Language change and evolution in Online Social Networks

Daniel James Kershaw, BSc (Hons), MRes

Highwire Centre for Doctoral Training

School of Computing and Communications

Lancaster University

Submitted for the degree of

Doctor of Philosophy

October, 2018

Declaration

The material presented in this thesis is the result of original research by the named author, Daniel James Kershaw, under the supervision of Dr. Matthew Rowe, Dr. Patrick Stacey and Dr. Anastasios Noulas; carried in the Highwire Doctoral Training Centre at Lancaster University.

This work was funded by the Engineering and Physical Sciences Research Council (EPSRC). Parts of this thesis are based upon prior publications by the author, listed below—all other sources, if quoted, are attributed accordingly in the body of the text.

1. D. Kershaw, M. Rowe, and P. Stacey, "Language innovation and change in on-line social networks", in *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, ser. HT '15, Guzelyurt, Northern Cyprus: ACM, 2015, pp. 311–314, ISBN: 978-1-4503-3395-5. DOI: 10.1145/2700171.2804449. [Online]. Available: http://doi.acm.org/10.1145/2700171.2804449

2. D. Kershaw, M. Rowe, and P. Stacey, "Towards modelling language innovation acceptance in online social networks", in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, ser. WSDM '16, San Francisco, California, USA: ACM, 2016, pp. 553–562, ISBN: 978-1-4503-3716-8. DOI: 10.1145/2835776.2835784. [Online]. Available: http://doi.acm.org/10.1145/2835776.2835784

3. D. Kershaw, M. Rowe, A. Noulas, *et al.*, "Birds of a feather talk together: user influence on language adoption", in *Hawaii International Conference on System Sciences*, Hawaii International Conference on System Sciences, Jan. 2017, ISBN: 9780998133102. DOI: 10.24251/HICSS.2017.225. [Online]. Available: http://hdl.handle.net/10125/41379

Funding Statement

This work is funded by the Digital Economy programme (RCUK Grant EP/G037582/1), which supports the HighWire Centre for Doctoral Training (http://highwire.lancaster.ac.uk).

Language Change and Evolution in Online Social Networks

Daniel James Kershaw

Abstract

Language is in constant flux, whether through the creation of new terms or the changing meanings of existing words. The process by which language change happens is through complex reinforcing interactions between individuals and the social structures in which they exist. There has been much research into language change and evolution [191], though this has involved manual processes that are both time consuming and costly. However, with the growth in popularity of Online Social Networks (OSNs), for the first time, researchers have access to fine-grained records of language and user interactions that not only contain data on the creation of these language innovations but also reveal the inter-user and inter-community dynamics that influence their adoptions and rejections. Having access to these OSN datasets means that language change and evolution can now be assessed and modelled through the application of computational and machine-learning-based methods. Therefore, this thesis looks at how one can detect and predict language change in OSN, as well as the factors that language change depends on. The answer to this over-arching question lies in three core components: first, detecting the innovations; second, modelling the individual user adoption process; and third, looking at the collective adoption across a network of individuals.

In the first question, we operationalise traditional language acceptance heuristics (used to detect the emergence of new words) into three classes of computation time-series measures computing the variation in *frequency*, *form* and/or *meaning*. The grounded methods are applied to two OSNs, with results demonstrating the ability to detect language change across both networks. By additionally applying the methods to communities within each network, e.g. geographical regions, on Twitter and Subreddits in Reddit, the results indicate that language variation and change can be dependent on the community memberships.

The second question in this thesis focuses on the process of users adopting language innovations in relation to other users with whom they are in contact. By modelling influence between users as a function of past innovation cascades, we compute a global activation threshold at which users adopt new terms dependent on exposure to them from their neighbours. Additionally, by testing the user interaction networks through random shuffles, we show that the time at which a user adopts a term is dependent on the local structure; however, a large part of the influence comes from sources external to the observed OSN.

The final question looks at how the speakers of a language are embedded in social networks, and how

the networks' resulting structures and dynamics influence language usage and adoption patterns. We show that language innovations diffuse across a network in a predictable manner, which can be modelled using *structural*, *grammatical* and *temporal* measures, using a logistic regression model to predict the vitality of the diffusion. With regard to network structure, we show how innovations that manifest across structural holes and weak ties diffuse deeper across the given network. Beyond network influence, our results demonstrate that the grammatical context through which innovations emerge also play an essential role in diffusion dynamics - this indicates that the adoption of new words is enabled by a complex interplay of both network and linguistic factors.

The three questions are used to answer the over-arching question, showing that one can, indeed, model language change and forecast user and community adoption of language innovations. Additionally, we also show the ability to apply grounded models and methods and apply them within a scalable computational framework. However, it is a challenging process that is heavily influenced by the underlying processes that are not recorded within the data from the OSNs.

Contents

Chapter 1

Introduction

Language is in constant flux, from the creation of new terms (innovations) to the subsequent adoption of new terms over time. Traditionally, large-scale assessment of language has been limited due to time-consuming sampling methods. However, the growth of OSNs provides a new opportunity to model and predict language evolution at a fine resolution over time.

Innovations in language can take many forms, from the short-term introduction of new words to long-term changes in grammar. However, it is the succession of small sequential adoptions of innovations over time that manifest as language change and/or evolution [45]. The process through which these individual innovations are created and adopted is not a simple process, but relies on the complex interplay of individuals and social structure to facilitate both their formation and propagation. Ultimately, this means that language and language innovation are a social phenomenon [74], heavily influenced by social structures, networks and norms. Thus, even within one formalised language (e.g. English), there are many variations used by different groups [33]. These variations in languages are heavily influenced by differences in social variables across populations, such as class [129] or geographical location [192]; these social variables subsume the network structures that mediate with whom individuals communicated [5]. Historic analysis of large-scale social variables, such as geographical location [148], identified that this structuring of social life around social variables both constrains and facilitates the diffusion of innovations, as these social variables subsume the communication points between communities that use variations in language [29]. This facilitation and hindering of the diffusion of innovation can be seen in Iceland, where there has been little change over an extended period of time compared to languages such as English [149], which was diffused extensively around the world through the spreading of the British Empire since the late 1700s.

Studying language through social variables only highlights the collective adoption and creation of language innovations, not the process that the individual goes through to adopt/accept new innovations. This process of change be seen on an individual level, with users accommodating their language to match

that of people in more powerful positions [191]. This effect of language adoption has been shown to occur within OSNs, as [115] showed that users in Wikipedia[1] change their language to that of the editors who are more powerful or authoritative than themselves. Users not only accommodate their language to individual users but also to a whole community that they want to join [53]. This adoption of language is also seen in offline environments such as the diffusion of language in New York between social classes at points of contact in the workplace [148], but only in terms of the lower classes adopting the language of the higher class. However, again, these studies focus on the high-level adoption of language, using small samples of data without explaining the process the user goes through to adopt a language.

As highlighted above, language change and evolution is a process that takes place over time; collectively, many innovations fail to become accepted and some aspects of a language die. To identify these changes in language on an individual basis (a given word), one can look at subsequent editions of dictionaries, as a dictionary documents the current state of a language along with the state of the language in the past. It is not only changes in meanings and spellings that are documented, but also the emergence and death of words. However, dictionaries are time-consuming to compile and require large amounts of manual work in first finding the innovations and changes in language, and then quantifying whether their existence (or death) means they should be included or removed from the dictionary. This ultimately means dictionaries capture only high-level understandings of language and not regional or community-specific variations. To standardise the process of assessing new words for inclusion d in a dictionary, a number of heuristics have been developed [18]. Again, these must be applied manually to words and are time-consuming as the words must first be identified. Additionally, OSNs is a prominent source of language, meaning that the amount of data that needs to be processed is continually increasing [118].

One limitation of these traditional studies is their reliance on manual and time-consuming data collection and analysis, resulting in small-scale studies that are a comment on language change and evolution in high-level terms, which can easily be explained by traditional social variables. With the recent growth of OSNs, the traditional boundaries that limited communication have been blurred, as individuals can easily communicate throughout the world using one of the many tools freely available (e.g. Twitter, Facebook or email). This has resulted in language change no longer being influenced only by the physical communities surrounding individuals, but additionally by the communities that individuals are a part of and communicate with online. This means that variations in language may no longer be confined by dominant features such as geographical boundaries, but rather the people with whom a user communicates online, compounding the challenges of studying language change at scale.

[1] https://www.wikipedia.com

1.1 Research Questions

Detecting and understanding the dynamics of language change and evolution can be a time-consuming process. However, the growth of OSNs and big data processing frameworks gives us the ability to look at the process of language change in a timely and cost-effective manner for the first time. Even though language and language change is a large subject area, ultimately, we are concerned with one overarching question: **How can we forecast language change in online social media platforms, and what are the factors that upon which the change depends?** This can be broken down into four underlying questions, which, when combined, aim to answer the one high-level research question.

1.1.1 Research Question 1

How do we detect language change in OSN?

This question focuses purely on language innovation (new terms), its usage patterns and how the growth in popularity of a language innovation can be quantified. This will investigate both users' and communities' adoption of language innovations, quantifying adoption not only in its core form (the same spelling) but also through variations in form and the context in which word are used.

To achieve this, we computationally operationalise heuristics proposed by [18] and [144], which are used to assess the growth and death of terms in language. These heuristics are used traditionally to assess whether a new word should be included in a directory, by assessing terms across a number of categories, such as the popularity of the word to the contexts in which the word is used. By applying these heuristics to multiple abstractions of communities across datasets from Twitter and Reddit, we show how language innovations are accepted at different levels of community abstraction, identifying how innovations have both geographical and topical locality.

1.1.2 Research Question 2

What is the role of social constructs in language innovation adoption in OSN?

Social structure and the individual user are brought together within the second question. Individuals make the decision (conscious or not) to adopt or reject an innovation, whether from exposure to it from their neighbours or from media they have consumed. Can this process of user decision making (to adopt an innovation, or not) be extracted from the traces left on social media? If so, what does this reveal about the process and pressures that cause the adoption of an innovation?

In answering this question, we draw on the work of [86] to model user adoption of language innovations as a function of a learnt global threshold, which, once breached by exposure to an innovation, will lead to the user adopting the innovation. In predicting user adoption of an innovation, we treat the process as a two-stage machine-learning process. First, we learn the influence between users as a function

of past user interactions and then as values to compute the joint pressure on the user to adopt a new innovation. As this framework relies on explicit relationships between users, we extract social interaction networks from both Reddit and Twitter, again, with abstracts representing inter-user and inter-community relationships.

1.1.3 Research Question 3

How does network structure influence the diffusion of language innovations?

In contrast to the previous questions, the third research question looks at the social structures, as represented through the network topology, and how this influences individuals' and communities' adoption of language innovations. The motivation for this question comes from wanting to understand the structures that influence language adoption, as [87] showed that innovation diffusion is heavily influence by the position a user takes in a network, and that the individual's position in a community's structure will influence their ability to be receptive or resistant to innovations. This will then identify how innovations move between individuals and communities, aiming to model community interactions and connections proposed in the work of weak ties by [87] and the social reinforcement of [187].

[36] proposed that a memes diffusion could be predicted by using features drawn from it's diffusion across a network, such as the average degree distribution of the diffusion. They formulated the question in a different way to other methods: instead of predicting the final size of the diffusion, they asked at each stage within the diffusion whether the final size would double the current size, as this would mean that the model is not influenced by unbalanced classes. This framework is applied in assessing the predictability of the diffusion of language innovations across both abstractions of Reddit and Twitter, showing differences in the dynamics of language diffusion between users and communities of users. The results indicate a reliance on structural holes and weak ties to achieve larger diffusions of invocations.

1.1.4 Research Question 4

How to perform the research at scale

As highlighted previously, this research will rely on large amounts of data generated and collected from OSNs. The large volume of textual data that will be used is defined as 'big data', as stated by [118]. Thus, the final question, which is less academic in nature, looks at how to implement the systems used throughout this research at scale. The motivation for this comes from the movement of research from small curated datasets to large samples of generated data or real interactions; thus, there are challenges in applying theoretical/traditional models and methods to new large-scale datasets.

This thesis shows how we can apply common machine-learning and data-modelling methods, used across the three previous research questions, with big data technologies such as Apache Spark, Hadoop and HBase. In doing so, we demonstrate that we are no longer limited by the size of the data and that

we can develop systems that can be easily applied to other similar datasets.

1.1.5 Overview

Across the three core research questions, this thesis focuses on language innovation adoption (within the wider context of language change and evolution) across OSNs. However, each question focuses on different aspects of quantifying and predicting the rate and reasons for innovation adoption, not only on an individual basis but also at the community level.

This is first achieved by detecting the change in popularity of new words by quantifying their rates of change (Question 1.1.1). When focusing on the individual, we will predict when users are going to adopt a new word or term based on historic diffusion patterns (Question 1.1.2). Finally, when focusing on a whole population, we show how predictable the diffusion of the innovation is across the network (Question 1.1.3). However, as content within OSN is generated at a high speed, the size of the data produced is on the scale of 'big data'. For this reason, the overarching Question 1.1.4 focuses on how methods within each question are implemented.

1.2 Problem Relevance

Language is embedded in every aspect of human life, from communicating with work colleagues to consuming media or writing emails. As language change is constant in much the same way as changes in fashion, aspects of language come and go with trends and pressures. For this reason, the application of this thesis is not constrained to pure academia but is also applicable to industry (such as security and marketing) and practitioners alike, all of which have interests in understanding the current and future states of language.

Assessment of language innovations, variations and large-scale changes have and still receive large amounts of attention, particularly from the fields of linguistics and social computing. Studies take many forms, from understanding the use of metaphors in cancer communities [173] to understanding the dynamics of using either colloquial or formal language in messages [153]. However, the majority of studies focus on small, curated datasets such as communications within a company or group of people. Therefore, the identified results have limited repeatability across different datasets and are highly dependent on the context from which in the data was collected. As this thesis works within the realm of big data, its applicability to academia is testing the ability to take known small-scale models, such as innovation curves and threshold models [86], which have been shown to work on small, curated datasets, to large-scale, noisy social media data that are more representative of real-world communication and interpersonal processes.

However, the applicability of this work is not limited to the scalability of known theories. Exist-

ing NLP systems have been shown to have reduced accuracy due to language variation and small scale innovations, thus results can be used to improve existing systems whos accuracy's have degregated in reaction to increased usage of noisy social media data [75]. This change and variation in language causes problems for NLP systems such as sentiment analysis or POS taggers, where models are unable to deal with new phrases or usage of words. This means that, over time, the accuracy of the models decreases. The ability to predict and model the ways in which language changes will help future NLP systems to adapt to the changes in language they encounter, by becoming aware of the contextual signals that indicate the change process.

However, the impact of this thesis is not only limited to the realms of academia; the findings could also find application in the marketing and security sectors. The resulting impact can be broken down into two classes: message optimisation and message monitoring.

The former is predominately of interest to marketers who are concerned with communicating a message to a particular audience. For example, using the wrong language or targeting the wrong audience could result in the failure of a multimillion-pound campaign. However, the rate of change of language and the speed at which innovations are adopted means that a once 'on point' message may now look old and antiquated. Thus, there is a need for marketers to understand and pre-empt language adoption and change, thereby reducing the risk of a message or campaign failing or not performing to its potential. However, it is not only understanding the language but also understanding the processes that effect the adoption of innovation for a user; thus, understanding to whom a particular message should be communicated (as identified in question 1.1.2 and 1.1.3) will also minimise the risk of a campaign's failure.

The opposite of optimising a message for an audience is understanding the messages within a community, which is a concern in the security and surveillance industries. These industries have been seen to combat the increase in the number of threats seen online, from child grooming to terrorist activity. However, as found in the literature, methods are challenged by the dynamics of language as the language used by communities changes quickly. For this reason, models need to be retrained, though this could miss the early signs of key community changes. Being able to preliminary predict changes in communication patterns could pre-empt terrorist activities, or identify new terms use by child groomers online.

1.3 Thesis Outline

The remainder of the thesis will be divided into two parts: Part I will bring context to the work, with Chapter 4 initially introducing the methodology and data used in the research and prepossessing stages applied to the data in Chapter 5. The three research questions will be explored in Part II (Chapters 6, 7 and 8), with the research questions 1.1.4 detailed within each of the other three Chapters.

Finally, in Chapter 9, the three research Chapters are brought together, highlighting the results and how, collectively, they answer the four research questions. The conclusion, and future direction, are discussed in Chapter 10.

1.4 Publication List

During the PhD, I have had the opportunity to publish three articles originating from the research for this thesis. These papers form the basis of 2 chapters within this thesis; Chapter 6 is an extended version of [112], utilising larger datasets and expanding on the computation methods and implementations used. In addition, Chapter 7 is an extension of [109], again, introducing the computation methods and implementations used. All work used across the two chapters was original research led by myself, with guidance from Dr. Matthew Rowe, Dr. Patrick Stacey and Dr. Anastasios Noulas.

Publications Related to the Thesis

1. D. Kershaw, M. Rowe, and P. Stacey, "Language innovation and change in on-line social networks", in *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, ser. HT '15, Guzelyurt, Northern Cyprus: ACM, 2015, pp. 311–314, ISBN: 978-1-4503-3395-5. DOI: 10.1145/2700171.2804449. [Online]. Available: http://doi.acm.org/10.1145/2700171.2804449

2. D. Kershaw, M. Rowe, and P. Stacey, "Towards modelling language innovation acceptance in online social networks", in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, ser. WSDM '16, San Francisco, California, USA: ACM, 2016, pp. 553–562, ISBN: 978-1-4503-3716-8. DOI: 10.1145/2835776.2835784. [Online]. Available: http://doi.acm.org/10.1145/2835776.2835784

3. D. Kershaw, M. Rowe, A. Noulas, *et al.*, "Birds of a feather talk together: user influence on language adoption", in *Hawaii International Conference on System Sciences*, Hawaii International Conference on System Sciences, Jan. 2017, ISBN: 9780998133102. DOI: 10.24251/HICSS.2017.225. [Online]. Available: http://hdl.handle.net/10125/41379

Other Publications

In addition to the three publications related to the PhD, I have participated in a number of collaborations across Lancaster University. This has resulted in a further five publications, covering topics such as the implications of geo-location mobile application for users in their local surroundings [30], to text processing on historic corpora [88].

1. D. Kershaw, M. Rowe, and P. Stacey, "Towards tracking and analysing regional alcohol consumption patterns in the uk through the use of social media.", *WebSci*, pp. 220–228, 2014. DOI: 10.1145/2615569.2615678. [Online]. Available: http://doi.acm.org/10.1145/2615569.2615678

2. A. Gazzard, M. Lochrie, A. Gradinar, *et al.*, "From the board to the streets: a case study of local property trader", *Transactions of the Digital Games Research Association (ToDIGRA)*, vol. 1, no. 3, 2014

3. P. K. Stacey, D. Kershaw, N. Puntambekar, *et al.*, *Managing big data analytics projects*, 1st ed. Achamore Books., 2017, p. 110. [Online]. Available: https://dspace.lboro.ac.uk/2134/24465

4. I. Gregory, P. Atkinson, A. Hardie, *et al.*, "From digital resources to historical scholarship with the british library 19th century newspaper collection", *Zurnal Sibirskogo federal'nogo universiteta. Seria: Gumanitarnye nauki*, vol. 9, no. 4, pp. 994–1006, 2016

5. C. Butterworth, D. Kershaw, J. Devine, *et al.*, "Presenting the past: a case study of innovation opportunities for knowledge dissemination to the general public through pervasive technology", in *21st Annual Meeting of the European Association of Archaeologists*, 2015, pp. 391–392

1.5 Invited Talks

During the PhD, I have been invited to present the research relating to this thesis to both academics and businesses around the country. This has given me the ability to solicit feedback, not only on the methods but also the impact of this work on real-world audiences.

1. **June 2017** - Lancaster Summer Schools in Corpus Linguistics and other Digital methods 'Large scale NLP', *Lancaster University*

2. **November 2016** - Big Data and Language Diffusion, *Myongji University, South Korea*

3. **November 2016** - Language change and adoption in online social networks, *Myongji University, South Korea*

4. **July 2016** - Lancaster Summer Schools in Corpus Linguistics and other Digital methods 'Large scale data mining ,'Mining social influence from language diffusions', *Lancaster University*

5. **September 2016** - Is this talk goin', to be flek bruh?!, *BrightonSEO*

6. **April 2016** - Invited Lecture, Information Management (MA), *Loughborough University*

7. **April 2016** - Invited Lecture, Information Management (BA), *Loughborough University*

8. **May 2016** - Invited Talk, Language Change and Online Social Networks, *Sheffield University*, *Open University*, *Bath University*, *Loughborough University*, *Lancaster University*
9. **February 2014** - Invited Lecture, E-business (MSc), *Lancaster University*
10. **January 2014** - Invited Lecture, Data Science 101 (MSc), *Lancaster University*
11. **September 2014** - Invited Talk, Tweeting whilst drunk, *BrightonSEO*

1.6 Press Clippings

Aspects of the research within this these have received both national and international press attention. Below is a limited selection of the articles that have been published by external press organisations:

1. **January 2016** - New Scientist Article: Tweets and Reddit posts give snapshot of our changing language
2. **January 2016** - Metro Article: Tweets and Reddit posts give snapshot of our changing language
3. **November 2015** - British Library Labs Awards 2015: Research category award-winning project

Part I

Background and Methodology

Introduction

When assessing language in OSN, one must acknowledge the underlying social dynamics that influence the formation of social systems. For this reason, the literature review is broken down into two distinct chapters. Chapter 2 focuses on reviewing the social literature theory that influences the research questions and assumptions we apply to the social networks. Chapter 3 focuses on the computational analysis that has been applied within the fields of social computing, sociolinguistics and sociology.

This work is heavily influenced by structuration theory [74], which separates the individual and the social structure, and analyses the dynamics between the two. The three research questions individually aligning with the separation of the individual innovation (research question 1.1.1), the social structures that influence the propagation of innovations (research question 1.1.2) and the interplay of the individual and the social structure in the process of innovation adoption (research question 1.1.3). The application of structuration theory will be explored further in Chapter 2.

Chapter 2

Language Change and Social Structure

In a social system in which individuals have agency over their actions, change, whether in the social system or the actions that the system facilitates, is a constant. Language change can be observed in individual users' choice of language, be this the pronunciation they choose or new words they decide to adopt. The first part of the literature review focuses ultimately on fundamental theories of *language*, *language change* and the *social structures* in which language is formed. The literature is drawn from across disciplines, including *linguistics*, *sociology* and *information systems*.

2.1 Language and Change

Language is central to human life, allowing for communication of ideas, thoughts and emotions between individuals or groups, across both time and space. However, communication acts are formed by more than just word formations, rather consisting of all sounds, sights and smells that collectively exist in and around the act between speakers (with *formalised* language, commonly understood words and sounds only taking up part of the act). Defining acts of communication as the collective result of multiple stimuli does not limit the definition to just humans but also includes animals. However, there is a distinct difference between human and non-human communication, which is predominantly related to the ability to apply and learn the *structural form* of language (such as the existence of grammar) [45].

Chomsky believed that language (and thus the ability to learn and form sentences) is innate to a user [97], through the existence of a set of common logical rules to which all users have access, allowing them to produce and understand the sentences. However, defining language through a universal grammar is highly restrictive, limiting user variation in *morphological*, *phonological*, *syntactic* or *semantic* form, then breaking the grammar [97]. To overcome this restrictive definition, [45] proposed that, instead,

language is a population of *utterances* defined within and by a speech community. Utterances differ as they are an 'actual occurrence of the product of human behaviour in communicative interaction', e.g. a sequence of sounds [45]. The resulting definition of an utterance differs from the generative definitions by not limiting a language to a set of 'correct' reproductions of sentences and allows the incorporation of variation in language.

Additionally, a collection of utterances differs from a sentence (a collection of words), in that utterances are bound by the context in which they are used, which is in contrast to a sentence in which the meaning is constant no matter where or when it is consumed or created. Then, defining language as a set of utterances results in the associated grammar being the set of rules that each speaker holds and uses for the reproduction of language in their own context. [45] stated that the speakers of a language, then, do not and cannot know the whole grammar of a language, but rather possess a sub-population of the grammar that has been acquired and interpreted from the people with whom the speaker communicates. This means that [45]'s definition of grammar breaks away from the formal generative/universal theory, whereby grammar will generate all possible ontologically correct sentences that could ever be produced. However, this is not to say that 'formal' grammar (such as that maintained by central bodies, e.g. Académie Française) does not have a place in language. However, it acts to guide the language populations through the existence of a common 'global' interpretive scheme from which individuals can draw their language.

Additionally, Chomsky [38] believed that the human capacity for the reproduction and comprehension of grammar was separate from the repository of the grammar, due to the innate nature of the human ability to form language. Again, this definition results in a language that is resistant to change, stemming from the belief that humans and/or the individual do not have the capacity to change. However, from a social context, the grammar and language produced are believed to be one and the same, due to their existence within a duality, both influencing each other, in much the same way as the duality of structure [74]. This conflict between the two theories of language can be seen in the three misconceptions that [58] believed exist about language and language change as a whole:

Language is biologically fixed - The evolution of human linguistic capability is in direct relation to human biological evolution [45]. However, human evolution is not connected to language evolution, as human evolution is not believed to constrain the development of variations of languages around the world.

Language does not change - Aspects of language can be seen to change quickly, such as *phonics*. However, more central components such as *grammar* and *structure* change at a slower pace. This can be seen in generational language changes, whereby variations in languages can be attributed to children adapting a variant of language from their parents and passing it on to following generations [47].

Everyone is not in full command of their language - Humans go through a number of phases of language acquisition throughout their lives, with acquisition not stopping when a child becomes an adult. Throughout life, new idioms, social contexts or meanings are introduced; thus, the language learning process never stops [58].

Building on the definition of language by [45], the set of utterances are utilised by individuals to generate their *personal interpretive scheme* of language. This results in a language defined by the groups who speak that given set of utterances rather than the whole population of the base language, e.g.ẽveryone who speaks English. However, the association of a user and their language is not random, but influenced by others with whom they interact. [142] showed that individuals do not cluster randomly but rather due to similarities between personal characteristics (or social variables). Therefore, language is then defined by the social background and social network of the user(s), meaning that it is defined by the users that use a 'sociolect' through its reliance on the social backgrounds of users:

> ...a variety of lect is thought of as being related to its speakers' social background rather than geographical background [193]

Thus, a language can be assessed through the social variables of the sub-culture or class of people [128] within the community; with social variables including identity, gender, sex, interest or religion, to name a few [127]. A major social variable that influences a user's language is the location in which they reside [129], as this one variable subsumes a number of others such as defining the social network and class of a user.

However, individuals are influenced by many combinations of their social variables. This means the user can speak with the identity of many different *lects*. This effect of multiple identities can be seen through the layering of social variables resulting in the nesting of *lects* within larger language communities. However, the effect of nesting of social variables can lead to the relative isolation of language communities, which then allows for the extensive reinforcement of norms within the language (through only being exposed to their own language); additionally, this leads to a community's identity forming through the language that they collectively speak [149].

Even though language (*lects*) are defined by community usage, [58] showed that language is in constant flux due to the interactions of users, communities and social variables. [45] proposed a theory of *utterance selection* as a model for language change, drawing influence from the theory of evolutionary genetic change through sexual reproduction as a reference point for the development of the model. [45] proposed that, much like a gene being replicated and modified though reproduction, language also changes through user replication, though instead of genes, [45] proposed that, within language, it is a *lingueme*. A lingueme is the element that is reproduced and modified in the language, thus variation in the structure of the lingueme allows for variation in the meanings of the utterances. The total set of liguemes within a

population of speakers is called a lingueme pool, which bares resemblance to the gene pool. However, [45] stated that the term 'pool' indicates a loose structure within the population of the lingueme, but there is in fact a high degree of structure, which comes from the linguistic lineages and structure of the utterances. [45] also noted that language change is a two-step process: *innovation* and *propagation*. An innovation (new utterance and, by extension, word) within the evolutionary model is an altered replication of a lingueme; that is, the process of incomplete replication or modification of the utterance within the original context.

Propagation/normal replication is the reproduction of the utterance in it's original form. When this is reproduced with intent, then it is classed as convention, though if there is no intention in using the word, it is then classified as a convention as the lexical item is entrenched within the speech community.

Innovation/altered replication is the process of innovation; this is either intentional or accidental. When it is intentional, this could be for a number of reasons, such as to express oneself better. However, it could be that the speakers, if they innovate, could indicate that they are correcting their understanding within the context of the conversation.

Through the use of the term *language change*, attention is not only drawn to the difference in the states of a language at two points in time, but also gives an in-depth look at which components within the language have altered and the reasons for these alterations. This is done by separating the language change into structural (the structure of the language) and non-structural components (the social aspects of language), which then allows for the explanation of linguistic variation that cannot be explained solely by the structure of the language itself. Such changes are therefore not, so-called, 'free' variations but may in fact be correlated with extra-linguistic social features, such as social class, age and gender [67], [141]. However, [149] differentiated between *language change* (the high-level change in language across a whole population over time) and *speaker innovations* (the individual alterations to a language that a user makes). [149] proposed that it is the summation of all successful *speaker innovations* that ultimately cause *language change*, as language is spoken and defined by more than one user. [149] believed that there are three stages in *speaker innovation* causing *language change*:

1. The speaker innovates in their language with intentional or unintentional alterations; however, these are *not* replicated by fellow speakers.

2. The innovation is only replicated by people in the same community, thus only partial diffusion is achieved.

3. The innovation diffuses beyond the community in which it originated; when observed at a global level, this could be classified as part of *language change*. The extent of the diffusion is dependent

on social variables and the number of connections between communities.

However, there are many misconceptions about language change; one of the main ones is that language changes due to the collective force of a population aiming for a collective goal [45], e.g. an innovation is created to fill a lexical gap. But rather, it is survival of the fittest much in the same way as with genetics, with adoption being the mechanism of survival. This then means that language change is a *probabilistic* process, not a *deterministic* one [45]. Thus, consensus human thought is removed from the process of language change. However, this definition removes the human consensus from the process of language change, though it does not exclude it from use in the process of preserving or innovating within a language.

2.2 Language and the Internet

Historically, acts of communication have taken one of two distinct forms: *spoken* communication or *written* discourse. Each of these is afforded different characteristics within the situations they are used [49]. First, spoken communication (sounds between individuals or groups) exists in both the time and place in which it is created (*space bound*), meaning the speaker and receiver(s) need to be within the same time-bound act in which the sound is created. By requiring both the speaker and receiver to be in the same communication act, the speaker has the ability to collect immediate feedback, allowing them to adapt the message depending on how well it is understood. Due to the speed of production, a speaker has limited ability to plan what and how something is going to be said, thus limiting the ability to produce a 'correct' sentence but still producing understandable language, resulting in increased usage of non-standard language (e.g. in the north of England, the dropping a 'h', such as 'ungry for hungry). Alternatively, written discourse is only time bound, with the meaning existing in the context in which it was written by the author. When generating written content, authors must anticipate the audience (who may read the text in a month's time or in 10 years) for whom they are writing as they will not be able to change the message when it is read. However, unlike spoken communication, written discourse allows for an extended thought process, which allows the author to develop long, semantically balanced sentences, along with using long, complex words that are never used in spoken language [49]. The difference is not just in how the communication is formed but also the situations in which it is used. Written discourse is used for recording facts and permanent records, whereas spoken language is used for interpersonal communication due to the temporary and personal aspects.

However, unlike written and spoken mediums, communication found online is a hybrid of the two forms. [49] believed that written text, even though the dominant form, such as that used online (e.g. text in Tweets and Facebook status updates) is produced and consumed as a *'mixed medium'* as its production and consumption follows the norms of both mediums. On one level, formal language can be found in

highly edited locations online, e.g. Wikipedia or news outlets, though it is OSNs that facilitates/forces written language to become like spoken language. Traditionally, in an offline setting, people had time to contemplate what and how a message was communicated; online, there is the expectation (especially on high-speed communication platforms such as IM and OSN) that a user responds quickly as a delay in response additionally purveys a message to the person waiting for the response (this pressure to respond quickly is amplified through features such as read receipts). However, unlike spoken communication, where users receive immediate visual feedback to a message, writing an Email, Tweet or IM message, results in no visual feedback for the author (even though the conversation is bound by time). Thus, when the user has written a message, they must hope there is no miscommunication. To overcome these pressures, which challenge the ability to be expressive through the modification of formal language, users have taken to using cues normally found in spoken communication. These manifest themselves as expressive lengthening through the use of multiple characters, e.g. *soooooo*, to the inclusion and development of emojiss or emoticonss to signify emotions, or coining new terms to signify new concepts, all in order to be more efficient within the constraints of Computer Mediated Communication (CMC).

The merging of the two forms can be seen in Twitter [158], where there is an increased usage of non-standard language (use of expression lengthening, abbreviations and emoticonss). Additionally, [79] showed that, within conversation chains in Twitter (users mentioning and speaking between each other), there is an increased usage of colloquial and regional dialects. However, even though the increase in non-standard language is beneficial to users (as it increases their efficiency in communication), it reduces the accuracy of NLP systems [62] as they cannot handle the increasing long tail of innovations. This means that a lot of traditional NLP applications need to be modified with new training sets to deal with OSN data.

2.3 Language and Social Structure

Language is the result and enabler of both social structures and personal interactions. Thus, when investigating language change, we must do so in the context of networks, social structures and interactions. With language change ultimately coming from individual *reproduction* and/or *misreproduction* of utterances, these (mis)reproductions are facilitated by the changing social systems/structures and differences between *personal interpretive schemes (personal grammars)*, facilitating the change and propagation of the innovations. Combining language change and the dynamics of social structures and interactions mirrors 'structuration theory' [73], in which the actions of users and the social structure (the *rules* that constrain actions) coexist with each other. As both the structure and the users coexist, the reproduction of acts by the user is in turn constrained by the social structure. Through the structuration framework, [73] stated that, when analysing social reality, one cannot separate the relationship between the

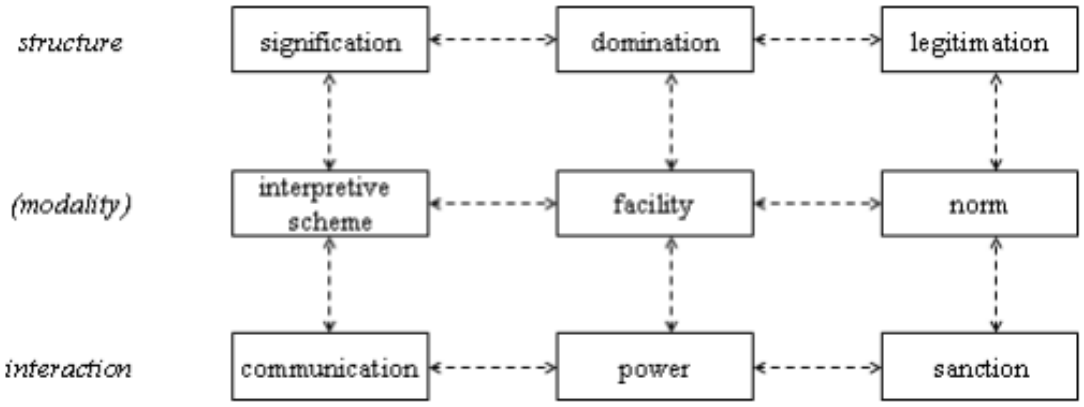


Figure 2.3.1: Giddens duality of structure, from [74]

actors and the social forces that are affecting the user.

By defining how *structure* and *agency* cannot be separated, [73] coined the term 'duality of structure', thus bringing attention to the interplay between the two. This means that the structure that a user exists within is defined by the actions a user reproduces, but also by the actions which they mis-produce. Additionally, through the repeated reproduction (by individuals and collections of users), the norms and values of users and communities emerge (or are solidified). These norms are then reinforced through social acceptance, allowing and guiding users to interact with each other. However, the social structure is not a global or physical entity, but rather defined by the user's memory:

> ...exists only as memory traces, the organic basis of human knowledge, and as instantiated in action [74]

Ultimately, the structure that individuals believe they exist within is defined by the rules and actions that the individual uses to reproduce the (their) social structure.

The dimensions of structure from which a user interprets takes three varying forms: *signification*, *domination* and *legitimation*, each representing different dimensions of language and the process of language change:

Signification (meaning) - This is the structure given by the (*personal*) interpretive scheme from which a user draws their actions; this can be the semantic scheme of words, or semantic codes of language.

Domination (power) - Production of power comes from the ability to control resources. Resources can take two forms: *tangible*, such as money or property, or *intangible*, such as positions of authority such as a manager in a workplace, or control over a faculty, such as language or intelligence.

Legitimation (norms) - The structure that provides the rules and norms in memory from which the user draws in order to authenticate the actions in the social context.

Each dimension, however, does not exist in isolation. Figure 2.3.1 represents each of their interactions (the dimension's *modality*) between *structure* and *intentions* of users. *Modality* transforms the *interactions* into *structures*, such that the structure of legitimation is achieved through the norms of a society that, when broken, means that there are sanctions.

When the framework is applied to language and language change, the following meanings can be inferred:

Signification for the continued usage of an innovation (and, ultimately, the long-term language change). The signification (meaning) comes from the variation across users' *personal interpretive schemes*. These inconsistencies between interpretive schemes allow for users intentionally and unintelligently to create and use innovations within their communications by having them grounded in a common frame of reference.

Domination within language change comes from the ability of the speaker to control the resource. Resources, however, are normally physical (taking forms such as money or property). However, language resources take intangible forms, in the form of the attention the user receives in the network.

For an innovation to be successful, it needs to spread across networks of users. However, not all users receive the same attention or have the same influence over their peers. Thus, dominant users (those who have influence and receive attention) by controlling their language (choosing whether to adopt an innovation), they then control the resource (exposure of the innovation to other users), which the innovation needs in order to replicate across the language system.

Legitimation within language change takes two forms, giving legitimacy to both the innovation itself and the user(s) of an innovation. As in domination, the legitimacy of an innovation comes from who adopts it (whether important people in the network such as celebrities or journalists) and from the number of users who ultimately use it; this could be said to be proportional to the amount of resource that is assigned to an innovation. Whereas a user's legitimacy as an important person in a community or to a language is assessed through their ability to change their language and adapt to new innovations.

An individual's actions (and agency over their actions) cannot be treated as discrete decision processes that are independent of each other, but must be thought of as a continuous flow of actions that is under constant assessment by individuals and structure alike [73]. The actions that a user chooses to perform come from needing to constantly redefine their existence within (and against) the social structure (to reassure the self-need to maintain their *ontological security*). This need to redefine comes from the *reflexive monitoring* of their actions and the structure.

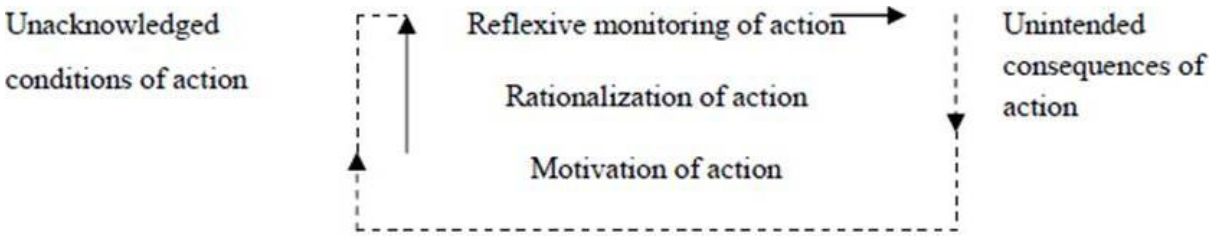


Figure 2.3.2: Stratification model of an agent, from [74]

Figure 2.3.2 represents the continuous flow of reflexive monitoring that a user performs on themselves and others. The framework of replication can be applied to the reproduction and modification of language (by a user and a community) by facilitating the separation of a user's intentions from their subsequent consequences (positive and negative) of (mis)reproduction of language. This flow can be demonstrated by an individual speaking English; their intention is to speak or communicate with fellow users, though the unintended consequence is the reproduction and thus reinforcement of the social practice of English. Thus, when a user unintentionally innovates (as stated by [45]), their unintended consequence is the mis-reproduction of the language; then, through reflexive monitoring, they can either correct their reproduction or, if there is no unintended action (they did not detect the mis-reproduction) from the receiver, then proceed with their actions. However, if the individual intends to innovate with their language, their intention is not to innovate but rather to use their resources (grasp of language) as *power/domination* to cause a change. This separation of agency and intention is intentional; the agency of a user is not the intention of the user, but rather the ability of the user to perform actions. This then relates to the ability of a user to have the *power* and *resources* to perform the action.

Ultimately, the core concept of *structuration theory* when applied to language is the power that a user possesses to transform the social world around them. Power within the social world allows members to wield control over the action of others who are internal or external to their social network [24]. For [74], however, it was the belief that power is the *transformative* capacity of people to change the world around them; power is seen to be legitimate and, at the same time, it can be repressive. When assessing language and power, two distinct sub-concepts emerge: *solidarity* and *prestige*. Solidarity contrasts power, as it is the asymmetrical relationship between two people (perceived to be) of the same power within their social circles [24]. Prestige, in contrast, builds on the concept of power; it can be seen in the social-class-based system where relative social circles have varying amounts of perceived power (power may not be measurable but rather perceived by individuals in the network). Prestige in language has been shown in a number of sociolinguistic studies within the US and UK, where more prestigious classes have the "standard" language and classes below have varying forms of the vernacular language [46]. There is evidence for the spreading of lexical innovation the other way, from the less prestige classes upwards; this is classed as negative prestige or covert prestige. Early studies into language change were in-depth

ethnographic studies of communities. However, it was highlighted that these studies never addressed the question of why certain innovations spread and others did not [201]. Solidarity can also be seen in language through users accommodating their language to that of a person of a lower power status [74].

Applying structuation theory [74] to language change and innovation allows for the structured separation of the agency over change and the structure in which change happens. It also allows us to understand that the innovation is not the intention of the user, but rather an expression of power that is used in their agency to affect change in the social structure. Subsequently, applying this framework to language change, in particular language change in an online context, allows for the structured separation between the innovation, the agent (user) and the social context.

Social Networks

The literature introduced so far has focused on the theoretical analysis of language and how social systems are interpreted and represented in the conceptualisation of language change and evolution from both the individual and system levels. However, how have social structures and networks been measured to quantify these (and other) theoretical arguments and perceptions?

Individuals are social beings, navigating the world through multiple identities and consequences of their actions [74], resulting in individuals living/existing within multiple networks of individuals, locations and communities [105]. These networks can represent an individual's social network, signifying their relationships at work or at home, or geographical networks based on the locations that they move between. Although structures found within these networks represent the predictable patterns of social life [77], there are difficulties within network science in terms of the ability to describe and express meaningful networks. [16] stated that these difficulties originate from the 'many different and rather complex networks'.

Even though network and interaction patterns are challenging to model, [181] identified that all networks can be decomposed into three core characteristics: *randomness* within building the network, *heterogeneity* in how diverse the link distribution is, and *modularity* dependent on the architecture of the graph. When applied to the human interaction network, [200] believed that the network topology can be characterised as:

Globally sparse : The number of edges in the network are far from saturated

Local clustering : The structure of the edges reveals high local clustering

Average path length : The structure of the edges means that the path between nodes is many times smaller than the number of nodes in the network

These three characteristics are good at describing the structure of the network, but not necessarily the reasons behind the structure. [180] later drew on social theory in the definition of social networks:

Reciprocation : Relationships between users are rarely in one direction, but rather both people depend on each other [150], thus relationships are reciprocal.

Homophily : As proposed by [104], people are more likely to connect within people that are similar to themselves, adhering to the analogy 'birds of a feather flock together' [66], [142]. This effect has been seen to exist both online and offline [4], [56]. This ultimately means that users with similar social variables will have a high probability of having connections between each other, or being members of the same group.

Transitivity : Similar to homophily, this looks at how far friendships in a network may extend, working to the analogy 'friends of my friends are my friends' [164]. This effect in a network can be seen by the higher probability of edges being formed in a network to nodes that are one hop away, thus closing the triangle.

Degree differentials : As seen in scale-free and power law networks, the degree of distribution of users is not the same; some users have a high degree whereas the majority have a low degree. The inequality in degrees results in social networks experiencing the Mathew effect, where the 'rich get richer' [132]; this results in preferential attachment in the network when new nodes join.

Hierarchies within networks express the power or standing of a node within the structure. This is expressed in directed networks that have low transitivity. In a directed network, this is identified though the relatively low number of three cycles [55]. Through analysis of the directed graph of multiple social networks, [91] indicated that, when the network is small, there is little hierarchy in the network due to people knowing more of the network. However, as the network grows, hierarchical structures emerge as people become leaders and gain power within the network.

As stated by [200], networks are characterised by sparse connections, local clustering and short average path lengths. The combination of these three characteristics results in networks feeling 'smaller' to the individuals in the network, as one is able to access a large proportion of the network in a small number of steps, creating a 'small world' effect. The small world effect has been shown in varying types of network; the seminal study of the phenomena was performed by sociologist [146], who mailed random letters to people in the Sates of Omaha, Nebraska and Kansas with instructions to pass on the letters to a person they thought would know a certain doctor in Boston, MA. Of the letters that were sent, only 64 made it to the doctor, with an average path of 5.5 people.

To understand the process of forming small world networks, [200] compared a number of real-world networks to two toy (simulated) networks. They found that the real-world networks had similar clustering coefficients as lattice networks [1]. However, they also had the short average path lengths that are found

[1] A graph that, when laid out in grid form, expresses a tiling effect

in random networks.[2] [200] proposed that, by increasing the randomness of a lattice network, they could maintain the clustering coefficient, but decrease the average shortest path. This transaction resulted in the toy network's properties following the observation of real small world networks. [200] then proposed that real-world networks are classified as small-world networks when $L \geqslant L_{random}$ and $C \geqslant C_{random}$, where L is the average shortest path, and C is the clustering coefficient.

> [a] common property of many large networks is that the vertex connectivities follow a scale-free power law distribution [16]

An additional core characteristic of real-world networks is a degree distribution that follows a power law. The existence of a power law distribution influences the preferential attachment of nodes in the network (Mathews effect), where new nodes in the network want to attach to nodes with high degrees as this will give them greater access to the network. Thus, networks are said to be 'scale-free' networks if they follow preferential attachment when new nodes join the network; as the network grows, nodes with a greater number of edges grow at a faster rate than those with less [16].

The internet is a prime example of preferential attachment, where the expansion of new connections is placed within well-connected nodes and routers, with the aim of increasing the connection speed and reducing the number of hops that the traffic needs to take [215]. However, these models of the networks remove the idea that a node can have a maximum number of connections. In friendship networks (interpersonal networks), this has been stated to be limited to a number of maximum connections that a person can maintain at once. Additionally, preferential attachment results in scale-free graphs being characterised by their topology having a 'highly' connected hub structure; thus, some members of the network would have a larger degree of centrality. However, 'hubs' would indicate that they are seen as core to the structure of the network, and maintain a viable graph [17]. These structures can be seen in OSNs like Twitter and Facebook, where the distribution of degrees follows a power law [177] where the majority of users have repetitively few connections, though a minority are highly central, connecting communities and users together.

Individuals organising their networks around a scale-free and small-world structure means that this is the network that language innovations have to diffuse across and over. Thus, if drawing on the average of users only being separated by 5.5 connections, this means that all languages in the world are separated only by 5.5 users [177]. However, as users additionally cluster together into community structures, each with different degrees of distribution, each user can influence the diffusion of content to a varying extent. Therefore, for an innovation to diffuse, it is not only about the innovation itself but also where it is used and who uses it.

[2] Networks whose edges are randomly assigned

Innovation Diffusion

The networks mentioned earlier aid with the diffusion of innovations, content and ideas as they link individuals and/or organisations together. The example of HIV at first only affecting certain communities within each city shows that segmentation within social networks can both enable and contain diffusion [21]. This segmentation could come from a number of different causes, such as:

- Geographical locations of different networks limiting the communication between different members; this was shown to be the case in Japan when predicating buying habits for cars [212].
- Industrial trade routes not only allow for trade between cities and countries [23] but also the cross-influence of cultural forces that come with the trade.
- User membership of the social network, which can be seen in [13], who showed that strong community membership hindered innovations diffusing outside of the community.
- The cultural importance of the innovation to a network; [68] identified that, if the innovation is deemed a moral hazard, then there is a greater need for stronger ties to help the information propagate through the network.

This segmentation may limit the rate of diffusion within the whole network, containing the innovation in one place or allowing it to spread to further parts of the network depending on the nature of the network segmentation, with examples such as trade routes affecting diffusion over a large scale (both time and location) and community membership being small scale (affecting a location's population immediately).

