On Measuring Social Dynamics of Online Social Media

Ian David Wood

Oct 2016

A thesis submitted for the degree of Doctor of Philosophy
of the Australian National University

Australian
National
University

To Otto, who made the future a reality.

Declaration

The work in this thesis is my own except where otherwise stated.

Ian Wood

Acknowledgements

This work would not have been possible without the help of many people. I cannot hope to include everyone here, and the journey has been long, with the early stages lost in the mists of time. Nonetheless, I would like to thank the following for their advice, support and encouragement: Henry Gardner, Rob Ackland, Dirk Van Rooy, Roddy Dewar, Scott Sanner, Richard Jones, Michael Dalvean, Liz Reiger, Otto Wood, Nausica Garcia Pinar, Sabrina Caldwell, Omid, Narjess Azfaly, Ardy Hadad, Brenda Martin, Mum, Dad, Jenny, Ian, Peter, Wendy and Misai.

Abstract

Due to the complex nature of human behaviour and to our inability to directly measure thoughts and feelings, social psychology has long struggled for empirical grounding for its theories and models. Traditional techniques involving groups of people in controlled environments are limited to small numbers and may not be a good analogue for real social interactions in natural settings due to their controlled and artificial nature. Their application as a foundation for simulation of social processes suffers similarly.

The proliferation of online social media offers new opportunities to observe social phenomena "in the wild" that have only just begun to be realised. To date, analysis of social media data has been largely focussed on specific, commercially relevant goals (such as sentiment analysis) that are of limited use to social psychology, and the dynamics critical to an understanding of social processes is rarely addressed or even present in collected data.

This thesis addresses such shortfalls by: (i) presenting a novel data collection strategy and system for rich dynamic data from communities operating on Twitter; (ii) a data set encompassing longitudinal dynamic information over two and a half years from the online pro-ana (pro-anorexia) movement; and (iii) two approaches to identifying active social psychological processes in collections of online text and network metadata: an approach linking traditional psychometric studies with topic models and an algorithm combining community detection in user networks with topic models of the social media text they generate, enabling identification of community specific topic usage.

Contents

Chapter 1

Introduction

This chapter presents an overview of the context of and motivation for this thesis and briefly summarises its core contributions.

The study of how societies operate is at once highly challenging and highly rewarding. Societies are a complex and poorly understood phenomenon. We make decisions and do things collectively. Although sociologists and social psychologists have been studying how collective decision making happens and how it evolves, social processes on large scales of populations and time are still poorly understood. Cultural evolution remains a mystery [Pennebaker and Lay 2002]. It is important to realise the sheer scope of this challenge — the complexity of human social systems, arguably, rivals any other system we study bar cosmology and life itself.

Nonetheless, the modern proliferation of data about the individuals in our societies, the evolving tools and techniques available to examine this data, and the rapid evolution of computer hardware enabling our ability to process it, all show enormous promise in advancing our knowledge. Though there are limitations to this endeavour and the quantity of useful information that can be extracted from large scale social noise is not clear [Ruths and Pfeffer 2014], the opportunities have only begun to be realised and it is apparent that much more can be achieved.

Current research attempting to unravel social phenomena using quantitative methods focusses on agent based computational models that are grounded in theories of social phenomena and their dynamics. Due to the difficulty in measuring actual social phenomena, grounding such models in actual groups and societies is done in a very limited way. The core focus of the research described in this thesis is the collection of richly dynamic data specific to particular social groups and the search for methods to identify and measure relevant social processes within that data.

This chapter is organised as follows: Section 1.1 presents an overview of the research contributions presented in this thesis. Section 1.2 reviews the current status of social simulation research, highlighting the need for stronger empirical grounding. Section 1.3 discusses the breadth and scope of socially generated data that are available today, noting that data used in research typically lacks the rich dynamics and community focus that would be necessary to empirically ground simulations of social processes and introducing the techniques of Chapter 3 as a remedy to the situation. Section 1.4 presents current directions in the analysis of social media data, highlighting the need for analysis focussed on social processes, as would be needed to empirically ground social simulation. Topic models are presented as an avenue to detect those processes and the techniques of Chapters 5 and 6 are introduced as initial steps toward their use in that way. Section 1.5 presents a vision and research agenda thus motivated and briefly reviews the contributions of this thesis in that context.

1.1 Overview of Research Contributions

This thesis presents 3 main research contributions. Firstly, in Chapters 3 an approach to identifying a specific community operating over social media is developed, and system for collecting richly dynamic data from such a community operating on Twitter is presented. This approach and system are used to collect data from a community of people experiencing eating disorders (the Twitter "pro-ana" community — see Section 3.2) over a period of nearly 3 years. Chapter 4 provides an overview and some initial analyses of the collected data.

The second contribution, developed in Chapter 5, is an approach to linking topic models of text data connected with a particular community to traditional psychological studies where standard Likert scale questionnaires are used in combination with carefully orchestrated free text responses from survey subjects. Using such an approach, results of such realtively small scale, detailed studies can be applied in much larger collections of community communications.

The third contribution, developed in Chapter 6, develops a novel approach to combine analysis of community structures in a network of social media users with topic models over their communications. A strong correspondence between a selection topics and communities is identified, proving characteristic communication patterns within those communities. As argued below and in Chapter 2, such patterns are good candidates for socially generated and propagated entities such

as social norms and identities (see Section 2.1 for a discussion of such entities).

These contributions are discussed in more detail in Chapter 7 along with a discussion of how these contributions can be applied in future sociological and social psychological research.

1.2 Grounding Simulations of Social Processes

Empirical grounding is a substantial challenge to research in social simulation. Psychologists have been studying people's behaviour for many years, both in groups and as individuals, but this has largely been done in a laboratory setting or through surveys and questionnaires. People's behaviour in real social contexts is thought to differ from how they behave in a laboratory and how they project themselves in questionnaire responses[1]. Sociologists and economists have also studied social behaviour for a long time, however their research has been either qualitative in nature or involving only collective statistics, without fine-grained measurements of individual behaviours and qualities. Until recently, measuring people's social behaviour "in situ" on a significant scale has been impossible. Such a task is the central challenge of modern social simulation.

Although the findings of social psychology, sociology, and economics provide internal structure for social simulation models, and although aggregate features and statistics can often be measured, providing some level of validation, direct grounding of modelled qualities of individuals and their communications is difficult and remains largely elusive [Windrum et al. 2007]. Data collected on individuals for grounding agent based models typically consist of inferred statistics from surveys of small sets of people randomly sampled from a larger population [Hassan et al. 2008], however such data lacks information on the fine dynamics of social interactions and detailed social structures. Social systems are inherently very complex, with a large number of diverse interacting processes. Without empirical data at least approaching the dimensionality and complexity of modelled systems, there is a great danger of overfitting (where the models fit the data, but do not generalise) and of missing important factors and processes.

Data collected from online social media promises new opportunities for empirically grounding agent based social models. Though some work in this direction exists (e.g.: [Garcia and Schweitzer 2011; Wise 2014]), to the best of my knowledge the direct measurement of social constructs that drive collective decision making

[1] Anecdotal evidence from conversations with social psychologists.

has not been investigated. Such identification is one of the core motivational goals of this thesis.

1.3 Social Data

The emergence and large scale uptake of online social media has opened enormous opportunities for the study of society and social processes. Combined with new machine learning and data mining techniques (in particular text analysis) the opportunities are great. This is particularly noticeable when one considers the apparent near ubiquitous uptake of social media by young people [Jones 2013; Go-Globe 2013] — there is good reason to believe that the rapid growth in social media usage will continue and that the already notable proportion of social interaction occurring on-line will only increase in years to come.

Analysis of social media data has received attention in the study of sociology and psychology. For the greater part, however, methods applied largely involved aggregate statistics and in-depth subjective analyses (see Section 2.2). Attempts to identify psychologically relevant patterns attributable to individuals require more sophisticated tools.

Recent research into data mining and machine learning has developed tools that hold some promise to achieve this (see Sections 2.3 and 2.4). A great deal of this research has, in fact, been directed at social media data, however it has been largely focussed on direct commercial applications such as sentiment and emotion analysis (Section 2.1.2) and has been largely naive in its modelling and investigation of psychological processes. Data sets for this research typically contain limited dynamic information (e.g.: data collected over a very short period of time) and are typically either generic and community wide or focussed by particular commercial aims, rather than focussed on particular social groups. Chapter 3 provides a remedy to this situation, presenting a methodology to target particular social groups, an approach and system to collect richly dynamic data from the targeted group and a substantial data set retrieved from the anorexia and eating disorder community operating on Twitter.

1.4 Mining Socio-Cultural Signals

Alongside the explosion of data available for study in recent years (in all walks of life, not only social media) has been research into data mining and machine

learning. A swathe of techniques have been developed to efficiently extract otherwise hidden patterns and information from large data sets. Text mining, the identification of patterns and extraction of information from large collections of text, is one area of application of these techniques that has been widely used.

There has been much research done on mining aspects of the 'meaning' of texts in the fields of computational linguistics and natural language processing. Bayesian graphical model approaches have been successful in this area [O'Connor et al. 2013]. What I attempt here has a rather different, though related, goal: I aim to detect actions and objects related to social processes. Many of these techniques draw on the distributional hypothesis of how we learn, perceive and communicate meaning [McDonald and Ramscar 2001], where meaning and experiential context are inextricably linked. By this hypothesis, the meaning of a portion of text can be identified, at least in part, by its lexical context within that text.

In this thesis I further draw on social representation theory from the field of social psychology [Billig 1991; Bauer and Gaskell 1999; Sammut et al. 2015] which posits the existence and importance of socially generated and propagated entities (representations) to the study of human social behaviour. Combining these ideas with the distributional hypothesis, one is drawn to the idea that social entities such as group norms and identities can also be identified by lexical contexts in text communications. Given that intra-group communication contains indications of shared socio-linguistic entities (such as norms, identities, conventions etc.) and given that the distributional hypothesis holds for cognition of those entities, one would expect those entities to be associated with particular lexical contexts, that is, loosely speaking, collections of words and phrases that co-occur.

Topic modelling (see Section 2.4) is a prominent approach for unravelling the semantics (meanings) expressed in collections of texts. In essence, topic models identify linguistic contexts whose semantics can then be interpreted by the words with substantial representation within those contexts. Topic models are a widely used and rapidly evolving set of data mining tools, finding many applications around document classification, search, and summarising key content in collections of text documents. The ability of topic models to uncover contextualised patterns in text collections presents itself as a real possibility for identifying and measuring social entities. This is especially true if a community uses social media as its primary communication medium as may be the case for the anorexia and eating disorder community operating on Twitter (see Section 2.2.4).

In Chapter 5 I develop an approach to link survey based psychology research

with topic models in order to infer psychological characteristics of the authors of a collection of texts. As an example, results of a recent study of identity salience in young women is combined with a topic model of Twitter eating disorder and pro-anorexia data, demonstrating the ability of topic models to identify linguistic contexts of psychological relevance.

In order to study the dynamics of social groups and the symbols and beliefs they entail, it is important to identify the members of those groups. Social media network data sets can be expected to represent many overlapping and interwoven social groups, and much work has been done to identify those groups from network characteristics. Chapter 6 presents an approach to combine detected groups in a network of social media contributors with topic models from the social media texts that those contributors create, attributing detected topics to groups. The presented approach works well with large data sets due to the substantial dimensionality reduction provided by the network detection algorithm and topic model. It is applied to a snapshot of the collected eating disorder and pro-anorexia Twitter data, revealing very distinct group topic usage.

1.5 Visions and Contributions

This research works from the larger vision of empirically grounded models of social processes capable of tracking and potentially predicting the evolution of social constructs that lead to collective decisions in our societies. Such models would enable research into social psychology and sociology in unprecedented ways and have the potential to provide deeper understandings of how our societies make decisions and function in general.

The sheer scope of this endeavour cannot be understated. Groups and societies are enormously complex entities and many of the relevant processes needed to understand them occur within peoples' minds, a place into which we have no window. This vision is grand, and it remains to be seen to what extent it can be achieved. Nonetheless, new data sources from online social media (among others) show some potential for a beginning of such grounding.

Empirical grounding has two fundamental requirements: the ability to gather data containing some imprint of the processes being modelled, and the ability to then identify and measure aspects of that data which constrain the modelled processes. The contributions of this thesis take small steps toward both of these requirements — the collection of data likely to contain traces of social processes

and the analysis of that data to identify and quantify socio-linguistic contexts as candidates for relevant social entities.

Chapter 2

Relevant Literature

In this chapter I review the literature on research efforts related to the contributions in this thesis. There are four areas that I review:

Section 2.1 reviews relevant research into human behaviour. This includes several aspects drawn from both the social psychology and sociology literature. Section 2.1.1 looks into the study of social groups and the role played by social entities such as norms and identities. Section 2.1.2 looks into the study of psychology as expressed in text — that is, the links between the psychology of an author and the text they produce.

Section 2.2 reviews various areas of relevant research using data collected from online social media. Section 2.2.1 outlines some of the difficulties and caveats when working with social media data. Section 2.2.2 looks at research into hash tag usage, noting in particular its role as a group meeting place. Section 2.2.3 mentions two studies of Twitter data of particular interest, one measuring stress levels of authors and the other using Twitter data to ground an agent based model. Section 2.2.4 briefly reviews work into online pro-anorexia and eating disorder communities.

Section 2.3 reviews relevant literature around the analysis of networks. Section 2.3.1 looks at research into community detection in networks. Section 2.3.3 briefly reviews work on the dynamics of social networks.

Section 2.4 presents a substantial review into Bayesian topic models. Section 2.4.1 introduces Latent Dirichlet Allocation (LDA), perhaps the most famous and first Bayesian topic model. Section 2.4.2 describes approaches and literature around topic model evaluation and diagnostics. Section 2.4.3 reviews data pre-processing approaches used in practical applications of topic models. Section 2.4.4 presents a non-exhaustive but lengthy summary of some variants

and extensions to LDA.

2.1 Human Behaviour

This section looks into human behaviour research relevant to this thesis. This includes group structure and function (Section 2.1.1), and research around methods for identifying and measuring psychological features in text (Section 2.1.2).

2.1.1 Social Groups and Entities

Humans are social animals. Forming, interacting with and acting in groups is a fundamental part of human behaviour. When we operate as part of a group, we naturally form and adhere to social norms (rules of conduct), identities (what it means to be part of a group) and other socially constructed and transmitted entities (such as values and beliefs).

Social Representation Theory: In the study of social psychology, the theory that pertains most closely to the existence and nature of such entities is Social Representation Theory [Billig 1991; Bauer and Gaskell 1999; Sammut et al. 2015]. This theory has not received a great deal of attention in social psychology research since its advent in the 1990s, but many have argued that it is a necessary part of a complete theory of social psychology, as it is capable of considering culture and ideology, which are essentially impossible to study with traditional laboratory studies [Nafstad and Blakar 2012].

In social representation theory, socially constructed and transmitted representations exist as shared conceptualisations. Sammut et al. [Sammut et al. 2015] present ...

> ... three defining characteristics of representations — the cultivation in communication systems; structured contents that serve various functions for the communication systems; and their embodiment in different modes and mediums. In social milieus, systems of communication (representations) evolve and circulate. This is referred to as the process of symbolic cultivation. Representations are embodied in one or more of four modes: habitual behaviour, individual cognition, informal and formal communication.

Such representations form a common context for social interactions and are often strongly connected to ones sense of self as well as defining and operationalising social groups. Examples include social norms, shared ideals, rituals (both formal and informal) that provide a sense of belonging to a social or cultural group and many more. The social entities referred to in this thesis equate to such representations. Those that can be measured in online social media pertain mostly to communication, though habitual behaviour may also leave a mark on a person's interaction with online social media.

Of further relevance here is the distributional hypothesis [Harris 1954; Firth 1957; McDonald and Ramscar 2001] which posits that meaning as perceived by humans is created and defined by cognitive context — we learn the meanings of things through observations/exposure to manifestations of those things in some cognitive context, and hence those meanings are inextricably connected to the contexts in which the concepts are encountered. This supports the idea that socially generated and propagated entities in communities that communicate via online social media can be recognisable as frequent lexical contexts, especially when interaction within a group occurs mostly online.[1]

It is interesting to note that in the case of highly abstract concepts, the manifestations may consist of the contexts alone without specific concrete markers (such as particular words or objects). This means that social entities that are highly abstract may still be identifiable and measurable among social media texts, even in the absence of specific words or phrases that define them.

Frames: A related concept to the social entities discussed above are "frames" [Fillmore 1976; Fillmore 1982]. Widely used in the field of sociology[2], a frame is a stance or perspective, usually with associated language constructs, that is used by and defines a social group. A frame fulfils a role both as a social identity (I am a group member and valuable because I believe in this frame) and as conforming behaviour (those who express our frame are with us). A frame can be seen as a particular type of social entity or representation, typically with a concrete manifestation and strong connection to a social identity.

One recent study of particular interest here posited that the "Topics" in topic models (see Section 2.4) could be used as proxies for "frames" [DiMaggio et al.

[1] Note that in natural language processing, the distributional hypothesis is typically applied to interpret the meaning of words from their lexical contexts — here we seek the meaning of contexts in a broader sense, with possibly multiple lexical expressions.

[2] As of October 2015, Google scholar counted 744 and 2023 citations of Fillmore's 1976 and 1982 articles respectively.

2013]. This is similar to the supposition that topic model topics can represent social entities and can act as a tool for their identification and measurement. Though this study implicitly associated frames with topic model topics, drawing on the conceptual similarity between them, the authors nonetheless draw interesting conclusions that agreed with perceived shifts in thought and culture surrounding the studies target themes (US government arts funding).

Dunbar's Number: Though not directly linked to the primary research agenda here, observations of collected data are suggestive that Dunbar's number may be playing a role, prompting a potential avenue for future research (see Section 4.1, second paragraph).

Based on studies of the size of the neocortex in primate brains, Robin Dunbar, in the early 1990's, recognised a link between the size of the neocortex in primate brains and the maximum size of social groups [Dunbar 1992]. Extrapolating this to humans, he suggested that for humans, the maximum size of social groups is about 150 (with 95% confidence interval from 100 to 230) [Dunbar 1993] and presented numerous examples of human social organisation that fall within these bounds. That study also hypothesised that language evolved primarily as an efficient method of social bonding that enabled larger group sizes than primates, for whom social grooming is the primary method. Many more recent studies have verified this range including groups in social media and specifically Twitter [Gonçalves et al. 2011; Dunbar et al. 2015].

Another line of enquiry into the size of social groups that pre-dated Dunbar's work looked at the number of acquaintances individuals have with people from a set of non-social minority groups (such as people whose name is Christopher or diabetes sufferers). They found a remarkably consistent pattern following a skewed distribution with a mean of about 291 (median 235) [Killworth et al. 1984; McCarty et al. 2001]. The mode of this distribution is close to 150. Adding further granularity to these numbers, several recent studies have found relationships between the size of parts of the brain and social network size [Dunbar 2012].

2.1.2 Language and Psychology

An increasing number of studies indicate that statistics of word usage provide a reliable measure of individual differences [Chung and Pennebaker 2013]. The words people use seem to not only provide information about themselves, but also about their context, and the audience that they are addressing. Psychometric

analysis can provide indirect measures of demographic variables (social status, age, sex), but also social motives (coercion, deception). It can can inform us whether an author is emotionally involved instead of detached, and even allows to infer a number of clinically important psychological traits (such as extraversion and neuroticism).

Analysis of text to find clues to the psychology of the person who wrote or said the words has a rich history. Many studies have been done that correlate the use of certain types of words or language forms with aspects of personality and a person's psychology. Several tools for automated analysis of texts have been created based on the results of these studies and our understanding of human psychology. As a rule, these tools rely on the frequency with which certain carefully selected classes of words are used in the text, though some structural statistics are also used. The most important tools are listed below:

LIWC — Linguistic Inquiry with Word Count Linguistic Inquiry with Word Count [Tausczik and Pennebaker 2010a] is one such tool that has been particularly widely used. Its approach is very simple: it calculates the frequencies of 77 classes of words and punctuation and 3 structural statistics in each text. These frequencies and statistics are known to correlate with certain aspects of a person's psychology and state of mind.

LIWC includes 32 word categories that indicate psychological processes (e.g.: affect, such as positive and negative emotions; social, such as family and friends; cognitive such as insight and causation), 22 that indicate linguistic processes (e.g.: adverbs; negations; swear words), 7 personal concern categories (e.g.: home; religion; work; leisure) and 3 paralinguistic dimensions (fillers; assents; nonfluencies) [Skowron et al. 2011].

PYM Norms — Paivio, Yuille and Madigan Another important tool is commonly referred to as the PYM norms, named after the original authors: Paivio, Yuille and Madigan [Paivio et al. 1968] and later extended by Clark and Paivio [Clark and Paivio 2004]. These norms grew from the theory of cognitive psychology. A large number of words (2311) are rated as indicators of various cognitive processes such as imagery, concreteness, meaningfulness, and familiarity.

POMS and GPOMS — Profile of Mood States POMS (Profile of Mood States) [McNair et al. 1981] is a tool developed for clinical psychology in which

patients rate a series of statements about mood on how well they describe their feelings. Many of these statements are single words. Bollen et al [Bollen et al. 2011] adapted and extended POMS-bi norms for large scale text analysis with the aid of Google N-grams data.

Sentiment and Emotion Analysis There has been a recent surge of interest in the machine learning research community to identify peoples' attitudes toward particular entities. Driven by a business need to measure consumer sentiment, a wide range of tools and approaches have been developed for measuring sentiment in (mostly online) text. These approaches typically use keywords to locate references to the item of interest (eg: "google" or "apple" or a book/movie title) then attempt to asses the sentiment of the surrounding text on a positive/negative scale. The sentiment scale could be discrete (eg: the values {-1,0,1}) or continuous (eg: a real number in the interval [-1, 1]). The sentiment scale is typically one dimensional, however there are situations when a person may experience both positive and negative sentiment at the same time, and some approaches use two separate scales, allowing for this possibility (e.g.: [Thelwall et al. 2012]).

Another related area that is gaining interest is affective computing — the detection of human affect (emotion) in text or other media such as voice or video. This has had particular success with voice and video media, however detection of emotion in text has received less attention, in part due to difficulties in providing a ground truth (as human emotion annotations often show poor inter-annotator correlation [Mohammad and Alm 2015]). Using user generated annotations in the form of hash tags on Twitter is a promising approach to improve this situation [Mohammad and Kiritchenko 2015]. A similar approach was effective at identifying author sentiment utilising hash tags and text smileys in twitter data [Davidov et al. 2010].

Though of some relevance, sentiment and emotion analysis measure psychological states, they do not have direct application to the contribution presented in this thesis, which is focussed on the detection of social psychological entities through topic models. In future work investigating candidate detected entities, emotion and sentiment analysis may however play a role.

2.2 Social Media Data

The growing ubiquity of the Internet and online social media (blogs, chat rooms, Twitter, Facebook etc..) provides an unprecedented opportunity to study people

in an unobtrusive way. The quantity of data available is somewhat daunting. For example, a blog and social media dataset from a one month period in 2011, used in the ICWSM 2011 Data Challenge, amounted to some 3 TB (3000 GB) of text [K. Burton et al. 2011]. Analysis of such large corpora of texts requires clever automated techniques and perhaps a dash of high performance computing.

2.2.1 Working with Data from Micro Blogs

Data from Twitter and other micro-blogs present specific challenges. Due to the restrictions on the length of tweets and the cumbersome nature of typing, language usage is often abbreviated [Eisenstein 2013], often breaking grammar rules and using community specific abbreviations, smileys, text art and other creative forms. The fact that texts have few words/tokens also poses difficulties, with one study suggesting that aggregating tweets can improve analyses [Hong and Davison 2010].

One important factor to keep in mind when data mining social media data is self selection bias: users of different social media platforms and different groups of people who use those platforms come from specific subgroups of the society at large, and comments made on social media platforms are filtered by a person's particular desire to communicate. It is important to consider these biases when attempting to draw conclusions beyond the literal scope of the set of users whose data is in your study and the specific social context that it represents [Ruths and Pfeffer 2014].

2.2.2 Hash Tag Usage

Hash tags are widely used in social media platforms as a means to add a "tag" to the text they are producing. Typically, a hash tag is simply a hash symbol "#" preceding a word, for example *#happy*. It can be said that the initial intention of hash tags was as a means of categorising and identifying types of content and its semantics. This is certainly true for social bookmarking sites such as Delicious[3]. Users of Twitter, Tumblr and similar micro-blogging sites have created other uses for hash tags. For example, tags such as *#fail* convey a judgement about something, *#meh* conveys a sense of boredom in response to something and *#happy* conveys an emotion. Their most typical role is still to convey thematic information about the content of the tweet, but emphasis, emotion and many

[3]https://delicious.com/

other expressions are commonplace, limited only by the creativity of the Twitter community.

Of particular interest here is the use of hash tags as a virtual meeting place and to address and define a particular community [Java et al. 2007; Huberman et al. 2008; Paul and Dredze 2011; Bruns and Burgess 2011; Himelboim 2014]. This, and the observation that people tweeting on "#pro-ana" and related tags appear to form such a community, are key to the data collection methodologies presented in Chapter 3.

The advent of group chats on Twitter has been noted in recent years [Cook et al. 2013; Budak and Agrawal 2013]. Group chats are groups of Twitter users that meet at specific times and hash tags on a regular basis. Many such chats are deliberately organised by institutions such as in education [Budak and Agrawal 2013] and depression/mood disorder support groups, but there is also a growing number of relatively spontaneous chats organised primarily by a passionate interest such as movies or skiing [Cook et al. 2013]. This is a refined version of the use of hash tags to define a community, and coordinated meetings between individuals and sub-groups will likely spontaneously occur within communities organised around hash tags.

There has been much research applying topic models to Twitter data. The extra meaning attached to hash tags has been used in some studies to improve topic models of tweet collections in several ways. Perhaps the simplest is to pool tweets by hash tag, which has been found to perform well for document clustering tasks and to improve topic coherence [Hong and Davison 2010; Mehrotra et al. 2013]. In other approaches, hash tags are used as labels in a supervised topic model [Ramage et al. 2010] and incorporated into the generative model [Lim et al. 2013]. Both approaches showed improvements in an author recommendation task. Where Ramage et al. focussed on characterisation of tweets according to substance, status, style and social dimensions, Lim et al. showed how their model could be used for automatic topic labelling and demonstrated improvements in clustering tasks and topic coherence.

2.2.3 Other Relevant Twitter Studies

One recent study has measured psychological stress in multi-modal social media data (text plus images and interaction metadata) using a deep, sparse, neural network ("deep learning") [Lin et al. 2014]. Hash tags of words in LIWC categories (Linguistic Inquiry with Word Count — see Section 2.1.2) related to stress

were used as a ground truth. The accuracy of these hash tags for identifying stress was verified by human annotation. Their system achieved good predictions of physiological and "other" stress (F1-measures[4] of 0.71 and 0.81 respectively) but performed poorly on affection, work and social stresses (F1-measures of 0.34, 0.54 and 0.50). Unstressed posts were detected very well (F1-measure of 0.999). Lin et al. tested support vector machines and three neural network models, with all performing similarly. An important take-home message from this work is the value of multi-modal analysis, and in particular including features extracted from images, for detecting psychological states.

2.2.4 Online Pro-Ana Commuities

In this section I briefly outline research and media commentary on the online pro-anorexia movement and explain reasons why this community is of particular interest for studying social processes in an online setting.

The online "pro-ana" (pro-anorexia) movement has received much attention in the study of clinical psychology (e.g.: [Homewood and Melkonian 2015; Sheldon et al. 2015; Yeshua-Katz and Martins 2013; Casilli et al. 2012]) and in popular media (e.g.: [McColl 2013; Times ; Tribune ; Reaves 2001]). The movement has received criticism on the grounds that it encourages eating disorders [Rouleau and von Ranson 2011; Jett et al. 2010; Cohen 2007] and acclaim on the grounds that it provides social support for an otherwise stigmatised condition [Yeshua-Katz and Martins 2013; Dias 2013; Csipke and Horne 2007]. Media depiction of unreasonably thin women and general support of the "thin ideal" have been cited as a potential factor leading to the increase in eating disorder diagnoses [Derenne and Beresin 2006].

As can be seen, there is an ongoing debate about the role of the online (and in particular social media) pro-anorexia movement. Though beyond the scope of this thesis, the collected data described in Chapters 3 and 4 as well as the methodologies developed in later chapters provide opportunities for answering some of the questions in that debate.

Of further relevance to this work is the social stigma attached to eating disorders such as anorexia in the wider society [Yeshua-Katz and Martins 2013]. Such stigma suggests that people with these conditions will be discouraged from

[4]The F1-measure of a predictor of a particular class or labelling is the harmonic mean of its precision (proportion of predictions that are correct) and recall (proportion of actual cases that are predicted). It provides an overall sense of the predictors quality.

communication about their condition within their immediate, physical, community. An anonymous venue such as online social media provides an environment relatively safe from that stigma, thus it may be expected that social media communities formed around eating disorders such as anorexia have very little offline communication. This means that a full picture of the group-wide communications of such communities can be obtained.

A couple of other studies are worth mentioning here: A grounded analysis of themes of communication in social media eating disorder communities has been undertaken [Juarascio et al. 2010]; A comparison of language usage between pro-anorexics and recovering anorexics was performed, identifying differences in self-presentation [Lyons et al. 2006]; Analysis of pro-anorexia social media usage have been proposed for building qualitatively-informed agent-based models [Tubaro and Casilli 2010].

2.3 Network Analysis

In many areas of science, the study of network structures and properties can yield useful and profound results, and this is particularly true of the study of social networks, where a substantial literature has been developed [Scott 2012]. In the course of studying complex systems, networks are typically an effective way of representing relational data. Examples can be found among social systems (friend and acquaintance networks), ecological systems (food webs), biochemical systems, technological systems (e.g.: power networks, the internet), logistic systems, communication systems and the list goes on. Networks of interest are often very large, with thousands or millions of nodes and millions or billions of edges (links between nodes). Typically there is extra meta-data associated with network nodes or edges [Fortunato et al. 2013]. Extracting useful information and insights from such data is an important and non-trivial task.

2.3.1 Identifying Communities

Identifying communities or clusters in networks — groups of nodes that are well connected internally and less connected externally — is an important and active research topic. Community structures can provide direct insights into hidden relationships (e.g.: functional groups in protein networks [Ou-Yang et al. 2014]) and give a higher level view of large and complex networks that may be more amenable to human interpretation [Fortunato 2012].

Much work on community detection in networks has used a model where communities are disjoint, so that no community overlap is possible [Fortunato 2010]. Here the problem is posed as finding a partition of the network, where each partition is considered a community. A range approaches to this formulation exist with a corresponding rich literature [Tang and Liu 2010]. Let us consider an undirected network (or equivalently "graph") consisting of a collection of "nodes"[5] and the "edges" that connect them[6]. For example, individual Twitter users could be the nodes, and mutual Twitter follower relations the edges.

One approach to non-overlapping community detection of note, and which has been widely used in the research literature, is often referred to as "modularity clustering". This approach considers the difference between the number of edges connecting pairs of network nodes and the expected number of edges given the degree distribution. This quantity, summed over node pairs within a group of nodes is known as the groups "modularity". If m is the total number of edges, d_i is the degree of node i (the number of edges connected to node i), A_{ij} is the network's adjacency matrix ($A_{ij} = 1$ if there is an edge between nodes i and j, 0 otherwise) and C a collection of nodes, the modularity Q_C of C can be written:

$$Q_c = \frac{1}{2m} \sum_{i,j \in C; i \neq j} (A_{ij} - d_i d_j / 2m) \tag{2.1}$$

Modularity can be either positive or negative, with positive values indicating the possible presence of community structure. Modularity clustering seeks to partition the network such that the sum of the partition modularities is maximised [Newman 2006].

A second approach of note draws on an information theoretic view of the problem, seeking to optimise the information about the original network contained in the cluster representation [Rosvall and Bergstrom 2007]. This approach was found to perform as well as modularity clustering on evenly distributed clusters, but outperformed modularity clustering (and other approaches) on data where the community structure in the network was uneven, with some small and some large communities.

In many cases, and, in particular, with social networks, communities can be expected to overlap, and recently a substantial body of work has been devoted

[5] The term "vertex" is often used also.

[6] For simplicity, we consider only undirected unweighed edges and only one possible edge between each pair of nodes, however methods also exist for more general networks.

to overlapping community detection [Xie et al. 2013]. Of particular note is the mixed membership stochastic block model [Airoldi et al. 2009], a Bayesian model for overlapping community detection, and the efficient inference algorithm for this model developed in [Gopalan and Blei 2013]. This model and inference technique can be said to be state of the art, showing notable improvements on various standard data sets and being applicable to massive networks in reasonable time.

2.3.2 Twitter Networks

Network metrics of Twitter data have revealed varying and quite different characteristics. For example, a study of early Twitter data (2009) from Singapore found 72% reciprocity (friend links that "follow back") [Weng et al. 2010] whereas another 2009 study that obtained a nearly complete snapshot of the Twitter network found on average only 10% reciprocity [Cha et al. 2010].

Kwak et al. [Kwak et al. 2010] attempted to crawl the entire Twitter follower graph. They collected user profiles by snowball sampling [Biernacki and Waldorf 1981] from famous blogger Perez Hilton[7] also adding users who mentioned trending topics[8] during the period of data collection (approximately one month), claiming to have crawled "the entire Twitter-sphere" — not an entirely unreasonable assertion, since any Twitter account they missed can be said to not have contributed to public discourse through either mentioning popular topics or connections in the follower network. They present a number of general statistics on the friend/follower network and other user, tweet and network characteristics. Of particular note are their findings that a user's retweet count does not correlate with the number of followers nor PageRank [Page et al. 1999] in the follower network; there is a non-power-law follower distribution, short effective diameter, and low reciprocity in the follower network. All of these are unexpected in the light of known characteristics of human social networks [Newman and Park 2003].

Huberman et al. [Huberman et al. 2008] investigated the relation between the Twitter follower network and the mention network. When one posts a tweet, it is possible to include another user's user name preceded by an "**@**" symbol, in which case the other user is alerted to the tweet. This is known as "mentioning" and was taken as a closer indication of true social networks than the follower network. They found that real social networks (approximated by the mention network) are

[7] Perez Hilton had over a million followers at the time of the study.

[8] Trending topics are hash tags, keywords and phrases identified by Twitter as "trending", meaning that they are currently present in many tweets

subgraphs of the follower network. In another example, [Jurgens 2013] found that the location of nearby users in the mention network to be useful in inferring the location of users for which no direct location information is available.

Retweets, when a twitter user reproduces a tweet verbatim (possibly with some small addition of their own) provide another opportunity to extract a user network from a collection of tweets. A retweet network is typically thought of as a network of information flows. A number of studies have utilised retweet networks. For example in [Conover et al. 2011], community structures in a retweet network were found to be good predictors of political alignment.

Several other Twitter studies described in the following section (Section 2.3.3) looked at aspects of network dynamics.

2.3.3 Dynamics of Social Networks

Social network dynamics has been established as playing an important role in coordinated action. In a controlled study by Rand et al. [Rand et al. 2011], people preferentially added social ties to cooperative people, and broke them with uncooperative people. A substantial body of work exists on link prediction in networks (social or otherwise) [Lü and Zhou 2011; Wang et al. 2014]. Almost all of this work focusses on the prediction of new future links or missing links, with no attention given to predicting the dissolution of links or presence of spurious links [Wang et al. 2014], though two recent studies defy this pattern.

Kwak et.al. [Kwak et al. 2011] studied "unfollow" behaviour of 1.2 million Korean speaking Twitter users. They took daily snapshots of the user's friend lists and attempted to identify factors leading to unfollow events — when a user has been following another users tweets, but decides to no longer follow them. Factors that were found to be significant were reciprocity in follow relationships (users unfollowed users who did not follow them), the duration of the relationship (users unfriended users they had not been following for long), the followee's informativeness (if a user has retweeted or favorited followee's tweets) and overlap of the user and followee's friend and follow lists. They also conducted a survey of 22 Twitter users to determine their motivations for unfollowing, identifying frequent tweeters, tweeters of uninteresting topics and tweeters of mundane details of their lives to be significant motivators.

Another study that looked at unfollow behaviour [Xu et al. 2013] also looked at the Korean Twitter network. They took four snapshots of nearly 700,000 Korean Twitter users friend and follower lists. Focussing on ordinary users in tightly

knitted groups they found that relational properties such as mutual following and followers in common reduce the likelihood of unfollowing. They found that unfollowing tends to be reciprocal — if someone unfollows you, you are more likely to unfollow them in return. They found no evidence that common interests and informativeness of interactions impacted unfollow behaviour. Their work suggests that there may be many diverse types of Twitter user groups where the impact of relational and informational factors may differ.

2.4 Topic Models

Much of the content in this Section appears in my publication [Wood 2013], though greatly expanded here. Some concepts here are used later, but others go beyond the scope of the thesis but are considered here for completeness and as a context for future work.

In the context of the ideas presented in this thesis, topic models serve as a candidate for the detection and quantification of social constructs relevant to the dynamics of social processes (see Chapter 5).

Probabilistic topic models originated from the search for automated techniques to identify the semantic content of texts without any pre-conceived notion of what the semantic content may be. The intuition is that texts covering similar themes will tend to use similar vocabulary — one expects, therefore, that counts of individual words in such texts will be strongly correlated, and it is these correlations that are of interest. Another perspective on this task is that of finding a low dimensional representation (topics) of high dimensional data (very many different words) that maintains the characteristics of interest (the semantics).

Many methods proposed to date for finding text semantics use a "bag of words" model to represent texts: a list of word counts. A text is typically represented as a vector of word counts. The resultant vectors have very high dimension (the size of the vocabulary of the corpus under consideration). Simply calculating the correlation matrix for such data is infeasible as the computational and memory overheads would be enormous and there would be many more parameters than data, rendering the result meaningless. Topic models were developed to overcome these limitations and provide meaningful and tractable summaries of text corpora.

The term "topic model" is often used to describe models where a "topic" is represented as a list of word proportions (a number for each word in the corpus

vocabulary) which are non-negative and add up to one. In probabilistic models, such topics are taken to represent multinomial distributions over words. Such models typically provide a list of topic proportions for each document in the corpus, taken as a multinomial over topics in probabilistic settings. The idea is that a combination of a documents topic proportions and the word proportions of those topics should approximate the actual word counts in the document. The exact process by which that combination is made depends on the model in question.

Latent Semantic Indexing (LSI) [Deerwester et al. 1990], also known as Latent Semantic Analysis (LSA), was an early attempt at a topic model. It used singular value decomposition (SVD) on the matrix of document word count vectors. In essence, this decomposition reveals a number of topics (lists of word proportions — these are the left singular values) equal to the rank of the document-word matrix and associates a "strength" with each topic (the singular value itself). Typically only the strongest 300 or so topics are used, depending on the corpus [Bradford 2008]. In the literature, these topics are often referred to as "latent semantic spaces".

Probabilistic Latent Semantic Indexing (PLSI) [Hofmann 1999] attempted to find a probabilistic analogue of such spaces (i.e.: multinomials) via maximum likelihood. Here each document was modelled as a mixture of multinomials, each multinomial approximating a semantic context or "topic". A generative model essentially the same as Latend Dirichlet Allocation (see Figure 2.1) without the Bayesian priors (α and β) was used, and a maximum likelihood estimate of latent variables θ (topic distributions for each document) and ϕ (word distributions for each topic) was typically inferred with an expectation-maximisation (EM) algorithm. Though PLSI performed reasonably well, and better than previous methods, it suffered from overfitting and did not generalise well [Blei et al. 2003]. Latent Dirichlet Allocation (below) was better able to predict previously unseen data and was also found to better match human assessments on word association tasks [Griffiths and Steyvers 2003].

2.4.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [Pritchard et al. 2000; Blei et al. 2003][9] uses a similar model to PLSI, but applies Bayesian inference. Dirichlet priors (conjugate to multinomials) are used, which greatly simplifies the structure of the posterior and, thus, inference procedures. Gibbs sampling [Griffiths and Steyvers 2004] or variational methods [Blei et al. 2003] are typically used to estimate the posterior. More recently, an approach using *statistical recovery* has been developed that is orders of magnitude faster, making a relatively simple separability assumption [Arora et al. 2012]. These Bayesian methods produce models that generalise far better than previous maximum-likelihood approaches.

One way to envisage LDA is to imagine solving a puzzle. You start with many jars (documents) containing coloured marbles (words) and a collection of bags (topics). You need to distribute the marbles into the bags, trying to ensure that each bag contains mostly marbles of only a few colours. There is another restriction, however: you need also to make sure that each jar has most of its marbles in only a few bags. Usually this problem has no good solution — if you satisfy one requirement, the other doesn't do very well. If, however, some of the documents (jars) cover the same semantic topic (which we represent by a small set of colours/word types), you can put all the words (marbles) characteristic of that literal topic into the one bag. Those documents (jars) then look "pretty good". Similarly, if several semantic topics are referred to in a document, you can put most of that document's words into bags representing those topics and do "pretty well".

There is a tension here between the two requirements — a reluctance to have marbles of more than a few colours in a bag and a desire to put most of a document's marbles into only a few bags. The balance between these is governed by two parameters of the LDA model, typically labelled α (less topics for a document) and β (less words for a topic). The third parameter for LDA is the number of topics.

LDA uses Bayesian inference — that is, it proposes a parametrised generative model for the texts and then seeks the most probable parameter set given the text under investigation[10]. In order to make this inference, a 'prior' distribution over

[9] The same model was invented independently in the fields of population genetics [Pritchard et al. 2000] and text analysis [Blei et al. 2003]. Both papers have been highly influential, with 12369 and 9056 citations respectively (Google Scholar Aug. 2014)

[10] True Bayesian inference seeks the full probability distribution over model parameters, however this is often impractical, and the single most probable set of parameters is estimated

possible parameter values is provided. These inferred parameters are typically referred to as the model's *latent variables* or *hidden variables*, whereas the term *parameters* typically refers to some parametrisation of the prior distribution.

Given data D, latent variables Θ and prior $P(\Theta)$, by Bayes rule we have:

$$P(\Theta|D) = \frac{P(D|\Theta)P(\Theta)}{P(D)} \tag{2.2}$$

We seek values for Θ that maximise $P(\Theta|D)$. Since $P(D)$ does not vary with Θ, $P(D)$ can be seen as a normalisation constant and does not need to be calculated.

The generative model proposed by LDA consists of a fixed number, T, of topics, each represented as a multinomial distribution over words, and a multinomial distribution over topics assigned to each document. Implicit here is a corpus structure that has a fixed number, D, of documents and a fixed length, N_d, for each document d. Each word-position in each document is then filled by first choosing a topic from the containing document's topic distribution, then choosing a word from that topic's word distribution.

The prior used for for the word-topic and topic-document multinomials is the Dirichlet distribution. This is a natural choice as it is conjugate to the multinomial distribution — the posterior distributions are also Dirichlet, greatly simplifying posterior estimation. With appropriate parameter settings, the Dirichlet priors also encourage sparsity — probability is concentrated around multinomials with most of their entries near zero. This is almost always desired for the topic-word distributions (topics with few words), but for the document-topic distributions it is sometimes natural to encourage topic diversity (where each document contains a broad mixture of topics).

