire others who similarly find real value in social media to pursue this path in order to help information get to users more effectively and accurately. The humble designs and results discussed here build off of a vast amount of practical and theoretical work that has come before me, and I hope that my work can serve as a launch pad for others interesting in advancing the field even further.

REFERENCES

Adomavicius, G., & Tuzhilin, A. (2005). Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering, 17* (6), 734–749.

Adomavicius, G., Bockstedt, J. C., Curley, S. P., & Zhang, J. (2013). Do recommender systems manipulate consumer preferences? A study of anchoring effects. *Information Systems Research, 24* (4), 956–975.

Alba, D. (2014, December 12). *Instagram's £22 billion valuation eclipses Twitter*. Retrieved March 31, 2015, from http://www.wired.co.uk/news/archive/2014-12/22/instagram-eclipsing-twitter.

Bawden, D., & Robinson, L. (2009). The dark side of information: overload, anxiety and other paradoxes and pathologies. *Journal of Information Science, 35* (2), 180–191.

Beevolve. (n.d.). *An Exhaustive Study of Twitter Users Across the World*. Retrieved March 24, 2015, from http://www.beevolve.com/twitter-statistics.

Berghel, H. (1997). Cyberspace 2000: Dealing with information overload. *Communications of the ACM, 40* (2), 19–24.

Billsus, D., & Pazzani, M. J. (2000). User Modeling for Adaptive News Access. *User Modeling and User-Adapted Interaction, 10* (2–3), 147–180.

Burke, R. (2002). Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction, 12* (4), 331–370.

Chorley, M. J., Colombo, G. B., Allen, S. M., & Whitaker, R. M. (2015). Human content filtering in Twitter: The influence of metadata. *International Journal of Human-Computer Studies, 74*, 32–40.

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association, 79* (387), 531–554.

Cosley, D., Lam, S. K., Albert, I., Konstan, J., & Riedl, J. (2003). Is seeing believing?: how recommender system interfaces affect users' opinions. *Proceedings of the SIGCHI conference on Human factors in computing systems,* 585–592. ACM.

Cremonesi, P., Garzotto, F., & Turrin, R. (2012). Investigating the persuasion potential of recommender systems from a quality perspective: An empirical study. *ACM Transactions on Interactive Intelligent Systems (TiiS), 2* (2), 11:1–41.

Deller, M., Ebert, A., Bender, M., Agne, S., & Barthel, H. (2007). Preattentive visualization of information relevance. *Proceedings of the international workshop on Human-centered multimedia*, 47–56. ACM.

Dunbar, R. (1992). Neocortex size as a constraint on group size in primates. *Journal of Human Evolution, 22* (6), 469–493.

Dunbar, R. (2011). How many "friends" can you really have? *Spectrum, 48* (6), 81–83.

eMarketer. (2014, September 18). *Facebook's US Ad Revenues Outpace Users' Average Daily Time Spent on the Site*. Retrieved March 31, 2015, from http://www.emarketer.com/Article/Facebooks-US-Ad-Revenues-Outpace-Users-Average-Daily-Time-Spent-on-Site/1011215.

Facebook. (2014, February 19). *Facebook to Acquire WhatsApp*. Retrieved March 31, 2015, from http://newsroom.fb.com/news/2014/02/facebook-to-acquire-whatsapp.

Facebook. (2015, January 28). *Facebook Reports Fourth Quarter and Full Year 2014 Results*. Retrieved March 31, 2015, from http://investor.fb.com/releasedetail.cfm?ReleaseID=893395.

Felfernig, A., Friedrich, G., Gula, B., Hitz, M., Kruggel, T., Leitner, G., Melcher, R., Ripena, D., Strauss, S., Teppan, E., & Vitouch, O. (2007). Persuasive recommendation: serial position effects in knowledge-based recommender systems. *Second International Conference on Persuasive Technology*, 283–294.

Friedrich, G., & Zanker, M. (2011). A taxonomy for generating explanations in recommender systems. *AI Magazine, 32* (3), 90–98.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM - Special issue on information filtering, 35* (12), 61–70.

Grisé, M.-L., & Gallupe, R. B. (1999). Information overload: Addressing the productivity paradox in face-to-face electronic meetings. *Journal of Management Information Systems*, 157–185.

Gunawardana, A., & Shani, G. (2009). A Survey of Accuracy Evaluation Metrics of Recommendation Tasks. *The Journal of Machine Learning Research, 10,* 2935–2962.

Harman, J. L., O'Donovan, J., Abdelzaher, T., & Gonzalez, C. (2014). Dynamics of human trust in recommender systems. *Proceedings of the 8th ACM Conference on Recommender systems,* 305–308. ACM.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS), 22* (1), 5–53.

Indratmo. (2010). *Supporting exploratory browsing with visualization of social interaction history.* Ph.D. Dissertation. University of Saskatchewan, Saskatoon, Canada.

Ipsos. (2013, January 8). *Socialogue: The Most Common Butterfly on Earth Is the Social Butterfly*. Retrieved March 31, 2015, from http://ipsos-na.com/news-polls/pressrelease.aspx?id=5954.