Network structures and user interaction facilitate the diffusion of information and content between users, at both local (between users) and global (between communities) levels. However, the relative position of users and, in turn, communities in the local and global structure affect their opportunities to access information from different parts of the network. Individuals at either extreme can be strongly embedded into a community, only communicating with users internal to the community, or sitting on the periphery with the ability to make connections in other communities. The position of a user affords them varying costs and benefits in information and innovation diffusion across the network. [87] believed the connection that a user maintains could be classified as having either 'weak' or 'strong' ties based on the level of communication across the edges, believing that the weak ties of a user only carried a small proportion of the information for a user, though this information is unique as it will have originated from other communities in the network. [87] believed that, across the edges where the nodes had less crossover between their users, there would be less communication. However, when there was communication, it would be important as the message would be diverse and interesting. This hypothesis of weak ties has been tested on multiple social networks, including Twitter [85], [102], [204], the Enron email corpus [204] and phone networks [204].

[13] showed that, even though communication internal to a community is reliant on the strong internal ties, the diffusion of new content is dependent on the existence of many weak ties between communities allowing information to flow across the network. The dependency of information diffusion on weak ties is identified by [87] and [28], whereby users who have many weak ties exist in a 'brokerage' position between communities, as they sit in the 'structure holes' of the network. This hypothesis has been tested on a number of social networks, with the results confirming the importance of users in a brokerage position in the network; this has been shown in both Twitter communities [204] and Facebook [57]. Alternatively, users can be found in a bonding position. This happens across strong ties in homogeneous groups such as family networks, helping with the building of trust and norms. However, these strong internal bonds hinder the individual's access to external information sources, allowing internal norms to be reinforced, in turn allowing them to develop in different directions and at different rates to those external to their community.

Language is dependent on communities, and diffuses between users and communities when innovations are replications. Labov [129] proposed that the diffusion of language happens internal to strong connections that people have within a network, stating that diffusion accrues due to the frequency of contact between users, which is encouraged within groups due to the open channels of communication. However, external to the community, these channels close and are thus discouraged. However, [148] identified that language diffuses across weak ties, as individuals control contact between groups of users. By assessing the diffusion of language between two isolated communities in Northern Ireland, [148] showed that the only way that language was diffusing between the two communities was through the limited contact that the communities had in a shopping centre; thus, the diffusion was happening across the weak ties between the two groups. The effect of weak ties was also seen by comparing the historic development of Icelandic and English. [149] proposed that English changed through the British population's numerous weak ties developed throughout the globe by a mobile population. Thus, English was influenced by other languages. However, for Iceland, whose population was less mobile, innovation diffused locally over strong ties. However, this reinforced the language, with little change happening over time.

> ... linguistic change is slow to the extent that the relevant populations are well established and bound by strong ties, whereas it is rapid to the extent that weak ties exist in the population [149]

However, even though a language may have access to structural holes to spread, [148] believed that the size of the community in which the innovation originated affected the final diffusion of the innovation external to a community. A larger community would have a greater proportion of weak ties that connected to more communities external to the community [194]. Additionally, it would mean that the community had a greater ability to resist change, thus allowing the language to be maintained for a longer period of time.

Thus, the diffusion of language and the speed at which changes happen can be attributed to two conflicting properties of network structures: strong internal bonds and weak ties. Strong internal bonds allow for the reinforcing of language and norms, whereas the external weak ties allow for the innovations to diffuse and for a community's language to be influenced by external sources.

Diffusion of Innovations

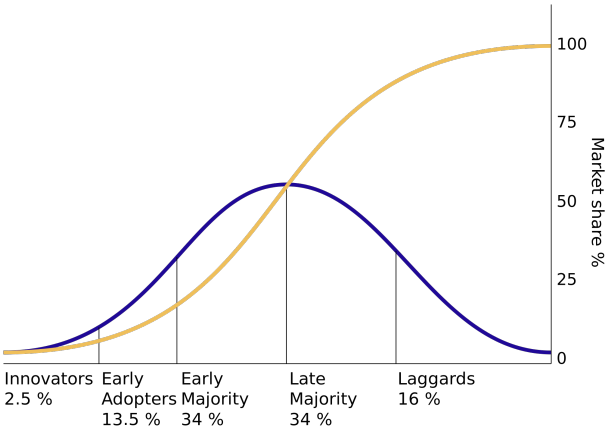


Figure 2.3.3: Rodgers diffusion of innovation - adopter categories

Diffusion of innovations is a process that happens over time, with each individual in the diffusion making the choice of whether to adopt an innovation or not. [196] defined the process as:

> [t]he process in which a few members of a social system initially adopt an innovation, then over time more individuals adopt until all (or most) members adopt the new innovation [196]

Rodgers further defined innovation itself as:

> . . . an idea, practice, or object that is perceived as new by an individual or other unit of adoption [167].

Examples of innovation that have diffused across a network and have been studied include fax machines and pharmaceuticals in the US, across networks of doctors [167]. The position of a user in a diffusion, such as at the beginning (being one of the first to adopt an innovation) or end of a diffusion process can suggest characteristics that express how open a user is to accepting innovations. [167] showed that a user's *innovativeness* (preference to adopt a new innovation) could be segmented into five distinct categories, defined by a user's Time to Addoption (TTA). Each category is defined through the standard deviation from the mean time of adoption, with TTA defined as the difference in the time of a user adopting an innovation (τ_u) and the time at which the innovation was created τ_i.

$$tta_u = \tau_u - \tau_i \tag{2.3.1}$$

The mean time-to-adoption is represented as $\bar{tta}$, with the standard deviation of the set of values represented as σ.

Innovator - Those users willing to experience new ideas. They are willing to bear the cost of potential failure of the innovation, to a greater extent than other people. They act as gatekeepers of the innovation, bringing innovation into a community. This is similar to Granovetter's theory of weak ties, where users that bridge structural holes in the network bring innovation into a community. Their time-to-adoption is relatively small in comparison to the rest of the community, formally being defined as $ino = \{u \in U | \tau_u < \bar{\tau} - 2\sigma\}$

Early Adopters - Following the innovators, Rodgers showed that it was then the leaders of a community that adopted next, classifying them as the early adopters. Rodgers believed that the leaders of a community were highly connected within the local network. This follows the work of Granovetter, who believed that *bonding relationships* within community structures indicated the opinion leaders of a local community. Rodgers identified that, as early adopters were leaders within the network, their opinions were highly regarded by the community and were significant in the further diffusion of an innovation due to the user acting as a role model to other users. The set of early adopters (ea) can be expressed as: $ea = \{u \in U | \bar{\tau} - 2\sigma \ < \tau_u < \bar{\tau} - \sigma\}$

Early majority - Unlike early adopters, these are users who are not opinion leaders within their communities, though they are active members of the community with good intentions, and the choice of adoption is deliberate in response to their interaction. They adopt the innovation before the other half of their peer group. The set of early majority (em) users is defined as: $em = \{u \in U | \bar{\tau} - \sigma \ < \tau_u < \bar{\tau}\}$

Late majority - Compared to early majority users, late majority users have a tta greater than the mean. This is due to users holding out until the majority of their social connections (or peers) have adopted the innovation. This is because they believe that the cost associated with the innovation's adoption may be high. However, they succumb to pressure from their peers. The set of users making up the late majority is defined as: $ino = \{u \in U | \bar{\tau} < \tau_u < \bar{\tau} + \sigma\}$. This class of users incorporates $\frac{1}{3}$ of users. Rogers stated that, 'the late majority feel that it is safe to adopt'.

Laggards - Rodgers believed that these users are more sceptical and risk averse, and to adopt an innovation they need to have seen a user adopt it along with them then maintain the innovation. It is also the user's position in the network that limits their ability to gain the information and knowledge about the innovation as they cluster with fellow sceptics; therefore, the scepticism is reinforcing. Laggers are thus defined as: $lag = \{u \in U | \tau_u > \bar{\tau} + \sigma\}$.

The benefits of categorising users *innovativeness* based on TTA allows for audience segmentation,

thus allowing for the comparison of early and late adopters' features such as personal and network characteristics. However, assessing only the TTA does not provide insight into the individual dissension process of the user, but rather how long it takes them to make the diffusion. However, TTA categories are assessed on the user's time to action in relation to the whole social system; though, as stated by [196], a user's actions are in relation to their immediate network and less influenced by the whole social system.

As highlighted earlier, individuals (Section 2.3) have agency over their own independent actions, such as the decision to adopt a new technology, belief or action. However, in a networked world, actions are influenced by those around us, whether from observing users mentioning new technology or adopting new beliefs. However, the amount of exposure/influence at which a user adopts an innovation varies through the individual user's cost-benefit analysis.

> Threshold models of collective behaviour postulate that an individual engages in behaviour based on the proportion of people in the social system already engaged in the behaviour [86]

[86] believed that each user has an arbitrary threshold ϕ_u at which, when breached, the user would perform an action or adopt an innovation. To model this, a set of users in a toy population were each assigned a threshold (ϕ_u), which was drawn from a probability distribution $f(\phi)$, where ϕ aimed to subsume all the variables and influences that make up a user's decision process. The threshold of each user (ϕ_u) represents the proportion of users in the whole system that need to be activated before u becomes active. $f(\phi)$ also represents the average value threshold across the whole network, along with the heterogeneity in the populations; thus, increasing or lowering this value would indicate the network having a high or low threshold.

Diffusions are a time-dependent process; thus, [86] proposed that, at the initial time step (t_0), there would be a set of users (a_0) that started the innovation for no apparent reason (a random jolt to the system). As the diffusion process progresses through time ($t+1$), the number of active users increases and the number of inactive users decreases. The fraction of the inactive population that becomes active at time t is the set for which $a_t \geq \phi_t$. As the model proposes that the threshold is distributed normally, the rate of users joining follows an S curve, initially speeding up as it reaches the median threshold, then slowing down.

The current maximum threshold of the community is defined through the use of a Cumulative distribution function (CDF):

$$F(a_t - 1) = \int_{\phi=0}^{a_t-1} f(\phi)d\phi \tag{2.3.2}$$

Thus, a threshold indicates that the decision to adopt is unique to the user, and a user with a lower threshold would engage earlier, whereas a higher threshold would mean that they needed to have a greater exposure to the process. A classic example given is the formation of riots: there will be a number

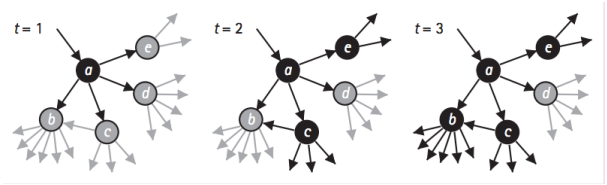


Figure 2.3.4: Content diffusion at subsequent time period only to nodes that immediately consent to the infected node, from [86]

of people always rioting with a very low threshold as they do not conceive the cost; as more people join, the cost (risk) of being arrested reduces, thus attracting more people over time. However, this example would indicate that the influence on the threshold comes from across the whole network. [86] stated that it was rather based on the personal connections that a user has, giving examples of professional networks and the adoption of birth control.

However, the assumption that a user has access to all the information in the network, and thus makes decisions based on global activity, does not hold up when applied to the decision to buy a new technology or adopt an innovation. This is due to innovation adoptions not necessarily being observable to the network or the only benefit of adopting an innovation coming from knowing people who have also adopted an innovation such as a fax machine. Thus, during the decision process, a user attempts to quantify the risk and uncertainty in adopting a new item relative to their local network, not the global network, attempting to learn the experiences of their friends. [196] believed that the threshold should be measured in relation to the neighbours in the social network and not the whole network, as proposed by [86].

[196] modelled a user's decision process as only being dependent on number of activated users within their ego network, not the whole system. By first generating a toy network that models who is influenced by whom, with the degree of nodes in the network following a normal distribution, this means the network follows a small-world pattern. Similar to [86], [196] randomly assigned an individual user threshold (ϕ_u) drawn from a distribution $f(\phi)$, with the value indicating the proportion of neighbours (k) that have to be activated before u activates. Thus, [196]'s diffusion model is defined by three parameters: ϕ, N and k. To simulate a diffusion at the initial step t_0, a number of nodes *randomly* activate, which would be thought to be the innovators within the network. A user becomes active only if the proportion of their neighbours $\frac{1}{k_a}$ are active, not the number of users in the whole network, with the current user threshold increasing by $\frac{1}{k}$ for each additional active neighbour (see Figure 2.3.4). Thus, it is not only the distribution of the threshold that affects the diffusion of content but also the structure of the social network and where in the network the diffusion starts.

Thus, instead of comparing based on a global measure, [196] used the individual adoption threshold (ϕ_u) as an indicator of *local innovativeness*. Using the adoption threshold instead of TTA allows assessments to be made on a local level; thus, users with lower thresholds engage in the collective behaviour

(adopt innovations) of their local network quicker than users within a high threshold. However, the power comes from the contrasts that can be made between the two models; if a user has a high threshold (ϕ_u) but a low TTA, then they are innovative in relation to the social system, but not their local system. Whereas, if they have a low threshold (ϕ_u) and a high TTA, then they are innovators in their local network, but not at a global level, which could indicate that they are not part of the dominant community and thus lack access to information.

In relation to the diffusion of language and the ability of users to adopt language innovations, the understanding of the diffusion process between both the user and communities is an important one. As highlighted in the previous section, at a high level, it is the language community's access to weak ties that influences the diffusion. However, on an individual level, each actor (user) has a decision process that is not only influenced by the people around them but also fellow members of their network. Thus, for language to diffuse, the users must be willing to have low threshold to adopt and experiment with the innovations, whereas a high threshold would mean that there would be little diffusion within their language and thus little change.

2.4 Summary

Part one of the literature review has focused on grounding the thesis in empirical work ranging from sociology and linguistics. Initially, we highlighted what language is and how language change and language innovation accrue. Drawing on [191], we highlighted that language is ultimately defined by the community who speaks the language, and that the innovation can be both intentional and accidental. However, the innovations themselves (and by extension, the component language, e.g. utterance) can be thought of in much the same way as genes, with survival be seen as a probabilistic process.

Language, as seen through the structural lens of [74] as the social structure in which users find themselves, is a duality, with the user causing change through their reproduction of the structure. This can then be applied to language, where the social structure affects how users talk; but through reproduction of the language, users can cause the language to change [46]. Concepts such as power and domination can be seen in users adopting the language of more powerful users within the network. However [148], during the process of language change, it is in fact a community's weak ties that influence the speaking of a language, with more mobile communities being influenced by the many other communities with which they come into contact. However, [129] highlighted the importance of social variables and the frequency of contact to the process of language formation internal to a community. Allowing communities to develop specialised communication patterns that differentiate themselves from external community members.

Connections can be drawn between the process of language change and innovation diffusion over a

population [167], as [86] proposed that individuals have an innate threshold at which they adopt an innovation. However, [196] highlighted that it is in fact users within an individual's immediate network who influence an innovation's adoption. However, for language to diffuse between communities, it is the weak ties and structural holes that mediate and manage the diffusion of innovation, whereas the strong internal bonds reinforce language and language norms.

Overall, language change is a process that happens to both the individual and the community, working in the dynamics of the duality of structure, through users' (mis)reproduction and (mis)interpretation of language and contexts. It is through these dynamics that we will model language innovation diffusions and the process of language changes across OSNs. Each of the three core research questions focuses on an individual aspect of the process, from individual innovations (question 1.1.1) to the interplay of users and language within the social structures (question 1.1.3). By grounding this thesis in the existing literature, we aim to answer the questions with rigour, which means that we can discover whether the theories are applicable in modern online contexts.

Chapter 3

Social Computing

The field of 'social computing' focuses on the interaction of human social behaviour and computational systems [198]. For this thesis, the following definition is adopted: 'computational facilitation of social studies and human social dynamics as well as the design and use of ICT technologies that consider social context' [198]. The following literature review will focus on the user dynamics within OSNs, and not the design of social systems themselves.

This literature review will summarise and critique recent and relevant work in the context of language change and social dynamics as seen in the study of OSNs. As in part one of the literature review (Section 2), the following review will be broken down into three distinct sections:

- **Detection of language innovation and change** (Section 3.1) - How have researchers quantified language change in the past? How has language change been detected? For what have the changes in language been used?
- **Diffusion of information and content** (Section 3.3) - How does information and language diffuse around OSN, and how have these dynamics been used in predicting the extent to which the information/language will spread or become viral?
- **Power and dynamics** (Section 3.2) - Users influence the actions of each other. Thus, how have these integrations been mined in quantifying the user's power within the network? Additionally, how are these power dynamics used to model and predict the actions of users?

3.1 Innovation Detection

As evidenced in Section 2, language is in constant flux, whether from the introduction of new words to regional or global changes in meanings. However, how are these changes initially detected, which data sources are used, what metrics are used, and what do the results show about language and the process

Table 3.1: Example innovations and relative frequency, as from [6]

Formation Type	Relative Frequency	Example
Composite	64	Think Tank
Shifts	14	Hardhat, meaning construction worker
Shortening	10	Jag for Jaguar
Loanwords / Borrowing	7	Macho from Spanish
Blends	5	Chunnel for channel tunnel
Unknown	1	Cowabunga

in which innovations happen and are accepted by a population? The following section highlights the state-of-the-art work in innovation detection, classification and modelling.

Linguists and lexicographers involve themselves in the identification and analysis of language change and variation over time and across cultures and populations. Traditional methods applied in identifying changes in language, however, are manual, time-consuming and expensive, involving manual annotation, synthesis of results and collection of data sources. However, the resulting dictionaries are treated as a snapshot of language at a moment in time. To aid lexicographers in the process of determining whether a new word should be included in a dictionary, a number of heuristics have been proposed; two of those widely cited are Barnhart's VFRGT [18] and Metcalf's FUDGE [144]. Both are used to assess and identify whether new innovations introduced into a community will be maintained or lost, and thus should be recorded in a dictionary. However, both were developed to be used by lexicographers with a scoring method; thus, the value assigned is at the discretion of a scorer, based on their assessment across a number of categories, such as how unobtrusive a word is and the endurance of the concept associated with the word. [34] attempted to automate these two heuristics by applying them to a 10-year historic corpus of Chinese newspapers, with the aim of tracing and predicting the inclusion of new terms in a dictionary. This was achieved by first manually applying Metcalf's scoring framework to a candidate set of innovations, and then using an SVM model to predict whether the word would appear in a dictionary. By using a predictive model, they identified that categories such as unobtrusiveness do not play a significant role in word adoption within the Chinese language. However, the work had limitations, as language in newspapers is inherently more formal and thus is not a representative sample of language. In addition, manual annotation was used to identify feature weights for varying word forms, thus limiting the scale at which this work can be performed, as well as the reproducibility of the work.

As highlighted by both Barnhart [18] and Metcalf [144], language innovations appear in varying forms. Therefore, in an attempt to classify the different types of language change, [6] assessed 1,000 words found in the 1973 editions of Barnhart's book of new words [211]. Through manual annotation, [6] identified six distinct classifications (Table 3.1). Automation of the classification of innovation and terms has been attempted, though the forms of innovation that appear in online discourse are greater than those outlined in [6], which excluded both expressive lengthening and representing images in text such as emojis and emotions. By using an ensemble of morphological features, [138] classified OOV words (with

regard to English) found on Twitter into six distinct classes: *emoticons*, *expressive word lengthening*, *expressions*, *abbreviations and shortenings*, *proper nouns*, and *word merging*. The features used were broken down into four broad categories: *lexical*, *content*, and *context*, with a combination of these used in an SVM model that achieved 80 % accuracy. However, POS tags for the innovation was one feature in the classifier used, though the POS tagger used [76] contained specific categorises for emojis and abbreviations and other innovations. Therefore, this leads to the question of whether the classifier only learnt the categories of the POS tagger.

Users' use of OOV allows them to be more expressive in their communication. However, it poses challenges to traditional NLP systems that have been developed and/or trained on standardised corpora. This challenge is due to the large number of OOV that is contained within the online discourse, and is not limited to non-standard punctuation, capitalisation, spelling, vocabulary and syntax. These challenges have been seen in the reduction in the accuracy of POS taggers and tokenisers from 97 % in the *Wall Street Journal* corpus to 85 % in a Twitter dataset, with similar results seen for NER systems in which accuracy reduces from 86 % to 44 % [14]. [3], [93], [94], [96] proposed a number of normalisation strategies to correct 'ill-formed' words back to their 'correct' form. [96] used morpho-phonemic similarity within word context to determine the 'correct' word, whereas [136] showed that, by using ML techniques, we can identify common word transformation patterns such as 'birthday' to 'bday', dropping 'irth'. [157] clustered words together to mitigate variation in spellings, and [75] applied a normalisation strategy resulting in the increased accuracy of a heterogeneous tagger.[1] However, the language that people classify as 'bad' is in fact highly expressive of the social variables of the users who create the text, such as age, gender and geographical location [14], [62]. Thus, 'correcting' language by removing OOV reduces the value of the data, meaning that the research gains less insight into the users and contexts in which language is used.

Language change is not only the emergence of new words but also the change in usage over time. Previous work only classifies the morphology or high-level semantic change, missing variations through amelioration and projection. [43] showed that changes in the semantic orientation of words can be identified by computing the Point-wise mutual information (PMI) of a word against co-occurrence with lists of known positive and negative words. However, language is highly dependent on the local community that uses the work, such that words such as 'sick' could have varying projection depending on the community, e.g. meaning something is good or bad.

To utilise the 'bad' language found online, but still utilise traditional NLP tool chains, a number of specialist tools have been developed to use the stylistic variations found in online text. [76], [157] developed a custom POS-tag set for Twitter that included tags for emoticons, URL and abbreviations. Along with using the position of tags in a tweet as a training feature, they achieved a 25 % error

[1] http://gate.ac.uk/wiki/twitter-postagger.html

reduction rate compared to the Stanford POS tagger.[2] However, the application of this tagger has not been demonstrated on other social media data.

However, we are not only concerned with identifying OOVs online but also understanding their meaning and origins of the meaning. [44] focused on identifying the source words of new lexical blends; through the use of a Twitter dataset, they used regular expressions to identify users defining new terms, e.g. 'slang (expression—phrase) for' and blend heuristics to detect the potential source words that form the new term. The results identified tweets containing both the source words and the word blend, with the source words being selected through a number of simple morphological heuristics. Expanding on this work, [42], [44] extended the feature set for determining source words of word blends by including the likelihoods of words appearing in the same context, and the probability of the blend accruing based on the phonological patterns of the source words. By the inclusion of the additional features and the use of a perception-based classifier, they demonstrated improvement in source word detection; though, again, this was limited by the quality of the underlying data, with the system still detecting a high number of false positives.

It is not just the introduction of new words but also shifts in meaning over time. These shifts are demonstrated in the historic variation of 'gay', from meaning 'happy' during the 18^{th} century, to the current meaning in relation to the LGBT movement. [123] showed that the changes in meaning associated with words can be assessed over extensive time periods using the Google N-Gram dataset [3]. They proposed three measures to assess changes in language, first, by modelling the probability of a word in each time period, and second, by quantifying the changes in the syntactic usage of a word by POS tagging each word then measuring the distribution of tags in each successive time period as the Jensen–Shannon divergence (JSD) between tag distributions. The final method measured semantic shifts in words by comparing the embedding of words in each time period. This was achieved by first learning the embedding per time period, then, for each word in each model, computing the distance between the embedding spaces of time period t and time period 0, thus measuring the change from the beginning of the corpus. The significance of the change at each point (t) in the time series was quantified through the use of a *mean shift model*, with the value being classified as significant when comparing the *mean shift* to a bootstrap sampled random baseline model. The combination of the three measures was able to identify significant changes in words such as 'gay'. However, they treated the language in each dataset as one entity, ignoring the community structure that influences the formation and evolution of language. Thus, the results indicate language change at a high level, offering no explanation or understanding of the community dynamics that influence the changes.

When talking about language change and evolution, we cannot ignore the growth in the use of emojis and emoticons in online communication. Emoticons are textual representations of pictures, predomi-

[2] http://nlp.stanford.edu/software/tagger.shtml
[3] https://books.google.com/ngrams

nantly faces, that help add an expressive context to a message (e.g. :-) and ;-/). These were followed by emojis, which are pictorial representations that have been incorporated into ASCII code. However, assessing the context in which emojis are used has shown that, as they are implemented differently on various devices, the meaning users wish to portray varies and can be lost in translation [159]. Whereas [159] showed that, as the prevalence of emojis increased, the popularity of emoticons decreased. This was assessed through the use of a causal inference framework that assessed the difference between users in a control group (users who had not used emojis) and those in a treatment group (users who, in March 2014, did not use any emojis, but had used at least five in March 2015). Users in each group were then matched on their usage of emoticons as a proportion of their whole corpus $\frac{\#ofemoticons}{\#numberoftokens}$ prior to the treatment. The treatment (introduction of emojis) was thus quantified as the effect on the emoticon usage rate after the treatment. They found that, indeed, with the introduction of emojis, the rate of emoticon usage reduced to 0.14 % compared to 0.30 %. Even though they controlled for the introduction of emojis, a number of other treatments were over looked, such as the device from which a tweet originated, as emojis on computers are more difficult to produce compared to mobiles, and some smartphones do not support them.

As stated earlier, the variation across user language is not independent of the user's social factors (variables), such as age, race and gender, all of which have been shown to have a strong influence on the language of a user. This was initially identified by [129], who assessed the language of different social classes in New York, though similar effects in user language can be seen across OSNs. [163] showed that the gender of a user can be predicted from features such as emoticon/emoji usage and topics being discussed. However, the ground truth of gender was inferred by which accounts the users followed, e.g. following a fraternity would class the user as male. However, the user of an inferred ground truth could lead to misclassification of the user, along with the features used to classify the user being used to predict their class, e.g. a user following men's health may talk about men's health. Alternatively, [5] showed that we could learn the latent attributes for the user from their immediate network, as users cluster together due to homophily and shared interests. However, again, the ground truth was inferred: gender was inferred from the name of the user, whereas a user's age was inferred from fellow users wishing them happy birthday, thus limiting the true accuracy of the system. For less granular latent variables such as political interest (which, in the US, is a binary label), the classification accuracy increased if they included features of the users within their ego network. This again aligns with homophily and social reinforcement.

Variation in language does not only exist on an individual scale but in a global context too; this can be seen in the fact that 2,000+ languages are spoken around the word, and also in the variations in English across different regions and countries. On a local scale, in the UK, this can be seen in the variations in the meaning of 'Tea' and 'Dinner', or the regional variations dropping of 'h' in certain

terms. To identify location-specific words, [61] compared the language of a region to the language of the whole country. This was initially achieved by using logistic transformation to smooth regional word counts with region-specific words identified when ranked by the ratio between regional counts and global counts. However, the words that were highlighted might not be considered as language change or innovation as, even though they were region-specific words, they were predominantly dominated by sports team names and regional food. Additionally, the results did not highlight the dynamics around how language may move between locations. Instead of attempting to identify language regions through the use of generative models, [100] looked at geographical variations in Twitter language. Through the use of kernel smoothing and PCA-identified language variance across the US, there was a large variance in the language used, though much of this, again, was placed down to specific geographical terms and pronominals such as sports teams and food preferences.

However, as stated by [129], the demographics of each region are highly variable. Thus, [155] used geo-tagged tweets and associated them with the demographic features attached to the given location, such as the percentage of African-Americans and the percentage of family households. By then developing a latent variable model, which combined the demographic data along with the text within the tweets, 9 clusters of language dependent on demographic variables were identified. One cluster identified the prominence of Spanish words with the prominence of a Hispanic demographic, along with the phonetic shortening of words being associated with higher than average income and more children. However, as stated, the demographics of users of Twitter are skewed from the demographics of the US. Even with the attempt to look at language and identity, the results are limited due to the discrepancy between the demographics of the users and the communities. [63] applied a similar method with the development of a multi-level generative topic mode that took into account the GPS location of the user. Again, this combination of language and location only highlighted variations in sports and food terminology. However, the model was able to identify the words within the same group, so one could infer the set of sports teams. [15] stated that the meaning of a word was dependent on the geographical context in which it was used, but stated that recent advances in NLP systems, such as `word2vec` or other systems, focused on using language across the whole population. By extending the `word2vec` model through the inclusion of *situated* information such as the state in which a tweet originated, [106] was able to develop a model that, when given a query word, was able to highlight the contextually similar words for each region, e.g. 'wicked' meaning 'cool' in Massachusetts and 'evil' in Kansas. However, the model was evaluated quantitatively by comparing the model to alternates, one of which was trained to represent the global language, and the second being a model for each location, trained only on the local data. Additionally, unlike [63], we cannot identify clusters of words that are substitutions for each other; therefore, to identify the 'similar' words, another level of analysis has to be applied.

Variation in a language is not limited to one country, which can be seen through the global number

of languages that share aspects of English. Again, through the application of word embedding, [122] was able to identify variation in meanings across cultural boundaries (different countries that share the same base language, such as England, the US and Australia) through the detection of statistically significant word embedding. This was achieved by learning the embeddings of words in regions in comparison to the global embeddings, with the results identifying words such as *freshmen* and *touchdown*, which have known contextual cultural variations.

In this section, we have focused on the innovation itself, how researchers have quantified its acceptance into a language (through inclusion in a dictionary) and how attempts have been made to automate meaning identification of these new innovations. In quantifying the acceptance of innovations, traditional grounded heuristics are still to be relied upon, though these have been applied in computational contexts but still rely heavily on human raters. However, across the two main heuristics, commonalities appear in that acceptance of an innovation is not based only on the quantities in which the innovation is used but also in maintaining constant meaning independent of the context in which it is used. The meaning can change over time, but it must still have a collective understanding independent of the context in which it is used.

In identifying the meaning of innovations, similar measures (as seen in the acceptance of innovations in language) can be seen across the identified research. Initially, this can be seen in the volume (frequency) in which the innovation is used, but, additionally, in the importance of context in which it is used. The importance of context can be seen in the geographical assessment of meanings of language where regional dialects and phrases are detected. This is not to say that these are not innovation, but it highlights the importance of context, location and community in assessing language change and innovation in OSN.

3.2 Innovation Adoption

The language that individuals use is a result of the language of the people with whom they come into contact (Section 2.3). However, some individuals have a greater effect on a user's language than others. The following sections look at how influence in language has been studied and how this can be applied to language change and evolution across a whole network.

As within any social systems, users in OSN are not all equal, as the influence and power that users have over each other varies. However, due to the concept of power and influence being implicit in nature, there are challenges in quantifying it. Researchers have quantified the power and influence of users and communities in a number of ways, ranging from structural measures of a user's position (through the application of measures such as page rank to degree of centrality) to modelling the power over users' actions as a function of cascades (chains) of actions. [1] looked at modelling power and influence as a function of language, using the Enron corpus to mine language in conversational chains.

Alternatively, [83], [84] proposed that cascades of information reveal influence and [202] used network structure to identify topic influences.

The language used in conversations is highly revealing of power dynamics between users, whether from the patterns of communication to users accommodating their language. In OSNs, users can be thought of as existing in a number of different roles: explicitly defined, such as a moderator on Reddit, or implicitly, as a user that answers questions for other users, each filling different positions of power and influence with the network. By assessing these roles within the editor chat logs of Wikipedia, [51] showed that, through the user conversation chains and usage of domain-independent language features (e.g. personal pronouns), we can identify different seniorities of editors. The results further indicated that low-ranked editors adapted (accommodated) their language to that of the senior editors, whereas administrators (senior editors) rarely altered their language. The datasets also allowed them to model changes in user status over time, which showed that a user's likelihood of accommodating their language to that of lower-ranked users decreased. However, as this work tracks domain-independent language features, and only uses standardised language, it would not detect the power dynamics within conversations that use a large proportion of innovations. However, even though these roles (positions of power and influence) evolve over time, they can transcend OSN. [26] identified that these implicit user roles transfer with the user as they move between communities. Assessing the language profiles of users and their interaction patterns showed that users identities/roles transcended communications, such that a user fulfilling the 'answering' role in one subreddit would be highly likely to answer questions within other subreddits. However, as these users are highly active within their communities, they have a strong internal bonding that limits their access to language innovations across structural holes. This would suggest that their language is more 'normalised' to that of the communities of which they are a part, and they do not 'play' with language from other communities.

However, accommodation of language does not only happen between individual users but also due to the collective pressure of a community upon an individual, especial when a new user enters the network. [53] showed that, as users enter OSN, their language profile moves towards that of the community, whereas, before leaving (churning), their language diverges away. Evidence was given from two online beer communities, where users talked about beer and brewing techniques, with users' language conforming to that of the community in the early stages. However, these beer communities use highly specialised language that is not seen across other networks when talking about beer and techniques; thus, a convergence in language profile may not be due to power dynamics but rather to them using more specialised terms to describe flavour profiles.

Another explication of language change and variation observed within OSN is *audience design*. This is when the user chooses their style of language depending on the audience the message is intended for, with the aim of making the message more 'genuine' to the receiver of the message. [188] assessed

audience design by modelling communication that was internal and external to users' dominant community membership on Twitter. The results indicated that users modulate features in their language depending on who they are communicating with, e.g. when communicating with a new community, they alter their language to increase the impact of their message. This effect of audience design was also seen by [158], who looked at the intended audience of tweets and the form of language, innovations and phrases used within the given tweets. The results showed that directed tweets (specifically mentioning a user) contained a greater number of OOV terms (innovations), with 'standardised' English language used when the message was intended for a larger audience (such as the use of a conference hashtag). This suggested that, for a message to be seen as genuine by a community, it needs to be in understandable 'normal' language, and interpersonal messages can play with language more. Although, as highlighted in Section 2.2, the pressures of communication online affect the language used, the research did not take into account the time constraints in directed messages as these represent an active conversation where time is important in responding. Thus, a greater pressure to respond could result in a greater number of mistakes as the user does not have time to correct their spelling.

The cited studies do not look at quantifying the influence/power that a user has but rather show the effect of power and influence on users' language. As highlighted by user accommodation of language, users' actions are influenced by the individuals around them, with influence being defined as 'the power individuals have over the actions of each other'. Research into influence within and across OSN has modelled the influence of individuals as a measure of user *activity* or *popularity* within a given network. Measures of user influence can be seen in [166], who proposed that the influence of a user can be broken down into three components: their in-degree influence, their re-tweet influence (proportion and quantity of tweets re-tweeted), and the mention influence (the number of mentions they receive). By using these three measures, they were able to identify public figures and celebrities due to their high number of mentions and in-degree metrics. However, the most re-tweeted content originated from news organisations. They concluded that a user's in-degree metric is a measure of their popularity rather than their influence, with the ability to cause people to re-tweet an action quantifying influence, to a greater extent. However, these measures were ultimately based on a structural definition that only represents the potential the user has to influence fellow users' actions, and not which actions they have caused.

Alternatively, [12] identified that the size of the information cascades (number of re-tweets a tweet receives) that a user initiates correlates with their follower number. Over a two-month period, they mined the diffusion of URL over Twitter, with the resulting influence measure representing the user's ability to seed cascades of information. The influence to cause a cascade takes one of two measures: first, *local influence*, which is proportional to the number of neighbours that post the URL after them, and *total influence*, which is the *log* average size of all cascades they seeded. By then using this as a ground truth (the number of followers the user has), they then used cross-fold validation and a regression

tree model to predict the influence of a user from a number of structural and activity features of the user. This showed that there was a strong correlation between the number of followers of a user and the influence that they have. However, this analysis did not look at the nature of the content that was diffused across the network, and presumed that there was no external influence affecting the actions of a user. Additionally, when quantifying the influence of a user two measures of where derived; *local* and *total influence*. Though, both measure the same effects, additionally if a user has a large number of followers, then one would think that they would have a large local influence.

From here, influence is defined as 'The effect of one user over another to change behaviour'. This definition moves beyond the measures of activity and popularity and identifies how influence is a user affecting change in another user's actions. These changes in user actions could be from their choice of language to their adoption of a new technology or a change in opinion. Again, through the use of cascades of information, [84] proposed a system called GuruMine, a system that not only looks at the size of cascades that the user initiates, but also the number of nodes in the information cascade that are added due to the user adopting the innovation. Additionally, unlike [12], [84] limited the time that a user can be classified as influencing (being influenced by) the cascade. This meant that the influence of a user was not indefinite but rather finite. Thus, for a user propagation graph g for action a, they only have influence for time t. In addition to this, a user is only said to be a leader if the size of the propagation graph $g_{u,a,t}$ has more than a certain number of nodes. Building on this, they stated that a user could be classified as a leader or a tribe leader depending on the number of users they had significant influence over for a prolonged period of time. Building on the mining cascades of action (this time, tagging photos on Flickr), [83] proposed that influence was a function of joint actions between neighbours within the network. However, not all joint actions may indicate influence between users as time between actions may be significant and thus the influence to perform the action has come from elsewhere. Thus, an action or information must propagate within an average time window (computed as the average time of inoperative between the two users being assessed). This also meant that by modelling the influence between users, rather than across the whole network, we could quantify the influence between individual users. Once these influence values had been learnt, the resulting model allowed for us to quantify the influence a user is feeling to perform an action. Thus, through the application of Receiver operating characteristic (ROC) analysis, we could learn a global threshold which when breached users perform actions.

However, even though a user many be influential in their immediate network, are they 'leaders' on a global scale? [82], [84] proposed that a 'tribe leader' is a user who starts a significant number of cascades compared to the whole network. However, one issue with the proposed approach of modelling influence between users is the inability to distinguish between the social influence between users and the homophily of user association. When looking at what causes a user to perform an action, we must distinguish social

influence from homophily, as a user's action may not happen due to their influence but rather based on their latent attributes; distinguishing social influence from homophily is similar to distinguishing correlation and causality, with a classic example being the correlation between the increase in beer and diaper sales at the retailer Walmart. To distinguish social influence from homophily, [175] proposed a match sample estimation framework, with similar frameworks being proposed in [9], [126]. By identifying infected nodes then comparing them to similar uninfected nodes, [175] was able to distinguish between these two effects, with the results indicating that homophily can account for up to 50 % of perceived behavioural contagions. However, as with much of the research that focused on online communities, the choice of a user to adopt an action is binary, with little explanation of external pressures to which the user may have been subjected.

Traditionally, homophily followed boundaries of gender, race and class [5]. In recent times, this has not been the case within OSN. [56] showed that social ties form between users based on topics and interests rather that demographics. Although not all users form ties based on the same variables, by classifying users based on the structure of their ego network (classifying users as *generators*, *mediators* and *receptors*), they determined that generators (bots and automated accounts) formed connections based on the activity of users, moderators (normal users) relied on location and interactiveness, whereas receptors (celebrities, users within a high in-degree) relied to a greater extent on location and topical interest. However, user attributes had to be self-declared, which is limiting within this research as only 14 % of users revealed their location, of which only 1.5 % appeared to be relevant.

Influence is not only expressed by one person on another, but also by groups and communities on each other (as stated by [87], [196]). This can be seen in [65], who modelled influence over the geographical landscape as the adoption of trending hashtagss or phrases on Twitter. A fully connected direct network of metropolitan areas in the US was generated, with the edge weights being proportional to the number of times a hashtags or phrase appeared in location X before location Y. However, a fully connected graph gives limited information, so statistically insignificant edges were pruned, resulting in no significant edges of information flow between nodes. The results showed that, when assessing influence across a geographical network, we can see the effect of both local and global forces. Similarity on a local level was measured as the Jaccard similarity between the set of trends in each location. By then using a hierarchical cluster, we can see that there are four distinct trend regions in the US. This allowed for the ranking of the trend setters and trend followers. However, as with [64], in modelling the diffusion of language across geographic landscapes, the models showed strong influences of large metropolitan areas on the east and west coasts, with the diffusions appearing to follow airline travel patterns.

3.3 Innovation Diffusion

The previous two sections have looked at identifying the innovations and understanding the processes and influences that cause users to change their language and adopt language innovations. The following section focuses on the methods and features used in predicting the rate, depth and speed of diffusions in OSN. This is different to the previous two sections, as this section looks at the collective actions in innovation adoption across networks.

Traditionally, tracking the diffusion of information relied on identifying ideas, thoughts and political movements over time to see how users adopted them. However, with the growth of large networked datasets, the tracking of diffusion has included news items [134], memes [169], links [12] and topics in blogs [90]. Though, the methods and techniques of analysis relay mainly the same users in a network becoming *active* or *inactive*, with the *active* nodes affecting the rate of *contagion* between users. This information does not just diffuse through a network in a random manner, but rather relies on action by the users, the collective actions of communities and the structure of the network in which it diffuses. Similarities have been drawn between information diffusions and the way in which viruses spread through populations, by simulating information cascades. [207] showed that viral content (hashtag) diffuses though a network as a complex contagion being affected by social reinforcements and homophily maintaining a diffusion within a community, rather than a simple contagion such as a disease from which one exposure could be enough to adopt. They showed that the content that became 'viral' moved beyond spreading like a simple contagion, with the diffusions not being affected by the community structure. However, they did not distinguish between the effect of homophily and social reinforcement. Using paired sampling, [9] identified that half of a contagion is due to social reinforcement, with the other half being a product of homophily in the network.

However, each user within a network makes the choice (active or inactive) to pay attention to items of information, as the amount of attention that a user has is finite. This limited attention can be seen in the time that users give to a limited number of topics. [203] quantified the total amount of attention that a user has for topics through the use of information (Shannon) entropy across the set of hashtags that they used in their posts (with the hashtags representing the number of topics in which a user is interested). This measure showed that, as the network grows (as new users join Twitter), so does attention (increasing its collective attention as the number of hashtags increases). However, the attention of each user stays the same as their number of active hashtags stays constant. This limited attention of each user hinders the diffusion of information a set of users collective attention could be full, thus blocking the diffusion, with the meme only being adopted if the extrinsic attributes correlate with items already within the interest (and thus attention) of the user. However, the interests of the user appear to be correlated with the user's position within the given network. [203] assessed users' topic diversity, showing that more popular users within Twitter (higher follower counts) had a smaller topic diversity, indicating that they

talked about one or two narrow topics, whereas less popular users had a large diversity of topics in which they were interested. Building on the diversity of topics of the user and attention of the network, [205] identified that the cumulative topic diversity of early adopters is an indicator that a message will have a large diffusion. However, the content in the message that is being transmitted must be focused, thus allowing for users with different topics of interest to understand and adopt the message.

Networks not only exist between users but also between geographical location though people moving between locations with users taking information with them. As stated earlier, the language people use is connected to the geography. Using latent vector auto-regressive modelling, [64] showed that they were able to identify high-level diffusion patters across the US. Similarly, [65] attempted to assess the propagation of trending topics on Twitter across the geographical landscape in the US. The topics were mined across a period of time from each distinct location in the Twitter database (Twitter customises trending topics to the location of a user). To assess the flow of information flow/diffusion across the geographical network, a fully connected directed weighted graph was generated, where the edge weights were proportional to the number of topics that had come before a given node; thus, a higher weight indicated more information travelling from one node to another. However, in much of the research that looks at information diffusion across the geographical landscape in the US, a strong east-west influence is detected, which bears a strong correlation to US air transport networks [174]. [148] showed that it is across these transportation links that language diffuses as the links represent weak ties and structural holes within the network.

Memes do not only exist in one network but rather jump from one social network to another, and even into the blog and news spheres. Using hierarchical edit distance clustering, [134] showed that we can track variation of quoted phrases across traditional online news outlets and the blog-o-sphere. The results showed that the peak activity (across blog and news websites) for a phrase appears on average 2 hours after the peak on traditional media. It should be noted that the use of edit distance and clustering of content is inherently challenging, with a large number of false positives. However, this highlights the important of perceived influence from the source of the information on the diffusion of information. This can be seen in language change, where users adopt features and topics from other users who have greater influence in the network [45].

[135] showed that the majority of these information cascades in the blog-o-sphere take the same shape, with the majority having a small number of nodes with limited depth; only rarely do cascades become large and viral. However, information diffuses across multiple trees (with multiple sources); thus, [121] proposed that looking beyond the signal tree and instead looking at the forest of trees increases the accuracy of predicting the ultimate diffusion of information across a social network. However, a cascade is ultimately the result of the interplay between user and group action. [214] modelled this interplay in the effect on the long tail diffusion of information. This was achieved by separating out diffusions of

information into three distinct levels: *macro* (user actions), *meso* (community distributions) and *micro* (global long tail distributions). By separating out the distribution at multiple levels within the network, they were able to theorise about the dependence between each, showing that growth that is internal to a community could predict growth in the number of activated communities. However, they treated each network as a static entity and not a dynamic process, which means that, even if users moved between communities, their action would be mis-counted.

Within language, conversational chains mined from social networks can be classified as diffusions of information. [37] assessed the dynamics of conversation chains within subreddits, using three measures to quantify the actions of both the users and whole chain in predicting participation and growth in the chain: *volume* (the number of posts and users in the conversation), *responsiveness* (the time between posts) and *virality* (the probability of a new user attracting more users). When comparing the structural growth of a conversational chain over time (the structure of a chain is measured by the Wiener Index (WI)), the results indicated that larger conversation chains had a greater relevance in between nodes (messages between users). However, the slower the response time, the more complicated the language used, potentially indicating that users are more considered in their responses (see Section 2.2). However, it would appear that the large chains are dominated by a small collection of active users who are core to the subreddit, and thus heavily bonded within the community. Thus, for information to diffuse between communities, it would be on the reliance of the rare users within the community as they span the structural holes of the network.

A user's location in a network can aid or hinder the diffusion of information; by quantifying the role of users as their position within a network, [81] modelled and revealed that users acting as bridges within and external to the community are more likely to be used as an information source; the users who are the bridges are the most active, but rather represent the influence a user has in the positions they take within a network. This bridging effect can also be seen in multilingual users who act as bridges between language communities on Twitter [92], [116]. Unlike [81], users who are both bilingual and non-native English speakers had a greater influence compared to similar monolingual users or bi-lingual users with English as their first language. This could be a result of the sampling within the study or could suggest that, as English is the dominant language within Twitter, it is easier to identify bilingual non-native English speakers. Intriguingly, [149] proposed that the reason for the spread of English is the mobility of English-speaking nationals, whereas, now, the diffusions of English may be a factor of the mobility of foreign nationals.

Information and innovations diffuse across a network, moving from one user to the next depending on many varying factors. This continuous flow of user actions (the process of adopting and rejecting the innovation) can be seen as a cascade of information or actions across the network. A cascade can also be thought of in the form of a diffusion tree, which is defined as the time ordered sequence of connected

nodes that mention the same item of information [135]. [121] proposed that an information cascade within a network does not have to have only one point of origin (and thus be modelled as a tree) but can have multiple independent origins (and can thus be modelled as a forest). This was achieved by extending the Wiener Index (WI) to be the average across all trees within the whole cascade. Additional metrics were introduced to quantify the distance between trees, with the aim of seeing whether they were within the same section of the network. This comes from the networks that are observed online as having numerous external influences that affect the actors within the observations. For language diffusions, this could be seen in language adoption, where multiple users travel or interact with other users in the same distant locations, each starting a diffusion in their own area of the language network.

Within the field of social computing, when predicting the diffusion of the content, we are proposing the question, 'is this content going to go viral?'. However, the meaning of 'going viral' depends on each researcher's own definition, though, ultimately, it is an equation to predict the final size of the diffusion (number of times the innovation is used or number of users who adopt it). Predicting if a diffusion is going to go viral and the final size are two distinct questions, one using a threshold at which a diffusion is classified as viral and the other being the prediction of a number. [207] predicted whether a meme would go 'viral' with the two classes of 'viral' or 'not viral' defined as a percentage threshold based on all final diffusion sizes of active users or number of usages in relation to the whole network (e.g. a viral meme is within the top 5 % of all final diffusion sizes). The model was implemented using a random forest classifier after 50 observations of the hashtag, drawing on features such as the concentration of the hashtag within a community or the time between sequential usages, with the aim being to predict whether the final diffusion would be above the given threshold. [206] also used percentage thresholds to classify whether the content would go 'viral'. However, instead of using the first N observations, they used all observations within a set time period from the first observation. When the threshold was set to 50 % (i.e. if the diffusion would be in the top half of all diffusions), the predictive models only achieved results that were marginally better than a random baseline. As the threshold increased, the accuracy of the model increased; however, this could have been the model learning from the distribution of the examples and not the features themselves.