This process is often represented with a plate diagram such as Figure 2.1. In the diagram, boxes represents collections of documents — K topics, M documents and N words (ideally, N would be subscripted as it varies between documents, but this is usually omitted). Circles represent individual entities: α and β are parameters for the Dirichlet priors, θ the topic mixture for a document, ϕ a topic, Z

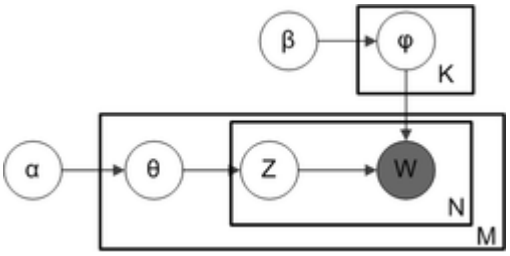

Figure 2.1: LDA Plate Diagram

instead — this approach is termed "Maximum a posteriori estimation" or "MAP".

a topic chosen from θ and W a word chosen from Z. W is grey, indicating that it is an observed variable (the only one in this model). Similar plate diagrams are often used to describe the model intra-dependencies of LDA variants and adaptations.

Estimation

As with most Bayesian models of modest complexity, direct calculation of maximum a posteori (MAP) values for θ and ϕ is not feasible. Instead one must employ a method that estimates these values. As noted earlier in this section, most of the literature employs either Gibbs sampling [Griffiths and Steyvers 2004] or variational methods [Blei et al. 2003]. Below I describe in somewhat more detail the Gibbs sampling approach.

Gibbs sampling is a method for estimating multivariate probability distributions that originated in statistical physics [Geman and Geman 1984]. It is a Markov chain Monte-Carlo (MCMC) method, meaning it produces a series of values that constitute a Markov chain whose stationary distribution is the probability distribution being sought. After a "burn-in" period during which the Markov chain settles down to a close approximation to its stationary distribution, the values can be taken as good estimates for samples from the target distribution. Gibbs sampling achieves this by sampling from each of the model's latent variables in turn, keeping the others fixed. A collapsed Gibbs sampler, as used in LDA estimation, first integrates out some of the models variables. For LDA estimation, θ and ϕ are integrated out, thus we need only sample word topic allocations z.

The full derivation of the Gibbs sampling update equations for LDA utilises relatively standard mathematical techniques which I do not present here. The resulting update equations are as follows. Write j for a particular topic index, vocabulary size W, T topics and n_j^d, n_j^w the counts of words assigned to topic j in document d or word w respectively. The $-i, j$ subscripts indicate that word i of the corpus is left out of the count. A '$\cdot$' in place of a super- or sub-script indicate a sum over that index (from [Griffiths and Steyvers 2004]).

$$P(z_i = j|\boldsymbol{z}_{-i}, \boldsymbol{w}) \propto \frac{n_{-i,j}^{wi} + \beta}{n_{-i,j}^{\cdot} + W\beta} \frac{n_{-i,j}^{di} + \alpha}{n_{-i,\cdot}^{di} + T\alpha} \tag{2.3}$$

In practice, i is fixed each time we apply equation 2.3 to sample z_i, thus the

second numerator is fixed and we can simplify thus:

$$P(z_i = j|\boldsymbol{z}_{-i}, \boldsymbol{w}) \propto \frac{n^{wi}_{-i,j} + \beta}{n^{\cdot}_{-i,j} + W\beta} \left(n^{di}_{-i,j} + \alpha\right) \tag{2.4}$$

Once we are satisfied that the model has converged to a reasonable level, we can calculate posterior values for document topic distributions θ and topic word distributions ϕ as follows.

$$\phi^{(w)}_j \approx \frac{n^w_j + \beta}{n^{\cdot}_j + W\beta} \tag{2.5}$$

$$\theta^{(d)}_j \approx \frac{n^d_j + \alpha}{n^d_{\cdot} + T\alpha} \tag{2.6}$$

2.4.2 Topic Model Evaluation and Diagnostics

Though topic models are powerful devices for identifying thematic patterns in collections of text documents (e.g.: [Griffiths et al. 2004; Blei 2012]), the models generated are not guaranteed to accurately represent real and meaningful patterns in the data. Even with random data, a model will be produced. Further, it is typically the case that a model will contain some "good" topics and some "poor quality" topics [Chuang et al. 2013; Mimno et al. 2011] (see *Human Assessments* below for interpretations of topic quality). Another study [Poldrack et al. 2012] observed that models with differing numbers of topics resolved concepts in the corpus with differing granularity. Thus there may not be a single "best" model, and which to choose may depend on the particular aims of the study at hand.

A more appropriate approach to assessing topic models is to utilise assessments that are tailored to the specific application under consideration. For some applications, such as document clustering and information retrieval, one can use assessment methods that do not directly relate to the semantic quality of the model (e.g.: [Blei and Jordan 2003; Wei and Croft 2006; Hörster et al. 2007; Perina et al. 2010; Jagarlamudi et al. 2012; O'Connor et al. 2013]). Typically such approaches utilise extra prior knowledge and/or human assessments, often including standard annotated data sets (e.g.: [Griffiths and Steyvers 2003]). In general, however, assessment of the semantic quality of topic models has largely

been limited to qualitative human judgements [AlSumait et al. 2009]. Recent work has identified several automated metrics that correlate well with such assessments as well as approaches for "posterior predictive checking" of topic models (see *Automated Model Diagnostics* below).

Topic Model Hyper-parameters

LDA, in its original form [Pritchard et al. 2000; Blei et al. 2003], requires the analyst to choose values for hyperparameters (the number of topics and Dirichlet sparsity parameters for word-topic and topic-document priors). With experience, reasonable choices can be made, although automated techniques have been developed that perform well and are generally seen as preferable. Two important approaches have been developed: (i) estimation of optimal Dirichlet parameters and (ii) so called 'non-parametric models' that essentially infer the number of topics (with a hyper-parameter determining how this is done; See Section 2.4.4).

Wallach et al. [Wallach et al. 2009; Wallach 2008] investigated inference of Dirichlet hyperparameters from data. They built models that inferred values for the Dirichlet priors of topic-document distributions α and word-topic distributions β, also experimenting with the standard symmetric priors and asymmetric versions of each. They found that the best configuration, in terms of data likelihood in the posterior, was to estimate an asymmetric topic-document prior α and a symmetric β. Allowing β to be asymmetric did not significantly improve likelihood. They argued that an asymmetric α allows some topics to be distributed fairly uniformly among documents, capturing very frequent words, whereas a symmetric β ensures that topics are distinct. Others have also noted that their hyperparameter optimisation performs well in terms of semantic coherence [Chuang et al. 2013] and in comparison to a grid search for optimal parameters [Asuncion et al. 2009]. It is interesting to note that one study found that optimising Dirichlet hyperparameters did not improve model quality, however they manually removed general and uninformative words from the data prior to estimating their final models, which may have obviated the need [Talley et al. 2011].

Held-Out Likelihood

In the machine-learning community, a typical approach for assessing the quality of a data model, is to test the likelihood of a held-out-subset of the data. A convention in language modelling is to use the *perplexity* of the held-out data as

a proxy for likelihood — it is monotonically decreasing in the likelihood of the held-out data [Blei et al. 2003].

A large proportion of papers relating to LDA use perplexity on held-out data as the measure of model quality. It has, however, been observed that perplexity does not necessarily capture human assessment of topic quality [Chang et al. 2009] and that models with more topics than indicated by optimal perplexity can better identify more granular concepts [Talley et al. 2011]. It is difficult to argue that the model with the best perplexity necessarily captures the types of structures we are interested in. Indeed, much of the effort in applied topic modelling attempts to recognise and remove modelled structures that are judged as noise or 'junk' (e.g.: [Chuang et al. 2013; Talley et al. 2011]). A model that captures such 'junk' structures well, but other, more interesting, structures poorly, may exhibit a good perplexity score, but is clearly suboptimal in real terms.

Perplexity is the inverse of the geometric mean of per-word likelihoods. Given held-out documents D and denoting the total number of words in the held-out documents as $|D|$:

$$\text{Perplexity}(D) := -\exp\left(\frac{\sum_{w\in D}\log p(w|\text{model})}{|D|}\right)$$

Note that calculating $p(w|\text{model})$ exactly is usually intractable as it is necessary to integrate over all possible topic-word assignments, however there are a number of ways to estimate it: for example see [Wallach et al. 2009].

Human Assessments

Application-focussed human assessments are always possible, however they require many man-hours (often from experts) and care must be taken to control for individual and systematic human biases. Several such studies have, nonetheless, been made, with some interesting observations on the ability of Bayesian topic models to find semantically-meaningful topics in text.

A powerful approach is to construct a multinomial model of topics based on human-identified semantics in a body of text. Chuang et al. [Chuang et al. 2013] recruited the help of domain experts to construct a comprehensive and exhaustive[11] model of the topics in a corpus of research literature. This model was then compared with LDA topic models generated with a comprehensive sweep of parameter settings (number of topics and concentration parameters for word-topic

[11] Respondents were requested to provide an exhaustive categorisation.

and topic-document priors). They identified several types of low-quality topics (fused, junk or missing) as well as good quality or 'resolved' topics, and observed all types in all models with proportions varying with parameter settings. A number of other studies have also observed that that LDA topic models typically contain a number of topics of poor quality [Hall et al. 2008; Talley et al. 2011; Poldrack et al. 2012]. Another common observation is that increasing the number of topics can resolve topics of increasing granularity [Poldrack et al. 2012]

Many studies aimed at producing high-quality models of specific document corpora have used post-hoc assessments of topic relevance to judge model quality (e.g.: [Hall et al. 2008; Talley et al. 2011; Wahabzada et al. 2012]). Dirichlet hyperparameters are typically either optimised during estimation (using methods such as in [Wallach et al. 2009]) or chosen by the authors. Similarly, the number of topics is often an ad-hoc choice by the authors that is not clearly explained in the publication (e.g.: [Hall et al. 2008]).

One study of successful NIH[12] grant applications [Talley et al. 2011] is of particular interest. This study combined diagnostic heuristics (see Subsection 2.4.2 below) with expert assessment by the authors. Preliminary models were found to contain topics that had "relatively uniform and distinctively low document allocations" (a diagnostic of semantically poor topics), and these topics were consistently found to contain general, non-research terms. Using these topics as an aid, the authors manually constructed a list of ~1200 uninteresting words, which were removed from the data for subsequent models. They also manually constructed lists of acronyms and commonly used bigrams and phrases. The result of both activities was to greatly improve assessed model quality.

Many studies have used topic models to tackle established estimation problems for which standard annotated data sets have been constructed (e.g.: [Griffiths and Steyvers 2003; Griffiths et al. 2004; Fei-Fei and Perona 2005; Boyd-Graber et al. 2007]).

It is worth noting that the paper that first introduced Gibbs sampling for LDA [Griffiths and Steyvers 2004] used subject category meta-data to qualitatively verify a topic model of PNAS[13] abstracts. The authors used a Bayesian approach to select the number of topics, then selected topics that were 'diagnostic' of each subject category (those with the highest average topic proportions) and made qualitative observations about topic semantics and category relationships.

[12]The American National Institutes of Health
[13]Proceedings of the National Academy of Sciences of the USA

Automated Model Diagnostics

Given the expense and potential impracticability of human assessments of topic models (e.g.: models with thousands of topics), there have been several attempts to construct automated methods that can separate topics judged by humans to be uninformative or incoherent from those judged as meaningful semantic units.

Types of Low Quality Topics Chuang et al. [Chuang et al. 2013] identified "junk" or "fused" topics in their analysis ("fused" meaning topics representative of two or more expert-identified topics). Mimno et al. [Mimno et al. 2011] asked domain experts to characterise ways in which topics could fail to represent semantically coherent units, resulting in a more detailed assessment, identifying several ways a topic can be "fused" as well as topics that are almost coherent. They summarised responses thus:

- *Chained*: words used in different contexts pull the contexts into one topic (these are "fused" topics).
- *Intruded*: either merged "chimera" topics ("fused") or a few words intruding into an otherwise coherent topic.
- *Random*: few sensible connections between words ("junk").
- *Unbalanced*: words are connected, but combine very general and specific terms ("fused").

Mimno later summarised diagnostic tools he had developed based both on this study and other practical experiences in topic modelling [Mimno 2012a]. These tools are implemented in the popular Mallet topic modelling toolkit [McCallum 2002].

- *Topic size*: Topics accounting for very few words in the corpus are more likely to be random, capturing words not accounted for by other more coherent topics.
- *Within-doc rank*: Ideally, a "good" topic should have significant presence in relevant documents, and little or no presence in others. In [Talley et al. 2011] topics with uniformly low rank were used to identify uninformative words, which were later removed from the analysis.

- *Similarity to corpus frequency*: A topic whose word probabilities match corpus frequencies is not very informative and tends to represent the remainder of the corpus after useful topics have been identified.

- *Locally frequent words*: Words appearing with high frequency in a small selection of documents, for example character names particular to a single chapter of a novel, often tend to dominate a topic. Such topics provide little extra semantic information, but can be recognised by their occurrence in very few documents but with high document proportions.

- *Co-doc coherence*: A measure of semantic coherence, and the main output of [Mimno et al. 2011]. Explained in more depth in the following paragraphs.

Topic Coherence Measures A number of studies have been made attempting to automatically assess topic semantic coherence with the aim of predicting human judgements. Two strategies for measuring human judgements have been proposed, both of which use short lists of the most probable words for each topic. The first asks assessors to rank a topic on a Likert scale (typically 1–3) [Newman et al. 2010]. The other draws on the observation that topics often contain significant "intruder" words that do not fit the semantics of the other words in the topic, posing a task where assessors attempt to detect a word chosen to not fit the semantics of the topic which has been deliberately added to the topic representation [Chang et al. 2009].

An important distinction between the coherence measures that have been introduced lies in the use of a large external reference corpus. Word relationships determined from the external corpus can be used to assess the thematic quality of topics. Newman et al. use pointwise mutual information (PMI) in the reference corpus between a topic's top words [Newman et al. 2010]. An approach creating a "semantic vector space" from reference data and using similarity metrics in those spaces performed marginally better than previous coherence measures [Aletras and Stevenson 2013].

Mimno et al. proposed a coherence measure based on the log conditional probability of a topic's top words [Mimno et al. 2011] using probability estimates based on word frequencies in the corpus data. This measure was found to be particularly good at detecting "chimera" topics [Mimno 2012b]. Unsurprisingly, this method significantly underperformed methods that used an external reference corpus [Aletras and Stevenson 2013].

Musat et al. [Musat et al. 2011] used WordNet [Miller 1995] to measure topic cohesion (how related topic words are) and specificity (how general is the common hypernym of words in the topic) and ranked topics by a weighted sum of these values. They found that their metric correlated very strongly with agreement among human assessments on a word intrusion task though they did not quantitatively compare their method to previously proposed coherence methods. Their work extends naturally to labelling topics with WordNet category labels.

With the exception of the WordNet-based method of Musat et al. [Musat et al. 2011][14], these coherence measures have been recently compared [Lau et al. 2014][15] together with the addition of a variant on Newman et al.'s method that uses normalised pointwise mutual information (NPMI [Bouma 2009]). Two training corpora and two reference corpora were used, one each from NY Times articles and Wikipedia[16]. Coherence measures were compared to human assessments via a word intrusion task [Chang et al. 2009] and a Likert scale. When assessing each model as a whole, all measures matched human assessments very accurately with the single exception of Newman et al.'s PMI based method, which performed only moderately well on one of the two models. The two human assessments were also very highly correlated at the model level. Assessing individual topics was identified as a more difficult task, with human assessments showing only moderate consistency[17]. Automated methods all correlated moderately well with human assessments, slightly worse compared to the word intrusion task, some slightly better compared to the Likert scale task. Lau et al. [Lau et al. 2014] also presented a technique to automate the topic intrusion task. This measure matched the human assessments of the same task very accurately when averaged over each model, though underperformed on individual topics.

Rosner et al. [Rosner et al. 2014] introduce a set of coherence measures drawing on results from scientific philosophy. Their group later proposed a framework for representing and combining coherence measures [Röder et al. 2015] which they evaluated on several, publicly-available data sets and with several proposed coherence measures[18]. Using their framework, they identified variations on existing measures that outperform all previous measures. Of note in this work is

[14]Musat et al. were cited but their method was not included in experiments.

[15]For fairness, probability estimates for Mimno et al.'s model were computed using the reference corpus.

[16]The reference corpora were substantially larger, though with overlap.

[17]Pearson correlations of .75 between subgroups for the word intrusion task and .63 in a 1 vs. rest analysis of the Likert responses

[18]Musat et al. [Musat et al. 2011] were not included in this study either.

the notion of "indirect confirmation measures" (a confirmation measure being a generalised conception of confirming the relatedness of sets of words). Indirect measures are able to detect words that semantically support each other (e.g.: different brands of car) but rarely coincide in documents by considering other words supported by both (e.g. words such as muffler, wheel and drive).

Posterior Predictive Checks (PPC) are a method for assessing if a latent variable model fits the data [Rubin 1984; Gelman et al. 1996]. A *discrepancy function* is identified which measures effects that are important to the task at hand. After Bayesian estimation of the posterior distribution, the behaviour of this function on the data is compared to the distribution of behaviours predicted by the posterior — a discrepancy indicates that the model is lacking. This predictive distribution is typically estimated by evaluating the discrepancy function on numerous samples taken from the posterior.

A variation of this process is to use a *realised discrepancy function* [Gelman et al. 1996] — a function of the data *and latent variables* of the model. Such a function induces an ordinary discrepancy function by integrating out the latent variables, however it can also be used to assess properties expected of the latent variables. This is particularly useful when MCMC (Markov Chain Monte-Carlo) methods such as Gibbs sampling have been used to estimate the posterior, as realisations of the latent variables are estimated also.

Mimno and Blei [Mimno and Blei 2011] presented two realised discrepancy functions for testing topic models. Both functions challenge the assumption that words drawn from a topic are drawn independently of the document into which they are placed. They measure this by the mutual information[19] between words W_k assigned to a particular topic k and document indices D_k of those words. This can be done at two levels: for whole topics they consider the mutual information $MI(W_k, D_k|k)$ between word assignments and corresponding document indices; for individual words they consider the *instantaneous mutual information*[20] $IMI(w_k, D_k|k)$ of assignments of a particular word w to each topic k and its document assignments for that topic. If the independence assumption holds, these values should be low (in the limit of infinite samples, zero).

If the independence assumption holds (i.e.: $P(w|d, k) = P(w|k)$ for words w,

[19]The mutual information of two random variables is a measure of the mutual dependence between the two variables. More specifically, it quantifies the "amount of information" that can be obtained about one random variable through the other.

[20]Instantaneous mutual information is the same as mutual information where one random variable is the occurrence of a particular instance.

documents d and topics k), these values are related:

$$MI(W_k, D_k|k) = \sum_w P(w|k)IMI(w, D_k|k) \tag{2.7}$$

In essence, these functions measure the presence of extra structure in the data beyond topic assignments. High mutual information values indicate that words assigned to a topic are *not* evenly distributed among documents — an effect that is not captured by the model is at play, causing extra structure in the data that the model cannot explain.

Other PPC's for topic models are also possible. Mimno and Blei suggest the use of *pointwise mutual information* between words [Newman et al. 2010]. Another study [Grimmer et al. 2011] predicted document keywords (the most probable words given the posterior and document) and verified this against human expert keyword choices.

Often PPC's are applied to the same data used to estimate the posterior. In this case, there is a danger of overfitting — a model that fits the data well but does not generalise. One cannot argue that such a model represents general, repeatable real data generation processes, as there is no assurance that the model is not a model of a particular instantiation of random effects. The typical approach to counter this is to test with data that was not used to infer the model. By comparing the distribution of *p-values*[21] from a collection of synthetic data sets with the uniform distribution (using a Kolmogorov-Smirnov test), Mimno and Blei deduce that their PPC method is not overfitting.

Other Assessment Methodologies Another approach to assessing individual topics is to detect insignificant "junk" topics. Al Sumait et al. [AlSumait et al. 2009] present several heuristic measures based on observations of types of uninteresting and junk topics in previous studies.

Sterckx et al. [Sterckx et al. 2014] use a supervised model [Ramage et al. 2009] as a gold standard, assessing models for their ability to capture topics in the supervised models. They also improve on the technique in [Chuang et al. 2013] for comparing a learned model to a gold standard model by investigating the kurtosis of the distribution of cosine similarities between learned topics and the human generated topics.

In a recent analysis of 19th century novels, Jockers et al. [Jockers and Mimno

[21] The posterior probability of a discrepancy less than the observed value.

2013] used variations on statistical permutation tests [Smucker et al. 2007] to make arguments about differing usage of themes between male and female authors. These tests essentially entailed shuffling words or authors many times to estimate expected proportions of e.g. male author usage of a particular topic, then comparing actual values to this distribution. If the actual values are highly unlikely, then the hypothesis that e.g. the topic is not disproportionately used by male/female authors is not supported. Similar tests can be done with other document metadata and, as with posterior predictive tests, such observations can be used to modify the generative model to incorporate metadata to which the model is sensitive.

Topic Model Presentation The typical approach for presenting topic models is to list the 10 most probable words from the topic. This approach biases the presentation of words that have a high frequency in the corpus as a whole, and is particularly troublesome when very high frequency words are allowed in the model, as these words typically have high probability in many topics and are thus presented as indicative of those topics despite not being particularly "special" to those topics. Similarly, low frequency words can have low rank, even if they are very specific to the topic.

Blei and Lafferty [Blei and Lafferty 2009] proposed an approach to counter this effect inspired by the well-known word scoring metric for information retrieval, term frequency–inverse document frequency (TF-IDF) [Salton and Buckley 1988]. In Equation 2.8, t represents a specific topic, w a word or token, ϕ_w^t the probability of w in topic t and K the number of topics in the model.

$$\text{term-score}_w^t = \phi_w^t \log\left(\frac{\phi_w^t}{\left(\prod_{j=1}^{K} \phi_w^j\right)^{\frac{1}{K}}}\right) \tag{2.8}$$

Here, a word's score in a topic is scaled according to the ratio of its probability in the topic to the geometric mean of its probabilities in all topics. A word whose probability is high in this topic relative to others will be promoted. Conversely, low probability relative to other topics will produce a negative score.

This measure depends only on topic-word probabilities and thus is, in a sense, independent of the corpus used for estimation. For assessing semantic content of new documents, this may be desirable. On the other hand, for summarising

the semantic content of a corpus, one might expect the added information in topic-document probabilities within the corpus to be important.

Chuang et al. [Chuang et al. 2012] introduced two measures of a word's significance in a topic model. First, the "distinctiveness" of a word w they define as the Kullback-Leibler (KL) divergence [Kullback and Leibler 1951] of overall topic probability $P(t)$ from topic probability given a particular word $P(t|w)$ as a measure of word informativeness given a topic model:

$$distinctiveness(w) = \sum_t P(t|w) log_2 \frac{P(t|w)}{P(t)} \tag{2.9}$$

The KL divergence of $P(t)$ from $P(t|w)$, or $D_{KL}(P(t|w) \parallel P(t))$ (also known as "information gain" or "relative entropy") measures the quantity of information (in bits) needed to specify $P(t|w)$ given knowledge of $P(t)$.

Second, Chuang et al. define the "salience" of a word w as its *distinctiveness* weighted by word probability:

$$salience(w) = P(w) distinctiveness(w) \tag{2.10}$$

These measures utilise knowledge of document topic proportions in order to estimate $P(t|w)$ and $P(t)$, though in some settings it may be possible to estimate these quantities in a different way. When assessing the semantic content of a corpus, this is an advantage as a topics prominence in the corpus is informative for such an enquiry.

Another approach to presenting topics is to identify and use meaningful topic labels. For example, [Aletras et al. 2014] tested topic representations using sets of topic words, textual phrases and images in a document retrieval task, finding that textual phrases performed better than keyword lists and images.

2.4.3 Data Pre-Processing

Human generated text is an extremely complex data form. Techniques for text analysis necessarily make substantial assumptions/approximations to the structure of the text in order to be at all tractable. In the case of topic models, actual word frequency and co-occurrence distributions may differ significantly from the typical Dirichlet priors and from distributions resulting from the model structure, leading to topics that may not faithfully represent the sought after document semantics. For example, large vocabularies and zipf distributed word

frequencies may allow topic models to detect word co-occurrence relations that are not strictly semantic in nature, reducing the utility of resulting models.

For these reasons, a pre-processing step is often used to remove aspects of the text that significantly skew or add substantial noise to the analysis results. The purpose of many of the approaches below is an attempt to obtain a smaller, denser vocabulary to capture more focussed semantics in the topics. Other approaches attempt to ameliorate statistical anomalies in co-occurrence and word frequencies.

Though the following list focuses on pre-processing steps typically used for topic modelling, most of these techniques are also used in other text analysis approaches (particularly bag-of-words based approaches). Typical pre-processing steps for topic modelling are:

- *Tokenisation*: processing text to identify the tokens or words is not always trivial, especially with social media data such as tweets.
- *Removal of "stop words"*: these are words that are frequent but carry little or no information relevant to the task at hand.
- *Word stemming*: verb tenses and plurals may be considered to contain no relevant differences in meaning, so suffixes such as '-ing' and 's' are often removed.
- *Named entity recognition*: look for constructs such as 'Barak Obama' or 'Mr Foo' and replace them with a single token, also attempting to equate that token with the same 'entity' in other forms (eg: 'President Obama' or 'Mr A. Foo').
- *Duplicate document detection*: duplicate or near-duplicate documents often fail to satisfy the (usually statistical) assumptions of text analysis algorithms, and can produce unexpected or spurious results. Detecting and removing duplicates may improve analyses.

A number of other pre-processing techniques have been used in the literature. Talley et al. [Talley et al. 2011] identified vacuous topics by their strikingly uniform distribution over documents without any documents strong in the topic. Words from these topics were removed from the data. They also created a vocabulary of ~600 acronyms and ~4200 commonly used bigrams and phrases.

Another approach is to restrict the vocabulary to words known to be of specific interest. Poldrack et al. [Poldrack et al. 2012] restricted the vocabulary to words found in a domain specific ontology when analysing a collection of related research

papers. Restricting the vocabulary in this way can be useful for summarising documents with a known conceptual framework, but prevents the discovery of novel conceptual associations with unexpected words.

The opposite — *not* restricting vocabulary with standard methods such as stop word removal is also an option. This approach has been used in connection with adaptor grammars in [Wong et al. 2012], where stop words were included in an extension of standard topic models in the task of identifying the native language of authors of English text for whom English is a second language. The approach has also been used by Thelwall et al. [Thelwall et al. 2012], where standard machine learning approaches were used for sentiment detection. They did not remove stop words in their analysis, noting that they are potential indicators of sentiment. Due to the psychometric intent of the models used in Chapters 5 and 6 and the clear relevance of "function words" (such as articles, pronouns) as indicators of social psychological processes [Chung and Pennebaker 2013], stop words were not removed.

O'Connor et al. [O'Connor et al. 2010] provide a good example of tweet tokenisation. Their tokeniser treats "hash tags, @-replies, abbreviations, strings of punctuation, emoticons and unicode glyphs (e.g. musical notes) as tokens". I have employed a similar approach in Chapters 5 and 6.

Though not commonly noted in the topic modelling literature, repeated subtexts can have a detrimental effect on topic model performance [Cohen et al. 2013], an observation I have also made when modelling the data presented in this thesis. A simple approach to reducing this problem is to remove repeated texts. It can be argued that in this way little thematic information is lost (since the text originals remain). This approach was applied in the models from Chapters 5 and 6, where retweets were removed. It should be noted, however, that retweets often contain small amounts of extra text, which is lost when they are removed.

A recent topic model includes duplicate sub-texts in the generating model [Cohen et al. 2014]. Their model is specific to the patient record context (groups of notes drawing text from a single 'root' patient record per group), though it could be generalised to other duplicate structures such as retweets with relatively little effort. This approach is superior to simply removing near-duplicate documents, as the remaining, non-duplicated words are also modelled. I have not developed such a model, however it would be of interest for future work, especially given the importance of retweets as indicators of community acceptance of the tweets contents.

2.4.4 Topic Model Variants

There is a rich literature of variations and adaptations of the original LDA model which continues to grow to this day. I have attempted to provide a substantial list here, describing the most important advances and a number of specific models indicating the breadth of applications, however the literature is extensive and growing continually and this list is in no way exhaustive.

Time Series LDA

Griffiths and Steyvers [Griffiths and Steyvers 2004][22] used linear trend analysis to identify topics whose prevalence changed over time (out of 300, 54 increased significantly, 50 decreased significantly, both with $p = 0.0001$) and linked the most significant with events and changes in the appropriate disciplines. They noted that including time in the generative model would be a more appropriate approach. One study attempting to identify topic shifts over time [Hall et al. 2008] employed a post-hoc time analysis rather than using LDA variants that incorporate changing topics.

Several models have been proposed that incorporate document time stamps in the generative model. The first such model is the Dynamic Topic Model (DTM [Blei and Lafferty 2006b]), where chained logistic normal distributions are used as priors for document-topic and topic-word multinomials.

The Topics Over Time model (TOT [Wang and McCallum 2006]) has fixed topic-word and document-topic distributions over time (using the standard Dirichlet construction), but also generates document time stamps via a Beta distribution for each topic. Estimation is achieved by Gibbs sampling similar to LDA (for topic and document distributions) and the method of moments for topic-time Beta parameters, applied each Gibbs iteration.

The Dynamic Mixture Model (DMM) by Wei et al. [Wei et al. 2007] also has fixed topic-word distributions, but conditions the topic distribution for each document on that from the previous one: the prior for the topic distribution of a document is a Dirichlet distribution parametrised by the previous document's topic distribution. The Dirichlet distribution requires one more parameter, which is provided as precision parameter common to all documents. The authors intention was to model time series data such as diffusion processes or water flow in water distribution networks. Their model outperformed SVD and standard

[22] The paper that first introduced Gibbs sampling for LDA

LDA in this setting. Though they cite the DTM, they do not compare it to their model.

Pruteanu et al. [Pruteanu-Malinici et al. 2010] used single topic assignments per document (as opposed to per word assignments in most models), fixed time topic-word distributions (as TOT), and time varying document-topic distributions which are a mixture between a distribution drawn from a Dirichlet prior (as in standard LDA) and the previous document's topic distribution. A Beta distribution[23] is used as a prior for mixture proportions. The assumption is that the corpus is a series of documents, each related to the previous one, as in a diary, news headlines or chapters in a book.

Non-parametric models incorporating time have also been proposed (e.g.: [Li et al. 2012]).

Non-Parametric Topic Models

A perceived shortfall of standard LDA is the necessity of specifying the number of topics. Non-parametric topic models[24] essentially attempt to obtain topic model parameters in a Bayesian manner.

The first work on this family of models introduced the Hierarchical Dirichlet Process (HDP) [Teh 2006][25] where a Dirichlet Process is provided as a prior to the Dirichlet Processes from which words are drawn. This formalism allows a theoretically infinite number of topics, however only a finite number have non-negligible contribution through the action of a concentration parameter.

The Sparse Topic Model (sparseTM) [Wang and Blei 2009] extends the HDP by decoupling sparsity and smoothing in the model. In the HDP model, posterior inference prefers a large number of sparse topics to explain the observed words, however this is at the cost of local smoothness in the per-topic word distributions resulting in less-smooth document predictive distributions. The sparseTM addresses this by introducing a Bernoulli variable for each term and topic, allowing terms to be excluded from topics. This Bernoulli variable then provides desired sparsity, allowing the Dirichlet Process to provide greater smoothing among the remaining words. The sparseTM was able to consistently provide simpler models with better perplexity than HDP models.

[23] The Beta distribution is the same as a 2-dimensional Dirichlet distribution.

[24] These are sometimes referred to as "hierarchical" models as the priors for the modelled parameters can introduce new parameters, which can in turn be modelled introducing futher parameters etc...

[25] Generalising LDA is but one application of this approach, for example modelling shared topics across corpora can be achieved through an extra DP layer across corpora.

A review of hierarchical Bayesian non-parametric models was conducted in [Teh and Jordan 2009] appearing in the book "Bayesian Nonparametrics" [Hjort et al. 2010]. Many subsequent non-parametric topic models have been, and continue to be, proposed including many models mentioned in the remainder of Section 2.4.4.

Models Accounting for Linguistic Properties of Text

It has long been established that word distributions in many languages (including English) follow a power law. In linguistics circles, this is known as Zipf's law. Priors for word frequencies derived from the Dirichlet distribution do not produce this effect. Sato and Nakagawa [Sato and Nakagawa 2010] introduced parametric and non-parametric topic models based on the Pitman-Yor distribution (also termed the two parameter Poisson-Dirichlet distribution) [Pitman and Yor 1997] which address this problem. This is of particular interest when attempting to capture psychological characteristics of text that relate to the use of common words such as pronouns and articles.

Another known characteristic of text is word "burstiness": if a word appears once in a document, it is likely to appear several more times [Katz 1996]. Doyle and Elkan [Doyle and Elkan 2009], introduced a model utilising the Dirichlet compound multinomial [Madsen et al. 2005] that accounts for this phenomenon. Inference for this model is, however, slow, limiting its wider utility.

Utilising "table indicator sampling" [Chen et al. 2011], a new inference technique for models with Pitman-Yor process priors, and through a computational trick incorporating a word burstiness component with little computational or memory overhead, Buntine and Mishra [Buntine and Mishra 2014] were able to combine both these advances in a single model with very competitive computational and memory overheads.

Models Incorporating Network Information

A number of models have been proposed that include network structures.

Early in the development of Bayesian topic models, a model was presented that generated links between documents (citations in their example) as well as document words (abstracts in their example) [Erosheva et al. 2004]. Utilising a "bag of links" representation, links were generated in a similar way to words, utilising the same document-topic distribution as used for word generation and a separately inferred topic-link distribution. This model is unusual in that it does

not specify a prior over topic-word and topic-link probabilities, treating them instead as fixed model parameters.

In the Relational Topic Model (RTM) [Chang and Blei 2009; Chang and Blei 2010], links between documents are generated with probability dependent on a weighted sum of the per topic products of normalised document topic allocations (the weights are parameters in the model and another scalar intercept parameter is added to the sum). They used two candidate functions to calculate link probabilities, a sigmoid (so each link is inferred by logistic regression) and an exponential (where link probabilities increase exponentially with similarity in topic proportions), finding that the exponential form performed substantially better. The Dynamic Relational Topic Model (dRTM) [Wang et al. 2011] is a non-parametric model that extends the RTM, including document time stamps and a random effect term to account for spurious links. Another generalisation of relational topic models can be found in [Chen et al. 2013]. Here the link probability function includes products of topic allocation counts between different topics, allowing topic interaction effects. They also introduce regularised Bayesian inference and present an efficient Gibbs sampling algorithm (see the paper for details).

A model, named "Topic-Link LDA" [Liu et al. 2009], published shortly after the Relational Topic Model, generates links between documents as well as a (latent) matrix of author community affiliations. Again, document links and words are observed. Words are generated as in standard LDA, and document links are generated according to a sigmoid function applied to a weighed sum of three factors: the dot product of topic proportions (similar to the RTM with equal weights, though topic proportions are not normalised), the dot product of author affiliations and a scalar intercept parameter. The authors note that an exponential function could be used in place of the sigmoid, as was explored in the RTM.

Further developments of topic models incorporating network data have been done by Lim et al. [Lim et al. 2013]. They incorporated a Gaussian network-generation model of Twitter follower networks into an author-topic model with an innovative use of hash tags as special tokens that acted as a prior on word usage. The model used a non-parametric Pitman-Yor prior on words and hash tags. Their model performed extremely well, however was not scalable. Later, a similar model that excluded the network information but added a sentiment lexicon performed well at detecting opinions about products [Lim and Buntine 2014b].

A number of models have attempted to incorporate citation networks into

topic models of scientific publications. Examples include [Nallapati et al. 2008; Lim and Buntine 2014a].

Finally, community detection combined with topic modelling has been attempted. Duan et al. [Duan et al. 2011] developed a full Bayesian model incorporating both a stochastic block model for community detection and hierarchical Dirichlet process for topic detection. In this model, all of an author's documents are assigned to just one community (hence they do not overlap) and its scalability is questionable. Li et al. [Li et al. 2012] present a different approach to combined community and topic detection by utilising extra thematic metadata — hash tags (Twitter data) and publication venue (citation data). The Twitter follower network was not utilised. In their model, communities (not documents) have topic mixtures and topics generate both words and hash tags/venues. Another approach used US Senate votes to indicate links (Senators are linked if they voted the same way), where the network for each vote is generated by topic specific group assignments and topics generate the bill under vote. Subjective analysis suggested the model was finding meaningful alliances. This model also used non-overlapping groups.

Informed Priors

Sometimes extra semantic information is available or presumed to be present in the data which one would wish to incorporate into the model as prior information. As always with Bayesian modelling, it is best to evaluate to what extent the model and its priors are applicable to the data and task at hand. Indeed, in Chapter 5 I observe that topic regularisation can break an independence assumption needed for the presented methodology. I would argue that one should exercise caution when applying ad-hoc priors, and apply posterior predictive checks or other means of verification that estimated models represent true and meaningful structures in the data, and are not merely an artefact of the applied prior.

One approach to using lists of related words that you expect to convey specific semantic information is to use the lists to form initial values for topic-word allocations prior to model estimation. If one expects a topic model to have multiple local optima in the posterior, this approach may help to choose the optima closer to the expected semantic relations. It was used by Hall et al. [Hall et al. 2008] but to my knowledge, no systematic analysis of its effect has been carried out. Jagarlamudi et al. [Jagarlamudi et al. 2012] incorporate such information into the generative model itself. Using seed words derived automatically from la-

belled data, chosen through high information gain of candidate seed words in the class of documents with a specific label. They report significant improvements in a document clustering task. The "Labelled LDA" model [Ramage et al. 2009; Ramage et al. 2011] utilises semantic labels on documents directly, associating specific topics with each label and promoting or enforcing those topics in labelled documents.

Drawing on observations from investigation of topic coherence, Newman et al. [Newman et al. 2011] used pointwise mutual information (PMI) between words in a reference corpus to provide a prior on topic-word probabilities. They developed efficient semi-collapsed Gibbs sampling procedures in which sampling word-topic allocations were interleaved with MAP estimation of topic-word probabilities. The stated aim of their work was to improve topic model quality for small and/or noisy data. Experiments derived from 3 standard data sets established the efficacy of their work as measured by PMI (as a proxy for topic coherence), perplexity and human assessments. Wahabzada et al. [Wahabzada et al. 2012] developed a variation of the same approach that uses variational Bayes estimation.

Other LDA Variants

Since the introduction of LDA, there have been many other variations, incorporating other relevant meta-data or structures expected therein and a number of other improvements to the basic model. Below is a summary of some of the main contributions that do not fit neatly into the categories described earlier in this section. Again, it does not attempt to be exhaustive, but instead to provide examples of what has been and can be done.

An early example is the author-topic model [Rosen-Zvi et al. 2004], in which each author has a mixture of topics and documents are generated by topics drawn from their author. This model has inspired other extensions and adaptations (e.g.: [Lim et al. 2013]).

The hierarchical topic model [Blei et al. 2004] and its non-parametric successor "hierarchical LDA" (hLDA) [Blei et al. 2010] provide topic hierarchies[26]. To capture topic correlations, the logistic normal distribution has been used to replace the Dirichlet [Blei and Lafferty 2006a]. Recent work by O'Connor [O'Connor et al. 2013] has proposed a latent Dirichlet multinomial model to discover 'semantic frames' — types of events and their participants.

[26]This model has a hierarchy of topics, not to be confused with Hierarchical Dirichlet Process models [Teh 2006] which have hierarchy in model parameters.

Several proposed models incorporate multi-word entities [Wallach 2006; Wang et al. 2007], and another fused an existing Hidden Markov Model for language syntax with a topic model [Griffiths et al. 2004]. Multi-word entities have also be explored at the text tokenisation stage [Lau et al. 2013], finding that using n-gram tokens improves topic coherence. A polylingual topic model [Mimno et al. 2009] (which assumes each document has several translations) has also been developed.

Repeated subtexts have been observed to cause problems for topic models, as there is a tendency for topics to form that simply reflect the repeated text. A recent work tackles this problem for the modelling of medical case notes [Cohen et al. 2014].

A model named "Correspondence LDA" (CORR-LDA) [Blei and Jordan 2003; Mimno and McCallum 2008] was developed to jointly model image features and caption text. Later, a generic model based on Dirichlet Multinomial Regression (DMR) was proposed, incorporating arbitrary extra document features [Mimno and McCallum 2008]. There are also models incorporating various document relations [Roberts et al. 2014; Du et al. 2012].

2.5 Summary

This chapter has presented a review of several areas of literature relevant to the vision and approaches developed in this thesis. Starting with a review of relevant work from psychology and sociology to motivate this thesis and draw connections between text and the psychology of its author, I went on to describe background work on social media data and Twitter in particular and finally presented overviews of existing work on community detection in networks and Bayesian topic modelling.

In Section 2.1.1, social representation theory combined with the distributional hypothesis of human cognition were presented as a motivation for the core premise of this thesis — that social entities such as norms and collective identities may be identifiable by contextual patterns in the communications of social groups. Research into the manifestation of an author's psychological state in the words they use (Section 2.1.2) enriches that premise, providing tools to investigate the role and meaning of detected entities.

One of the contributions of this thesis is a methodology and tool for the targeted collection of social media data from a particular social group. Section 2.2 provides an overview of relevant work on social media data, providing background

information on social media, and in particular Twitter, usage and establishes its role as a social communication medium and the role of hash tags as meeting places for social media users. A brief review of the pro-anorexia presence on Twitter (Section 2.2.4) establishes the presence of that particular community and indicates the level of interest in the psychology research community for developing an understanding of anorexia and eating disorder related social media usage.

A brief review of research in community detection in networks, relevant to Chapter 6, is presented in Section 2.3.1. A discussion of research into network dynamics of social networks, motivating the collection of dynamic network data in Chapter 3, is presented in Section 2.3.3

Finally, a substantial review of the Bayesian topic modelling literature is presented in Section 2.4. This review goes beyond what is needed to support the topic models used in the Chapters 5 and 6, however provides insights into future applications of those methods with more sophisticated topic model variants.

In the next chapter, I develop a methodology for collecting social media data from a particular targeted community, discuss lessons learned from the design and development of a data collection system around that methodology and describe its application to the pro-anorexia and eating disorder community operating on Twitter.

Chapter 3

A Data Collection System for Dynamic Twitter Data

This chapter summarises the data collection techniques and primary data set used in this thesis. The content in this Chapter has been published in [Wood 2015a].

The study of social processes requires data that is both richly dynamic and relevant to some coherent social context. To date, data for research into online social media lacks rich dynamic information for some aspects of the data (e.g.: Twitter follower network dynamics) and often is not focussed on particular social groups or other social contexts.

This chapter presents a data collection strategy that identifies a social context and a system that collects richly dynamic data from that context. The strategy and system have been used to collect a substantial number of tweets and related (dynamic) network data from the "pro-ana" (pro-anorexia) Twitter community. This data is the primary data set used in this thesis.

This chapter is organised as follows: In Section 3.1, I present an overview of the motivations, innovative methodologies and the resulting data collection system. In Section 3.2, I discuss communities defined by hash tags and the approach used to identify tags from the Twitter "pro-ana" community. In Section 3.3, I discuss the approach for sampling the dynamics of community network and user data. In Section 3.4, I discuss software architecture and design as well as several technical challenges that may be of wider interest. In Section 3.5, I present statistics from the collected data giving some indication of the volume of dynamic user and network information collected. In Section 3.6 I talk about possible system enhancements and related future research directions and finally in Section 3.7, I summarise my findings.

3.1 Overview

The study of social processes in on-line social media content is a relatively new and rapidly growing endeavour. Many social media platforms provide a public API (Application Programming Interface) which can be used for the targeted collection of data from perceived communities, however existing software for this purpose focusses on a "snapshot" of the community and its communications, and ignores important aspects of its dynamics. I present an approach to identify a set of tags or keywords that identify and define an on-line community and a data collection system designed to capture tweets and the dynamics of Twitter user profile and friend/follower lists. This approach and system were used to collect a data set currently spanning 2 years and 10 months (including 3 Christmas periods) from the "pro-ana" (pro-anorexia) and eating disorder Twitter community — over 1.2 million tweets, 300 thousand users and 200 thousand images.