Ipsos MediaCT. (2014, March). *Social Influence: Marketing's New Frontier*. Retrieved August 6, 2015, from http://corp.crowdtap.com/socialinfluence.

Kilgore, T. (2015, March 20). *Facebook's stock-market valuation tops $230 billion*. Retrieved March 31, 2015, from http://www.marketwatch.com/story/facebooks-stock-market-valuation-tops-230-billion-2015-03-20.

Laporte, L., van Nimwegen, C., & Uyttendaele, A. J. (2010). Do people say what they think: Social conformity behavior in varying degrees of online social presence. *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries,* 305–314. ACM.

Lapowsky, I. (2013, October 4). *Ev Williams on Twitter's Early Years*. Retrieved March 31, 2015, from http://www.inc.com/issie-lapowsky/ev-williams-twitter-early-years.html

Losee, R. M. (1989). Minimizing information overload: the ranking of electronic messages. *Journal of Information Science, 15* (3), 179–189.

Lü, L., Medo, M., Yeung, C. H., Zhang, Y.-C., Zhang, Z.-K., & Zhou, T. (2012). Recommender Systems. *Physics Reports, 519* (1), 1–49.

Lynley, M., & Edwards, J. (2014, February 3). *This Was the First News Story Ever Written About 'TheFacebook.com,' from the Site's Birth 10 Years Ago*. Retrieved March 31, 2015, from http://www.businessinsider.com/first-news-article-about-facebook-2014-2.

Mackinlay, J. (1986). Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics (TOG), 5* (2), 110–141.

Maes, P. (1994). Agents that reduce work and information overload. *Communications of the ACM, 37* (7), 30–40.

Malhotra, N. K., Jain, A. K., & Lagakos, S. W. (1982). The information overload controversy: An alternative viewpoint. *The Journal of Marketing*, 27–37.

McNee, S. M., Lam, S. K., Konstan, J., & Riedl, J. (2003). Interfaces for Eliciting New User Preferences in Recommender Systems. *9th International Conference, UM 2003, Proceedings*, 178–187.

McNee, S. M., Riedl, J., & Konstan, J. (2006). Being accurate is not enough: how accuracy metrics have hurt recommender systems. *CHI '06 Extended Abstracts on Human Factors in Computing Systems,* 1097–1101. New York: ACM.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63* (2), 81.

Nagulendra, S., & Vassileva, J. (2013). Providing Awareness, Understanding and Control of Personalized Stream Filtering in a P2P Social Network. *19th International Conference, CRIWG 2013, Proceedings*, 61–76.

Nanou, T., Lekakos, G., & Fouskas, K. (2010). The effects of recommendations' presentation on persuasion and satisfaction in a movie recommender system. *Multimedia systems, 16* (4–5), 219–230.

O'Brien, T. (2014, October 17). *The spirit of experimentation and the evolution of your home timeline*. Retrieved March 31, 2015, from https://blog.twitter.com/2014/the-spirit-of-experimentation-and-the-evolution-of-your-home-timeline.

Oppenheim, C. (1997). Managers' use and handling of information. *International Journal of Information Management, 17* (4), 239–248.

O'Reilly, C. A. (1980). Individuals and information overload in organizations: Is more necessarily better? *Academy of Management Journal, 23* (4), 684–696.

Oyeka, I. C., & Ebuh, G. U. (2012). Modified Wilcoxon Signed-Rank Test. *Open Journal of Statistics, 2* (2), 172–176.

Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You.* Penguin UK.

Pu, P., & Chen, L. (2006). Trust building with explanation interfaces. *Proceedings of the 11th international conference on Intelligent User Interfaces,* 93–100. ACM.

Pu, P., Chen, L., & Hu, R. (2012). Evaluating recommender systems from the user's perspective: survey of the state of the art. *User Modeling and User-Adapted Interaction, 22* (4–5), 317–355.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). GroupLens: an open architecture for collaborative filtering of netnews. *CSCW '94 Proceedings of the 1994 ACM conference on Computer supported cooperative work,* 175–186. New York: ACM.

Ries, E. (2011). *The Lean Startup: How today's entrepreneurs use continuous innovation to create radically successful businesses.* Random House LLC.

Rogers, E. M., & Agarwala-Rogers, R. (1975). Organizational communication. *Communication behaviour*, 218–239.

Sagolla, D. (2009, January 30). *How Twitter Was Born*. Retrieved March 31, 2015, from http://www.140characters.com/2009/01/30/how-twitter-was-born.

Sanghvi, R. (2006, September 5). *Facebook Gets a Facelift*. Retrieved March 31, 2015, from https://www.facebook.com/notes/facebook/facebook-gets-a-facelift/2207967130.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2000). Analysis of recommendation algorithms for e-commerce. *EC '00 Proceedings of the 2nd ACM conference on Electronic commerce*, 158–167. New York: ACM.

Schick, A. G., Gordon, L. A., & Haka, S. (1990). Information overload: A temporal approach. *Accounting, Organizations and Society, 15* (3), 199–220.