However, [208] proposed that, instead of using a threshold (binary classes), we could instead segment the range of diffusion sizes and treat the prediction problem as a multi-class classification problem. Each meme is assigned a diffusion class based on the final diffusion, with the class definitions being log scaled bins (e.g. $\log_{10}|T| \pm 0.5$ or $\log_{10}|A| \pm 0.5$), resulting in classes of increasing size. This method resulted in imbalanced classes as the majority of diffusions were in the range of $\log_{10}|T| \pm 0.5 \approx 2$ ($31.623 to 316.228$), with the results performing worse than the random baseline. The effects of imbalanced classes were also seen in [207], discussed in the previous paragraph.

The reason for imbalanced classes comes from cascade sizes following a power law distribution. This

means that a larger proportion of diffusions only have a small number of node activations, and a relatively small proportion having a large diffusion size. Thus, moving the threshold to a higher value (i.e. going 'viral' only accounting for the top 0.1 % of diffusions) means that the distinction between values will become more distinct, resulting in higher accuracy. This questions whether using a binary or multi-label classifier is the correct method in assessing and predicting the diffusion of content across networks.

We can also predict the final size of a cascade rather than predicting the diffusion class in which it may be included. This can be seen in [185], who proposed a regression model against the number of views of content within the first 30 days since creation. This showed a correlation of growth over time, though it did not use any features from the network, only highlighting that popular content becomes more popular. Similarly, [195] proposed a hybrid regression model that used a number of features mined from the content in Twitter, such as the re-tweet ratio and the number of followers that a user has. The value for which the model was optimised was Mean Square Error (MSE) between the predicted and actual size of the cascade. However, the hybrid model failed to take into account the power law of cascade size, thus the MSE of 3.5 achieved across features sets could, in reality, be much larger for larger diffusions.

Instead of predicting or classifying the final size of popular content cascades, [36], [121], [135] aimed to understand the predictability across the entire lifetime of a cascade, and utilise the power law within the model. They proposed that predicting the future cascade can be thought of as a binary classification problem, where, after observing k re-share actions, we can predict whether the cascade will be above or below the mean end cascade size for all cascades with at least k re-shares. As has been shown, cascades' final size follow a power law distribution, which means that, for an $\alpha = 2$, half of the remaining cascades will have a final size less than the mean, thus creating near-balanced classes. Alternatively, this can be thought of as only 50 % at any point doubling in size. Similar to [208] and [207], features used across these models included measures of the structural and temporal dimensions of a diffusion. When assessing the individual features and their influence on the model, we could see that, as the cascade sizes increased, there was a change in the influence of varying features. For smaller cascades, the root sharer (source) of the content had a large influence on the final size, but for larger cascades it was content features and who else had seen/shared the content. This change in feature importance can also be seen in [207], where the community influences the popularity of the content diffusion.

As show in this section the diffusion of content, the assessment and prediction can take many different forms. [207] showed that community structure heavily influence the diffusion of meme, though for meme that diffuses extensively the community structure plays a reduced role in the reasons for its diffusions, as the diffusions process moves from a complex to a simple contagion. Though ultimalty the diffusion of content is dependent on the actions of users, with [203] showing that users have a limited attention and that what they do addopt have to both come from a source with a narrow topic profile but also

be inline with their topic profile. Though in predicting the size of diffusions a number of methods have been developed, from simplistic applications of regression models to classification models that had be developed with the power law distribution in mind. Though all of the models used features mind from both the topology of the network and temporal spaces of the diffusion. Traditionally, content tracking across OSN has been limited to defined systems, such as hash-tags, images or phrase.

3.4 Summary

Part two of the literature review has focused on the technical aspects of social computing, how theories have been applied and how these can be used in relation to language innovations and language change as seen across OSN. This has resulted in this chapter being broken down into three sections, each reviewing a distinct component of language innovation and change in the context of social computing and OSN.

The first section (Section 3.1) focused on how language innovations are identified and how their meaning and sources have been identified using pointers from within the data from which they were mined. Across the identified work, a number of heuristics have been developed and applied, which first allow human raters to assess which words (or word forms) should be included in a dictionary, with aspects of the heuristics being used to train classifiers. However, this has not been taken further back in the pipeline in automating the identification, still relying on identification through a manual process before being rated. To further automate the process of innovation understanding, [41] again applied grounded knowledge and statistical models in identify the source words from word blends. However, due to the amount of noise within the OSN data, there was a large number of false positives within the results. It was also shown that the meaning of the innovation is contextual to where it is used, and that this should be taken into account when assessing the meaning of innovations.

The second section (Section 3.2) looked at how users come to adopt content and actions within OSN (not language innovations), looking particularly at the power dynamics between users and how this can be modelled within OSN. Users have been shown to accommodate their language to the more powerful users around them as they want to mimic the leaders in order to gain their attention or respect. The influence that users have on each other can be modelled as a function of cascades of these actions, which is in contrast to measures that use topology features (e.g. the degree of the user). However, measuring influence as a function of actual cascades is a greater evidence of influence, as structural measures only giving evidence of popularity not influence. The process of following more powerful/influential users within the network can also be seen in users performing actions or adopting new technology, with influence measures being used to predict when a user is going to adopt an action. However, there is the challenge in distinguishing whether the adoption is due to social pressures (e.g. social reinforcement) or homophily, with the results indicating that social influence, at times, only accounts for 50 % of the

reasons for users adopting an action.

In the final section (Section 3.3), we looked at the diffusion of content (memes and information) across networks. This took a number of forms, from understanding the processes involved within the diffusions, to predicting the final size of the cascade. Content that 'goes viral' has been shown to spread like a simple contagion, with the process being affected less by the community structure and social reinforcement than the majority of diffusion processes. Users have limited attention; therefore, when adopting actions, they must decide if it aligns with their interests and if they have spare attention to give (as each user has a finite amount of attention to give). The number of topics to which a user gives attention influences their ability to generate cascades; if the user has a narrow range of interests, they have a constant message to their follower, thus giving them a greater chance to diffuse. The methods used to predict the final size of content diffusions vary, and a number of binary classifications (is the content going to go viral) or multi-label classifications are used across the literature. However, the issue with using such methods is that they have unbalanced classes due to the power law distribution of final diffusion sizes. Alternatively, it has been shown that we can utilise the power law distribution and rephrase the question as, 'will the diffusion double in size?', which means that we have balanced classes in the model.

Across the three sections, we can see that there are many different facets that need to be considered when assessing and predicting content/innovation diffusion across a network. These include how we identify and understand the innovation, to the complex interactions between user(s) and community (network) actions. In assessing language and language change, the challenge lies in taking the grounded theories and applying them to the methods influenced by the literature within this section. In doing so, we aim to show that language change can be first detected, then modelled on both an individual adoption level and network-level diffusion.

Summary

In the two sections of this literature review, we have highlighted both theories of language change, along with the efforts within the field of social computing to detect and model the dynamics of language within OSN.

Within Chapter 2, we highlighted that language and language change are dependent on the communities in which the language is used, as this allows a common interpretive scheme to be used within the community for communication. However, it is the social structures and the interactions of the individuals in the system that facilitate language change and evolution (intentionally or not). This can be seen in large-scale historic changes in English that have been attributed to a geographically mobile society, allowing English to influence and be influenced by communities from around the world. However, the reason for users adopting language change and innovation is an internal process dependent on many unobserved variables, such as prestige and power.

In the second part of the literature review (Chapter 3), we focused on social computing, looking at how language has been used to model the interactions of individuals and communities, but also how researchers have dealt with the increasing noise coming from the growth of OOV. In focusing on OOV usage within OSN, one can see that a number of strategies have been used to deal with them, such as normalisation or misspellings and identifying the source words of blends, though this removes some of the value of the data. Although a number of methods were highlighted in modelling content diffusions and user action adoption across social networks, these ultimately tracked explicitly (well-defined) information across the network; one can draw comparisons with the diffusions of language innovations.

Ultimately, across these two chapters, we have shown that there is a gap in dealing with language change and evolution within the field of social computing. Instead of thinking of language as separate from the structure of the social network, we must think of users' usage of language as being what defines and helps form the social structures, and the social structures help form the language. In this thesis, we apply the theories and methods identified here to OSN datasets to identify and model language change across a number of social structures.

Chapter 4

Research Methodology

4.1 Introduction

The following chapter introduces the epistemological stance and research methodologies that are applied in order to answer the research questions outlined in Chapter 1. The methodology developed for this thesis is heavily influenced by the manner in which language change is theorised (through the separation of the innovation, the user and the network), resulting in a pluralist methodology and framework being applied. The foundations of the pluralist methodology can be seen in Chapter 1, which introduced the research and separated it into four distinct questions, three of which are focusing on one of the three components of language change.

However, it is not only the nature of the social systems that must be take into account but also the nature of the data that is used when formulating the methodology. Big data is used throughout this work; thus, a number of stances first need to be outlined. In doing so, section 4.2 explicitly defines the scientific paradigms and the epistemological stance (as influenced by big data), as well as how these have modified the methodological framework developed for this thesis.

A pluralist methodology and the use of big data results in a generalised research framework (section 4.3) being implemented across the three research questions (1.1.1, 1.1.2 and 1.1.3). Using this framework, a commonality between the methods of the three questions emerges, such as data source and pre-processing pipelines. These are developed in sections 4.4 and 4.5, which includes the extraction of the social networks and identification of *speaker innovations*.

Finally, as language is innately human and the data collected is generated by users and individuals, ethical considerations are detailed in Section 4.6. These ethical view-points critique the public and private perceptions of the data producers (users), and uses the relevant framework in reporting the results in an ethical manner.

Table 4.1: The four scientific paradigms, from [118] and [98]

Paradigm	Nature	Form	When
First	Experimental science	Empiricism; describing natural phenomena	pre-Renaissance
Second	Theoretical science	Modelling and generalisation	pre-computers
Third	Computational science	Simulation of complex phenomena	pre-big data
Fourth	Exploratory science	Data-intensive; statistical exploration and data mining	Now

4.2 Epistemology and Paradigm

The ever-increasing use of big data in answering social science questions challenges the traditional epistemological and theoretical perspectives taken in social and scientific research [117], [118]. Ultimately, this change has come from the access to research data, which traditionally has been 'data sparse', relying on sampling methods and techniques and curated datasets. In recent times, there has been a move to 'data-rich' studies (for which the data is collected from the whole population, or no (sub)sampling methods are applied). This movement has allowed the researcher to collect and process not only a sample of data but the whole dataset in its entirety. A prominent example of this movement can be seen in the emergence of 'digital humanities' [118] (the intersection of computing and the traditional humanities fields), giving the ability to access not one novel at a time but a whole library. This movement of accessing whole populations of data has caused fundamental changes, leading to the proposition of the *fourth* scientific paradigm by [98].

By tracking trends in scientific methods, [98] identified the existence of four paradigms of scientific method, each characterised by the form that the research method takes. Each paradigm evolved due to both the limitations of current methods and the development of new techniques. This evolution can be seen in the movement from the second to the third paradigm, with the introduction of computational power that allowed for the simulation of the theoretical models defined in the second paradigm.

This thesis sits within both the *third* and *fourth* paradigms due to its reliance on big data, its use of statistical and data-mining methods, and the use of computational science to simulate and extract diffusion processes. Placing this thesis across the two paradigms allows the research to not only look at one unit of change, or one area of a social network, but, through the development of custom systems' *generalised models*, to access language change across whole systems and multiple networks.

The growth of big data has led to a number of authors arguing about the return of 'empiricism', with a number of academics believing that big data 'speaks for itself' and that there is no need for theory-influenced methods in the scientific process [8], [183]. [183] further states that results from big data are the absolute truth and need no further interpretation. This is due to big data algorithms not being effected by research bias, along with the data sampling the whole of a system ($n = all$) [162]. However, empiricism is not the stance taken in this research, for a number of reasons, including data collect and sampling, as well as the way in which language is theorised. Language is around us, used in every act and

in every location; therefore, one cannot collect a complete sample of language that has not been influenced by unseen factors. Even though the thesis is about language online, these issues still exist when sampling language across the internet. Therefore, the sample collected for this research comes from a sub-sample of society as it is constrained by users who use online social media. Additionally, language is not innate or quantifiable; thus, the methods developed are influenced by the researcher's perceptions of language and the theoretical stance that they take on language change. Thus, the methodological stance of this work is that of a *post positivist*; this advocates methodological pluralism, asserting that findings are approximate rather than absolute truths, through the application of quantitative/experimental methods that demonstrate rather than confirm or develop theory. Thus, the truth that is revealed by the study is not the absolute truth, but rather an approximate truth due to the limitations of the datasets and models developed. However, the truth in the results sits in the wider context of the research and the thesis. Therefore, this work at it's *interpretive* in nature as it suggests that the truth in the results is contextual, aiming to generate new theory through the application of hermeneutics or phenomenology.

Additionally, the application of *interpretivism* in this thesis plays into the idea that big data systems and research cannot be separated from the social context in which they were designed and developed. The social context, therefore, affects every aspect of data collection in this thesis. As the data collected about each user has been collected by a person (the development of the OSN), the manner in which the users interact with the OSN has been designed by a person (again, the developer); thus, the interaction is constrained by the system and ultimately by the designer. Thus, one cannot have an objective view of a system, and can only assess the data and results in the context of the system. [118] further believes that one cannot apply empiricist methods and then ignore the research from the past century, as seen with physicists applying models to human populations. This comes from the understanding that, even though a pattern exists within the data, one should not take it as a result or full meaning simply for results' sake, without framing it in the wider research context.

4.3 Research Framework

Building on the epistemological stance, the following section introduces and builds a research framework that is applied across this thesis and to each of the three research questions. The framework is *hypothetical-deductive* in nature, and each question is answered independently of the others. However, aspects of the process applied to each question will have commonality across all questions (figure 4.3.1).

The global framework (figure 4.3.1) allows the research to be executed in a sequential (vertical axis, allowing the research to be conducted in a logical order) and concurrent (horizontal axis, allowing different stands of the research to be conducted at the same time) manner. Running the research in concurrent strands allows the development of three independent but interconnected research questions that sit within

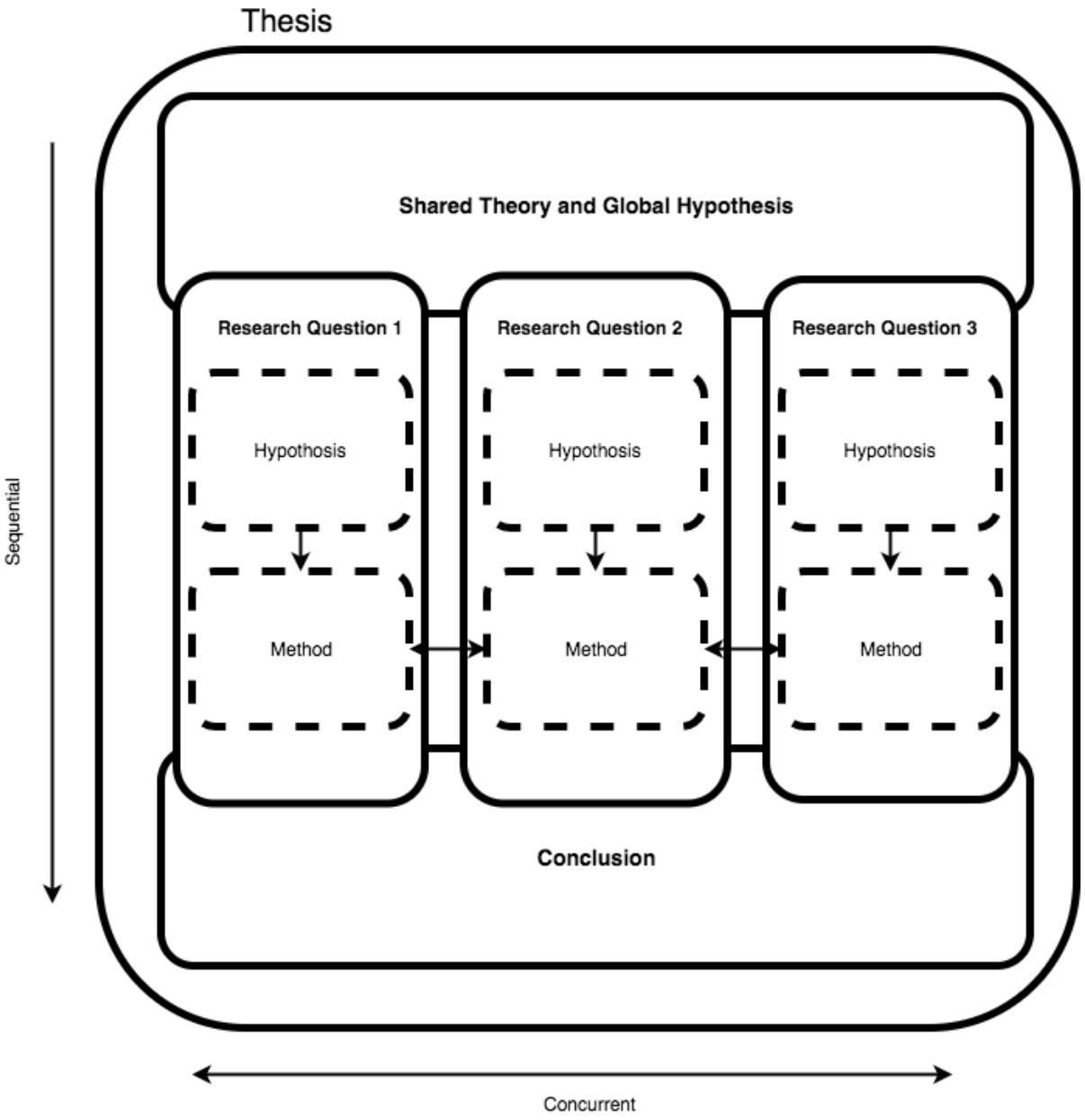


Figure 4.3.1: Applied research framework

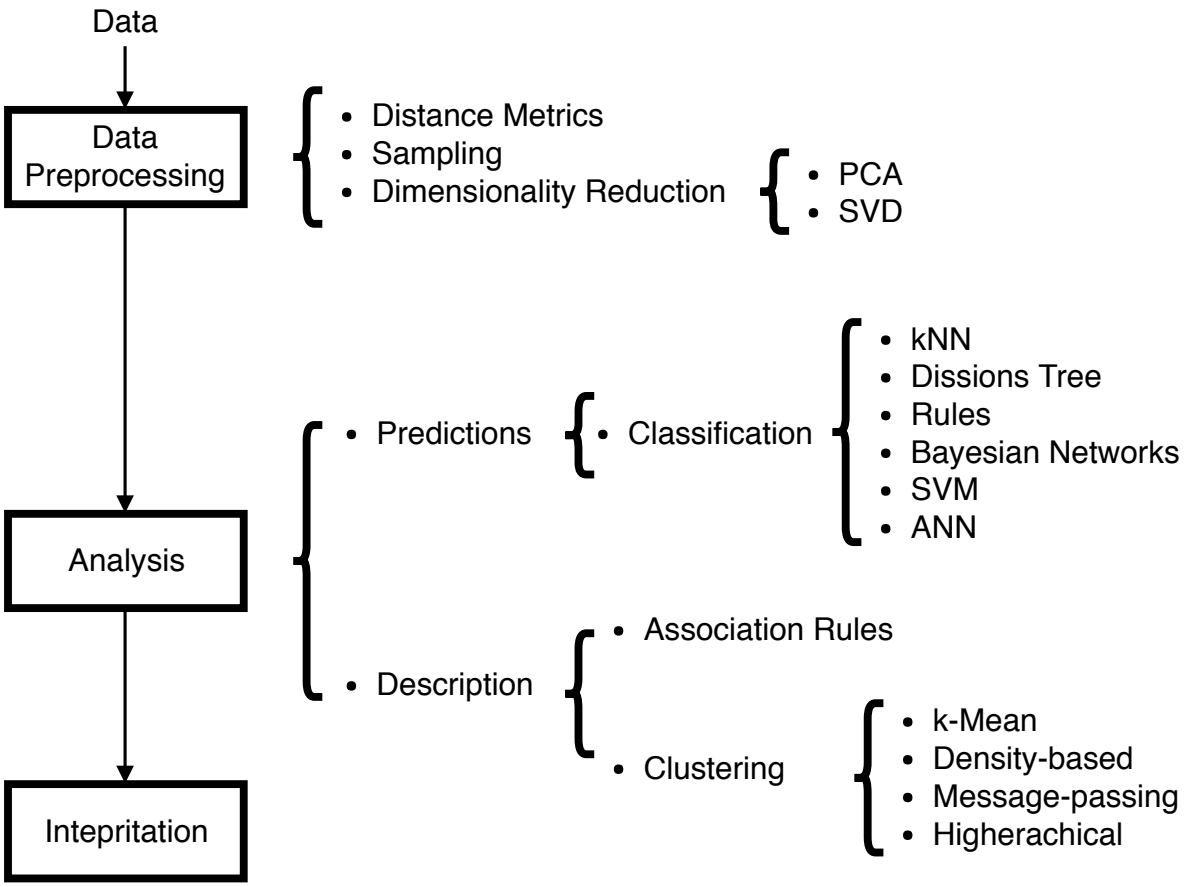


Figure 4.3.2: Data Mining Pipeline, from [7]

the winder context of the thesis.

The framework introduces a common shared and global hypothesis, which influences and guides the hypotheses and methods applied to each of the three research questions (1.1.1, 1.1.2 and 1.1.3). With each question focusing on one aspect, there is limited redundancy in the subsequent hypotheses and methods. In addition, the framework allows the development of independent methods and validation for each question. However, each of the individual validations and conclusions will influence the final discussion and conclusion with regard to the overarching research question and theory as a whole.

For the three research question the methods will follow the framework proposed by [7] (figure 4.3.2). [7]'s framework separates the methods applied in the data-science pipeline into three distinct stages: data pre-possessing, data analysis and interpretation/validation.

Pre-processing - Initially, the data collected from each of the OSN is not in a form in which it can be analysed. Therefore, a number of stages must be applied to extract the relevant information from the datasets, such as user community membership in the social network, or assigning POS tags to each word. These methods can be in the nature transformation or dimensionality reduction, thus making the data easier to process at later stages.

Analysis - This stage takes the hypothesis about language change and develops and applies methods to the given pre-processed data in order to answer the given question. It is at this stage that the analysis of language change will take one of two distinct forms: either analysis of the data in a *descriptive* manner, or taking the data and developing *predictive* models. Applying descriptive analysis to language change involves the use of data-mining techniques such as clustering and measures to describe how language has changed. Alternatively, *predictive* techniques can use historic language change and descriptive measures to train models, giving the ability to predict either what can be classified as language change, or whether a single speaker innovation will cause language change.

Interpretation - The final stage within the pipeline is the validation and verification of the results within the context of the research questions, hypotheses and wider body of research. This stage limits its involvement of computational methods, but realises the qualitative and interpretive analysis of the results, discussing them within the acknowledged limitations of the methods developed and data collected, as well as identifying where they sit within the broader context of language change. The validation from each of the three research questions within the global framework will be triangulated with the aim of answering the global overarching question.

Even though the methods are developed independently of each other and applied in isolation (as identified in figure 4.3.3 and 4.3.2), there are areas of commonality across the three methods. The shared methods are predominately found within the pre-processing of datasets in aspects such as the extracted social networks shared across the three research questions. These shared pre-processing stages are built upon in Chapter 5. In addition, structuring this work in a number of embedded and shared frameworks allows for the research to be performed in a manner that is understandable, deductive, grounded and replicable. This allows a collective validity of the three research questions to be established, as well as allowing them to answer the overachieving global hypothesis.

4.4 Datasets and Data Collection

The premise of this thesis is to identify, model and predict the diffusion of *language innovations* and ultimately quantify *language change* across/through OSNs. The prevalence of CMC/OSNs can be seen in the ever-increasing usage of OSN as a core component of inter-personal communication (both personal and professional) [120], with them aiding in relationship building and community formation. This is typified by Facebook[1], which is, by far, the dominant social network in the world, with 1,600,000,000 active users [59], allowing users to connect with friends around the world, share pictures and create events, to name a few features. However, depending on the need of a community, specialised social

[1] https://www.facebook.com

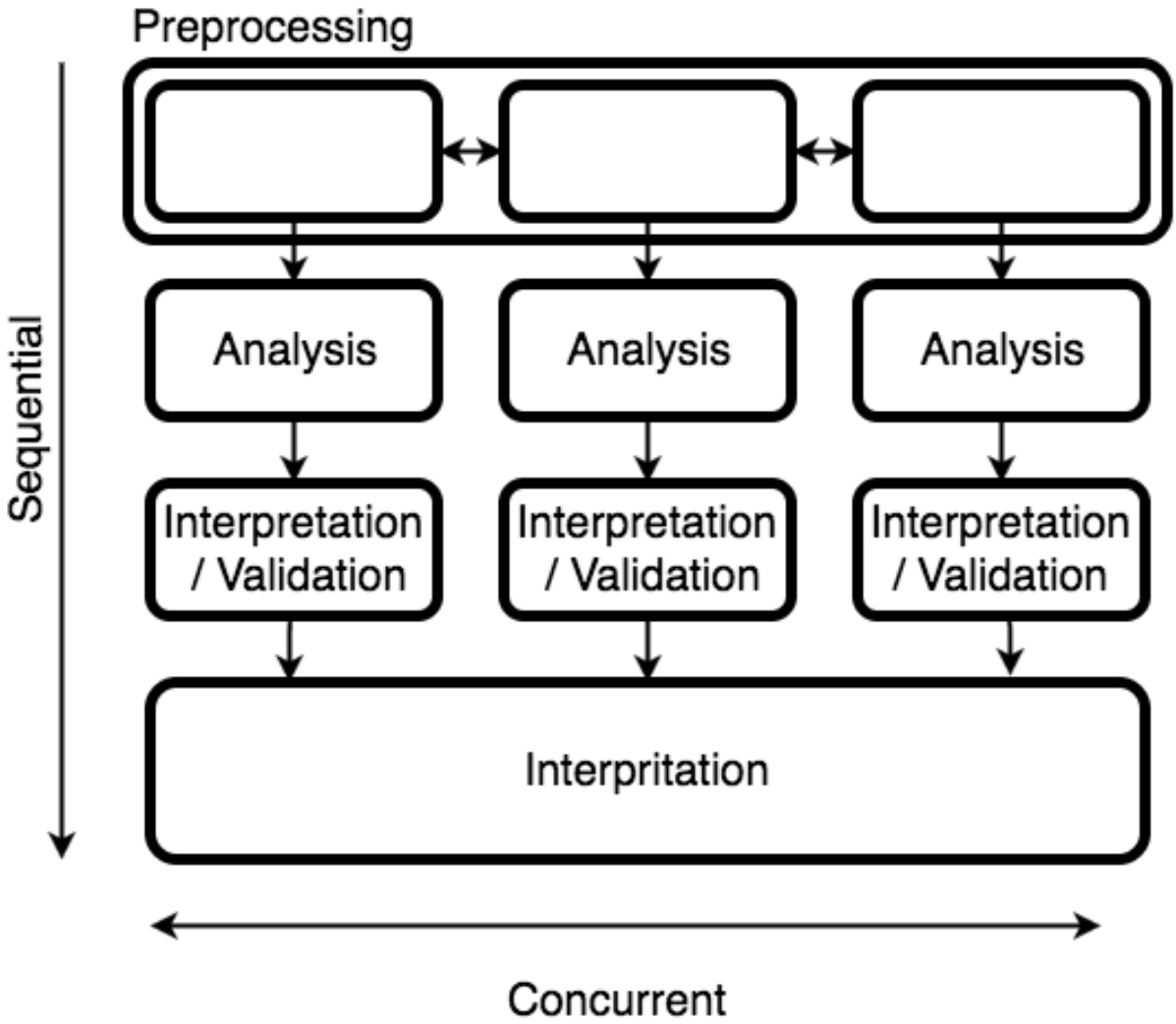


Figure 4.3.3: Diagram represents the modified version of the data-processing pipeline found in figure 4.3.2, processing is shared across all three research questions, each question has distinct analysis and interpretation, then a final collective interpretation.

networks have been developed, from Naked Wines[2], for sharing a passion for wine, to Etsy[3], for buying and selling home craft products.

For the first time, researchers have had the opportunity (to an extent, subject to limitations on API access) to access samples of data stored within OSN; this data contains high-granularity user interactions with the system and with each other, such as those representing user friendships, knowing which users are tagged in the same photo, or knowing which users have been in the same location at the same time. This availability of OSN data has allowed researchers to develop large-scale predictive models that attempt to understand social life and predict outcomes of events. For example, Twitter has been used to predict flu outbreaks [10], [35], [50], [131], [133]; regional alcohol consumption has been modelled in the UK [111]; online beer communities have been used to assess users' language accommodation [53]; and Facebook has been used to show how information evolves [2].

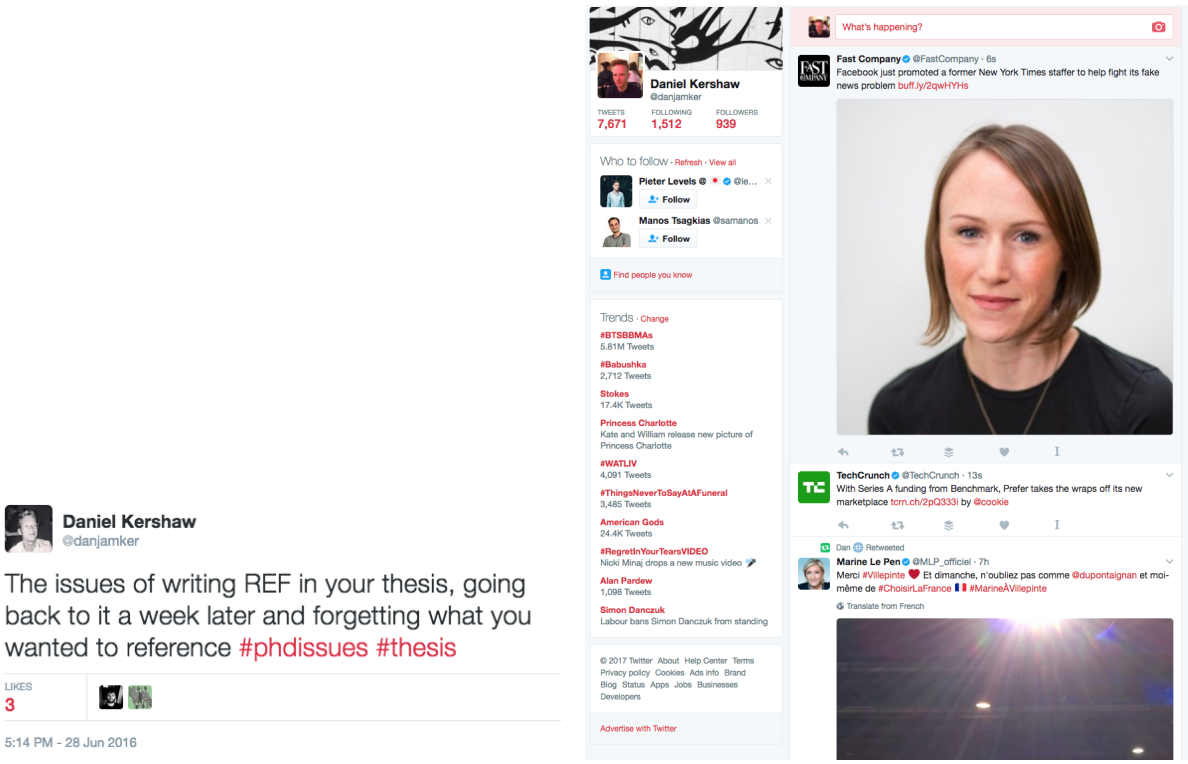
With the aim of using OSNs as the data source to assess individual and community actions, as well as the interplay between the two, this thesis focuses on two dominant OSNs as primary data sources: Twitter (section 4.4.1) and Reddit (section 4.4.2). They have been chosen due to their popularity and ease of data access; the content generated is predominantly in written form, and users interact at high speeds with the two OSNs, making the content highly dynamic to the world around the user.

4.4.1 Twitter

Twitter[4] is a micro-blogging platform that was launched in 2007 and currently has upwards of 310,000,000 users in the USA, the UK and globally [5]. Twitter allows users to *broadcast* short snippets of information to each other, known as 'tweets', without necessarily needing or wanting for a response [99]. These snippets of information or *tweets* consist of short messages (140 characters) that can contain links, photos, emojis and text. Users can 'follow' each other, allowing user content to fill their activity feed ('timeline'). However, relationships on Twitter do not have to be reciprocal, thus creating a directed social network graph.

Within tweets, users can mention each other through the inclusion of an '@' tag followed by the username of another user (e.g. @mrsstephenfry), thus bringing the message to the tagged user's attention. Tweets can also be tagged with additional 'hashtags' (e.g. #blacklivesmatter). The reasons for tagging tweets range from attaching the tweet to an ongoing event or bringing the tweet to the attention of a topic community [25]. Additionally, tweets can be re-tweeted, bringing the tweet to the attention of the re-tweeter's followers who may not have seen the information the first time; re-tweets can be identified through the addition of 'RT' at the beginning of the tweet. An example tweet can be seen in Figure 4.4.1a, along with an example of a Twitter timeline 4.4.1b.

[2] http://www.nakedwines.com
[3] http://www.esty.com
[4] http://www.twitter.com
[5] http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/



(a) Example tweet that contains hash tags 'phdlife' and 'thesis' along with the number of users who have liked it, and the time it was created.

(b) Example Twitter timeline. Trending topics can be seen on the left, recommendations on who to follow at the top left, and tweets in chronological order in the centre

Figure 4.4.1: Example components from Twitter

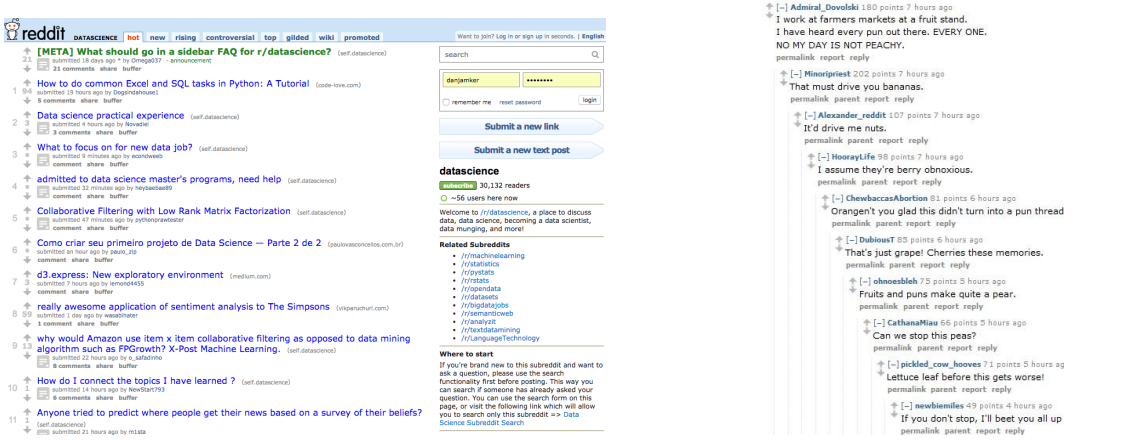
However, even though Twitter was designed as a social network for people to interact with each other, the reasons people give for using the systems vary; these reasons can be broken down into two distinct groups: *content generators* (frequent brief updates about personal life activities) and *content consumers* (people-based RSS feeds). The former are users who want to share content about their personal lives and maintain fast interpersonal contact with their friends, thus enabling them to maintain active social connections even if they are not in the same place. The latter are people who gather information from their social network, but do not interact; this can be seen in those who use Twitter as a news source [125], following companies or just friends in the network.

Users initially accessed Twitter through its website, though users now predominantly interact with Twitter through dedicated applications on their smartphones, which allows them to communicate while on the move. Additionally, when users tweet from a mobile device, the location from which they have tweeted is embedded in the tweet, and Global Positioning System (GPS) coordinates give a tweet a spatio-temporal property, binding the content users have generated (text in tweets or photos) to a specific location and time.

Twitter is used around the world, with the largest user base coming from the US, which accounts for 141.8 million users; however, the UK is fourth when ranked in terms of registered users, with 32.3 million, behind the US, Brazil and Japan [137]. Additionally, within the UK, 80 % of tweets now come from smartphones compared to a global average of 71 % [32].

Twitter was thus chosen as a data source for the following reasons:

- **Popularity** - Wide adoption across the UK
- **Text-based communication** - The dominant form of communication within the network is through written text, be this status updates or comments to other users.
- **Geographical bound** - As reviewed in section's 1 and 2.1, the language of a user is influenced by the geographical landscape in which they exist. This comes from users' connections being influence by the users they are close to and their ability to connect with other communities. The inclusion of GPS coordinates from tweets originating from mobile devices connects the language within the tweet to a location; this then enables the analysis of language innovation across the geographical landscape in the UK.
- **Time bound** - Each tweet is produced at one specific time; thus, in connection with location, one can assess changes in language over time.
- **Social network** - Language changes due to exposure to the language of the user's friends and acquaintances (see section 2.1), so the social network can be extracted from users mentioning each other, or from their follower networks.



(a) Front page of a subreddit

(b) Example thread within a subreddit

Figure 4.4.2: Example components of Reddit

- **Ease of access** - Access to a public API that allows tweets to be collected in bulk with the collection constrained by a given set of query parameters.

4.4.2 Reddit

Reddit[6] can be thought of as an online forum, which, unlike Twitter, is structured into self-governing communities (sub-reddits), where users can post content for the community to discuss. Reddit was created in 2005, and, as of January 2016, it has 19,000,000 unique visitors a month, with an estimated 6 % of all online adults in the US using Reddit [59]. The network is self-governing, allowing users to form subreddits around topics or movements (such as r/soccer of r/AskThe_Donald), where users then post content, links or information for other users to discuss. The structure of Reddit's subreddits bears a resemblance to traditional online forums, with subreddits being thought of as rooms with users creating new threads. Over time, users can also be promoted to 'moderators', who govern the subreddits and have the power to ban users and remove content that violates the subreddits and Reddit's guidelines.

However, unlike forums, threads can fork off into multiple sub-threads, creating a tree structure. An example of a subreddit post and the resulting chain of comments can be seen in Figure 4.4.2b. Unlike Twitter, there are no limits on the number of characters that can be used in each post. Each thread and post has the ability for users to 'up' and 'down' vote posts and comments, which affects the content that is shown on the homepage of each subreddit, as well as what is shown on the homepage of each user.

Like Twitter, users can 'friend' each other, though, again, it is a non-reciprocal relationship. In addition to befriending fellow users, users can subscribe to subreddits that interest them, which populates their newsfeed with content generated within the subreddits. Thus, from these explicitly defined relationships, there are two distinct forms of explicitly defined networks: the user-to-user relationship,

[6] http://www.reddit.com

Table 4.2: Dataset description

	Reddit	Twitter
Posts	1,054,976,755	111,067,539
Users	10,528,521	1,696,630

and the user-to-subreddit relationship.

Much like Twitter, Reddit has significant importance in online cultures, and research finds many recognisable memes and phrases originating from various subreddits. This culture has been embraced by the mainstream, especially an event called Ask me anything (AMA), which is when a user answers questions. This has attracted people such as US president Barack Obama to participate in Reddit AMA. However, Reddit is also known for a number of controversial events, the most notable being gamergate, which highlighted misogyny in online cultures [20] and the banning of a number of controversial and racist subreddits [168]. These controversies and their reporting in mainstream media highlight the importance of Reddit as an online and offline cultural force.

Thus, the reasons for using Reddit to model and assess language change in OSN are:

- **Perceived to be less public** - Twitter is highly public and used in professional life (users attempt to manage their professional reputations); however, Reddit has a level of anonymity that allows users to engage in language that might not be seen in the public sphere.

- **Continuous data production** - As with Twitter, data on Reddit is being produced constantly, and is produced in reaction to online and offline events.

- **Topic-based structure** - Unlike Twitter, Reddit is constructed around topics, which allows language and communities to form around them. This enable the study of the impact of the topic/-constructed community on the language.

- **Ease of access** - As with Twitter, there are public APIs that allow simple and quick access to the data, allowing users, communities and the text to be identified.

4.4.3 Data Collection

When conducting research using data from OSNs, one of the main challenges is the collection and acquisition of the data. This come from companies (data owners) protecting their data as it is fundamental to their business operations. However, to encourage developers to integrate systems with their OSN, they release public APIs that allow limited access to data within the system. The following section explains the techniques used to acquire the data, as well as some basic metrics about the data collected.

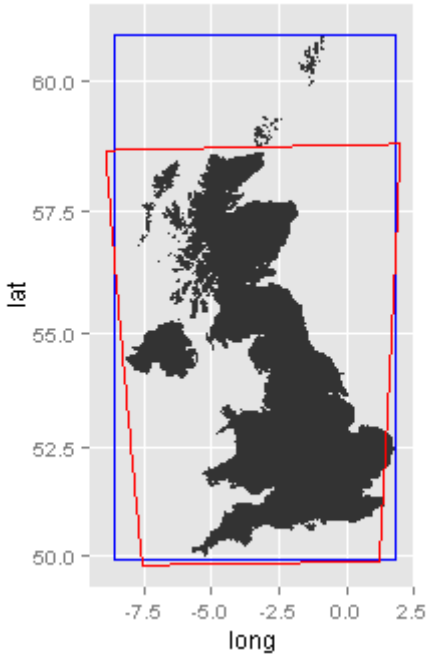


Figure 4.4.3: Bounding box from around the UK where tweets were collected

Twitter

Twitter has a publicly accessible API[7] that allows developers to interact with user data (allowing third-party applications to manage user accounts) and to mine in real time the current Twitter stream (Firehose). To access information, such as user follower lists or historic tweets, one simply calls a REST API endpoint, which returns the relevant information. However, as with Reddit, there is limited to a number of which can be made to the API per-minute. To over come this one can use the streaming API[8] to collect tweets (which are returned within a given search query, e.g. containing the hashtag '#Obama') in real time as users post them. The streaming API works by maintaining one open connection with the Twitter servers for them to post data down; this is unlike the REST API, which requires the user to constantly call the different endpoints. However, like the REST API, the streaming API has rate limitations, only delivering up to 10 % of all tweets that have been created at that moment in time, and not the complete sample.

This research is interested in language and language innovation. As shown in [64], [154], language is dependent on geography, so the tweets collected had to originate from the UK. This is achieved by passing a set of coordinates to the Twitter API in the form of a bounding box. This means that any tweet delivered through the streaming API had to originate from within the pre-defined area bounding box (see Figure 4.4.3). The collection of tweets started from September 5, 2014 and concluded on June

[7] https://dev.twitter.com/
[8] https://dev.twitter.com/streaming/overview

9, 2015 resulting in a total of 111,067,539 individual tweets.

To implement the collection of tweets, a simple Python script was created, which used the tweepy[9] module to interact with the Twitter API. This allowed the coordinates of the bounded box to be passed as parameters to the Twitter API, and the resulting tweets to then be appended to a file on the server. This ran consecutively for the whole collection period on one machine, restarting when the machine crashed, and periodical moving the output into Hadoop Distributed File System (HDFS) on the research cluster.

Reddit

Like Twitter, Reddit has a publicly accessible REST API[10], which allows developers to programmatically interact with the underlying systems from the view of a user (the API has been designed for front-end developers); thus, what a user can do on webpage (e.g. submit content, up-vote and down-vote posts) can all be done though the API. However, as with Twitter, the number of times the API can be called in quick succession is limited to 10 per minute from the same IP address; this is to stop automated bots flooding Reddit pages with spam posts. However, for this work, the limit on the API makes the collection of a large sample from Reddit challenging; therefore, a distributed system is developed that uses multiple IP addresses to circumvent the calls per minute limitations. Additionally, unlike Twitter, where one sends a query to the streaming API and new posts are returned, mining Reddit requires a number of steps to find recently added posts in the OSN.

First, one must identify the recently 'active' boards on Reddit by querying https://a.4cdn.org/boards.json; this returns a list of the most recently active boards from across the network and is updated periodically. From this, each board is then mined. However, requesting the content of one board could then require numerous calls to the API as the is data structured in such a way that it returns only one page (as viewed on the web interface) of results at a time; therefore, large boards that span multiple viewable pages require multiple requests, all of which would break the API limit.

To circumvent the API restrictions, a distributed solution has been developed that coordinates a number of servers, each with a different IP address, to mine different active boards on Reddit concurrently. The system was developed around task/job queues, where a task (representing which board is to be mined) is added to a queue; a worker then gets the task from the queue and then mines the board. The current set of active threads across all boards can be found at https://a.4cdn.org/boards.json. Each board ID is then placed in the task queue to be mined by a worker on a different machine. Using task queues circumvents the rate limits through a horizontally scalable distributed framework. Figure 4.4.4 presents the high-level implementation of the task queue. Additionally, using multiple machines means that, when large threads that require multiple requests are mined, they do not cause a pause in the

[9] http://www.tweepy.org/
[10] https://www.reddit.com/dev/api/

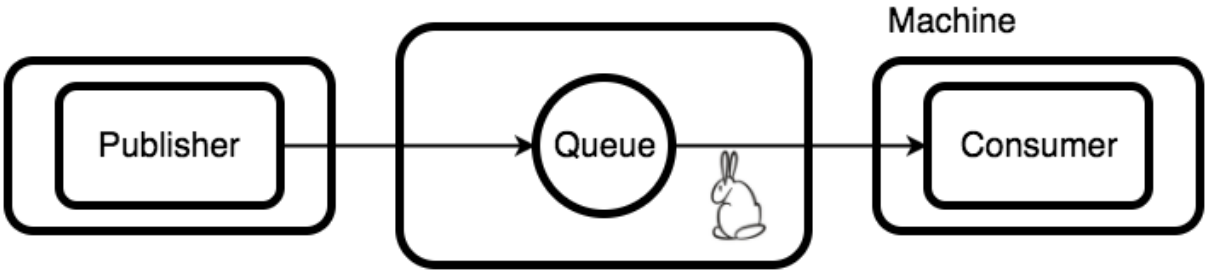


Figure 4.4.4: Diagram representing the Reddit distributed mining system. The miner poles the active boards feed from Reddit, placing the ID of active boards onto the RabbitMQ queue. A consumer then retrieves the ID from the message queue and subsequently mines the board from Reddit. There is one consumer per machine (each machine has one IP address), and the mined board is then saved to a MongoDB instance.

collection of other threads.

The task queue is implemented using RabbitMQ[11], which is a broker software that implements the Advanced Message Queuing Protocol (AMQP). Python's Celery Framework[12] was used in addition to RabbitMQ to manage and monitor task execution across the system. To interact with the Reddit API, Python's PAWN[13] module was used. This provides a Pythonic interface for the Reddit API. In addition, PAWN manages the 10 calls per minute, blocking requests once the limit is reached.

Developing a distributed mining system that uses messaging queues meant the system could scale across a cluster of 20 machines. Thus, running the systems over nine months allowed 90 Gb of data to be collected. This comprised 3,108,844 users and 73,528,954 posts. However, in comparison to the Twitter dataset, the quantity of mined data is significantly smaller, with only 73,528,954 posts from Reddit compared to 111,067,539 from Twitter. Additionally, when comparing the mined Reddit dataset to statistics released by Reddit[14], we can see that only a fraction of Reddit posts have been sampled.

Fortunately, in June 2015, an extensively mined public dataset was released, consisting of approximately 99 % of all comments/posts that are publicly available on Reddit [15]. The size of this dataset is many magnitudes larger than the dataset that was collected through the system that has been developed. Thus, the data used in this thesis is the found data and not the collected data, due to the significantly large coverage of Reddit and the extended time period over which the data was collected.

4.4.4 Limitations

The mined data has a number of limitations in terms of what can be represented and inferred. As with any online data source, there are issues and limitations with the representative nature of the data. The demographics of users of both Twitter and Reddit are not in line with the general population of the

[11] https://www.rabbitmq.com/
[12] http://www.celeryproject.org/
[13] https://github.com/j2labs/pawn
[14] https://redditblog.com/2015/06/23/happy-10th-birthday-to-us-celebrating-the-best-of-10-years-of-reddit/
[15] https://archive.org/details/2015_reddit_comments_corpus

USA, the U.K. or any country in the world, with research showing that the demographics of the user base are skewed to a younger, white male demographic [59], [60]. A number of attempts have been made to apply smoothing techniques to the raw data to bring the representation of users in line with regional demographic data [64]. However, applying similar smoothing methods to the data for this thesis would be complex as the Reddit data is not bound to one country. Additionally, it adds a further layer of abstraction that could introduce errors in future analysis. Thus, for this thesis (as stated in Chapter 1), we focus on the language of the internet as a proxy for language used in the offline world.