Social psychology is a dynamic process — people enter and leave social groups and groups change and adapt their sense of identity and social norms. It is these dynamics that make our societies what they are, that generate and define human social fabric. In order to study and eventually make predictions about group behaviour, it is essential that we capture the dynamics of the socially meaningful features under study.

With the exception of one very recent paper [Myers and Leskovec 2014][1], data collection for the study of Twitter friend/follower network dynamics has been limited to a small number of network snapshots without finer grained timing [Xu et al. 2013; Hutto et al. 2013; Rainie 2014] and focus on large scale and/or short term effects, without consideration of specific Twitter communities.

What I propose and demonstrate here addresses both these short-falls in data collection. Working from the observation that user generated tags in social media can define a community [Yang et al. 2012; Starbird and Palen 2011], I propose an approach analogous to adaptive cluster sampling [Thompson 1990] to iteratively expand one or two initial tags thought to be used by/define a community, resulting in a more complete collection of community tags. By collecting all the socially generated texts that contain one or more of the tags, nearly all community communication through a social media platform can be collected whilst minimising the number of collected texts that do not represent such communica-

[1] Their data contains precise timing for friend/follower network changes. They do not describe how they obtained the data, however the first author was an intern at Twitter prior to publication.

tions. The approach was applied to the "pro-ana" Twitter phenomenon and its social context, revealing a set of tags that almost exclusively identified "pro-ana" and eating disorder related tweets.

Collecting all tweets from a set of hash tags is a relatively simple task — Twitter provides API's specifically for the task. Collecting fine-grained information on the dynamics of Twitter user profiles and friend/follower relations is not so straightforward, however. Twitter only provides API's that return a snapshot of current friend/follower relations and user profiles, and does not provide information on when changes occur. A similar situation exists for many other social media platforms. In order to collect information on the dynamics of friend/follower relations and user profiles, snapshots are collected for active users each time they tweet. In this way, rich dynamic information for frequent contributors is obtained whilst not wasting limited resources to collect information on infrequent or one-off contributors.

The system is designed to be efficient and robust in many other ways. For example, storing full snapshots would be redundant, and would quickly produce an overly-large data set, thus only changes to friend/follower relations and user profiles are stored. Reliability and robustness were seen as important so as to capture dynamics without interruption. The data collection system was designed to be robust to various technical outages (extreme tweet rates, Twitter API outage, network outages, hardware and software failures) with minimal impact on the continuity of data collection. Further, for data security, data is stored in a database replicated across geographically dispersed sites and regular backups are automatically performed.

3.2 Adaptive Sampling for Search Tags

Hash tags are used in Twitter and other micro-blogging sites as a way to organise, emphasise and otherwise colour posts. A hash tag is simply a word with the hash character "#" prepended, such as *#diet*. They allow users to specify aspects of their posts that they consider important and to direct their posts to what they feel is an appropriate audience [Yang et al. 2012; Starbird and Palen 2011]. It is this second point that I attempt to harness in order to collect the output of a hypothesised community around "pro-ana" (pro-anorexia) and eating disorders.

This study attempts to focus primarily on people who are experiencing or have experienced eating disorders and are participating in hypothesised communities

operating on Twitter. The aim is to gather data that may, with appropriate analyses, shed light on the such communities and their roles in the development, continuation and curing of eating disorders. There are many discussions on Twitter around eating disorder themes and indeed the "pro-ana" phenomenon itself, however these discussions are not the focus of this study.

Following the intuition behind adaptive sampling [Thompson 1990], a search query for collecting tweets was selected through an iterative process identifying hash tags used by people with eating disorders. At each step a sample of tweets was collected containing the tags from the previous step, then a set of relevant new hash tags was selected from frequent tags in that sample.

In this way, one might expect that a significant proportion of communities of people with eating disorders can be identified, as one would expect that such communities of any size would typically contain individuals who explore eating disorder relevant tweets beyond the community, and would very likely find the tags I have identified — in this case, those individuals would likely tweet with tags for the external community as well as those I identified, enabling my methodology to recognise the external community and add it's tags to the query. There may, however, be communities in the public Twitter sphere that the methodology is unable to detect, for example, communities with strong boundaries to the 'out group'. For the purposes of this thesis, I have not pursued this issue further, working instead with the detected tweets and presumed communities that created them.

In the first iteration, a brief study of tweets and Tumblr[2] posts containing *#proana* revealed related tags *#thinspiration #anorexia #bulimia* and *#promia*[3]. The *#thinspiration* tag is not used exclusively by people with eating disorders (though it is sometimes hard to tell if the author of a *#thinspiration* tweet has an eating disorder), but was deemed important to include as it directly relates to the "thin ideal" and is used extensively by people with eating disorders.

For the second iteration, tweets were collected from August 18–22 2012 on these tags, a total of 1182 tweets. Hash tags were counted in the collected tweets, and those with more than 3 occurrences were considered. See Appendix (A) for a complete list of the tags found in the sample. The majority of the tags appeared to have much wider usage than the community of interest (e.g.: *#diet*). A quick manual check by conducting a Twitter search on each tag confirmed these suspicions, and those tags were discarded. Manual checks on the remaining tags

[2]www.tumblr.com

[3]Short for bulemia nervosa, an eating disorder related to anorexia nervosa.

revealed many more with wider usage.

For each tag, a judgement had to be made about the number of irrelevant tweets compared to the number of new tweets not collected by other tags in the query. In general the distinction was clear, and no compromise was necessary. However a few, such as *#depression* and *#selfharm* were discarded despite identifying a small number of relevant tweets not identified by other tags. It was deemed preferable to maintain a higher degree of data relevance and a smaller query. Similarly, the tag *#thinspiration* was retained due to it's relevance and wide usage by users contributing to communities of people with or having had eating disorders, despite apparently collecting many tweets from people not contributing. Table 3.1 lists some typical tags that had wider usage with descriptions of the wider contexts in which they were used.

Description	Tags
diet and weight loss	*#diet #weightloss*
body shape and parts	*#skinny #thin #supermodelbody #legs #bones*
abstract terms associated with eating disorders	*#perfect #motivation #recovery #perfection #starve*
used to discuss eating disorders from outside the community	*#anorexia #eatingdisorder #eatingdisorders*
fitness and exercise	*#fitspiration #fitspo #gym*
depression and its symptoms not specific to eating disorders	*#selfharm #depression #cutting*
with other meanings irrelevant to eating disorders	*#ana #ed #mia #purge #liquiddiet*

Table 3.1: Non Personal Eating Disorder Tags

This process applied to all the remaining tags resulted in a core of 14 tags identified as used almost exclusively by people with eating disorders, as well as capturing nearly all tweets in the sample that were deemed produced by those people. Tweets with these tags were collected for a further 4 days from September 15 to 19 2012 and the analysis repeated, however no extra tags were identified. Table 3.2 lists the final set of selected tags.

#proana #promia #anasisters #bulemia #bulimic #ednos #edproblems #hipbones #thingsanataughtme #thinspiration #thinspo #abcdiet #thighgap

Table 3.2: Tags Selected For Search Query

3.3 Collecting Dynamic Twitter Data

Tweets collected with the Twitter APIs contain a snapshot of the tweeting user's profile information, giving some information about user profile changes a user may have made since the last observation of the users profile[4]. However the profile snapshot does not contain the lists of friends and followers of the user, and to the best of my knowledge historical data on changes in the friend/follower network are not commercially available[5].

Several previous studies of Twitter friend/follower network dynamics have used a small number of network snapshots without finer grained temporal information [Xu et al. 2013; Hutto et al. 2013] and collected data for a fixed set of users, without focus on particular communities. One very recent study of note [Myers and Leskovec 2014] was able to obtain a large data set containing precise timing for friend/follower network changes, however they do not describe how. Their data is sufficiently large that it would contain comprehensive information from many social groups, however it spans just one month.

The Twitter APIs do not provide a feed containing network changes as they happen, however they do provide a REST (Representational State Transfer) interface for collecting a snapshot of a user's friends and followers lists. The data collection strategy presented here uses that endpoint to poll a user's friend/follower lists each time they tweet. In this way the dynamics of the friend/follower network of users active in the data is recorded with similar frequency to their tweets and the user profile data contained therein.

It should be noted that the dynamic data thus collected is not complete. A user who watches tweets in the data set, but themselves tweets rarely, will only be polled on the occasions they tweet — profile changes and follow/unfollow actions made between recorded tweets by users not active in the data are only captured in aggregate (though follow/unfollow actions will be detected if the recipient tweets). The temporal data, therefore, has an element of sampling error, a systematic lag, which is particularly pronounced with infrequent tweeters. Nonetheless, one

[4] Although user data embedded in tweets can be stale.

[5] A recent Quora post claims that the TwitterCounter service provides this for up to one year.

might hope that changes to both the user profile and friend/follower lists which are relevant to social processes within the hypothesised pro-ana Twitter community will have a high probability of occurring near the time of each tweet to that community (i.e.: each tweet within my data set). We should, however, be careful when analysing behaviour of infrequent tweeters, as there may be a bias in the recorded dynamics of their data.

3.4 Algorithms and Technical Challenges

This section outlines the design of the dynamic Twitter data collection system and describes some of the challenges that had to be faced in order to create a stable system, robust to the the vagaries of tweet data flow, that efficiently utilises the restricted data bandwidth provided by the Twitter APIs. Over the course of two years of continuous data collection, a stable system has been built that is robust to Twitter outages and sudden increases (by orders of magnitude) in data volume, and efficiently utilises the narrow data bandwidth for collecting friend/follower data.

3.4.1 Overall Architecture

The system uses a multi-threaded architecture, enabling asynchronous HTTP requests to the various Twitter API end points and media URLs (see Figure 3.1). Communication between threads is achieved with thread-safe queues. A tweet collection thread regularly polls the *search/tweets* REST API with the query tags shown in Table 3.2. Each tweet and its meta-data are stored in the database and the authors user id is added to the friends, followers and user profile queues. Any media URLs and corresponding media entity ids are added to the entities queue.

The system's main thread (the initial thread when the system launches) initially constructs the shared, thread-safe queues and launches the other threads. It then monitors the other threads, restarting them if they crash and gracefully coordinates system shut-down when requested. To help debug frozen threads, as can happen in early development of multi-threaded systems if thread locks are not properly released, the main thread also responds to operating system QUIT signals, dumping a stack trace of each thread. As a further precaution, a unix *cron* script (which is run regularly by the operating system) monitors the system as a whole and relaunches it if an hour passes without activity. The system also

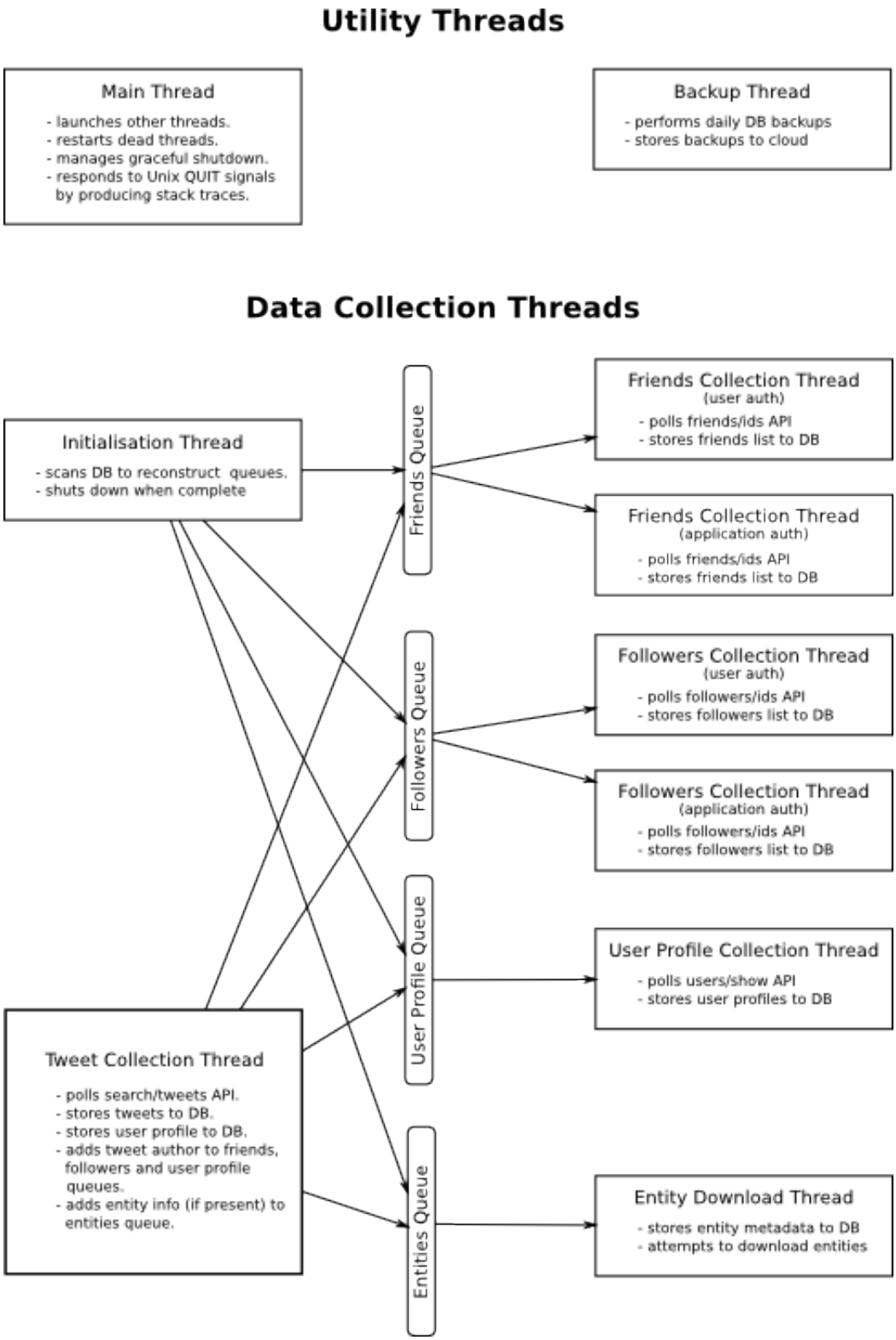

Figure 3.1: Overall Architecture. Arrows indicate data passed between threads.

has a simple thread which initiates regular database backups.

The initialisation thread reconstructs the data collection queues by scanning the tweet and user databases for partial and out of date data (e.g.: friend or follower lists that have not been polled since the last tweet from the user). Once the scan is completed, the initialisation thread shuts itself down. Other data collection threads are launched and run synchronously with initialisation.

The tweet collection thread regularly polls the *search/tweets* API endpoint, collecting all tweets that contain tags from Table 3.2. Collected tweets and embedded user profile data are stored to the database and then the tweets authors' id is placed in the friends, followers and user profile queues and any media entity meta-data in the entity queue. On system restart, all tweets since the last collected tweet are requested, however Twitter does not guarantee that some will not be missed. Our experience indicates that for this query, tweets are usually accessible for several days and down times of that order do not result in lost data. This is probably not the case during heightened tweet rates, however.

Twitter has two modes for accessing its REST APIs: with user authentication and with application only authentication. With application only authentication, you cannot perform tasks on behalf of a user (which is not needed here), but you are given separate API rate limits. For a data collection application such as this, using both forms of authentication essentially doubles the rate limit — in the case of polling friend and follower lists, this is significant. Thus four threads were used for collecting friends/follower information (each of friends or followers with each of user and application only authentication). When storing friend or follower list information, any changes to previously recorded lists are stored in the database and the lists updated. Similarly, when new user profile information is obtained from tweets or collected by the user profile thread, changes are stored alongside the new data.

Data is stored in a replicated MongoDB instance. MongoDB was chosen due to its easy deployment, easily modified schema, easy replication and because the native format of stored data is JSON, the same as that returned by the Twitter APIs. Three main collections are maintained: tweets, user profiles and media meta-data (see Table 3.3). The Nectar research cloud[6] was used to house database replicas and for reliable storages of database backups.

[6]http://nectar.org.au/research-cloud

Collection	Twitter Data	Added Meta-Data
tweets	tweet data	– how and when it was collected
entities	media entity data	– tweets that contain the entity – a history of any changes to its data – download attempts/success
user profiles	user profile data	– history of profile changes – friends and followers lists – a history of friend/follower list changes – when and how the data was last polled – number of tweets collected from this user

Table 3.3: Database Collections

3.4.2 Polling friend/follower lists — the main bottleneck

Tweet rates are highly non-linear, with relatively low rates most of the time, however occasionally, the rate increases by an order of magnitude, and very occasionally by many orders of magnitude. During these "tweet storms", the download requirements of the system often exceed the Twitter API rate limits, causing the data collection queues to grow. The four public Twitter REST API endpoints used to collect data; tweets by query string, user profiles, user friend lists and user follower lists; all have data rate limits[7] which were on occasion exceeded. Prioritisation schemes were developed to ensure timely collection of more important data.

Collection of tweets for the pro-ana and eating disorder community query fell behind during the highest tweet rates, however due to the ability of the tweet search endpoint to retrieve past tweets, the pro-ana/eating disorder query did not apparently lose data as a result.

The user profile endpoint can poll 100 users per query with 180 queries per 15 minutes. This high rate quickly caught up with even the most extreme "re-tweet storms", and it was sufficient to prioritise users whose previously stored data was oldest (unseen users first, ties resolved by user id).

The API endpoints for friends and followers lists poll only one user per query and 15 queries per 15 minutes. Also, each query returns a maximum of 5000 friend/follower ids — occasional users with millions of followers require hundreds of queries. Frequent 'moderate' re-tweet storms often took days to clear the queue, and extreme events could take weeks. This substantial delay was considered unacceptable.

[7] https://dev.Twitter.com/rest/public/rate-limits

Investigation of the re-tweet storms indicated that the majority of tweeting users had no other tweets in our data, especially for the more extreme events. Thus a strategy was implemented where users with at least one other tweet in our data were given priority. Of those, the user whose friend/follower data was oldest (i.e.: least recently polled) was given priority. With this strategy, more frequent tweeters were quickly re-polled, while the queue of less interesting one-tweet users can take many days to eventually clear. A newly seen user who tweets again before having been polled is moved to the front of the higher priority queue ('never' is considered least recent). In an attempt to get a snapshot of at least some one-tweet users friend/follower lists at the time they tweet, the most recently seen first-tweet users are polled first. In both priority schemes, rare ties are resolved by lexical order of user names.

User seen before?	Priority	Which user first?
Repeat User	First Priority	Least recently polled
New User	Second Priority	LIFO queue[8]

Table 3.4: friend/follower lists polling priority scheme

3.4.3 Image Collection

Many tweets, and especially tweets in this data, contain images. Twitter includes media URLs in tweet meta-data, and assigns a unique id to each image. When a tweet containing images is collected, the images meta-data is stored in the database including a link to the tweet. If the image with that Twitter image id has not been downloaded yet, it is downloaded and stored as a file on disk. A simple 'cron' script is used to backup the stored images to the servers running the database replicas. It is common for duplicate images to be assigned different Twitter image ids. The system does not attempt to identify such duplicates.

3.4.4 Other Technical Challenges

Tweets can be directed to a recipient Twitter user. Collecting the friend/follower lists of recipients was also attempted, however it soon became evident that recipients were frequently celebrities with millions of followers, causing extra burden on the already stretched follower API endpoint. Users of interest that are part of the pro-ana/eating disorder community would be tweeting regularly, and their

[8]Last In First Out queue

friends and followers lists would be polled regularly anyway, so it was decided that polling tweet recipients should be abandoned.

To test the relative reliability of the Twitter streaming and tweet search APIs, a separate process received tweets via the streaming API and stored them in an extra database collection. This collection was monitored by the main program, and any extra tweets were copied to the main tweet collection. I found that no tweets would have been lost without the streaming API data, so this part of the system is unnecessary.

During system development, a bug in the python http library and difficulties coordinating thread locks were identified from stack traces generated by the main thread in response to Unix QUIT signals. Early in development, data loss was avoided by automatically restarting the system via an hourly Unix cron script when further bugs triggered by infrequent combinations caused the system to crash. The system has now been running continuously for over a year without any of these problems.

During the first year of data collection, Twitter announced that it was making significant changes to its APIs, and especially to rate limits and the ways they are reported and applied. The system attempts to utilise its rate limits as fully as possible without exceeding them (which can prompt Twitter to block the application for a time), so the API changes required substantial adjustment to the rate limit monitoring logic. There were also a few changes to the meta-data for tweets and users. This did not directly require changes to program logic, however in order to keep database consistency, some logic was added to update old format records.

At the time of initial development, Tweepy[9] was chosen for access to the Twitter APIs. Unfortunately, at that time Tweepy did not have support for application only authentication. Twython[10] did however, and since both packages present the Twitter API in a similar way, it was not difficult to add threads that utilised this capability.

3.5 Dynamics of Collected Data

Figures 3.2 and 3.3 give some indication of the scale of captured user profile and friend/follower list dynamics. It can be seen that for tens of thousands of users, more than a hundred friend and follower list changes have been recorded and

[9] http://www.tweepy.org/
[10] http://twython.readthedocs.org/en/latest/

for thousands of users, more than a thousand changes recorded. Similarly, for hundreds of users, hundreds of profile changes have been captured, and for many thousands of users, tens of changes. Our goal of capturing friend/follower list and profile dynamics has succeeded.

As with many aspects of social phenomena, the number of changes to friend and follower lists as well as the number of user profile changes per user follow an a distribution similar to a power law (linear plot in a log-log scale graph). There is a notable change in exponent (slope of the plot) leading up to around 1000 changes for friend and follower lists. This may be related to Twitters policy restricting increases in user friend lists beyond 2000 friends (see Section 4.1). It is remarkable that the four most prolific users recorded over 100,000 changes in their friends lists — over approximately 1000 days of data collection, this means that these users made 100 changes per day on average. It would be interesting to investigate these users more closely and attempt to understand what drives such behaviour (or if the users in question are human or automated).

In considering if a profile had changed for Figures 3.2 and 3.3, automatically generated profile attributes such as the number of tweets, followers, favourites, etc. . . were not considered. Changes in *https* versions of profile image URLs were also not considered, as it was noted that Twitter provides these from different web domains depending on the way in which the user information is collected (embedded in a tweet or via the *user/search* API endpoint). Since both methods were used to collect profile data, this resulted in many recorded changes to these URLs where the actual images did not change.

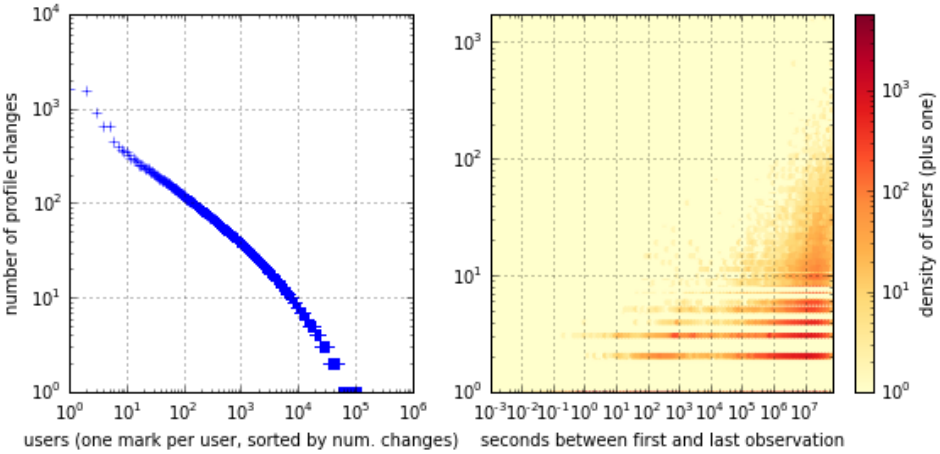

Figure 3.2: Number of user profile changes and user observation windows (one mark per user).

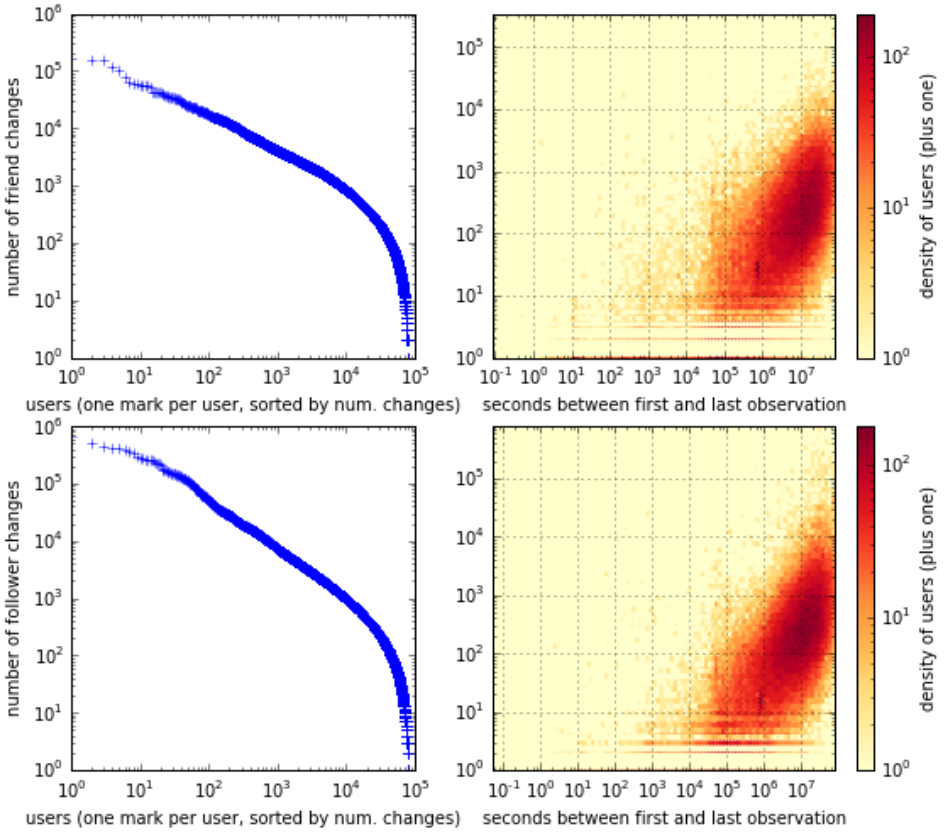

Figure 3.3: Number of changes to friends (above) and follower lists (below), and user observation windows (one mark per user, right).

3.6 Discussion and Future Work

One question that may be of interest for prospective users of the system is it's scalability. With the pro-ana community, the system was successful at collecting follower network dynamics of core users with only very occasional delays in collecting follower lists. For other communities with a significantly greater number of users and tweets, this situation could be expected to deteriorate — the frequency of "tweet" storms capable of disrupting and delaying follower data collection would increase rapidly with the average tweet rate, and for high volume communities, long delays of several days, weeks or even longer between a core user tweeting and the next poll of their friend/follower lists could be expected.

It should be noted that the collection of tweets was not compromised by high tweet volumes from the pro-ana community, and substantially more prolific communities would be possible to trace. Further work on identifying, prioritising and regular polling of "core" users could prospectively allow effective collection of dynamic network data from substantially larger communities. Several potential strategies to achieve this outlined below.

The experiences from identifying the pro-ana Twitter community and collecting their public data have raised several research questions relating to the methodology and its implications as well as a number of possible strategies to improve the efficacy of data collection.

The methodologies presented in Sections 3.2 and 3.3 suggest several questions worthy of further investigation.

- The final search query from tags in Table 3.2 represents a balance between a wide net and saturating the Twitter API limits. The choice was made to keep the query small in order to maintain a high degree of relevance in the data at the expense of not collecting a small number of relevant tweets. It is an interesting feature of the "pro-ana" phenomena on Twitter that the choice was quite clear, and that little compromise was needed. It would be of interest to investigate other potential Twitter communities to see if their boundaries can be so clearly delineated.

- In Section 3.3 I mentioned the systematic temporal sampling error inherent in the data collection process. A valuable addition to research into social media as a metric for social processes would be the study of this and related sampling errors (e.g.: self selection bias).

- For a longitudinal study of an online community, it may be appropriate to

revise the list of hash tags found to identify the community on a regular basis (say, each month or quarter), as tag usage may change over time, with new tags adopted by the community and old tags losing favour or being co-opted by other communities.

- As noted in Section 3.2, it is possible that there are other communities operating on Twitter that are not identified by this methodology. It would be of interest to attempt to estimate the extent and number of such communities. This could be done by surveying people known to have or have had eating disorders and by a more substantial (subjective) survey of random tweets, tweets with tags more distantly related to eating disorders and tweets from Tiwtter users seen to be experiencing eating disorders.

Below are a number of opportunities to strategically increase the amount of data collected. The main aim for most of these is to increase the resolution of dynamic information and to gather extra data peripheral to the community under consideration. One would want to carefully assess inherent biases in such strategies, noting that the lack of information does not necessarily reflect lack of activity.

Though collection of friend and follower lists is the main bottleneck for data collection (see Section 3.4), the friend and follower list API endpoints are often not fully utilised when collecting data for the pro-ana community, with data rates substantially lower than API rate limits. Following are several possible strategies for effective use of the friend/follower polling capacity.

- Recent research into the follow behaviour of Twitter users indicates that retweets often stimulate a cascade of follows and unfollows [Myers and Leskovec 2014], particularly when there are a large number of retweets. Thus a strategy could be devised to re-poll users who are exposed to retweets (i.e.: those who follow those who retweet and the retweeters themselves) for a period of time after the retweet event.

- Another approach could be to gather extra data from users of particular interest. Such users could be apparent members of a particular social group, perhaps through analysis of the friend/follower, re-tweet and/or user mention networks. The system could regularly re-poll friend/follower data and user profiles for interesting users in periods of spare API bandwidth and/or at the expense of timely data collection from less interesting users.

- A simpler, naive approach could be to keep all users in the priority queue for polling friend/follower lists. The current priority strategy could be used, with lower priority given to users that would otherwise not have been in the queue (those who have been polled since their last tweet). A priority strategy between those extra users would have to be devised, probably incorporating an increasing interval between polls as the period since a users last activity increases.

The *search/tweets* REST API is also not fully utilised by the pro-ana query, and substantially more tweets could at times be collected.

- Tracking tweets by all individuals in the data would quickly become intractable, however it could be valuable to collect more or all tweets by identified interesting users.

- A simpler strategy of collecting all tweets from users for a certain time since their last tweet could also be valuable.

Since the creation of this software, Twitter has introduced several new API endpoints that could be integrated into the data collection strategies.

- Of particular interest is the *friendships/show* Twitter API endpoint, which returns information about the relationship between two Twitter users and has a high rate limit of 180 calls per 15 minutes. Given a technique to regularly identify users of particular interest, their user relationships could be polled more frequently.

- Another Twitter feature that may be of interest is "lists". Users can create and join lists, and use their list membership(s) to filter the tweets that appear in their Twitter feeds or manually view tweets from list members. It is reasonable to assume that community members may use lists to more easily identify community members. List membership of users in the pro-ana data follows a typical power law related distribution, with about 40% of users members of some list (Figure 3.4).

3.7 Conclusions

Capturing data on dynamic aspects of social media communities is important for the study of online social behaviour. Systems designed to capture data from

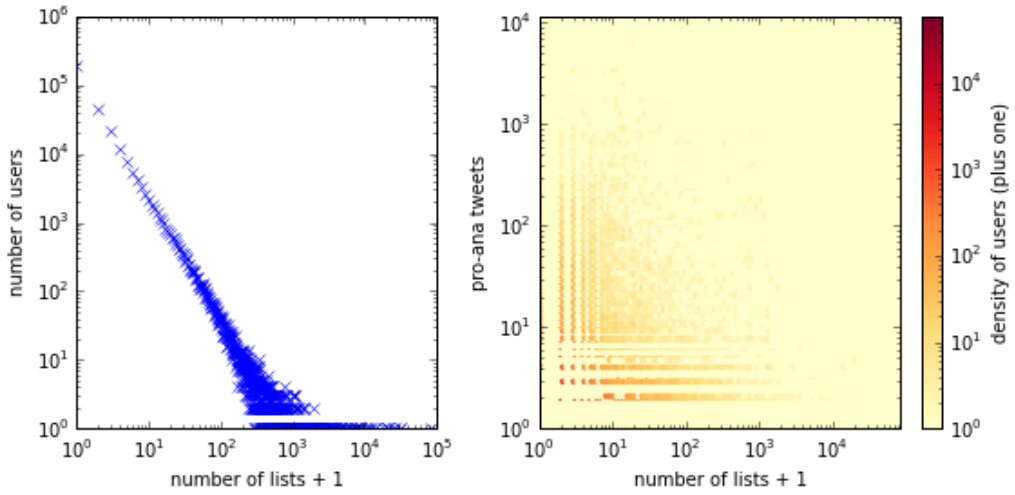

Figure 3.4: Number of users and number of pro-ana tweets vs. list memberships. 1 has been added to the number of lists so that 0 lists appears as 10^0.

Twitter and other social media typically lack the ability to capture important dynamics, such as changes in the social network. Further, approaches to collect data from targeted communities have been ad-hoc and typically rely only on the authors intuitions.

I have developed a methodology similar to adaptive sampling for identifying social media communication from a targeted community and constructed a system that captures network and user profile dynamics from Twitter communities thus identified. The system is very robust to the bursty nature of tweet streams, network problems and other difficulties associated with online data collection.

I identified a set of Twitter hash tags that is used almost exclusively by the Twitter "pro-ana" and eating disorder community and subscribers to the "thin ideal", and used the system to collect a nearly unbroken record of tweets, user and network data from that community covering 2 years and 10 months: over 1.2 million tweets, 300 thousand users and 200 thousand images.

In the next chapter, I present an overview and some preliminary analyses of the collected data.

Chapter 4

Preliminary Data Analyses

This chapter focusses on the collected pro-ana Twitter data, aiming to provide an overview of the data and report a few preliminary investigations. It asks more questions than it answers and is intended to give an initial indication of the richness of the collected data and present a few potential directions for further investigation.

This chapter is organised as follows: Section 4.1 provides some general statistics and observations of the collected pro-ana Twitter data. Section 4.2 presents some challenges and approaches to working with the uncertainties in the dynamic network information. Section 4.3 presents a simple approach to estimating the expected lifetime of a link in the data, attempting to account for the uncertainties. Section 4.4 discusses the creation of network snapshots, utilising the estimated link lifetime as a binary cutoff. Finally, Section 4.5 presents a preliminary grounded analysis of the data by clinical psychologists specialising in eating disorders as well as a more subjective appraisal of themes likely to be present in the data based on clinical experience.

4.1 Overview

As of 30 January 2015, the data contained 1,283,875 tweets, 296,483 users and 307,723 unique image ids (though many are in fact duplicates). There were 1,616,199 friend and follower list snapshots and 1,655,280 user profile changes. Like many social phenomena, hash tag usage presents a straight line in a log-log plot, suggesting a possible power law relation, as did the number of followers and friends (Figure 4.2) though follower and friend counts were only weakly correlated (Pearson's r 0.12). The sharp cut-off in the friends vs. followers plot and the

hiatus in the friends plot are due to limits imposed by Twitter: you are allowed to follow up to 2000 Twitter accounts, after which the ratio of the number of friends to the number followers must not cross an unpublished boundary. An unofficial source suggested that the ratio is 110%, which Figure 4.2 attests to. Plots of the number of tweets per user, both in our data and overall are also suggestive of a power law (Figure 4.3).

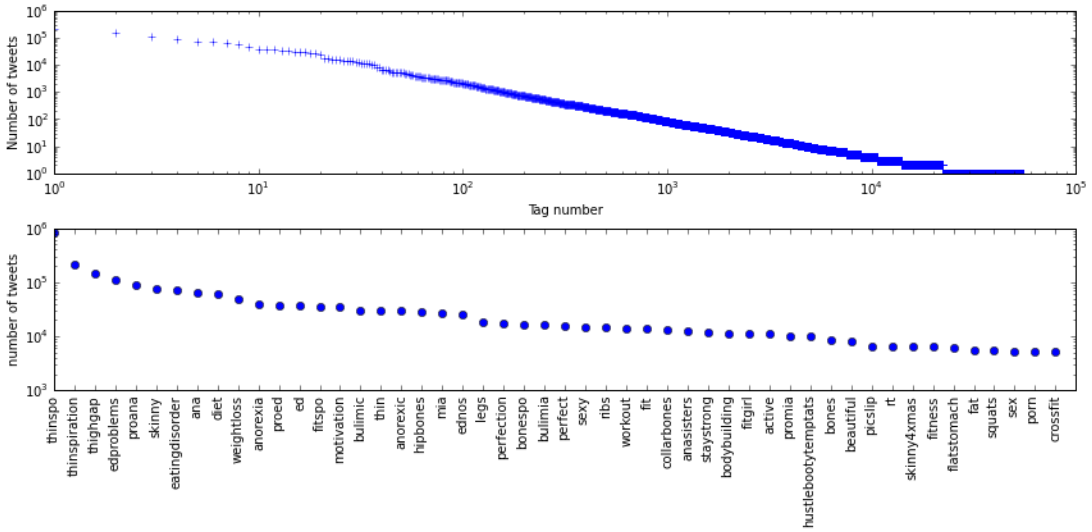

Figure 4.1: Tag frequencies (converted to lower case, one mark per tag).

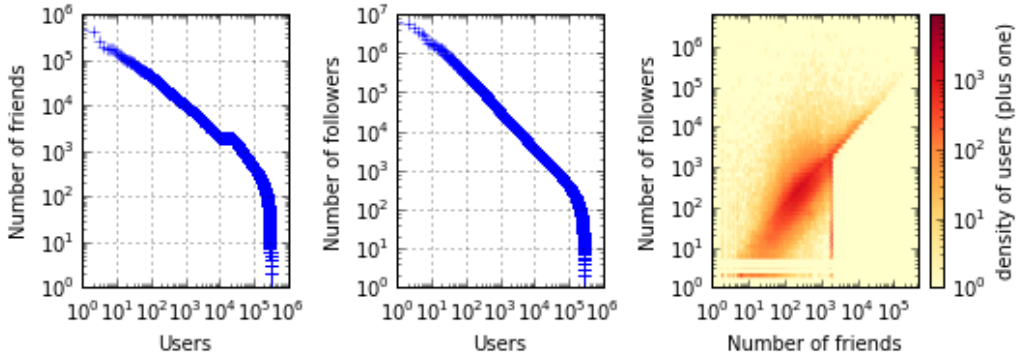

Figure 4.2: Number of followers and friends of users (one mark per user). Note: Irregularities in the left and right plots are due to limitations placed by Twitter on the number of people a user can follow.

Hash tags related to "thinspiration" (e.g.: *#thinspiration*, *#thinspo*, *#fitspo* etc...) dominate the data, with 73% of tweets[1]. Retweets and images also

[1]Such tweets typically contain images of people, mostly thin women.

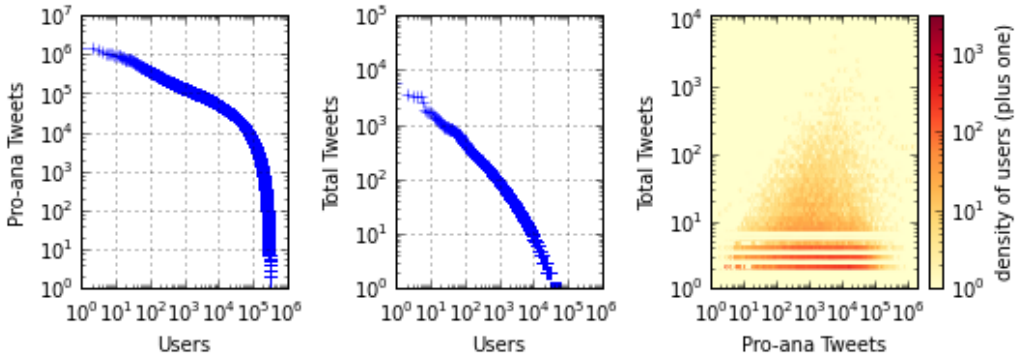

Figure 4.3: Number of tweets per user: in search query, overall and the correlation between the two.

account for a significant portion, with 57% of collected tweets retweets and 71% containing images. Thinspiration tweets account for 89% of the images, 80% of the retweets contain images and 76% of retweets contain thinspiration tags.

4.2 Working With Partial Network Data

During data collection, exact time stamps for follow link creation and deletion were not possible to obtain — the Twitter API simply does not provide that information. As a result the link data consists of snapshots of friend and follower lists taken at irregular intervals depending on when the linked users tweeted on the search query. Full snapshots were not stored as it would require substantial redundant storage of information. Instead only the friends and followers that changed since the last snapshot for the polled user were stored. Note also that these snapshots are partial with respect to the whole network: only links to or from the polled user are recorded.

From this data, we can identify lower and upper time bounds for friending events (link creation) by the link's absence in one snapshot and presence in the following snapshot, and similarly unfriending events (link destruction) by the link's presence in one snapshot and absence in the following snapshot. In practice the situation is slightly more complex: we must consider recorded link changes for both users in the link, and we must also consider polls of each of the user's friends and followers lists where no change was recorded. To see this second point, imagine the following sequence of events: user A was not friend of B at some point in time and that fact is noted in the records of both A's and B's

follower list changes. Next, imagine user A friends[2] user B, then user A tweets (so we have a record of the friending in A's history of changes). Next, user A unfriends user B and finally user B tweets — B does not record a change to her follower list since A was not a friend of B the last time B's followers were recorded. If we look only at recorded changes, we see A and B are not friends (the initial state), then A becomes a friend of B (A's record). Given that we have no further tweets from A or B, we only see that A subsequently unfriended B if we also consider that B tweeted and recorded no change in her followers list (hence at that point A and B were not friends).

In Figure 4.4 (left) we see that the number of observed changes to links is suggestive of a power law distribution. However in Figure 4.4 (right) we see that the number of changes per observation[3] does not follow an apparent power law, suggesting that the power law like relation in the left plot may be largely due to the such a relation in the number of link observations per link (due to that of the number of tweets — Figure 4.3). We can also see a significant number of links with the number of changes per observation close to one. This suggests that there are likely many link changes that were not observed, though the relation is complex, since observations happen just after a user tweets, which is clearly not independent of when users make changes to their follower lists and when others follow or unfollow them. In any case, the numbers in Figure 4.4 are underestimates — such highly repeated friending and unfriending has not, to my knowledge, been observed in the Twitter research literature.

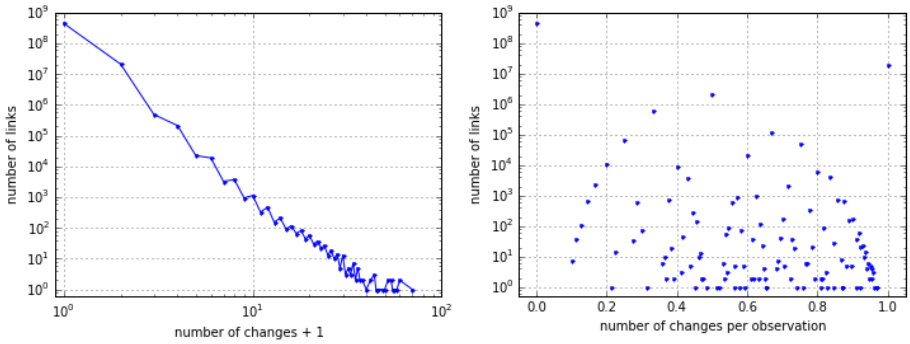

Figure 4.4: Links with multiple recorded creation/destruction events in the period December 2012 to July 2015.