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., & Lehman, D. R. (2002). Maximizing versus satisficing: Happiness is a matter of choice. *Journal of Personality and Social Psychology, 83* (5), 1178–1197.

Sherr, I. (2015, January 6). *Facebook's WhatsApp tallies 700M monthly active users*. Retrieved March 31, 2015, from http://www.cnet.com/news/facebooks-whatsapp-messaging-service-tallies-700-million-monthly-active-users.

Shi, S., Largillier, T., & Vassileva, J. (2014). Rings: A Visualization Mechanism to Enhance the User Awareness on Social Networks. In J. Kawash (Ed.), *Online Social Media Analysis and Visualization,* 99–127.

Shneiderman, B. (1996). The eyes have it: A Task by data type taxonomy for information visualizations. *Proceedings, IEEE Symposium on Visual Languages,* 336–343. IEEE.

Simon, H. A. (1956). Rational Choice and the Structure of the Environment. *Psychological Review, 63,* 129–138.

Sinha, R., & Swearingen, K. (2002). The role of transparency in recommender systems. *CHI'02 extended abstracts on Human factors in computing systems,* 830–831. ACM.

Sriram, B. (2010). *Short text classification in Twitter to improve information filtering.* Master's Thesis. The Ohio State University.

Suh, B., Woodruff, A., Rosenholtz, R., & Glass, A. (2002). Popout prism: adding perceptual principles to overview+ detail document interfaces. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 251–258. ACM.

Swearingen, K., & Sinha, R. (2001). Beyond algorithms: An HCI perspective on recommender systems. *ACM SIGIR 2001 Workshop on Recommender Systems 13,* 1–11. ACM.

Thomson, L. C., & Wright, W. D. (1947). The colour sensitivity of the retina within the central fovea of man. *The Journal of Physiology, 105* (4), 316–331.

Thorson, E., Reeves, B., & Schleuder, J. (1985). Message complexity and attention to television. *Communication Research, 12* (4), 427–454.

Tintarev, N. (2007). Explanations of recommendations. *Proceedings of the 2007 ACM conference on Recommender systems,* 203–206. ACM.

Tintarev, N., & Masthoff, J. (2007). A survey of explanations in recommender systems. *Data Engineering Workshop, 2007 IEEE 23rd International Conference on Data Engineering (ICDE),* 801–810. IEEE.

Tintarev, N., & Masthoff, J. (2010). Designing and Evaluating Explanations for Recommender Systems. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender Systems Handbook,* 479–510. Springer US.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*. 12, 1 (Jan. 1980), 97–136.

Twitter. (2015, February 5). *Twitter Reports Fourth Quarter and Fiscal Year 2014 Results*. Retrieved March 24, 2015, from https://investor.twitterinc.com/releasedetail.cfm?ReleaseID=894844.

Twitter. (n.d.). *About Twitter, Inc.* Retrieved March 24, 2015, from https://about.twitter.com/company.

Tyler, S. K., & Zhang, Y. (2008). Open Domain Recommendation: Social Networks and Collaborative Filtering. *4th International Conference, ADMA 2008, Proceedings*, 330–341.

Upbin, B. (2012, April 9). *Facebook Buys Instagram for $1 Billion. Smart Arbitrage.* Retrieved March 31, 2015, from http://www.forbes.com/sites/bruceupbin/2012/04/09/facebook-buys-instagram-for-1-billion-wheres-the-revenue.

Waldner, W. J., & Vassileva, J. (2014). A Visualization Interface for Twitter Timeline Activity. *RecSys '14 Joint Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS'14),* 45–52.

Waldner, W. J., & Vassileva, J. (2014). Emphasize, don't filter!: displaying recommendations in Twitter timelines. *RecSys '14 Proceedings of the 8th ACM Conference on Recommender Systems,* 313–316. ACM.

Wang, Y. (2010). *SocConnect: a social networking aggregator and recommender.* Master's Thesis. University of Saskatchewan, Saskatoon, Canada.

Wang, Y., Zhang, J., & Vassileva, J. (2010). Towards Effective Recommendation of Social Data across Social Networking Sites. *14th International Conference, AIMSA 2010, Proceedings*, 61–70.

Webster, A., & Vassileva, J. (2006). Visualizing personal relations in online communities. *Adaptive Hypermedia and Adaptive Web-Based Systems: 4th International Conference, AH 2006, Proceedings*, 223–233.

Webster, A., & Vassileva, J. (2007). The KeepUP recommender system. *Proceedings of the 2007 ACM Conference on Recommender systems,* 173–176. ACM.

Yahoo Finance. (n.d.). *FB Key Statistics*. Retrieved March 24, 2015, from http://finance.yahoo.com/q/ks?s=FB+Key+Statistics.

Zuckerberg, M. (2006, September 5). *Calm down. Breathe. We hear you.* Retrieved March 31, 2015, from https://www.facebook.com/notes/facebook/calm-down-breathe-we-hear-you/2208197130.