Additionally, there are limitations in the methods used to collect the data. Tweets are collected using the Twitter streaming API, which limits what is collected to an upper bound of 10 % of the global firehose. However, when collecting tweets with a bounded box, even though one may get all tweets with GPS points within the box, the limitation comes with the number of users who tweet with GPS coordinates. The number of tweets that contain geotags is relatively small, with only 6 % of tweets having a location feature[16]. Thus, the data collected will be magnitudes smaller than the actual number of tweets being produced. In addition, the sample across the country is focused on metropolitan areas where there is increased mobile penetration and larger populations.

Limitations also apply to the type of data collected in relation to the aim of this thesis. The data collected represents the language of the users and the locations in which it is spoken; however, language is used within many social contexts, represented through social interactions and relationships, which is not explicitly represented in the collected data. Even though these are defined explicitly in the social networks, the information defining users' social relationships is not publicly available through the Twitter or Reddit API. Even though these networks are not explicitly defined, section 5 uses a number of techniques that extract the structure of underlying social networks from information contained within each post.

Even though there are limitations in the data collected, which have been acknowledged, the nature and size of the data should allow us to assess the language and language changes of a large population for the first time.

4.5 Research Methods

As highlighted in the research framework (section 4.3), this thesis consists of three distinct research questions, requiring three independent but interconnected methods. The methods developed for each question aim to be consistent with previous research methods, as well as being reproducible and generalizable, allowing them to be applied to other datasets not used in this research. For this reason, the pipeline/flow proposed by [7] (introduced in section 4.3) will be used to structure the methods of each

[16] https://pressroom.usc.edu/twitter-and-privacy-nearly-one-in-five-tweets-divulge-user-location-through-geotagging-or-metadata/

research question. Each method will be broken down into three distinct stages: *pre-processing*, *analysis* and *interpretations*.

Thus, the following section introduces and discusses the methods used in each of the three questions. Each research question will be broken down to introduce the hypothesis and to give a high-level view of the models developed and the forms of validation used to assess the model. Fuller descriptions of each method are found in the research questions' related chapters (6, 7 and 8).

4.5.1 Innovation Detection

The first research question focuses on detecting the growth in popularity of new terms or phrases, which collectively form *language change*. Traditionally, lexicographers have developed a number of heuristics that are applied manually to new words and innovations to assess if they should be included in a dictionary. However, these processes are time-consuming and costly, requiring extensive manual annotation and analysis. Therefore, the hypothesis for the first research question is: *Through the application of methods influenced by lexicographical techniques, new terms (language innovation) can be tracked and, thus, significant growth can be classified as acceptance in language change.*

Modelling

Traditionally, the collective acceptance of language change has been assessed through lexicographical heuristics such as VERGHT [18] and FUDGE [144]. These heuristic models assess components of language change such as the frequency of use of a word across different genres of media, to the morphological forms of an innovation. However, traditionally, this application focuses on language at one point in time and not variations over a constant period of assessment. This thesis thus develops computational variations of VERGHT [18] and FUDGE [144] and applies them to the time series data. The model consists of three measures:

User and term frequency An increase over time would indicate an acceptance into people's vocabulary

Diversity of usage Users are using the word in a variety of contexts, thus users are accepting it into their everyday vocabulary

Convergence of context (meaning) There is more than one fixed meaning; thus, multiple communities are potentially using the word

As the measures are applied to each time series, the issues becomes how to assess/rank the acceptance of an innovation. Quantifying acceptance is achieved by ranking each word's time series through quantifying it's growth rate by fitting a Spearman's rank. This then enables the comparison and ranking

of different words in a time series. However, as language is defined by community usage, the models developed are applied over corpora, which will enable the analysis of language growth as it starts in small communities and slowly grows in the communities around it, finally becoming accepted at a global level.

Validation

Studying and modelling language poses challenges in terms of the validation and verification of results. This is due to *language change* being a sociological phenomenon, with no 'gold standard' or ground truth to say exactly how and why language changes, and if there is, it is not in a machine-readable format for the data we are using. For the first research question, apparent validation and verification of the model will be performed using a heuristic assessment of the results, drawing on the author's own knowledge of the language system and the context in which the systems have been generated, relating the findings to real-world examples.

4.5.2 User Innovation Adoption

Language change is the collective actions of individuals adopting *language innovations*. However, as proposed by [148] and [129], the adoption of an innovation by a user is dependent on their neighbours. As reviewed in section 2.3, propagation of innovation across a population has been modelled in two forms, initially through the development of *collective attention* models where the user can observe the whole network [86]. However, as proposed by [196], users only have access to the users around them and thus can only be pressured into adopting an innovation from their immediate neighbours.

The hypothesis for this research question is: *User adoption of language innovations can be predicted by exposure from friends or neighbours.*

Modelling

This question looks at the ability to predict user innovation adoption based on exposure from the neighbours of a user within their ego network. However, the attention that a user gives to each of their neighbours is not equal, with some users gaining more and some less. For this reason, the model must take into account this variation in influence in a user choosing to adopt a language innovation. This method is thus broken down into two steps: learning the influence between users, and predicting the adoption of language innovations based on exposure from their neighbours (this model is adapted from [83]).

Influence is mined through cascades of past actions: if a user constantly adopts the actions of another user a significant amount of times then one could say that this indicates that there is a strong level of influence, that influence only exists between neighbours in the network, and this has a value of 0 to 1.

The pressure that is then on a user to adopt an innovation is the collective influence from their neighbours who have already adopted the innovation.

Validation

Unlike the validation of research question 1.1.1 (section 4.5.1), this is a prediction task. Therefore, a quantitative evaluation can be made through the application of holdout validation. The dataset is split into training and testing sets (80 % and 20 %), based on the number of innovations; 80 % will be used in the training the model, with the remaining used to predict innovation adoption.

The result generated from the model is a threshold value that represents the pressure that has been applied to the user to adopt an innovation. The aim is then to learn a general threshold across the whole of the network that signifies when a user is going to adopt an innovation. This is achieved using Receiver operating characteristic (ROC) analysis, which represents the pay-off between true and false positive rates by varying the general threshold. Thus, the metric used in assessing the accuracy or the effectiveness of the model is Area under curve (AUC). A value of 1 would indicate that the model perfectly predicts if a user is going to adopt an innovation, and a value of less than 0.5 would indicate the model performs less well than a random baseline model.

4.5.3 Innovation Propagation

The diffusion of language is a process that happens over time, across a network of users and communities. However, to what extent does the structure of the network affect how language innovations diffuse? Can features within the network be used to predict the success of an innovation diffusion across a network? Drawing on previous work, [148] proposes that language spreads between communities and users due to a user's access to weak holes. In later work, [207] proposes that vital content within OSNs could be predicted from community features within the first n observations of a diffusion.

Thus, the hypothesis for this research question is: *The extent of diffusion of a language innovation can be predicted from the network and structural features of a network and diffusion.*

Modelling

The aim of this research question is to predict the extent to which a *language innovation* will diffuse across a network. This takes the form of predicting the number of unique users of an innovation and the number of usages of an innovation. However, this does not predict the final size, but rather predicts, at each stage of an innovation diffusion, whether the final size will be above or below the median of all diffusions that have at least n usages (n is the point at which the prediction is made).

The features extracted focus on the path that the diffusion takes across the network, such as the average constraint of the diffusion and the average degree of all users who have used the innovation.

These will then be used in conjunction with a logistic regression, aiming to predict if the diffusion will double.

Validation

Similar to section 4.5.2, quantitative validation will be used. This will take the form k-fold Cross Validation (CV) used to calculate the accuracy of the model on data which it has not already seen. This is achieved by separating the datasets into k equal randomly assigned groups, then holding one out while the model is trained. The held-out data is tested in the model. This is performed k times until each set has been held out, with the estimated accuracy of the model being the average across all held-out tests. As the prediction is binary, f-messure is used as the metric for performance, through the assessing both the precision and recall of the test set.

4.5.4 Summary

This section has introduced the three hypotheses, as well as the three methods and validations that are used, each of which aims to assess a different aspect of language change. The methods implemented across the three research questions come from the fields of data mining and machine learning; however, the validation that is performed on the results performs only part of the final analysis. As stated in section 4.3, the results of each of the three research questions will be fallible on their own; however, as a body of work and as a contribution to the field, the results will sit within the context of each other and the broader field. Thus, each research question will be tested through quantitative means; however, the conclusion of the thesis will be a qualitative analysis of the quantitative results in the context of the field.

4.6 Ethics

As with any research conducted using social agents, ethical considerations must be taken when developing the study. Research using data collected from online resources is no exception. Ultimately, the ethical impact of this work will be limited as the research is removed from the generation of the content that is to be analysis. However, this is not to say there are no ethical implications; one of the main ethical points is understanding the separation/merging of the public and private sphere for the user who generates the content for the platforms of Reddit and Twitter used in this research.

> Privacy is not about control over data nor is it a property of data. It's about a collective understanding of a social situation's boundaries and knowing how to operate within them. In other words, it's about having control over a situation. It's about understanding the audience and knowing how far information will flow. It's about trusting the people, the situati[on],

> and the context. People seek privacy so that they can make themselves vulnerable in order to gain something: personal support, knowledge, friendship, etc. [218]

The ethics of conducting research using data generated by users in online social network such as Twitter and Facebook have recently received growing attention, looking at the ethics surrounding internet research, mainly as people question the concept of informed consent of the user/content generator in using 'their' data. The debate around informed consent and publicly accessible data has led to a number of studies looking into the ethics of using online social media data from the perspective of the content generators [218].

The data collected from Twitter and Reddit is generated by human subjects. It is generated in specific locations (geographical or areas of a website) about a particular topic. Thus, it could be argued that there is a need for the consideration of the content generator at all times, and that consent should be gained from each user from whom data is mined [218]. However, in this research, the number of users contained in the dataset are in the millions, making this impossible to achieve [139]. [170] acknowledges that contacting each user would be an unmanageable task and that it is better, if there is no informed consent, that data should be displayed with depersonalised information, with a recommendation that systems should have filters in place to remove such information automatically.

Additionally, the debate around informed consent comes from the separation of the public and private sphere within OSN (such as Reddit and Twitter). The debate comes from the conflict between the perceived privacy of the content generator and the openness of the systems. To a researcher, the OSNs may be publicly viable; however, the members who generate the content believe the interactions exist within their private sphere, away from public observation and tampering. Thus, for this research, even though Twitter and Reddit are already within the public sphere, it may not mean that the data can be classed as 'public' [218].

> Individual and cultural definitions and expectations of privacy are ambiguous, contested, and changing. People may operate in public spaces but maintain strong perceptions or expectations of privacy. [139]

However, for simplicity, a distinct line can be drawn between the public and private sphere by a username and password being required. Thus, sites such as Facebook can be seen as existing in the private sphere; however, for this research, no username or passwords were required to access the data, and, for simplicity, we treat the data as publicly viewable. However, to respect the privacy of the users, no identifiable content will be published and the raw data will not be accessible after the research.

The main ethical considerations around social media research are the understanding of what is public and what is private, and also understanding the lack of informed consent for 'public data'. To address this, a number of steps will be taken.

- Data will only be collected from publicly accessible sources
- Data collected will not be released into the public domain
- Results that are reported will be aggregates and descriptive statistics, such as accuracy of classifiers
- When modelling social networks, all usernames will be stripped and replaced within a random ID

4.7 Technical Implementation

The data collected is 1 Tb+ in size; therefore, realistically, it is not possible to store and process the data on one machine, or write custom code to combine network storage and processing across a number of machines. For this reason, big data frameworks are used extensively throughout this work. This has a number of benefits: first, it lowers the barrier to achieving results, as we do not need to develop the entire analytic system; second, it allows for the research to be reproducible as we can take the developed code and run it into the same framework and dataset on a different cluster.

The two main framework users are Apache Spark and Apache Hadoop. These are two industry standard big data frameworks that are used for data ingestion and data processing, and can be extended to solve machine-learning and graph analytic tasks at scale.

Hadoop is an open-source Apache JAVA framework that allows for the distributed storage and processing of large datasets across a number of machines. The system is highly influence by Google's distributed file system [72] and Google Map Reduce [39].

Hadoop is split into two components: Hadoop Distributed File System (HDFS) and MapReduce. HDFS is the distributed file system that runs across the cluster, combining the available disk space across all machines into one consolidated block. Additionally, HDFS manages the replication of blocks of data across the different nodes, resulting in two benefits: first, if one node fails, there are then multiple copies of a block across the cluster; second, keeping multiple copies of data on different nodes allows for the optimisation of job placement on the cluster, by placing the processing on a different node rather than moving the data around.

The MapReduce portion of the framework deals with the distributed processing of data across the cluster. MapReduce is a programming model that is split into two stages: `Map()` and `Reduce()`. `Map()` applies the same operation to items of data within the dataset. This can be anything from transformation to the data, filtering items, or sorting data. The data is outputted in the formation of `(key, value)`. `Reduce()` is an aggregation function that combines the output of the maps that all have the same `key` value.

Spark is similar to Hadoop. However, even though Hadoop allows for the distribution of processing across a cluster, there are a number of limitations and shortcomings of the system. The main benefit is

Listing 4.1: Spark Word Count

```
JavaRDD<String> textFile = sc.textFile(``hdfs://...'');
JavaRDD<String> words = textFile.flatMap(new FlatMapFunction<String, String
    ↪ >() {
  public Iterable<String> call(String s) { return Arrays.asList(s.split(``
      ↪ '')); }
});
JavaPairRDD<String, Integer> pairs = words.mapToPair(new PairFunction<
    ↪ String, String, Integer>() {
  public Tuple2<String, Integer> call(String s) { return new Tuple2<String,
      ↪  Integer>(s, 1); }
});
JavaPairRDD<String, Integer> counts = pairs.reduceByKey(new Function2<
    ↪ Integer, Integer, Integer>() {
  public Integer call(Integer a, Integer b) { return a + b; }
});
counts.saveAsTextFile(``hdfs://...'');
```

the ease with which the user can create applications on top of Hadoop; a simple word count application can take 100+ lines of code, and more if we want to do more than one cycle of analysis; e.g. for multiple `map` and `reduce` stages, we would have to produce multiple Hadoop jobs and manually chain them together, thus increasing the completion time for the user and increasing chances of errors within the system.

Spark is API-centred, and relies on datasets being stored in memory while processing (instead of being read from and written to disk). These make applications easier to write and quicker to execute, allowing for a more iterative data science and exploration process. Additionally, we are not constrained to just `map` and `reduce` with multiple functions such as `filter` and `flatmap`.

Applications such as Apache Mahout[17] and Apache Giraffe[18] have been developed on top of Hadoop to bring scalable machine learning to the framework, but these are developed externally to Hadoop and can be complex to set up. However, Spark has embraced the expandability with the inclusion of native libraries such as MLLib (for machine learning) and GraphX (for graph processing).

Across this thesis, a combination of the two systems will be used when processing datasets as each has its own benefits. Spark is used for the integration and distribution of existing Java applications (such as deploying OpenNLP across a cluster), and Hadoop is used for the bulk processing of large datasets. Hadoop and Spark will be used in conjunction with additional systems such as HBase and Hive for the storage and querying of intermediate results, and for the execution of SQL-style queries on Hadoop.

[17] https://mahout.apache.org
[18] https://giraph.apache.org

4.8 Summary

This chapter has outlined the broad methodology applied to understand, predict and model language change in online social networks.

From a theoretical perspective (section 4.2), this work will follow a *post positive* epistemological stance. This means that the work aims to triangulate the results within a wider research context, through the development of prediction and ranking models. Even though this means that no one section is conclusive about language change, the collective results will allow for a quantitative assessment across the collective qualitative results.

To model language change, Twitter and Reddit have been chosen as data sources, due to their prominence in online culture and ease of access. From a practical perspective, the research is framed as three independent methods, which allows custom methods to be developed for each of the three questions.

The first method focuses on the detection of language changes by implementing known lexicographical models. The results are quantified by apparent validation through the experience of the researcher. The second research method focuses on the development of general threshold models to predict user addition of language based on exposure, verified through hold-out validation. The final research method utilises network and diffusion features with linear regression and binary classification to predict the extent to which an innovation will diffuse. The following three chapters (6, 7, 8) formally develop the methods further, as well as expanding on their implementations and the results from the methods.

As identified in section 4.3, there are commonalities in the pre-processing sections, such as network generation and innovation identifications. These are discussed in detail in section 5, and also in each of the methods chapters.

Chapter 5

Data Prepossessing and Social Networks

5.1 Introduction

As introduced through the shared framework and methodology in Chapter 4, there are aspects of the methods that are common across each of the three research questions in this thesis. Thus, the following chapter details how the social network and its user interactions are extracted from the mined data (Section 5.2). In addition, we define how each dataset can express both micro and macro user interactions. Also, as this thesis looks at language change, we must identify what we classify as a *language innovation*, which is explained in Section 5.3.

5.2 Social Networks

Language change and speaker innovation propagation exist within the context of social networks, whether professional networks in the workplace or social networks in the context of personal lives. However, in this research, the focus is on the OSNs of Twitter and Reddit as these allow for the permanent recording of interpersonal relationships and communication.

Reddit and Twitter are both formed around users connecting and communicating with each other, expressing social relationships, and allowing content to diffuse. However, as identified by [148], social networks not only express the interactions between users but also the interactions of communities and users, and between communities and other communities. The definition of what constitutes a community varies, from the classical structural definitions based on clustering coefficients, to defining communities based on users' commonality in geographical locations.

As shown in [129] and [148], language diffuses not only between users but also between communities.

	Reddit	Twitter
User Interaction - (*micro*)	Users interact by commenting on each other's posts forming trees of comments in each subreddit, thus the relationship between users is defined by their interactions through commenting on each other's posts	Users can mention each other through the inclusion of each other's usernames in a tweet, thus the relationship is defined by their interactions through mentions.
Community Interaction - (*macro*)	Users post comments within subreddits, but users move between subreddits as their tastes and interests change over time, thus this network represents the relationship between subreddits as a function of user movement.	Users generate tweets in towns and cities across the UK, but users also move between locations, thus taking language with them. This network represents the relationship between locations in the UK as a function of users moving between them.

Table 5.1: Network definition summary

Thus, this research models the networks that users exist within through two abstractions: *micro* and *macro*. *Micro* represents the interpersonal interactions allowing for the expression of relationships between users, whereas the *macro* network representation signifies the relationships between the places or communities where users have interacted and generated content, such as a geographical location or an area of a website.

The methods used to generate these two network abstractions across the two datasets varies. The remainder of this section introduces the basic graph notation used through this thesis (section 5.2.1), as well as detailing how *micro* (section 5.2.2) and *macro* (section 5.2.3) graphs are extracted from the Twitter and Reddit datasets.

5.2.1 Graph Notation

A graph G contains a set of vertices V and a set of edges E, where $E = \{e_{i,j} | i, j \in V\}$. Graph ($G$) can either be directed or undirected; for an undirected graph, the edges are identical in either direction, $e_{i,j} = e_{j,i}$, whereas, in a directed graph, the edges are not equal, $e_{i,j} \neq e_{j,i}$, with edge $e_{i,j}$ representing an ordered tuple identifying the relationship $i \rightarrow j$.

The number of edges and nodes in the graph is represented as $|E|$ and $|V|$, respectively, with the degree of vertex v accessed through function $d(v)$. A directed network in-degree is represented as $d_{in}(v)$ and out-degree $d_{out}(v)$, where $d(v) = d_{in}(v) + d_{out}(v)$. $N(v)$ represents the neighbours of node v in an undirected graph, whereas $N_{in}(v)$ and $N_{out}(v)$ returns the set at the end of the incoming and outgoing edges.

In addition to the basic features of the graph (edges E and vertices V), there are two additional features per edge: edge weight ($w_{i,j} \in W$), and edge creation time ($\tau_{i,j} \in T$). The weight of an edge ($w_{i,j} \in W$) is a non-negative numeric value, which represents the strength of the relationship between nodes i and j. In addition to weight, each edge is given a time stamp ($\tau_{i,j} \in T$) that represents the time

Table 5.2: Network description

Network	Nodes	Edges	Power Law	Communities
Twitter Geo	2,910	436,849	3.398	14
Twitter Mention	283,755	329,440	3.004	39,767
Reddit Comment	861,955	2,402,202	4.134	36,885
Reddit Subreddit	15,457	142,285	1.511	407

at which the edge between the two nodes was created.

Additionally, a graph G can be represented in an adjacency matrix A. This is a $n \times n$ matrix, which represents the structure of a graph, with the nodes within the network being the column and row indices. The relationships between nodes is represented by the values of each cell, such that, if an edge exits from i to j, then a 1 is placed at the corresponding cell of row i and column j ($a_{i,j} = 1$). For a directed graph $a_{i,j} \neq a_{j,i}$; however, for an undirected graph, they are the same $a_{i,j} = a_{j,i}$.

5.2.2 User Interaction

Language is defined by the users that use it, with the speaker innovations diffusing between people that come into contact with each other. These points of contact can be defined by the interactions between users within Twitter and Reddit.

Thus, we define the social networks as a function of interaction between users. These user interactions can be represented in a directed graph G with the vertices V representing the users and the edges E representing the relationships.

Twitter

Twitter, as stated previously (Section 4.4.1), allows users to 'follow' each other, filling their timelines with tweets from the people they follow. Researchers have used the reciprocal follower relationship used to indicate a 'relationship' between users. However, mining follower relationships at scale is challenging due to the data being protected behind rate-limited API. However, direct user interactions can be mined by identifying when users mention each other in their tweets.

In tweets, users can 'mention' each other though the addition of an '@username', which brings the tweet to the attention of the mentioned user. Users can then reply to these tweets by including the original username, thus causing a chain of messages or conversations between sets of users. Given the challenge of extracting the follower relationships, the relationships between users are defined as a function of the interaction that is recorded through users 'mentioning' each other.

Whereas the follower/followee relationship will indicate an explicitly defined relation, using the mention allows a strength to be given to the relationship as a function of the users' interactions. This means that we can distinguish between active (high user interaction) and passive relationships (no user

interaction).

The features of the graph are defined as follows:

G_{tm} - The directed graph represents the mention relationship between users

V - Each user in the Twitter dataset

E - $e_{i,j}$ represents if i has mentioned j

W - The number of times i has mentioned j

T - The time that i first mentioned j

From here on, we talk about the graph G_{tm} in the form of an adjacency matrix A_{tm}. The relationship between users ($a_{i,j}$) can be expressed in an adjacency matrix (A_{tm}) as:

$$a_{i,j} = \begin{cases} 1 \text{ if } |mentions(i,j)| > 0 \\ \text{else } 0 \end{cases} \tag{5.2.1}$$

The function $mention(i,j)$ returns an ordered list of each time user i has mentioned j; e.g. $\left\langle \tau_0^{i,j}, \tau_1^{i,j} ... \tau_n^{i,j} \right\rangle$. However, a user can be mentioned who is not in the datasets; thus, for a relationship to exist, both users (i and j) must be in the set of users ($i \& j \in V$).

The weight of an edge is defined as the number of times that user i has mentioned user j, such that:

$$w_{i,j} = |mentions(i,j)| \tag{5.2.2}$$

In addition to edge weights (W), the creation time of the edge is defined as $\tau_{i,j}$; this is the time at which user j was first mentioned by i, such that:

$$\tau_{i,j} = \min(mentions(i,j)) \tag{5.2.3}$$

Reddit

As with Twitter, the explicitly defined relationship between users is not readily available from the dataset. Thus, as with Twitter, the relationship between users is defined as a function of two users' interactions. Users interact (similar to Twitter) with each other by commenting on each other's posts, creating chains of conversations that can contain multiple users. However, a message can have multiple responses (child comments), causing the thread to take the form of a tree as the conversation branches, which can formally be described as a directed acyclic graph.

Graph G_{rc} represents the Reddit user interactions graph, in which each user is represented within the set of vertices V. For user i (such that $i \in V$), the outgoing edges represent the users (i) whose posts

they have commented on ($i \rightarrow j$ reads that i has commented on j). Alternatively, user i's incoming edges are from the users who have commented on i's post ($j \rightarrow i$, reads j has commented on i).

The features of the graph are defined as follows:

G_{rc} - The directed graph represents the mention relationship between users

V - Each user in the Reddit dataset

E - $e_{i,j}$ represents if i has mentioned j

W - The number of times that i has mentioned j

T - The time that i first mentioned j

The network thus can be defined through an adjacency matrix (A_{rc}):

$$a_{i,j} = \begin{cases} 1 \text{ if } commented\_on(i,j) \\ \text{else } 0 \end{cases} \tag{5.2.4}$$

The function $commented\_on(i,j)$ returns the list of times that i has commented on j, such as $\left\langle \tau_0^{i,j}, \tau_1^{i,j} ... \tau_n^{i,j} \right\rangle$.

Alternatively, the set of outgoing neighbours can be defined as:

$$N_{out}(i) = \{owner(parent(z)) : z \in posts(i)\} \tag{5.2.5}$$

Where $posts(i)$ returns the set of user i's posts, and $parent(z)$ returns the set of parent posts of z, with the author of the parent posts being accessed through $owner(z)$.

The incoming edges can subsequently be defined as the set of users who created child posts on comments that i created:

$$N_{in}(i) = \{owner(child(z)) : z \in posts(i)\} \tag{5.2.6}$$

Where $child(z)$ returns the child post of z.

As with the Twitter network, the weight of an edge $w_{i,j}$ would be the number of times that i has commented on posts of j:

$$w_{i,j} = |commented\_on(i,j)| \tag{5.2.7}$$

With the time of the edge $\tau_{i,j}$ being the time of the first interaction.

$$\tau_{i,j} = \min(commented\_on(i,j)) \tag{5.2.8}$$

However, the direction of information flow between users is in the opposite direction of the relationship; thus, with a user commenting on a post, the information is flowing from the post to the user commenting.

Implementation

As stated throughout this work, one of the challenges is the ability to process large quantities of raw data, with Twitter containing 111,067,539 tweets and Reddit 1,054,976,755 posts. The size of the data thus limits the use of traditional systems, such as relational databases or systems such as `R` on one machine. Therefore, the majority of preprocessing is performed across Hadoop, allowing for scalable execution of code. Thus, all systems used to implement the networks use tools such as Apache Hive and Map Reduce, which can scale processing using familiar query languages.

Apache Hive[1] is a data-warehousing system that is built on top of Apache. Hadoop[2] allows for structured querying of large datasets using Hive Query Language (HQL) (a variant of SQL), which is then translated into Hadoop Map Reduce jobs. Additionally, HQL can be used to query collections of JSON documents, thus can used to query raw tweets and Reddit posts.

Twitter First, the complete set of tweets is loaded into Hive's meta table store.

```
CREATE TABLE tweets ( json string );
LOAD DATA INPATH '/twitter/Twitter-Combined' INTO TABLE tweets;
```

This loads all tweets into one table called `tweets`, where each row is a single JSON object representing an individual tweet. To then generate the Twitter mention graph (G_{tm}), we must first create a list of all users within the dataset (users who have generated the tweets); these will represent the nodes within the network (V).

```
SELECT distinct get_json_object(tweets.json, '$.user.id') as id
 FROM tweets
```

The `get_json_object(tweets.json, '$.user.id')` allows Hive to query the JSON object through the use of JSONPath[3] expressions. The query `$.user.id` returns the username of the user who created each tweet.

In addition, we need to extract which users' tweets have used mentions; this is different from extracting the username as a tweet can mention more the one user. JSONPath allows for the extraction of arrays.

[1] https://hive.apache.org/
[2] https://hadoop.apache.org/
[3] JSONPath allows for the transformation of JSON objects in the same way the XPath can be used on XML

```
SELECT get_json_object(tweets.json, '$.user.id') as source, ms as target,
    ↪ CAST(UNIX_TIMESTAMP(get_json_object(tweets.json, '$.created_at'),'EEE
    ↪  MMM dd HH:mm:ss z yyyy')*1000 as timestamp) as created_at
from twitter.tweets
LATERAL VIEW explode(split(regexp_replace(get_json_object(tweets.json, '$.
    ↪ entities.user_mentions[*].id'), '\\[|\\]', ' '),',')) x AS ms
```

First `get_json_object(tweets.json, '$.entities.user_mentions[*].id')` extracts an array of usernames that are mentioned in a tweet. However, this returns an array of usernames, or a `string`. This is then split using `regex` into an array structure that HQL can process. At this point, the `explode` function then creates virtual rows, each containing a value from the array, along with the original JSON, thus the username of the author of the tweet and the username of the user they have mentioned. However, usernames need to filter out those who have been mentioned who are not users within the dataset (they have not created any tweets in the dataset). This is achieved through a `LEFT SEMI JOIN`, which is a join that only returns the records from the left table, thus removing all users who are mentioned but who are not in the dataset.

Listing 5.1: Twitter mention graph HQL query

```
CREATE TABLE twitter.mentions_5 as
SELECT edges.source as source, edges.target as target, min(edges.created_at
    ↪ ) as created_at, count(1) as weight
FROM(
 select get_json_object(tweets.json, '$.user.id') as source, ms as target,
    ↪ CAST(UNIX_TIMESTAMP(get_json_object(tweets.json, '$.created_at'),'EEE
    ↪ MMM dd HH:mm:ss z yyyy')*1000 as timestamp) as created_at
 from twitter.tweets
 LATERAL VIEW explode(split(regexp_replace(get_json_object(tweets.json, '$.
    ↪ entities.user_mentions[*].id'), "\\[|\\]", ""),',')) x AS ms
 ) as edges
 LEFT SEMI JOIN (
   SELECT distinct get_json_object(tweets.json, '$.user.id') as id
   FROM twitter.tweets
   ) AS users ON users.id = edges.target
 GROUP BY edges.source, edges.target
```

Finally, to count the number of times a user has mentioned a fellow user, along with the first time a mention happens, a `groupby` is used across the username and target user. Thus, the weight of the

relationship is the size of the group, and the creation time is the minimum date-time.

Reddit A similar method was taken to construct the Reddit Comment graph (G_{rc}). The Reddit data was loaded into a one-column Hive table, with each line containing the entire JSON object representing a post.

Listing 5.2: Load Reddit into HIVE

```
CREATE TABLE reddit ( json string );
LOAD DATA INPATH '/reddit/Reddit-Data' INTO TABLE reddit;
```

Unlike Twitter mentions, in which each tweet could mention many user, Reddit posts feature a one-to-one mapping of a child to a parent post (though this would be a one-to-many relation when treated as the parent, as there could be many child posts). This comes from the chain nature of conversations within Reddit, where the relationship that is being extracted is the relationship between the author of the child and the author of the parent post.

First, a temporary table is created containing the `user-name`: `post id` and parent `post id`;

```
SELECT get_json_object(json_table.json, '$.author') as author,
    ↪ get_json_object(json_table.json, '$.subreddit') as subreddit,
    ↪ get_json_object(json_table.json, '$.name') as post, get_json_object
    ↪ (json_table.json, '$.parent_id') as parent, CAST(cast(
    ↪ get_json_object(json_table.json, '$.created_utc') as int) * 1.0 as
    ↪ TIMESTAMP) as time
    FROM json_table
    WHERE get_json_object(json_table.json, '$.author') != ''[deleted]''
```

Performing a self-join based on the `post id` and the `parent post id` results in consecutive posts being joined, representing the relationship between the authors.

Listing 5.3: Reddit comment graph HQL query

```
CREATE TABLE user_reddit_comment_network AS
SELECT target.author as source, source.parent as target, min(target.author)
    ↪ as created_at, count(1) as weight
FROM (
    SELECT get_json_object(json_table.json, '$.author') as author,
        ↪ get_json_object(json_table.json, '$."subreddit"') as subreddit,
        ↪ get_json_object(json_table.json, '$.name') as post,
        ↪ get_json_object(json_table.json, '$.parent_id') as parent, CAST(
```

```
        ↪ cast(get_json_object(json_table.json, '$.created_utc') as int) *
        ↪ 1.0 as TIMESTAMP) as time
    FROM json_table
    WHERE get_json_object(json_table.json, '$.author') != ''[deleted]''
    ) AS source
JOIN (
    SELECT get_json_object(json_table.json, '$.author') as author,
        ↪ get_json_object(json_table.json, '$."subreddit"') as subreddit,
        ↪ get_json_object(json_table.json, '$.name') as post ,
        ↪ get_json_object(json_table.json, '$.parent_id') as parent, CAST(
        ↪ cast(get_json_object(json_table.json, '$.created_utc') as int) *
        ↪ 1.0 as TIMESTAMP) as time
    FROM json_table
    WHERE get_json_object(json_table.json, '$.author') != ''[deleted]''
    ) AS target
    ON source.parent = target.author
GROUP BY source.parent, target.author
```

Again, by using a `grouby` method, we can compute the weight of the edge and the time of the edge by counting the size of the group and the minimum time of the edges.

5.2.3 Community Interaction

The second set of networks aims to model the interaction between locations/communities within each network by mining the paths that users take in navigating predefined locations within the network. The idea of location or communities in either datasets varies; for Twitter, it can be thought of as the user moving around the geographical network, with tweets being assigned locations within the network through reverse engineering their respective GPS coordinates. Whereas with Reddit, each post is submitted to a subreddit; therefore, instead of being a physical location, the network models the interactions of subreddits in the network.

To extract these networks, a general framework is applied to each dataset, aiming to model the interactions between these communities and locations. People, as shown by [152], travel around locations in common short hops interspersed with random large hops, with the shorter hops coming from user movement around places in a local vicinity (travelling to and from work), and the random long hops being holidays or seeing family. Thus, to quantify the interaction between communities, we first asses the interaction of each user with given communities and locations.

As with the graphs defined in section 5.2.2, a graph is defined as $G = (V, E, W, T)$. The locations that a user can visit are defined as the vertices ($v \in V$), with the graph additionally defined as an adjacency matrices A with the dimensions $V \times V$. The aim is to capture the user's transaction between locations, i.e. where a user moves consecutively.

$$a^u_{i,j} = \begin{cases} 1 \text{ if } |moves(u,i,j)| \rangle 0 \\ \text{else } 0 \end{cases} \tag{5.2.9}$$

Where the function $moves(u, i, j)$ returns a list of order times that user u moves from community $i \rightarrow j$. The weight ($w^u_{i,j} \in W^u$) of the edge is then defined as the number of times a user has moved between the location ($|moves(u, i, j)|$), and the time of each edge ($\tau^u_{i,j} \in T^u$) is the first time the user u has transitioned from $i \rightarrow j$, $\min(moves(u, i, j))$.

However, the method described only computes the matrices for each user (A^u); to compute the whole graph for the whole population, we sum across all the matrices.

$$A = \sum_{u \in U} A^u \tag{5.2.10}$$

For the weights matrix (W), again, it is the addition across all user matrices:

$$W = \sum_{u \in U} W^u \tag{5.2.11}$$

However, for the time matrix (T), the value for the first traversal will be the minimum across all users traversing between the two locations:

$$t_{i,j} = min(\{t_{i,j} : T^u, u \in U\}) \tag{5.2.12}$$

This is a generic method, which first computes an individual user's sequential interactions with the community structure of the network. However, this varies from a number of methods used in the past that have been based on a fully connected graph of past interactions with communities [65]. However, accumulating all interactions into one fully connected graph (between all communities) presumes that users have the chance of transitioning from any location to the next. This then means that vital information is lost, indicating a change in user habits, or exposure from other communities.

The following sections now describe in greater detail how the communities (vertices) are defined for both Twitter and Reddit datasets, as well as how the user interactions within these communities are defined and extracted.

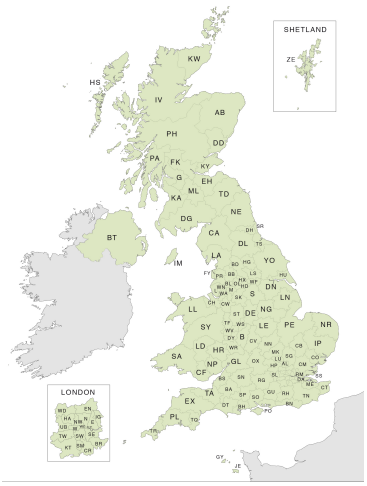


Figure 5.2.1: The UK postcode map shows the boundaries of each of the low-level postcodes within the UK; the colour also represents regional groupings of postcodes such as Scotland and Northern Ireland

Twitter

Data collected from Twitter had to originate from within the boundaries of a predefined bounding box (section 4.4.3) of the UK. This means that each tweet will contain a set of GPS coordinates, representing where in the UK each Tweet was created. Thus, the aim is to develop a network of communities with the aim of modelling the user interaction within the landscape of the UK.

The UK can be split into hierarchical regions (zones) using the UK postcode system. This hierarchical taxonomy breaks the UK into distinct regions, such as the North West, but additionally into small postal regional regions, e.g. LA, each nested within each other, e.g. North West $\rightarrow$ LA $\rightarrow$ LA1 (see figure 5.2.1). One aim of this thesis is to assess language and geography; clustering tweets into postcode location rather than using the GPS coordinates allows for a simple and hierarchical abstraction of the UK geographical landscape that is representative of population density and regional structures.

Each low-level postcode (e.g. LA1) can be represented by a centroid (which represents the central point in the postcode area), with each tweet then assigned to a postcode from which it has the shortest distance. This is not the most accurate method, as not all centroids are equality spaced; thus a tweet could be at the edge of a very large post code, but be closer to the centroid of another. However, due to copyright restrictions, the boundaries of postcodes are not publicly available [4], so using the centroid is the best alternative.

When computed using the method outline above, for the network G_{tp}, each node ($v \in V$) is a centroid in a postcode p ($p \in P$), with an edge ($e_{i,j}$) representing the number of user traversals from postcode i to postcode j. This can be represented as an adjacency matrix A_{tp}, with the cell $a_{i,j}$ representing the edge from i to j ($e_{i,j}$). The value of $w_{i,j}$ represents the number of times a user has moved from i to j.

The network is defined as follows:

[4] https://www.theguardian.com/technology/2009/sep/23/post-office-database-copyright-leak

G_{tp} - The directed graph represents user movement between postcodes

V - The set of postcodes within the UK

E - $e_{i,j}$ represents if users have moved from postcode i to j

W - The number of times users have moved from i to j

T - The time that the first user moved from i to j

Reddit

Community structure in Reddit, compared to Twitter, is easier to infer. This is due to the community (subreddits) being explicitly defined rather than implicitly inferred. Therefore, the network can be generated by mining user traversals across subreddits, thus treating a subreddit in much the same way as a postcode. Therefore, network G_{sr} consists of a set of nodes V such that $v \in V$, where v is a subreddit within Reddit. The adjacency A and weight W matrix represent the users that have sequentially moved between the two Reddit, and how many times. As before, the adjacency (A) and weight (W) matrix is the summation of individual user adjacency (A_u) and weight (W_u) matrices, with the time matrices as the global minimum.

The network is defined as follows:

G_{rs} - The directed graph represents user movement between subreddits

V - The set of subreddits within Reddit

E - $e_{i,j}$ represents if users have moved from subreddit i to j

W - The number of times users have moved from i to j

T - The time that the first user moved from i to j

Implementation

As when implementing the extraction of micro networks, Hive and HQL will be used again. However, a number of additional stages have to be applied in assigning the initial community tag (translating GPS to postcode) to each tweet before extracting the network.

Twitter, unlike previous network construction challenges, generating G_{pc} cannot be achieved in one HQL job. This is due to each tweet not containing the postcode in which it originated, but rather a set of GPS coordinates indicating the location from which it was tweeted. Therefore, to identify the postcode from which the tweet originated, we must translate the GPS coordinates to a given postcode.

As mentioned earlier, there are issues in gaining access to the coordinates of the boundaries of each postcodes, though we can gain easy access to the centroid (central point) of each postcode.

Therefore, given a set of known centroids (longitude and latitude) of postcodes, a number of strategies could have been used to identify the nearest to a given tweet. These could have included comparing the distance between each centroid and a tweet to find the closest; however, as before, this would have been costly to perform with a computational complexity of $o(N)$. Instead, the centroids of each postcode are loaded into a KD-tree [119], which is then queried to identify the nearest centroid to a set of GPS coordinates. A KD-tree is a binary search tree implemented in k-dimensions; it allows for the storage and querying of continuous data types (not strings). Its computation complexity of a look-up is $O(\log N)$, storage $O(n)$ and search $O(\log n)$. A KD-tree is constructed by recursively splitting the sequence of data points at the median across one dimension, with the following level being split on the next dimension. This causes the k-dimensional space to be split into multiple regions.

To identify the nearest postcode to a GPS point, a nearest neighbour search is used; the tree is recursively searched using a left/bottom technique computing the distance between the node and the query point. Each sub-tree of a node is searched, though one side is disregarded as it will be further away from the node due to data being split along the median. One modification, however, has to be made to the algorithm to deal with the type of points that the structure is using. The surface of the earth is not flat but curved, so using the simple `cosine` distance is not appropriate; therefore, to calculate the distance within the KD-tree, the Haversine distance is used:

$$d = 2r \arcsin\left(\sqrt{\text{hav}(\varphi_2 - \varphi_1) + \cos(\varphi_1)\cos(\varphi_2)\,\text{hav}(\lambda_2 - \lambda_1)}\right) \tag{5.2.13}$$

φ_2 and φ_1 are the latitudes of points 1 and 2 in the radian, with φ_2 and φ_1 representing the longitude in the radians. r is the radius of the sphere; thus, for earth, it is $6{,}371 * 10^3$ m.

One challenge of using KD-trees is that the majority of implementations are designed to run on one machine, which is not scalable for this quantity of data. To the pattern above in Spark, a solution would be to learn the KD-tree to each mapper in either a Spark or a Hadoop job; however, this would mean learning the KD-Tree N times, thus reducing the speed of the distributed applications.

Apache Spark, however, has a number of features that can be used to minimise the overheads of learning the KD-tree by being able to share the model across mappers. To share models in Spark, we can use the `broadcast` function available within the framework; this takes a model (implemented in JAVA), serialises it, and then copies it into the memory of the other executors within the cluster (Figure 5.2.2 visually represents the `broadcast`). There are limitations to this as, when it has been broadcast, the data within the object cannot be updated and exists only in read mode. Additionally, larger objects may slow the execution of an application as the JVM attempts to more aggressively manage the memory.

In tagging each tweet with a postcode, the output for the Spark job is then a number of CSV files containing the ID of the tweet and the associated postcode. These CSV files are then loaded into Hive, the same as the previous networks.

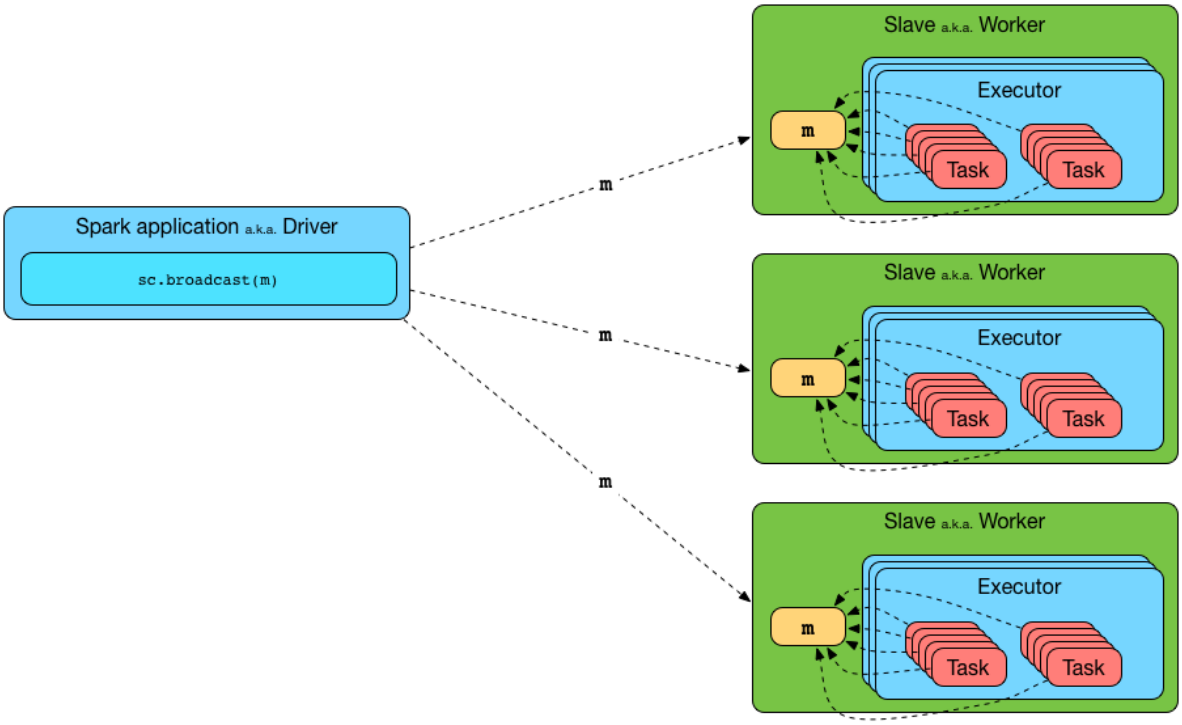


Figure 5.2.2: Spark broadcast allows for variables to broadcast across all executors in the cluster, which is achieve through serialisation. The KD-tree is thus learnt on the master nodes and then broadcast across the cluster within one copy for each executor, to which the tasks then have access.

```
CREATE TABLE twitter_postcode ( tweet_id string, username string, postcode
    ↪ string );
LOAD DATA INPATH '/twitter/Twitter-Postcode-Data' INTO TABLE
    ↪ twitter_postcode;
```

As introduced in section 5.2.3, the aim is to first connect consecutive user movements together, then sum across all users as the global transaction matrix. The order of tweets can be inferred from the tweet ID assigned by Twitter as these are generated consecutively across all tweets.[5]

We must assign the rank in which each user generated a tweet; this is achieved through the application of the `rank()` over a list of user tweets, which is achieved by partitioning the data by the username:

```
SELECT Rank() over (Partition by t.username Order By t.id)
```

Then, to connect consecutive tweets, a self-join is made, where the join criteria is based on the username and the current rank minus one.

Listing 5.4: Twitter geo network HQL implementation

```
SELECT source.postcode AS SOURCE,
```

[5] https://dev.twitter.com/overview/api/twitter-ids-json-and-snowflake

```
        target.postcode AS taget,
        count(1) AS weight FROM
    (SELECT Rank() over (Partition BY t.username
                         ORDER BY tweet_id) AS rank, t.username, t.tweet_id,
                             ↪ t.postcode) AS SOURCE JOIN
    (SELECT Rank() over (Partition BY t.username
                         ORDER BY tweet_id) - 1 AS rank, t.username, t.
                             ↪ tweet_id, t.postcode) AS target ON source.rank
                             ↪  = target.rank
GROUP BY SOURCE,
          target
```

Through the application of a `group by`, the collective number of times that users move from one postcode to the next is counted. In this section, there has been the notable exclusion of time; this is because the presumed iterations between locations have not started during the data collection, but are rather historic and existed before the data collection started.

As with Twitter, generating the macro Reddit network aims to represent the user movement across subreddits within Reddit. A similar method as used for Twitter is implemented. However, as each post contains the subreddit in which it originated, the location is already assigned, so we can use the JSON from the beginning.

Initially, a temporary table is generated that contains the authors of a post, the sub-reddit, and the time that the post was generated. Again, through the applications `Rank()` across user posts ordered by time, we can infer the sequence of movements across subreddits:

```
SELECT Rank() over (Partition by rp.author Order by rp.time) as rank, rp.
    ↪ author, rp.subreddit, rp.time
      FROM (
         select get_json_object(json_table.json, '$.author') as author,
             ↪ get_json_object(json_table.json, '$.subreddit') as subreddit,
             ↪  CAST(cast(get_json_object(json_table.json, '$.created_utc')
             ↪ as int) * 1.0 as TIMESTAMP) as time
from json_table
WHERE get_json_object(json_table.json, '$.author') != ''[deleted]''
```

Finally, a self-join in performed, with the right table's rank being subtracted by 1. Then, through the use of a `group by` on the source and the target subreddit, the number of user traversals is computed, and the first time that this movement happened is the $min(time)$ within the group.