[2] "Friending" refers to voluntarily choosing to follow the tweets of another user. If A "friends" B, subsequently B appears in A's friends list and A appears in B's followers list.

[3] here an "observation" is an opportunity to observe a change — that is, when the friends/followers lists of one end of the link are polled and a previous record of the link exists.

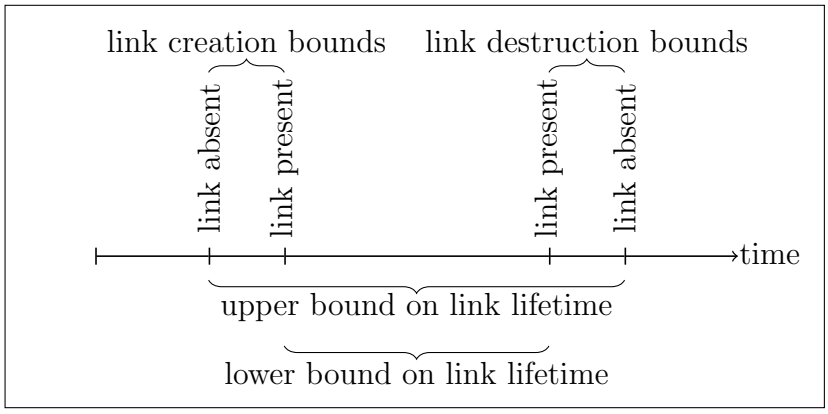

Figure 4.5: Illustration of link event and lifetime bounds.

4.3 Follower Link Life Cycles

Understanding the uncertainty in presence or absence of links given our observations will be important when attempting to analyse the dynamics of the follower network. In a first analysis of full follower network snapshots over the period of observation, it became apparent that many links had remained unobserved for a long period. It is reasonable to assume that some of those links may have been subsequently broken. A strategy to account for links for which we have no fresh observations in a principled way was needed. To enable such a strategy, an analysis of observed link life times and link dissolution patterns was undertaken. Such an analysis is also of interest in its own right and to my knowledge has not previously been attempted.[4]

One possible strategy to account for stale links (those which have not been observed for some time) is to introduce link weights related to the estimated probability that they have been since removed. However, some network analysis techniques, including those performed in Section 6, do not admit link weights. In this case, a binary approximation (presence or absence) of link survival must be used. A reasonable cut-off would be the median of the inferred link lifetime distribution: links unobserved for longer than the median would be discarded.

As a first step, links for which we do not have upper and lower bounds for both a creation event and a destruction event were discarded. This precludes users who have 3 or less tweets recorded in the data, which is desirable since such users are not significant contributors to any social groups present. Of greater

[4] [Myers and Leskovec 2014] looked at bursts of link creation and destruction related to retweet events but did not present a survival analysis and [Xu et al. 2013] examined factors related to unfriending, however had only 4 whole network snapshots to work with.

concern are stable links extant for the whole period of observation or links for which one of the bounds on its creation or destruction events would have been just outside the observation period. Removal of these links introduces a bias that we can control for, however: It is less likely that both creation and destruction events are observed for longer lasting links than those that are broken quickly — the effective time window in which we can observe both ends of a links lifespan is the total observation time *minus* the links lifespan. To be more precise, a link can be included only if the initial and final records of a links absence (with presence recorded in between) lie within the observation period. To correct for this bias, the number of links of a given age should be scaled by $\frac{w}{w-l}$ were w is the length of the observation window (the total amount of time over which data was collected) and l is the upper bound on the links lifetime (the time between the lower bound on the links creation and the upper bound on its destruction).

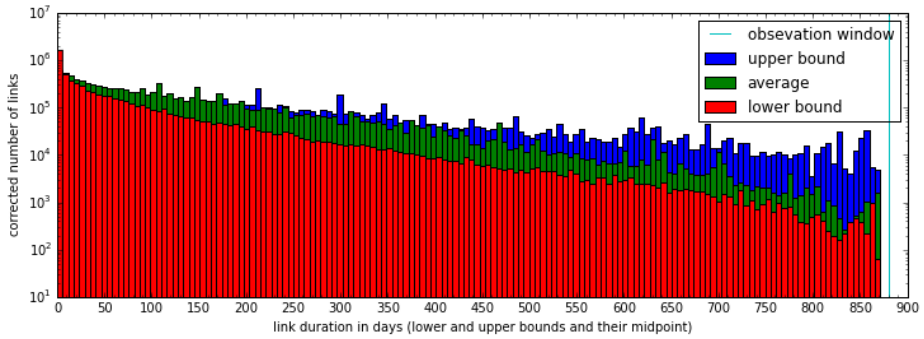

Figure 4.6: Corrected link lifetimes.

A histogram of the resulting link lifetimes are presented as a histogram in Figure 4.6. Observing the near linearity of the corrected histogram heights with log-scale y-axis, we can expect a reasonable fit from an exponential. This is typical of lifetime data with constant hazard rate (the probability a link will be broken at any given point in time). A least squares estimation in $\log_{10}$ of the number of links versus midpoints between upper and lower bounds on link duration can be seen in Figure 4.7. The exponent (slope of Figure 4.7), intercept and Pearsons-r correlation coefficient are provided in Table 4.1.

The high level of correlation indicates a very good fit, suggesting that link lifetimes in this data exhibit a constant dissolution rate — the probability of a given link dissolving in a given time period is very close to constant. This is an interesting observation, as one might expect time-relative effects such as younger links being more prevalent in a user's mind and so more dynamic, appearing and

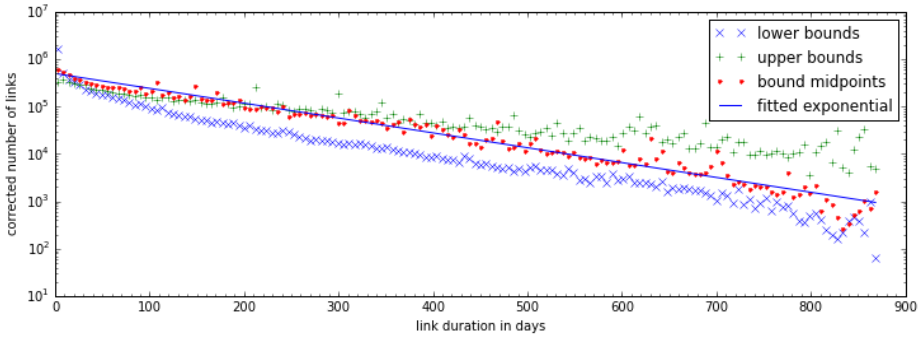

Figure 4.7: linear fit in log space to link lifetimes

exponent	-0.00313
y-intercept	5.696
correlation	-0.980

Table 4.1: Linear regression and correlation coefficients with units in days and log_{10} of the number of links

disappearing with greater frequency, where a link that is older may be forgotten and left untouched.

The estimated exponent in Table 4.1 equates to a constant attrition probability of 0.0072 per day. As a binary approximation (presence or absence) of link survival, I employed the median of the inferred exponential[5], shown in Table 4.2, as a cut-off — links that had not been observed for more than this value were taken to no longer exist.

Median link age
96.11 days

Table 4.2: Cutoff for binary approximation of link survival.

The above analysis is essentially a survival function estimation, as is often done in medical trials, and there are more sophisticated statistical methods [Miller et al. 1981; Radke 2003] that are able to utilise extra information from censored data[6] that could have been applied. Though such an analysis would be of interest in its own right, for the purposes of approximate correction of long-unobserved links however, it was considered unnecessary since a good estimate could be achieved

[5]Note that the median is the inverse of the attribution probability multiplied by $ln(2)$.

[6]In this situation, we have *interval* censored data, where we only have bounds on link lifetime, and possibly only one bound.

with a more prosaic approach presented here.

In estimating expected link lifetimes, I considered removing links for which we do not have a reasonable level of certainty in addition to requiring upper and lower bounds on its lifetime. That is, links for which the difference between the upper and lower bounds on the link's lifetime (the uncertainty in the links duration) is large. Figure 4.8 shows the number of links with a given duration uncertainty in days (upper plot) and the ratio between As can be seen in Figure 4.8, the lifespan of a significant number of the sampled links is highly uncertain. As a simple heuristic, I investigated links with uncertainty less than 10% of the observation window. This was chosen in preference to a relative uncertainty cut-off, as we wish to investigate dynamics at the scale of the observation window, not relative to link lifetimes. As can be seen in Figure 4.9, however, this approach introduced a complex bias, retaining more links with lifetimes close to zero or the observation window than with central values, hence it was abandoned.

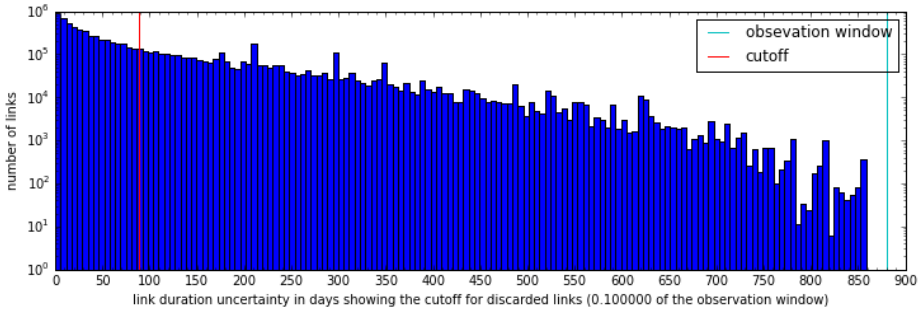

Figure 4.8: Histogram of the uncertainty in link duration.

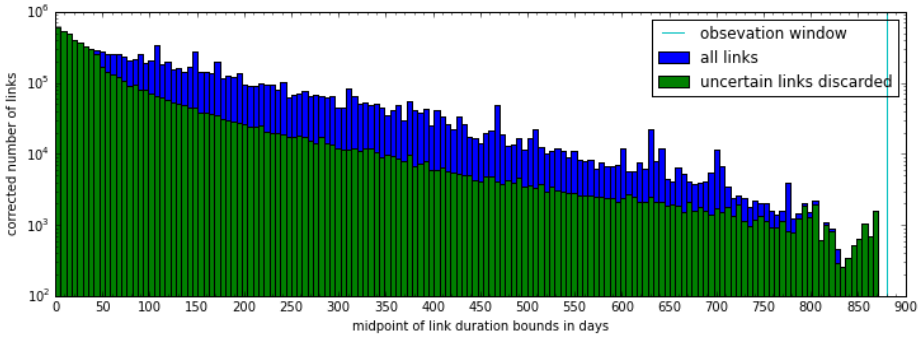

Figure 4.9: Corrected link lifetimes — all links vs. subset without discarded links. Data points are the midpoint of upper and lower link lifetime bounds.

4.4 Network Snapshots

In this section, I discuss the extraction of follower network snapshots and present an overview of the snapshots thus extracted. An anomaly due in part to the data collection strategy, where extraneous events can cause the sudden influx of users not related to the communities of interest in the data, is discussed along with its wider implications for analysis of the data. The primary aim is to prepare data for the application of community detection algorithms. Assuming the communities of interest in the data form detectable communities in the follower network, this could be used as a filter to focus only on those communities, excluding extraneous users.

One avenue for analysis of network dynamics is to make comparisons between snapshots of the network at regular time intervals. This approach renders each snapshot amenable to existing community detection algorithms and other analysis techniques for static networks. Dynamics of inferred properties (e.g.: communities) can then be studied by comparing results between snapshots. In some cases, inferred properties for one snapshot can be used as a starting point or prior for analysis of the next. Where possible, incorporating link creation and destruction times in the analysis algorithm would be preferable, but to my knowledge at this time such methods for dynamic community analysis do not exist.

Only mutual links were included when extracting snapshots for two reasons. First, a mutual link is more likely to represent a social connection than a link that is not reciprocated. Non-reciprocated links often represent a "fan club" or information source. Ignoring them also reduces the prominence of highly popular users, though not entirely as some popular users follow many of their followers. The second reason is simply pragmatic and two-fold: the full network is simply too large for community detection algorithms to handle in reasonable time and many require undirected networks as inputs (in particular, those used in Chapter 6).

An initial naive approach to extracting weekly network snapshots resulted in a roughly linear network size over time. This was unexpected as overall tweet activity had remained relatively constant. Investigation confirmed that there were many links that had not been polled for a long time, some since the beginning of data collection, two and a half years before. Separating "stale" links, older than the empirical estimate of median link lifetime (Table 4.2), from "fresh links", observed more recently than the empirical median lifetime, produced a more favourable picture where "fresh" network size declines slightly over time after an initial ramp up (Figure 4.10).

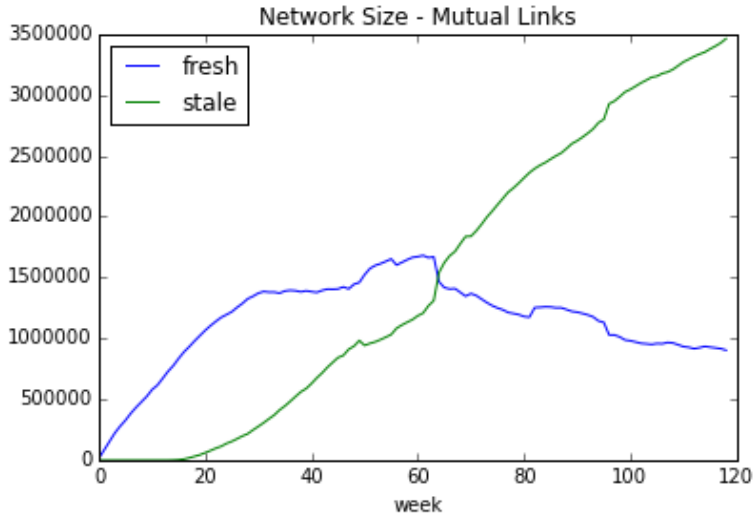

Figure 4.10: Network sizes for weekly snapshots showing fresh, recently observed links and stale links last observed more than 96.1 days before.

The small peak around 50 weeks corresponds to a particular event recorded in the data: the American ABC news ran a Twitter "chat" (online discussion) on eating disorders, using one of the search tags (#thinspo) in a tweet to introduce the discussion.

> ABC tweet: *"#EatingDisorder Twitter chat begins in 15 min w/ @drrichardbesser + experts. Will cover #anorexia #bulimia #thinspo. Use #abcDRBchat."*

Though the ABC had over 2 million followers at the time they tweeted, this substantial list was not recorded as it was too large for the database to fit into a single user object. Only 16 tweets in the data set contain the tag *#abcDRBchat* (indicating direct involvement in the chat) and the ABC only tweeted once, however there was a surge of activity on the data collection search tags, and that surge is what can be seen in Figure 4.10. Notice that the width of the peak corresponds to the cut-off time (96.11 days) for a link to be discarded if neither end has authored a tweet containing a search tag. It can be seen that about 300 thousand new follow relations appeared over a relatively short time and that the users that contributed those links promptly stopped using the searched hash tags. Another similar, but weaker, event can be seen around 82 weeks.

This event nicely highlights the need to be cautious when drawing conclusions from data collected in this way. The "fringes" of the data are very noisy and only partially observed. One of the aims of the community detection work in Chapter 6

is to filter out such fringe data, retaining core users that represent coherent social groups and their interactions.

Another thing to notice in this event is the drop in the number of "stale" links (those last observed more than 96.1 days before the snapshot). This would seem to indicate long lasting links to/from users who tweet infrequently on the search query tags, and who tweeted in response to the event.This highlights a deficiency in the binary cutoff for temporarily unobserved links, which apparently removed these links prematurely. Possible improvements could be made through a more sophisticated link survival analysis incorporating extra user, tweet and link metadata which may be able to identify such links and allow them to persist.

The gradual decline in snapshot size (the "fresh" networks in Fitgure 4.10) should not be taken as a decline in the popularity of eating disorder topics in online social media and in the community at large. Equally (if not more) likely is that the venue for eating disorder discussion has moved to another social media portal.

It is interesting that snapshot sizes, and the number of "stale" links for the full (directed) network are almost exactly proportional to corresponding mutual link statistics, albeit with slightly different proportions (Figure 4.11). The mean and standard deviation of per week proportions are given in Table 4.3.

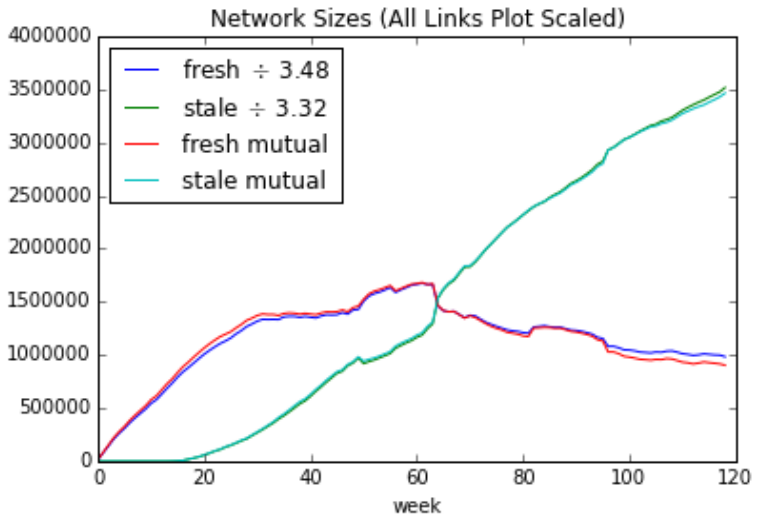

Figure 4.11: Mutual network sizes and scaled full network sizes evolve almost identically.

These proportions relate to "followback" rates — when a user responds to someone following them by following the other user in return. There is an eti-

	Mean Ratio	SD
fresh links	3.48	0.16
stale links	3.32	0.07

Table 4.3: Mean and standard deviation of per week ratios of all vs. mutual links for fresh (recently observed), stale (not observed recently) links. Note that the difference is not significant.

quette surrounding followback, as shown by many tweets complaining of users who do not followback, however the etiquette is not universally adopted. Many users are quite selective of who they follow, often not following back. This is perhaps especially the case for very popular users. The constant ratio is suggestive of robust behavioural statistics surrounding followback etiquette, however it would be well to control for confounding variables such as user popularity (number of friends/followers), network statistics such as betweenness, and other socially relevant variables before drawing firm conclusions or spending significant effort orchestrating further controlled experiments.

4.5 Grounded Analysis of Tweets

In order to form some interpretations of the possible meaning of interesting topics in Chapter 5, Section 5.7.2, I sought the help of clinical psychologists experienced with eating disorder patients and engaging in eating disorder research.

In collaboration with Dr Elizabeth Rieger and Dr Kristen Murray, both clinical psychologists specialising in eating disorders, a preliminary analysis of themes in the data relevant to eating disorder psychology was undertaken. First, the data was summarised using topic models and an analysis akin to grounded theory performed. Second, a random selection of several hundred images from the data were assessed for thematic content. These observations were compiled and several extra themes were identified, drawn from recollections of the data and clinical experience.

In the first phase, topic models with 20 and 50 topics were used to summarise the data. The data was pre-processed in a similar way to Section 5.6 — retweets were removed, named entities identified, constructs such as smileys (e.g.: "(-:"), hash tags and URL's were kept intact and remaining punctuation characters considered as individual tokens (see Section 5.6 for more details). Topic models were inferred using standard LDA ($\alpha = 0.05N/DT$, $\beta = 0.01$, N words and D

documents in corpus, T topics). These values of α allocate 5% of the probability mass for smoothing, and β encourages few topics per tweet.

Each topic was presented as a one page list of the most probable word tokens from the topic, with the physical size (font size) of the words proportional to their topic probability. Topic order was randomised for each collaborator. A one page list of tweets was presented along side the top words, sampled according to the tweets topic probability in the model, with words assigned to the topic highlighted. Example tweets were independently drawn for each collaborator. Appendix D shows a copy of the 20-topic model as shown to a collaborator.

Collaborators were then asked to identify themes in the example tweets and topic top words and results collated and compared in a group discussion. Themes identified in the sampled images were added, and a final summary of the identified themes was then collated and ratified with collaborators. The resulting themes are presented in Appendix C.

In general, collaborators found the data rich with themes familiar from their work. About half of the topic model topics were found to express very clear meaning, and about another third were assessed as clear, but somewhat confused. All sampled images were seen to relate strongly to psychological themes around anorexia, and in particular the "thin ideal".

4.6 Discussion and Future Work

Richly dynamic longitudinal data such as this presents many opportunities for future analysis. Research into social network dynamics, research into the pro-anorexia online movement and how it relates to stereotypes projected in traditional media and research into the evolution and propagation of memes would all find utility in this data. In Section 7.2 I describe these possibilities in more detail.

Apart from the general utility of the data, this chapter identified a number of specific observations worthy of further investigation. An apparently constant ratio between the number of reciprocated and unreciprocated links was observed, suggesting statistically robust "followback" rates. It would be interesting to further examine this relation in the data, as dicussed towards the end of Section 4.4.

The plots of the number of friends and followers (Figure 4.2) both show a change in slope around 100 to 200 friends/followers. This suggests that there may be a change in the processes that lead to the relationship. It is interest-

ing that the region in which the slope changes roughly corresponds to Dunbar's number[7] which has been found to hold in Twitter @reply networks [Gonçalves et al. 2011]. We could conjecture, for example, that in the latter part of the plot with less connections, links are formed primarily between true social connections, whereas in the earlier part, with more friends/followers, link formation is driven by active seeking of connections beyond a person's social ties. Interestingly, several comprehensive studies of the Twitter follower network did not reveal such a distinctive change in the slope [Kwak et al. 2010; Krishnamurthy et al. 2008], perhaps indicating that the pro-ana data is more social in nature than is typical on Twitter.

Another aspect of the data that deserves mention and further attention is the substantial number of collected images, which have been largely left out of analyses presented in this thesis. Most modern social media platforms provide opportunities for sharing images and video which are heavily utilised by social media users, and analysis of those data streams is an active area of research.

During the development of this thesis, consideration was given to image analysis, and meaningful extraction of image features was considered feasible with relatively minor adaptations to existing methodologies. These features could then be utilised as "tokens" in text analysis or extra metadata for other analysis methods.

Perhaps the simplest attainable feature is de-duplication — many images in the data are near-identical duplicates. Another whole-image feature that is relatively simple (though potentially expensive) to attain is identification via services such as TinEye[8], which can reveal image sources and useful metadata such as the names of any celebrities depicted in the image. Many celebrities in the image data are revered as examples of the "thin ideal".

Beyond image level features, a number of within-image features were considered detectable with reasonable or at least useful accuracy. Off-the-shelf tools exist for face and body part detection, and many features particular to the data, such as thigh gaps, collar bones and general skinniness are likely detectable using standard machine learning and image analysis techniques alongside a few hundred hand-labelled images.

The presence of spam in social media data is often a concern for their analyses. Examination by myself of data samples and topic models, and during the grounded analysis discussed in Section 4.5 all revealed apparently low levels of

[7] Dunbar's number is often cited as 150, but the research indicates between 100 and 200.
[8] http://tineye.com

spam in this data, though characterisation of what "spam" is in this context may not be clear. Due to the apparently low levels and the ability of topic models to effectively filter spam, which tend to be drawn into topics of their own due to repeated patterns in their text, this has not been investigated further here, however an analysis of potential spam could be in order in the context of deeper analyses of the data.

4.7 Conclusions

This chapter presented some preliminary analysis and overall statistics of the collected pro-anorexia and eating disorder Twitter data and developed principled approaches to extract and utilise the captured network and user profile dynamics that are used in subsequent chapters and will be useful for the many further analysis opportunities the data entails.

A preliminary subjective analysis of the data is presented, confirming the relevance of the data set to eating disorder research and encouraging deeper and more sophisticated analyses intended to understand better the role of social media in the lives of people with eating disorders and in the propagation and cure of the malady.

Apart from establishing the relevance of the collected data and providing a bridge between data collection and analysis, this chapter presented several interesting observations of apparently robust behavioural statistics that are worthy of further investigation.

The next chapter explores an approach to combine topic models with more traditional, intervention based psychological research in order to measure social psychological phenomena and processes in large socially generated data sets.

Chapter 5

Topic Models as a Quantitative Tool

This chapter presents an approach to measuring social psychological phenomena in large socially generated data sets by combining topic models with more traditional, intervention based psychological research. The content of this chapter has been published in "Social Computing, Behavioral-Cultural Modeling, and Prediction 2015" [Wood 2015b], extended here with a deeper analysis of salient topics.

Despite a growing body of research into computational models of social psychological processes, direct empirical grounding for these models remains an elusive goal. This is largely due to the difficulty of measuring modelled characteristics of social groups. This chapter presents a methodology combining topic models with traditional psycho-linguistic research as a first step towards addressing that difficulty. The method is applied to the data described in Chapter 4, a collection of over a million tweets from the Twitter 'pro-anorexia' community, in combination with a recent study of gender salience in women.

This chapter is organised as follows: Section 5.1 presents an overview of the methodology and its motivation. Section 5.2 presents posterior predictive checks and their application in connection with the methodology. Section 5.3 describes the use of word frequencies as a tool for measuring meaningful characteristics of text and goes on to explain how topic models can be used to contextualise those meanings. Section 5.4 reviews some recent work associating identity salience with certain LIWC word classes. Section 5.5 presents an adaptation to topic model regularisation intended to improve model relevance. Section 5.6 presents the steps taken to prepare the pro-ana tweet data for topic modelling. Section 5.7 presents

results of the application of the methodology to the pro-ana data. Section 5.9 reviews the methodology and results and presents some opportunities for further work.

5.1 Overview

Traditional techniques for measuring social psychological phenomena require controlled interventions and/or intensive expert annotations, restricting the number of individuals that can be assessed, and with the added difficulty that any interventions may perturb the very processes under study.

In cases where a substantial portion of group interaction can be captured as text, notably communities that operate over online social media, analysis of those texts and associated metadata promises an avenue for direct and unobtrusive observation of relevant traits of individuals and their communications.

Recent advances in data mining have produced new methods for extracting information from large collections of text data. However applying those techniques to reveal social psychological features, and in particular as an empirical grounding for computational models, remains largely unexplored. The methodology presented here attempts to make a modest start to this endeavour by combining topic models with word frequency based psycholinguistic methods where the combined frequency of lists of words is linked to characteristics of interest. These lists could be generated by expert analysis of document samples or chosen from word frequency based tools such as LIWC [Tausczik and Pennebaker 2010b] (Linguistic Enquiry with Word Count) combined with research that links LIWC word classes with characteristics of interest. In the example presented in Section 5.7, I use LIWC word categories thought to be linked to the salience of personal or gender identity [Dann 2011].

Topic models are then applied to identify patterns of word co-occurrence (topics) across a collection of group interactions recorded as text. The discovered topics contextualise the frequency based tools, identifying other words that occur alongside those indicated by the tool. Posterior predictive checks [Mimno and Blei 2011] are used to ensure the identified topics accurately represent true structures in the data.

A novel modification to a topic regularisation technique [Newman et al. 2011] with the aim to focus topics on words from the lists was also applied. This technique allows to provide a prior on word associations within topics. Given

word lists of interest, the hypothesis was that providing a prior in which word pairs from those lists are more likely to appear together in topics would result in topics better focussed on the characteristics of interest. This was not borne out, however, as giving the prior more than trivial strength resulted in inferred models that were not faithful to the data (as measured by posterior predictive checks). The trivial prior strength models were not notably different to equivalent standard LDA models.

5.2 Posterior Predictive Checks

Though topic models are a powerful tool, they provide no guarantees on the accuracy of the inferred model. Posterior predictive checks [Mimno and Blei 2011] are a mechanism for testing Bayesian models. A discriminant function of input data is chosen to capture some quantity of interest. The inferred model is then used to generate many artificial data sets, and the values of the discriminant function on these data sets provide an estimate of probable function values. If the value of the discriminant function applied to the real input data is improbable, the generative model and/or prior has failed to capture relevant structure in the data. Mimno et.al [Mimno and Blei 2011] proposed the mutual information between word allocations to a topic and the allocations of those words to documents (Equation 5.1) as a discriminant function.

$$MI(W, D|k) = \sum_{w,d} P(w, d|k) \log \frac{P(w, d|k)}{P(w|k)P(d|k)} \tag{5.1}$$

The intuition is that the probability of a word given its topic allocation should be independent of the document in which it falls. If this is the case, the stated mutual information will be zero. Due to sampling errors this can only be expected for corpora of near infinite size. In practical settings some small positive value will result from random fluctuations.

Important here is the interpretation of failure of this discriminant function. Topic-document mutual information that is higher than expected indicates that words assigned to a particular topic are not evenly distributed among documents. Conversely, values within the expected range indicate that words are as evenly distributed as can be expected, and so topic probabilities can provide a reasonable estimate for word frequencies among words assigned to the topic. If we restrict the discriminant function in turn to words from each word list of interest, topics

with acceptable values make reasonable proxies for the frequencies of words in that list.

5.3 Word Frequencies as Metrics

Word frequencies are often used as proxies for underlying meanings and themes in document collections. Indeed, topic models can be seen as a more sophisticated word frequency based approach and the widely used "bag of words" model concerns itself only with word frequencies, discarding other structural and syntactic information in text. Simpler approaches such as LIWC [Tausczik and Pennebaker 2010b] utilise lists of words, the frequencies of which, when summed together, have been found to correlate with particular themes and meanings. The methodology proposed here assesses topic models by their ability to resolve words from word lists deemed to be relevant to the enquiry at hand, that is, to produce topics with high probabilities for those words (typically a small number of such topics).

It should be noted that word frequency based tools, and in particular LIWC, largely rely on relatively long documents to obtain statistical significance and their applicability to short texts such as tweets is questionable. Topic model topics from tweet corpora, however, represent a relatively large number of words distributed over many tweets. If one can argue that words assigned to a particular topic are associated with some common psychological state, such as might be expected from discussions within a particular social group on a particular topic, then one can argue that the topics word frequencies have similar psycho-indicative characteristics to word frequencies in longer documents from single authors. This point will be discussed in more concrete terms later in Section 5.7.2.

When using topic models to assess questions of interest in a particular line of enquiry, there are typically experts at hand who are able to identify relevant themes in a sample of documents. I present two approaches to utilise this expert knowledge. First, I provide labels for sampled documents, from which word lists predictive of those labels can be estimated (for example via logistic regression), and second, identify specific words and their contexts that are indicative of relevant themes.

As well as expert derived word lists, there may be word correlation results from previous research of specific relevance as well as more generic tools, such as LIWC, that are relevant. Several such tools for psychology research are discussed in Section 2.1.2.

5.4 Identity Salience

Research in psychology has found that there are typically many facets to a person's sense of identity. The social and cognitive context in which a person acts determines which of these facets are active or 'salient' at any given time. In terms of group dynamics, the identity salience of group members during interaction is an important factor in the determination, propagation and reinforcement of the groups social identities. These, in turn, can have a significant impact on the decisions of and opinions formed within the group.

A recent study [Dann 2011][1] investigated identity salience in 142 young women, mostly undergraduates at the Australian National University. Respondents were first given a priming task designed to make either their gender or personal identity salient. Several self-report psycho-metric tasks and a writing task on dieting and weight loss were then performed. The psychometric measures confirmed that the priming task had succeeded. In-sample logistic regression on LIWC scores from the writing task was able to predict the prime condition in 73.9% of cases. This study has some relevance to the pro-ana Twitter data as the pro-ana and eating disorder community consists predominantly of young women, and diet and especially weight loss are significant topics of discussion.

5.5 Topic Model Regularisation

In an attempt to focus topics on words from the word lists of interest, a novel adaptation to a topic model regularisation technique [Newman et al. 2011] was used. The regularisation technique essentially replaces the usual Dirichlet prior for topic-word probabilities ϕ by a structured prior that favours known word associations given in a matrix $\mathbf{C}$.

$$P(\phi_t|\mathbf{C}) \propto \left(\phi_t^T \mathbf{C} \phi_t\right)^{\nu} \tag{5.2}$$

Optimising the log posterior with respect to ϕ and adapting the LDA Gibbs update results in the following update equations (see [Newman et al. 2011] for more details):

[1] On reanalysis of the data used in this study, I was not able to reproduce it's results and the author was not available to confirm the exact details of the procedures used, so it's validity must be considered questionable. Nonetheless, it serves as a valid example of how the methodology can be applied.

$$\phi_{w|t} \leftarrow \frac{1}{N_t + 2\nu}\left(N_{wt} + 2\nu\frac{\phi_{w|t}\sum_{i=1}^{W}\mathbf{C}_{iw}\phi_{i|t}}{\phi_t^T\mathbf{C}\phi_t}\right) \tag{5.3}$$

$$P(z_{id} = t|x_{id} = w, \mathbf{z}^{\neg id}, \phi_{w|t}) \propto \phi_{w|t}(N_{td}^{\neg id} + \alpha) \tag{5.4}$$

In the original approach, word associations are drawn from a large, relevant reference corpus. Here we have no reference corpus, and instead use words from the corpus under investigation and choose artificially strong associations between words from the same list, leaving other word pairs 'unassociated'. The word associations in [Newman et al. 2011] are given in a matrix of word dependencies represented by the pointwise mutual information (PMI) between words in a reference corpus.

$$\text{PMI}_{ref}(w_i, w_j) = \log\frac{P_{ref}(w_i, w_j)}{P_{ref}(w_i)P_{ref}(w_j)} \tag{5.5}$$

In our case, we calculate PMI values for word pairs in our corpus with the (false) assumption that words from the same list of interest always co-occur. In this case $P(w_i, w_j) = \min(P(w_i), P(w_j))$ and

$$\text{PMI}_{corp}(w_i, w_j) = -\log\left(\max(P_{corp}(w_i), P_{corp}(w_j))\right) \tag{5.6}$$

5.6 Data Preparation — Pro-Ana Tweets

Data preparation is an important step in topic modelling analysis, and can have profound effects on modelling outcomes (see Section 2.4.3).

For the purposes of this study, a "document" was taken to be a single tweet. Though some studies have found that agglomerating tweets by hash tag or author has a positive impact on document clustering and topic coherence (as measured by PMI) [Mehrotra et al. 2013; Hong and Davison 2010], the property we seek is different: to resolve a particular set of words in a uniform way. Standard LDA topic models were found to perform well though it would also be interesting to explore tweet pooling strategies.

Retweets and replies were removed as they arguably provide little extra thematic information and the repeated text can significantly skew topic models (see Section 2.4.3). Models with retweets included contained several "retweet topics" that reflected the contents of a small number of highly retweeted tweets, a

situation deemed undesirable.

Tweets were tokenised by standardising numerous text emoticon ('smiley') forms, isolating punctuation as individual word tokens (these are a LIWC variable of some interest) and converting mixed case words to lower case (all caps words, known in social media as SHOUTING, are indicative of emotional intensity and so were retained as is). Url's, #tags, @mentions and apostrophised words (e.g.: "didn't") were left unchanged. Url's and @mentions are often removed or standardised, however common url's are indicative of community concerns and frequent mentions indicative of community membership and conversations and were deemed relevant here. Apostrophised words are often separated (e.g.: "didn't" into "did" and "n't" or "not"), however LIWC does not do this separation, so they were left whole.

A list of named entities was generated with Stanford NER and verified by hand (there were a small number of spurious entities and several that referred to the same person) and multi-word entities were encoded as single tokens.

Further pre-processing included removal of word tokens appearing less than 5 times[2] and removal of tweets with less than 3 word tokens. Removing very infrequent words and very short documents is a common approach for topic modelling as they contribute little to models with a moderate number of topics (as studied here), in many cases simply add noise, and can increase computation time significantly. The thresholds were a conservative choice balancing model complexity against relevance of removed tokens and documents, erring on the side of data retention.

The resulting corpus consisted of 262736 documents and a vocabulary of 18713 distinct word tokens.

5.7 Experiments

The eight most significant LIWC variables in the logistic regression mentioned in Section 5.4 were chosen as a proxy for differentiating between personal and gender identity salience (see Appendix B). LIWC variables are calculated by the sum of word frequencies in natural language for the list words associated with each variable. Hash tags in tweets have particular roles typically (but not always) outside of word use in natural language. Even hash tags that are in the LIWC dictionary carry extra meaning and cannot be considered equivalent to words

[2]Word tokens from LIWC word classes under study were retained.

used naturally. For these reasons, when calculating LIWC scores for topics, hash tags and non-word tokens were removed and word frequencies were calculated relative to the remaining words only.

Twenty six models were estimated, thirteen with 50 topics and thirteen with 20 topics, five each using standard LDA ($\alpha = 0.05N/DT$, $\beta = 0.01$, N words and D documents in corpus, T topics) and eight each using regularised LDA (α as for standard LDA, $\nu = 0.01V/2$, V words in corpus vocabulary). These values of α allocate 5% of the probability mass for smoothing. The choice of ν reflects some equivalence to the choice of β in that the denominator of the update equation for ϕ is the same. This value encourages tweets to exhibit few topics.

All models performed acceptably in the posterior predictive checks, with topic-document mutual information falling within the span of 100 simulated corpora in all but 0.3% (50 topics) and 1.2% (20 topics) of topic/LIWC/model combinations. The ordinary LDA models performed somewhat worse, with an average 47.6 (out of 400) and 12.8 (out of 160) topic/LIWC combinations falling outside the middle 80% of simulated values (20 and 50 topic models respectively). The regularised model averaged 29.6 and 8 outside 80% of simulated values.

5.7.1 Model Assessment

To assess the difference between the models with respect to identity salience, we used the conditional entropy of the predicted probability of a salient personal identity in each tweet, given word-topic allocations.

$$H(\texttt{salience}|\texttt{model}) = -\sum_{\texttt{topic}\ t} P(t) \sum_{\texttt{salience}\ s} P(s|t) \log_2 P(s|t) \qquad (5.7)$$

This quantity measures the amount of extra information (in bits) needed to obtain the predicted probabilities, given word-topic allocations. The difference in encoded information about salience between the regularised and standard LDA models was not significant ($p = .37$ and $p = .21$ for 20 and 50 topics respectively in 2-tailed Kolmogorov-Smirnov tests). As expected, the 50 topic models encoded more information (median 0.73) than the 20 topic models (median 0.77, K-S test $p < 10^{-6}$). The total entropy of personal vs. gender salience is very close to one (> 0.99). Thus the topic assignments alone account for about a fifth of that information (50 topics) or a quarter (20 topics) of it.

Other choices of ν with both stronger and weaker regularisation were also attempted, however all models with lower entropy than standard LDA also had

unacceptable posterior predictive checks. There are several possible explanations as to why regularisation was found to be ineffective at this task. First, [Newman et al. 2011] did not perform posterior predictive checks — it may be that regularisation of this type will *always* perform poorly in such checks. Perhaps more likely, however, is that the applied prior on word relations is artificial and does not correspond to actual patterns in word usage. If this is the case, it may be possible to construct prior PMI relations that are more natural or perhaps a better idea is to abandon this approach.

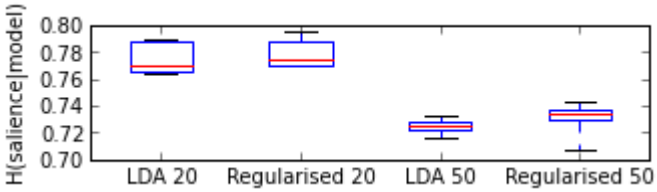

Figure 5.1: Conditional entropy of predicted salience given model. Whiskers = min/max.

Figure 5.2 indicates that in these models, about a third of the topics exhibit very high personal salience probability, much higher than the average 0.51 for documents in the corpus. Those topics represent a coherent context with high personal salience, demonstrating the utility of this approach. Again, the regularised models are more or less equivalent to the standard LDA models.

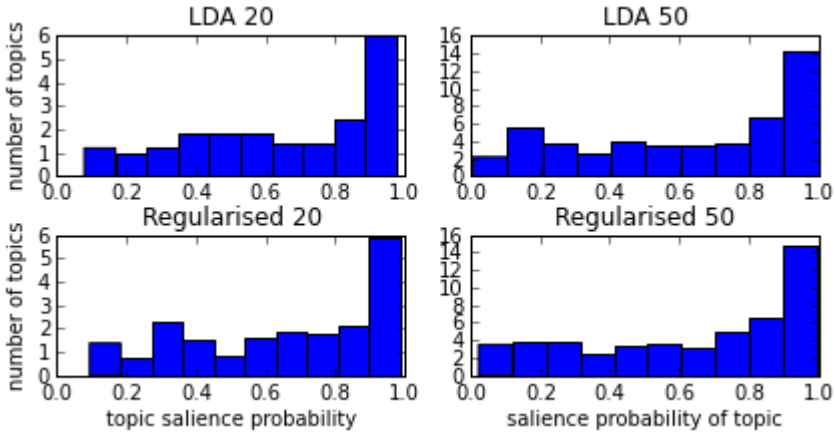

Figure 5.2: Number of topics with high personal salience averaged over 5 standard LDA models and 8 regularised LDA models for each of 20 and 50 topics.

5.7.2 Analysis of Salient Topics

This section presents a qualitative analysis of topics with high estimated probability of personal over gender salience from one of the 20-topic models and highlight some caveats on their interpretation. Though I draw some weak conclusions about the individual and social psychology of authors in the community, I leave stronger conclusions to further study.

First I would like to investigate the logistic model for differentiating gender from personal identity salience. Appendix B includes a table of the eight most significant logistic coefficients of the model. The coefficients of the remaining LIWC classes are very small and do not contribute significantly to the model. One could interpret personal identity salience as a kind of soft default in the model: the model has many LIWC classes with negative coefficients that indicate gender salience (suggesting the model is able to identify gender salient text) but few positive coefficients and a positive intercept value (suggesting the model identifies personal salience largely by the *absence* of gender salience indicative words). We will see that this default characteristic played a role in the apparent personal salience of some topics. The two positive coefficients do, however, allow for positive identification of personal salience, as we will later see with topic 4.

Figure 5.3 presents a visualisation of that model created with the termite topic model visualisation tool [Chuang et al. 2012]. Termite uses a balance between the significance of words in each topic, and their ability to distinguish between topics, to choose a set of words that can provide an overview of the semantics captured by the model. Words are ordered in an attempt to present groups of words from each topic as well as present common phrases where possible. The interactive version of Figure 5.3 as well as a visualisation of the model with non-words removed and an example 50-topic model can be found at `http://cs.anu.edu.au/~Ian.Wood/termite/20-run1-LIWCvocab/public_html/`. Summaries of the words with greatest contribution to the logistic model of salience probability for each topic are also presented in Appendix B.

On inspection of the visualisation, it can be clearly seen that topics 1 and 13 are dominated by collections of hash tags. Less obvious in Figure 5.3, though easily identified in the interactive version, is that topics 6 and 8 are also dominated by hash tags.