Listing 5.5: Reddit Subreddit Graph HQL query

```
CREATE TABLE user_traversal_network_2014 AS
SELECT utr.subreddit as source, utr1.subreddit as target, count(1) as
    ↪ weight, min(utr.time) as created_at
FROM (SELECT Rank() over (Partition by rp.author Order by rp.time) as rank,
    ↪  rp.author, rp.subreddit, rp.time
      FROM (
        select get_json_object(json_table.json, '$.author') as author,
            ↪ get_json_object(json_table.json, '$.subreddit') as subreddit,
            ↪  CAST(cast(get_json_object(json_table.json, '$.created_utc')
            ↪ as int) * 1.0 as TIMESTAMP) as time
 from json_table
WHERE get_json_object(json_table.json, '$.author') != "[deleted]"
        ) as rp
      ) AS utr
JOIN (
  SELECT r.rank-1 as rank, r.author, r.subreddit, r.time
  FROM
   (SELECT Rank() over (Partition by rp.author Order by rp.time) as rank, rp
       ↪ .author, rp.subreddit, rp.time
      FROM (
        select get_json_object(json_table.json, '$.author') as author,
            ↪ get_json_object(json_table.json, '$.subreddit') as subreddit,
            ↪ CAST(cast(get_json_object(json_table.json, '$.created_utc') as
            ↪  int) * 1.0 as TIMESTAMP) as time
 from json_table
WHERE get_json_object(json_table.json, '$.author') != "[deleted]"
        ) as rp
      ) as r
  ) as utr1 ON utr.rank = utr1.rank AND utr.author = utr1.author
 GROUP BY utr.subreddit, utr1.subreddit
```

5.2.4 Backbone Extraction

When extracting networks from social data, we can end up with a graph that has limited representation of the network; that is, the extracted networks are highly connected as they track every interaction. However, not all edges have the same significance in representing the network, such as node v having all edges with a weight of 100, except for one, which has a value of 1. We can then say that the edge with a weight of 1 is less significant than the others, and we can then believe that the less significant edges are less important to the nodes and the nodes' communication pattern.

Thus, to reduce the number of edges but retain the structure of the network, statistically insignificant edges in the network are removed, based on their edge weights. This is achieved through the application of the backbone extraction algorithm [174]. This method preserves the core structure of the network by removing the statistically insignificant edges of the network. This is achieved by comparing given the edge weights to a `null` model that believes that edge weights are uniformly distributed at random. The edge significance ($\alpha_{i,j}$) for an undirected network is computed as follows [156]:

$$\alpha_{i,j} = (1 - \frac{w_{i,j}}{\sum_{v \in N(i)} w_{i,v}})^{|N(i)|-1} \tag{5.2.14}$$

The function $N_{out}(i)$ returns all the neighbours of i, with $w_{i,j}$ representing the raw weight of edge ($e_{i,j}$). However, as we can see, this significance is computed on the outwards edges of a network, though it is not as simple for undirected networks where an edge has 2 significance scores. The edge is then removed by taking into account either the average ($\frac{\alpha_{i,j}+\alpha_{j,i}}{2}$) of the scores.

For this work, edges are classified as significant if $\alpha_{i,j} > 0.05$. [156] showed that this is the optimal value for edge pruning in social networks, allowing the network to maintain the core small network characteristics, though with the sparser number of edges making it easier to analyse.

Implementation

Edge significance can be computed in a simple HQL query, allowing for the use of big data technology. For this, we presume the table representing the directed network consists of three columns: `source`, `target` and `weight`:

```
SELECT m.source as source, m.target as target, m.weight as weight, (m.
    ↪ weight / c.s) as norm_weight, power((1-(m.weight / c.s)),(c.cou-1))
    ↪ as alpha, c.cou as total_edge_count, c.s as total_edge_sum_weight, m.
    ↪ created_at as created_at
FROM network as m
JOIN (
  SELECT source as source, count(target) as cou, sum(weight) as s
```

```
    FROM network
    WHERE source != target
    GROUP BY source
    ) AS c ON m.source = c.source
WHERE m.source != m.target
```

This works by first computing a temporary table that represents the total edge weight of a node (`group by` on node). This is then joined to each of the nodes' edges (rows), with the resulting significance score computed within in-built functions.

5.2.5 Community Detection

Social networks are characterised by short average path lengths and high clustering coefficients, ultimately representing the small world effect [199]. This form of network structure is heavily influenced by homophily, where individuals connect to users who share similar interests or bonds. The combination of these two effects results in dense areas within the network that can be identified as communities of nodes, all sharing common interests and connections. More formally, *communities* are defined as regions in a network with an increased node density, which can be measured by clustering coefficients.

Community structure can be seen in each of the four networks, through the uneven degree distribution, suggesting the existence of community structure (Table 5.2). However, the communities within each of the extracted networks represent different interaction of nodes; for the user-based networks (Twitter mentions and Reddit comments), it represents collections of dense inter-user interactions, whereas for the macro network, clusters/communities represent collective user interactions with wider geographical locations or subreddits.

We aim to explicitly define to which community within a network a node belongs. However, in real life, one user may belong to many different communities; therefore, defining network membership through topological and content-based features can be challenging. For this thesis, a nodes community membership was detected though the application of the *Lounvain modularity* method [22].

The Lounvain method aims to extract communities from the network structure in a two-phase process that aims to maximise the network's modularity across the edges internal and external to each community [151]. This is implemented using a *greedy* recursive algorithm approach split down into two phases; the initial phase maximises the modularity of the community, and the second generates a network of communities, followed by maximising the modularity between clusters of nodes.

Modularity of the network is assessed as follows:

$$Q = \frac{1}{2m} \sum_{i,j} \left[w_{i,j} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \text{where } m = \frac{1}{2} \sum_i k_i \tag{5.2.15}$$

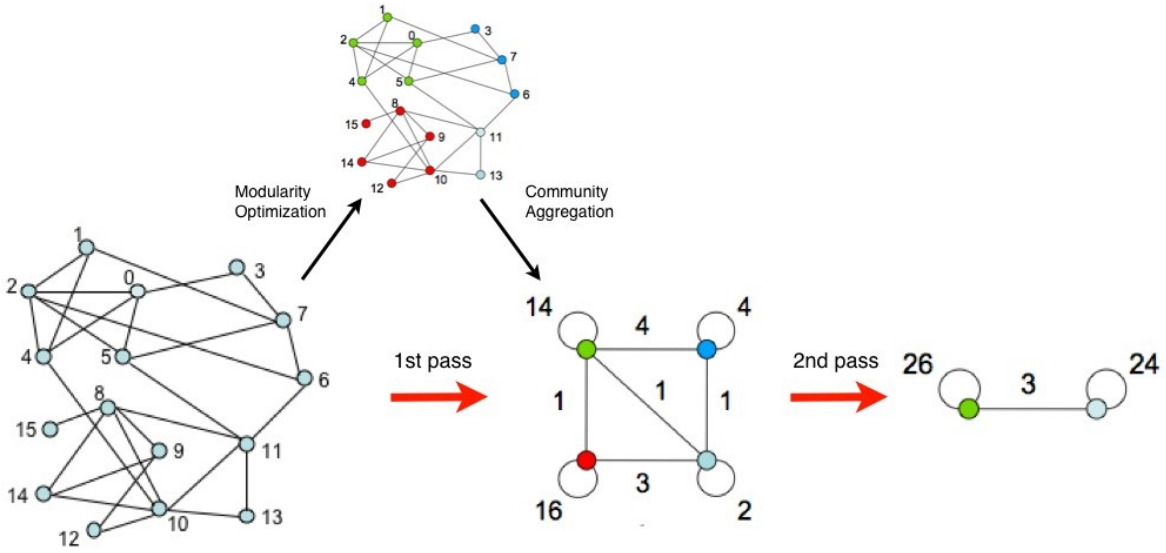


Figure 5.2.3: Louvain Method, source [22]

$w_{i,j}$ is the weight of edge $e_{i,j}$, with m representing the total weight across all edges in the network, c_i and c_j identifying the community membership of node i and j, with the Kronecker delta $\delta(c_i, c_j)$ returning 1 when the two nodes are in the same community, otherwise the value is 0. k_i and k_j is the total weight of edges from i and j. m is half the sum of all the edge weights across the whole network.

On the first iteration, each node is assigned its own community label, thus the number of communities is the same as the number of nodes in the network. Modularity is then computed across the network, with each node being within the community of a neighbour, with the largest positive change in modularity being carried into the next iteration. This process continues until there is no change in the network's modularity. The second phase then combines nodes in the same assigned communities into one node, with all edges within the community that end in the same target community being aggregated. The previous method outlined for phase one is then reapplied, until there are no changes in modularity between communities of nodes. At this point, a hierarchical community structure has been extracted.

Implementation

Unlike previous methods used in generating the networks (sections 5.2.2 and 5.2.3) and computing the significance of edges (section 5.2.4), there are challenges in implementing Louvain at scale using big data technology. This challenge comes from Louvain modularity requiring iterative recursive passes until the maximum modularity is attained in both phases. Thus, implementing a solution using tools such as Hive and HQL is impractical and impossible.

For ease, the original Python module, as release by the authors of the original paper [22][6], is used. All

[6] http://perso.crans.org/aynaud/communities/

that is required is a CSV of the generated network; this is then transformed into a Python NetworkX[7] object. By then using the Python `community` module, community membership is assigned to each node within, which can then be saved to a CSV file for later use.

5.3 Innovation identification

The aim of this thesis is to detect and model language change. However, to identify language change, one must first classify what is 'normal language'. When assessing diffusion of language, [64] used a hand-curated list of innovations to track, and [124] used a model across all words in a corpus. However, we are interested in new words, words that have never been seen before; thus, we aim to remove 'known' words from the datasets and focus on the 'new' words remaining. The removal of known words requires the identification of the gold standard of language used to identify formalised/accepted words in the English language. A number of sources could been used, from the Oxford English Dictionary (OED) to all the words contained within Wikipedia.[8] However, for simplicity and ease of access, the British National Corpus (BNC) was used as the gold standard for English.

The BNC [11] is a 100-million sampled corpus of written and spoken English that was compiled in the 1990s and is freely accessible for commercial and research usage. However, there are limitations in using the BNC, mainly because it was compiled in the 1990s, as language has changed slightly since then. These changes can be seen in the number of words that have been accepted into common language that are not in the BNC, such as 'email' and 'internet'.

Only removing words that are in the BNC still leaves a lot of undesired noise and items that may not be considered as innovations, such as URLs and emojis; therefore, these are removed too through the use of `regex` and pattern matching. Additionally, [62] shows that there is a prevalence for expressive lengthening within social media, so the 'same' word may appear within multiple forms by varying the number of repeated characters, such as *soooo* and *so*. This final stage reduces sequences of three or more repeated characters down to two. The remaining words are treated as the innovation to be assessed in the reset of the research.

Implementation

Before filtering out known words and assessing the new words, a number of steps must first be performed: tokenisation, POS tagging and light normalisation. These initial stages will take the paragraphs or sentences within each tweet or Reddit post and separate them into individual words, along with associating semantic tags to each given word.

[7] NetworkX is a Python package for the creation, manipulation and visualisation of network structures, https://networkx.github.io/

[8] http://corpus.byu.edu/wiki/

Tokenisation is the process of splitting a stream of characters (text) into their distinct word forms; there is a wide set of open-source tools that can be used, such as CoreNLP[9] and OpenNLP.[10] However, as identified by [64], traditional NLP tools have limited accuracy in noisy social media text, mainly due to social-network-specific features within the text, such as mentioning in Twitter through the inclusion of '@', which traditional tokenisers could have been treated as part of a malformed email address. Thus, for this work, each tweet and Reddit post is tokenised using a specially designed tokenizer for noisy social media data such as Twitter, which deals with features such as hashtags and emoticons [64].

Following tokenisation, each post is then POS tagged; this is the assignment of grammatical classes to a word dependent on the context of the word. Again, there are many popular taggers, such as those implemented in OpenNLP and CoreNLP. However, again, as shown by [62], traditional taggers' accuracy is reduced when used on noisy data. Therefore, again, we use a [75], which is a specially designed POS tagger with a custom hand-annotated Twitter training set. An added benefit of using [75] is that the traditional tag set has been expanded to induced tags for OSN-specific features such as emojis, hashtags and expressive lengthening.

The two datasets in this research are large in size, so an NLP pipeline is implemented in Apache Spark. In the solution, a `mapper` is used to apply the same sequence of NLP steps to each document (Tweets or Reddit posts). To limit the overheads of loading, the necessary models for the tokenisers and POS taggers `mappartition` is used, which allows for the same model to be used across multiple documents. Through the use of Apache Spark, the whole of the 1 Tb+ Reddit dataset can be tokenised and tagged within six hours.

5.3.1 Timings

As highlighted in Section 4.7 the choice to use big data technologies was driven by the need to process large amounts of data (data sets of 1Tb+) in a timely manner. To better understand to what extent Spark, and the number of nodes in a cluster effect the processing time of data-sets we will run a number of the prepossessing jobs on a sample of the datasets (Twitter) timing the execution and varying the number of nodes. These tests used a sample of the Twitter data set (100,000,000 Tweets) and was run against the code to identify the innovations within the text of each Tweet (detailed in Section 5.3). This means that each Tweet is tokeniser, tagged and then filtered of known words. The output of this job is then the list of innovations used throughout this thesis.

Each node in the set-up contains a 4 core Xeon processor and 32 Gb of RAM. The code will be run on a cluster containing 1, 2, 3, 4 and 5 nodes.

As one can see that the time taken to process the data reduces as the number of nodes increases. Though, as one doubles the nodes one does not see a halving of the execution time. This is because of

[9] http://stanfordnlp.github.io/CoreNLP/
[10] https://opennlp.apache.org/

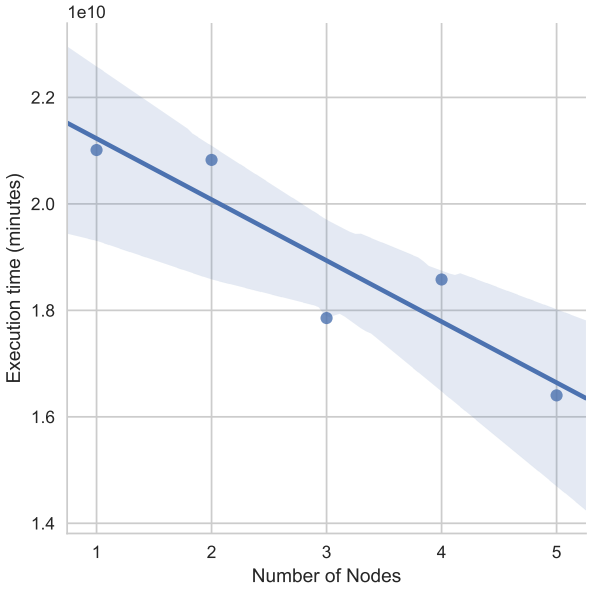


Figure 5.3.1: Time performance results of Spark, running job the innovation detection on sample of data varying the number of nodes.

increases over heads from the JVM as the amount of data which is needing to be moved between machines (shuffled) increases. Though within the limited sample there appears to be a linear relationship between the number of nodes and the time taken to finish the computation, similar results are also see in processing large scale log data in [140].

5.4 Summary

This section has introduced the core preprocessing stages that are applied to the raw data. This is done in order to extract the network structures and identify the innovations that will be analysed within the three core research questions.

First, the four networks were introduced, with the micro network (Section 5.2.2) representing user interactions, and the macro networks (Section 5.2.3) representing the interaction of aggregations of users by abstracting the networks by their geographical or subreddit structures. The networks can be extracted by querying the raw data sources, which can be implemented in HQL.

However, mining all the data means that the resulting networks are dense, hiding the core network structures and interactions (Section 5.2.4). Thus, the number of edges in each network is reduced through the removal of insignificant edges, which is achieved through the application of backbone extraction. Backbone extraction is a simple significance test on the edge weight, which can be implemented in a simple HQL query.

To assign communities membership to each node (Section 5.2.5) within the four networks, a number of scalability challenges are encountered. Thus, when implementing a Louvain modularity, a single node

implementation using NetworkX and Python must be used.

Finally, we introduced the method of classifying language innovations, the unit we aim to model, and predict their growth and adoption. These are words that have not been seen before. This was achieved by the removal of words that appeared within the BNC, along with the use of regular expressions and specialised POS taggers to identify URL and emojis/emoticons.

This section ultimately introduced in detail the preprocessing stages that are common across the three research questions (sections 1.1.1, 1.1.2 and 1.1.3). In each of the three research chapters, aspects of preprocessing will be reintroduced. However, the manner in which they are implemented is discussed in less detail, with the focus on the data, and the research applied to the preprocessed data.

Part II

Contributions

Chapter 6

Detection of Language Innovations

6.1 Introduction

Recent word innovations with roots in OSN include: 'fleek', 'Brexit', 'Grexit' and 'Ubered'. Some readers may or may not understand the meaning of some of these words, yet they are becoming codified into the English language through their recent addition to a number of dictionaries, such as:

- *bants (also bantz)*, **pl. n.**: (Brit. informal) playfully teasing or mocking. Remarks exchanged with another person or group; banter.
- *beer o'clock*, **n**: an appropriate time of day to start drinking beer.
- *brain fart*, **n.**: (informal) a temporary mental lapse or failure to reason correctly.
- *Brexit*, **n.**: a term for the potential or hypothetical departure of the United Kingdom from the European Union (joining of 'British' and 'Exit').

Due to this nature of language change, the purpose of this chapter is to present the results of the work on detecting and ranking language innovations in OSN. To date, language innovation research has been restricted to manual work, which is both time-consuming and costly. This research enhances the capacity for innovation detection through the operationalisation of heuristics in a scalable manner, drawing on the heuristics proposed by lexicographers [144] and [18]. By applying them to large textual datasets from two OSN (Reddit and Twitter), our results show the ability to detect innovations within a community and at multiple levels.

Previous research has gained tremendous insight into online and offline activities of users from data generated on OSN, from tracking flu outbreaks on Twitter [133] to modelling user 'churn' in online forums [53]. Innovations in language have started to gain attention, with [42] determining source words of word blends and [64] modelling the diffusion across the geographical landscape of new words. However,

none of these works has shown the ability to detect the emergence of innovations across a large period of time, or across multiple datasets.

Analysing OSN data requires the handling of large-scale datasets and associated challenges; in view of this, the contributions of this work are as follows:

- **Word innovation acceptance models though computational means**: An approach that takes grounded heuristic models from the fields of linguistics and lexicography and applies them through computational approaches.
- **Identification of local and global innovations**: We show that the models that have been developed show innovation acceptance on local and global levels within their respected networks.
- **Multiple network analysis**: We find that innovation detection can be applied across different types of OSN.
- **Large-scale corpus analysis / Analytical tools**: We demonstrate the ability to perform large-scale textual analysis within minimal preprocessing at scale using state-of-the-art scalable frameworks.

Identification and understanding of innovations in language is not merely of academic interest but has impact in other fields. By understanding the changes in language and variations across networks, marketers should be able to focus tailored messages to specific online communities, identify key phrases that resonate within the community, and anticipate the need to modify a message in conjunction with language changes. This importance in understanding OSN by marketers has been seen in the ever-growing body of work that aims to understand the influence and diffusion of content within networks ([165]).

The remainder of this chapter is organised as follows: section 6.2 identifies and discusses the state-of-the-art work in innovation detection by computational means, along with a discussion of language as a feature of OSN research. The innovation acceptance models that are operationalised through the use of computational means are introduced in section 6.3. Section 6.6 explains the technical implementations of the innovation acceptance models, along with the issues involved in processing large quantities of data. Section 6.7 explains the experiments conducted, and the discussion and conclusion are presented in section 6.8.

6.2 Related Work

The following section will identify relevant work across the fields of NLP and OSN research, both of which have seen a growth in interest in the usage of non-standard language.

Variation in language can be seen throughout the world; this phenomenon has been studied though the use of geo-tagged tweets, with [79], [80] identifying geographical variations in language in both English and Spanish. Whereas [100] was able to identify dialect regions in the US, e.g. Upper Midwest and Lower Midwest. [61] was able to predict the diffusion of innovations across the US as a function of census data and influence probabilities. However, clusters of tweets in metropolitan areas within the US were very large, with results only demonstrating the diffusion of words such as abbreviations (e.g. 'idk' - I don't know) and shortening (e.g. 'you' to 'ur'), which is connected to a younger demographic.

Language, however, is not independent of social factors such as age, race and gender, with [160] and [163] showing that these factors have a strong influence on the language of the user. [163] showed that the gender of a user can be predicted from features such as emoticon/emoji usage or topics discussed. Whereas [5] showed that including the attributes from a user's local neighbourhood can increase the accuracy of user attribute prediction.

The variation of user language over time can also be seen in specialised OSN; [53] showed that, as users enter a community, their language moves towards that of the community. However, before leaving, their language diverges; this can be used to predict if a user is going to leave a community. However, the dataset used contained highly specialised language, and the convergence was merely users using the more specialised terms within the community.

The power dynamics between users can be identified by assessing the likelihood of one user accommodating the language of another user ([51], [52], [190]). [51] analysed user roles within the network, convocation chains and domain-independent language features (e.g. personal pronouns) from the Wikipedia editors' chat logs. The results showed that lower-ranked users adapted their language to that of the senior 'admins', whereas the 'admins' altered their language very little. By also tracking users as their seniority increased, the study showed that their propensity to accommodate the language of lower users decreased. However, as this work tracks domain-independent language features, and only uses standardised language, it would not detect the power dynamics within conversations that represent a large proportion of innovations.

As stated earlier, users are highly selective in the language they use, not only on an inter-personal basis, but also when attempting to broadcast a public message on Twitter. [158] showed that users vary their language dependent on the intended audience size of a tweet. By classifying intended audience size (is the message directed at a small or large group of people?) based on certain features of Twitter messages such as if the tweet contains a hashtag '#' would indicate a wide audience. The results indicated that focused messages to small audiences increase proportions of non-standard forms (innovations) compared to messages with a larger audience that had a higher usage of standard form words.

Depending on the innovation type (abbreviation, word blend, etc.) a number of solutions have been proposed for determining the meaning of innovations. For word blends, [42] assessed the morphological

and phonological characteristics of known word blends and then attempted to use these features to correctly identify the source words for identified new word blends on Twitter. Although the results were marginally better than a random baseline, limitations appeared to stem from sparsity in identifying the meaning from the same dataset. Additionally, to deal with acronyms, [171] proposed a solution that uses dynamic programming and a set of rules to identify candidate combinations from the web. Even though they were able to correctly identify the source words of a number of acronyms, the reliance on search engine results as a corpus limits the repeatability of the research and brings into question the validity of the search engine results.

However, the meaning of words does not stay constant over time, nor context of use. Through the use of neural nets and deep learning, large-scale semantic changes have also been shown in the Google N-gram corpus and social media datasets [123]. However, even though they identified words such as 'gay' as having changed significantly over time, there was limited explanation as to what may have caused the changes.

6.3 Language Innovation and Acceptance

The following section introduces the grounded models that will be used to guide the methods within this research.

Within the fields of linguistics and lexicography, there is a long tradition of assessing the current state of language, as well as determining the emergence of norms and innovations, both of which are within the scope of this research question. Seminal studies into language change and variation originated from Labov in New York during the 1960s, [129]. Labov focused on language variation across social classes. The downside of these studies, as with much field research, is that it is time-consuming compared to the approach we purport here. However, the recent commoditisation of computing power and the availability of large corpora such as the Google N-gram dataset have enabled the exploration of historic language with greater ease (c.f. [123]).

Nonetheless, it should be noted that the existence of an innovation within a community does not mean that it has been accepted into the language of said community. For it to be accepted, an innovation must first achieve a certain level of usage or be acknowledged within the community. However, to determine at what point an 'innovation' has been accepted into a language by a community is a challenging task due to the wide range of factors that acceptance of an innovation encompasses; for instance: does frequency influence acceptance, or does acceptance rely on a collective understanding of the concept to which the innovation refers?

To assess the acceptance of a change in a language, for example, whether a word should be added to a dictionary, a number of heuristic tools have been developed by lexicographers. Two widely cited tools are

Barnhart's VFRGT [18] and Metcalf's FUDGE [144]. Both were developed to assess and predict whether innovations introduced into a community would be maintained or lost. However, both were developed to be used by lexicographers using a scoring method; thus, the value assigned is at the discretion of the scorer, and therefore subjective.

Metcalf stated that, for a word to be accepted into a dictionary, it must first fulfil five points that are measured on a 0 to 2 scale; thus, the higher the score, the higher the probability a word will have entered into a language. The five points are:

- (**F**) Frequency of the word - how often does the innovation appear
- (**U**) Unobtrusiveness of the word - the word should not be used for an exotic reason
- (**D**) Diversity of users and situations - the variation of situations and users using the innovation
- (**G**) Generation of other forms and meaning - if the word starts to influence other innovations, then it has an increased chance of success
- (**E**) Endurance of the concept to which the word refers - in reference to historic meanings of the word

Barnhart proposed a similar method, again, identifying time and frequency as key components, but also placing emphasis on the formations of words and the genre in which they are used:

- (**V**) Number of forms, including variation in spelling and/or derived forms
- (**F**) Frequency of the word
- (**R**) Number of sources, e.g. newspaper, magazine, books
- (**G**) Number of genres the word is used within - news, poetry, spoken, blogs
- (**T**) Time span of the word

At the core, both measures rely on time and frequency. However, the difference between Metcalf and Barnhart's methods are that Metcalf's is user centric when looking at the diversity of users' word usage, whereas Barnhart focuses on the generally of usage and sources as a proxy for the users. These heuristics will be used as a base for the model that we propose.

6.4 Methods

Having defined which grounded methods will influence this work, the following section will explore the methods developed to operationalise the heuristic models, as well as introducing the datasets to which the methods will be applied. The approaches come predominantly from the fields of data mining and

	Reddit	Twitter
Unique Words	2,942,555	526,342
Posts	1,054,976,755	111,067,539
Users	10,528,521	1,696,630
Communities	121,373	3,046
Innovations	2,712,629	373,217
Days in Dataset	880	283

Table 6.1: Data Set Description

NLP. Additionally, due to the size of the data being processed, these methods must be adapted with the use of distributed computing frameworks.

6.4.1 Data

As highlighted in the previous sections, and also in the preprocessing (Chapter 5) and methodology (Chapter 4) chapters, the growth in popularity of OSN means that there is the ability to mine large samples of textual data that represent the language that users use, as well as representing individual users' interactions. For this research, we use samples from two popular social networks, Twitter and Reddit.

Twitter (as introduced in Section 4.4.1) is an OSN that allows people to share 'tweets'; these are posts that are limited to 140 charters. When a tweet is posted, followers of the user will see the tweet appear on their timelines, and these can then be re-tweeted or commented on. Studies have shown that people tweet for a number of reasons, including communicating with friends, gathering information, and even stress relief [103]. This data was collected using the Twitter streaming API, which allows filters to be applied in order to indicate what should be returned to the server. For this study, a bounding box was applied, which means that only tweets with GPS coordinates deriving from the UK were mined, as we are focusing on the English language.

The second dataset is a sample taken from Reddit (introduced in section 4.4.2) from the beginning of 2013 to mid 2015. Reddit is a self-governing online social network structured into 'subreddits' that encapsulate a topic or community. Content is posted to each subreddit, and users can comment, as well as 'down' or 'up' vote each thread (and post), affecting the thread's prominence within the subreddits. Reasons for using Reddit as a data source are that a) the user base is highly active and the popularity of the site is relatively high [59] (though this is skewed to a younger demographic, much the same way as Twitter), and b) the majority of comments and posts on the site are public. Reddit is different from Twitter; Twitter tends to be used for self-broadcast purposes, whereas Reddit is used to a greater extent for discussion and debate.

6.4.2 Data Grouping

The following section explains how we prepare/group the data, allowing for analysis at different times and community granularity. For data, we mean each individual post mined from either Twitter or Reddit.

Group by Time

To group the data by time, the function $weekofyear(e)$ returns the week the tweet or Reddit post was created. We thus define time groupings as deriving a set as follows:

$$E_k = \{e : weekofyear(e) = k, e \in E\} \tag{6.4.1}$$

Where k is the number of weeks since the first item was collected within each dataset, and E is the set of all entities (Reddit posts or Twitter tweets).

Group by Community

Each of the two datasets represents highly distinct user interaction patterns. For Twitter, this is the user interaction with the geographical landscape in the UK, and for Reddit, it is the user interaction between subreddits. As highlighted earlier, the language that individuals use is heavily influenced by the individuals around them, and, ultimately, the communities to which they belong. Thus, this section introduces how communities are defined in each network, and how these are used to assess language change.

As the Twitter dataset is geographically bound within the UK, we can use geographical features as the foundations for defining communities within the network. However, as highlighted in section 5.2.3, there are issues in aiming to extract communities from the data; therefore, the UK postcode system is used as the basis for clustering tweets into distinct communities. An added benefit of using postcodes as a basis is that there is a pre-existing hierarchical taxonomy, allowing for the translation from low-level postal areas, e.g. LA1, to high-level regions within the UK, e.g. North West. Thus, a tweet that originates from the postcode LA1 appears in the LA1 set, the LA set (of which LA1 is a member), which is itself part of the North West set, which in turn is part of the national set.

For Reddit, the identification of communities of users and posts is significantly simpler. Each post must be generated within a subreddit, all of which are explicitly defined in each entity. Thus, communities are classified as subreddits for this research question.

Even though the community structures are different, comparisons can be made and hypothesised. Lower community levels (postcode and subreddit level) will result in more specialised language that is affected to a higher degree by the community, showing a greater level of nationalisation. However, on a global level, the specialisation across communities could be amalgamated, revealing the 'standard'

language of the community.

To group entities (e) into their given communities, the function $communities(e)$ is used. This returns the set of entities within the given community:

$$E_r = \{e : c = r, c \in community(e), e \in E\} \tag{6.4.2}$$

Where E is a complete set of entities (tweets or Reddit posts), and $e \in R$ is the set of all possible communities.

6.5 Operationalisation

At the core of this research, we aim to operationalise the acceptance models proposed by Metcalf and Barnhart (section 6.3), to identify the existence and acceptance of language change and innovation within OSN. Operationalising FUDGE and VFRGT in a technical sense means identifying a number of variables/features that can be extracted that are believed to subsume the properties of each heuristic.

We propose that the FUDGE and VFRGT models can be adapted into three metrics:

- **Variation in Frequency** - The change in user usage and word frequency over time
- **Diversity in Form** - The variation in the number of forms of an innovation over time
- **Convergence in Meaning** - Does the meaning used across a network converge over time and is the convergence maintained?

The premise of this work is to identify the acceptance of innovations; therefore, first, a word must be classified as an innovation or not. To achieve this, we classify a word as an innovation if it does not appear within the BNC [11]. The BNC was chosen as the baseline of English language as it is one of the most comprehensive studies to sample the language in recent times, sampling not only newspapers but also books, written communication and oral discourse. This can be seen in Table 6.1.

6.5.1 Generalised Framework

We first propose a generalised framework that is used in conjunction with the three metrics introduced above in assessing language change over time. The premise of the metrics introduced in section 6.5 is that they are applied to each time period across the dataset. Though, as these metrics are applied to each time period how to one quantify and rank changes (between innovations) across subsequent time period too ultimately model innovation acceptance.

To rank each word ($w \in W$) for each metric ($m \in M$), we initially propose that language changes in a monotonic fashion (changes in such a way that it either never decreases or never increases); thus, we

aim to quantify an increase or decrease in each of the three measures over time. To quantify the rate of change, we propose that, by fitting a Spearman's rank correlation to a word's time series for a given metric (w_m) set, we will quantify the extent to which the change is increasing or decreasing, with the value ($\rho_{w,m}$) of each w_m in the range of -1 to 1.

$$l = (X_t : t = 1, 2, ...len(w_m)) \tag{6.5.1}$$

$$\rho_{w,m} = SpearmansCorrolation(w_m, l) \tag{6.5.2}$$

Where l is an ordered vector of increasing numbers indicating the time period since the beginning of data collection. w_m is the result of each time period with the metric (m) applied. With the computed result, $\rho_{w,m}$ represents the Spearman's rank correlation value for the given word (w) time series of metric (m).

Additionally, we are aiming to sample statistically significant language changes for each metric. For each metric ($m \in M$), we have a set of values ($\rho_{w,m} : w \in W$); thus, we make the assumption that the set of $\rho_{w,m}$ such that ($\rho_{w,m} \in R_m$) is normally distributed. Presuming that distributions of $\rho_{w,m}$ have a normal distirbution, statistically significant changes are those that fall above or below the 95 % confidence interval of the distribution. Thus, if a word is above the upper 95 % confidence interval, it is increasing in popularity; if it is below the bottom confidence interval, it is classified as decreasing. As stated, this model is applied to each of the three metrics ($m \in M$) and acts as a generalised framework, sampling statistically significant changes within a language.

6.5.2 Variation in Frequency

Both Metcalf and Barnhart stated that, at the core of accepting an innovation is the frequency and endurance of usage. To operationalise this, the relative frequency of a term will be used as a proxy of its popularity, and the endurance is measured as the relative frequency over time.

First, we will use a uni-gram model of the language for each time period; thus, the relative frequency of the word within a time period is:

$$F(w, t) = \frac{|w \in C_t|}{|C_t|} \tag{6.5.3}$$

Where C_t is a bag of words at time t, with $|C_t|$ returning the number of words in the bag for time period (t) and $|w \in C_t|$ returning the frequency of the word (w) within the bag of words (C_t).

To convert this into a time series ($\bar{x}_w$), function $F(w, t)$ across all time periods (t) in the dataset (C):

$$\bar{x}_w = \{F(w,t) : t = [0...n]\} \tag{6.5.4}$$

However, measuring relative frequency is susceptible to users who may use the same 'innovations' multiple times in a given time period. This will result in the appearance of the innovation's popularity increasing across that network, which may not be the case. For this reason, we introduce a second measure that, instead of counting distinct instances of a word, counts the number of distinct users in a time period using the given word.

$$F(w,t) = \frac{|user(w \in C_t)|}{|C_t|} \tag{6.5.5}$$

Where the function $user(w)$ returns a list of distinct users that have used w within the given dataset.

6.5.3 Diversity of Form

Building on variation in frequency, variation in form is key to the acceptance, as it shows that users feel that concepts will be understood if conveyed through varying the morphological form. Variation in form can take a number of modes, from dropping a letter to attaching a prefix or suffix. To assess variations in form, we look at two different methods: addition of a prefix or suffix, and alternate spellings within a given edit distance.

Assessing the addition of a prefix or suffix allows us to see if there is morphological variation in the term. To implement this, two lists of common prefixes and suffixes were taken from the Oxford English Dictionary (OED); these include: 'ing', 'homo' and 'hetero'. As with the previous measure, we aim to compute a time series representing the probability of prefix of suffix addition for each time period.

To determine the probability of suffix addition for a word (w) at a given time period (t), we apply the following function:

$$MS(w, C_t, S) = \frac{\sum_{s \in S} |endswith(w, s, C_t)|}{|C_t|} \tag{6.5.6}$$

with the probability of suffix addition

$$MP(w, C_t, P) = \frac{\sum_{p \in P} |beginswith(w, p, C_t)|}{|C_t|} \tag{6.5.7}$$

MS and MP are functions that take a word (w), a bag of words (C) for a time period (t), and a list of prefixes (P) or suffixes (S), respectively. The functions $startswith(w, p, C_t)$ and $endwith(w, s, C_t)$ both return a list of all instances of words starting or ending with the prefix or suffix within C_t.

To turn these metric into a time series, the functions are applied to each time period in the datasets

as:

$$\bar{P}_w = \{MP(w, C_t, P) : t = [0...n]\} \tag{6.5.8}$$

and:

$$\bar{S}_w = \{MS(w, C_t, S) : t = [0...n]\} \tag{6.5.9}$$

The second proposed method clusters words together by using 'edit distance' instead of fixed prefix and suffix additions. This has a number of advantages over merely looking for the addition of common prefixes and suffixes, as it will allow for the identification of common misspellings of innovations, and identification of 'innovations' that may be misspellings of 'standard' words.

Edit distance calculates the number of inserts, deletions and substitution operations that have to be performed to transform one string into another. Thus, to transform the word *apple* to *apples*, we only have to insert an 's', resulting in a value of 1; whereas, turning *football* into *ball* would have a value of 4 due to the deletion of four characters.

To apply edit distance, we use the following function:

$$E(w, C_t, d) = \frac{|editdistance(w, d, C_t)|}{|C_t|} \tag{6.5.10}$$

Where $E(w, C_t, d)$ is a function that takes a word (w), a bag of words for a given time period (C_t) and a maximum edit distance (d). The function $editdistance(w, d, C_t)$ returns all instances of words within the corpus (C_t) that have a maximum edit distance from w of d.

Again, to turn this into a time series (E_w), this will be applied to all the time periods within the corpus (C).

$$\bar{E}_w = \{E(w, C_t, d) : t = [0...n]\} \tag{6.5.11}$$

6.5.4 Convergence in Meaning

The final measure aims to quantify the collective meaning associated with an innovation across a population of users. As an innovation initially enters a system, it may have a diverse set of associated contexts, but, as users interact to a greater extent, the meaning(s) converge into a collectively understood context. Thus, we aim to assess the convergence in collective meanings of innovations by comparing similarities in word contexts.

Traditionally, systems such as WordNet have been used to identify he synonyms of words. WordNet ([147]) is a large linguistic graph database that represents the synonyms within its structure, giving us the ability to query it for similar words. However, in this research, we are looking at words that have

never been seen before, so it does not make sense to use such systems as our words will not be included. For this reason, we propose a new method that relies on word co-occurrence and the embedding of the word within its own dataset, to suggest potential synonyms.

In the fields of data mining and NLP, there has been an increasing body of work using neural-network-based techniques for learning the vector representations of words; these have been used for a number of tasks such as POS tagging and machine translations ([217]). To detect similar words in corpora, [145] proposed `word2vec`, which is an unsupervised method of learning the embedded dimensions of word vectors by maximising the likelihood that words are predicted from their context.

Both datasets used in this research can be clustered on a time (t) and by community (c); thus, we propose learning the embeddings of words across communities and time periods. This means that, for each time period (t) within each community (c), we apply `word2vec` to each bag of words ($C_{t,c}$); this results in an embedded model ($(W2V)_t^c$) specific to the community (c) and time (t). Then, for each innovation, we query each model ($(W2V)_t^c$) for the top 100 ,similar words. This results in a list of similar words for each community (c) for a given time period (t). Thus, to then determine if communities within the same time period have 'similar' word representation, a Jaccard Similarity Index (JSI) will be applied across all community pairs (cartesian product, $C \times C$) within a given time period (t). Finally, the set of coefficients for each time period will be averaged out, giving one metric per time period.

$JSI(A, B)$ is the function that takes two sets of words and returns the resulting Jaccard similarity score.

$$JSI(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{6.5.12}$$

$W2V_{t,c}$ represents the `word2vec` model for community c at time t. This can be accessed through $S_t^c(k, w)$, which is a vector of words (of length K, dependent on how many you want to return - the default is 100) for the time period (t) for community (c), for word (w).

$$S_t^c(k, w) = function(c, w, t, k) \tag{6.5.13}$$

To compute the average JSI for a given time period (t), a Cartesian product is taken across all communities (C), with the JSI computed for each combination, then divided through the number of community combinations. This is computed using the following function:

$$f(w, t, C) = \frac{\sum_{i,v \in C, i \neq v} JSI(S_t^i(k, w), S_t^v(k, w))}{|C|\,(|C| - 1)} \tag{6.5.14}$$

The goal of this method is to assess the extent to which the different communities in the dataset use the innovation (i) in the same context (embedding). Thus, in a given time period, if the value is 0 then it would indicate that the context in which the innovation is used is significantly different, whereas a

value nearing 1 would indicate near identical contexts across all communities.

A modification has been made to this method due to preliminary research that produced limited results; this was put down to a limited sample being returned from the $W2V$ model due to sparsity across a number of communities. Thus, to improve this, we could have increased the number of words returned. However, this reduced significantly the JSI as the number of overlapping works decreased. A second approach was taken, whereby each word in the vector of similar words (as returned from function $S_t^c(k, w)$) is also queried against the model. This second order search method means that the space on which the JCI is applied is 10 times the original size, with results indicating a greater overlap than achieved by just expanding the number of the original query.

6.5.5 Limitations

The three methods proposed do not cover all the categories discussed in the VFRGT and FUDGE frameworks. However, we believe that the categories that have been removed (e.g. genera and source) lay outside the boundaries of this research, and would have removed the focus from the core research questions.

A second limitations is the lack of a grounded truth against which the results of this work can be assessed. We could have used the respective changes in the Online Social Network (OSN) as a baseline; however, the data collected is what is currently being used in language. Since the OED is only retrospectively compiled, in order to detect the successive changes, we would have to look for historical data from one or more years ago. An alternative is the Urban Dictionary [1], which is a crowd-sourced online dictionary focusing on internet slang and terminology, with creative and/or offensive definitions. However, the API available for the dataset does not give time of entry or the order in which the word definitions were assigned or changed, thus limiting its usefulness as a baseline for the entry of a word into the dictionary. However, as we develop a system that removes the human variable from a tool for curating a dictionary, the need for a grounded truth is minimised.

It should be noted that the Twitter dataset had sampling issues during collection, meaning that the data was corrupted between April and May 2015.

6.6 Computational Methods

The following section will explore the computational methods used to implement the operationalised metrics discussed in the previous section. The size of data being used could be classified as 'big data' (as discussed in section 4.4); therefore, to process the data in a timely manner, the methods are implemented in such a way as to be easily scaled across a cluster of computers.

[1] Urban Dictionary

Key	Column Family: comunitylevel1	Column Family: comunitylevel2			
	Column: Reddit.com	Column: AMA	Column: Politics	Column: Movies	...
Fleek	[0,2,1,1]	[0,2,1,1]	[0,2,1,1]	[0,2,1,1]	...
Lol	[0,2,1,1]	[0,2,1,1]	[0,2,1,1]	[0,2,1,1]	...
...	...	...	...	...	...

Table 6.2: HBase scheme design

6.6.1 Technical Set-up

Recently, there has been much development in the field of scalable computational systems. One of the first popular systems was Apache Hadoop [210].[2] However, in recent times, this has been overtaken by Apache Spark [216].[3] Both have been shown to handle large quantities of data, though Hadoop is IO bound due the HDFS being used for hot storage; however, Spark maintains intermediate results in memory, allowing for fast access speeds.

Spark also has a number of benefits compared to Hadoop when writing applications; this comes from Spark's API being more flexible and easier to programme compared to Hadoop, in which each step in an application must be defined as its own job.

6.6.2 Methods

As stated earlier, the operationalisation of measures has been developed in such a way as to allow them to scale across a cluster of computing nodes. This was achieved using Spark, along with a number of distributed data stores.

Variation in Frequency

Computing the relative user and raw frequency of an innovation for each time period and community level is relatively simple. The issue ultimately arises from processing the data in one day across different community abstractions, as this requires large aggregation stages that are highly memory intensive. However, this is very slow and when we implemented this process, Apache Spark crashed. For this reason, we utilise the time series features of Apache HBase to store intermediate stages of the processing pipeline, allowing for incremental processing of data across both community and time.

Apache HBase is a column-orientated database that can be distributed across a number of machines. Within its design, we can apply time stamps to each cell, thus using it as a time series store for a given word. For example, the key for each row is the 'innovation', the column family is the level of the communities, e.g. national, regional, postcode, and each individual 'column qualifier' is each community that has been processed. Each cell can have multiple time-stamped versions, used to store each value

[2] https://hadoop.apache.org
[3] https://spark.apache.org

within the time series; a value is given the time stamp of the day in which it occurred. To then compute the Spearman's rank across each, a second stage (again a Spark job) iterates though each cell, retrieves the time series for the given cell, and computes the correlation.

Diversity of Form

The second set of metrics attempts to cluster words based on their morphological form, the first quantifying prefix and suffix addition, and the second clustering through edit distance. Each of the two versions of the metric requires slightly different implementations due to scalability issues. Thus, in this section we explain how we solve the issues of large-scale edit distance calculations, and then how we use the results to cluster words together.

The main pattern is the ability to cluster multiple forms into one original base word, e.g. cooking → cook. To implement this, we first compute two word lists, one of which contains all of the words within given datasets, and one that contains only the innovations that are going to be assessed. The challenge is that we need to find the related morphological forms in the set of all words from the words in the set to be assessed and combined.

For prefix and suffix addition, this is relatively simple: for each innovation, the prefix or suffix is added, and the complete world list is searched for the modified innovation. Ultimately, each innovation must be compared to each word within the candidate list.

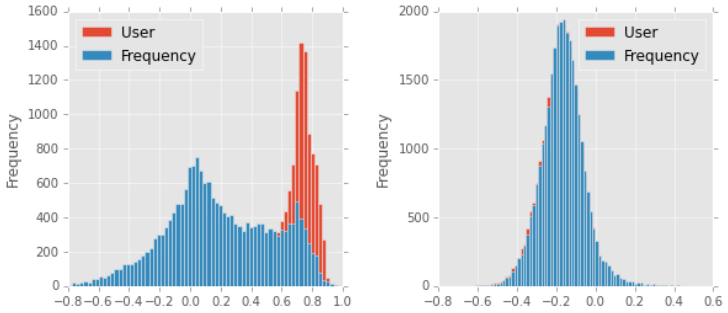
For the second metric (which utilises edit distance), a different approach must be taken due to scalability issues. The main issue comes from computing the edit distance, as this is proportional to the product of two strings. If a similar method to prefix and suffix addition is used, then the edit distance needs to be computed between each innovation and all words in the dataset; thus, computing the edit distance for a large dataset is proportional to $O(n^2)$, where n is the number of words. Thus, for simplicity, an alternative implementation is utilised, namely PostgreSQL[4], to identify strings within a given distance. By using the `fuzzystrmatch` and `pg_trgm` packages that ship with the database, we create an index on the word list based on the string similarity, then each innovation is queried against the database, returning all words within an edit distance of three.

Convergence in Meaning

The final method to be implemented is convergence of meaning. As stated earlier, measuring the semantic nature of an innovation is challenging, so we aim to measure the convergence in meaning through the commonality in context. As stated earlier, we are aiming to access the context of words across a geographical landscape by using time-bound `word2vec` models.

The `word2vec` model has been implemented in a number of languages; however, for this research, we

[4] http://www.postgresql.org/



(a) Reddit (b) Twitter

Figure 6.7.1: Spearman's distribution, user and word

will be using a JAVA implementation provided in MLLib (the Spark machine learn framework), which allows the processing power of the whole cluster to be utilised, not just the power of one. The cluster then cycles though each time period in each region, training the models and then querying each innovation against the learnt model.

6.7 Experiments

The following section discusses the application of the computation methods developed in section 6.4 to the datasets introduced in section 6.4.1.

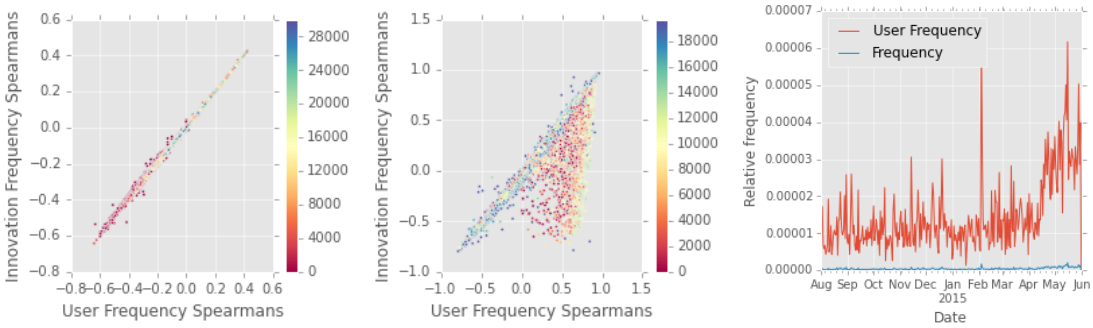
6.7.1 Variation in Frequency

As stated in the methods, variation in frequency looks at assessing the normalised word count of each innovation, as well as the normalised user frequency. In this section, we examine the results, as well as presenting some explanations of what is being seen.

Figure 6.7.1a shows the distribution of Spearman's rank from the global level of Reddit for both the normalised word and user frequency metrics. As we can see, both distributions are bimodal; however, the user distribution's left peak is significantly higher than the frequency, and frequency has a higher peak at a Spearman's value of roughly 0.05. These two distributions are significantly different from that of Twitter (Figure 6.7.1b), where both distributions are normal, with neither being significantly different from each other, though both are negativity skewed.

To sample the bimodal Reddit distribution, we applied two Gaussian, and then sampled from each respectively. The Twitter distribution (Figure 6.7.1b), however, is normally distributed for both methods; thus, we applied the sampling method as discussed in section 6.5.1.

By looking at innovations on the global level of both Reddit and Twitter, we can see the influence of online and gaming cultures in the prominent growth of '*ghc*', '*csgo*' and '*failfish*'. However, when looking at the normalised user frequency, the highly active communities become more subdued, showing that



(a) Twitter (b) Reddit (c) Innovation 'nutz'

Figure 6.7.2: Comparing word frequency against user frequency

Word	ρ	Definition
csgo	0.948870	A abbreviated name of the video game "Counter-Strike Global Offensive"
bruh	0.903385	term of endearment between two males
failfish	0.891613	Witness a huge fail
ghc	0.765274	"G: Game - H: Hacks - C: Cheats"
jihadis	0.533707	Jihadism is used to refer to contemporary armed jihad in Islamic fundamentalism.
lolit	0.353371	the act of laughing and agreeing to a statement

Table 6.3: Reddit user frequency sample from upper confidence interval

these words are not used by the majority of the network, and that phrases such as *'bruh'* and *'tumblrina'* increase at a significant rate.

When we look at the results from across varying locations within Twitter and different subreddits, we can see that the growth of innovations could be related to the topic and structure within these sub-communities. This is noticeable in the subreddits; in the politics subreddit, growing innovations indicated political discourse, e.g. *'blogging'*, and gaming uses gaming-related abbreviations, e.g. *pcmr*. However, when sampling within each community from the unique user usage distribution, we see little difference, with similar words appearing in each sample.

Table 6.4: Twitter user frequency sample from upper confidence interval

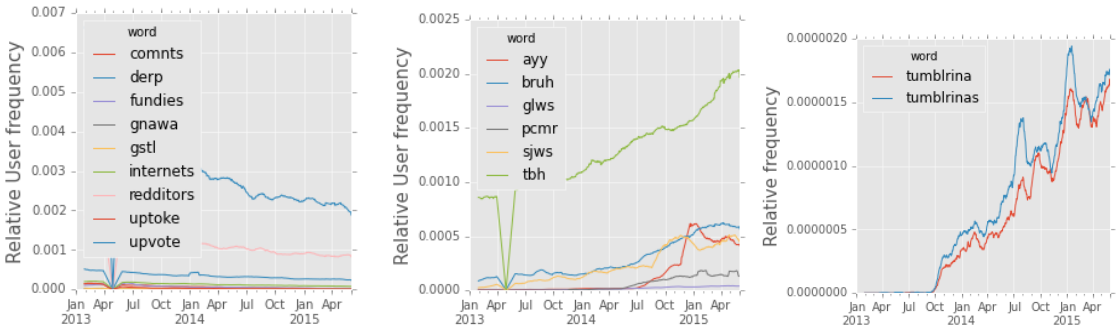
Word	ρ	Definition
ubered	0.357206	"To have used an Uber car", "To have been ripped off by uber"
fleeky	0.345077	"Eyebrows on fleek", "Eyebrows on point"
grexit	0.309802	"grexit, an abbreviation for Greek exit", which refers to Greece's potential withdrawal from the eurozone [5]
csgo	0.131103	"A abbreviated name of the video game Counter-Strike Global Offensive"
tfw	0.111941	"That feeling when"

Table 6.5: Top five innovations with the highest ρ from popular subreddits

Community	Frequency	User frequency
AMA	commenter, sfw, nsfw, faqs, lmao	commenter, stw, faws, nsfw, gotta, tbh
Funny	rekt, sjw, lmao, ayy, bruh	sjw, lmao, clickbait, tifu, rekt
AskReddit	multireddits, remindme, bruh, sjq, buzzfeed	bruh, sjw, askreddits, darude, textbook
News	sjq, gamergate, tumblr, ebola	isil, tumblr, gamergate, sjw, bruh
Politics	rehosted, blogging, politifact, midterms, millennials	subreddit, brownback, politifact, midtermsm, sjw
Movies	babadook, foxcatcher, moviegoers, ultron, nightcrawler	babadook, foxcatcher, snowpiercer, ruffalo, nightcrawler
Gaming	rekt, pcmr, amiibo, csgo, ayy	rekt, pcmr, amiibo, lmao, fpss

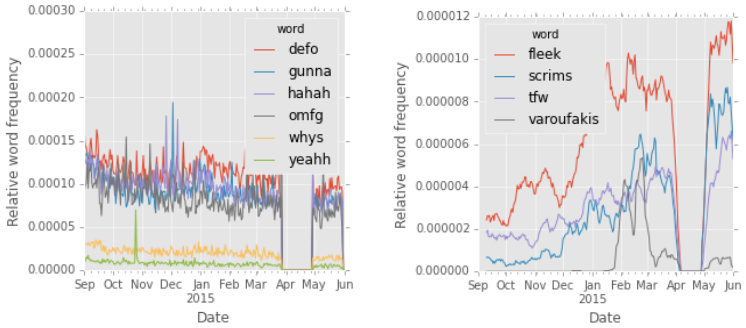
Table 6.6: Variation in frequency examples from UK regions

Region	Top 5 Normalised Frequency	Normalised user frequency
South East	fleek, splatoon, scrimming, demanda, awg	slamdunk, whato, gfin, scrims, gdfr, fetty
Greater London	reelected, deez, csgo, ubered, llah	fetty, fleeky, bugzy, blurryface, ketty
North West	pcars, cooldown, blurryface	blurryface, bugzy, dubsmash, beutiful, repunk
Northern Ireland	deez, scrims, tfw, lbgt, shinee, fleek	gfin, yikyak, gya, superfruit, blurryface
Scotland (S & C)	fleek, blurryface, rida, lebron	ryze, gfin, fetty, tiga, ipd
Wales (N)	daar, scrims, bruh, pbuh (Peace Be Upon Him), smfh	ukippers, sjw, hina, rns, svu
Wales (S)	flook, depay, gsw, dwp, tnx	allahha, fetty, lstd, gsw, blurryface



(a) User freq decrease (b) User freq increase (c) Tumblerian increasing

Figure 6.7.3: Reddit innovation examples



(a) Word freq decrease (b) Word freq increase

Figure 6.7.4: Twitter innovation examples

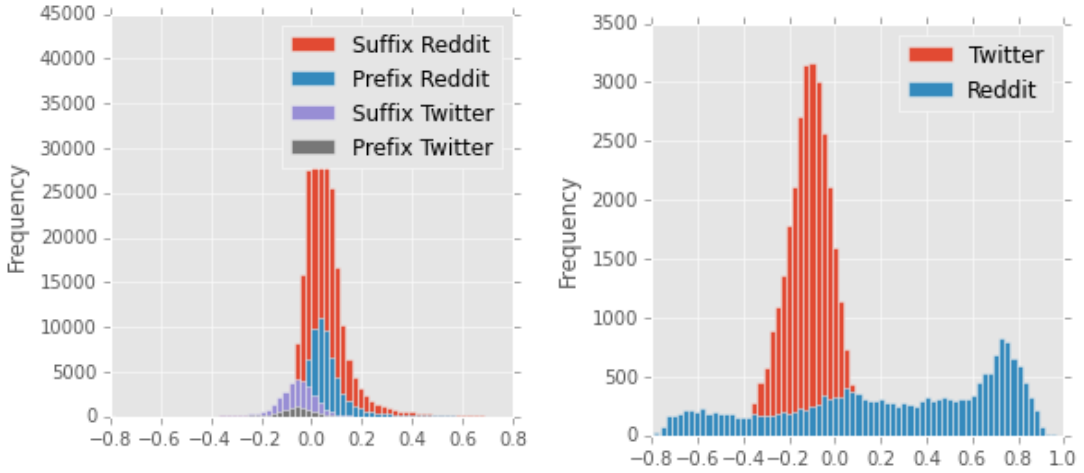
Regional variations can be seen within the Twitter dataset (Table 6.6); for example, the term *'ubered'* can be seen to increase on the global level, and also on the regional level in Greater London. This could be due to Uber[6] operating heavily in the capital city; thus, as the service becomes more dominant, the brand name will enter the vernacular in much the same way as people now use the term 'to Google' when referring the use of a search engine. This is compared to what could be classified as colloquial terms appearing in the region *'Wales (South)'* such as *flook* and *tnx*, which might have no meaning within different regions. However, unlike Reddit, in which these appear to be a correlation with the topic of the subreddit, meaning un-shared terms, we can see numerous terms appearing across the regions, e.g. *'fleeky'* and *'blurface'*.

Insights can be gained by comparing the distributions of user and word frequency Spearman's rank (Figure 6.7.2a and 6.7.2b). Figure 6.7.2a shows that, as the Spearman's rank of word frequency increases, so does the rank of unique user usage. However, when comparing the distributions for Reddit (Figure 6.7.2b), we can see an interesting trend; much like Twitter, there is a linear relationship between both the user and word frequency. However, there is a second trend, whereby low- and medium-frequency words have a higher rank across user usage than for frequency; when we sample from this second trend line, we get words such as '*nutz*', '*darkspaw*', '*sarkozy*', '*fng*', '*oud*' and '*banii*'. However, when we look at the raw time series of each metric, we see that this trend could be explained due to the user frequency increasing at a greater proportion than the overall word frequency, as seen in Figure 6.7.2c.

6.7.2 Variation in Form

Variation in form, unlike variation in frequency, aims to assess varying morphological forms of an innovation, identifying users who are using the innovation in varying forms. Two methods to assess variation in form were proposed: the initial method looked at the probability of prefix or suffix, whereas the second clustered words together into distinct sets through the use of edit distance.

[6]Uber develops, markets and operates the Uber mobile app, which allows consumers with smartphones to submit a trip request, which is then routed to Uber drivers who use their own cars.



(a) Twitter and Reddit, prefix and suffix distribution

(b) Twitter and Reddit edit distance distribution

Figure 6.7.5: Variation in form

The first experiment was run using the probability of prefix or suffix. This meant that innovations such as *'cooldown'* would have been assessed on their transformation into *'cooldowns'* or *'uncooldown'*. As we can see, the distributions (Figure 6.7.5a) show that the Spearman's rank distribution for all suffix and prefix additions in both datasets is normally distributed, with Reddit being positively skewed and Twitter negatively. The differences in frequency between both datasets is due to the size of the innovations lists from the different sampling windows, and the removal of words with no prefix or suffix addition. However, we can see that there is a greater probability of suffix addition. When sampling from the upper confidence interval of each of the four distributions, we get innovations such as *tumblrina* (see Table 6.7) with varying suffix and prefix addition. Similarly, Twitter shows the existence of timely morphological forms of innovations, such as *'jockalypse'* in reference to the Scottish referendum in the UK in May 2015, and also the growing popularity of *vapeing*. However, due to the large sampling period of Reddit, we are able to see longer growth periods, e.g. *tumblrina*.

However, modelling variations by prefix and suffix addition raised a number of issues; for the term *jkt* in Reddit, the term *lijkt* is believed to be a variation; however, it is in fact a Dutch word meaning *'you appear'*. This form of error was seen across all distributions.

The second measure proposed (equation 6.5.10) worked by clustering words to an innovation using edit distance, treating them all as the same word. As stated in section 6.6, this was implemented using lookup tables that were built with an edit distance set to three. One of the issues with using edit distance was that short words were highly likely to end up in large clusters with unrelated words as three edits made a larger difference compared to longer words.

Figure 6.7.5b shows the Spearman's rank distribution for a cluster's edit distance. Unlike the distribution of prefix and suffix, the distribution of edit distance ρ values have a greater resemblance to user

Table 6.7: Example suffix addition

Network	Word	Example
Twitter	fleek	fleeking, fleeks, fleeked
	vape	valpes, vapeing
	bugz	bugzy
Reddit	tumblrina	tumblrinas
	clickbait	clickbaitable, clickbaited, clickbaiter, clickbaiting,clickbaitness, clickbaits, clickbaity
	multireddit	multireddits, multiredditing

Network	Word	Variations
Twitter	meninists	meninist, meninism, feminists, meninists
	dubmash	dubsmash, dubsmashes, dubmash
Reddit	casuals	casuel casul casull
	brutha	brita brotha bruh

Table 6.8: Example edit distance

and frequency distribution (Figure 6.7.2b). For this reason, we apply the same strategy of fitting two Gaussians to the Reddit data.

Table 6.8 shows a sample of high growth innovation clusters from both Reddit and Twitter. We can see innovations that would not have been detected by prefix and suffix addition, such as *'meninists'*, which is derived from *'feminists'*, and is used to mock feminism in most instances.

6.7.3 Convergence in Meaning

The final measure aimed to assess the convergence in meaning of innovations across communities within the OSN, by assessing the similarities in the contexts in which innovations are used. This is achieved by training a `word2vec` model per community per time period, then querying the model with each innovation.

At any point in time, the convergence or divergence of the word is assessed using the Jaquard similarity between contextual words from the `word2vec` model for each community (section 6.6). Following the previous two metrics, we apply the general framework for sampling statistical significant convergence in meaning.

Figure 6.7.7a shows the distribution of values, again, with samples taken from the upper (Fig 6.7.7c) and the lower (Figure 6.7.7c) confidence intervals. Both figures show slight but notable increases and decreases in the assessed metric, indicating a potential convergence or divergence of meaning over time. However, to understand what is going on within the model, we must look at the content containing the source words, e.g. what is being seen might not be a convergence of meaning but rather on content or topic.

Looking at words such as *twitpic* and *ebola*, respectively, we can explain what might actually be seen for some of the innovations. Twitpic was a service[7] used to distribute pictures and images on Twitter. However, the popularity of twitpic decreased over time due to Twitter introducing a similar service. Therefore, as people posted less pictures from the twitpic application, terms such as '*tweeitpic*' and related words were used less in 'similar' situations.

The process of declining popularity could also explain why '*ebola*' is classified as a meaning that is diverging. The convergence in meaning at the beginning of the time series could be put down to the increased media attention and discussion at the beginning of the data collection period in relation to the Ebola outbreak in West Africa. This spike in popularity due to events surrounding the Ebola outbreak can be seen in Figure 6.7.6. However, this could be seen as a convergence in context, as stated in [71], indicating that it is a 'timely' word that could represent the cultural significance at the point in time [176].

Success in the method, however, can be seen by manually inspecting the convergence in meanings

[7]Twitpic was a mobile application and website that allowed users to post picture to Twitter.

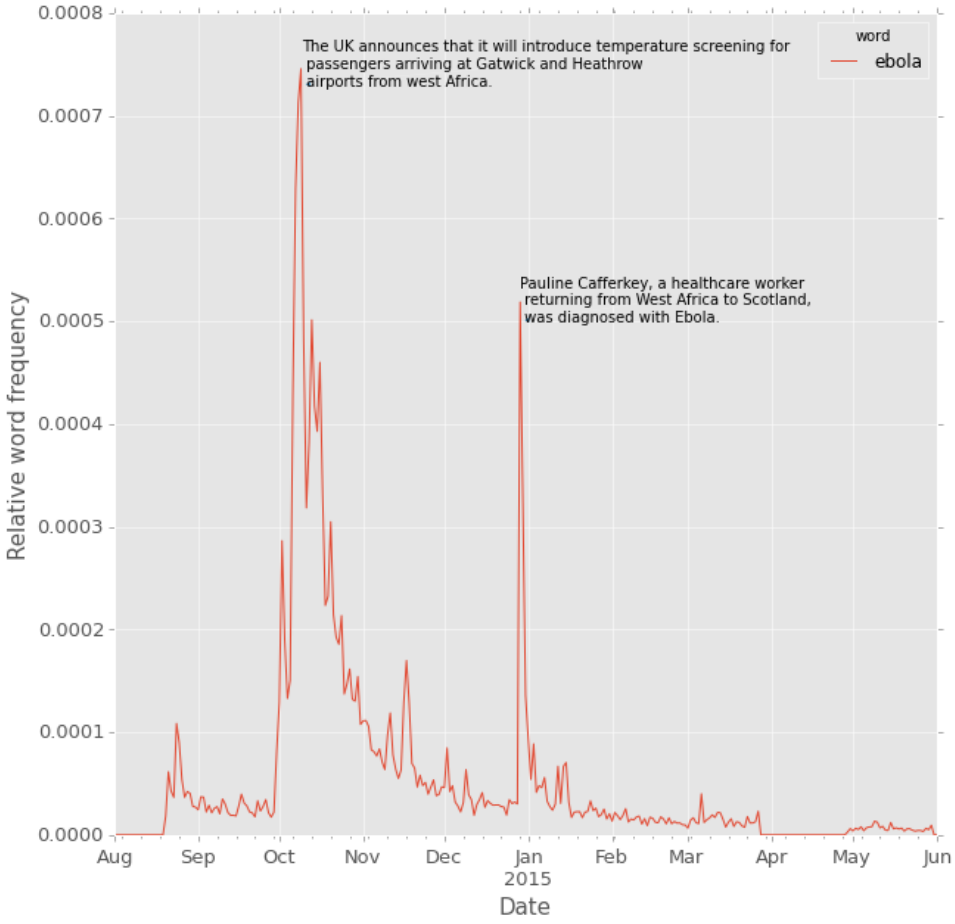
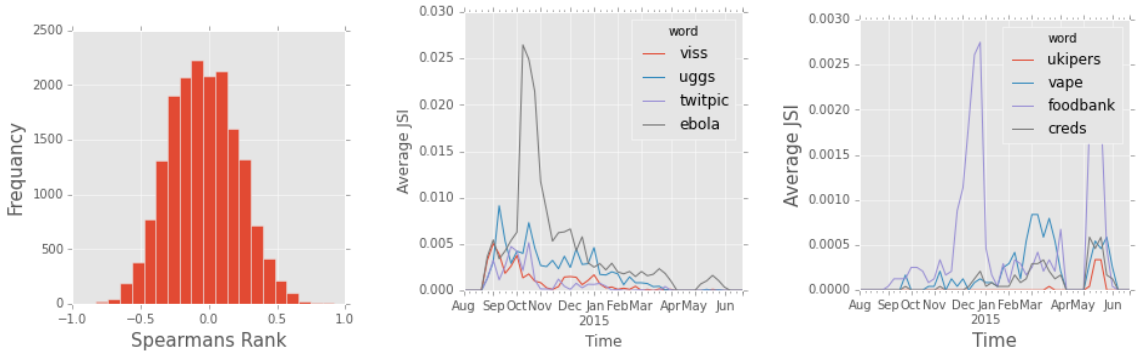


Figure 6.7.6: Word frequency for ‘ebola’ with annotated news events



(a) Spearman's distribution (b) Meaning convergence (c) Meaning divergence

Figure 6.7.7: `word2vec` community meanings

over time. We can see one confirmation of the method in *vape* increasing over time, which, in the offline world, entered people's vocabulary at the beginning of 2015.

6.8 Discussion and Conclusion

In this final section, we discuss the methods used and results achieved in relation to whether they answer the four research questions.

Using statistical methods, sampling techniques and computational models, we have been able to show that language used within online social networks is indeed constantly changing. This change comes from the extended use of innovations over time, such as *'fleeky'*, or the use of time-sensitive words such as *'ebola'*. We found that a better measure than absolute frequency is relative user usage in conjunction with frequency, as this limits the impact of bot accounts, revealing innovations where the unique user usage grew at a greater rate than raw frequency, such as '*nutz*'. These metrics, when utilised with the varying granularities of networks, showed the ability to identify region and community-specific innovation growth. These community-specific innovations were seen in the topic-specific innovations within subreddits and words that appear to be geographically bound on Twitter, e.g. '*ubered*' in Greater London and '*midterm*' in the subreddit '*politics*'.

In addition, by assessing the morphological variation of words (through prefix and suffix addition, and clustering with edit distance), we were able to show the growth of multiple morphological variations over time across both networks, such as '*vape*' and '*clickbate*'. However, issues were encountered from short innovations being only three edits away from other short innovations, and the addition of prefix or suffix causing the word to conflict with a foreign language.

Results in this chapter bare resemblance to those seen in other works which have looked at the emergence of language innovation and/or change in language using OSN data as a source. The results seen in the analysis of variation in frequency (Section 6.7.1) correlate with results is seen in [89], both

identifying the emergence of innovations (OOV) such as *fleek*, *fuckboi* and *rekt* on Twitter. Though, different data sets underpin these two works, where as the work in this thesis focuses on Reddit, and Twitter usage in the UK where as [89] focused purely on language found from Twitter in the USA. Similarity in results indicated a number of interesting points, firstly that the English language across the globe appears to innovate/influence in a similar manner to each other, and the methods developed in both works can be used across geographical boundaries. Additionally, when looking at the variations of the innovations by taking into account its multiple forms (Section 6.7.2) similar results can be seen in [89] who identify growth/emergence of terms as a collections of variants. However, they did not cluster OOV together to create one canonical form, instead they reported each terms growth individually manually combining them in a qualitative evaluation. This meant one could see both the growth of *fuckboi* and *fuckbois* individual and not combined into one metric. Though, [89] did not look at the regional variation in language, though results of such analysis can be see in [101]. [101], instead of applying time series analysis to each region within the USA, [101] looked modelling the variant-preference e.g. quantifying the variation in using *Mom* or *Mum* in different regions. Even though the methods used in [101] were not the same as those in this Chapter, similarity in results can be seen as this Chapter was also able to detect regional variations which could be clustered into distinct lexical regions. Finally when looking at the results representing the convergence in meaning (Section 6.8) a number of similarities seen in [122] and [61]. Though these instead looked at again the variation in meanings e.g. *Mom* vs *Mum* though the methods used word embeddings in much the same way. Results showed similarities to those in this Chapter as a lot of sports teams where identified as they would have similar embeddings as the contexts in which they are used within would be similar.

Even though one of the questions was to generate a simple model, a limitation could have been that the models developed were over-simplistic, e.g. the use of Spearman's rank to assess the increase in innovations. This was due to the measure being suitable for low ranking words appearing at the end of the data collection period, as these achieved a higher rank than some words that increased slowly over time. This could be overcome with the implementation of a sliding window, with which we could assess for seasonal growth and decay of innovations, along with assessing the variations in growth at differing stages in the innovation's life cycle.

When assessing the convergence of meaning, we indeed showed increases and decreases when the time series are ranked using Spearman's ranking. However, upon further analysis of the data, what is being shown might not be truly a convergence of meaning but rather an indicator of other processes (context) happening around the use of an innovation or word. Such detections of context can be seen in the increased discussion of *Ebola*, or the declining usage of services such as *twitpic*. In future work, instead of looking at common similar words per time period, we could identify he similar words using word-net, finding common 'in vocabulary' words to describe the innovation.

6.9 Data Access

All data and code created during this research are openly available from Lancaster University data archive at D. Kershaw, *Towards modelling language innovation acceptance in online social networks - dataset*, http://dx.doi.org/10.17635/lancaster/researchdata/46, 2016. DOI: 10.17635/lancaster/researchdata/46

Chapter 7

Predicting Innovation Adoption

In the previous chapter (6), we focused on detecting language innovation by modelling the growth and death of new and old terms by operationalising a number of heuristics traditionally used in assessing words for inclusion in dictionaries. The results indicated that we can discover new innovations that come into a language; however, the underlying noise in OSN data affects the modelling process, such as false positives from new organisations mass tweeting about a particular topic, e.g. Ebola. The proceeding chapter will move on from looking at the individual innovation to looking at the user dynamics and how we can utilise the diffusions of innovation to model the influence between users and thus predict when a user will adopt an innovation based on the number of exposures they have received.

7.1 Introduction

'I'll brb' and *'how did you vote in the Brexit'* are all examples of words and phrases that can be classified as recent linguistic innovations. There is, therefore, a tension between the need to be expressive by using language innovations and being understood, as highlighted by linguist David Crystal: 'Although many texters enjoy breaking linguistic rules, they also know they need to be understood' [48]. Individuals not only use/develop language innovations to be expressive, but also to aid in the (re)production of community identity. Therefore, community structures are at work that enable and constrain innovation adoption; ultimately, a user and a community must either collectively adopt or reject a new word (language innovation).

This cyclical interaction between users and communities is characterised by the social network to which a user belongs. However, this relationship by no means determines whether a user will adopt an innovation in view of such social structures; at the level of the individual users, some people are more willing to change and be influenced than others. This can be modelled as a user-specific threshold that, when breached, indicates that a user will adopt an innovation [86], [196]. The challenge is how we learn

the influence exerted on a user, and the threshold at which a user adopts a new term.

The purpose of this chapter is to show that influence between users can be determined from mining innovation cascades, ultimately showing how user influence can be used to infer the user language adoption threshold, and to show how the influence between users is dependent on global and local structures of the networks. This is ultimately summarised in the following meta question: *Given the creation of innovation words, to what extent can their adoption be predicted, and to what extent does the structure of a social network influence influence?*. This is a more detailed version of the question for this chapter, which is *'How does network structure influence the diffusions of language innovations?'*.

The contribution of this work is as follows:

- **Modelling influence between users or communities by mining language innovation cascades**: We show the ability to learn influence from mining historical language diffusions from macro (between communities) and micro (between users) interactions.
- **Learning global adoption thresholds**: We show that, at a global level, there is a general language adoption threshold that can be learnt through the use of Receiver operating characteristic (ROC) curves.
- **Language adoption across word forms**: We show that users adopt language with little variation dependent on the innovation's POS tag.
- **Variation of influence based on network structure**: We show that network structures are highly influential at the time at which an innovation is adopted, though less influential in long-run adoption.

The implications of this work are not only confined to the realms of academia, but will impact both the business and security communities.

In a globalised world, language innovations or changes in a broader context pose a number of challenges, be it the alienation of users due to miscommunication, or individuals not understanding region-specific words. Further to this, changes can hinder the ability of foreign language learners to adopt a language due to the changing meaning of words, or can impede collaborative work across cultures ([31]) due to different business jargon. However, understanding which users and communities have greater influence on a language will allow foreign language teachers to pre-empt new words entering a language, or companies to place greater emphasis on communication in the language that has the greatest influence in a given region of a network.

However, there is a dark side to the internet, seen in the prevalence of trolling, hate crime, and online child predators. On Safer Internet 2016, Nicki Morgan, the former Education Secretary in the UK, stated that the challenges of keeping children safe online include understanding the terms that children use in

online communication: 'These are all terms that didn,Äôt exist when I was young, and I suspect I,Äôm not alone in needing them explained'. [1] However, even though the government launched a website [2] detailing the words used and their respective meanings, the true value will come from pre-empting language change and generating a dynamic list of current and future terms, allowing parents to stay one step ahead of their children.

The remainder of this chapter is organised as follows. Section 7.2 highlights the state of the art. Section 7.3 introduces the datasets. An explanation of the construction of networks is presented in section 7.3.1, and what we class as an innovation is discussed in section 7.3.2. The measures and methods applied are explained in section 7.4, with 7.6 summarising the results. Section 7.7 outlines the final contribution of this work and possible future directions.

7.2 Related Work

Research into language and Online Social Network (OSN) has attracted studies across a diverse set of academic disciplines. In this section, we will focus on user influence, information diffusion, the effects of social structures and language change within OSN.

Much debate has been had surrounding the nature of language and the prevalence of 'bad' language that has developed in OSN. From a technical point of view, the excessive variation reduces the accuracy of traditional NLP tools such as POS taggers and NER systems [14]. However, [62] states that normalisation strategies on social media data removes value from text, as the language that people use contains information about the author and their community. Variations in the language used being are used to predict user features such as age and gender [172], location of user [95] and social habits [111].

Language is intrinsically connected to geographical locations, leading to the ability to predict a user's location from text [95]. However, over time, as users interact and move around the landscape language, innovations will ultimately diffuse. [61] showed that, through the use of `logit` transformations and relative ratios between specific locations to global frequencies, we are able to sample location-specific terms such as *'bogus'* in the upper-east of the US, and *'lbvs'* in the upper-mid-west in the US. Studying the geographical variates in language was taken further by modelling the diffusion process of new words; [64] used stochastic modelling to infer the diffusion network of new words across the US. The results showed that language moves between cities with similar demographics and size; however, similar results were also seen when studying user movement patterns in Flickr [19], which attributed the user patterns to the US air network. However, this could be attributed to the small number of innovations tracked and their potential correlation to a highly mobile college demographic.

Influence is traditionally defined as 'getting people to change their attitudes and behaviours' [104].

[1] Nicky Morgan: We simply can't know everything our children are doing online - 09/02/2016
[2] Online teen speak - Parent Info

[86] proposed that each user has an adoption threshold. When the collective influence of the network breaches the threshold, the user then changes their behaviour. However, [196] stated that it is only those in the immediate neighbourhood of a user who influence actions, not the network as a whole. In addition, each user's threshold (at which they change) varies as people make different active dissensions to adopt change; a low threshold suggests a user adapts to change quicker, whereas those with a higher threshold are more conservative.

Building on the user threshold models proposed by [196] and [86], [83] proposed that user *influence* can be modelled as a function of past action propagations (tagging the same photo on Flickr), with the influence then decaying over time. Even though the results achieved high accuracies in predicting user actions, this was only tested on one limited dataset. Influence of a community on a single user can be seen in [53], which showed that users adapt their language to that of community as they join. This effect can also be used to predict when a user is going to leave a community as their language diverges away from the global language model. However, this research was performed only on a small specific 'beer community', where there would be a naturally higher convergence as users used more 'technical' terms.

However, it is not only predicting who will adopt an action but also how many people will adopt the same action. By modelling the diffusion of memes, [205] identified that, for a meme to spread, it is not only the initial popularity of the content that is important (as stated in [185]) but also who initially uses a meme, with initial users needing to have a set of diverse topics and interest. [207] proposed that diffusions are highly dependent on the network structure. By comparing simulated and real-world diffusion, they identified the effects of *homophily* and *social influence*. However, they did not differentiate between the two effects, which have been shown to auto-correlate. [9] proposed through matched pair sampling that we are able to distinguish between *homophily* and *social influence*, showing that homophily accounted for 50 % of the persuasive behavioural contagion.

Drawing on the related work, this chapter shows how users and communities influence each other's language, as well as the effect of a network's structure on inter-user influence. This is achieved by applying a known influence framework across multiple network abstractions.

7.3 Methods

One of the limitations of previous research is the reliance on one dataset/social network. Therefore, in this work, we draw on two distinct networks: Reddit and Twitter. Even though both social networks are highly popular[3][4], they both can be conceptualised in different ways. Twitter is a personal broadcast network that allows users to express messages and emotion without necessarily getting a response. Alternatively, Reddit is content-focused and structured into self-governing *subreddits* that have particular

[3] PewResearCenter - 6% of online adults are Reddit users
[4] PewResearCenter - Mobile messaging and social media 2015

Table 7.1: Dataset description

	Reddit	Twitter
Unique Words	2,942,555	526,342
Posts	1,054,976,755	111,067,539
Innovations	2,712,629	373,217
Days in Dataset	880	283

topics which draw user attention and comments.

Twitter has been used extensively in academic research due to its relative widespread adoption and the availability of a publicly accessible API that provides up to a 10 % sample from the global firehose. For this study, we bound the results returned from Twitter to those having originated from within the UK, thus limiting the sample to tweets that only contained GPS coordinates within the UK. Even though studies have shown that only 4 % of tweets contain GPS tags[5], the sample that was collected from September 2014 to June 2015 contained 111 million tweets. The second data source is an 18-month dump (January 2013 to June 2015) of comment posts from Reddit. This data comes from a larger data release that spans the entire existence of Reddit from conception in 2007 through to mid 2014.[6]

7.3.1 Networks

One of the issues that arises when studying OSN is the need to infer network structures; whereas on Facebook and Twitter, we can infer friendship through a user being a 'friend' or a reciprocal following, this information is challenging to collect and is guarded by the respective companies. Thus, the research has extracted relationships between users from retweets [161] and mentions [203]. For this work, we aim to learn the influence between users or communities through two abstractions of networks from each data source, one representing the interactions between users (*micro*), with the second modelling interactions between communities (*macro*); this allows us to contrast different concepts of influence.

Ultimately, the networks will take the form of a *directed social graph*, where the graph (G) is defined as a quad $G = (V, E, T, W)$, containing vertices $v, u \in V$ and edges between the vertices $(v, u) \in E$, denoting an outward connection from v to u. The quad also includes the time (T) when the edge was created, while $w \in W$ denotes the weight of a given edge. Edges are only added over time and never removed, and there are also no self-looping edges. For greater detail on the implementation of the networks, please refer to section 5.2. The remainder of this section will give an in-depth review of the networks used.

[5] PewResearCenter - Location-Based Services
[6] Data-set available on the Internet Archive

Table 7.2: Network description

Network	Nodes	Edges	Power Law	Communities
Twitter Geo	2,910	436.49	3.398	14
Twitter Mention	283,755	329,440	3.004	39,767
Reddit Comment	861,955	2,402,202	4.134	36,885
Reddit Subreddit	15,457	142,285	1.511	407

Micro

At the micro level, we will model the graphs through user interactions. Within Twitter, users interact with each other in a number of ways. However, the predominant form is by mentioning fellow users in tweets (through the inclusion of the '@' symbol and a username). We use this to build a user-to-user graph, where a relationship from user $v \rightarrow u$ is inferred if u *mentions* user v; the edge time ($t_{u,v}$) is when this interaction first happens, with the weight ($w_{u,v}$) being the total number of times u mentions v. Both users u and v must also exist within the dataset.

Similarly, within Reddit, users comment on each other's posts, forming a chain of interactions. Thus, we define a relationship between users if user u comments on a post of user v, thus forming an edge $u \rightarrow v$. The time ($t_{u,v}$) of the edge would be the first time this happened, and the weight ($w_{u,v}$) would be the number of times user u commented on a post of v.

Macro

Even though users may not comment or interact with each other, this does not mean that they are not exposed to each other's information by observing the network; [186] showed that the majority of users in OSN lurk and only interact sporadically. Collectively, a group of users may also exert influence over other collections of users; for this reason, we cluster together content (posts and tweets) generated within the same communities (subreddits or postcodes), and generate an edge between these nodes by extracting the users traversing across the network *between* the nodes.

Similarly, within Reddit, users interact and move around different subreddits depending on their current interests or in reaction to popular content. A similar method to that detailed above can be applied to extract the interaction between subreddits. The weight between two nodes is the number of users moving consecutively from one subreddit to the next, with the associated edge time being the first time a user first moved between the two.

Graph Filtering

However, not all edges are significant, as a user who has been mentioned once is not as important as a user who has been mentioned 100 times. For this reason, we extract the backbone network by filtering edges that are not statistically significant using the backbone extraction algorithm [156]. This also makes

the processing of each network easier as it removes sparse edges, reducing the number of connections from which influence can come. On top of filter edges, we apply a Louvain modularity community detection [54] to each of the four networks to identify sets of nodes that exist structurally within the same group. We use the identified communities of nodes to later assess influence between communities in on each others language adoption.

7.3.2 Innovations

The premise of this work is to predict the adoption of innovations; thus, we must first classify what is and what is not a language innovation. For this work, we will be stating that an innovation in language is a word that does not appear within the British National Corpus (BNC) [11]. The BNC was chosen as the baseline for British language as it is the most comprehensive study of British English in recent times, taking its sources not only from books, but also newspapers, written communication and oral discourse transcripts. Though, this research is only interested in the emergence and diffusion of new innovations; therefore, only innovations that appeared after the first month of data collection are used. However, on initial manual inspection, a large number of innovations were used only by one user (predominantly bot accounts); for this research, innovations had to appear over 10 times and be used by more than 10 users.

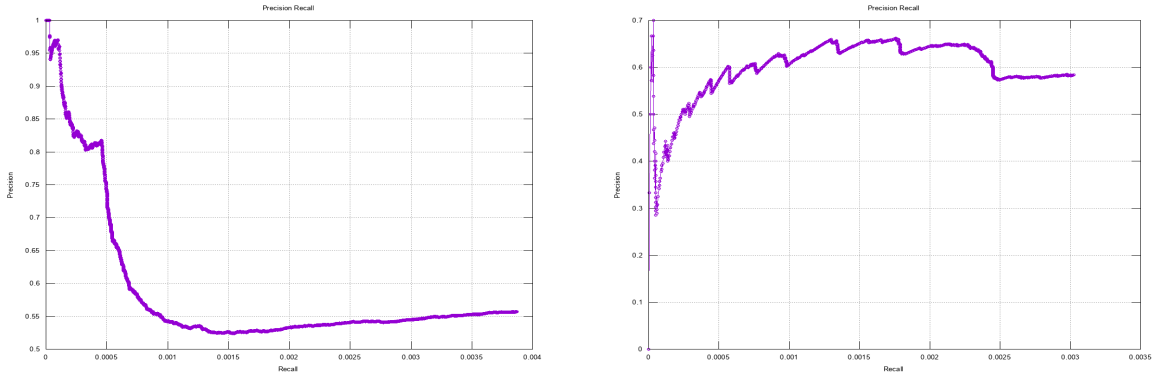
The function of words in communication varies; for this reason, we break the analysis down into distinct classes of words. This is achieved by initially POS tagging each dataset, but each innovation being assessed can have multiple classes. For simplicity, the class assigned to each innovation is the class with the highest count. This is implemented using TwitterNLP [76], due to its ability to deal with noisy social media data.

7.3.3 Comparative Evaluation of Datasets

As has been highlighted in the literature review (Section 3.3) social media data, as used in this thesis, has been used to model many other forms of information diffusion; be this from tracking memes across Twitter in [208] to News items as headlines over time in [135]. To understand the quality of data we are using to model the influence and diffusion's of innovations (OOV) between nodes in the network, we are going to run a sample of the data-set through an existing model in an attempt to see if similar results can be achieved.

[78] showed that one could extract network structure though monitoring the diffusion of content by modelling its cascade through a probabilistic model. The method developed works by first defining a contagion transition model which describe how likely two nodes are to infect each other in sequence, which is then used to build a model which describes if the cascade (the complete diffusion of one item of information) will follow a particular cascade tree pattern. This final tree model is then used to estimate a (near-)maximum likelihood network which is the network which maximises for a tree to accrue.

We run the code released with the paper[7] on a subset of the macro Reddit and Twitter data sets. This means that within the sample of data there are 2,911 nodes and 6,819 unique language innovation cascades for Twitter and 4,007 nodes and 3,431 unique language innovation cascades for Reddit. They are then formatted into the input format required for the code base. The parameters chosen where $\alpha = 1.0$, $\rho = 10^{-9}$ and $\beta = 0.5$ this is the same as the parameters using in [78].



(a) Twitter-Geo Network

(b) Reddit Traversal Network

Figure 7.3.1: Precision recall plots when running a sample of cascades through `netinf` to extract the actual network structure.

Figure 7.3.1 shows the pay-off between precession and recall as the value of k varies. k is the computed maximum likelihood between two nodes within the extracted graph. The data set which was used in [78] was not released, this means it is challenging to compare the results. However, the reported break-even point though was with a break-even point of 0.44, this was a 30 % improvement over their pre-computed base line.

As one can see recall on both models train on our data the recall is low, this can be attributed to only a sample of all cascades being used which does not contain all the nodes in each of the network. Thus the model was not able to extract the edges for nodes which where not observed in the cascades. However for the edges it was able to extract the precision was high. Focusing on the Twitter-geo network network, `netinf` was able to extract with a high procession a small amount of the network, though this deteriorates as more of the network was extracted. Alternatively, the Reddit traversal network though maintained a high precision as the recall increased, this would indicate that as more of the cascades when consumed by the model more of the network could be extracted at a higher quality. This difference in results between the two network types comes down to the quality of the underlying data sets and how they were constructed, Reddit traversal is representative of all movements between sub-reddits, where as Twitter-geo is contracted from a sample of data which could have biases in distibution (see Section 4.4 for more information on data sets).

Even though there are variation in results, in addition to low recalls, this indicates that cascades

[7] https://github.com/snap-stanford/snap

follow a pattern which is correlation to the network structure. This means that one can extract the network from the diffusion of content. For this Chapter it means that for this chapter one should be able to extract a global threshold at which people use an innovation as the diffusion of innovations is connected to the network structure.

7.4 Operationalisation

People accommodate their language to that of the people around them; thus, this research aims to predict when people adopt new terms in response to exposure from their neighbours. We propose that, as shown in [196] and [86], there is a mean general threshold across a population (σ), which represents the '*joint influence*' at which individuals within a network then adopt an innovation. We, therefore, use the framework proposed by [83] to model influence between users as a function of previous join actions (propagations), which are then used to compute the pressure applied to a user to adopt an innovation.

[83] state that influence between users $i_{v,u}$ (influence of v on u) can be learnt as a function of previous joint actions that have propagated between the two users (from v to u). After learning the influence, we can use the joint probability across all active neighbours of user u to express the current joint influence (i_u) on u to adopt action a. To predict if a user is going to adopt an action a, the joint influence i_u will need to be higher than the threshold σ, such that $i_u > \sigma$; if they adopt with a value less than the threshold, this will be classified as a true negative.

This breaks the analysis down into two distinct stages: learning the influence between two users (section 7.4.1) and then predicting user adoption of terms (section 7.4.2).

As discussed in section 7.3, we define a social graph (G) as a set of directed edges ($(v, u) \in E$), each of which has a time stamp (t) of when the edge was formed. Thus, to learn the influence between users ($i_{u,v}$, where $i_{u,v} \in [0, 1)$), we first define a number of basic measures: O_v and O_u are the number of distinct innovations used by users v and u respectively; alternatively, $O_{v|u}$ is the number of distinct innovations v or u have used (i.e. the union of their vocabulary). The number of propagations of innovations between users is defined as O_{v2u}, which is the number of innovations that were first used by v and then by u, thus $t_v(o) < t_u(o)$ (with function $t_v(o)$ and $t_u(o)$ returning the time that the innovation was used by each user). However, propagation cannot occur if an edge between users has not yet been created, meaning that propagation must also fulfil the following $e_t(v, u) < t_v(o) < t_u(o)$, where $e_t(v, u)$ returns the creation time of the directed edge from v to u.

7.4.1 Learning Influence

We now define four measures that quantify the influence ($i_{v,u}$) of user v on u. Each measure is based on the values above, aiming to quantify influence in different ways, as a proportion of different aspects of

historic diffusions.

Bernoulli

We first state that influence is proportional to the fraction of innovations that have propagated from user v to u as a fraction of *all* the innovations that v has used:

$$p_{v,u} = \frac{O_{v2u}}{O_v} \tag{7.4.1}$$

Thus, if all the innovations that v used end up being used by u, then the value will be 1.

Jaccard

Alternatively, influence is proportional to the number of innovations that have propagated (O_{v2u}) out of the union of all innovations used across the two users ($O_{v|u}$):

$$p_{v,u} = \frac{O_{v2u}}{O_{v|u}} \tag{7.4.2}$$

This means that, if all of v's innovations propagate but u uses a large amount of other innovations as well, then the value will be lower than that computed in equation 7.4.1.

Partial Credits

However, when users adopt a new term, it could be said that each of their neighbours (who have used that innovation before) all have an equal part to play in the user adopting a new term; therefore, it could be said that they share equal credit:

$$credit_{v,u}(o) = \frac{1}{\sum_{w \in S} I(t_w(o) < t_u(o))} \tag{7.4.3}$$

Where S is a list of activated neighbours of v (e.g. users connected to v who have adopted the innovation before), with the function I acting as an indicator function that returns 1 if the neighbour w has used the innovation before u.

We then modify equations 7.4.1 and 7.4.2 to incorporate the partial credit definition (equation 7.4.3). Instead of influence being defined as the number of propagations, it is instead defined as the average credit per innovation used by v or:

$$p_{v,u} = \frac{\sum_{o \in O} credit_{v,u}(o)}{O_v} \tag{7.4.4}$$

Or the average credit used across the union of all innovations used across users v and u:

$$p_{v,u} = \frac{\sum_{o \in O} credit_{v,u}(o)}{O_{u|v}} \tag{7.4.5}$$

7.4.2 Computing Joint Influence

The measures in equations 7.4.1, 7.4.2, 7.4.4 and 7.4.5 aim to quantify the influence between users. These metrics can be used to predict user adoption of new terms by computing the joint influence exerted on the user by active neighbours.

The joint probability ($i_u(S)$) can be computed by utilising the *monotonic* and *sub-modular* nature of the influence probabilities:

$$i_u(S) = 1 - \prod_{v \in S} (1 - i_{v,u}) \tag{7.4.6}$$

Where S is the set of active neighbours of node u.

However, the influence that a user exerts might not be constant, with the influence decreasing over time after they themselves have adopted the innovation. A reduction in influence between users may be due to a number of reasons, from the Tweets dispersing from a user's Twitter timeline, or users not coming into contact again on Reddit.

Therefore, we attempt to model the decay of influence between two users as a function of the average time of propagation ($\tau_{v,u}$). The decay takes two forms: a *discrete* form, where the influence stays consent for a set amount of time, or a *continuous* form, where influence decay happens exponentially.

First, we define the average propagation time of an innovation between two users v and u. This is defined as $\tau_{v,u}$:

$$\tau_{v,u} = \frac{\sum_{o \in O_{v,u}} (t_u(o) - t_v(o))}{O_{v2u}} \tag{7.4.7}$$

With $O_{v,u}$ being the set of innovations that have propagated from v to u, and $t_u(o)$ being the time that u adopted o. As before, O_{v2u} is the number of actions propagating from v to u.

To model the decay of influence in a basic form, we allow a user to have influence over another only for the length of $\tau_{v,u}$. This is to say that after a user u is exposed to an innovation by v, the influence will reduce to 0 once $\tau_{v,u}$ has elapsed; thus, the influence window is $[t_v, t_v + \tau_{v,u}]$. At the point when a user's influence reduces, the joint probability (equation 7.4.6) can be updated with the following equation:

$$i_u(S, w) = \frac{i_u(S) - i_{w,u}}{1 - i_{w,u}} \tag{7.4.8}$$

Where S is the set of active nodes before w becomes inactive, and w is the node that has become

inactive. This means that the whole probability does not have to be recomputed at the search step, rather just updated.

In reality, influence does not just vanish, rather it diminishes over time. Therefore, for the final variation, instead of the influence being fixed for a set amount of time, it decays exponentially:

$$i_{v,u}^{t} = i_{v,u}^{0} e^{\frac{-(t-t_v)}{\tau_{v,u}}} \tag{7.4.9}$$

With $i_{v,u}^{t}$ being the influence from user v on u at time t. Thus, the maximum influence will be when $t = 0$. Therefore, the new joint probability function is:

$$i_{u}^{t}(S) = 1 - \prod_{v \in S} (1 - i_{v,u}^{t}) \tag{7.4.10}$$

As stated at the beginning of this section, the aim is to first learn the influence probabilities, then to use these learnt values to infer the current global threshold, which represents the mean threshold that needs to be breach for a user to adopt an innovation. To achieve this, we use 80 % of the data to train the model, with the remaining 20 % used to test the ability to predict innovation adoption. As mentioned above, each user must breach the global threshold $i_u(o) > \sigma$ for the user to adopt the innovation. To infer this individual threshold, we use ROC curves to determine the optimum trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR), as users may have been exposed to an innovation and not adopted it.

7.4.3 Measuring Network Effect

We want to assess the extent to which the network structure affects users' adoption of new language. To assess the effect of local network structures and the timing of innovation propagation, the underlying network will be randomised. Randomising the network will allow us to quantify the pressures to adopt an innovation that come from sources external to the network (section 7.4.3).

Secondly, we measure the influence of densely connected regions of the network on the language adopted by users. These densely connected regions can be seen to represent communities of users who communicate to a greater extent with each other; we want to see if users internal to these densely connected regions exert more influence than users who are external to the cluster (section 7.4.3). This will distinguish between inter- and intra-community effects.

Random Network

To understand the effect of network structures, we shuffle the four networks, with the aim of randomising the edges, along with the edge times. However, social networks are defined by their degree of distribution; therefore, even though we shuffle the edges, we aim to maintain the degree of distribution. As proposed

in [197], instead of shuffling the edges, we iterate over each edge, randomly selecting an alternative edge and swapping the source of each edge, thus maintaining the degree of distribution. With these four new graphs, we then *relearn* the influence measures across the new networks, and make new predictions.

We expect there to be a significant reduction in accuracy in the ability to learn the global threshold (σ); this should be seen through a reduction in AUC. However, we do not believe there to be a complete reduction, with the remaining accuracy representing the influence external to the network that cannot be quantified through mining data from the social networks.

Community Influence

In social graphs, users and communities cluster together; thus, we aim to see if influence between nodes is greater *internally* or *externally* to these communities (collection of densely connected nodes). As each network has been classified into distinct communities by [22], this means that intra-community edges ($E_{\curvearrowleft}$) are those that cross community boundaries, whereas inter edges ($E_{\circlearrowleft}$) are edges within the same community.