Topic 13 in particular has very little weight in normal words (tokens that are not hash tags, smileys, punctuation etc. . .), with nearly all the probability in hash tags, the vertical bar ("|") and the word "PIC". In Appendix B we can see that

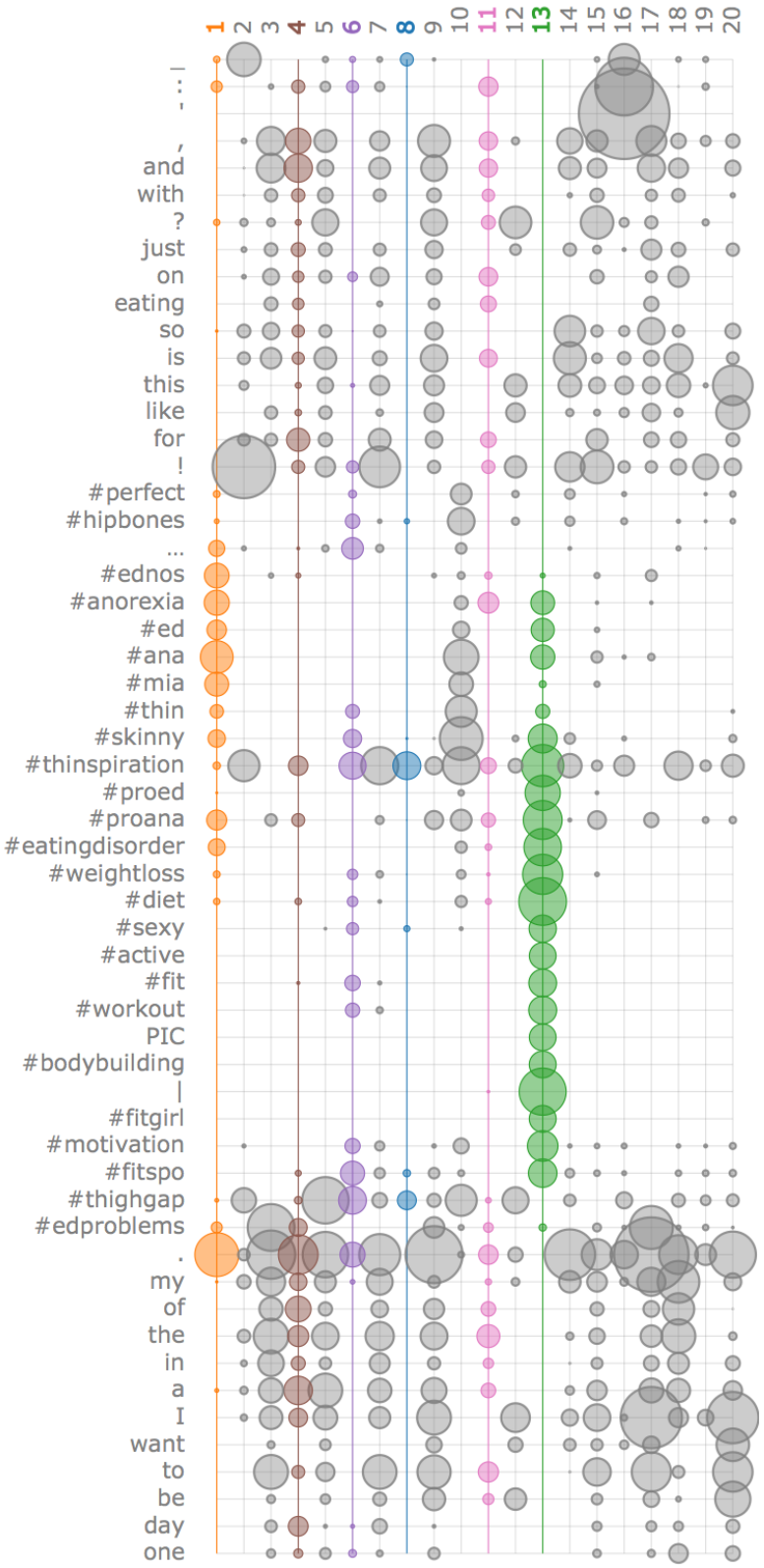

Figure 5.3: 20 topic model with probable personal salient topics highlighted. Interactive version at http://cs.anu.edu.au/~Ian.Wood/termite/

this topic has only one word (with only two occurrences) considered in the logistic model for salience probability, and it is the *absence* of words that gives this topic high estimated probability of personal salience, since only the intercept value of the logistic model plays a significant role. The absence of sentence punctuation such as full stops and commas, and the extreme lack of diversity among actual words, this topic most probably does not contain sentences and other discourse. The estimated probability must therefore be considered suspect, as the context is quite different to the identity salience study, there are very few word types to draw conclusions from and the logistic model arguably ignores any semantics that the topic may indicate, which perforce must be encoded in non-words such as hash tags.

In the case of topics 1 and 6, this argument appears less strong. Though about 90% of the tokens assigned to these topics are non-words, they nonetheless contain a diversity of words in the remaining 10% and non-trivial probability in sentence punctuation such as full stops. Topic 8 does not contain notable proportions of punctuation, though a diversity of other words is present. It would be well to investigate tweets strong in topic 8 and assess to what extent they could be said to fit the context of discussion on the topic of diet.

Of perhaps greater interest is topic 4. The tags present in this topic are also informative (in order of significance within the topic): "#edproblems", "#thinspiration", "#proana", and to a lesser extent "#diet", "#thighgap", "#fitspo"[3] and "#ednos"[4]. The association between these tags indicates particular themes seen as important or relevant to people using this topic in their tweets. These themes resonate with several of the themes identified in the expert assessment reported in Section 4.5 and Appendix C. #edproblems and #ednos specifically refer to personal difficulties relating to eating disorders. #thinspiration and #fitspo (short for "fit-inspiration") are motivational. #proana could have several interpretations, such as identifying with a community or motivation, but clearly relates some affinity to eating disorders and anorexia. #thighgap[5] is a symbol of the "thin ideal". #diet reflects a concern over controlled eating, most probably with an eye to weight loss, and occurs with similar prominence to the words "eat" and "eating". The presence of #diet within the more significant tokens suggests that the context of this topic is similar to the writing task used in the gender

[3] "fitspo" is short for "fitness inspiration".

[4] "ednos" refers to "Eating Disorder Not Otherwise Specified", a term used for clinical classification [WHO 2015].

[5] An open space between a person's thighs when their feet are together — there are many images of this in the data.

salience study. It is also interesting that this topic contains many function words and simple punctuation (“.”, “,”, “!”, “?”) and several verbs. This may suggest that tweets strong in this topic may contain short full sentences, however examination of a sample of such tweets reveals tweets that are primarily just hash tags, and that the sentence-like tokens are more likely a (low frequency) theme that is merged into this topic.

Pronouns in topic 4 are mostly in the first person singular (“I”, “I’m”, “my”, ...), with a small presence of impersonal pronouns (“it”, “it’s”) and a distinct lack of any other pronouns. It is interesting to note that personal pronouns do not have predictive power in the logistic model of personal vs. gender salience from [Dann 2011]. In past studies, high first person singular pronouns frequencies have been associated with honesty, depression, low status, personal and emotional communications, and informal language [Tausczik and Pennebaker 2010b]. However, one must be careful to match the context of those studies with the current context. For example, though status may play a role in the Twitter pro-ana community, it seems unlikely that it would reveal itself in pronoun usage in a model of such low granularity. On the other hand, personal and emotional communications seem plausible in this context. To identify the actual role of topic 4 pronouns, expert analysis of a sample of tweets strong in topic 4 would be required.

In regard to the high estimated personal salience probability for topic 4, almost all the positive contributions are from words that relate to eating (LIWC variable “ingestion”) with a small contribution from words relating to insight. The only LIWC variable with a negative contribution that is not represented is “influence” (it has a trivial contribution). Impersonal pronouns and causality words (because, effect, hence, ...) make a small contributions and the others — inclusion words (and, with, include, ...) exclusion words (but, without, exclude, ...) and negation words (no, not, never, ...) all contribute substantially (see Appendix B). The presence of most of the contributing LIWC classes as well as the hash tag #diet support the idea that the context of this topic may be similar to that of the gender salience study, and that the model is indeed detecting personal identity salience.

5.7.3 Caveats

In this section I describe some of the more technical pitfalls and difficulties encountered during the development of the analyses presented here

Topic 8 is an interesting example of unexpected text processing behaviour:

In the first attempted analysis, the tokeniser treated underscores in hash tags and user mentions as separate tokens, resulting in a large number of underscore tokens. To further complicate the situation, when counting the number of words (as opposed to non-word tokens) attributed to a topic, the underscore ("_") was counted as a word (as opposed to punctuation) due to the default characterisation of word characters in perl regular expressions (which include underscores). The number of words attributed to a topic was used in calculating salience probabilities, so this had a small effect on some salience scores. 1067 underscore characters are assigned to topic 8, which increased the number of words (as opposed to non-words) for topic 8 from 5124 to 6191, a 20% increase. Topics 1 and 6 also contained notable numbers of underscores, though only 4% and 2% of words respectively. Topic 8's regression coefficient remained dominated by the regression model intercept, however, changing only from 3.13 to 3.10, and thus the resulting salience probability only changed from 0.958 to 0.957. The effect on topics 1 and 6 was an order of magnitude smaller. If, however, we were interested in topics with low estimated personal probability, the word frequencies dominate in the regression calculation, and a substantial shift could occur. For example, topic 16 shifted from 0.48 to 0.32 with a 20% increase in words if the underscore is included.

Two of the topics with high estimated personal salience have non-trivial probability assigned to colons (:) and underscores (_). In Twitter, one can "reply" to a tweet, in which case the produced tweet typically has the form

> **@***my_user_name:"replied tweet text" my additional text.*

If a reply is not recognised by the pre-processing filter and removed (for example, a tweet made before Twitter provided metadata indicating replies and which the user edited the reply text), and a user name containing an underscore was not recognised by the mention filter, the resulting corpus document would contain at least one underscore and colon. Many such tweets would result in frequent co-occurence of these characters, and tend to cause them to also co-occur in one or more topics. Such topics would include words often used in these replies that may not have a strong relationship elsewhere in the corpus. This is an indication of one way a topic may represent non-semantic structural features and easily be mis-interpreted as representing general semantic relations. For example, it is generally a good idea to investigate sample documents with strong representation in a topic before drawing too many conclusions.

It is worth noting that not all words are present in the Figure 5.3, and that

some words with non-trivial probability in one or more topics may not have made it into the 100 words chosen by termite to represent the model. For example, in topic 13, the words"see" and "more" can be found to have substantial representation in the interactive visualisation for which non-words were excluded (found at the URL mentioned above) but these words do not appear in Figure 5.3.

5.8 Discussion and Future Work

In Section 5.2 I state that a favourable posterior predictive check for a topic indicates that words are as evenly distributed as can be expected, and claim that thus topic probabilities can provide a reasonable estimate for word frequencies among words assigned to that topic. A more rigorous statistical assessment of that claim and quantification of its uncertainties would be in order.

The observation that regularisation can break the independence of topic and document word allocations (as tested by the posterior predictive checks presented here) likely extends to other supervised topic models also. In Section 2.4.4 I introduce several other methods for providing prior information about topic structure [Jagarlamudi et al. 2012; Hall et al. 2008; Ramage et al. 2009; Ramage et al. 2011]. If these models also break the independence of topic and document word distributions, the general utility of such supervised models must be brought into question — do the models reflect "true" structure in the corpus, or merely the prior provided? Measuring this effect and developing unbiased tests tailored to the intended application of such models would be needed to establish model credibility.

Conversely, if those approaches to topic supervision are found not to adversely effect, or to minimally effect, the independence of topic and document word distributions, they would be good candidates for improving the presented approach to measuring psychological (or other word frequency correlated) features.

The analysis in Sections 5.7.2 and 5.7.3 is intended as an indication of the type of analyses that can be done to interpret the results of the presented approach. A deeper analysis including expert review of tweets strong in presumed salient topics and associations between topics found to be good candidates for salience and particular users, groups, hash tags etc. . . would be of interest to both the social media and eating disorder research communities.

5.9 Conclusions

This chapter develops and demonstrates a methodology for combining topic models with word frequency based psychometric tools, providing useful contextualisation and a measure of the features those tools detect. Results such as this can help to provide insights into the psychological processes active within a group as well as provide some measure of their activity.

Though the psychological study used to provide a psychometric proxy was small and arguably distant from the context of people tweeting in the Twitter eating disorder and thinspiration community, this study serves as a useful illustration, paving the way for future studies combining more traditional psychological questionnaires, elicited text responses and online social media data.

Topic regularisation as a means for model supervision was found not to improve the method due to its adverse effect on the independence of topic and document word distributions (as measured by posterior predictive checks).

The next chapter introduces a method for combining a topic model and overlapping network community model drawn from the same data set, associating individual documents with communities and estimating topic mixtures for each community.

Chapter 6

Community Topic Usage

This chapter presents a Bayesian model to identify community topic usage in data combining documents and a network of their authors. The content of this Chapter has been published in the proceedings of the workshop "Topic Models — Post Processing and Applications" at CIKM 2015 [Wood 2015c], presented here with minor additions.

Members of social groups share some purpose, beliefs or other common human features, and one would expect those features to appear as common language markers. One premise of this thesis is that such common language use can be detected via topic models.

In the presented model, overlapping communities are identified using standard network community detection algorithms and document topics using standard topic models. The model then associates those topics with communities, balancing community topic coherence with author community affiliation.

This chapter is organised as follows: In Section 6.1 I present an overview of the background, motivation and relevant literature to the model. In Section 6.2 I describe and develop the model, including the conjugate prior to the Dirichlet distribution. In Section 6.3 I present an algorithm based on Gibbs sampling for estimating the posterior. In Section 6.4 I describe the data set and contributing topic and community detection models used as an example in this study. In Section 6.5 I develop two metrics for assessing model quality. In Section 6.6 I present results showing the that the model succeeds in its aims. In Section 6.7 I discuss the results and their implications, and indicate some research questions that may be of interest for future work. In Section 6.8 I summarise the contribution.

6.1 Overview

Several studies have found that communities in the Twitter follower network can act as a kind of forum on particular topics of discussion [Java et al. 2007; Huberman et al. 2008; Himelboim 2014]. In this scenario, tweets intended for such a forum would reflect those topics, whereas tweets by the same users that are intended for other audiences would show distinct topical content. It is the aim of the work presented here to distinguish the intended audience (in terms of follower network communities) of each tweet and in this way estimate the topics used by those communities. One would expect that many Twitter users would be members of/contribute to multiple communities, thus one would expect such communities to be overlapping [Java et al. 2007].

Approaches to linking social media texts with network communities have been studied previously. Java et al. [Java et al. 2007] performed overlapping community detection on the full Twitter network and identified coherent themes in key terms used by some inferred communities, though they were not clear on how the key terms were identified and did not provide numerical measures of such coherence. There were about 94,000 Twitter users in April 2007 when they performed their study, thus scale was less of an issue than today (in 2015 there are over 300 million Twitter users).

Duan et al. [Duan et al. 2011] developed a full Bayesian model incorporating both a stochastic block model for community detection and hierarchical Dirichlet process for topic detection. In this model, all of an authors documents are assigned to just one community (hence they do not overlap) and its scalability is questionable.

Li et al. [Li et al. 2012] present a different approach to combined community and topic detection by utilising extra thematic metadata interpreted as a publication venue. They applied their model to Twitter data, where hash tags served as venues, and scientific publications, where conferences and journals served as venues. The Twitter follower network was not utilised. In their model, communities (not documents) have topic mixtures and topics generate both words and hash tags/venues. The Twitter data analysed was intended as a summary of hot topics over a 2 month period, in contrast to the data utilised here that intends to capture interactions within a restricted set of Twitter communities over a longer period. In such a social data set, follower links are of great importance, as they represent a significant conduit over which interactions are possible.

Earlier, Li et al. [Li et al. 2010] combined the results of community detection

and topic modelling and applied the resulting synthesis to social bookmarking data. The community model they applied, however, did not produce overlapping communities so a naive approach to inferring community topic proportions was effective. Inferring overlapping communities, as is presented in this chapter, is a more difficult task.

The approach presented here is similar to Li et al. [Li et al. 2010], performing community and topic modelling independently and later combining the results, however overlapping communities are accommodated. This requires a method for attributing each users topic usage to the communities in which she participates.

A naive approach for attributing topic usage to communities is to distribute each author's topic usage (averaged over authored documents) proportionally to the author's community affiliations. This, however, does not take into account that authors will communicate differently when intending their communication for different communities.

To accommodate such a possibility, a novel Bayesian model is developed that attempts to identify the intended audience of such communications. It is assumed that the intended audience of each document is a single community. In this way, the topics of discussion used by the detected communities can be inferred by averaging topic proportions of attributed documents. The model balances the coherence of community topic usage against the similarity between each author's document allocations and community affiliation by modelling community topic usage with a Dirichlet distribution and using author affiliation as a prior on document allocation.

The Dirichlet distribution modelling community topic usage is given a fixed concentration parameter leading to a well defined conjugate prior. Though this prior is computationally expensive, the already reduced dimensionality in both topics and communities make a tractable algorithm feasible, even for large data sets. The model is applied to a network snapshot of the pro-ana Twitter data from September 2014 (see Sections 4.4) and topic model of tweets up to the same date (see Section 4.1), resulting in clearly distinct topic usage by most detected communities.

6.2 Document Assignment Model

Assigning documents to their author's communities is done according to two premises: the proportion of an author's documents in a community should reflect the author's proportional community membership and the topic proportions of documents assigned to a given community should be similar. This is operationalised by the following generative model. A authors, C communities, T topics and N documents are modelled.

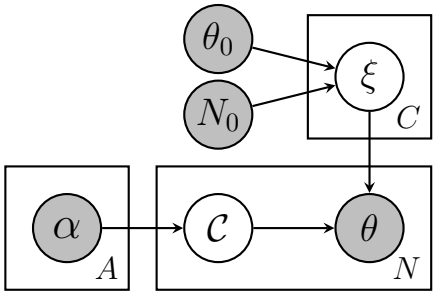

Figure 6.1: Generative Model

Document community assignments $\mathcal{C}$ are generated by a fixed multinomial whose probabilities are the document author's community membership proportions α. For each document d assigned to community c, topic proportions θ_d are drawn from a Dirichlet distribution (with parameters ξ_c) for that community. A conjugate prior for the ξ is provided, parametrised by N_0 and θ_0 (see Section 6.2.1 for the construction of the prior). The model is summarised in Figure 6.1. Grey nodes indicate observed or pre-set values.

The probability of assignment of document d with author a_d to community c, and the probability of a documents topic distribution θ_d are as follows:

$$P(d \in c) = \alpha_{a_d c} \tag{6.1}$$

$$P(\theta_d | d \in c, \xi_c) = B(\xi_c)^{-1} \prod_{t=1}^{T} (\theta_{dt})^{\xi_{ct}-1} \tag{6.2}$$

6.2.1 A Conjugate Prior For Dirichlet Distributions

The Dirichlet distribution is a member of the exponential family of distributions, and as such has a (conjugate) prior with a relatively simple, constant-dimensional Bayesian update. Given the equation for the T dimensional Dirichlet distribution with parameters ζ

$$P(\theta) = B(\zeta)^{-1} \prod_{t=1}^{T} \theta_t^{\zeta_t - 1} \tag{6.3}$$

$$B(\zeta) = \frac{\prod_{t=1}^{T} \Gamma(\zeta_t)}{\Gamma(\sum_{t=1}^{T} \zeta_t)} \tag{6.4}$$

where B is the beta function, it is easy to write down a candidate conjugate prior and corresponding posterior update after evidence $\{\theta_1 \dots \theta_M\}$:

$$P_\pi(\zeta) \propto B(\zeta)^{-N_0} \prod_{t=1}^{T} (\theta_{0t})^{\zeta_t - 1} \tag{6.5}$$

$$P(\zeta|\theta_1 \dots \theta_N) \propto B(\zeta)^{-(N_0+N)} \prod_{t=1}^{T} \left(\theta_{0t} \prod_{m=1}^{M} \theta_{mt} \right)^{\zeta_t - 1} \tag{6.6}$$

here, n ranges from 1 to N and t from 1 to T. The values for N_0 and θ_0 can be interpreted in terms of hypothetical prior observations: N_0 being the number of prior observations and θ_0 the element-wise product of those observations.

Note that due to the $\Gamma(\sum_{t=1}^{T} \zeta_t)$ term in $B(\zeta)$, this only defines a probability if $\sum_{t=1}^{T} \zeta_t$ is bounded. We could however multiply this candidate by an arbitrary function of ζ and it would remain a conjugate prior (ie: have convenient posterior form and update). For example we could choose to multiply by $\Gamma(\sum_{t=1}^{T} \zeta_t)^{-\sum_{t=1}^{T} \zeta_t}$ and the resulting function would have bounded integral (and could thus define a probability). For the purposes of this study, however, we chose instead to fix $\sum_{t=1}^{T} \zeta_t$.

For convenience we will express $\zeta = \Xi\xi$ with fixed concentration parameter $\Xi = \sum_{t=1}^{T} \zeta_t > 0$, a scalar, and $\sum_{t=1}^{T} \xi_t = 1$, $\xi_t \geq 0$. We can now write down the full probability of the model. Taking $\mathcal{C}_d$ to represent the allocated community for document d (so $\alpha_{a_d \mathcal{C}_d}$ is the prior probability of that allocation) and N_c the

number of documents allocated to community c, we have:

$$\begin{aligned}
&P(\mathcal{C}, \theta, \xi | \alpha, \theta_0, N_0) \\
&\quad = \prod_{d=1}^{D} P(d \in \mathcal{C}_d) P(\theta_d | d \in \mathcal{C}_d, \xi_{\mathcal{C}_d}) \prod_{c=1}^{C} P(\xi_c | N_0, \theta_0) \\
&\quad \propto \prod_{d=1}^{D} \alpha_{a_d \mathcal{C}_d} \left(B(\Xi \xi_{\mathcal{C}_d})^{-1} \prod_{t=1}^{T} (\theta_{dt})^{\Xi \xi_{\mathcal{C}_d t} - 1} \right) \\
&\qquad\qquad \times \left(\prod_{c=1}^{C} B(\Xi \xi_c)^{-N_0} \prod_{t=1}^{T} (\theta_{0t})^{\Xi \xi_{ct} - 1} \right) \\
&\quad = \prod_{c=1}^{C} B(\Xi \xi_c)^{-(N_0 + N_c)} \left(\prod_{d \in c} \alpha_{dc} \right) \prod_{t=1}^{T} \left(\theta_{0t} \prod_{d \in c} \theta_{dt} \right)^{\Xi \xi_{ct} - 1}
\end{aligned} \tag{6.7}$$

6.3 Estimation

To obtain a maximum a posteriori (MAP) estimate for document-community associations and community topic distributions, we use a modified Gibbs sampling algorithm not dissimilar to that used in [Griffiths and Steyvers 2004]. The method iterates between sampling from the posterior distribution of document-community associations and MAP estimation of ξ with those associations fixed.

To sample document community allocations, we need the conditional probability of a documents community membership given the current value of ξ. Omitting inconsequent conditional dependencies and terms independent of d and c, we obtain:

$$\begin{aligned}
P(d \in c | \xi, \theta) &\propto P(d \in c | \xi) P(\theta_d | \xi) \\
&\propto \alpha_{dc} B(\Xi \xi_c)^{-1} \prod_{t=1}^{T} (\theta_{dt})^{\Xi \xi_{ct} - 1}
\end{aligned} \tag{6.8}$$

For the MAP estimation of ξ we need its conditional probability given current document allocations. Again omitting inconsequent dependencies and terms independent of ξ_c and c, and writing θ_c for the set of topic proportions for documents in c, we obtain:

$$
\begin{aligned}
P(\xi_c|\mathcal{C},\theta) &\propto P(\theta_c|\mathcal{C},\xi_c)P(\xi_c) \\
&\propto \left(\prod_{d\in c} B(\Xi\xi_c)^{-1}\prod_{t=1}^{T}(\theta_{dt})^{\Xi\xi_{ct}-1}\right) \\
&\quad\times\left(B(\Xi\xi_c)^{-N_0}\prod_{t=1}^{T}(\theta_{0t})^{\Xi\xi_{ct}-1}\right) \\
&= B(\Xi\xi)^{-(N_0+N_c)}\prod_{t=1}^{T}\left(\theta_{0t}\prod_{d\in c}\theta_{dt}\right)^{\Xi\xi_{ct}-1}
\end{aligned}
\tag{6.9}
$$

Estimates for ξ_c were obtained from Equation (6.9) via numerical optimisation. With fixed Ξ and due to the logarithmic convexity of the Gamma function for positive real numbers, this expression can be seen to be logarithmically concave, thus numerical optimisation of its log can be expected to behave reasonably, as was found to be the case.

Neither Equation (6.8) nor (6.9) scale well, however due to the already reduced dimensionality of the input data through topic modelling and community detection algorithms, it has proved tractable on large data sets.

6.4 Data Set

Collected tweets up to December 2014 (see Chapter 3) and an inferred follower network snapshot from the end of that month (see Section 4.4) were used for the analyses.

Text Data and Topic Model: Retweets were removed and Tweets were tokenised by standardising numerous text emoticon forms, isolating punctuation as individual word tokens and converting mixed case words to lower case (all caps words were retained). Url's, #tags, @mentions and apostrophised words (eg: "didn't") were left unchanged. Further pre-processing included removal of word tokens appearing less than 5 times and removal of tweets with less than 3 word tokens. See Section 5.6 for more discussion on these choices. This resulted in a corpus of 262,736 documents and a vocabulary of 18,713 words.

A standard latent Dirichlet allocation [Blei et al. 2003] topic model with 20 topics was inferred for the resulting corpus. The LDA Dirichlet prior on topic/word probabilities was set to $\beta = 0.01$. Writing N for the number of

words, D the number of documents in corpus, T the number of topics, the parameter for the LDA Dirichlet prior on document/topic probabilities was set to $\alpha = 0.05N/DT$. This value allocates 5% of the probability mass for smoothing.

Network Data and Community Model: I consider only mutual follower links as they indicate greater likelihood of mutual interaction, a feature we would expect of social communities. Degree one nodes were removed as they are highly likely to have many connections outside the observed data, and thus be members of communities that are not otherwise represented. If the hypothesis that topic usage is driven by community membership, these authors would often use discussion topics not typical of the communities of interest, however they would be assigned to the same communities as their single connection, and would thus dilute the observable topic coherence of those communities.

The resulting network has 40,242 nodes and 441,655 edges. Overlapping communities were inferred using a mixed membership stochastic block model [Gopalan and Blei 2013; Airoldi et al. 2009]. This model is able to make an unsupervised estimate of the number of communities, and with this data, it estimated 183 communities. Visualisation of the network with community observations (Figure 6.2) reveals a relatively small number of fairly distinct, well separated communities with the remainder highly interconnected. A rough visual count reveals approximately 18 noticeable communities plus a region without clear community membership. One should take such a rough estimate as a lower bound on the number of coherent communities. Using this as a guide, I chose to make three models, one with the inferred 183 communities, one with 20 communities and a third with 50 communities.

Combination After performing network and topic analysis, the results of the two analyses were combined. Only users who appeared in both were retained, that is, users who had at least one tweet retained for the topic model as well as two (recently observed) mutual follower links. This resulted in 15,515 users and 133,851 of their tweets.

Note that the inputs for the topic modelling and community detection analyses are completely separated, and so the only sources of dependence between them are possible relationships in the data between tweet texts and author networks.

This approach was preferred to performing the analysis on the reduced data for two reasons: first, to avoid any chance of dependencies between the two models, and second, because excluded users from one model could provide useful

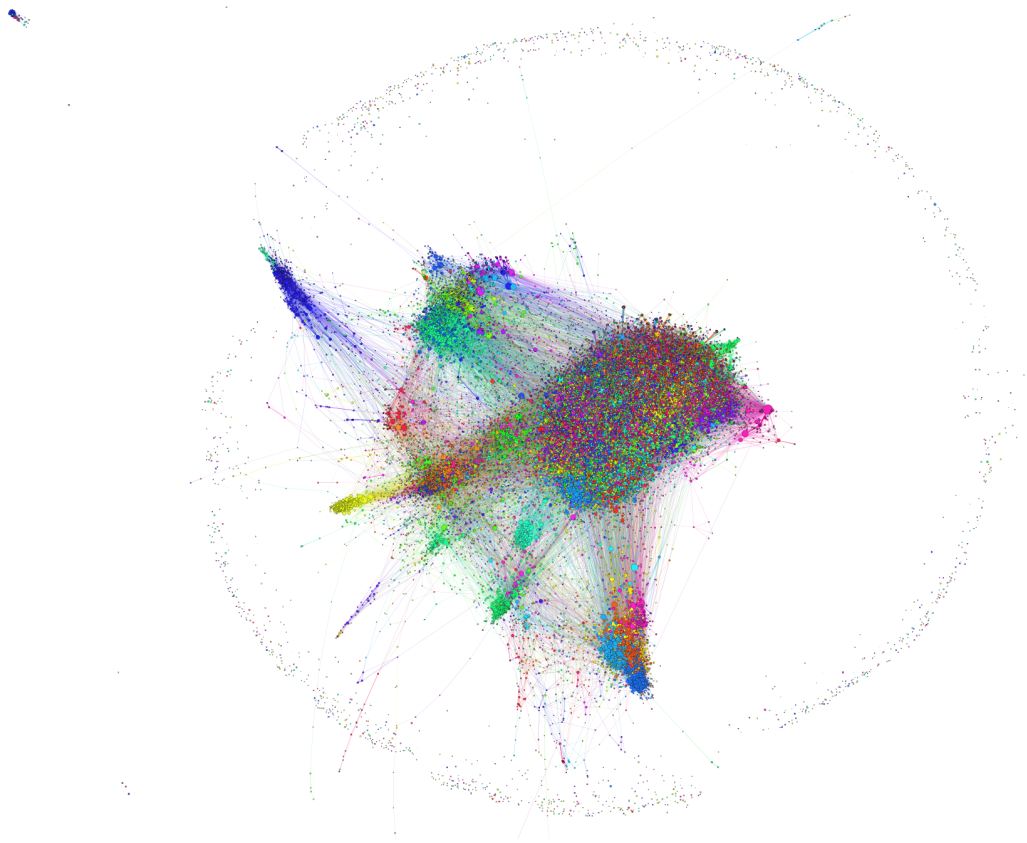

Figure 6.2: Network Visualisation with 183 Communities. Colours represent different communities, node size indicates bridgedness [Nepusz et al. 2007]. Note few distinct, separated communities and many highly interconnected communities.

context in the other. E.g.: a user with few tweets may have connections that are informative of community structures and vice versa.

6.5 Metrics of Model Quality

In this unsupervised setting, comparison to a ground truth is impossible. The numerical metrics below attempt therefore to assess the efficacy of the model in terms of the model's goals: maximising community topic coherence and respecting author community membership. The metrics were applied both to estimated models and to naive document allocation via community membership proportions (α) alone. Results are presented in Table 6.1.

Community Topic Coherence To capture how effective the models had been at resolving coherent community topic proportions, the conditional entropy of community allocations $\mathcal{C}$ given community topic proportions $\mathcal{T}$ was employed.

$$H(\mathcal{C}|\mathcal{T}) = -\sum_{c=1}^{C} P(c) \sum_{t=1}^{T} P(t|c) \log_2 P(t|c) \tag{6.10}$$

This quantity captures the amount of extra information (measured in bits) needed to obtain the community document allocations given knowledge of community topic proportions. If the documents associated with a community are faithful to the topic proportions of that community, you would expect this to be low. On the other hand, if they have a diversity in their topic mixes, much extra information would be needed to identify them.

Taking N_a to be the number of documents from author a and recalling N represents the total number of documents, α_{ac} the affinity of author a for community c, and θ_{dt} the topic proportions of document d, probabilities for naive allocation can be written as follows:

$$\begin{aligned} P_{\text{naive}}(c) &= \sum_{a=1}^{A} P(a)P(c|a) \\ &= \sum_{a=1}^{A} \frac{N_a}{N} \alpha_{ac} \\ &= \frac{1}{N} \sum_{a=1}^{A} N_a \alpha_{ac} \end{aligned}$$

$$
\begin{aligned}
P_{\text{naive}}(t|c) &= \frac{P(t,c)}{P(c)} \\
&= \frac{\sum_{a=1}^{A} P(a)P(t|a)P(c|a)}{\sum_{a=1}^{A} P(a)P(c|a)} \\
&= \frac{\sum_{a=1}^{A} \frac{N_a}{N}(\frac{1}{N_a}\sum_{d\in a}\theta_{dt})\alpha_{ac}}{\sum_{a=1}^{A}\frac{N_a}{N}\alpha_{ac}} \\
&= \frac{\sum_{a=1}^{A}\alpha_{ac}\sum_{d\in a}\theta_{dt}}{\sum_{a=1}^{A}\alpha_{ac}N_a}
\end{aligned} \tag{6.11}
$$

For the estimated models, I use MAP estimates of document allocations for community probability and expected values of the posterior community Dirichlet distributions, which are just their parameters ξ_c, for conditional topic probabilities.

$$
\begin{aligned}
P_{\text{estimated}}(c) &= \frac{N_c}{N} \\
P_{\text{estimated}}(t|c) &= \xi_{ct}
\end{aligned} \tag{6.12}
$$

To assess individual communities, we can also calculate the entropy $H(c|\mathcal{T})$ for some community c:

$$
\begin{aligned}
H(c|\mathcal{T}) = -\Bigg[& P(c)\sum_{t=1}^{T} P(t|c)\log_2 P(t|c) + \\
& P(\neg c)\sum_{t=1}^{T} P(t|\neg c)\log_2 P(t|\neg c)\Bigg]
\end{aligned} \tag{6.13}
$$

Again it is useful to compare entropies from a naive model and an estimated model. We already have formulae for $P(c)$ and $P(t|c)$ (Equations 6.8 and 6.9). For a naive model, we have:

$$
\begin{aligned}
P_{\text{naive}}(\neg c) &= \frac{1}{N}\sum_{a=1}^{A} N_a(1-\alpha_{ac}) \\
P_{\text{naive}}(t|\neg c) &= \frac{\sum_{a=1}^{A}(1-\alpha_{ac})\sum_{d\in a}\theta_{dt}}{\sum_{a=1}^{A}(1-\alpha_{ac})N_a}
\end{aligned} \tag{6.14}
$$

and for the estimated models, we have:

$$
\begin{aligned}
P_{\text{estimated}}(\neg c) &= \frac{N - N_c}{N} \\
P_{\text{estimated}}(t|\neg c) &= 1 - \xi_c
\end{aligned}
\tag{6.15}
$$

Faithfulness to Author Community Membership There can be a tension between respecting author community affinities and creating coherent community topic distributions. A model that produces excellent community topic distributions may require documents to be allocated in different proportions to their author's community affinities.

To assess this disparity, we use the Hellinger distance [Bhattachayya 1943] between estimated author community affinities calculated from document assignments and the actual affinities used as inputs to the model. Kullback-Leibler divergence [Kullback and Leibler 1951] was also considered, however this leads to uninformative infinite divergences if the estimate for a community is zero and the actual affinity non-zero. The Hellinger distance can be written thus:

$$
\begin{aligned}
&H(P_\alpha(c), P_{\text{estimated}}(c)) \\
&\quad = \frac{1}{\sqrt{2}} \sqrt{\sum_{c=1}^{C} \left(\sqrt{P_\alpha(c)} - \sqrt{P_{\text{estimated}}(c)} \right)^2} \\
&\quad = \frac{1}{\sqrt{2}} \sqrt{\sum_{c=1}^{C} \left(\sqrt{\sum_{a=1}^{A} \frac{\alpha_{ca} N_a}{N}} - \sqrt{\frac{N_c}{N}} \right)^2}
\end{aligned}
\tag{6.16}
$$

Community membership of authors in the Twitter follower network is an indication of who they listen to. The model presented here makes the assumption that documents are divided between those communities in similar proportions to the number of links to those communities, but this may not be the case. The links represent the mix of sources of tweets that a user sees (passive communication), whereas the documents assigned to a community represent tweets intended to be seen by that community (active communication). Proportions of active and passive communication may not always coincide. For example, other users followed for interest as sources of information are unlikely to be considered as targets for published tweets.

As such, we may not necessarily expect complete symmetry between listening (represented here by follower links and α) and speaking (represented by tweets

183 Communities

	Naive	$\Xi = 30$	$\Xi = 100$	$\Xi = 600$
Entropy	3.78	2.81	2.27	1.99
Hellinger	0	0.103	0.154	0.176

50 Communities

	Naive	$\Xi = 30$	$\Xi = 100$	$\Xi = 600$
Entropy	3.84	3.23	2.84	2.62
Hellinger	0	0.161	0.158	0.162

20 Communities

	Naive	$\Xi = 30$	$\Xi = 100$	$\Xi = 600$
Entropy	3.94	3.76	3.53	3.38
Hellinger	0	0.108	0.116	0.146

Table 6.1: Conditional Entropy $H(\mathcal{C}|\mathcal{T})$ (Equation 6.10) and Hellinger Distance $H(P(c|\alpha), P(c|\mathcal{C}_{\text{estimated}}))$ (Equation 6.16)

and their allocation to communities), and low similarity may be acceptable.

6.6 Results

Models were inferred for several values of Ξ and compared to naive document allocation via community membership proportions α alone using the metrics presented in Section 6.5. Initial experiments suggested values up to $\Xi \simeq 600$ appeared to perform well whilst greater values lead to numerical instabilities. Models were thus estimated with $\Xi \simeq 600$ and two smaller values, $\Xi = 100$ and $\Xi = 30$, for comparison. Results are summarised in Table 6.1.

As expected, the larger value of Ξ produced more resolved community topic proportions (lower entropy scores) and were less faithful to the community membership information inferred from the network data (Table 6.1). The increased distance to community membership information was however small compared to the improved resolution of topics for communities, and some deviation from community membership is acceptable (see Section 6.5), thus higher values of Ξ should be preferred.

The implementation used for experiments presented here assigned batches of 100 documents between estimations of ξ_c for communities whose membership had changed. Estimations of ξ were done with scipy.optimize.minimize using the "Nelder-Mead" method [Singer and Nelder 2009], a hill climbing simplex method.

This method does not allow caching and control of the initial simplex, thus small perturbations of community membership require a similar number of function evaluations (in the order of 800) to uninformed starting points. Substantial improvements in run times could be achieved with control of the starting simplex and simple heuristics for required precision at different stages during estimation. Execution times were not insubstantial (approximately 3 hours to converge for all configurations on a 16 core 2.3Ghz machine) but only around 30 full iterations (over all documents). We could expect at least an order of magnitude improvement with more intelligent and integrated numerical optimisation of ξ. It should also be noted that bounds checking of ξ was performed within the optimised function to prevent evaluation outside the simplex. Such checks would be better placed within the optimisation routine itself.

6.7 Discussion and Future Work

On inspection of the community topic allocations, it was found that approximately half the communities had more than half the probability mass concentrated on just one topic in all models. The individual community entropy scores (Equation 6.13) give a good indication of the level of concentration, the more concentrated having notably lower entropy scores. Figure 6.3 shows community allocations for 50 communities and $\Xi = 100$. Similar patterns were found for other models.

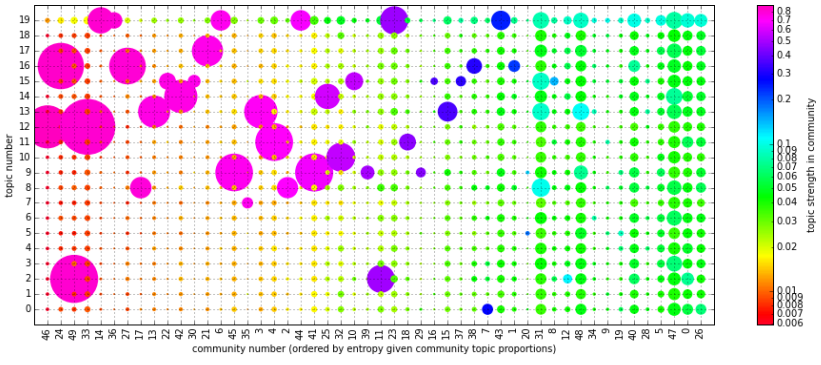

Figure 6.3: Community topic allocations with 50 communities, $\Xi = 100$. Data point area proportional to expected sum of document topic proportions for a community. Communities ordered by Equation 6.13.

One might suspect that the models relative freedom to allocate documents to

communities, combined with a strong prior enforcing community topic coherence, could be sufficient to enforce such clear associations between topics and communities. If communities were strongly correlated, the model would be free to move documents between correlated communities without compromising the community membership prior, whereas low community correlations would restrict that freedom.

Working again with the 50 community model, community correlations (over documents, given document/community probabilities α from the network community model) were calculated followed by Principal Component Analysis (PCA) on the correlation matrix. The correlation data is intrinsically high dimensional, with the maximum proportion of variance for a component being just 7.1%, more than half the components accounting for $> 1.9\%$ and all but one accounting for $> 1\%$. Hierarchical clustering using correlations as a similarity metric supported this observation, revealing just two clusters with internal correlations ≥ 0.2 (communities 19, 32, 46 and 14, 44) with the greatest correlation in the data at .385 between communities 32 and 46.

Given that, in Table 6.1, the Hellinger distances are low (indicating that most documents have *not* been forced to communities where their authors have low membership), one is drawn to conclude that the inferred community/topic associations are indeed intrinsic to the models input data (the original topic and community models).

Another potential explanation of the results could be the existence of inherent dependencies between the topic and community models. The models were, however, trained on completely separate data sets (tweet texts for topics, author links for communities), so the only source of dependencies are potential relationships between authors texts and their follower link choices.

The clear associations between topics and communities thus testifies to the efficacy of the community detection and topic modelling algorithms and supports the hypothesis that communities in the mutual follower network often define fora for discussion or sharing along a particular theme. None the less, further investigation of the model, perhaps using synthetic data sets with known properties may be appropriate to allay such suspicions.

The main bottleneck in the presented algorithm is numerical estimation of the posterior community topic proportions. Substantial improvements in performance could be achieved by tailoring the optimisation algorithm to the specific setting. Possible extensions of the model include estimation of the scaling parameter for community topic proportions and introducing a parameter to moderate

the relative strength of author community membership and community topic coherence during inference.

The principles used and design of the algorithms presented here are a step to understanding the relations between community structures and topic usage with the future aim of developing a joint model of community detection in author networks and topic modelling of document content. Recent work with Bayesian community detection models [Gopalan and Blei 2013] and sampling techniques for topic models [Li et al. 2014] shows promise for eventual combination of community detection with Bayesian topic models in a scalable algorithm. The work presented here can act as a motivator and guide to such further development.

6.8 Conclusions

This chapter presents an approach to identifying Twitter communities and their topics of discussion. Existing efficient methods for community detection and topic modelling are leveraged and a novel Bayesian model and inference algorithm are developed to associate tweets with communities in which their authors participate.

A Dirichlet distribution is used to model community topic usage, and a conjugate prior for the Dirichlet distribution is developed. A modified Gibbs sampling procedure incorporating alternate sampling of document/community allocations and MAP estimation of community topic Dirichlet distributions is used to estimate the posterior. The MAP estimation step requires costly numerical optimisation, however due to the already reduced dimensionality of the problem (from text topic modelling and network community detection), this remains tractable for reasonably large data sets.

The model is applied to a collection of 262,736 tweets and 441,655 user follow relations collected from public tweets related to "pro-anorexia" and eating disorders. A substantial improvement of community topic coherence is demonstrated relative to a naive approach that utilises author community membership alone. Results show very distinct community topic usage for more than half the communities. This is a strong result, supporting the hypothesis that communities in the mutual follower network often serve as fora for particular themes. Though the possibility of inherent bias in the model, leading to the potential identification of patterns not intrinsic to the data, has been addressed in for this data set, a deeper and more systematic analysis of such biases would be worth investigating.

The clear associations between topics and communities support the conclusion that the detected communities represent social groups operating on Twitter. This work, therefore, provides a foundation for the analysis of topic usage by social groups and opens the way to investigation of the social significance of those topics, such as their role as symbols of social identity, norms or other social constructs.

The next chapter concludes the thesis, summarising the contributions and presenting both a larger research vision and concrete directions for further work towards that vision.

Chapter 7

Conclusion

This chapter concludes the thesis, summarising the important results, presenting opportunities for future work and discussing this work in the context of a larger vision for developing our understanding and knowledge of collective human conduct.