To compare the influence that exists internal and external to communities, we compute the average intra influence ($\overline{i \curvearrowleft}$) and average inter-community influence ($\overline{i \circlearrowleft}$) across the networks.

$$\overline{i \circlearrowleft} = \frac{\sum_{e \in E\circlearrowleft} w(e)}{|E \circlearrowleft|} \tag{7.4.11}$$

Where $w(e)$ returns the influence of the given edge (e); this is divided by the number of internal edges ($|E_c \circlearrowleft|$).

Similarly, computing the intra-community influence sums the influence of all external edges, then divided by the number of external edges.

$$\overline{i \curvearrowleft} = \frac{\sum_{e \in E_c\curvearrowleft} w(e)}{|E_c \curvearrowleft|} \tag{7.4.12}$$

If the community structure has limited effect on the influence internal or external to a community, we would expect there to be limited difference between $\overline{i \curvearrowleft}$ and $\overline{i \circlearrowleft}$. Whereas a larger inter value would indicate that the community structure is having an effect on limiting the distribution of influence, with influence being affected by concepts such as structural trapping [207]. If the internal influence was smaller than the external, then the nodes within access to the *structural holes* would have a disproportionate influence on the diffusions of innovations, confirming, as postulated by [148], that structural holes are highly influential in language change.

7.5 Computational Methods

One of the inherent challenges within the nature of this work is the challenge of processing and analysing large/big datasets. Thus, to implement the systems that have been introduced highlighted in Sections 4 and 5, a custom solution is developed based around Apache Spark for processing the data (Section 7.5), and Apache HBase for storage of intermediate results (Section 7.5).

Storage

One of the challenges in working with network data is the need to store the data in an accessible manner, but also deal with the large number of edges and nodes. The Reddit comment network has 861,955 nodes and 2,402,202 edges, which, when stored in an in-memory data structure such as a dense matrix, can limit the performance, especially when additional matrices are needed for edge values such as weight and time.

As stated in section 5.2, each social network can be represented as an adjacency matrix A, which is a $n \times n$ matrix, where column i and row j represent the directed relationship (outward edge) from $i \rightarrow j$, with $a_{i,j} \neq a_{j,i}$ representing a directed relationship. A matrix can also be used to represent additional values between nodes, such as the time at which relationships were created ($\tau_{i,j} \in T$), and the weight of the relationship ($w_{i,j} \in W$). Additionally, the majority of the computed values introduced in section 7.4 can also be represented in the form of matrices, such as O_{v2u}, where the row would be v and the column u.

For this reason, Apache HBase[8] is used to store the respective matrices in a permanent and queryable format that can be accessed from a number of machines. HBase itself is a distributed NoSQL database designed for storage of large/big sparse datasets, which can be represented in the form of a table. HBase allows data to be accessed through row keys, with data being stored in columns. Additionally, columns can be grouped into families of columns through the use of a `column family` key. This allows the same column key to be used across multiple data points in the same row.

HBase, thus, can be used to store each matrix in a scalable and accessible manner. For the adjacency matrix A, the row and column values are stored in the rows and columns within HBase. Additionally, each matrix has identical dimensions $n \times n$; thus, the same columns and rows. Therefore, instead of creating multiple HBase tables, we can store each within a different column family. By storing values across different column families, instead of having to perform multiple `get` requests across multiple tables, one `get` is performed that results in multiple column families.

However, one of the limitations of using HBase is that cells can only be accessed through row keys. This is ideal for undirected networks in an adjacency matrix $a_{i,j} = a_{j,i}$. However, for directed networks $a_{i,j} \neq a_{j,i}$, the cell represents direction from $i \rightarrow j$. The network stored in HBase is directed, with

[8] https://hbase.apache.org

Column Family	v_{in}			A_{v2u}				A_u				A_{voru}		
Node	gif	funny	pic	gif	funny	pic	game	gif	funny	pic	game	gif	funny	pic
gif				1			4	3				4		4
funny					1	2			2			4		4
pic					4	4				3			4	

Table 7.3: Example HBase table

the row values representing the outward edges for the node (with the node represented as the row key). Inward edges for a node are then represented by accessing the columns with the same key. However, accessing values (the inward edges) through row keys introduces excessive overhead, as we cannot simply access a column, but rather have to iterate over each row, selecting the column value that is of interest. This is computationally expensive as it requires a `get` request on each row. To reduce the complexity and speed up the time taken to query the data, one main optimisation is implemented, by storing two copies of the matrix, one being the transposition of the other, all within the same table, though separated into different column families. This means we can access both inward and outward edges through the use of the `row key`, speeding up access time.

As before, to limit the number of `get` requests on the table, the transposition of the table is stored in its own column family. This, however, does mean that each matrix is stored in the HBase twice, though this is a negligible increase in storage costs compared to the performance increase achieved.

Processing

As stated earlier, Apache Spark allows the computational tasks to be distributed across a number of machines in a scalable and parallel manner. Scaling an application is achieved by breaking a process down into tasks that can be performed in parallel on each item of data. As stated, the data that is being processed is large in nature and cannot easily be storage or processed on one machine; therefore, a big data approach to this research is taken in distributing the processing of the data.

The aim of the proposed model (section 7.4) is to learn the four versions of influence across the three variants of time. Each measure of influence is comprised of the same set of values that is computed from each innovation diffusion across the given network. Therefore, to compute the raw values and learn the influence between nodes in a scalable manner, the method proposed in the original paper [83] is adapted to utilise Apache Spark and HBase. As with the original paper [83], processing is split into two phases: phase one (algorithm 7.1) learns the raw values O_{v2u}, O_u and $O_{o\&u}$, and phase 2 (algorithm 7.2) computes the respective influence values based on the values from phase one.

The values in phase one can be learnt in a `map-reduce` format, with each `mapper` taking as the input a complete diffusion of an innovation in the form of a time-ordered list of username and time pairs. The majority of the learnt values can be seen as incremental counts across diffusions (section 7.4.1). Therefore, to implement the proposed framework in a `map-reduce` pattern one can `emit` a `<key,value>`

each time a value should be incremented. Thus, for O_{v2u}, when v uses o before u, the key value pair $< O_{v2u}, 1 >$ will be emitted; then, in the `reduce` phase, the values are summed together across different innovation diffusions. The output of the reduce phase is then loaded into HBase, allowing the values to be accessed in phase two.

The majority of measures are incremental whole numbers; however, measures such as $\tau_{v,u}$ (the average time of adoption between u and v) are not. This require averages to be computed, with the results in large floating point numbers, which HBase finds challenging to store. To circumvent the inability to store floating point numbers in HBase, the array of values (time differences) is stored, and then the average is computed when the cell is accessed.

We now walk through how the values in phase one are learnt using Apache Spark (algorithm 7.1). Multiple mappers run across multiple nodes in a cluster, each of which has access to the social network stored within HBase (line 2), and takes a complete diffusion as an input. Line 3 iterates over the ordered list for a given diffusion, checking to see if v performed the action before its neighbour u (line 6). If so, then, the relationship existed between them before as they both used the innovation (line 10), the following three lines then emit the values to increment O_{v2u} and $\tau_{v,u}$. Line 13 then emits the value that $O_{v\&u}$ incremented, which represents the credit that v has in u adopting the innovation (o). Line 10 adds node v to a list of parents, which is used in lines 15 to 17 to compute the partial credit associated with u adopting o.

Phase two (algorithm 7.2) is similar to phase one (algorithm 7.1), but focuses on computing the influence values that are dependent on the window of average diffusion time $\tau_{v,u}$, such as the partial credit $credit^{\tau_{v,u}}_{v,u}$ and the count of cascades of innovation between two users A_{v2u}. Similar to phase one, aspects of phase two can be thought of as being incremental counts, so can be summed within the `reducer`. Line 9 shows that O_{v2u} is learnt for a second time; whereas, for the value learnt in phase one (algorithm 7.1, line 8) counted any successive usage. The value learnt in phase two is dependent on a window of average diffusion between the two users ($\tau_{u,v}$), which is partially learnt in phase one (algorithm 7.1, line 9), and accessed from HBase. Lines 14 to 21 additionally recompute the partial credit that is associated with each parent, dependent on the average diffusion time between the parent and the node ($credit^{\tau_{v,u}}_{v,u}$). Again, the reduce class then sums the respective metrics values together, outputting the resulting quadrupedal to be stored in HBase.

The two phases (algorithm 7.1 and 7.2) learn the values needed to now compute the influence between users, based on 80% of the dataset. The evaluation stage now takes these raw values for each of the three time forms, *static*, *continuous* and *discreet* (equations 7.4.9, 7.4.9 and 7.4.9), across the four variants of influence between users (equations 7.4.1, 7.4.2, 7.4.5, and 7.4.4).

As with learning the influence values between nodes, the evaluation phase is implemented as a number of Spark jobs. This allows the collective pressure on users to be computed across a cluster of machines,

Algorithm 7.1: Learning Phase 1

Data: Each mapper receives a complete innovation diffusion in the [userid, time], with the docid being the name of the diffusion

Result: List of quads that represents (row, column family, column, value), which then can be loaded into HBase

```
1  Class Mapper(docid a, diffusion d)
2      n ← network
3      for row u, t_u ∈ d do
4          parents ← []
5          for row v, t_v ∈ parents do
6              if t_v < t_u then
7                  if e_{u,v} ∈ n then
8                      EMIT(metric O_{v2u}, 1)
9                      EMIT(metric τ_{v,u}, t_u − t_v)
10                     parents + v
11                 end
12             end
13             EMIT(metric O_{v&u}, 1)
14         end
15         for p ∈ parents do
16             EMIT(metric credit_{v,u}, (1/len(parents)))
17         end
18     end
19 Class Reducer(metric m, counts [c_1, c_2, ...])
20     EMIT(metric m, count sum([c_1, c_2, ...]))
```

thus speeding up the application. As before, the input into each job is the complete diffusion in the form of a list of tuppels, again consisting of `<user, time>` pairs, representing an innovation diffusion. The basic premise of the system is to compute the join influence on a user who has been exposed to the adoption of an innovation (o); thus, as user u adopts an innovation, the join influence is updated on all the neighbours $S(u)$.

Again, the input of each `mapper` is a complete diffusion in the form of an ordered list of tuppels containing the username and time of adoption. Line 4 iterates through the diffusion, initially checking to see if the user has been exposed to the innovation before; if so, the user is set to have adopted the innovation (line 6), and if not, the user has then innovated with the innovation (line 8). Then, each neighbour $S_in(n)$ is iterated again, and their respective joint influences are set or updated (line 10 to 18).

The three states that a node can be in are:

Innovator They used an innovation with no exposure to the innovation before

Performed They used the innovation after exposure from their local network

Algorithm 7.2: Learning Phase 2

Data: Each mapper receives a complete innovation diffusion in the [userid, time], with the docid being the name of the diffusion

Result: List of quads that represents (row, column family, column, value), which then can be loaded into HBase

```
1  Class Mapper(docid a, diffusion d)
2      n ← network
3      current ← []
4      for row u, t_u ∈ d do
5          parents ← []
6          for c ∈ current do
7              if e_{u,v} ∈ n then
8                  if 0 < t_u − t_v < τ_{u,v} then
9                      EMIT(metric O_{v2u}, count 1)
10                     parents + v
11                 end
12             end
13         end
14         for p ∈ parents do
15             s ← []
16             for p ∈ parents do
17                 if v < u&t_u − t_v < τ_{v,u} then
18                     s + p
19                 end
20             end
21             EMIT(metric credit_{v,u}^{τ_{v,u}}, count 1/len(s))
22         end
23         current + u
24     end
1  Class Reducer(metric m, counts [c_1, c_2, ...])
2      EMIT(metric m, count sum([c_1, c_2, ...]))
```

Not Performed They did **not** use an innovation even though they had been exposed to it before

However, this implementation is limited to the static time model; therefore, a number of modifications need to be made to compute the continuous (equation 7.4.9) and discrete (equation 7.4.8) time models. This requires the re-computation of the joint probability upon each exposure; however, to achieve this, each node must remember its historic exposures. Therefore, upon further exposures, the influence has to be updated in respect of τ and each of its neighbours.

The output of each time model is in the form of a CSV file containing data on whether the user has adopted an innovation, the largest join influence exerted on the user, the name of the node, and the innovation itself. These are not stored in HBase as they are no longer needed in a queryable form and are at a size at which they can be loaded into a Python script to compute the ROC to determine the

Algorithm 7.3: Evaluating - Static Time

Data: Each mapper receives a complete innovation diffusion in the [userid, time], with the docid being the name of the diffusion.

Result: List of quads that represents (row, column family, column, value), which then can be loaded into HBase.

```
1  Class Mapper(docid a, diffusion d)
2      n ← network
3      results_table ← [(user, messure, value, performed)]
4      for row u, t_u ∈ d do
5          if v ∈ table then
6              results_table[u] ← perfored
7          else
8              results_table[u] ← 0, innovator
9          end
10         for v ∈ S_out(u_t) do
11             p_{v,u} ← set
12             if u ∉ results_table then
13                 results_table[u] ← p_{v,u}, notperformed
14             else
15                 results_table[u] ← update(results_table[u], p_{v,u})
16             end
17         end
18     end
```

threshold of adoption.

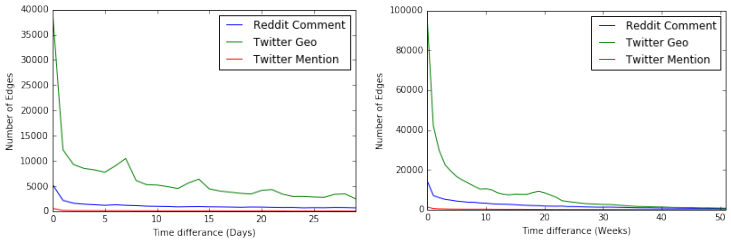
7.6 Experiments

The following section outlines the main findings and results from the experiments that have been performed. The results have been split into two separate sections: section 7.6.1 discusses the results of learning influence and using it to predict innovation usage; the effect of network structure is discussed in section 7.6.2.

7.6.1 Innovation Prediction

Figure 7.6.1 shows the diffusion time of innovations across two separate granularities, time measured in days (figure 7.6.1a) and weeks (figure 7.6.1b). We can see that the majority of the innovation diffusions (between nodes) happen within the first five days of exposure. However, within the Reddit comment network, there appears to be a regularity in peaks at 7 and 14, potentially indicating the existence of an underlying process in users' access patterns of Reddit.

To assess the accuracy of using the influence measures in learning the activation global activation



(a) Day (b) Week

Figure 7.6.1: Innovation diffusion frequencies

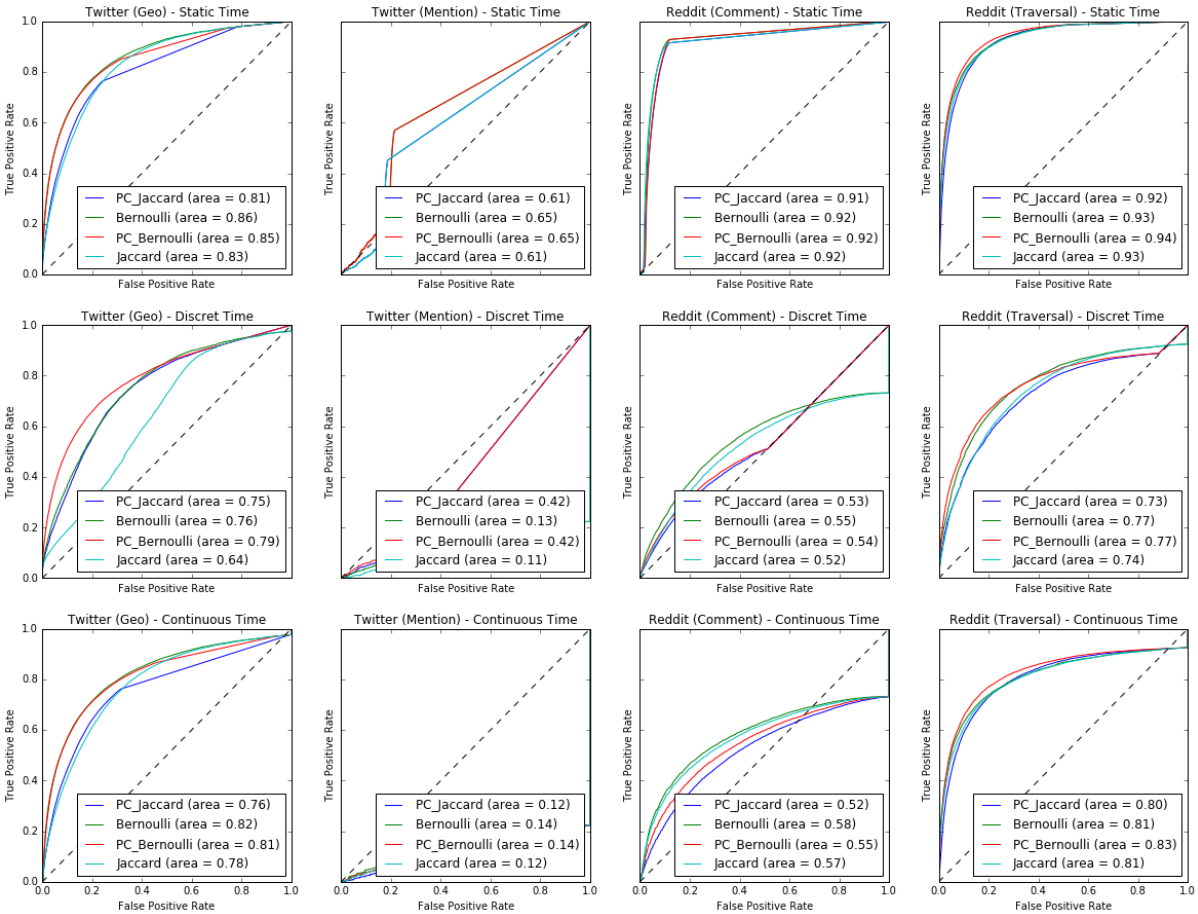


Figure 7.6.2: ROC curves for each of the four networks, using the three different influence models

Algorithm 7.4: Evaluating - Static Time

Data: Each mapper receives a complete innovation diffusion in the [userid, time], with the docid being the name of the diffusion.

Result: List of quads that represents (row, column family, column, value), which then can be loaded into HBase.

```
1  Class Mapper(docid a, diffusion d)
2      n ← network
3      results_table ← [(user, messure, value, performed)]
4      for row u, t_u ∈ d do
5          if v ∈ table then
6              results_table[u] ← perfored
7          else
8              results_table[u] ← 0, innovator
9          end
10         for v ∈ S_out(u_t) do
11             p_{v,u} ← set
12             if u ∉ results_table then
13                 results_table[u] ← p_{v,u}, notperformed
14             else
15                 results_table[u] ← update(results_table[u], p_{v,u})
16             end
17         end
18     end
```

threshold (see section 7.3), we only focus on users that have been exposed to an innovation, and not users who have used it without being exposed (this is due to the need to predict an innovation's adoption based on exposure from other nodes; if there is no exposure, then we will not be able predict the adoption). As stated earlier, this prediction challenge is a binary classification, with the adoption being predicted if the joint probability $i_u(o)$ is greater than the activation threshold (such $i_u(o) > \sigma$).

To learn each global activation threshold σ, we use ROC analysis. ROC analysis models the trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR) by varying the threshold σ, with the aim of finding the threshold at which a user has the largest number of true positives and true negatives, with the higher AUC representing a more accurate threshold. Figure 7.6.2 shows the ROC curves for each of the four networks, across the three time modes, and the four measures of influence. As we can see, the results across all datasets are varied, with the static models proposed in section 7.4.1 being able to predict with AUC highs of 0.92, though there is no discernible difference between the four variations of modelling influence. The introduction of decay functions appears to have reduced the accuracy across all models, resulting in some performing with an accuracy less than a random model ($AUC = 0.5$).

This lowest accuracy can be seen to the greatest extent in the Twitter mention network. This stems

Algorithm 7.5: Evaluating - Continuous Time

Data: Each mapper receives a complete innovation diffusion in the [userid, time], with the docid being the name of the diffusion.

Result: List of quads that represents (row, column family, column, value), which then can be loaded into HBase.

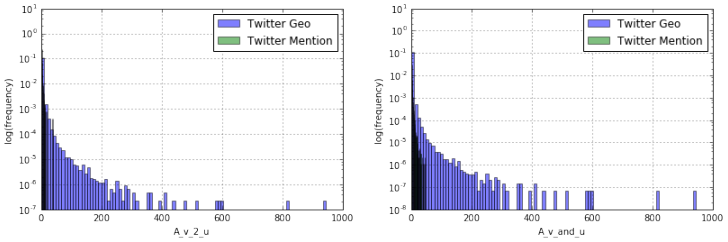
```
1  Class Mapper(docid a, diffusion d)
2      n ← network
3      results_table ← {user : (value, performed, time)}
4      for row u, t_u ∈ d do
5          if v ∈ table then
6              results_table[u] ← 0, perfored, t_u
7          else
8              results_table[u] ← 0, innovator, t_u
9          end
10         for v ∈ S_out(u) do
11             if v ∉ results_table then
12                 results_table[u] ← 0, Never
13             end
14         end
15     end
16     for (useru, valuep_u, performedi_u, timet_u) ∈ results_table do
17         sorted_parents ← {time : (node, value, performed)}
18         for (userv, valuep_v, performedi_v, timet_v) ∈ results_table do
19             if i ≠ Never&e_{u,v} ∈ N then
20                 sorted_parents[t_v] ← (v, 0, i_v)
21             end
22         end
23         for (time, node, value, performed) ∈ sorted_parents do
24             computep_{v,u}
25         end
26         tmp ← {time : (user, value)}
27         for dt ∈ sorted_parents do
28             tmp[dt] ← dt
29             intermediate ← 0
30             for dt2 ∈ tmp do
31                 minutes ← dt2 − dt
32                 power ← minutes/τ_{v,u}
33                 intermediate ← update(intermediate, value ∗ e^{power})
34             end
35         end
36     end
```

Algorithm 7.6: Evaluating - Discrete Time

Data: Each mapper receives a complete innovation diffusion in the form of [`userid`, time], with the `docid` being the name of the diffusion.

Result: List of quads that represents (row, column family, column, value), which then can be loaded into HBase.

```
1  Class Mapper(docid a, diffusion d)
2      n ← network
3      results_table ← {user : (value, performed, time)}
4      for row u, t_u ∈ d do
5          if v ∈ table then
6              results_table[u] ← 0, perfored, t_u
7          else
8              results_table[u] ← 0, innovator, t_u
9          end
10         for v ∈ S_out(u) do
11             if v ∉ results_table then
12                 results_table[u] ← 0, Never
13             end
14         end
15     end
```



(a) O_{v2u} (b) $O_{v\&u}$

Figure 7.6.3: Twitter geo and mention network values for the number of propagations between users O_{v2u} and the joint number of innovations using $O_{v\&u}$

Table 7.4: AUC values for each given POS tag

	Reddit Comment		Reddit Traversal		Twitter Mention		Twitter Geo	
Tag	Bernoulli	Jaccard	Bernoulli	Jaccard	Bernoulli	Jaccard	Bernoulli	Jaccard
!	0.945,814	0.932,778	0.919,580	0.919,668	0.596,826	0.391,674	0.821,373	0.754,288
#	0.988,482	0.985,960	0.992,614	0.986,222	-	-	-	-
,	-	-	0.926,367	0.916,496	-	-	-	-
A	0.974,176	0.966,811	0.941,658	0.933,533	0.372,159	0.390,327	0.828,381	0.781,653
D	0.916,667	0.655,914	-	-	-	-	0.710,258	0.639,674
E	-	-	0.997,986	0.994,964	-	-	-	-
G	0.949,570	0.946,256	0.956,844	0.954,782	0.815,499	0.712,872	0.890,735	0.852,988
L	0.961,226	0.905,481	0.970,527	0.971,248	-	-	0.769,028	0.687,771
N	0.919,589	0.914,614	0.923,128	0.919,163	0.606,992	0.544,365	0.847,362	0.806,809
O	0.945,832	0.943,943	0.902,753	0.901,661	-	-	0.801,335	0.767,615
P	0.797,396	0.812,736	-	-	-	-	0.836,548	0.822,249
R	0.960,906	0.959,497	0.936,667	0.926,402	0.426,421	0.436,959	0.817,372	0.758,002
V	0.950,552	0.945,546	0.943,985	0.941,717	0.635,784	0.637,408	0.819,181	0.767,474
^	0.914,667	0.910,982	0.928,660	0.924,448	0.615,903	0.567,143	0.857,183	0.819,170

from the sparsity of the network and number of propagation (O_{v2u}), which is significantly smaller than that of, for example, the Twitter geo network (as visualised in Figure 7.6.3). However, the overall reduction in accuracy could be due to external unobserved processes that affect language adoption. This could be in line with [86], who stated that influence/exposure comes from not only the local connections but also the community as a whole.

Across all four networks, as the time models become more complex, the AUC reduces, though not to a large extent. This could be for a number of reasons, including the existence of external pressure which influences users language. Additionally, as the windowing strategy ($\tau_{v,u}$) computes the decay of influence as only happening between the two users, though influence to adopt an innovation may decay in relation to all users who have applied pressure; thus, $\tau_{v,u}$ may be too conservative, reducing the $o_u(i)$ too quickly when it is a function of all active neighbours.

As stated in section 7.3, innovations can be classified into different function sets (POS tags). Table 7.4 shows the AUC values for the two non-credit models (equations 7.4.1 and 7.4.2) in each of the four networks. Across the board, the values in Table 7.4 show high AUC, with noticeable improvement in the Twitter comment network. However, the majority of values are still less accurate than a random baseline ($AUC < 0.5$) for the Twitter mention network. Again, this could be due to the sparsity within the network. When looking at the function classes, we can see that abbreviations (G) and verbs (V) appear diffuse in the most predictable manner, potentially indicating that the words that describe the action of a user are more likely to be adopted. This can be seen in Chapter 6 with the growth of words such as *vape* and *vaping*. This can also be seen for acronyms such as *cyw* and *sjw*.

Unlike the static and discrete time models, there is only one maximum joint probability $i_u(o)$ when using the continuous time model (equation 7.4.9) for an individual user. This is the point at which there is the greatest amount of influence on the given user to adopt a term. To assess if the time of adoption is

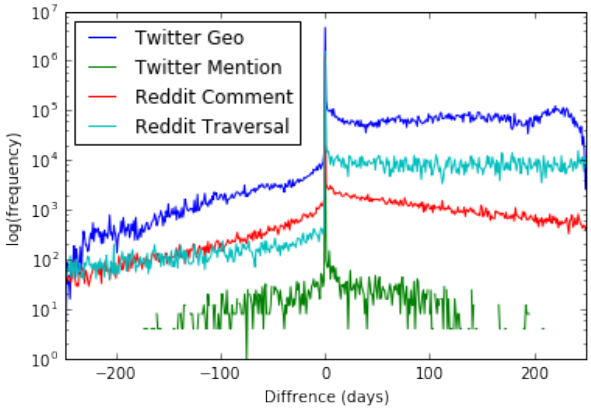


Figure 7.6.4: Time difference from global maximum of user adoption

near the time at which there was the greatest amount of influence, we compute the difference between the time that a user adopted a term and the time at which they experienced the greatest influence. This can be seen in Figure 7.6.4: values < 0 indicate that the individual used the innovation before the maximum influence, whereas values > 0 indicate the term was used after the maximum. There is a distinctive spike around 0, indicating that the value of the maximum is on the same day the innovation was used. However, the long tail to the right of the peak potentially indicates that there is a large proportion of users who delay their first usage. This long tail could also highlight that the model is decaying the join influence too quickly, or by the wrong proportions.

7.6.2 Network Structure

Influence to adopt language innovations comes from within the social network, from the people with whom a user communicates and events within the network that they observe. These interactions (communicating with each other) result in densely connected areas of the network that can be thought of as communities of nodes. For the Twitter mention and Reddit comment networks, these can be thought of as clusters of users, whereas, for the Twitter geo and Reddit subreddit network, they can be seen as cluster of areas. As highlighted in section 7.2, the bonds internal and external to the community structure influence the diffusion of language and content, with strong internal bonds reinforcing language internal to the community structures.

Figure 7.6.5 plots the distributions of influence across edges internal and external to each community. The aim is to see a statistically significant difference between internal and external influences, as we would expect users who are highly connected to each other to have a greater amount of influence on one another's language. If this effect was seen, then there would be evidence of structural trapping [207], where innovation and the diffusion of content is trapped by the community structure of the network, allowing for specialised languages and norms to develop internal to the community structure.

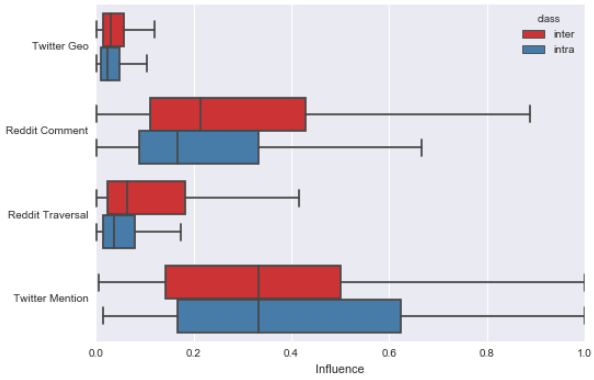


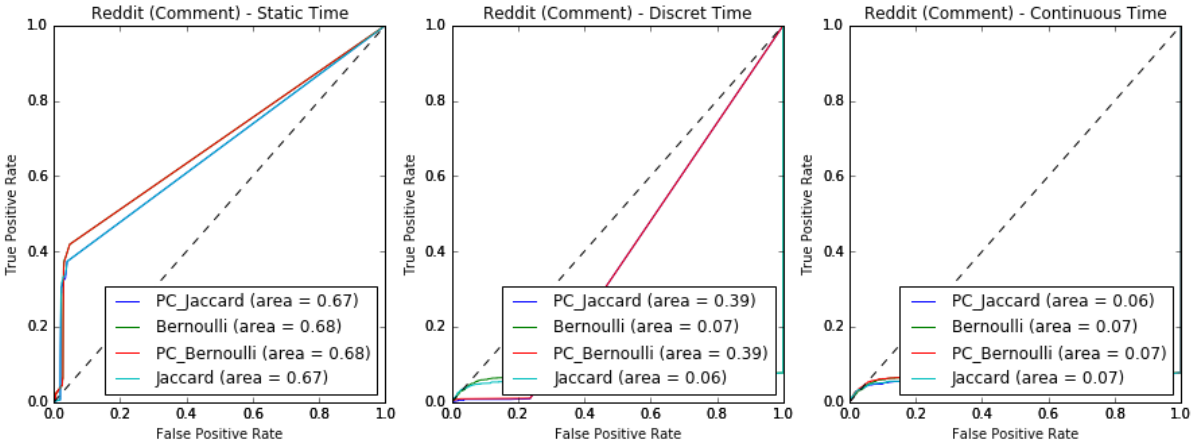
Figure 7.6.5: Influence distributions structures internal (inter) and external (intra) to the community of each network.

However, the results are inconclusive, showing that there are no significant differences in influences internal or external to the network structure. We can see that, internal to a community, the spread of influence is greater, with the majority of influence being higher across the four networks internal to a network. However, the sparsity of the Twitter mention network results in values with the greatest ranges, though with the same mean but higher external influence. Thus, influence may not be affected by structural trapping, as proposed in [207], showing that innovations diffuse internal and external to the community structure of each network.

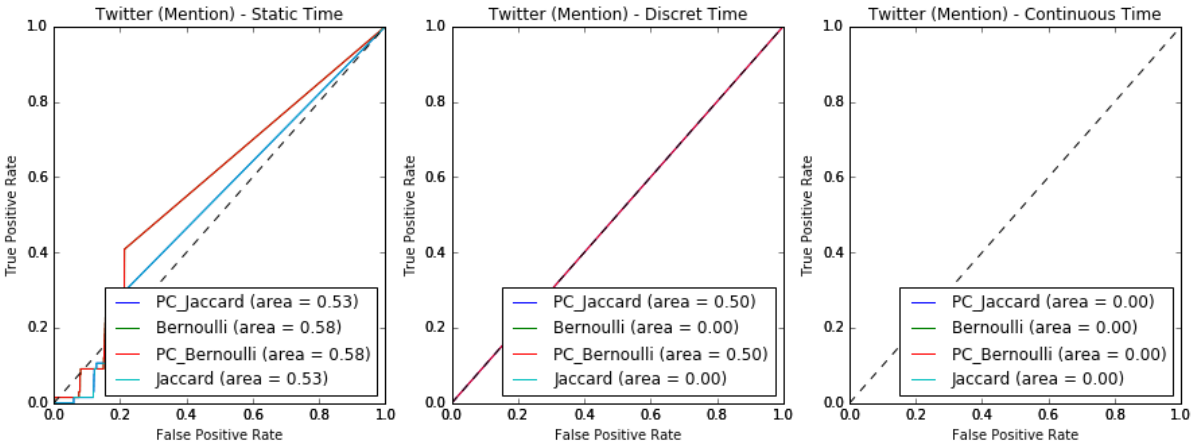
The extent to which language adoptions depend on influence originating from the network structure and which come from external sources cannot be measured. To assess this, we shuffle each of the four networks, as stated in section 7.4.3.

Figure 7.6.6 shows the results from shuffling the Reddit comment network and the Twitter mention network. As we can see, for the static time model of Reddit comments, there is roughly a 25 % reduction in AUC, whereas for Twitter, it is 20 %. However, when looking at the time-dependent models, the accuracy falls below that of the random baseline model with values as low as 0.07 %. When inspecting the values from the other networks, similar patterns can be seen, whereby the static time model has a reduction in accuracy, with the time-dependent models reducing to a larger extent. As [86] stated, influence might come not only from one's local acquaintances, but also from across the whole network. However, when assessing the decay models, it could be that a user's adoption of an innovation is dependent on the network, though the time at which the innovation is adopted is dependent on the local network.

Building on the analysis of network structures and their effects on innovation diffusion, a number of comments can be made between the two forms of network structure, micro (user interactions) and macro (community interactions). We can see that there is a higher accuracy in predicting when a community rather than an individual will adopt a new term. This could indicate that the collective influence on a group of users is greater than that of the individual, showing that it is not an individual that affects



(a) Reddit comment



(b) Twitter Mention

Figure 7.6.6: ROC curves learnt on the shuffled version of network abstraction

change but rather a collective. However, for research purposes, we classify an adoption as being when a term is adopted only once in a community. In reality, for a community to 'adopt' a given term, it would more likely to be when the majority of users in the community use the term, or when it has been used more than a certain number of times. Additionally, network structure does not influence the diffusion of content across the networks, with 50 % of the influence coming from sources external to the network.

7.7 Discussion and Conclusion

The novel contribution in this chapter lies in the implementation of the framework in the theoretical frameworks of innovation diffusion proposed by [86] and [196], through computational methods proposed by [83].

[86] and [196] proposed that, with each user, there is a threshold representing the amount of influence that they will succumb to in order to adopt an innovation. [83] proposed that this threshold (σ) should be learnt in a two-phase process, first by learning the influence between nodes as a function of historic innovation diffusion, then using these inter-node influence values ($i_{v,u}$) to compute the global threshold value (σ) and predict when a node is going to adopt an innovation (this is achieved using ROC curve analysis).

The results from the two initial phases showed that, across the two data sources and the two network representations (*micro* and *macro*), we can predict the process of users adopting language innovations (as $AUC > 0.5$). However, the nature of the mined network's influence the accuracy of the model can be learnt. This can be seen to a greater extent in the Twitter mention network, which is the sparsest of the four networks, in both network and innovation diffusion. This is due the manner in which the tweets were collected, as each at the end of an edge must have tweeted with GPS turned on; as only 4 % of all user have GPS turned on, this limits the number of edges in the dataset. Additionally, the manner in which the data was collected means that not all user interactions will have been captured, therefore only identifying a small part of a user's interaction online. However, for the denser networks, namely both Reddit networks, we can see that it is easier to predict the innovation adoption of a group of users (subreddit) than individual users. As highlighted, for a group of users to adopt an innovation, they only have to use the innovation once.

As proposed by [83], influence on a user to perform an action may deteriorate over time. To model this, we measure the mean time of historic exposure to adoption time, and use this to decay the influence, either in a discrete or continuous manner. However, in doing so, the AUC across the board reduced. This could be for a number of reasons; as [83] developed the decay model to see if users would tag a photo in a newsfeed on Flickr, meaning that a user had a limited window of opportunity in which to tag the photo before it left their newsfeed, with limited chances for re-exposure. For language, the innovation

lives with the user, so it will not vanish from a feed, and may last within the user's memory for a longer period of time, meaning that there may be a delay between being exposed to an innovation and using it. A potential adjustment to the model could be to use a slower decay function, or, instead of computing a decay function for each edge, computing a decay function based on an aggregation across all of the edges of a user.

In addition to predicting language adoption, we also showed that there is a potential underlying background process affecting the adoption of language. This is broken down into understanding the effect network structure has on a user adopting an innovation, and how community (collection of densely connected nodes) influences language adoption. By shuffling the edges in each of the four networks, we showed that the network structure produces about 75 % of the influence on a user to adopt an innovation, with the remainder coming from processes that could not be captured. This could come from a number of sources, either from a user browsing the network and from elsewhere on the internet, or from sources such as TV or radio. Additionally, our results show that there is little dependency on the community structure in the propagation of influence, potentially showing that, in an online world where users can move about freely, there is less constant membership within one community, as users change their membership frequently.

This chapter has outlined the processes that influence the language adoption of both individuals and collections of users (regions or subreddits). We showed that this is possible by learning a global threshold. However, this is one threshold for the whole network. Future work should focus on learning the threshold per user, which would mean that we would be able to distinguish between users who are innovative with their language (by having a low threshold) and users who are conservative. This would mean that we could apply more of the innovation diffusion work of [86] and [196].

Intra-regionally, we have shown that language diffuses across geographical landscapes in a similar manner to that online. This potentially shows that users are forming communities online that mirror the strength and influence of historic geographical communities [148]. This raises questions about what it is that has a greater influence on user language adoption, whether geographical communities or online communities. This work cannot distinguish between these two influences.

7.8 Data Access

All data and code created during this research are openly available for further use. The datasets are available from the Lancaster University data archive at D. Kershaw, M. Rowe, A. Noulas, *et al.*, *Birds of a feather talk together: user influence on language adoption - data set*, http://dx.doi.org/10.17635/lancaster/researchdata/99, 2017. DOI: 10.17635/lancaster/researchdata/99 with the code hosted on github D. Kershaw, M. Rowe, P. K. Stacey, *et al.*, *Birds of a feather talk together: user influence on*

language adoption - code. DOI: 10.5281/zenodo.1216050. [Online]. Available: https://doi.org/10.5281/zenodo.1216050.

Chapter 8

Predict Language Diffusion

Due to processing and time constrains issues one network, Reddit Comment, had to be removed from the results of this Chapter. In particulate these are Figures 8.6.4, 8.6.5 and 8.6.6.

In the previous two chapters (6 and 7), we have focused on modelling the growth of language innovation and predicting when a user will adopt an innovation based on influence from their neighbours. The results show that we can operationalise known heuristics of language acceptance, and that users adopt language from other users around them. However, the results for predicting user adoption indicate that upto 50 % of the influence that a user receives could come from other areas of their ego network. This final research chapter will look at the diffusion of innovation not at the individual user level but across the whole network, and whether the diffusion of innovations across social networks is predictable.

8.1 Introduction

Language is in constant flux, from changes in pronunciation and variations in meaning, to the introduction of new words or terms. These changes can be seen in the world around us, from recent innovations (referring to new words or phrases that are intentionally or unintentionally created by a user [45]) such as *fleek*[1], or historic variations in meanings, e.g. *gay*. At the core of language change (the process of variation and innovation over time in language) is the understanding that users manipulate their language and the language around them to achieve their own goals, whether to exchange information or to attract attention. These individual *speaker innovations* can be anything from a shortening or lengthening of words, e.g. *'casue* or *sooooo*, to bringing attention to the word through the introduction of foreign words to make the speaker appear sophisticated, e.g. *schadenfreude*, or the formation of completely new words to draw attention to a new concept, e.g. *email*.

[1] adjective, to mean flawlessly styled or groomed

However, the creation of an innovation (new words or a variation of an existing term) does not cause *language change*; rather, it is the collective adoption of multiple *innovations* that ultimately causes change in a language over time. The process of language change, and, ultimately, the diffusion of new innovations, has been studied widely using traditional linguistic methods, with results identifying numerous factors influencing the diffusion processes; [18] claims that new words (innovations) diffuse due to their ability to be used in a variety of situations and contexts, whereas [148] proposed that the final range of the diffusion of an innovation is dependent on community structure and access to structural holes. Even so, [129] postulated that social class and geographical boundaries (a user's social variables) define the language that they use and thus defined boundaries along with innovations would travel across.

In recent times, significant effort has been made into understanding how information diffuses across OSN; this has included simulating memes across networks to predict the vitality of the content [207], to identifying who the important players are in aiding the diffusion of content [121]. However, there has been limited work on understanding and quantifying how language and language innovations diffuse across network OSN, with studies focusing rather on discrete units such as memes or URLs.

This chapter is summarised in one over-arching question: *Given the creation of speaker innovations (a new term), to what extent can their diffusion across a network be predicted?*. Our contributions can be summarised in the following points:

1. **Predictability of *speaker innovation* diffusion**: Using a selection of features extracted from the network structure, the context of innovation usage and the temporal adoption dynamics, we demonstrate that the final diffusion size (final number of users) of new words within a social network is highly predictable.

2. **Dependency of diffusion on structural holes**: We further show that diffusion of language in social networks is highly dependent on the presence of structural holes, with weak ties critically influencing the diffusion process.

3. **Factors influencing language change differ depending on the networks**: We show that the diffusion of language is governed by different factors depending on whether the diffusion process is studied on an aggregated group level or on an interpersonal (user-to-user) level.

More specifically, we demonstrate the predictability of *language change* by observing the *structural*, *temporal* and *grammatical* features of each word innovation.

This work draws on the methods proposed by [36], but, instead of predicting the final size of a diffusion, we ask whether, after n steps, the innovation will diffuse more or less than a typical (*median*) innovation with n observations (turning the problem into a binary classification task). As we explain in detail in Section 8.6.1, this formulation offers practical benefits in terms of both training diffusion models and assessing their performance in a manner than generalises across different prediction settings.

In terms of features exploited for the mode and to study language diffusion, we consider the following broad families (with detailed mathematical definitions provided in Section 8.4):

- **Baseline:** Drawing on [207], we consider a number of baseline features (e.g. number of adopters of a new term) that are indicative of the spread of diffusion on the network's surface. These popularity-aware baselines set a bar for other models in the prediction task, with results from the models using these features performing well across the board in terms of the ability to predict the virility of language innovations.

- **Temporal:** Diffusion of language takes place over time in much the same way as memes diffuse across OSN [185], [208]. We therefore devise a number of metrics that exploit the temporal adoption patterns of innovations. These take into account the growth rate of a diffusion relative to its start time, and also the time patterns emerging across consecutive user adoptions, such as the inter-adoption time variance.

- **Network topology and community structure:** We design a number of features that incorporate topological information on diffusion patterns across the network. These include measures that exploit characteristic information pathways followed by a diffusion (e.g. furthest distance between two active nodes in the network) and characterisations of node topology (e.g. degree centrality of all the active nodes). We focus on the role of structural holes in language diffusion and examine the decisive role of weak ties in the diffusion process.

  Further, as [208] and [148] suggest, community structure plays a decisive role in information diffusion. As a result, we measure the diffusions of innovations across both individual users (*micro*) and the aggregation of users (*macros*). This results in measures assessing the fraction of activated communities in the network (e.g. communities where an innovation has been adopted) and diverse adoption frequency patterns across communities, but also intra-community (between community) diffusion patterns.

- **Grammatical:** Finally, for words to be 'accepted', [18] proposed that they must have the ability to be used in varying contexts and situations. Examples of this can be seen in words such as *Google* being used as both a *noun* and a *verb*. Thus, as an innovation is accepted more broadly (by achieving a larger diffusion), we would expect the diversity of grammatical tags (used as a proxy for contexts) to increase. To this end, we consider the impact of the grammatical diversity of an innovation in the diffusion process.

The implications of this work lie both internal and external to academia. For the first time, this work allows us to assess and predict the diffusion of speaker innovations (new words) across a number of online social network OSN, between both users and aggregations of users. By understanding how language

diffuses through nodes of varying bonding and bridging statuses in the network, law enforcement can focus activities on particular members of a community to understand how messages are changing and evolving over time, and how we can identify new terms that could be monitored on social media to identify the future activity of groups. Additionally, the developed methods could also be used by marketers who want to maintain a language for a campaign that that is pre-emptive of language changes in the general population or specific communities. From an academic perspective, these methods could be used within digital humanities by quantitatively assessing the historic diffusions of language across international trade or migration networks, with historic methods ranging on the more qualitative side [45].

The remainder of this chapter is organised as follows. Section 8.2 focuses on related work. The datasets that are used in this study are discussed in Section 8.3.1. The models and features used within the models are described in Sections 8.3.2 and 8.4. Section 8.6 introduces the three experiments performed, along with their respective results.

8.2 Related Work

Language change (and variations in language over time) has and is receiving an increasing amount of attention from the field of computational social science, from detecting the growth of new words [89], to the predictive modelling of language diffusions across the US [64]. However, limited work has been conducted on predicting the diffusion of a language through social networks, in particular quantifying the effect of different factors (such as network topology, linguistic and temporal aspects) on the diffusion of language innovations.

Diffusions of language innovation (and, ultimately, language change itself) can be thought of more broadly as information diffusions, which can be categorised as either the study of when a user is going to adopt an action or more broadly how many users, collectively, are going to adopt a set of actions. In this work, we focus on the latter. In predicting user adoption of actions, [83] focused on modelling influence between users as a function of past diffusions of actions (tagging the same image of Flickr). This led to a general threshold being learnt (drawing on the work of [86]) across a network in which a user performed an action based on the influence and exposure of their neighbours.

Studies on predicting collective user attention have included predicting the number of video views on YouTube [185], the number of up-votes on a Reddit thread [179] and the number of re-tweets a tweet receives [69]. [208] identified that, in the case of memes, their initial growth rate is indicative of the future size of a diffusion. This correlation was additionally identified by [185], who showed that the number of views of videos on YouTube can be predicted through a linear relationship of views between the start and end of a month. The diffusion of textual content was assessed by [189], who showed that there is a large dependency not only on the message itself but also on the language used in communicating the

Table 8.1: Dataset statistics

	Reddit	Twitter
Unique Words	2,942,555	526,342
Posts	1,054,976,755	111,067,539
Innovations	2,712,629	373,217
Days in Dataset	880	283

message. However, not all users within a network have the same effect on the diffusion of information; [213] identified that the position of a user who spans structural holes influences content diffusion to a greater extent.

Ultimately, the approaches in predicting the *collective actions* of users varies, taking either the form of *regression* (as seen in [185]), or a *classification* task. A number of methods have been developed that aim to classify a diffusion as 'viral' (as seen in [208], [209]) or not. However, what is classed as 'going viral' varies. [208] separated all final diffusion sizes into log scale bins, with the classification being the bin the final diffusion fell into, whereas [205] classified 'viral' as being in the top $X\%$ of final diffusion sizes (ranging from 50 % to 0.01 %). **Cheng:2014kmb** proposed that treating the challenge of classification as simple 'viral' threshold implies an over-reliance on large but rare diffusions. They proposed that we should focus instead on the predictability of a diffusion across its whole life. This is done by predicting at each step of a diffusion whether the final size of it will be above or below the median of all diffusions of similar size. Following the empirical evidence in our analysis, which suggests that the diffusion of language innovations follows a similar distribution in terms of size, we adopt a similar methodology. We justify our choice in detail in Section 8.6.1.

8.3 Methods

The following Section (8.3.1) will introduce the datasets used within this chapter, as well as explaining how the social network is extracted from them. In Section 8.3.2, basic notation is then introduced, which is used to describe the social networks and the diffusions that happen across them. We finally introduce the general framework that is used in assessing the predictability of innovation diffusions across the social networks in Section 8.3.3.

8.3.1 Datasets

We now introduce the datasets and network formulations used throughout this work. Additionally, we define what we classify as a language (speaker) innovation, which is the unit of analysis when assessing language change.