This chapter is organised as follows: Section 7.1 briefly reviews the contributions presented in this thesis. Section 7.2 presents an overview of concrete research directions relevant to the theme and larger vision of this thesis. Section 7.3 concludes the thesis, presenting the larger vision motivating this work and locating the contribution and concrete research directions in that context.

7.1 Contribution

In this thesis I identified a substantial gap in current research into modelling social processes, namely the lack of direct empirical grounding with actual social groups, and made a modest start to fill that gap. To empirically ground a model, one needs to make measurements. Social phenomena occur both in the minds of the individuals that make up the group or society in question and in their communications. Direct measurement of peoples minds is challenging, if not impossible, however we can succeed in capturing the interactions that facilitate those phenomena and in this way gain insights into the internal processes involved, paving the way to a deeper understanding of the phenomena as a whole.

The advent of socially generated data such as online social media presents an opportunity to capture an (at times large) proportion of social communications, with the hope of identifying the social phenomena and processes that moulded those communications in greater detail and with greater fidelity than has previ-

ously been possible.

The volume and complexity of available data is vast. This presents challenges both to identifying coherent social contexts within the available data and the identification of conceptual and linguistic entities that represent social processes operating in those contexts. This thesis has presented initial investigations into both of these challenges, developing methods to collect community focussed data, identify socio-linguistic contexts within that data and to link traditional, questionnaire-based psychological surveys and experiments to those contexts.

Chapter 3 presented a methodology for targeting social media data collection at a particular social group, a pragmatic approach to collect richly dynamic communication and network data from that group and a data set collected from the eating disorder and pro-anorexia community operating on Twitter.

To target a group of interest, the methodology starts with a set of contextual markers (e.g.: hash tags) thought to be associated with the group. Further markers are added to the set by iteratively collecting a sample of social media data and identifying new markers that identify data relevant to/generated by the group of interest, differentiating it from data external to or irrelevant to the group. This was applied to the Twitter pro-anorexia and eating disorder community, resulting in a set of hash tags used almost exclusively by that group.

Collecting richly dynamic data from Twitter presents many challenges due to limitations Twitter places on data collection rates and the inability to directly track changes to user profile and friend/follow data. In Chapter 3, a strategy is developed to maximise observations of relevant changes within those limits and in Chapter 4 approaches to utilise the resulting partial dynamic information are presented. Experiences gained through the collection of data from the pro-anorexia and eating disorder community over nearly 3 years are reported, with many solutions to pragmatic challenges applicable to further data collection efforts from Twitter and other online social media sources are embedded in the resulting data collection system.

The resulting data set itself is unique, providing a near continuous record of group communications during the collection period and records of a substantial number of user profile and friend/follower network changes. To the best of my knowledge, no other data sets with such rich information on network dynamics exist in the public research literature. Testifying to its uniqueness, high levels of friending (link formation) and unfriending (link destruction) were observed for some links, a behaviour, to the best of my knowledge, not previously observed in the twitter research literature.

In Chapter 5 a methodology is developed to link linguistic contexts (represented by topic models) with psycho-linguistic features revealed by survey data combining established Likert-scale psycho-metric questionnaires and targeted free text expression. The ability of a topic model to coherently identify linguistic contexts is tested with relevant posterior predictive checks. The combined survey data is used to connect measured psychological features with linguistic markers in the form of word or word-class frequencies (e.g.: LIWC [Tausczik and Pennebaker 2010b] word classes).

Though the collected data is well focussed on the group of interest, it nonetheless contains users and tweets that can not be considered part of the group as well as complex subgroup structures. This can be expected in social media data in general. Working from the premises that subgroups can be identified through community structures in the friend/follower network and that topic models such as Latent Dirichlet Allocation (LDA) can identify themes of communication relevant to processes of social influence and evolution, Chapter 6 presents a method for attributing topic usage to groups and thus detecting those communication themes and associated social entities. The outputs of existing network community detection algorithms applied to the follower network, and topic models applied to Tweet texts, are combined with a Bayesian attribution model. The resulting synthesis shows strong associations between groups and topics for the pro-ana and eating disorder data, supporting the supposition that topics can represent socio-linguistic contexts and paving the way for identification and tracking of those contexts and how they develop and propagate.

What I have presented in this thesis are techniques intended for the detection of socially and psychologically relevant entities as represented in community and linguistic contexts, establishing statistically the validity of those techniques. It was considered beyond the scope of this thesis to assess in depth the candidate entities thus discovered from the perspective of their possible psychological or sociological meaning. I would like, now, to discuss how such assessment could be carried out.

Chapters 5 and 6 present approaches to identify topic model topics that, respectively, correspond to psychological features measured in survey data or are used with some exclusivity by particular groups (and hence candidates for social representations or entities[1] particular to those groups). Before stronger conclusions about the social and/or psychological significance of those topics can be

[1] In the sense of social representation theory — see Section 2.1.1

drawn, a number of checks are in order. Primarily, a sample of texts in which the candidate topics are prominent (preferably including some contextual information such as other texts related by time/group/conversation etc...) should be assessed by domain experts to ascertain first that they are not just some statistical anomaly to which meaning is difficult to ascribe, and second that some meaning of interest and significance can be identified. Assessing candidate topics using topic quality criteria such as those discussed in Section 2.4.2 may also be prudent, especially in cases where there are a large number of candidates and poor quality topics may be discarded without expensive expert appraisal.

I would like to stress the importance of such direct appraisals when using statistical tools to study social and psychological phenomena, as the outputs of complex inference tools such as those presented here may not always represent the features you might expect after casual evaluation.

The primary benefit of identifying topics that after appraisal appear to indicate particular social entities or psychological features is one of scale — expert appraisal of a relatively small sample of texts allows one to draw conclusions about a much larger number of texts and their associated authors and groups of authors. Given richly dynamic data such as has been collected in Chapter 3, one may investigate the dynamics of how such entities and features are adopted and transmitted and how they effect dynamic group processes such as acquiring and losing members, evolution of group identities etc...

7.2 Further Work

In each chapter I presented opportunities for further work specifically relevant to the content and goals of that chapter. I will not reiterate those opportunities here, but focus instead on the larger context of understanding and modelling social processes operating through online social media.

The focus of this thesis has been on collection of socially rich data and the development of novel technical approaches to identify social contexts and psychological markers in socially generated text and network data. This work was motivated by the need to develop empirical grounding for models of social processes such as the formation and evolution of social norms, identities and other socially transmitted and maintained entities. The work has proceeded on the premise that such entities can be detected through contextual patterns in social media data representing group communications.

Sociology and Psychology Research with Existing Tools and Data: A substantial body of data was collected from the pro-anorexia and eating disorder community operating on Twitter (Chapter 4.5). This data is itself a rich and largely un-utilised resource — there are many avenues to further analysis in relatively conventional ways that promise to reveal interesting and provocative observations. A few examples include:

- A general investigation of topics of discussion within the community would be of great interest to the research community investigating the psychology of eating disorders and clinical psychology of their treatment.
- A deeper analysis of communities represented in the data would be both of great interest (particularly when linked to discussion topics) and would allow further verification that the data collection approach successfully identified such communities as well as quantify the proportion of data not attributable to them.
- The investigation of processes of network dynamics (e.g.: involving link formation and dissolution) and factors that influence these processes.
- In Section 4.4 I observed an anomalous increase in the extracted network size related to an on-line discussion by the American ABC News and it is clear that other external events have had a significant impact on the data and community. Those events and the communities response are of considerable interest in their own right. A fruitful avenue for research with data such as this would be to link it to events either manually identified or automatically detected in other more traditional media (such as news or celebrity antics). Linking community responses to attempted interventions such as government advertising intended to change community behaviour would also be of particular interest.

Further Development of Socio-Metric Tools and Rich Social Data Further application of the technical approaches from Chapters 5 and 6 is also a fruitful avenue of further research. This could include:

- Chapter 6 draws on existing topic model and community detection algorithms, linking topics to communities. Using the approach on weekly network snapshots to track the progress of detected topics (and the social entities they may represent) linked to the coming and going of individual

users from groups. This in itself represents a substantial research agenda with many challenges and opportunities including:

  - collection of more data sets with rich dynamic information, and in particular the possibility of obtaining annotated or community-tagged data where those annotations/tags can operate as a kind of ground truth.
  - developing further arguments and methods to establish the link between lexical contexts and specific social entities and their roles.
  - experimenting with existing community detection and topic modelling algorithms for these models and potentially developing new innovative combined topic-community or task-specific models. A great many topic model variants and community detection algorithms exist, many of which have potential for application in this context.
  - investigating links between social roles (as identified by topic usage, network features, influence patterns etc...) and user profile characteristics and dynamics.

- Applying detected dynamics of social entities as grounding for relevant agent based models (for example the model of [Van Rooy et al. 2014]). This also is a substantial research agenda requiring carefully thought out arguments and examples to establish the probable links between measured, modelled and actual social entities as well as the model suitability to this context and interpretation of model predictions.

- Designing and implementing further medium to large scale surveys linking psychometric questionnaires to text features (along the lines of [Dann 2011]) targeted at psychological qualities relevant to models of social processes. This could be an important step toward establishing the validity of social entity detection methods presented here.

- In Section 4.6 I mentioned the potential to extract features from the substantial number of images collected in the pro-anorexia and eating disorder data. Social media data in general often contains images and video. Identifying objects, extracting meaningful context and measuring human affect in such data is already an active area of research. Such detected entities features and emotions would be useful additions to social media text data

for the detection of social entities. Adaptation and innovation in such computer vision and audio analysis techniques tailored to the detection of social entities also represents an interesting research direction.

Caveats: I have presented a bold and broad research agenda related to the motivating vision of this thesis and extending its contribution. In research such as this, where the object under study is difficult to define and measure, it is important to realise the limitations and possible pitfalls associated with the conclusions and conjectures that can be attained. I wish to discuss some of these general issues.

Social media data represents a subset of the communications of a subset of the society. Even if social media usage were to become nearly ubiquitous across all ages and social strata, there will always be an element that does not participate fully or at all, and there will always be communications that are not detectable, either offline or private. One must, therefore, be careful to generalise any conclusions made from data such as that presented in this thesis. The grand vision of understanding how societies make decisions must be approached with caution, with due care and attention given to the missing data and its potential impact on any conclusions drawn.

Another way in which the data may be incomplete comes from the way in which context effects the expression of social entities (norms, identities, symbols etc...). Social entities such as norms and identities are activated by social and situational contexts. If those contexts do not arise during data collection, the entities will not be expressed. This is a double edged sword: it can help isolate particular entities, but some may not be well represented and/or may be dormant. Extrapolation into situations and contexts where those entities are more strongly activated may fail, and perhaps in a spectacular way.

We must be wary of "dirty data". It is impossible to verify that all texts in a large corpus are genuine. Advertising and spam, as well as errors made by data collection software may result in spurious texts and other features existing in the corpus. Though these will typically have distinct styles and will likely be discernible from the features of interest, and though the data discussed in this thesis appears to have low levels of spam (see Section 4.6), they may still infect our results. Verification of important results through other media or methods should be undertaken, despite any apparent large cost of such work.

7.3 Vision

This thesis is built on the vision of direct measurement of social processes as verification and an empirical grounding for computational and numerical models of those processes. With such grounding, it may, in time, be possible to build models capable of tracking and even predicting the evolution and development of the modelled social entities.

The application of such models has a substantial range of scope, from, for example, understanding more about social aspects of eating disorders and thus helping develop ways to treat and prevent them, to grand models that attempt to capture socio-economic processes of whole societies, probably build on vast data collection and processing efforts.

Though one must be careful about the scope of the information that can be collected and the limits to what it is possible to glean from it, such a grounding could ideally lead to prediction of some large scale phenomena, helping guide us as we make decisions about how to run our societies and move into the future.

This vision is a profoundly difficult project and will likely require generations of researchers to attain, even in part. Likely also it will change as it develops, becoming something quite different in method, purpose and application to what I present here. Nonetheless, it is a vision well worth pursuing, with the potential for great benefits along the way.

One of my greatest fears in work such as this is the potential for it to be turned to evil. For it to be used to manipulate people and societies to the benefit of a few and detriment of many, that things we hold of great value, such as personal freedoms, may be eroded by such application. I maintain hope in the face of these fears, however. I liken social forces to a tide. You can attempt to hold it back, but in time it will always overcome, and with violence if held too strongly. My hope is that this vision, were it to become reality, would show this to be the case, and would form the basis of a philosophy and politic in which mutual benefit and the good of all would be paramount.

Bibliography

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. 2009. Mixed membership stochastic blockmodels. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou Eds., *Advances in Neural Information Processing Systems 21* (2009), pp. 33–40. Curran Associates, Inc. (pp. 20, 106)

Aletras, N., Baldwin, T., Lau, J. H., and Stevenson, M. 2014. Representing topics labels for exploring digital libraries. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '14 (Piscataway, NJ, USA, 2014), pp. 239–248. IEEE Press. (p. 37)

Aletras, N. and Stevenson, M. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)-Long Papers* (2013), pp. 13–22. (p. 32)

AlSumait, L., Barbará, D., Gentle, J., and Domeniconi, C. 2009. Topic significance ranking of lda generative models. In W. Buntine, M. Grobelnik, D. Mladenić, and J. Shawe-Taylor Eds., *Machine Learning and Knowledge Discovery in Databases*, Number 5781 in Lecture Notes in Computer Science (2009), pp. 67–82. Springer Berlin Heidelberg. (pp. 28, 35)

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., and Zhu, M. 2012. A practical algorithm for topic modeling with provable guarantees. *arXiv:1212.4777 [cs, stat]*. (p. 24)

Asuncion, A., Welling, M., Smyth, P., and Teh, Y. W. 2009. On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09 (Arlington, Virginia, United States, 2009), pp. 27–34. AUAI Press. (p. 28)

Bauer, M. W. and Gaskell, G. 1999. Towards a paradigm for research on social representations. *Journal for the Theory of Social Behaviour 29*, 2, 163–186. (pp. 5, 10)

Bhattachayya, A. 1943. On a measure of divergence between two statistical population defined by their population distributions. *Bulletin Calcutta Mathematical Society 35*, 99–109. (p. 110)

Biernacki, P. and Waldorf, D. 1981. Snowball sampling: Problems and techniques of chain referral sampling. *Sociological Methods & Research 10*, 2, 141–163. (p. 20)

Billig, M. 1991. *Ideology and Opinions: Studies in Rhetorical Psychology.* SAGE. (pp. 5, 10)

Blei, D., Griffiths, T., Jordan, M., and Tenenbaum, J. 2004. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems 16*. Cambridge, MA: MIT Press. (p. 45)

Blei, D. and Lafferty, J. 2006a. Correlated topic models. *Advances in neural information processing systems 18*, 147–154. (p. 45)

Blei, D. M. 2012. Probabilistic topic models. *Commun. ACM 55*, 4, 77–84. (p. 27)

Blei, D. M., Griffiths, T. L., and Jordan, M. I. 2010. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *J. ACM 57*, 2, 7:1–7:30. (p. 45)

Blei, D. M. and Jordan, M. I. 2003. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (2003), pp. 127–134. (pp. 27, 46)

Blei, D. M. and Lafferty, J. D. 2006b. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning* (2006), pp. 113–120. (p. 40)

Blei, D. M. and Lafferty, J. D. 2009. Topic models. In *Text mining: classification, clustering, and applications* (2009), pp. 71–94. CRC Press. (p. 36)

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research 3*, 993–1022. (pp. 23, 24, 26, 28, 29, 105)

Bollen, J., Pepe, A., and Mao, H. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Fifth International AAAI Conference on Weblogs and Social Media* (may, 2011). (p. 14)

Bouma, G. 2009. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference* (2009), pp. 31–40. (p. 33)

Boyd-Graber, J. L., Blei, D. M., and Zhu, X. 2007. A topic model for word sense disambiguation. In *EMNLP-CoNLL* (2007), pp. 1024–1033. Citeseer. (p. 30)

Bradford, R. B. 2008. An empirical study of required dimensionality for large-scale latent semantic indexing applications. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08 (New York, NY, USA, 2008), pp. 153–162. ACM. (p. 23)

Bruns, A. and Burgess, J. E. 2011. The use of twitter hashtags in the formation of ad hoc publics. In *Proceedings of the 6th European Consortium for Political Research (ECPR) General Conference 2011* (University of Iceland, Reykjavik, 2011). (p. 16)

Budak, C. and Agrawal, R. 2013. On participation in group chats on twitter. In *Proceedings of the 22nd international conference on World Wide Web* (2013), pp. 165–176. (p. 16)

Buntine, W. and Mishra, S. 2014. Experiments with non-parametric topic models. In *20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (New York, USA, 2014). (p. 42)

Casilli, A. A., Tubaro, P., and Araya, P. 2012. Ten years of ana: Lessons from a transdisciplinary body of literature on online pro-eating disorder websites. *Social Science Information 51*, 1, 120–139. (p. 17)

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, P. K. 2010. Measuring user influence in twitter: The million follower fallacy. *ICWSM 10*, 10-17, 30. (p. 20)

Chang, J. and Blei, D. M. 2009. Relational topic models for document networks. In *International Conference on Artificial Intelligence and Statistics* (2009), pp. 81–88. (p. 43)

Chang, J. and Blei, D. M. 2010. Hierarchical relational models for document networks. *The Annals of Applied Statistics 4*, 1, 124–150. (p. 43)

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., and Blei, D. M. 2009. Reading tea leaves: How humans interpret topic models. In *Advances*

in neural information processing systems (2009), pp. 288–296. (pp. 29, 32, 33)

Chen, C., Du, L., and Buntine, W. 2011. Sampling table configurations for the hierarchical poisson-dirichlet process. In D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis Eds., *Machine Learning and Knowledge Discovery in Databases*, Number 6911 in Lecture Notes in Computer Science (2011), pp. 296–311. Springer Berlin Heidelberg. (p. 42)

Chen, N., Zhu, J., Xia, F., and Zhang, B. 2013. Generalized relational topic models with data augmentation. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence* (2013), pp. 1273–1279. AAAI Press. (p. 43)

Chuang, J., Gupta, S., Manning, C., and Heer, J. 2013. Topic model diagnostics: Assessing domain relevance via topical alignment. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)* (2013), pp. 612–620. (pp. 27, 28, 29, 31, 35)

Chuang, J., Manning, C. D., and Heer, J. 2012. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (2012), pp. 74–77. ACM. (pp. 37, 92)

Chung, C. K. and Pennebaker, J. 2013. Using computerized text analysis to track social processes. *Handbook of language and social psychology. New York: Oxford*. (pp. 12, 39)

Clark, J. M. and Paivio, A. 2004. Extensions of the paivio, yuille, and madigan (1968) norms. *Behavior Research Methods, Instruments, & Computers 36*, 371–383. (p. 13)

Cohen, D. 2007. The worrying world of eating disorder wannabes. *BMJ 335*, 7618, 516–516. (p. 17)

Cohen, R., Aviram, I., Elhadad, M., and Elhadad, N. 2014. Redundancy-aware topic modeling for patient record notes. *PLoS ONE 9*, 2, e87555. (pp. 39, 46)

Cohen, R., Elhadad, M., and Elhadad, N. 2013. Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. *BMC Bioinformatics 14*, 1, 10. (p. 39)

Conover, M. D., Goncalves, B., Ratkiewicz, J., Flammini, A., and Menczer, F. 2011. Predicting the political alignment of twitter users. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom)* (2011), pp. 192–199. (p. 21)

Cook, J., Kenthapadi, K., and Mishra, N. 2013. Group chats on twitter. In *Proceedings of the 22nd international conference on World Wide Web* (2013), pp. 225–236. (p. 16)

Csipke, E. and Horne, O. 2007. Pro-eating disorder websites: users' opinions. *European Eating Disorders Review 15*, 3, 196–206. (p. 17)

Dann, E. 2011. *The Thin Ideal, Female Identity and Self-Worth: An Exploration of Language Use*. PhD thesis, Department of Psychology, The Australian National University. (pp. 84, 87, 95, 122, 151, 158)

Davidov, D., Tsur, O., and Rappoport, A. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10 (Stroudsburg, PA, USA, 2010), pp. 241–249. Association for Computational Linguistics. (p. 14)

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. 1990. Indexing by latent semantic analysis. *JASIS 41*, 6, 391–407. (p. 23)

Derenne, D. J. L. and Beresin, D. E. V. 2006. Body image, media, and eating disorders. *Academic Psychiatry 30*, 3, 257–261. (p. 17)

Dias, K. 2013. The ana sanctuary: Women's pro-anorexia narratives in cyberspace. *Journal of International Women's Studies 4*, 2, 31–45. (p. 17)

DiMaggio, P., Nag, M., and Blei, D. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of u.s. government arts funding. *Poetics 41*, 6, 570–606. (p. 11)

Doyle, G. and Elkan, C. 2009. Accounting for burstiness in topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09 (New York, NY, USA, 2009), pp. 281–288. ACM. (p. 42)

Du, L., Buntine, W., and Jin, H. 2012. Modelling sequential text with an adaptive topic model. In *Proceedings of the 2012 Joint Conference on Empirical*

Methods in Natural Language Processing and Computational Natural Language Learning (Jeju /Korea, 2012), pp. 535–545. (p. 46)

Duan, D., Li, Y., Li, R., Lu, Z., and Wen, A. 2011. Mei: mutual enhanced infinite generative model for simultaneous community and topic detection. In *Discovery Science* (2011), pp. 91–106. Springer. (pp. 44, 100)

Dunbar, R. I. M. 1992. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution 22*, 6, 469–493. (p. 12)

Dunbar, R. I. M. 1993. Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences 16*, 04, 681–694. (p. 12)

Dunbar, R. I. M. 2012. The social brain meets neuroimaging. *Trends in Cognitive Sciences 16*, 2, 101–102. (p. 12)

Dunbar, R. I. M., Arnaboldi, V., Conti, M., and Passarella, A. 2015. The structure of online social networks mirrors those in the offline world. *Social Networks 43*, 39–47. (p. 12)

Eisenstein, J. 2013. What to do about bad language on the internet. In *HLT-NAACL* (2013), pp. 359–369. (p. 15)

Erosheva, E., Fienberg, S., and Lafferty, J. 2004. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences 101*, suppl 1, 5220–5227. (p. 42)

Fei-Fei, L. and Perona, P. 2005. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Volume 2 (2005), pp. 524–531. (p. 30)

Fillmore, C. 1982. Frame semantics. *Linguistics in the Morning Calm*, 111–137. (p. 11)

Fillmore, C. J. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences 280*, Origins and Evolution of Language and Speech, 20–32. (p. 11)

Firth, J. 1957. A synopsis of linguistic theory, 1930-1955. In *Studies in Linguistic Analysis* (1957), pp. 1–32. Blackwell. (p. 11)

Fortunato, S. 2010. Community detection in graphs. *Physics Reports 486*, 3-5, 75–174. (p. 19)

Fortunato, S. 2012. Community detection in networks. (p. 18)

Fortunato, S., Macy, M., and Redner, S. 2013. Editorial: Statistical mechanics and social sciences. *Journal of Statistical Physics 151*, 1-2, 1–8. (p. 18)

Garcia, D. and Schweitzer, F. 2011. Emotions in product reviews-empirics and models. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom)* (2011), pp. 483–488. (p. 3)

Gelman, A., Meng, X.-L., and Stern, H. 1996. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica 6*, 4, 733–760. (p. 34)

Geman, S. and Geman, D. 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-6*, 6, 721–741. (p. 26)

Go-Globe. 2013. Social media in china - statistics and trends [infographic]. (p. 4)

Gonçalves, B., Perra, N., and Vespignani, A. 2011. Modeling users' activity on twitter networks: Validation of dunbar's number. *PLoS ONE 6*, 8. (pp. 12, 80)

Gopalan, P. K. and Blei, D. M. 2013. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences 110*, 36, 14534–14539. (pp. 20, 106, 114)

Griffiths, T. L. and Steyvers, M. 2003. Prediction and semantic association. In S. Becker, S. Thrun, and K. Obermayer Eds., *Advances in Neural Information Processing Systems 15* (2003), pp. 11–18. MIT Press. (pp. 23, 27, 30)

Griffiths, T. L. and Steyvers, M. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America 101*, Suppl 1, 5228–5235. (pp. 24, 26, 30, 40, 104)

Griffiths, T. L., Steyvers, M., Blei, D. M., and Tenenbaum, J. B. 2004. Integrating topics and syntax. In *Advances in neural information processing systems* (2004), pp. 537–544. (pp. 27, 30, 46)

Grimmer, J., Shorey, R., Wallach, H., and Zlotnick, F. 2011. A class of bayesian semiparametric cluster-topic models for political texts. Technical report. (p. 35)

Hall, D., Jurafsky, D., and Manning, C. D. 2008. Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08 (Stroudsburg, PA, USA, 2008), pp. 363–371. Association for Computational Linguistics. (pp. 30, 40, 44, 97)

Harris, Z. S. 1954. Distributional structure. *extlessi extgreaterWORD extless/i extgreater 10*, 2-3, 146–162. (p. 11)

Hassan, S., Antunes, L., Pavon, J., and Gilbert, G. N. 2008. Stepping on earth: A roadmap for data-driven agent-based modelling. In *Proceedings of the 5th Conference of the European Social Simulation Association (ESSA08).* (2008). (p. 3)

Himelboim, I. 2014. Mapping twitter topic networks: From polarized crowds to community clusters. Blog post, Pew Research Center's Internet & American Life Project. (pp. 16, 100)

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. 2010. *Bayesian Nonparametrics*. Cambridge University Press. (p. 42)

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99 (New York, NY, USA, 1999), pp. 50–57. ACM. (p. 23)

Homewood, J. and Melkonian, M. 2015. What factors account for internalisation of the content of pro-ana websites. *Journal of Neurology, Neurosurgery & Psychiatry 86*, 9, e3–e3. (p. 17)

Hong, L. and Davison, B. D. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10 (New York, NY, USA, 2010), pp. 80–88. ACM. (pp. 15, 16, 88)

Hörster, E., Lienhart, R., and Slaney, M. 2007. Image retrieval on large-scale image databases. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, CIVR '07 (New York, NY, USA, 2007), pp. 17–24. ACM. (p. 27)

Huberman, B. A., Romero, D. M., and Wu, F. 2008. Social networks that matter: Twitter under the microscope. SSRN Scholarly Paper ID 1313405, Social Science Research Network, Rochester, NY. (pp. 16, 20, 100)

Hutto, C., Yardi, S., and Gilbert, E. 2013. A longitudinal study of follow predictors on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13 (New York, NY, USA, 2013), pp. 821–830. ACM. (pp. 50, 54)

Jagarlamudi, J., Daume III, H., and Udupa, R. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (Avignon, France, 2012), pp. 204–213. Association for Computational Linguistics. (pp. 27, 44, 97)

Java, A., Song, X., Finin, T., and Tseng, B. 2007. Why we twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, WebKDD/SNA-KDD '07 (New York, NY, USA, 2007), pp. 56–65. ACM. (pp. 16, 100)

Jett, S., LaPorte, D. J., and Wanchisn, J. 2010. Impact of exposure to pro-eating disorder websites on eating behaviour in college women. *European Eating Disorders Review 18*, 5, 410–416. (p. 17)

Jockers, M. L. and Mimno, D. 2013. Significant themes in 19th-century literature. *Poetics 41*, 6, 750–769. (p. 35)

Jones, K. 2013. Growth of social media infographic. (p. 4)

Juarascio, A. S., Shoaib, A., and Timko, C. A. 2010. Pro-eating disorder communities on social networking sites: A content analysis. *Eating Disorders 18*, 5, 393–407. (p. 18)

Jurgens, D. 2013. That's what friends are for: Inferring location in online social media platforms based on social relationships. *ICWSM 13*, 273–282. (p. 21)

K. Burton, N. Kasch, and I. Soboroff. 2011. The ICWSM 2011 spinn3r dataset. In *Proceedings of the Fifth Annual Conference on Weblogs and Social Media (ICWSM 2011)* (Barcelona, Spain, jul, 2011). (p. 15)

Katz, S. M. 1996. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering 2*, 01, 15–59. (p. 42)

Killworth, P. D., Bernard, H. R., McCarty, C., Doreian, P., Goldenberg, S., Underwood, C., Harries-Jones, P., Keesing, R. M., Skvoretz, J., and Wemegah, M. V. S. 1984. Measuring patterns of

acquaintanceship [and comments and reply]. *Current Anthropology 25*, 4, 381–397. (p. 12)

Krishnamurthy, B., Gill, P., and Arlitt, M. 2008. A few chirps about twitter. In *Proceedings of the First Workshop on Online Social Networks*, WOSN '08 (New York, NY, USA, 2008), pp. 19–24. ACM. (p. 80)

Kullback, S. and Leibler, R. A. 1951. On information and sufficiency. *The Annals of Mathematical Statistics 22*, 1, 79–86. (pp. 37, 110)

Kwak, H., Chun, H., and Moon, S. 2011. Fragile online relationship: A first look at unfollow dynamics in twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11 (New York, NY, USA, 2011), pp. 1091–1100. ACM. (p. 21)

Kwak, H., Lee, C., Park, H., and Moon, S. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web* (2010), pp. 591–600. (pp. 20, 80)

Lü, L. and Zhou, T. 2011. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications 390*, 6, 1150–1170. (p. 21)

Lau, J. H., Baldwin, T., and Newman, D. 2013. On collocations and topic models. *ACM Trans. Speech Lang. Process. 10*, 3, 10:1–10:14. (p. 46)

Lau, J. H., Newman, D., and Baldwin, T. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the European Chapter of the Association for Computational Linguistics* (2014). (p. 33)

Li, A. Q., Ahmed, A., Ravi, S., and Smola, A. J. 2014. Reducing the sampling complexity of topic models (2014). pp. 891–900. ACM Press. (p. 114)

Li, D., Ding, Y., Shuai, X., Bollen, J., Tang, J., Chen, S., Zhu, J., and Rocha, G. 2012. Adding community and dynamic to topic models. *Journal of Informetrics 6*, 2, 237–253. (pp. 41, 44, 100)

Li, D., He, B., Ding, Y., Tang, J., Sugimoto, C., Qin, Z., Yan, E., Li, J., and Dong, T. 2010. Community-based topic modeling for social tagging. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10 (New York, NY, USA, 2010), pp. 1565–1568. ACM. (pp. 100, 101)

Lim, K. W. and Buntine, W. 2014a. Bibliographic analysis with the citation network topic model (2014). (p. 44)

Lim, K. W. and Buntine, W. 2014b. Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon (2014). pp. 1319–1328. ACM Press. (p. 43)

Lim, K. W., Chen, C., and Buntine, W. 2013. Twitter-network topic model: A full bayesian treatment for social network and text modeling. In *NIPS2013* (2013). (pp. 16, 43, 45)

Lin, H., Jia, J., Guo, Q., Xue, Y., Huang, J., Cai, L., and Feng, L. 2014. Psychological stress detection from cross-media microblog data using deep sparse neural network. In *2014 IEEE International Conference on Multimedia and Expo (ICME)* (2014), pp. 1–6. (p. 16)

Liu, Y., Niculescu-Mizil, A., and Gryc, W. 2009. Topic-link lda: joint models of topic and author community. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09 (Montreal, Quebec, Canada, 2009), pp. 665–672. ACM. (p. 43)

Lyons, E. J., Mehl, M. R., and Pennebaker, J. W. 2006. Pro-anorexics and recovering anorexics differ in their linguistic internet self-presentation. *Journal of Psychosomatic Research 60*, 3, 253–256. (p. 18)

Madsen, R. E., Kauchak, D., and Elkan, C. 2005. Modeling word burstiness using the dirichlet distribution. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05 (New York, NY, USA, 2005), pp. 545–552. ACM. (p. 42)

McCallum, A. K. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu. (p. 31)

McCarty, C., Killworth, P. D., Bernard, H. R., Johnsen, E. C., and Shelley, G. A. 2001. Comparing two methods for estimating network size. *Human Organization 60*, 1, 28–39. (p. 12)

McColl, G. 2013. Anorexia underworld. *The Age*. (p. 17)

McDonald, S. and Ramscar, M. 2001. Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (2001), pp. 611–616. (pp. 5, 11)

McNair, D., Lorr, M., and Droppleman, L. 1981. *Profile of Mood States, POMS*. EdiTS, Educational and Industrial Testing Service. (p. 13)

Mehrotra, R., Sanner, S., Buntine, W., and Xie, L. 2013. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13 (New York, NY, USA, 2013), pp. 889–892. ACM. (pp. 16, 88)

Miller, G. A. 1995. Wordnet: A lexical database for english. *Commun. ACM 38*, 11, 39–41. (p. 33)

Miller, R. G., Gong, G., and Muñoz, A. 1981. *Survival analysis*. Wiley series in probability and mathematical statistics. Wiley, New York. (p. 73)

Mimno, D. 2012a. The details: Training and validating big models on big data. (p. 31)

Mimno, D. 2012b. Topic modeling workshop: Mimno. http://vimeo.com/53080123. (p. 32)

Mimno, D. and Blei, D. 2011. Bayesian checking for topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11 (Stroudsburg, PA, USA, 2011), pp. 227–237. Association for Computational Linguistics. (pp. 34, 84, 85)

Mimno, D. and McCallum, A. 2008. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Uncertainty in Artificial Intelligence* (2008). (p. 46)

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., and McCallum, A. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2* (2009), pp. 880–889. Association for Computational Linguistics. (p. 46)

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11 (Stroudsburg, PA, USA, 2011), pp. 262–272. Association for Computational Linguistics. (pp. 27, 31, 32)

Mohammad, S. M. and Alm, C. O. 2015. Tutorial on computational analysis of affect and emotion in language. In *Proceedings of the International Con-*

ference on Empirical Methods in Natural Language Processing (2015). (p. 14)

Mohammad, S. M. and Kiritchenko, S. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence 31*, 2, 301–326. (p. 14)

Musat, C. C., Velcin, J., Trausan-Matu, S., Rizoiu, M.-A., et al. 2011. Improving topic evaluation using conceptual knowledge. In *IJCAI* (2011), pp. 1866–1871. (p. 33)

Myers, S. A. and Leskovec, J. 2014. The bursty dynamics of the twitter information network (2014). pp. 913–924. ACM Press. (pp. 50, 54, 64, 71)

Nafstad, H. E. and Blakar, R. M. 2012. Ideology and social psychology: Ideology and social psychology. *Social and Personality Psychology Compass 6*, 4, 282–294. (p. 10)

Nallapati, R. M., Ahmed, A., Xing, E. P., and Cohen, W. W. 2008. Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08 (New York, NY, USA, 2008), pp. 542–550. ACM. (p. 44)

Nepusz, T., Petróczi, A., and Bazsó, F. 2007. Fuzzy clustering and the concept of bridgedness in social networks. In *Proceedings of the International Workshop and Conference on Network Science, NetSci*, Volume 598 (2007). (p. 107)

Newman, D., Bonilla, E., and Buntine, W. 2011. Improving topic coherence with regularized topic models. In *Advances in Neural Information Processing Systems 24* (2011), pp. 496–504. (pp. 45, 84, 87, 88, 91)

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10 (Stroudsburg, PA, USA, 2010), pp. 100–108. Association for Computational Linguistics. (pp. 32, 35)

Newman, M. E. J. 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences 103*, 23, 8577–8582. (p. 19)

Newman, M. E. J. and Park, J. 2003. Why social networks are different from other types of networks. *Physical Review E 68*, 3, 036122. (p. 20)

O'Connor, B., Krieger, M., and Ahn, D. 2010. Tweetmotif: Exploratory search and topic summarization for twitter. In *ICWSM* (2010).

(p. 39)

O'Connor, B., Stewart, B. M., and Smith, N. A. 2013. Learning to extract international relations from political context. In *ACL (1)* (2013), pp. 1094–1104. (pp. 5, 27, 45)

Ou-Yang, L., Dai, D.-Q., Li, X.-L., Wu, M., Zhang, X.-F., and Yang, P. 2014. Detecting temporal protein complexes from dynamic protein-protein interaction networks. *BMC Bioinformatics 15*, 1, 335. (p. 18)

Page, L., Brin, S., Motwani, R., and Winograd, T. 1999. The pagerank citation ranking: bringing order to the web. (p. 20)

Paivio, A., Yuille, J. C., and Madigan, S. A. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology 76*, 1, Pt.2, 1–25. (p. 13)

Paul, M. J. and Dredze, M. 2011. You are what you tweet: Analyzing twitter for public health. In *ICWSM* (2011). (p. 16)

Pennebaker, J. W. and Lay, T. C. 2002. Language use and personality during crises: Analyses of mayor rudolph giuliani's press conferences. *Journal of Research in Personality 36*, 3 (jun,), 271–282. (p. 1)

Perina, A., Lovato, P., Murino, V., and Bicego, M. 2010. Biologically-aware latent dirichlet allocation (BaLDA) for the classification of expression microarray. In T. M. H. Dijkstra, E. Tsivtsivadze, E. Marchiori, and T. Heskes Eds., *Pattern Recognition in Bioinformatics*, Number 6282 in Lecture Notes in Computer Science, pp. 230–241. Springer Berlin Heidelberg. Cited by 0008. (p. 27)

Pitman, J. and Yor, M. 1997. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability 25*, 2, 855–900. (p. 42)

Poldrack, R. A., Mumford, J. A., Schonberg, T., Kalar, D., Barman, B., and Yarkoni, T. 2012. Discovering relations between mind, brain, and mental disorders using topic mapping. *PLoS Computational Biology 8*, 10, e1002707. (pp. 27, 30, 38)

Pritchard, J. K., Stephens, M., and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics 155*, 2, 945–959. (pp. 24, 28)

Pruteanu-Malinici, I., Ren, L., Paisley, J., Wang, E., and Carin, L. 2010. Hierarchical bayesian modeling of topics in time-stamped documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*, 6, 996–1011. (p. 41)

Radke, B. R. 2003. A demonstration of interval-censored survival analysis. *Preventive Veterinary Medicine 59*, 4, 241–256. (p. 73)

Rainie, L. 2014. The six types of twitter conversations. (p. 50)

Ramage, D., Dumais, S. T., and Liebling, D. J. 2010. Characterizing microblogs with topic models. In *ICWSM*, Volume 10 (2010), pp. 1–1. (p. 16)

Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (Singapore, 2009), pp. 248–256. Association for Computational Linguistics. (pp. 35, 45, 97)

Ramage, D., Manning, C. D., and Dumais, S. 2011. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11 (New York, NY, USA, 2011), pp. 457–465. ACM. (pp. 45, 97)

Rand, D. G., Arbesman, S., and Christakis, N. A. 2011. Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences 108*, 48, 19193–19198. (p. 21)

Röder, M., Both, A., and Hinneburg, A. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (2015). (p. 33)

Reaves, J. 2001. Anorexia goes high tech. *Time Magazine 31.* (p. 17)

Roberts, M. E., Stewart, B. M., and Airoldi, E. M. 2014. Structural topic models. Working paper. export bibtex tagged xml. (p. 46)

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04 (Arlington, Virginia, United States, 2004), pp. 487–494. AUAI Press. (p. 45)

Rosner, F., Hinneburg, A., Röder, M., Nettling, M., and Both, A. 2014. Evaluating topic coherence measures. *arXiv:1403.6397 [cs].* (p. 33)

Rosvall, M. and Bergstrom, C. T. 2007. An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences 104*, 18, 7327–7331. (p. 19)

Rouleau, C. R. and von Ranson, K. M. 2011. Potential risks of pro-eating disorder websites. *Clinical Psychology Review 31*, 4, 525–531. (p. 17)

Rubin, D. B. 1984. Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics 12*, 4, 1151–1172. (p. 34)

Ruths, D. and Pfeffer, J. 2014. Social media for large studies of behavior. *Science 346*, 6213, 1063–1064. (pp. 1, 15)

Salton, G. and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management 24*, 5, 513–523. (p. 36)

Sammut, G., Andreouli, E., Gaskell, G., and Valsiner, J. 2015. Social representations: a revolutionary paradigm? In G. Sammut, E. Andreouli, G. Gaskell, and J. Valsiner Eds., *The Cambridge Handbook of Social Representations* (Cambridge, UK, 2015), pp. 3–11. Cambridge University Press. (pp. 5, 10)

Sato, I. and Nakagawa, H. 2010. Topic models with power-law using pitman-yor process. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (2010), pp. 673–682. ACM. (p. 42)

Scott, J. 2012. *Social Network Analysis*. SAGE. (p. 18)

Sheldon, P., Grey, S. H., Vickery, A. J., and Honeycutt, J. M. 2015. An analysis of imagined interactions with pro-ana (anorexia) implications for mental and physical health. *Imagination, Cognition and Personality*, 0276236615587493. (p. 17)

Singer, S. and Nelder, J. 2009. Nelder-mead algorithm. *Scholarpedia 4*, 7, 2928. (p. 111)

Skowron, M., Pirker, H., Rank, S., Paltoglou, G., Ahn, J., and Gobron, S. 2011. No peanuts! affective cues for the virtual bartender. In *FLAIRS Conference* (2011). (p. 13)

Smucker, M. D., Allan, J., and Carterette, B. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings*

of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM '07 (New York, NY, USA, 2007), pp. 623–632. ACM. (p. 36)

Starbird, K. and Palen, L. 2011. Voluntweeters": Self-organizing by digital volunteers in times of crisis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11 (New York, NY, USA, 2011), pp. 1071–1080. ACM. (pp. 50, 51)

Sterckx, L., Demeester, T., Deleu, J., Mertens, L., and Develder, C. 2014. Assessing quality of unsupervised topics in song lyrics. In M. d. Rijke, T. Kenter, A. P. d. Vries, C. Zhai, F. d. Jong, K. Radinsky, and K. Hofmann Eds., *Advances in Information Retrieval*, Number 8416 in Lecture Notes in Computer Science (2014), pp. 547–552. Springer International Publishing. (p. 35)

Talley, E. M., Newman, D., Mimno, D., Ii, B. W. H., Wallach, H. M., Burns, G. A. P. C., Leenders, A. G. M., and McCallum, A. 2011. Database of nih grants using machine-learned categories and graphical clustering. *Nature Methods 8*, 6, 443–444. (pp. 28, 29, 30, 31, 38)

Tang, L. and Liu, H. 2010. Community detection and mining in social media. *Synthesis Lectures on Data Mining and Knowledge Discovery 2*, 1, 1–137. (p. 19)

Tausczik, Y. R. and Pennebaker, J. W. 2010a. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology 29*, 1 (mar,), 24 –54. (p. 13)

Tausczik, Y. R. and Pennebaker, J. W. 2010b. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology 29*, 1, 24 –54. (pp. 84, 86, 95, 119)

Teh, Y. and Jordan, M. 2009. Hierarchical bayesian nonparametric models with applications. *Bayesian Nonparametrics 28*, 158. (p. 42)

Teh, Y. W. 2006. A hierarchical bayesian language model based on pitman-yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44 (Stroudsburg, PA, USA, 2006), pp. 985–992. Association for Computational Linguistics. (pp. 41, 45)

Thelwall, M., Buckley, K., and Paltoglou, G. 2012. Sentiment

strength detection for the social web. *Journal of the American Society for Information Science and Technology 63*, 1, 163–173. (pp. 14, 39)

Thompson, S. K. 1990. Adaptive cluster sampling. *Journal of the American Statistical Association 85*, 412, 1050–1059. (pp. 50, 52)

Times, N. Y. pro-anorexia search results. (p. 17)

Tribune, C. Articles about anorexia - tribunedigital-chicagotribune. (p. 17)

Tubaro, P. and Casilli, A. A. 2010. Pro-ana and pro-mia social networks. the promises of qualitatively-informed agent-based modeling. (p. 18)

Van Rooy, D., Wood, I., and Tran, E. 2014. Modelling the emergence of shared attitudes from group dynamics using an agent-based model of social comparison theory. *Systems Research and Behavioral Science*. (p. 122)

Wahabzada, M., Kersting, K., Bauckhage, C., Roemer, C., Ballvora, A., Pinto, F., Rascher, U., Leon, J., and Ploemer, L. 2012. Latent dirichlet allocation uncovers spectral characteristics of drought stressed plants. *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*. (pp. 30, 45)

Wallach, H. M. 2006. Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06 (New York, NY, USA, 2006), pp. 977–984. ACM. (p. 46)

Wallach, H. M. 2008. *Structured Topic Models for Language*. PhD thesis, University of Cambridge, Cambridge, UK. (p. 28)

Wallach, H. M., Mimno, D., and Mccallum, A. 2009. Rethinking lda: Why priors matter. In *NIPS*, Volume 22 (2009), pp. 1973–1981. (pp. 28, 30)

Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. 2009. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09 (New York, NY, USA, 2009), pp. 1105–1112. ACM. (p. 29)

Wang, C. and Blei, D. M. 2009. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In *Advances in neural information processing systems* (2009), pp. 1982–1989. (p. 41)

Wang, E., Silva, J., Willett, R., and Carin, L. 2011. Dynamic relational topic model for social network analysis with noisy links. In *2011 IEEE Statistical Signal Processing Workshop (SSP)* (2011), pp. 497–500. (p. 43)

Wang, P., Xu, B., Wu, Y., and Zhou, X. 2014. Link prediction in social networks: the state-of-the-art. *Science China Information Sciences 58*, 1, 1–38. (p. 21)

Wang, X. and McCallum, A. 2006. Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06 (New York, NY, USA, 2006), pp. 424–433. ACM. (p. 40)

Wang, X., McCallum, A., and Wei, X. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Data Mining, IEEE International Conference on* (Los Alamitos, CA, USA, 2007), pp. 697–702. IEEE. (p. 46)

Wei, X. and Croft, W. B. 2006. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06 (New York, NY, USA, 2006), pp. 178–185. ACM. (p. 27)

Wei, X., Sun, J., and Wang, X. 2007. Dynamic mixture models for multiple time-series. In *IJCAI*, Volume 7 (2007), pp. 2909–2914. (p. 40)

Weng, J., Lim, E.-P., Jiang, J., and He, Q. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10 (New York, NY, USA, 2010), pp. 261–270. ACM. (p. 20)

WHO. 2015. International classification of diseases 10th revision (icd-10). (p. 94)

Windrum, P., Fagiolo, G., and Moneta, A. 2007. Empirical validation of agent-based models: Alternatives and prospects. *Journal of Artificial Societies and Social Simulation 10*, 2, 8. (p. 3)

Wise, S. 2014. *Using Social Media Content to Inform Agent-based Models for Humanitarian Crisis Response*. PhD thesis, George Mason University. (p. 3)

Wong, S.-M. J., Dras, M., and Johnson, M. 2012. Exploring adaptor grammars for native language identification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12 (Stroudsburg, PA, USA, 2012), pp. 699–709. Association for Computational Linguistics. (p. 39)

Wood, I. 2013. Recent advances and applications of probabilistic topic models. In *Proceedings of the 33rd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt 2013)*, Volume 1636 (Canberra, Australia, 2013), pp. 124–130. AIP Publishing. (p. 22)

Wood, I. 2015a. A case study of collecting dynamic social data: The pro-ana twitter community. *Australian Journal of Intelligent Information Processing Systems 14*, 3. (p. 49)

Wood, I. 2015b. Using topic models to measure social psychological characteristics in online social media. In *Social Computing, Behavioral-Cultural Modeling, and Prediction*, Volume 9021 of *Lecture Notes in Computer Science* (Washington, DC, USA, 2015), pp. 308–313. Springer International Publishing. (p. 83)

Wood, I. D. 2015c. Community topic usage in social networks. In *CIKM Workshop on Topic Modelling, Post Processing and Applications*, TM '15 (Melbourne, Australia, 2015), pp. 3–9. ACM. (p. 99)

Xie, J., Kelley, S., and Szymanski, B. K. 2013. Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Computing Surveys (CSUR) 45*, 4, 43:1–43:35. (p. 20)

Xu, B., Huang, Y., Kwak, H., and Contractor, N. 2013. Structures of broken ties: Exploring unfollow behavior on twitter. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13 (New York, NY, USA, 2013), pp. 871–876. ACM. (pp. 21, 50, 54, 71)

Yang, L., Sun, T., Zhang, M., and Mei, Q. 2012. We know what @you #tag: Does the dual role affect hashtag adoption? In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 (New York, NY, USA, 2012), pp. 261–270. ACM. (pp. 50, 51)

Yeshua-Katz, D. and Martins, N. 2013. Communicating stigma: The pro-ana paradox. *Health Communication 28*, 5, 499–508. (p. 17)

Appendix A

Hash Tag Sampling Summary

Common hash tags from initial sample with 3 or more occurences with tags used in eventual search query indicated.

Count	Tag	Search Query?
731	#thinspiration	In Search Query
366	#anorexia	
196	#thinspo	In Search Query
142	#proana	In Search Query
127	#skinny	
118	#bulimia	
111	#diet	
107	#eatingdisorder	
97	#weightloss	
73	#ednos	In Search Query
46	#promia	In Search Query
45	#fitspiration	
41	#thin	
40	#ThingsAnaTaughtMe	
25	#supermodelbody	
25	#fitspo	
24	#ana	
20	#eatingdisorders	
19	#perfect	
19	#motivation	

Continued on next page

Count	Tag	Search Query?
19	#edproblems	In Search Query
18	#selfharm	
18	#depression	
16	#recovery	
16	#perfection	
15	#ed	
14	#food	
13	#workout	
13	#obesity	
13	#beautiful	
13	#anorexic	
12	#weighgtloss	
12	#mia	
12	#inspiration	
11	#hipbones	In Search Query
11	#anasisters	In Search Query
10	#legs	
10	#bones	
9	#summer	
9	#StayStrong	
9	#LisaBalash	
9	#FitFluential	
8	#myself	
8	#ILoveAna	
8	#healthy	
8	#fitness	
8	#embers	
8	#cutting	
8	#bodyimage	
8	#bikini	
7	#suicide	
7	#obsession	
7	#NoOneLovesAFatty	

Continued on next page

Count	Tag	Search Query?
7	#mentalillness	
7	#depressed	
7	#bulimic	In Search Query
6	#weightlose	
6	#summerbody	
6	#scary	
6	#recoverywarriors	
6	#NYC	
6	#gym	
6	#Guilt	
6	#fat	
6	#empathy	
6	#dontgiveup	
6	#Bulemia	
6	#anxiety	
6	#abuse	
5	#instagram	
5	#how	
5	#health	
5	#flatstom	
5	#FattyGetsFit	
5	#cutter	
5	#cut	
5	#bingeeating	
5	#abdomen	
4	#TurnsSidewaysAndDissapears	
4	#thighgap	In Search Query
4	#teamfollowback	
4	#ribs	
4	#overit	
4	#OverallSkinnyCunt	
4	#nomorechunkymonkey	
4	#lonely	

Continued on next page

Count	Tag	Search Query?
4	#lifesucks	
4	#kristenstewart	
4	#I	
4	#goaway	
4	#girl	
4	#getfit	
4	#feel	
4	#EatSomeFood	
4	#eati	
4	#done	
4	#delusional	
4	#CBB	
4	#blog	
4	#black	
4	#binge	
4	#anamia	
4	#AnaFamily	
4	#alone	
4	#abcdiet	In Search Query
3	#yoga	
3	#Woah	
3	#sad	
3	#Reserach	
3	#purge	
3	#obsessed	
3	#mirror	
3	#justdoit	
3	#journal	
3	#fuck	
3	#freebook	
3	#foodjournal	
3	#fashion	
3	#EDrecovery	

Continued on next page

Count	Tag	Search Query?
3	#bu	
3	#anaprobz	
3	#anadiary	

Appendix B

Topic Salience Summaries

This appendix summarises logistic regression coefficients from [Dann 2011] (see Section 5.4) and their application to the example topic model presented in Section 5.7.2. A new LIWC variable "Influence" introduced in [Dann 2011] and used in the model is at the end of this appendix. This LIWC variable is intended to capture the influence of the "thin ideal" as portrayed in public media.

A table is presented for each topic with estimated probability of personal salience over gender salience greater than 0.9. The table header contains the topic number, the number of words assigned to the topic and the total number of assigned tokens (both words and non-words).

The table body shows the 25 words with the largest absolute regression coefficients (scores) and the sum of the coefficients for the remaining words (indicated by **low score words**). For each of these, the table shows the word itself, the words' contribution to the regression equation (score), its LIWC category(s) and the number of occurrences of the word assigned to the topic.

The table footer shows the sum of the regression coefficients, the total number of contributing word occurrences and the estimated probability of personal identity salience over gender identity salience. The probability is derived from the sum s of regression coefficients and the logistic models intercept r by the logistic function $P(\text{personal}) = \frac{1}{1+e^{-(s+r)}}$.

Regression Coefficients	
LIWC Variable	Coefficient
insight	27
ingestion	17
influence	-2
impersonal pronouns	-22
inclusion	-22
causality	-27
negations	-29
exclusion	-46
Intercept	3.27

Topic 1: 7052 Words, 103886 Tokens			
Word	Score	LIWC class	Count
fat	0.459773	ingest	198
feeling	0.077448	insight	21
relate	0.047944	insight	13
remember	0.044256	insight	12
bulimic	0.034831	ingest	15
snack	0.027865	ingest	12
feelings	0.025816	insight	7
DINNER	0.023221	ingest	10
breakfast	0.023221	ingest	10
memories	0.022128	insight	6
low score words	0.113782	ingest	49
low score words	0.084824	insight	31
low score words	0.000000	cause	8
low score words	-0.006010	ipron	2
low score words	-0.006010	incl	2
low score words	-0.016391	influence	5
roots	-0.003688	cause	1
FORCE	-0.003688	cause	1
rootbeer	-0.003688	cause	1
compelling	-0.003688	cause	1
NEGATIVE	-0.003961	negate	1
isnt	-0.003961	negate	1
imagen	-0.009835	influence	3
changed	-0.011064	cause	3
thats	-0.045076	ipron	15
perfection	-0.059008	influence	18
everything	-0.114192	ipron	38
society	-0.137686	influence	42
nobody	-0.153258	ipron, negate	22
that's	-0.162273	ipron	54
what	-0.216364	ipron	72
make	-0.446251	cause	121
sum	-0.437039		796
probability	0.944539		

Topic 4: 134087 Words, 171381 Tokens			
Word	Score	LIWC class	Count
diet	0.134517	ingest	1061
ate	0.119937	ingest	946
breakfast	0.113851	ingest	898
eat	0.109414	ingest	863
lunch	0.108019	ingest	852
dinner	0.107639	ingest	849
eating	0.105103	ingest	829
water	0.093946	ingest	741
food	0.070872	ingest	559
pizza	0.068463	ingest	540
eaten	0.061236	ingest	483
low score words	0.908530	ingest	7166
low score words	0.128469	insight	665
low score words	-0.001432	influence	8
low score words	-0.049129	excl	128
low score words	-0.068418	incl	417
low score words	-0.081033	negate	348
low score words	-0.103097	cause	539
low score words	-0.121876	ipron	724
because	-0.033627	cause	167
this	-0.036260	ipron	221
that	-0.036916	ipron	225
made	-0.044501	cause	221
haven't	-0.048879	negate	226
it's	-0.051519	ipron	314
out	-0.060707	incl	370
it	-0.074653	ipron	455
or	-0.081648	excl	238
no	-0.142959	negate	661
but	-0.156779	excl	457
with	-0.182941	incl	1115
not	-0.256736	negate, excl	459
just	-0.423680	excl	1235
and	-0.835458	incl	5092
sum	-0.762251		30073
probability	0.924826		

Topic 6: 9181 Words, 112109 Tokens			
Word	Score	LIWC class	Count
inspiration	0.169378	insight	59
lunch	0.106645	ingest	59
dinner	0.079532	ingest	44
breakfast	0.077725	ingest	43
feeling	0.045933	insight	16
snack	0.041574	ingest	23
inspirational	0.037321	insight	13
meal	0.028921	ingest	16
low score words	0.140989	ingest	78
low score words	0.100478	insight	84
low score words	0.000000	cause	49
perfectin	-0.002552	influence	1
fashionably	-0.002552	influence	1
modelling	-0.002552	influence	1
MODELO	-0.002552	influence	1
PERFECTNESS	-0.002552	influence	1
OBEY	-0.002871	cause	1
shant	-0.003083	negate	1
trendy	-0.005104	influence	2
perfectness	-0.005104	influence	2
foundation	-0.005742	cause	2
PICK	-0.008612	cause	3
intento	-0.011483	cause	4
somethingabitdarker	-0.014460	ipron, excl	2
models	-0.017863	influence	7
results	-0.043062	cause	15
vs	-0.092929	excl	19
no	-0.138756	negate	45
this	-0.224561	ipron	96
sum	0.248012		688
probability	0.971253		

Topic 8: 5124 Words, 59355 Tokens			
Word	Score	LIWC class	Count
SECRETS	0.026167	insight	6
remembering	0.017445	insight	4
cooking	0.013730	ingest	5
candy	0.010984	ingest	4
snacking	0.010984	ingest	4
lesson	0.008722	insight	2
LEARN	0.008722	insight	2
whiskey	0.005492	ingest	2
dietin	0.005492	ingest	2
swallows	0.005492	ingest	2
low score words	0.026167	insight	36
low score words	0.024713	ingest	9
low score words	0.000000	cause	30
everybody	-0.003554	ipron	1
controla	-0.004361	cause	1
independ	-0.004361	cause	1
intent	-0.004361	cause	1
bosss	-0.004361	cause	1
causedd	-0.004361	cause	1
COME	-0.007107	incl	2
launch	-0.008722	cause	2
therefor	-0.008722	cause	2
thingirl	-0.010661	ipron	3
tv	-0.011630	influence	3
PERFECTION	-0.015506	influence	4
led	-0.017445	cause	4
perfecta	-0.027136	influence	7
guide	-0.054272	influence	14
perfect	-0.116298	influence	30
sum	-0.167642		186
probability	0.957074		

Topic 13: 8023 Words, 149028 Tokens			
Word	Score	LIWC class	Count
TRIGGERING	-0.006731	cause	2
sum	-0.006731		2
probability	0.963220		

Words in new LIWC class "Influence" [Dann 2011] ("*" indicates continuation by arbitrary letters)
cultur*
educat*
encourag*
expect*
ideal*
image*
influenc*
information
magazine*
perfect*
pressur*
societ*
source*
television*
image
tv*
advert*
attitude*
bombard*
campaign*
fad*
trend*
fashion*
guide*
media*
misinformation
model*
online
portray*
stereotyp*
television

Appendix C

Grounded Analysis of Tweets

This appendix summarises the results of a grounded analysis of a sample of the pro-anorexia and eating disorder twitter data described in Section 4.5. The following table lists themes identified as present in the collected tweets and images.

- Motivation connected with:
  - recovery
  - illness
  - commitment
  - ambivalence
  - thinspiration
  - interpersonal validation and support
- Body ideals and relations with the body
  - body dissatisfaction
  - critical view of body
  - various – and shifting – ways in which the thin ideal is exemplified such as thigh gap etc.
- Control
  - control over weight/shape/eating/exercise
- Body change and checking/comparison/measurement of this

  - weight loss
  - restriction
  - compensatory behaviors:
    - fasting
    - exercise
    - purging
    - calories
    - weight checking
    - pinching

- Eating and Food and how it relates to
  - perfectionism and control
    - calories, fasting
    - binges
    - failure (to maintain diet)
  - negative affect and negative self-assessment
    - failure (to maintain diet)

- Self-worth and self-evaluation
  - in terms of body and weight/control
  - general self-criticism

- Affect
  - emotions in general
  - distress
  - negative affect:
    - as a trigger for engaging in eating disordered behaviours
    - as a response to not meeting self-imposed high standards of control

- Coping Behaviours
  - self-harm

- Relationships and support

  - both personal and at the level of the sites themselves
  - Seeking interpersonal validation

- Perfectionism
  - attempting to attain the thin ideal
  - control over weight/shape/eating/exercise

- Inflexibility/rigidity
  - extremes in thinking,
  - lack of flexibility or openness,
  - lack of recognition of information counter to drives around body and thinness

Appendix D

Topic Model Summary

In the following pages is a summary of the 20 topic model discussed in Section 4.5. Presentation matches that shown to collaborators.

Page 0 - Topic 13

#edproblems
#thinspo
’
!
#thinspiration
you
your
:
when
#thighgap
this
the
#skinny
_
in
food
cold
to
being
up
;
because
#diet
#weightloss
&
#eatingdisorder
feeling
hunger
her
you’re
calories
eat
#ana
perfect
#anorexia
legs
always
#proed
yourself
out
#thin
having
stomach
#hipbones
?
and
#ed
every
purging
fat
,

Page 0 - Topic 13 Example Tweets

tweet id 319874259566596096, 19 topic words out of 22 modelled words
@sambarrow1987 sitting here on your own gives you plenty of thinking time! Why don't you just go be sick again? #youdoloveit #bulimic

tweet id 305683960837582849, 10 topic words out of 10 modelled words
My hands are freezing and I love it. #EDproblems

tweet id 365984057345064960, 13 topic words out of 22 modelled words
Analysing your body from every angle for hours on end in the mirror! Then breaking down and crying :'(#edproblems #realityproject

tweet id 315883159822671872, 7 topic words out of 7 modelled words
Pulling out tons of hair. #edproblems

tweet id 320970458948321280, 9 topic words out of 19 modelled words
Tired of being the (FAT) one. Wanna be the |skinny| one. #MyLife #EDproblems

tweet id 417336936798224384, 10 topic words out of 15 modelled words
"@thinsporexic: All the time hurts #EDproblems #EDthoughts http://t.co/ioW87zo3zp"..ay..

tweet id 300277559566942208, 13 topic words out of 13 modelled words
#EDProblems working out every day even though you're too weak to think properly

tweet id 298147140272799744, 2 topic words out of 6 modelled words
Aaaaah! Fasting headaches SUCK! #EDproblems

tweet id 378938080691892224, 11 topic words out of 12 modelled words
Every morning I wake up and the scale is lower. #EDproblems

tweet id 292986788232970242, 4 topic words out of 16 modelled words
I wish I had a thigh gap like that when I was lying down :(#thinspo http://t.co/znjxRgTA

tweet id 338466578268057601, 18 topic words out of 21 modelled words
lax lovin.... jk mentally preparing for the middle of the night awful cramps and sickness #edproblems #whatsrecovery #trying

tweet id 312353668844359682, 15 topic words out of 21 modelled words
"@crazytobe_ana: My stomach is so full it hurts on only 400 calories. #EDproblems" my life.

tweet id 323553458697404417, 5 topic words out of 19 modelled words
I think I'm the only girl ever to cry over the fact that my boobs are getting bigger #EDproblems

tweet id 347149268756336641, 5 topic words out of 10 modelled words
Sitting on the couch #swag #help #sobored #ugh #sigh #thighgap #kony2012 #freelohan #taxidermy

tweet id 366427746399817728, 9 topic words out of 10 modelled words
I'm a fat whale. End of story. #edproblems

tweet id 345167607697141760, 10 topic words out of 14 modelled words
Standing up & feeling like you're going to passout... #edproblems

tweet id 326308655454617600, 9 topic words out of 21 modelled words
This is what recovery looks like, had to slip up with the doc, huh Ana? #edproblems #reversethinspo #picslip http://t.co/hVItqzL11m

tweet id 297794067218194432, 11 topic words out of 12 modelled words
@Kajika_BoomCO you know you're #fuckedup when you love being hungry #EDproblems #TheEnd

Page 1 - Topic 11

is
#thinspo
perfect
my
she
she's
'
I
beautiful
this
so
favorite
to
:
perfection
favourite
miley
one
ultimate
#edproblems
body
oh
you
goal
#proana
thinspo
_
god
dream
.
such
me
gorgeous
kendall
be
kate
cyrus
#ana
taylor
it
want
fav
;
#thighgap
fave
#diet
that
?
#weightloss
on
&
ariana
have

Page 1 - Topic 11 Example Tweets

tweet id 291877075055431680, 2 topic words out of 11 modelled words
Keep this in mind when you want to binge. #thinspo http://t.co/UsFPOzcl

tweet id 404047039526232064, 1 topic words out of 7 modelled words
#FF the queen of Oreos and snapchats @gabrielabarkho #thighgap

tweet id 296024373008138242, 6 topic words out of 6 modelled words
She is so sexy #thinspo #fitspo http://t.co/i1N9Z8kP

tweet id 400750381346873344, 4 topic words out of 4 modelled words
she's actually perfect #thinspo http://t.co/71jPqm5rL7

tweet id 283265871684722690, 4 topic words out of 14 modelled words
#ReasonsToBeFit To have amazing bodies like ALL of these wonderful #VSAngels #motivation #thinspo #fitspo http://t.co/1TybIP8b

tweet id 430344260068470784, 3 topic words out of 16 modelled words
Parang nag-hehele si Teh Iza maghost.lol Anyways. She much better than the previous host noh. #thinspiration #BiggestLoserPH

tweet id 378904137066217472, 7 topic words out of 7 modelled words
She is so stunning and #perfect #thinspo http://t.co/aAne7MWV0H

tweet id 349851874033274880, 4 topic words out of 9 modelled words
This is my favorite #thinspo <3 http://t.co/MXBaLkFwTB

tweet id 400371004049866754, 4 topic words out of 6 modelled words
Perfection... #thighgap #thinspo http://t.co/HDZJtz5mCi

tweet id 299632228013588480, 4 topic words out of 4 modelled words
My #thinspiration @caradelevigne #stunner. http://t.co/f0XEn41N

tweet id 377010168309231616, 5 topic words out of 12 modelled words
Hashtag my body. 🙌✌ Okay let's do this!! #thinspiration #skinny http://t.co/wSLwvIHxPs

tweet id 328147972787404801, 4 topic words out of 4 modelled words
She's perfect. #Thinspo http://t.co/hAVSfYZZtZ

tweet id 396626022944100353, 3 topic words out of 3 modelled words
🙌 #thinspo #ArianaGrande http://t.co/W77UZ503IL

tweet id 310890702101946370, 3 topic words out of 10 modelled words
Omg. Can I have your body? #soskinny #thinspo http://t.co/KiXqZ6AJBm

tweet id 324059927142354944, 3 topic words out of 3 modelled words
#thinspo she's perf http://t.co/GaTK4fdP10

tweet id 402243197889298433, 14 topic words out of 17 modelled words
Omg this is seriously perfect... The ideal body ❤🙌 #thinspo #thinspiration #ana #anorexia #weightloss http://t.co/VXZibMpbtw

tweet id 285788833629495296, 5 topic words out of 22 modelled words
isnt she gorge..now dont you wanna have a body like this one day??..x #thinspo http://t.co/cFsV03aB

tweet id 328743331226062849, 5 topic words out of 6 modelled words
#thinspo one of my favorites actually http://t.co/jsqcFubNKu

Page 2 - Topic 12

;

&

amp

lt

3

gt

’

I

w

to

:

#foodaddiction

.

/

a

#healthyeating

and

#edproblems

!

is

you

?

addicts

$2

#proana

garver’s

felicity

@amazonkindle

#SALE

#skinny

so

#bulemia

99

me

33

learn

be

night

patricia

bacall

#kindle

#USA

#canada

like

#EU

#UK

the

#ana

it

-

kills

want

http://t.co/Pj1zcL7fSY

hurts

#thighgap

#diet

that

#weightloss

Page 2 - Topic 12 Example Tweets

tweet id 295343694117216256, 4 topic words out of 20 modelled words
Sitting here on a Saturday night reading LandArch blogs... and I'm not the least bit ashamed. #EDproblems #yesIsaidEDproblems

tweet id 311613554304360448, 3 topic words out of 25 modelled words
Summers right around the corner, so I gotta get my ass back in the gym & get in shape. #thinspiration… http://t.co/c3TsJdANGT

tweet id 361651696096665600, 3 topic words out of 3 modelled words
Sunday night #Thinspo http://t.co/FCcKfRiIFk

tweet id 340362710191587329, 6 topic words out of 14 modelled words
Good night skinnies <3 sweet dreams! :) #thinspo #ED #night

tweet id 335296118256726016, 22 topic words out of 22 modelled words
#Bulemia & #FoodAddiction Kills. #HealthyEating w/Patricia Bacall s Love Yourself Thin #Kindle http://t.co/Pj1zcL7fSY #canada #UK #USA #EU

tweet id 291662294390292480, 1 topic words out of 6 modelled words
More perfect legs! #thinspo #thinspiration http://t.co/P2H6Rdvl

tweet id 347061662651645954, 5 topic words out of 18 modelled words
"@hatingmyselfx_: "@uglyxfat: #thinspo http://t.co/hFvnbUq6vc" wow, that's perfect <3" OMFG THESE

tweet id 310658010697592833, 1 topic words out of 14 modelled words
Ill stop when I'm skinny. Ill stop when I'm beautiful. #proana #thinspo

tweet id 383166017964224512, 4 topic words out of 18 modelled words
#InternationalEDMeetup #represent! <3 from #Colorado! Anyone wana hv a physical meetup? #eatingdisorders #anorexia #ednos #bulimia #support

tweet id 283131824266424320, 5 topic words out of 8 modelled words
<3 #audrey #hepburn #thinspo #smallwaist #collarbones #skinny http://t.co/CG5BhFCd

tweet id 328672714980225024, 2 topic words out of 11 modelled words
I think thigh gaps are DISGUSTING #sorrynotsorry #thighgap #behealthy @nicolearos obeymelibaeex3 http://t.co/44R9YQQNvw

tweet id 305007382822547456, 5 topic words out of 22 modelled words
Just saw this on my facebook..omg, her legs are perfect <3 real life #thinspo ♥ http://t.co/NQ4eXl1qrW

tweet id 289508107816673280, 5 topic words out of 5 modelled words
#Thinspo <3 http://t.co/djt7BTb7

tweet id 377610812556595200, 5 topic words out of 14 modelled words
Lamb & sweet potato tangine! #lunch #healthy #healthyeating #fitspo #fit #thinspiration #6pack http://t.co/LCD8iBD07C

tweet id 290496536448733184, 3 topic words out of 4 modelled words
Leah & Dianna #Thinspo http://t.co/vc7rkY8F

tweet id 387195562065264642, 4 topic words out of 28 modelled words
Obsessing over 'before' pictures & measurements. Together I've taken over 100 of them. Does anyone else go to this extent? #edproblems

tweet id 399557542772486144, 3 topic words out of 24 modelled words
Insta girl uploads a picture of apple #eatclean #fitness #fitnessfreak #thighgap Me: *uploads pic of Ben & Jerrys* #getfat #happy #curves

tweet id 356760828663435265, 3 topic words out of 5 modelled words
Fashionable #Thinspo you'll love #skinny http://t.co/FIhbD2H4cV

Page 3 - Topic 17

!

#thinspiration

’

.

#thinspo

gym

:

@vickygshore

#edproblems

#proana

I

#skinny

meck

get

@lucy

work

#ana

@josiestweet

workout

for

want

summer

amazing

;

#weightloss

&

#eatingdisorder

haha

back

?

?

#diet

perfect

bod

#anorexia

legs

new

#proed

wait

#thin

run

off

working

x

#ed

if

i

looks

when

2013

xx

don’t

hard

eat

DVD

need

#anasisters

screen

please

Page 3 - Topic 17 Example Tweets

tweet id 319189811032371200, 5 topic words out of 20 modelled words
I will look like this when I go on holiday!! #thinspiration #skinny #love #hot #body #want #need #60…
http://t.co/DyERdTOBuR

tweet id 284280181345095680, 5 topic words out of 6 modelled words
New year, new me #thinspiration #zzzzzz #nochance #giveitamonth http://t.co/Dh4dRxnP

tweet id 287684876000456704, 5 topic words out of 8 modelled words
Dreaming of being this petite! #thinspo #fitspo http://t.co/4XTRqnqe

tweet id 344908736705474560, 3 topic words out of 6 modelled words
@Josiestweet #thinspo you're so perfect x http://t.co/ykXbC8Byku

tweet id 371336232548466688, 5 topic words out of 5 modelled words
#ThinSpo aha I'm so creative http://t.co/usxvT5hVnx

tweet id 328925165499072512, 18 topic words out of 19 modelled words
@jesswalker_93 has accompanied me to the gym, bit is sat on the weights eating jelly! #thinspiration!

tweet id 341795994360479744, 12 topic words out of 14 modelled words
Finally my fast diet recipe book has arrived! So excited! #fastdiet #fattynomore #thinspiration #5and2
http://t.co/Z15DoAPQ7s

tweet id 403442332961607680, 7 topic words out of 7 modelled words
Bye bye alcohol hello abs #letsdothis #thinspiration

tweet id 311181184833056768, 17 topic words out of 17 modelled words
going to start the healthy diet! need to loose weight for cousins wedding #2weeks #thinspiration

tweet id 367337920371257344, 14 topic words out of 14 modelled words
finally got my beach bod, come at me maga! ?#bikini #thighgap http://t.co/Wk1n2UmTvP

tweet id 349141541312417792, 4 topic words out of 8 modelled words
#ThighGap #HipBones #CollarBones #TheCurve. Work it baby xx http://t.co/1fOsJ0KSpO

tweet id 390599321366040576, 7 topic words out of 20 modelled words
If any girl said they weren't jealous they are just a plain upfront LIAR!!! @lucy_meck #thinspiration
http://t.co/6jyEX1QxVa

tweet id 348174688867991552, 1 topic words out of 6 modelled words
#thinspo #pale i love the bikini http://t.co/iXE678vDSe

tweet id 292931828657049600, 11 topic words out of 11 modelled words
Worked way too hard at the gym yesterday #ouch #thinspiration #achey ??

tweet id 328534511375769601, 9 topic words out of 10 modelled words
Always on my mind. #healthyactivelifestyle #thinspiration #running @ Ultra - Track Oval http://t.co/vcVkGXp9bJ

tweet id 335053852481044480, 11 topic words out of 12 modelled words
really enjoyed that gym session. Going back for more tonight #suckerforpunishment #hitthegym #thinspo

tweet id 289217456810106881, 3 topic words out of 4 modelled words
I need some #Thinspiration

tweet id 329876276041113600, 4 topic words out of 9 modelled words
Healthy brecky before the gym.... ?? #thinspiration http://t.co/UJw0wKkS1c

Page 4 - Topic 7

’

#thinspo
a
and
dinner
!
of
tea
breakfast
lunch
ate
for
had
:
chocolate
water
you
#skinny
pizza
_
salad
coffee
be
#thighgap
today
diet
cream
#ana
green
cheese
ice
this
,
eaten
want
;
coke
#weightloss
meal
&
#eatingdisorder
her
chips
apple
cake
chicken
drink
#edproblems
eating
soup
fruit
food
perfect
day
#anorexia
legs
cup
?
drinking

Page 4 - Topic 7 Example Tweets

tweet id 407346291967270912, 5 topic words out of 5 modelled words
Ughhhh blood work tomorrow #EDproblems

tweet id 407262624796377088, 4 topic words out of 12 modelled words
Ive defo eaten double my body weight in food today #instahotty #instalike #thinspiration

tweet id 424571679382061057, 5 topic words out of 9 modelled words
“@dynshmi: didn’t eat the whole day ahahahahahaha” #thinspo

tweet id 402521452303622144, 17 topic words out of 21 modelled words
Two days ago I ate pizza just fine, now I have anxiety over oven-roasted vegetables. #ednos #edproblems

tweet id 396047796463484929, 9 topic words out of 12 modelled words
The amount of candy and donuts I’ve just eaten □ #thighgap #halloween

tweet id 406546654561443840, 15 topic words out of 20 modelled words
I’m making a list, checking it twice, gonna find out if i can sort out my life #christmaslist #EDproblems

tweet id 310633021210509312, 11 topic words out of 19 modelled words
My fuckin roommate is heating up brownies to mix in with her ice cream. FUCK YOUUUUU. □□□ #EDproblems

tweet id 416208042699390978, 8 topic words out of 11 modelled words
my first meal of the day no joke #beslim #slim #sexy #hourglassbody #thighgap

tweet id 345495679801626624, 4 topic words out of 8 modelled words
A microwaveable sticky toffee pudding for lunch #hellyeah #thinspiration #jks #dontcare

tweet id 320243411900854272, 16 topic words out of 23 modelled words
Make that 223 cals today. Cut my half apple for lunch down to a quarter apple because FUCK food. #edfamily #anasisters #staystrong

tweet id 355081970725814273, 25 topic words out of 27 modelled words
So I’ve being doing to rainbow diet and today I’ve had an apple and a cucumber also went in a run but I feel so bloated #edproblems

tweet id 286827722653704192, 8 topic words out of 9 modelled words
Needing a wee right after u have been #EDproblems

tweet id 407435836305993728, 3 topic words out of 19 modelled words
seeing women who have had kids with perfectly flat stomaches, fills me with envy□. #greeneyedmonster #envy #getfit #thinspiration

tweet id 317683754380435456, 8 topic words out of 12 modelled words
I’ve eaten too much, time to purge. #purge #needtobeskinny #proana

tweet id 291617449412018176, 10 topic words out of 21 modelled words
i’m hungry but i’ve recently eaten so much shit and haven’t burnt it off so yay hunger be my friend #thinspiration

tweet id 324115970111201280, 8 topic words out of 11 modelled words
#eatinghealthy #thinspiration #summerbod... You’re a streak of piss anyway

tweet id 299891265330372608, 24 topic words out of 24 modelled words
#EDProblems having meal eg stew/ soup and having to look at recipes find the individual cal content of ingredients then add together :/

tweet id 388807242017148928, 6 topic words out of 12 modelled words
Haven’t eaten in 8 hours, neeeww clothes :$. #thinspo #thinspiration http://t.co/mlADLcMBIl

Page 5 - Topic 1

her
legs
#thinspo
bones
perfect
’
love
tiny
so
those
want
hip
.
arms
to
collar
are
:
ribs
omg
these
#edproblems
collarbones
beautiful
flat
waist
hipbones
the
she’s
!
stomach
you
?
a
#proana
of
_
chest
body
be
wow
for
⍰
in
;
#diet
#weightloss
lovely
on

Page 5 - Topic 1 Example Tweets

tweet id 298335648358735872, 1 topic words out of 3 modelled words
#thinspo amazing! http://t.co/rOVtibZf

tweet id 334449209401020417, 3 topic words out of 3 modelled words
Lovely bones #thinspo http://t.co/OZOHAQSRfC

tweet id 296337362852671491, 5 topic words out of 5 modelled words
Omg want this outfit #thinspo http://t.co/glazIcv5

tweet id 294672690961330176, 11 topic words out of 18 modelled words
@DelicateDoll123 hahahahahahahha clear, i saw her tweets! I love her! She is my #thinspo

tweet id 295924193374109697, 8 topic words out of 11 modelled words
She has a tiny body & cute outfit #thinspo http://t.co/OIjK9EfJ

tweet id 344670156275986432, 5 topic words out of 24 modelled words
I didn't count my calories today now i'm not going to be #thinspo at this rate. Ugh my collarbones dont even show :(

tweet id 290579285276758017, 5 topic words out of 12 modelled words
@c4t_scr4tches oh my god! Your legs are perfect! #Thinspo

tweet id 350439785686515713, 5 topic words out of 5 modelled words
her legs! #thinspo #legs http://t.co/yHmE1tbZeN

tweet id 375649918243065856, 6 topic words out of 6 modelled words
I need those collarbones. #thinspo http://t.co/krueLySMjh

tweet id 394985855213142017, 4 topic words out of 4 modelled words
Skinny and beautiful #thinspo http://t.co/bzArbjDPXW

tweet id 283705912181538816, 11 topic words out of 27 modelled words
#Thinspo #idol Demi is perf. I want her hip bones. I want to be able to wear these clothes and feel comfortable. Ok? http://t.co/xdNRgO3z

tweet id 289043257718083585, 5 topic words out of 27 modelled words
Making a fat home made veggie soup with the flat for tea! Apparently day three is the hardest but so far so good! #healthy #thinspo

tweet id 424203174569074688, 11 topic words out of 15 modelled words
@RochelleTheSats your figure is perf, I'm so jealous I need a stomach like that ??? #thinspiration http://t.co/ArPi0L6f2g

tweet id 334067402834857984, 4 topic words out of 8 modelled words
"@AnaRexxx: #thinspo ! http://t.co/8oHIbpgeyE" her stomach omfg

tweet id 338333446097215491, 9 topic words out of 9 modelled words
Her legs tho, I need them! #thinspo http://t.co/7KHaiYAMD8

tweet id 319499927300472832, 3 topic words out of 3 modelled words
#thinpso #thinspiration omg http://t.co/xAzq1klRde

tweet id 304726503474749442, 4 topic words out of 6 modelled words
http://t.co/psMPGP022L - Such perfect legs! #thinspo

tweet id 318669155039580160, 4 topic words out of 5 modelled words
Amazing summer legs #thinspo #thighgap http://t.co/dzCTt3gShR

Page 6 - Topic 6

/

—

:

•

@
@ana
I
@skinny
.
to
angelx
the
,
!
a
and
#edproblems
RT
@not
you
?
#proana
#skinny
my
@anorexic
of
@starvationarmy
be
for
chanel
mind
like
@the
#ana
in
s
i'm
beskinny
;
dreams
SA
#diet
#weightloss
on
&
have
#eatingdisorder
just
@idontknow
it
very
#anorexia
breathing
but

Page 6 - Topic 6 Example Tweets

tweet id 414203551791861760, 9 topic words out of 11 modelled words
“@not_breathing_: #thinspo her legs tho http://t.co/PzJDNlEHYM”

tweet id 357249037923131394, 4 topic words out of 6 modelled words
@Tams_1Dxxx so do I #Thinspiration

tweet id 306593976453062656, 5 topic words out of 5 modelled words
″@LaurenBrianna13: #Thinspiration ???? http://t.co/3LXT9HM536″

tweet id 378746168920248320, 13 topic words out of 15 modelled words
“@stop_eating_now: ″@StarvationArmy: #Thinspo #StayStrong http://t.co/EXE47uUYn5″ Beautiful.”

tweet id 314933004751491073, 11 topic words out of 11 modelled words
“@Thais_Kila_Moon: #thinspo #ReasonsToLoseWeight #motivation http://t.co/pUL0kEUCAe”

tweet id 375858607620751360, 4 topic words out of 21 modelled words
@_HUMANOID_CITY_ plz also follow @HexiiSexii @FairyKeilu @FireflyWinter & @dailyTOKIOHOTEL #thinspo (via http://t.co/X5bJvQH1P1)

tweet id 403671004058116096, 13 topic words out of 15 modelled words
“@starve2belovely: “@not_breathing_: #thinspo #hipbones http://t.co/NjQTnD1fqG”and nirvana oh god so perf”lovely

tweet id 403240283208712192, 15 topic words out of 25 modelled words
“@proana4lifexx: “@cut_wrists: “@chelseaBiffyC: #thinspo bones <3 http://t.co/35Ajd9QUoZ” pretty bones”” this is gonna be me

tweet id 374968534977499136, 5 topic words out of 18 modelled words
″@/SkinnierPlease: Can’t get over those perfect legs.. #thinspo http://t.co/t1XhuR49SX″ looks unhealthy tbh

tweet id 418315544996356096, 3 topic words out of 12 modelled words
″@2014fittie: #thinspo http://t.co/BIgJxwMjJB″ you’re doing it wrong...

tweet id 360823200030007296, 9 topic words out of 10 modelled words
″@ana_angelx: #thinspo http://t.co/36FTFIX2fn″ now that’s ugly

tweet id 342793655150583808, 4 topic words out of 4 modelled words
? RT @arieeolla: #ThighGap

tweet id 409460135988170752, 10 topic words out of 10 modelled words
″@_fatalwaysfat_: 5 #thinspo #thinspiration http://t.co/4hZwFH4Ff3″

tweet id 382172600270454785, 8 topic words out of 12 modelled words
″@anorexic_dreams: #thinspo her outfit is too cute http://t.co/laTNyG4DWi″

tweet id 291346424459825152, 7 topic words out of 15 modelled words
“@Iwilldieskinny: Me = crying My life #anasisters http://t.co/9aLRGoIv” Oh. My. Days.

tweet id 330153585667080195, 10 topic words out of 16 modelled words
″@SkinnyNowwww: “@ToBeBeautiful56: #Thinspo #thighgap http://t.co/a3JZfoxCCV” I want this.″ SO pretty!

tweet id 394147442473439232, 7 topic words out of 19 modelled words
″@Iwanta_thighgap: I want collar bones like that. ? #thinspo http://t.co/EuQ44VukRr″ :) soon enough.

tweet id 425667214574755842, 10 topic words out of 13 modelled words
“@why_food: I honestly love this. #thinspo #mydreambody http://t.co/E3ys3F1Myq”

Page 7 - Topic 10

#bulemia
.
#anorexia
loss
-
weight
’
eating
disorders
I
#eatingdisorders
#ipad
#followmejp
#sougofollow
#followall
#teamautofollow
#mustfollow
#teamfollowback
#followme
before
#followback
#follow
after

pics
#thinspo
disorder
~
/
anorexia
A
#EDNOS
on
this
#kids
an
edge
factor
via
goddess
epidemic
cutting
technique
blogger
basically
venus
conflicted
ladies
#recovery
you
#bulimia
pressure
#skinny

Page 7 - Topic 10 Example Tweets

tweet id 316691263036461057, 13 topic words out of 13 modelled words
Get #thighgap pics #ipad #follow #followme #followmejp #sougofollow #followback #teamautofollow #teamfollowback #mustfollow #followall 2465

tweet id 367725632836210688, 9 topic words out of 9 modelled words
MEW! Blog Post: #AlmostAnorexic Personal Review http://t.co/FJeT1QrUFr #eatingdisorder #ednos #osfed

tweet id 295618509923299328, 7 topic words out of 7 modelled words
Before and After Weight Loss #thinspo #21 http://t.co/bDjtnNuG

tweet id 299647642173059073, 9 topic words out of 19 modelled words
Re-pinned so much #thinspo on @Pinterest I got a PSA message about eating disorders.... #awkward

tweet id 359363527544279040, 4 topic words out of 7 modelled words
Jeensa from Eternal Desire - http://t.co/gNLWppZ3X5 #skinny #thin #thinspo

tweet id 343083545708871682, 3 topic words out of 5 modelled words
#thinspo #workout http://t.co/19oHyH3jCE /via @thatsneakyED

tweet id 330431529656389632, 2 topic words out of 12 modelled words
Anything other than vegetables to me is considered a binge. #edproblems

tweet id 316754014060544001, 13 topic words out of 13 modelled words
Get #thighgap pics #ipad #follow #followme #followmejp #sougofollow #followback #teamautofollow #teamfollowback #mustfollow #followall 3479

tweet id 355050702659264512, 3 topic words out of 15 modelled words
just found the most depressing instagram accounts ever.. :(#anorexia #bulimia #thighgap #sosad

tweet id 409942724008509440, 1 topic words out of 7 modelled words
“@whatisskinny: #thinspo #proana #thighgap #skinny http://t.co/3UdDEWtHlZ”

tweet id 348205770053140483, 11 topic words out of 11 modelled words
Excellent article on eating disorders by @UrbanMooCow A must read http://t.co/QDSy8KBk8o #anorexia #bulemia

tweet id 288761039359705088, 5 topic words out of 8 modelled words
From 232 to 148 Before and After Weight Loss #thinspo

tweet id 346844564100902912, 6 topic words out of 7 modelled words
The actress who plays Nikita is serious #thinspo http://t.co/60WkwREYih

tweet id 334778641202421763, 5 topic words out of 5 modelled words
Christine Quinn's Brave Choice http://t.co/AVso8q4Q11 #bulemia #anorexia

tweet id 397323659951738880, 6 topic words out of 8 modelled words
Love @instagram ~ http://t.co/r4FbLWRHMS #thinspo #bodyimage #psa (via @HuffPostCollege) http://t.co/QZIpEFBphO

tweet id 428184699651891201, 4 topic words out of 12 modelled words
Yes yes yes #thinspo #thinspiration #fitfeb -shared via the Android Thinspo App http://t.co/SBd5zcu31V

tweet id 294905188798590976, 7 topic words out of 7 modelled words
Before and After Weight Loss #thinspo #2 http://t.co/LNVaqqMj

tweet id 329077360043372544, 3 topic words out of 7 modelled words
@Huntermoore #thighgap #creeponme Instagram: kelseyski #thighgap #thighgap #thighgap http://t.co/6eVkVQSucg

Page 8 - Topic 2

/

:

@skinnyanorexic
to
#thinspo
I
@skinnierplease
@lookingforthin
the
#bonespo
my
@inmyskinnydream
a
@lanadel
and
#edproblems
,
@skeletalstoner
RT
90
@thiinspohelp
you
#proana
!
of
123
@thinspoloves
@delicatedoll
be
for
perfect
#thinspiration
#ana
in
it
@dreaminskinny
i'm
;
#diet
@tothinkthin
#weightloss
on
&
have
morning
#eatingdisorder
perfection
theme
_
#anorexia
but
@skinny4
look

Page 8 - Topic 2 Example Tweets

tweet id 432389482068078592, 9 topic words out of 12 modelled words
″@KidrauhlDenmark: “@dxstract: #thinspo #ribs #hipbones http://t.co/ZGcmREpqhI” OH MY TITS″

tweet id 395341086178021376, 7 topic words out of 11 modelled words
″@LukesSassQueen: Do it for... #thinspo http://t.co/8HJg3DdgWv″

tweet id 420349926280019968, 8 topic words out of 11 modelled words
“@SucidialSkinny: Perfect #thinspo http://t.co/JkfljfAazZ” can I be her please

tweet id 352226528853098496, 5 topic words out of 5 modelled words
#Thinspo theme today: Legs http://t.co/6kzrO6W017

tweet id 389824482539757568, 4 topic words out of 6 modelled words
″@StarvingDaily: #thinspo http://t.co/W2wjq1kRpc″ want

tweet id 401091182727798784, 8 topic words out of 8 modelled words
This is gross “@Thiinspohelp: #thinspiration http://t.co/1hVI1V5Xbq”

tweet id 375312097170554881, 6 topic words out of 25 modelled words
″@DelicateDoll123: #wheniwakeup I pee and then weigh myself. #EDproblems″ Every morning. If I miss, I freak out all day

tweet id 303451224244629505, 3 topic words out of 19 modelled words
“@ProAnaTip: An eating disorder is for life, not just until you reach your UGW. #EDproblems”

tweet id 401808869275729921, 15 topic words out of 17 modelled words
Incredible ″@paradeengordar: “@calmfocused: Amazing bod... RT @Anabonegirl: □ #thinspo http://t.co/FALwK6UurR””

tweet id 352200282786107392, 7 topic words out of 7 modelled words
“@lanadel90: #thinspo #7 http://t.co/WfhVlNRz9f”

tweet id 391647049118396416, 5 topic words out of 5 modelled words
″@HiAnorexia: #thinspo #motivation http://t.co/N1o9HE93VI″

tweet id 405113697993240576, 6 topic words out of 7 modelled words
“@ariarose123: http://t.co/7qS0Eks7g7” #thinspiration □□

tweet id 321403856238022657, 4 topic words out of 15 modelled words
″@LifeAnorexic: #legs #perfection #thighgap #thinspo http://t.co/e2h9lcKMZ8″thigh gaps are the prettiest!!

tweet id 331197737406169088, 8 topic words out of 13 modelled words
″@CalorieCutter: Karlie Kloss #thinspo http://t.co/TWJG9WkjFJ″ | Perfection at it's finest.

tweet id 369341632245559297, 6 topic words out of 16 modelled words
“@Beautyandbonees: Do it for the hipbones. #ana #mia #ed #skinny #thinspo #hipbones http://t.co/lnviKe5tHW”

tweet id 431581575105351680, 6 topic words out of 6 modelled words
″@perishingbones: #thinspo #thinspiration http://t.co/Pwe0tl5sga″

tweet id 432575311176540160, 13 topic words out of 14 modelled words
“@lmloosingmyself: “@Skinnybones91: This is me! #EDproblems http://t.co/bM3m1wS0Ae””

tweet id 313046875450269697, 2 topic words out of 9 modelled words
Perfection! #skinny #thinspiration #eatingdisorder #thinspo #proana #proed #EDNOS http://t.co/tKzpWIaJIq

Page 9 - Topic 0

I
#thinspo
′

.
#edproblems
#thinspiration
!
:
eat
i′m
#thighgap
but
feel
to
#skinny
fat
_
don′t
myself
you
;
as
hate
#diet
#weightloss
&
#eatingdisorder
am
her
hungry
i
perfect
not
can′t
legs
feels
are
it
#proed
purge
your
#thin
tastes
nothing
#hipbones
-
never

Page 9 - Topic 0 Example Tweets

tweet id 399026620976996353, 16 topic words out of 19 modelled words
Thinking I should do some crunches and squats. Not like I can sleep anyway. #EDproblems #insomnia #Ana

tweet id 326823815281782785, 3 topic words out of 25 modelled words
I purposely stop taking my pain pills from surgery so chewing hurts like crazy that way I have an excuse to not eat #edproblems #crazy

tweet id 340533416623759360, 17 topic words out of 34 modelled words
I had a great week...cancer markers great, got a great SW job...but feel so down cause i feel so damn FAT...#edproblems #mental #fatpig

tweet id 411488511162732544, 6 topic words out of 14 modelled words
I'm sorry for the #thinspo spam. Don't hate me. I'm weak.

tweet id 344233196331859969, 13 topic words out of 14 modelled words
Gained weight? Ew. Lost weight? Ew. Still fat. #EDproblems

tweet id 336223844245307392, 19 topic words out of 32 modelled words
I can binge on my birthday, right? 'cause I kinda have been all day... #shit I just realised I didn't make my birthday GW #fml #edproblems

tweet id 362549855547162624, 22 topic words out of 26 modelled words
if you could guarantee I wouldn't gain weight through eating 1hotdog I still wouldn't eat it because i'll feel too guilty& fat #edproblems

tweet id 315718270604963840, 19 topic words out of 23 modelled words
I'm too fat to even be allowed to tweet right now. Reading Skinny Bitch Guide until I fall asleep #ana #mia #edproblems

tweet id 427466103577903104, 10 topic words out of 19 modelled words
My leggings feel a little loose so that means I should go to Dunkin donuts alone right now right #THINSPO

tweet id 318220868100689920, 10 topic words out of 10 modelled words
I wish I wasn't so ugly. #Ana #Mia #Ednos

tweet id 296909460876713984, 6 topic words out of 6 modelled words
I'm addicted to fasting. #EDProblems

tweet id 314450429129936896, 10 topic words out of 13 modelled words
yup im #bulimic and #fat and #hopeless and #depressed welcome to my life

tweet id 286255271389577216, 20 topic words out of 29 modelled words
There's pie here. I fucking love pie. I would eat it... If I wasn't already fat. Send some #thinspo my way pls):

tweet id 338798392459284481, 7 topic words out of 9 modelled words
#thighgap I know I'm not funny, I try https://t.co/U4Wm49r7hj

tweet id 294287888831614980, 14 topic words out of 26 modelled words
Checked my BMI and still below a 17.5 was feeling fat I have to maintain the diagnosed #anorexic BMI or I fail #EDproblems #EDthoughts

tweet id 344065467218018304, 20 topic words out of 26 modelled words
Do you know what I love about life? ....yeah...no me either... :/ #edproblems #depression

tweet id 343979446946111488, 15 topic words out of 24 modelled words
I've strayed so far off the path of Ana...I need help. Im disgusting and I hate myself! #NoFood #MustNotEat #ProAna #AnorexicForLife

tweet id 318102619681853442, 6 topic words out of 24 modelled words
So mom basically just told me I would never be skinny, thanks for the #thinspo mom(: just the motivation I needed!

Page 10 - Topic 4

#skinny
#thinspo
#hipbones
.
#legs
#thin
#ribs
#perfection
#thighs
#thinspiration
#ana
#perfect
#thighgap
#girly
#beautiful
#collarbones
#skinny4
#bones
’
xmas
I
#mia
#flatstomach
#ED
!
#tiny
see
to
more
#motivation
:
#arms
my
the
,
#thinspire

Page 10 - Topic 4 Example Tweets

tweet id 292992048632582144, 10 topic words out of 10 modelled words
#perfect #thinspo #thinspiration #proana #ana #mia #anorexia #skinny #thin #bones http://t.co/KwndUL4k

tweet id 339481846121369602, 5 topic words out of 5 modelled words
#thinspiration #thinspo #inspiration #thighgap #flatstomach http://t.co/OdNS5hE4T8

tweet id 350430753919729666, 4 topic words out of 5 modelled words
#ribs #flat stomach #hipbones #thinspo http://t.co/KEaEBHXetQ

tweet id 356341729516994561, 3 topic words out of 3 modelled words
#thinspo #abs #legs http://t.co/SkRm2snpIk

tweet id 389204536851591168, 8 topic words out of 8 modelled words
#thinspiration #thinspo #thin #thighgap #collarbones #thinarms #ribs #hipbones http://t.co/BMXHe8qT9f

tweet id 315739039733002241, 4 topic words out of 4 modelled words
#thinpso #Hipbones #thighgap #ribs http://t.co/u9mOBiRHDA

tweet id 413916898821423104, 3 topic words out of 3 modelled words
#thinspo #hipbones #perfection #canihave http://t.co/QFUjMLoKNM

tweet id 372115790046429185, 6 topic words out of 7 modelled words
see more http://t.co/fZjqmZ1Xp1 #thinspo #skinny #thinspiration #girly #thighs http://t.co/CrviqQ1Z6z

tweet id 363464582817275904, 4 topic words out of 4 modelled words
#BeforeAndAfter #Fitspo #Thinspo #StayStrong http://t.co/2grvxv3iea

tweet id 390206553795874817, 3 topic words out of 3 modelled words
#thinspo #skinny #want http://t.co/Se1ngVbiQ2

tweet id 379922471434330112, 11 topic words out of 11 modelled words
#thinspo #thinspire #skinny #tiny #ana #ed #anamia #thyn #proana #promia #edfamily http://t.co/EnHEZqmOMD

tweet id 362309856138518528, 3 topic words out of 3 modelled words
#Thinspo #Thinspiration #Collarbones http://t.co/cX9oxhm449

tweet id 281989973236920320, 2 topic words out of 3 modelled words
#want #thighgap □ http://t.co/2VIfUu80

tweet id 290618862079193089, 3 topic words out of 3 modelled words
#thinspo #thinspiration #loveit http://t.co/ql6m8Ahu

tweet id 399696732348350464, 2 topic words out of 4 modelled words
Another night #thinspiration #bikini http://t.co/x0vLPu70Pu

tweet id 297230139413114880, 4 topic words out of 4 modelled words
#thinspo #thinspiration #beautiful #thin http://t.co/DObubKPb

tweet id 372855637639958528, 7 topic words out of 7 modelled words
see more http://t.co/3mHfuGmEm6 #thinspo #skinny #thinspiration #girly #thighs http://t.co/aZMH74RjLB

tweet id 319858550664986624, 9 topic words out of 9 modelled words
#thinspiration #thinspo #tattoo #ana #mia #skinny #thin #thinner #body #soul http://t.co/aB8uV1nyZL

Page 11 - Topic 18

…
#fitness
#fitspo
#fit
#fat
’
#healthy
I
#anorexic
!
#workout
#summer
#me
#thin
#love
#health
to
#skinny
#ana
my
#bulimic
#exercise
#depressed
the
,
#motivation
#thighgap
#girl
a
photo
and
this
#mia
#sexy
#cute
#selfie
is
#gym
#abs
#depression
#hipbones
#inspiration
#quote
#bulimia
#size
#fitspiration
you
?

Page 11 - Topic 18 Example Tweets

tweet id 280444212112158721, 2 topic words out of 5 modelled words
#thinspo #runway #fashion. #gorgeous #JoanSmalls http://t.co/awCon2RG

tweet id 428759352057798656, 2 topic words out of 11 modelled words
#jeans #eatingdisorder #skinny #thinspire #thinspo #thighgap #thinspiration #thin #ana #summer #SpainSkinnerGirls http://t.co/KQ50TDjeUZ

tweet id 420404208425521153, 1 topic words out of 4 modelled words
#Thinspo #bonespo Goodnight ♡ http://t.co/h5iE2UJO91

tweet id 347763599411920896, 3 topic words out of 4 modelled words
#thinspo #thinspiration #weightloss #anasisters http://t.co/o5k4f28jxg

tweet id 353641036658130944, 7 topic words out of 10 modelled words
#anorexic #weightloss #proana #eatingdisorder #skinny #hipbones #chestbones #fat #anorexicthoughts… http://t.co/x5UhdpAZSw

tweet id 281566225732100097, 10 topic words out of 10 modelled words
#skinny #skinnygirls #thinspo #thinspiration #fit #weigh #loss #exercise #workout #want http://t.co/kwoQFMIG

tweet id 302514893108695040, 14 topic words out of 18 modelled words
#JustMe... I wanna be #Perfect.. #ProAna #Ana #Mia #perfectBody #Skinny #Slim #Beauty #Art http://t.co/eYGQc37b

tweet id 321990568093839360, 2 topic words out of 18 modelled words
The more cute clothes I have to workout in the more motivation I have #looksexy #motivation #pretty #strong #thinspiration

tweet id 365225522483044352, 3 topic words out of 3 modelled words
#thinspiration #size8 ???? http://t.co/Ql025jAfPO

tweet id 340572530622472192, 1 topic words out of 3 modelled words
#thinspo #progress #weightloss http://t.co/s3VSM7fkJB

tweet id 429397633493663744, 9 topic words out of 12 modelled words
#beforeandafter #thinspo #thighgap #summer #EDproblems #anorexic #ana #bulimia #perfect #motivation #motivate #fitspo http://t.co/p3FE9ks7oz

tweet id 398077104547725312, 6 topic words out of 6 modelled words
#pink #skinny #perfect #abs #thinspo #thinspiration http://t.co/JzzcktwJCh

tweet id 283765673006596096, 8 topic words out of 11 modelled words
Super classy tonight. #me #fat #ana #ED #picslip #thighgap #collarbones #lbd http://t.co/oGZyHzKo

tweet id 355163296325828608, 3 topic words out of 3 modelled words
#thinspiration #motivation #healthy http://t.co/EQV9jiyuxh

tweet id 329395559725150209, 4 topic words out of 4 modelled words
#Thinspo #SelfHarm #Skinny #Cutting http://t.co/RXL08MkDGR

tweet id 396288805742714880, 1 topic words out of 4 modelled words
#thinspo #skinny #thighgap #thinspiration http://t.co/Ybf4lDsnWr

tweet id 334780075440500737, 4 topic words out of 6 modelled words
Gold dust skeleton. #thinspo #love http://t.co/0af86ObSAo

tweet id 349128159456800769, 1 topic words out of 5 modelled words
omg this please #thinspo #thighgap http://t.co/JZYdJ9s4jO

Page 12 - Topic 5

thigh
gap
#thighgap
#thinspo
′
thighs
a
#thinspiration
?
gaps
#edproblems
have
:
#proana
#skinny
touch
#ana
feet
apart
together
that
;
#diet
#weightloss
&
#eatingdisorder
your
girls
@huntermoore
perfect
#anorexia
doesn′t
exist
#proed
#picslip
bigger
between
#thin
about
#ed
don′t
ass
her
why
dat
obsession
#fitspo
day
no
eat
knees
big
you
lol
are
#anasisters
got
please
3
_
#motivation
#mia

Page 12 - Topic 5 Example Tweets

tweet id 391775324469743616, 1 topic words out of 13 modelled words
LOVE LOVE LOVE! RT @AdiosBarbie: No #ThighGap? NO PROBLEM! http://t.co/LTyuiXgypF #bodyimage

tweet id 291710550839287808, 9 topic words out of 18 modelled words
Nothing better than having skinny arms, legs, thighs and a stomach over summer. #thighgap #fastmetabolism #best

tweet id 342068051753193472, 8 topic words out of 8 modelled words
Spongebobs thigh gap is dope as fuck #spongebob #thighgap #mewant

tweet id 336827399834574848, 7 topic words out of 13 modelled words
My second name,Rose~a flower that cuts#thighgap #thinspo #crying

tweet id 419036923366477824, 3 topic words out of 7 modelled words
@NuttyMadam @PerezHilton wow @MileyCyrus has a #thighgap #Cool

tweet id 340389072575008769, 2 topic words out of 3 modelled words
My screensaver #thinspiration http://t.co/KMC0ZB7cD5

tweet id 405844897023795200, 13 topic words out of 24 modelled words
Argh how do u get skinny skinny thighs/legs without building any muscle or chub. #nicolerichie legs? □ #weight #fitness #thinspo #halpp

tweet id 359295518603636736, 4 topic words out of 14 modelled words
Her waist is so tiny and doesn't bulge out when she bends. #thinspo http://t.co/uToqTuGM4C

tweet id 306359856300953600, 12 topic words out of 17 modelled words
@AmyAnderssen1 http://t.co/cZzu2VL6hN Love this pic... feet together, thighs apart. Wow. #perfection #thinspiration #bigfakeboobs

tweet id 283331545975095297, 2 topic words out of 9 modelled words
Next christmas..... #Christmas #thinspo http://t.co/gpWVAuPn

tweet id 301502226705289216, 5 topic words out of 5 modelled words
progress? #fat #sorrynotsorry #sorryexcuse #thighgap http://t.co/5rehgwkx

tweet id 291749391184850945, 3 topic words out of 10 modelled words
Can I have this thigh gap please? #thighgap #thinspo http://t.co/h8scByrO

tweet id 387487064293269504, 9 topic words out of 10 modelled words
@leevin3 are you going for a thigh gap? #thinspo

tweet id 398848115391725569, 14 topic words out of 14 modelled words
Oh shit! I just noticed I don't have a thigh gap! #thighgap #dumbculture

tweet id 308960908422553601, 9 topic words out of 10 modelled words
come out come out wherever you are stupid bitch #hipbones http://t.co/wX5EXmfAs9

tweet id 357619371021041665, 3 topic words out of 26 modelled words
Trying to explain what you're going through, even if people have a slight insight in it, it just seems impossible to explain! #edproblems

tweet id 384856519390396416, 6 topic words out of 6 modelled words
mermaids dont have thigh gaps #thighgap

tweet id 303198866340200449, 4 topic words out of 5 modelled words
Thigh gap for dayyyzzzz #thinspo http://t.co/b3PHnZIy

Page 13 - Topic 19

#eatingdisorder
#weightloss
#diet
#proed
#proana
#thinspiration
#ed
#ana
#anorexia
#skinny
.
@proanam
’
I
#thin
!
#mia
to
:
#promia
my
the
#thighgap
,
a
and
this
sanctuary
#reasonstoloseweight
is
#bulimia
@ed
you
?
proanam

Page 13 - Topic 19 Example Tweets

tweet id 400304669509308416, 8 topic words out of 8 modelled words
#diet #weightloss #eatingdisorder #thinspo #proed #proana #thinspiration #skinny http://t.co/zp8enG8DRi

tweet id 311527329451110400, 4 topic words out of 11 modelled words
This is all I need #thinspo #weightloss #ana #ed #ribcage #bones http://t.co/E7C9O2L0t6

tweet id 362929802723856384, 3 topic words out of 10 modelled words
thinspo #proana #ana #thighgap #proed #ed #mia #promia #skinny #thinspo http://t.co/xpqBT14Ke5

tweet id 393326620292816896, 1 topic words out of 4 modelled words
#thighgap #perfect #thinspo #thinspiration http://t.co/4HtLTIRj1Y

tweet id 407376856455712768, 3 topic words out of 12 modelled words
#thinspo #thinspiration #thighgap #tumblr #Skinny4Xmas #decemberdiet #weightloss #motivation #proed #proana #ednos http://t.co/DC7OYZ8zxc

tweet id 392774815461163008, 9 topic words out of 15 modelled words
“@imdyinginside96: RT @ProAnaM: #diet #weightloss #skinny #anorexia #thinspiration #proed #proana #eatingdisorder http://t.co/BFIQS6KDMN”

tweet id 414289828344119296, 4 topic words out of 4 modelled words
#thinspiration #thinspo #ana #mia http://t.co/rd5kwSxKnt

tweet id 285896727465181184, 9 topic words out of 12 modelled words
″@ProAnaM: #diet #proana #thin #thinspiration #weightloss #eatingdisorder #eatingdisorderawareness #skinny http://t.co/PvUlnJmi″

tweet id 281906556897992706, 1 topic words out of 5 modelled words
Tonight was the night #promia

tweet id 395645254957006848, 4 topic words out of 7 modelled words
Starting saturday #abcdiet #ana #diet #r??quilibragealimentaire #thin #weightloss http://t.co/vwjDBQT3y7

tweet id 392406551635312640, 9 topic words out of 9 modelled words
#diet #weightloss #ed #anorexia #eatingdisorder #proana #proed #thinspiration #ana http://t.co/bWyxpmBBdc

tweet id 365428604152209408, 8 topic words out of 8 modelled words
#diet #proed #WeightLoss #thin #skinny #proana #eatingdisorder #thinspiration http://t.co/RzCT8Ffwyg

tweet id 311099572942409728, 8 topic words out of 9 modelled words
#weightloss #proana #skinny #proed #diet #EDproblems #thinspiration #Thinspo #eatingdisorder http://t.co/7wd1aJv97x

tweet id 394539996914257920, 7 topic words out of 8 modelled words
#thighgap #Ana #Mia #thinspo #weightloss #diet #eatingdisorder #thinspiration http://t.co/83Jg5SUb0w

tweet id 322308963142098945, 6 topic words out of 6 modelled words
#diet #weightloss #proana #beforeandafter #proed #thinspiration http://t.co/QNKqzCFR1s

tweet id 310266919339061249, 8 topic words out of 8 modelled words
#diet #weightloss #thin #thinspiration #skinny #eatingdisorder #proana #proed http://t.co/WOXmtWT0Nx

tweet id 329267551806844928, 10 topic words out of 15 modelled words
″MyMindKilledMe: RT @ProAnaM: #diet #thinspo #skinny #proana #thinspiration #proed #ana #eatingdisorder #weightloss http://t.co/ZRq3zmG1xU″

tweet id 316040314144059392, 1 topic words out of 5 modelled words
#thinspo #thinspiration x5 #thighgap http://t.co/H7POruIKqp

Page 14 - Topic 14

′

you

:

!

#thinspiration

people

are

it

#skinny

_

girls

not

what

they

#thinspo

#ana

we

if

don’t

that

;

#diet

#weightloss

&

#eatingdisorder

my

her

.

think

do

?

,

know

tag

#anorexia

legs

say

#proed

it’s

who

but

#thin

#hipbones

#ed

how

all

is

or

about

me

#thighgap

sad

make

pictures

themselves

you’re

Page 14 - Topic 14 Example Tweets

tweet id 382396565773488128, 12 topic words out of 20 modelled words
Well wouldn't ya know what I thought would be an all male comedy revue @TheParlorHW opened with @WhitneyCummings! Great set! #thinspiration

tweet id 424638265317027841, 11 topic words out of 21 modelled words
personally, I always assumed a #thighgap meant you were just free and loose with sharing your nether regions. |o=

tweet id 303384876353191936, 16 topic words out of 21 modelled words
What is wrong with me? #anafamily #anasisters #selfharm Is there such a thing as a person who doesn't cheat? http://t.co/I0zKAU9p

tweet id 225166057260847104, 3 topic words out of 7 modelled words
Yeeeess girls keep on going!!#thighgap http://t.co/rleGDJLR

tweet id 336984572573130752, 13 topic words out of 25 modelled words
"@ryanx28: "@TBertelsman: #thinspo is no longer available on Instagram. [?]" I know, it's devastating" why???

tweet id 376866361059704832, 10 topic words out of 10 modelled words
need this, not even joking. . . #thinspo http://t.co/P4kb8vBr3N

tweet id 342553555711959040, 4 topic words out of 4 modelled words
@EmmaHaff haha #thighgap better happen

tweet id 224808739868393472, 10 topic words out of 12 modelled words
Don't lie to yourself. Choose a realistic goal.. #Thinspo http://t.co/aN3XPlJb

tweet id 300595392289050624, 1 topic words out of 24 modelled words
#EDProblems having meal eg stew/ soup and having to look at recipes find the individual cal content of ingredients then add together :/

tweet id 353802982598512640, 2 topic words out of 7 modelled words
My minature thigh gap. #thighgap #sitingdown #fatty #fatspo http://t.co/wuWMujdOM0

tweet id 320930503966461955, 19 topic words out of 20 modelled words
Most girls my age are worrying about their boyfriends. I'm worrying about being fat. How lovely. #EDproblems

tweet id 302261929290891264, 8 topic words out of 17 modelled words
Idk who this is but this picture hangs proud on my mirror. #thinspiration she is beautiful ! http://t.co/DYzxxehd

tweet id 417850360816881665, 20 topic words out of 30 modelled words
People who unabashedly use the #proana tag <<< Either you dont understand the sickness of eating disorders or you actually HAVE the sickness

tweet id 411320757281501184, 6 topic words out of 13 modelled words
#Ana, #Proana, #Thinspo, #Thinspiration, #CollarBones,#PerfectBody, #SkinnyGirl http://t.co/3P3uEDjRsy

tweet id 311614474366894080, 11 topic words out of 11 modelled words
@CharliieSimmons @momardiagne you don't wanna be any kind of meat. #ano #thinspiration

tweet id 293218908947759104, 5 topic words out of 17 modelled words
#Beauty isn't a number on a scale. #EDProblems But we will listen. Share your story http://t.co/XaXjstd5

tweet id 354971177472630784, 21 topic words out of 26 modelled words
"@Brittney_Tara: People that hashtag themselves on instagram as hot, beautiful, sexy, etc. Just... NO! #instafail" #thinspo

tweet id 295868804729298944, 6 topic words out of 23 modelled words
#thinspiration for 2013!!! Just like what they say "DREAMS ARE FOR FREE" lol. I can do thiiiiiiis! http://t.co/YekNSO9t

Page 15 - Topic 3

#anasisters

?
follow
’
#anafamily
me
some
#thinspiration
you
please
:
help
anyone
#thighgap
(:
need
#edfamily
send
guys
for
any
)
#skinny
followers
(
account
tweet
if
skinnies
stay
someone
#replytweet
#edproblems
strong
#help
bio
read
xo
;
http://t.co/X5bJvQH1P1
#diet
#weightloss
thnx
&
goodnight
#miasisters
#eatingdisorder
motivation
her
tips
.
here’s

Page 15 - Topic 3 Example Tweets

tweet id 425034985431511041, 11 topic words out of 16 modelled words
@ED_Challenges_ I love you tweets and twitter page so helpful and #relatable #beautiful #edproblems

tweet id 364081482534354944, 3 topic words out of 19 modelled words
#Thinspiration have to say (without soundin like a conceited twat) i really like my legs in this..#GiveMeMyBodyBack http://t.co/xcXS290Zni

tweet id 319246402813308929, 9 topic words out of 9 modelled words
I need to talk please?!? #anasisters

tweet id 355904117677371395, 16 topic words out of 16 modelled words
Starting the abc diet tomorrow. Who wants to join me?? #abcdiet #ed #anasisters

tweet id 315446874633035776, 2 topic words out of 10 modelled words
? Ow. My feet. #tired #somuchwalking #EDproblems goodnight y'all

tweet id 369416629701971968, 4 topic words out of 12 modelled words
I hope some day that my #weight vil be perfect like #proana girls

tweet id 344691621020975106, 10 topic words out of 10 modelled words
Fasting @jesslovesanamia . We can do this!! #thinspo #motivation http://t.co/7ljub0ETah

tweet id 281510426494259200, 10 topic words out of 10 modelled words
Going to post some #Thinspo !! Enjoy :)) xx

tweet id 387462286249443328, 8 topic words out of 21 modelled words
I think it's time for me to give up on #thinspo #thinspiration #proana for like the rest of the year.

tweet id 316502745379123200, 5 topic words out of 5 modelled words
#thinspo uploading thinspo's for motivation http://t.co/QNWFrgP0Qa

tweet id 412873050976026624, 7 topic words out of 9 modelled words
Any Aussie people wanna talk? #selfharm #proana #thinspo #selfinjury

tweet id 357194242185117696, 7 topic words out of 7 modelled words
Thinspos? Anyone? #help #anasisters #EDfamily

tweet id 336396334275895296, 6 topic words out of 6 modelled words
Good morning #anasisters ! #staystrong today

tweet id 316955467861610496, 21 topic words out of 25 modelled words
I'm gonna do a 24h fast but for every RT I'll fast one hour more. Please help me and RT!! #anasisters http://t.co/ZP8mGd7P3j

tweet id 292376594452996096, 7 topic words out of 18 modelled words
So basically I'm really fat now. Any exercises to help lose weight? #anafamily #anamiafamily #EDProblems #edgirlprobs

tweet id 285894053718007810, 9 topic words out of 15 modelled words
#StayStrongfor1moreYear RT <3 to all the beautiful people with #SH and #EDproblems :)

tweet id 304819427566813184, 8 topic words out of 8 modelled words
Here's a goodnight #thinspo for you night. http://t.co/tVLxeXizhi

tweet id 369975903042150400, 3 topic words out of 12 modelled words
Give me a bony body and i'll be forever grateful . #edproblems

Page 16 - Topic 8

′

the
in
:
!
my
on
a
of
at
?
#skinny
was
_
phone
#proana
#ana
you
saw
watching
;
girl
looking
#diet
#weightloss
&
#eatingdisorder
secret
perfect
#anorexia
top
pictures
jeans
seen
#thinspo
want
best
clothes
wearing
#proed
#thin
picture
do
#ed
fashion
size
school
show
old
victoria′s
don′t
ever
eat
and
#anasisters
please
shorts
fit
3
while
#mia
#perfection

Page 16 - Topic 8 Example Tweets

tweet id 311217229935022080, 5 topic words out of 16 modelled words
@mcneill1 look at @PinkBoutiqueUK for outfits for #ibiza2013..they're amaze #thinspiration x x http://t.co/voP66m9JCs

tweet id 375235227779481600, 9 topic words out of 12 modelled words
I'm freezing!! And it's summer.... #EDproblems

tweet id 322068053741420546, 4 topic words out of 11 modelled words
@JoshuaNLemon u would look good wit one, would compliment yr #thighgap

tweet id 413471850585661440, 8 topic words out of 8 modelled words
I have a long way to go #thinspo http://t.co/BAyArhCLo8

tweet id 306065589112303618, 5 topic words out of 24 modelled words
#anafamily #AnaSisters #helpme .... What's the most amount if weight you've lost in a month? And how...

tweet id 283843639250145280, 10 topic words out of 15 modelled words
"@thinandsick: Found a picture I took a month ago but never posted http://t.co/xoutHiHb" #thinspo

tweet id 281196500288147456, 11 topic words out of 16 modelled words
apparently #Jesy from @LittleMixOffic became #bulimic after the sheer amount of comments about her #weight ... #celebsecrets

tweet id 289509702222626816, 15 topic words out of 19 modelled words
Going to pin Victoria secret models around my uni room. And maybe on the fridge.. #thinspiration

tweet id 334347461265068032, 11 topic words out of 16 modelled words
You never have to worry about running out of food on a model shoot. #thighgap

tweet id 342041119288340480, 14 topic words out of 14 modelled words
@rachelizondo witchoo double laced up all the way extra tight White girl chuck Taylor's no socks goofy looking ass #NoSwag #ThighGap

tweet id 286865575249076224, 15 topic words out of 25 modelled words
Watching skins for #thinspo effy is super gorgeous, I wanna wear my docs with just a top and fishnet tights. This is my goal

tweet id 299148598208253952, 16 topic words out of 26 modelled words
"@ranger_66: @eggbilby Kelly Brook is how a real woman should look, curves in all the right places and a local lass. #hothothot" #thighgap

tweet id 301327452234452993, 8 topic words out of 17 modelled words
so looking forward :)"@f00l_tobelieve: #thinspo What I will wear when I'm skinny http://t.co/caEBRo8V"

tweet id 326660263510155264, 13 topic words out of 15 modelled words
Is it weird that I run better listening to hot female pop stars? #thinspiration

tweet id 336526844234969088, 11 topic words out of 14 modelled words
need some new #thinspiration all the pictures I have are doing nothing now #fatgirlproblems

tweet id 345585551535128578, 7 topic words out of 8 modelled words
#thinspo that was all my current faves :) http://t.co/2hUOYyS8Dp

tweet id 380722195062087680, 5 topic words out of 12 modelled words
Gonna buy the neverfull mm when I loose 10 lbs. #inspiration #thinspiration

tweet id 309073937785049088, 9 topic words out of 13 modelled words
My all time favorite #thinspo I forgot who I got it from. http://t.co/qOHrBCAFRl

Page 17 - Topic 9

lbs
#abcdiet
’
day
#thinspo
2
pounds
1
)
(
lost
5
10
days
30
you
starting
.
today
#skinny
4
cals
_
tomorrow
week
/
6
500
20
100
#thighgap
lose
7
weeks
calories
8
hours
want
diet
months
;
&
#eatingdisorder
since
(:
her
kg
fasting
fast
like
#day
weight
perfect
minutes
i’ve
#anorexia

Page 17 - Topic 9 Example Tweets

tweet id 282973453039063040, 7 topic words out of 9 modelled words
145 pounds, weight loss starts today #proana #weightloss

tweet id 355022088890884097, 11 topic words out of 18 modelled words
Rush is in a month. I just can't be fat. I have to finish the #ABCDiet

tweet id 295434652661972992, 7 topic words out of 11 modelled words
#thinspo number 21 (old). collarbones and my legs http://t.co/XQbfIeZQ

tweet id 312312776435589120, 24 topic words out of 24 modelled words
fast started 4minutes ago. can't eat now for 24hours...so 8.15 tomorrow night #icandothis #proud #thinspo

tweet id 315334317159088128, 8 topic words out of 17 modelled words
#diet #weightloss #thinspiration #proed Posting on behalf of ProAnaM. She will be back next week :)
http://t.co/HXReurSCOS

tweet id 303734900287234048, 2 topic words out of 6 modelled words
Mmmmm.... #thinspo http://t.co/2gWrFra4

tweet id 285404420102049792, 3 topic words out of 7 modelled words
#Picslip working on my #thighgap #scars #cutter http://t.co/WoBxfePB

tweet id 322605572891504640, 11 topic words out of 11 modelled words
Just did 80 leg lifts on each leg (: #ThighGap #IWillBeSkinny #ProveThemWrong

tweet id 328636382664871936, 5 topic words out of 5 modelled words
Target for summer #letsgo #hourglass #thinspo ?? http://t.co/v5iNkJ1WNG

tweet id 312331135550119936, 15 topic words out of 17 modelled words
100 cal so far today...light headed and dizzy...:(#proana

tweet id 340576423343435776, 15 topic words out of 15 modelled words
4 betches best drop pure barre tomorrow I will be going to the 11am class #thinspiration #shake #waitlist

tweet id 311810906248069121, 8 topic words out of 27 modelled words
Day three of the #ABCdiet gonna fast till about 3 and then have this soup with noodles and veggies tht's 300 cals. Wish me luck dolls :)

tweet id 355372361299668993, 4 topic words out of 11 modelled words
37 days to get the #thighgap i can do this :) #goingtoshowyouwhatyoumiss

tweet id 309173702849884160, 13 topic words out of 21 modelled words
Fasting every other day for the next 2 weeks. Gotta look great for my dance partners wedding. #danceproblems #proana #edproblems

tweet id 423951706410332160, 13 topic words out of 13 modelled words
Tomorrow is going to be so difficult. :/ #ABCDiet Day 5.

tweet id 287024511688273920, 6 topic words out of 8 modelled words
really though who's done the #abcdiet before?

tweet id 286598110812192769, 8 topic words out of 10 modelled words
Setting myself up for the reverse diet. #wishmeluck #anasisters

tweet id 354298674349756416, 5 topic words out of 10 modelled words
Wearing laces today because summer finally started. #OOTD #PicSlip #Thinspo http://t.co/i8ezvm3mAu

Page 18 - Topic 16

be
like
this
I
look
want
can
’
#thinspo
to
please
i
her
wish
:
?
the
will
#edproblems
body
have
would
able
why
#proana
of
_
looked
wanna
wear
one
#ana
can’t
thin
let
http://t.co/mIvfbbDtU3
kill
do
;
!
#diet
#weightloss
&
#skinny
#eatingdisorder
i’d

Page 18 - Topic 16 Example Tweets

tweet id 319993463397548033, 7 topic words out of 19 modelled words
THIS is how I want my legs to look when I wear jeans. Perfect. #thighgap #thinspo #thinspiration http://t.co/CXKV1WnmXn

tweet id 321672751356186625, 3 topic words out of 12 modelled words
Doing it for the dainty yet strong legs! #thinspo #fitspo #finspo http://t.co/wdzc2lzKZw

tweet id 402921961120337920, 2 topic words out of 5 modelled words
please #thinspo #thinspiration #ed #ana http://t.co/rXcoTTLgdm

tweet id 373191470427021312, 11 topic words out of 12 modelled words
I will look like this one day... #thinspo #perfection http://t.co/gWB7dZX420

tweet id 294560438539149312, 7 topic words out of 15 modelled words
I'm not sure if I wanna be her or if I wanna fuck her #thinspo http://t.co/jOJ8IVnt

tweet id 396295460941332480, 8 topic words out of 14 modelled words
If I wore jeans like that my fat would be hanging out :(#thinspo http://t.co/dSPk0sNur3

tweet id 314781885240836096, 17 topic words out of 17 modelled words
Major #thinspo can I please please look like that?! I swear I'd be happy. http://t.co/lAfLjaL2Em

tweet id 287057592818208771, 9 topic words out of 18 modelled words
Want my hair this color and to be back this skinny in 2 months □ #unrealistic #birthday #tbt #thighgap #twig http://t.co/YEMtUWxY

tweet id 374362217363554305, 5 topic words out of 9 modelled words
"@DaintyAna: #Thinspo http://t.co/KeDhWRiJSz" can i please?!

tweet id 342479943965483008, 10 topic words out of 19 modelled words
"@imagirlwhocuts: I love this. #thinspo #ana #perfect http://t.co/RXG12GFEI7" I would kill to look like that..

tweet id 325686704621105154, 7 topic words out of 8 modelled words
Want to look like this #thinspo #getfit #motivation http://t.co/HBUnwy1TXd

tweet id 414785689663062016, 12 topic words out of 23 modelled words
WHEN I become a skinny bitch, nobody will be able to hold me & my bikini selfies back. #thinspiration

tweet id 300614647088820224, 8 topic words out of 8 modelled words
I want to be like this again #thinspiration http://t.co/nEYViwur

tweet id 374264354667245568, 5 topic words out of 6 modelled words
this body would be perfect #thinspo http://t.co/xzOyR9kIUJ

tweet id 317749795433357312, 5 topic words out of 22 modelled words
"@xxaanax: Don't you wanna be beautiful? #bath #thinspo #thinspiration #skinny http://t.co/OUa6htCs7F" I suddenly love my fat...

tweet id 427486348467916800, 4 topic words out of 4 modelled words
#thinspo I want this http://t.co/2BQ4D9YZTt

tweet id 313759617530347520, 2 topic words out of 15 modelled words
Okay so get home from school and #thinspo is all I can think about.

tweet id 313266535026282496, 3 topic words out of 5 modelled words
#thinspo #thinspoquotes I will. http://t.co/MB71ZJpoV4

Page 19 - Topic 15

!

*

THIS
MY
’
ME
YOU
IS
SO
#THINSPO
WANT
TO
OMG
THE
to
HER
LEGS
PLEASE
BE
AND
my
the
yes
CAN
IT
a
LOVE
LOOK
and
LIKE
BODY
is
DO
NEED
FOR
FUCKING
OH
HAVE
A
FUCK
FAT
PERFECT
you
NO
WHY
GAP
NOT
THIGH
ARE
YOUR
SKINNY
of
so
ALL
YES
JUST
OF
NOW

Page 19 - Topic 15 Example Tweets

tweet id 315381766808551424, 4 topic words out of 7 modelled words
I LOVE this. #thinspo #flexible! http://t.co/BazTEZ9yqj

tweet id 358351627910000640, 5 topic words out of 29 modelled words
Got told tonight I 'need to eat a burger & fries'. Fuck you you fat bastard eat a salad!!!! #EDproblems

tweet id 288091584359182336, 1 topic words out of 6 modelled words
#Thinspo #perfection #Ana #Mia #ME #FAT #sad #DIE

tweet id 411915569957724160, 3 topic words out of 23 modelled words
Every time I think ME & my LIFE are getting better, God reminds me that I'm destined for nothingness and #depression #EDproblems

tweet id 335251302105899009, 4 topic words out of 6 modelled words
@Alevieu AGGGGGGGGGGGGGGG QUE TAL PERRA LO DICE LA CABEZILLA DE #PROANA Y #PROMIA #anorexia #saveher #please

tweet id 363209977449549824, 10 topic words out of 13 modelled words
GUYS MY COUSIN IS SUPER SKINNY SHE'S LIKE MY PERSONAL #Thinspo (Deleting) http://t.co/Ro6nVa9UFK

tweet id 405238850911367168, 9 topic words out of 15 modelled words
Can I be her!!! Candice is wow!!! #thinspo #thinspiration http://t.co/8EqvmDOoIH

tweet id 297111691278295040, 3 topic words out of 4 modelled words
FUCK PERFECTION #Thinspo #Thinspiration http://t.co/3cabVZqZ

tweet id 383089728188203008, 6 topic words out of 6 modelled words
@xoxoshyy @Ryanna_hammond #thighgap!!!!

tweet id 410351538649837569, 3 topic words out of 3 modelled words
THIS! #thighgap

tweet id 383361714445881344, 6 topic words out of 24 modelled words
"@Skinnyasyou: Constant battle between eating to not pass out and not eating to pass out #edproblems" omg omg yes!! :(

tweet id 369502798548721664, 5 topic words out of 5 modelled words
DONE BEING FAT! #proana

tweet id 337659531377115137, 5 topic words out of 6 modelled words
OH WOW HER LEGS #thinspo #thin http://t.co/wPPY2sQz0F

tweet id 380783377567059968, 4 topic words out of 12 modelled words
Karlie Kloss' body is just SO oh ma gah! #thinspo http://t.co/cfMV5my41v

tweet id 324427768949530624, 4 topic words out of 7 modelled words
#thinspo :O this!!! http://t.co/yWrR6Jgetg

tweet id 285601169898864641, 1 topic words out of 4 modelled words
#Thinspo fucking perfect. http://t.co/EECqS9NJ

tweet id 366831622487699456, 3 topic words out of 9 modelled words
Omg i want this!!! #thinspo ♡ http://t.co/P4KQXyLtG5

tweet id 395628892956143616, 9 topic words out of 18 modelled words
"@solo_true: SHES SO PERFECT❤❓❤❓❤❓ MY GOAL #thinspo #proana http://t.co/b4lZ5pzpn8" w.t.f