Defining Speaker Innovations

In defining a language innovation, we must first define what is not an innovation in language. For this purpose, the British National Corpus (BNC) [11] is the gold standard of the English language, which is represented as a list of common English language words (for more detail on the BNC, see Chapter 5.3). Thus, the BNC is used to remove all 'accepted' terms from the mined social media datasets from Reddit and Twitter. However, social media data is inherently noisy [14], so a normalisation strategy is also applied. This reduces instances of charter repartition (e.g. changing *sooooooo* to *soo*). Additionally, a large number of innovations are only seen once; therefore, to reduce the number of diffusions that need to be computed, we focus on innovations that have been used at least 10 times. This post filtering leaves us with 2,712,629 innovations found on Reddit, and 373,217 on Twitter (see Table 8.1).

Social Network Data

Previous research on information diffusion offers limited examples of the models working on multiple datasets from varying OSN; thus, for this work the methods developed will be applied on two distinct social networks: Twitter and Reddit.

Users engage with either Online Social Network (OSN) for different reasons, with the networks themselves offering different features to the end user. Twitter is a micro-blogging platform that allows users to broadcast short amounts of information (thoughts, images, links, etc.) to their followers without needing a response. We exploit the fact that a fraction of tweets are geo-tagged, thus allowing for the exploration of the relationship between geography and language diffusion. Our data collection focuses on tweets that are generated in the United Kingdom region (though not all tweets generated in the UK are collected since only 4% of all tweets [2] are geo-tagged).

Reddit is more like a traditional online forum, with user activity focused around topic-specific content generation (links, posts, comments, etc.). These areas are typically organised in collections, commonly known as *subreddits*. The Reddit dataset is an 18-month public data dump containing all (roughly 1 billion in total) posts in Reddit from 2,007 to June 2,015.[3]. Our Twitter sample was collected from September 2,014 to June 2,015, and contains 111 million tweets in total (Table 7.1 shows a summary of basic statistics for both datasets).

For more information on the datasets used, please see section 5.2.

As previously mentioned, we study diffusion across two levels of network abstraction: one representing user-to-user interactions and the second between groups of users. The abstractions can be thought of as representing the *micro* and *macro* interactions within OSNs. *Micro* models interactions between users on both Reddit and Twitter. Alternatively, *macro* models the interaction of users across regions within

[2] PewResearCenter - Location-Based Services
[3] Dataset available on the Internet Archive.

Table 8.2: Network statistics

Network	Nodes	Edges	Communities
Twitter Geo	2,910	436,849	14
Twitter Mention	283,755	329,440	39,767
Reddit Comment	861,955	2,402,202	36,885
Reddit Subreddit	15,457	142,285	407

each network: for Reddit, this is subreddits and for Twitter this is across geographical locations within the UK.

For this work, each OSN network takes the form of a *directed graph*. In this setting, a graph G is defined as a triplet $G = (V, E, W)$, containing vertices $v, u \in V$ and edges between them, $e_{v,u} \in E$; the latter denotes an outward connection from v to u. The triplet also includes the weight of an edge $w_{i,j} \in W$, which is proportional to the number of interactions i has with j (e.g. i mentioning j in a tweet). Additionally, the function $N_{out}(i)$ returns the set of neighbours of node i that can be reached through the outward edges of i, and $N_{in}(i)$ for the inward edges. For more detail, see Section 5.2.

Micro

At the micro level, we model the graphs through user interactions. Twitter users interact with each other in a number of ways; however, the predominant form within this dataset is mentioning fellow users in tweets (through the inclusion of the '@¡username¿'). We use this to build a user-to-user graph, where a relationship from user $v \rightarrow u$ is noted if u *mentions* user v. The edge is assigned a weight that is set to be the total number of times u mentions v. Similarly, within Reddit, users comment on each other's posts, forming a chain of interactions; therefore, we define a relationship between two users u and v; thus, if user u comments on a post of user v, an edge is formed $u \rightarrow v$. The weight ($w_{u,v}$) is the total number of times user u has commented on a post of v.

Macro

Even though users may not comment or interact directly with each other, they still may be exposed to each other's content by observing the network or being part of the same group. Collectively, a group of users may also exert influence over other collections of users; for this reason, we cluster together content (posts and tweets) generated within the same collections (subreddits or postcodes). These collections are then treated as nodes, with associated edges between nodes being proportional to the number of users within the network who have consecutively moved between the nodes (e.g. user i tweeting in $LA1$ then tweeting in $LA2$ would generate the edge $LA1 \rightarrow LA2$). For Twitter, the aim is to model the interaction of users across the geographical landscape; therefore, nodes in this graph represent postcodes within the UK. Each postcode has a centroid, with tweets being assigned to a postcode based on the shortest

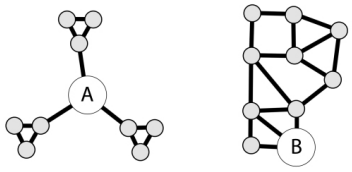


Figure 8.3.1: **A** then **B** are less constrained as they have access to multiple other nodes that are not connected with each other.

distance between the GPS coordinates of the tweet and the centroid of the postcodes. To then model the interaction between postcodes, the number of users that tweeted consecutively between two postcodes is counted (e.g. tweeting from `LA1` to `LA2`). This is then represented as the weight between two nodes ($w_{i,j}$). Similarly, within Reddit, users interact and move around different subreddits depending on their current interests or in reaction to popular content. Thus, the interaction of subreddits is modelled by users moving between subreddits. The weight between two nodes ($w_{i,j}$) is the number of users moving consecutively from subreddits i to j.

8.3.2 Definitions

The following section introduces the notation representing the diffusion of an innovation across a network, along with a number of basic measures used throughout this chapter.

Speaker innovations: A single innovation in a network is defined as o, such as $o \in O$, where O is the set of all innovations. Each network is generated from a series of posts $p \in P$, with $P(o)$ returning all the posts containing o. Posts are generated in sequence; therefore, the first n posts containing o are accessed with $P_n(o)$, with the sequence of posts appearing as $\langle p_0^o, p_1^o, ...p_n^o \rangle$. p_n^o. Posts are generated by nodes ($v \in V$, with v representing either *macro* or *micro* nodes) in the network, with the adopter/usage sequence of o is accessed through $V(o)$ with the first n in the sequence coming from $V_n(o)$, which returns $\langle v_0^o, v_1^o, ...v_n^o \rangle$. If node uses o multiple times then they appear in the sequence numerous times.

Topology: an important concept we explore in this chapter is that of the influence of the networks topology on innovation diffusion. In addition to grouping users on the basis of domains (e.g. subreddits on Reddit and geographic communities on Twitter), we use the Louvain method [22] to assign each node in a network community (collection of nodes). Each node is thus assigned to a community c based on the application of the Louvain method, the whole set of communities is represented as $c \in C$. A node's (v) community membership can be accessed through C_v, whereas all the nodes that are members of a community can be accessed through $C(v|c)$. Further, all posts (p) that use innovation o in a given community c are accessed through $P(o|c)$; similarly, for nodes, we have $V(o|c)$. $V_n(o|c)$ and $P_n(o|c)$ then return the first n items that appeared in c.

A node becomes 'activated' when it uses an innovation o one or more times. A community c is

classified as activated if there is one or more instance of an innovation appearing within the community, with the set of activated communities being defined as $C(o) = \{c, c \in C, |P(c|o)| > 0\}$. Additionally, the degree distribution of a node is defined as $d(v)$ and the neighbours of node v are accessed through $S(v)$.

Structural holes: Not all nodes within a network fulfil the same role; some place themselves internal to communities and appear on the periphery. This effect can be seen in strong internal bonds. Strong internal bonds (internal to a community) result in the reinforcement of communication, ideas and relationships internal to the group. However, with strong internal bonds, there are also weak external ties, which are the result of voids across the network structure where no or limited edges exist. These limit the ability for users to see/communicate with the rest of the network. These voids in network structure are also referred to as *structural holes*, even though they limit access to information. It also gives opportunities to users that span these voids as they gain 'better' access to a variation of ideas and innovations as they are exposed to communication from multiple communities. A key hypothesis for information diffusion, and hence language change, is that users who are able to bridge voids influence, to a greater extent, the final diffusion size of a language innovation and highly influence the process of language change.

[27] proposed that a user's access to structural holes can be measured by the degree to which a user's ego network is *constrained*. The more constrained, the less access they have to structural holes and thus less opportunities to foster the spread of a diffusion. We set $z_{i,j}$ as the constraint between two nodes i and j and mathematically define it as:

$$z_{i,j} = (w_{i,j} + \sum_{n \in N(j)} w_{i,n} w_{n,j})^2 * 100, \quad n \neq i, j \tag{8.3.1}$$

With $w_{i,j}$ (weight of an edge) representing the time and attention that i gives j. $N(j)$ is the set of users connected to j, with $w_{i,n}w_{n,j}$ computing the resulting strength of a connection up to one hop away. If there is no connection from i to q, then this will be 0.

The value is computed for each edge in the network. To then represent how constrained a node is as a whole, we use a *constrain index* Z_i, defined for a node i as:

$$Z_i = \sum_{j \in N(i)} z_{i,j}, \quad i \neq j \tag{8.3.2}$$

Grammatical tags: Innovations are used for a variety of reasons, whether to name an object or as a new interjection between users. The variation in contextual usages is assessed by the variation in grammatical tags ($q \in Q$) of words assigned by a POS tagger. The tag for each usage of an innovation (p_n^o) is accessed using the method $Q(p_n^o)$. Additionally, $Q(o|q)$ returns all the innovations o that have been tagged with q, whereas $Q_n(o|q)$ returns all the posts that contain innovation o that have been tagged with q.

8.3.3 Predicting Diffusions

Building on the method proposed by **Cheng:2014kmb** we use a binary classifier after n observations to predict whether the final diffusion size will be above or below the median of all diffusions that have achieved at least n observations. As stated by **Cheng:2014kmb** this means that the question being asked is *'whether the current diffusion will double in size'*. The underlying assumption is that, for a distribution that follows power law with an exponent $\alpha \simeq 2$, at any stage in the diffusion process (at any point n) only half of the innovation will achieve a final size greater than the median of all final diffusion sizes for all innovations at point n. The main premise of this model is that the diffusion has an exponent of 2, as this will mean that the median is $2 \times x_{min}$ of the final diffusion size of all that have reached at least n, with the model predicting whether the diffusion will double or not in size. This is proven using the following calculation:

$$\int_{x_{min}}^{f(x)} \frac{\alpha - 1}{x_{min}} \left(\frac{x}{x_{min}} \right)^{-\alpha} dx = \frac{1}{2} \Rightarrow f(x) = 2^{\frac{1}{\alpha - 1}} x_{min} = 2x_{min} \tag{8.3.3}$$

Where x_{min} is n as all diffusions being assessed will have at least n shares.

8.4 Operationalisation

The aim is to assess the predictability of a diffusion at each point (n), by predicting whether the final cascade size will be double that of the current point. The features mined are influenced by the theories on language diffusion and previous works referred to in section 8.1.

Basic

First, we present a number of baseline features that are indicative of the historic (that is up to step n) popularity of a diffusion.

Number of users: $|A_n(o)|$ quantifies the number of unique nodes that have used a given innovation. $A_n(o)$ returns a set of nodes up to (and including) the nth usage.

Number of activations: Each node can use an innovation more than once; therefore, we use $|P_n(o)|$ to represent the total number of usages of an innovation.

Surface: We consider the number of uninfected nodes that can be reached from infected nodes in one hop in order to assess the potential of the diffusion. The size of the surface is defined as:

$$|S_1(A_n(o))| = \{u, u \ni N(v); v \in A_n(o)\} \tag{8.4.1}$$

Where $N(v)$ returns the set of neighbours of node v.

Temporal

Diffusions naturally evolve over time, going through bursts and lulls in speed. t_n^o represents the time of a post p_n containing o.

Average step time: Step distance is the time between consecutive usages; therefore, the average step distance is defined as:

$$\bar{\Delta t_n}(o) = \frac{\sum_{i=1}^{n-1}(t_{i+1}^o - t_i^o)}{n-1} \tag{8.4.2}$$

CV step time: We define the variance across all consecutive time steps as:

$$C_v(\Delta t_n(o)) = \frac{1}{\bar{\Delta t_n}(o)}\sqrt{\frac{\sum_{i=1}^{n-1}(t_{i+1}^h - t_i^h - \Delta t_n(o))^2}{n-2}} \tag{8.4.3}$$

Average time to adoption: Additionally, it is not only the time between usages that may matter, but also the time since the original innovation's first use (the age of an innovation). We define $tta_n(o)$ as the difference between $t_n(o)$ and the original usage ($t_0(o)$). Formally, we have:

$$\bar{tta_n}(o) = \frac{\sum t_n(o) - t_0(o)}{n} \tag{8.4.4}$$

Network Topology

The position of a user in a network is important to the diffusion of innovations, with the following measures quantifying the structure of a diffusion across a network's topology.

Average step distance: The average step distance between consecutive usages of an innovation across the network is formally defined as:

$$\bar{d_n}(o) = \frac{1}{n-1}\sum_{i=1}^{n-1} d(v_i^o, v_{i+1}^o) \tag{8.4.5}$$

With function $d(v_i^o, v_{i+1}^o)$ returning the distance between the two nodes.

Diameter: The maximum distance between all activated nodes in the network within the first n usages of an innovation o, defined as:

$$D_n(o) = \max_{1 \le i \ne j \le n-1} d(v_i^o, v_j^o) \tag{8.4.6}$$

Average degree: Computes the average degree across the sequence of adopters:

$$\bar{D_n}(o) = \frac{\sum_{v \in V_n(o)} d(v)}{|V_n(o)|} \tag{8.4.7}$$

Average constraint: Building on Equation 8.3.2, the average constraint of the first n nodes is

calculated as:

$$\bar{Z_n}(o) = \frac{\sum_{v \in V_n^o} Z_v}{1 - n} \tag{8.4.8}$$

Community

Community (a collection of nodes within a network, not to be confused with the aggregated macros networks) influences the language that is used and plays a role on how it diffuses, through its ability to aid in norm formation through social reinforcement. Therefore, we define a number of features that exploit information about communities of nodes in each network.

Proportion of activated communities: $P_C(o)$ returns the proportion of communities in the network that have used the innovation o.

$$P_C(o) = \frac{|\{c; |C(o|c) > 0| \, ; c \in C\}|}{|C|} \tag{8.4.9}$$

Activation entropy: Assessing the proportion of communities that have been 'activated', though it does not give an indication of the extent to which each community has accepted the innovation. Therefore, by adopting the notion of *information entropy*, we quantify the spread of activated nodes across the community structure:

$$H_n^T(o) = -\sum_{c \in C(o)} \frac{|P_n(o|c)|}{n} \log(\frac{|P_n(o|c)|}{n}) \tag{8.4.10}$$

Usage entropy: Instead of using the number of nodes, we take into account the number of times that the innovation has been used in the community by any node. This is defined as:

$$U_n^T(o) = -\sum_{c \in C(o)} \frac{|V_n(o|c)|}{|V_n(o)|} \log(\frac{|V_n(o|c)|}{|V_n(o)|}) \tag{8.4.11}$$

The idea is that the higher the entropy, the greater the spread of the diffusion across the communities, thus indicating a greater acceptance of the innovation.

Grammatical

To measure diversity in context, we also use an entropy-based metric.

POS tag entropy: $Q_n^T(o)$ measures how tags of a given innovation are distributed. A large value will indicate high diversity in tags, whereas a small value will indicate focused tag usage:

$$Q_n^T(o) = -\sum_{q \in Q} \frac{|Q_n(o|q)|}{n} \log(\frac{|Q_n(o|q)|}{n}) \tag{8.4.12}$$

8.5 Computational Methods

One of the technical challenges of this chapter is computing a number of large metrics at each stage (n) within each innovation diffusion. This is second to the challenge of scaling this attainably to extract metrics at each stage and then use them to test and train a `logistic regression` model.

The aim of this work is to assess the predictability of the diffusion of speak innovations at each stage within the diffusion (n). Therefore, at each stage in an innovation's diffusion, all metrics need to be computed. Then, across diffusions, at the same vale of n, the metrics need to be aggregated, as we make the prediction with regard to n.

However, grouping all metrics with the same value of n poses a challenge, as, when n is small, the number of innovations will be large, meaning that sorting and storing large quantities of data on one machine can be a challenge. However, this store, sort and group pattern of separating the mining of the metrics and making the prediction means we can use the `map reduce` pattern. This additionally means proposed models and framework can then run on a Hadoop cluster, allowing for scalable execution.

Implementing the model in a `map reduce` pattern means that the feature extraction and model learning and testing can be separated into two distinct phases. The `mapper` phase iterates over a diffusion, and computes and emits the metrics at each new observation (n) of an innovation. The second phase, the `reducer`, then aggregates all the metrics that were observed at the same point (n), then trains and tests the predictive model, outputting the accuracy of the model.

Algorithm 8.1: Innovation diffusion - map reduce

Data: Each mapper receives a complete innovation diffusion in the [userid, time], with the docid being the name of the diffusion

Result: Accuracy of predicting the correct classification at n observations of the diffusion, based on the metrics computed from the mapper.

Class `Mapper(`*docid* a, *diffusion* d`)`
  network $\leftarrow$ []
  for $n \leftarrow 0$ **to** `len(`d`)` **do**
    update $d[n]$ **to** network
    EMIT$(n, (\texttt{metrics}(\text{network}), \texttt{len}(d)))$
  end
Class `Reducer(`OBASERVATION n, METRIC $([([c_1, c2, \ldots], n), ([c_1, c2, \ldots], n), \ldots]))$
  diffusion_class $\leftarrow \{n > median; n \in$ METRIC $\}$
  test_x $\leftarrow n[10\%]$
  test_y $\leftarrow$ diffusion_class $[10\%]$
  train_x $\leftarrow n[90\%]$
  train_y $\leftarrow$ diffusion_class $[90\%]$
  model $\leftarrow$`LogisticRegression(`train_x, train_y`)`
  EMIT$(n,$ `accuracy(`model, test_x, test_y`)`$)$

The input of each `mapper` is the complete diffusion of an innovation in the form of an ordered list

of username and time of innovation. Each `mapper` then iterates over this list (line 3), computing the metrics at point n in the diffusions, then emitting the metrics at that point along with the final size of the diffusions (line 5). The key that is emitted is the absolute position within the diffusion n, with the final size of the diffusion being stored in the value part of the key value pair.

Each `reducer` then receives all records with the same value of n, which are all the diffusions that have 'at least' n observations. This means that the number of reduce tasks is the maximum value of n across all diffusions. Testing and training of the model is performed within the same `reduce` task, which means that, first, the data must be split into a testing and training set (10% and 90%, respectively). The final value that is emitted from the reducer is then the accuracy of the model at predicting the class of the diffusion based on metrics observed at n usages of all innovations.

Line 2 applies a function to the list of final diffusion sizes (all of which will be larger than n, which is passed to the mapper). This results in an array indicating whether the value is above or below the median of all diffusion sizes within the mapper, returning a list of binary classifications, e.g. [1,1,1,0,0,1,0]). Line 3 to 6 separates the respective arrays into testing and training sets. Line 7 trains the logistic regression classifier, with the accuracy emitted from the reducer in line 8. The accuracy is keyed by the observation level that the metrics came from (n).

One of the reasons for implementing the system in a `map reduce` pattern is to utilise the Hadoop framework. Traditionally, Hadoop applications are implemented using Java; however, the code developed for this system was implemented using `Python`. Python was chosen for a number of reasons, though mainly due to Python's extensive toolset for data manipulations and machine learning libraries, such as pandas[4] and scikit-learn[5]. However, there are no natively implemented (and maintained) scalable compute systems implemented in Python, so it is challenging to deploy on a cluster of machines. However, we can run Python on top of Hadoop using an additional framework. Therefore, for this research, all code was implemented within the MRJob[6] framework. MRJob not only interfaces with Hadoop's `map reduce` functions, but also manages code distribution and dependency installation on each node across the cluster.

In Chapter 7, the social networks were stored in an external HBase database. However, HBase is good for one-off queries (e.g. querying what are the neighbours of a node) but not for interactive computations such as distance between two nodes. For this work, NetworkX[7] is used to manipulate and study the structure of diffusions across each network. NetworkX is a Python library that is designed for the exploration and analysis of network structures. The core package allows for the representation of many graph types, such as directed and undirected graphs, along with enforcing the filtering of self-looping and parallel edges. NetworkX allows a number of network properties and transformations to be computed

[4] http://pandas.pydata.org/
[5] http://scikit-learn.org/
[6] https://pythonhosted.org/mrjob/
[7] https://networkx.github.io/

across a graph, such as the distance between nodes or a node's page rank, as well as dynamically adding or removing edges and nodes.

When used within our application, line 2 loads in the complete network in each mapper from a file in HDFS. Then, by iterating over the diffusion, nodes within the network are updated to indicate whether or not they have been activated. This then allows for metrics to be computed using an up-to-date network representation in line 5, using functions defined in NetworkX.

8.6 Experiments

The premise of this work is to assess the predictability of language innovation diffusion and, ultimately, language change. At each step within an innovation diffusion (n), the aim is to predict whether an innovation's diffusion will reach a size that is *greater* or *smaller* than a typical diffusion (*median*) of similar magnitude. Additionally, we assess how models that exploit information on the temporal patterns, network and community structure or grammatical context perform against a set of popularity-aware *baseline* features. Our findings are organised into three sections:

- **Predictability of innovation diffusions:** Our goal is to assess whether or not the adoption of a new word in the network is predictable using the application of the method proposed by [135]. We assess how predictability varies across different information signals and varying network structures.

- **Effect of structural holes in language diffusion:** We demonstrate the impact of *structural holes* and *weak ties* in language diffusion considering the two levels of network abstraction, *macro* and *micro* (section 8.3.1).

- **The role of semantics in language diffusion:** We discuss how the semantic context of the language innovation matters in comparing the performance of a diffusion. This model is based on the grammatical diversity of word usage against models exploiting other information sources.

8.6.1 Predictability of Innovation Diffusion

First, by assessing the diffusions of speaker innovations, we show that language is predictable, using only information up to the current state of a diffusion (n), with the method used detailed in section 8.3.3.

In Figure 8.6.1, we plot the respective distributions of diffusion sizes for the four networks we consider. Our experiments have shown exponent values to be within a .2 interval from exponent 2 and, given this approximation, we assume a similar prediction setting to **Cheng:2014kmb** As the classes are split on the median value of the typical diffusion size, this assumes a binary classification (implemented through the use of logistic regression) task, where classes are balanced and, as a consequence, the performance

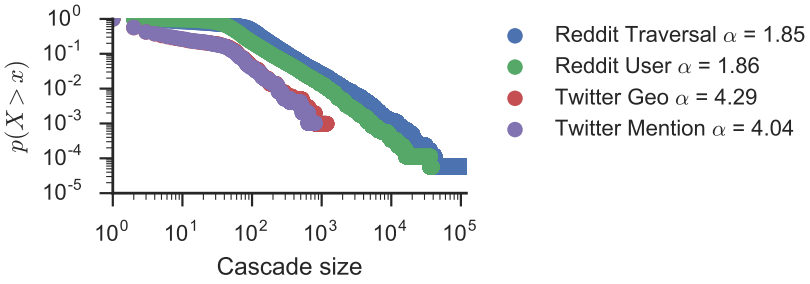


Figure 8.6.1: The complementary cumulative distribution function ($CCDF$) of diffusion sizes for each network.

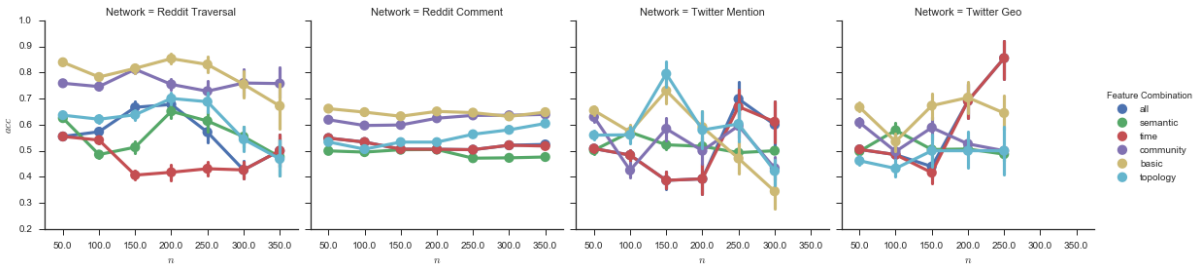


Figure 8.6.2: Average accuracy (acu) in predicting whether, after n observations, the final diffusion will be greater or less than the *median* of all diffusions as measured through the number of posts ($|P_n(o)|$).

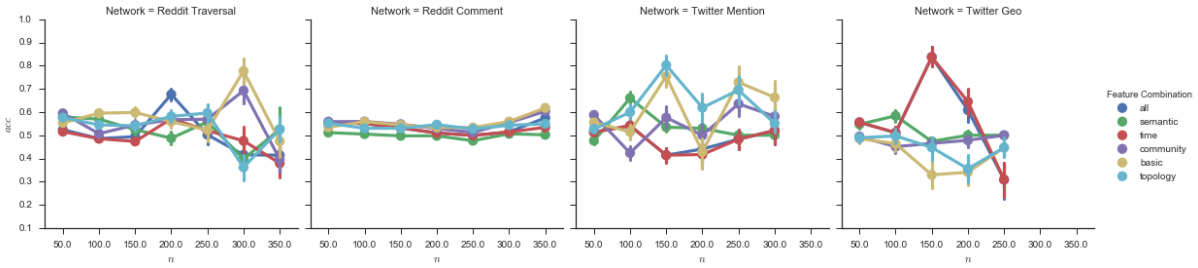


Figure 8.6.3: Average accuracy (acu) in predicting whether, after n observations, the final diffusion will greater or less than the *median* of all diffusions as measured through the number of unique users/nodes ($|A_n(o)|$).

of a random baseline will yield an accuracy of 0.5. Thus, a model's performance above this threshold implies that diffusion is predictable.

We assess two values across a diffusion: first, the final number of usages of an innovation ($|P_n(o)|$), and, second, the number of unique nodes that have used the innovation o ($|A_n(o)|$). We highlight that, unlike in **Cheng:2014kmb** we are not aiming to measure the size of a continuous cascade, but rather a diffusion across a network that many not have neighbouring activated nodes.

As our binary classifier, we use logistic regression with the evaluation being performed through 10-fold cross validation. During evaluation, we consider different models, one for each family of information sources (e.g. network topology, grammatical context). Finally, the accuracy score averaged across prediction tasks (of size n) is the metric employed to assess the effectiveness of each model.

Results: The model predicts, after n observations, whether the final size of the diffusion will be above or below the median value of all diffusions that have been used at least n times. Figure 8.6.3 represents the class prediction accuracy across all four networks. The Reddit networks (both user and traversal) achieve the highest accuracies consistently over the observation periods. Reddit's traversal accuracy peaks at 0.9 after 200 observations, with *basic* and *community* features performing the best. Similarly, the Reddit comment network's *community* and *basic* features consistently achieve the highest accuracy, though there is little variation over time, with values plateauing around 0.6.

The difference in accuracy between the *macro* and *micro* Reddit network could be attributed to the variation in network structures. The Reddit user network is significantly larger (see Table 5.2), which we would expect to be more expressive of language change. However, the content generated by each node is significantly less than the traversal network that aggregates content, which in turn could reduce/amalgamate a number of the external influences that affect user language adoption.

However, unlike Reddit, the predictability of the different models for Twitter networks appears to be less stable (potentially cause by its exponent, which cause imbalanced classes (see Figure 8.6.1)). However, when looking at the results, *topology* features perform best for early-stage diffusions (small vales of n) for Twitter mentions, achieving similar results to *baseline* and *temporal* factors. Across Twitter mentions metrics, there is significant variation in performance; this variation could be due to higher degrees of noise given the sparser nature of the underlying network structure (there are relatively few users who are mentioned in the dataset, with electively few edges). This sparsity can also be seen to affect the results in Chapter 7.

Twitter's geo network, which is built on regions of postcodes, again, appears to be a less predictable setting for language diffusion. Feature sets such as *all* and *community*-based features, within early stages of a diffusion, perform the best, with accuracies achieving a high of 0.75, though this quickly reduces below the 0.5 random baseline (Figure 8.6.3).

The large variation in accuracy (across feature sets and different levels of n) of the models could be

due to the structure of the network. The Twitter geo network has a maximum of 2,910 nodes, which means that at $n = 100$ then 5 % of the network would have been activated. Achieving repetitively large coverage in a short number of steps might not be an issue for other networks, though for Twitter the proportion of activity per postcode is skewed, with a large number of nodes concentrated in big cities that have large populations. This means that the distribution of content across the nodes is unbalanced, concentrating the effective network into a smaller subset of nodes, thus reducing the effective size of the network. This is compounded by the fact that only 4 % of the tweets are geo-tagged, and these are skewed toward more densely populated locations.

The predictability of the number of unique users (Figure 8.6.2) paints a similar picture. This is largely due to the fact that the number of users/nodes and the number of posts featuring an innovation are likely to be highly correlated. In the case of the subreddits network, accuracies hover above the random baseline 0.5, with *basic* features breaking the 0.6 boundary. For large diffusions, where $n > 200$, *basic* and *community* increase. In the case of adoption number, in the Reddit comment network, accuracy stays constant for different n values.

Temporal features across all networks frequently achieve accuracies below the 0.5 baseline. This is in contrast to research into the dynamics of meme and information diffusion online [185], [208], which identified that temporal aspects of a diffusion are highly important. However, [148] and [129] argue that language change is a slow and drawn out process, with some innovations spreading fast (like viral memes) as they are associated to events, but most taking an extended period of time as they are slowly integrated into people's vocabulary.

It is worth mentioning that the set of diffusions in the data collected is expected to be highly heterogeneous in nature, which, as with information diffusion in general, could be the cause of variations in model performance as well. Additionally, for some feature classes, accuracy decreases as diffusion size increases. This potentially indicates a change in the diffusion's form as it grows, which is in agreement with [207], who suggests that external influence on the diffusion process becomes more profound as it grows is size.

Overall, we have shown that *basic*, *community* and *topology*-based models appear to be the best predictors of language innovation diffusion. *Temporal* models have been less stable, potentially due to variations in the diffusion process. Additionally, the aggregation of content (*macro* networks) into the respective groups implies higher accuracy scores as some of the external influences are mitigated. However, the variation in the Twitter network's power law distribution results in classes being unbalanced, thereby affecting the accuracy.

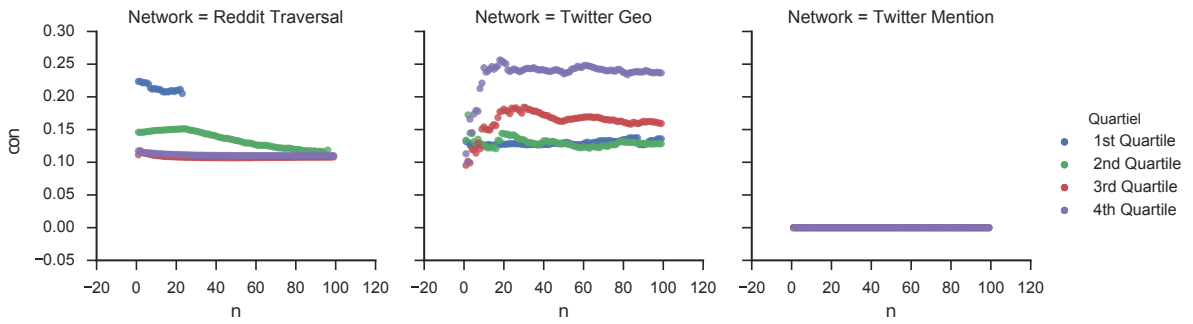


Figure 8.6.4: $\bar{Z_n}(o)$ computed on the first n observations of all cascades. The first quartile represents the smallest diffusions, and the fourth quartile the largest.

8.6.2 Effects of Structural Holes

[148] stated that weak ties and structural holes affect the diffusion of language innovations, as the users who bridge structural holes give the innovation access to different, far-reaching parts of the network. A node's (v) access to structural holes is quantified though a node's constraint Z_v, as introduced in section 8.6.2.

We would expect that larger diffusions would, on average, have a lower constraint compared to smaller diffusions. The idea is that nodes that are less constrained give the innovation access to the network, thus giving it a greater chance to diffuse access different communities. To quantify the effect of structural holes on diffusions, we first split each network's set of diffusions ($P_n(o)$) into four classes, based on their respective inter-quartile ranges of the final size of each diffusion ($|P_n(o)|$, classifying them from low to high diffusion ranges). The constraint of a node v is measured using Z_v (see equation 8.3.2), with the constraint of point n in a diffusion being the `mean` constraint of all nodes within the diffusion (see equation 8.4.8). The constraint at point n of a diffusion is clustered into its respective interquartile class, with the average then taken over all diffusion cases at each stage n.

The aim of this analysis is to see if larger diffusions, those within the upper inter-quartile ranges, have a lower constraint. If they do, this would indicate that one of the factors influencing their success in diffusion is access to structural holes. This would then ratify the beliefs of [148].

Results: Figure 8.6.4 represents the average constraint for diffusions after n observations across each of the networks. Each line (colour) represents differing clusters of final cascade size, as defined through inter-quartile ranges. Due to the high sparsity of the Twitter mention network, the majority of users had a constraint 0. However, for the Reddit traversal network, we can see clear differences in average constraint across the four final diffusion size clusters, even within the first 100 observations. Diffusions with a final size that falls into the first and second quartiles have the highest average constraint, whereas third and fourth have the lowest. As stated earlier, the lower the constraint of a node, the greater its access to structure holes, thus exposing the language innovations it uses in a greater proportion of the

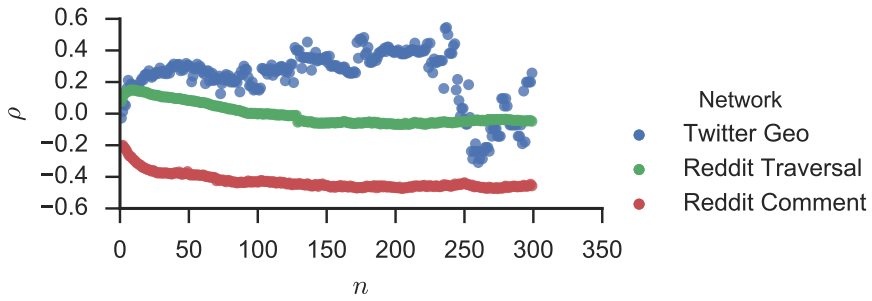


Figure 8.6.5: Spearman's correlation between the average network constraint and the community activation entropy.

network. As the diffusions with smaller sizes have on average a higher constraint, this indicates that the structure of the network could be constraining the potential for the innovations to diffuse.

Nevertheless, when considering the Twitter geographic a different picture is drawn. Despite the fact that, initially, all diffusion classes start with relatively low constraints, the fourth (largest diffusion quartile) quickly achieves the highest constraint values, whereas the first quartile (smallest diffusions) has the lowest average constraint. This appears counterintuitive, though we must take into account how the network is constructed. In the case of the Twitter geographic network, nodes that are heavily constrained are nodes in highly populated locations. These nodes produce a large proportion of tweets (thus a large usage of innovations), whereas less constrained nodes are in rural locations and thus produce fewer tweets and less activity. Intuitively, we would expect that the diffusion would start in the highly populated locations and travel to the rural ones, thus causing larger innovations to have a greater average constraint.

Additionally, as access to structural holes increases (by decreasing Z_n values), we would expect that the diffusion would start spreading to different communities. An innovation spreading across communities can be examined by measuring the activation entropy of a diffusion (Equation 8.4.10) or the proportion of activated communities (equation 8.4.9). Therefore, to assess this relationship between a diffusion's constraint and community usage, we quantify the correlation using a Spearman's rank. This measure of correlation (ρ_n) is between community entropy (E_n^o) and the constraint (C_n^o) after n observations across all innovations (o). A value nearing 1 would indicate that the higher the constraint, the higher the entropy in usage across communities, with -1 suggesting that the lower the constraint, the higher the spread across communities.

Figure 8.6.5 shows the variation in correlation across the four different networks. When considering the Twitter geographic network, the initial correlation indicates that, as community diversity increases, the constraint increases (in line with larger diffusions that have an increased constraint). For Reddit traversal (subreddits), there is initially a positive trend, indicating that higher constraint potentially indicates greater spread across the network. However, this correlation turns negative for larger diffusions,

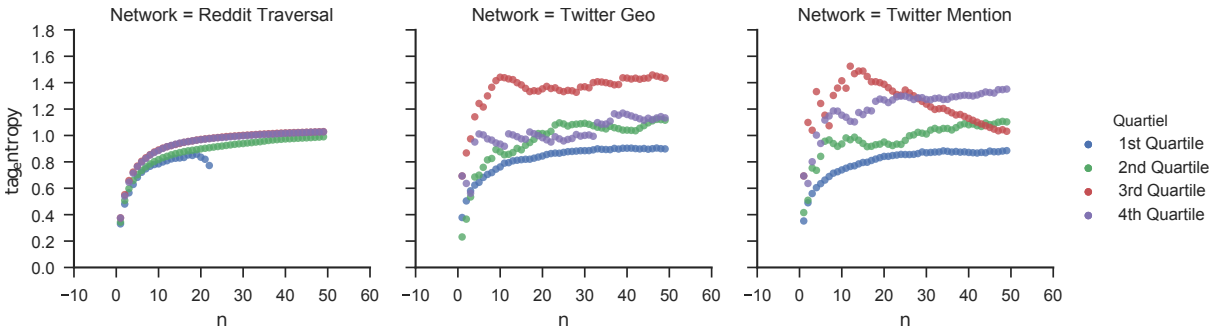


Figure 8.6.6: Average Part of Speech (POS) tag entropy $\bar{Q_n}(o)$ computed on the first n observations of all cascades. The first quartile are the smallest diffusions, with the fourth quartile representing the largest diffusions

indicating that an increase in access to structural holes gives greater access to a diverse set of communities. This trend is also seen, to a greater extent, in the Reddit user comments network.

A transition to a negative correlation (lower constraint, higher entropy) is what is expected by [148], as, through an increase in structural holes, the innovation has a greater chance to spread across the network, thus increasing the entropy across the network.

8.6.3 Language and Context

Finally, [18] suggested that, for an innovation to be successfully adopted by a community (measured by the proxy of adding it to a dictionary), it must be able to be used in a variety of contexts and situations. When applied to the work in this chapter, this means that, for innovations to achieve a larger diffusion, we would expect there to be a larger number of contexts in which it is used. For this work, we use the associated Part of Speech (POS) tag as a proxy for context, as an innovation being used as a *noun* or *verb* indicates users' willingness to play with an innovation. To quantify the varying contexts in which the innovation is used, we use tag entropy (see equation 8.4.12). A low entropy value would indicate that the innovation is used in the same context, whereas a high entropy would indicate the word is used in a variety of contexts. Again, innovations are clustered by binning them into their respective inter-quartile ranges based on their final diffusion size, with the average tag entropy (Q_n) at n being commuted within the cluster (the same method as in section 8.6.2).

Figure 8.6.6 represents the variation of entropy across the first n observations of innovations in each network. We can see that, across the two Twitter networks, there is a noticeable difference in mean tag entropy, with larger diffusions (the third and fourth quartiles) having a greater variation in assigned POS tags compared to smaller diffusions. This same pattern is additionally seen in diffusions across the Reddit traversal network, thus suggesting an increase in diversity in the first 100 usages of a term, which is a potential indicator that the innovation will ultimately diffuse the furthest and therefore have the greatest chance of causing language change.

8.7 Discussion and Conclusion

In this chapter, we examined the predictability of language innovation diffusion and language change, as well as quantifying the effects of structural holes of the diffusion of innovations and how, as diffusions grow, the grammatical contexts of an innovation vary.

In assessing the predictability of innovations across each of the four networks, the method proposed by [36] was adopted. [36] stated that we should not attempt to predict the *final size* of a diffusion (e.g. predicting that an innovation is used n times), but rather at each stage in the diffusion (n), we should predict if, out of all the innovations at that given point (n), the respective final size will be in the top 50 % of final diffusion sizes. Proposing this slightly different question results in a model that can be trained on balanced classes (as long as the exponent, α, follows a power law of 2; see section 8.3.1 for further explanation), correcting for some of the issues found in previous research (see section 8.2).

[36]'s method was applied across both *micro* (inter-user interactions) and *macro* (inter-group interactions) representations of social networks. Initially, the separation of network structures came from the understanding that language change not only originates from a single user but rather from the collection of users in which the language and innovation is used. At a high level, across these two network abstractions, the method shows that language innovations, to an extent, diffuse in a predictable and similar manner. However, each of the different social networks and network abstractions show varying results.

Additionally, [36]'s model was also applied to two measures that can quantify the size of the diffusion: the number of unique usages ($|A_n(o)|$) and the number of usages ($|P_n(o)|$). Focusing on the two different social networks (Reddit and Twitter), we can see distinct variation in accuracies, across both the number of unique node activations (Figure 8.6.2) and the number of innovation usages (Figure 8.6.3). We can see that, across the Reddit network, higher accuracies are achieved in predictions based on the number of usages of an innovation and not the number of nodes, with similar results seen across the Twitter networks. This potentially indicates that innovations are highly specialised to the communities that use them, with only a few spreading successfully.

However, as n increases, we can see differences in the variation of accuracies across both Reddit and Twitter. For Reddit, we can see limited variance in accuracy across the two network abstractions (and across the two respective measures) for varying feature classes. For Twitter, there are large swings from one extreme to the other, with large variances at later stages. These large variations are due to the size of the underlying networks, with the Reddit network having in excess of 861,955 nodes and the Twitter geo network limited to 2,910. Thus, an innovation being activated over 200 nodes across the Reddit networks represents only a small number of users, whereas for Twitter geo, it represents 10 % of the network activated (which is not seen in Reddit). Thus, when quantified in proportions, in the early stages of diffusion, as the number of activated nodes increase, the diffusion becomes more predictable. However, there is a point at which the predictability of the diffusion of an innovation decreases. As

mentioned earlier, both the Twitter mention and geo networks' results are limited in nature. For the Twitter mention network, these limited results originate from the sparsity of the network, due to the large number of nodes but limited number of edges. This effect of the network can be seen in both the use of *topological* feature classes, and the near 0 value for constraint (Figure 8.6.4).

This does not detract from accuracies as high as 0.9 in predicting if the diffusion size will double; although, depending on the feature classes used, this accuracy varies highly. Overall, the model was trained and tested on six different feature classes, each representing varying measurable aspects of the diffusion process: *basic*, *temporal*, *topological*, *community* and *grammatical*. When predicting the diffusion of an innovation, the different feature classes produce varying rates of accuracy across the four variations of networks. This difference can be seen in the Reddit traversal network (Figure 8.6.2), where the *basic* and *community* feature sets show the greatest accuracy in predicting if the diffusion size will double (highs of 0.85 and 0.80, respectively), and *temporal* the least (high of 0.55). This shows a potentially more important influence of the community and network structure on the diffusions of innovations than time. This addition validates the claims by [45], who states that language change and innovation diffusion happen over both long and short time frames, thus attempting to measure *temporal* features could result in a feature set with high variations in quantities.

By drawing on the grounded work of [148] and [149], we proposed that, by measuring the constraint of nodes within the diffusion, we could measure the effect of structural holes within the process of language change. The hypothesis was that, as innovations were used by less constrained nodes, then these would expose the innovation across structural holes, thus allowing them to diffuse to other parts of the network. As can be seen in Figure 8.6.4 for the subreddit and comment networks, we can see that diffusions with smaller final sizes do indeed have larger constraint. However, for the Twitter geo network, it is larger diffusions that have a greater amount of constraint, which comes from innovations being used in densely populated locations and not the sparsely populated locations that represent the structural holes within the network. Building on the constraint of nodes within a diffusion, we looked at the correlation between a constraint and the number of communities that had used an innovation (Figure 8.6.5). We showed that, for the Reddit networks, there is a negative relationship between the number of communities and the constraint of a diffusion. This, again, ratifies the work of [148], showing that, for innovations to diffuse across a network, they must have access to structural holes.

Finally, drawing upon the research first identified in Chapter 6, [18] proposed that a measure of an innovation's acceptance (whether to include it in a dictionary) is the diversity of context and variation in morphological form. Within this chapter, we used entropy of POS tags as a proxy for variation in context. We proposed that, as diffusion sizes increase, the tag entropy will increase as the innovations are used in an increasing number of contexts. As can be seen in the results (Figure 8.6.5), across the four network, smaller diffusions have a lower entropy as they grow. However, for larger diffusions, the results

are less distinguishable, potentially indicating a limit in the number of contexts in which an innovation can be used.

As highlighted, there are limitations to this work; for Twitter, these come from the distribution of innovations that affect the balance of classes in the prediction task, along with the sparsity of the Twitter mention network. Additionally, across both network abstractions and data sources, we cannot categorically say the diffusions cause language change, as this requires extended usage over time, which is not assessed in this model.

Overall, this chapter aimed to show how predictable an innovation's diffusion across a network was, not only at the user level but also at the group level. For the Reddit network, this was evident, with accuracies of 0.9. Additionally, we showed that, for an innovation to diffuse, nodes that had access to structural holes had to use the innovation. However, the results for the Twitter networks varied due to the sparsity of the mention network and the nature in which the geo network was constructed. However, across the board, we can see that innovations do diffuse in a predictable manner, with features such as *basic* and *community* being the best at predicting the diffusion of an innovation.

8.8 Data Access

All code created for this research Chapter is openly available on github at [108].

Chapter 9

Results and Discussion

As introduced in the research framework (Section 4.3), this thesis is characterised by one overarching research question, which is broken down into three distinct, independent research questions (with the fourth being a technical question). Each of the three research questions focuses on one aspect of measuring and modelling language change across OSNs. Ultimately, it is through the combination of the results of these three questions that we can then answer the overarching research questions: **How can we forecast language change in online social media, and what are the factors that innovation diffusion depend on?**

By assessing this overarching research question, we developed three sub- questions by means of the application of structuration theory [74]. Framing this thesis in structuration theory allowed us to break down language change into three distinct components that could then be assessed: the *innovations*, the *user* and the interactions of the two through *social structures*. This resulted in the three respective sub-questions and one technical question:

1. How do you detect language change in online social networks?

2. What is the role of social constructs in language innovation and use within online social networks?

3. How does network structure influence the diffusion of language innovations?

4. How can such research be conducted at scale?

The following chapter will revisit the literature within the context of each research question, highlighting how they are related to the respective results of each chapter. Additionally, we will highlight limitations in both the data collected and used, and the methods applied to each question.

9.1 Results Summary

The following section summarises the methods and results of each chapter in its ability to answer the given research question. Additionally, the results from each research chapter are critiqued in relation to the original research questions and related literature.

9.1.1 Research Question 1 - Detecting Language Innovations

The first research question focused on the innovation itself, and how we could detect and model the innovation over time. Previous research into language innovations has shown that innovation identification, disambiguation and modelling are of growing importance as OSN text dominates NLP tasks. Efforts have ranged from looking for the source words of new word blends in OSN [40], [42], to the use of generative models to model regional language variations [64]. The recent growth in interest can be seen as coming from two fields of thought: first, increasing accuracies of traditional NLP tools in relation to noisy data [75]; and second, trying to understand the innovations themselves as they contain vital information about the user and communities in which they are used. This first research question sits between these two fields of thought, by trying to understand the regional differences in language and how their growth can be detected.

This first research question (Chapter 6) focused on the detection of innovations and their subsequent growth across the two OSN. Its focus was on how to quantify and detect the growth of language innovations. To achieve this, *VERGT* [18] and *FUDGE* [144] were operationalised, and treated as a general time series framework to detect growth and death in language usage. The developed methods not only quantified the growth in popularity of language innovations, but also assessed variations in the morphological and contextual meanings of the innovations. However, this is not the first time that these heuristics have been adapted for use in a computational model, as [130] manually rated a number of innovations across the categories highlighted in the original framework and then used an SVM model to predict if the innovations would be added to a subsequent edition of a dictionary.

The two sets of heuristics led to the development of three computational methods that quantified change in the popularity and morphological form of innovations over time; these were: *variation in frequency*, *diversity in form* and *convergence in meaning*. Each measure attempted to subsume parts of the grounded heuristics. To rank the growth of the innovation (in relation to each metric), a Spearman's rank was fitted against the given metric, with a positive value indicating growth of the innovation and a negative value indicating the death of the innovation. Growth was seen as significant if it was above the 95^{th} percentile, with death being classified if the negative growth was in the bottom 95^{th} percentile.

The developed methods were applied to two different network abstractions (macro and micro) of Reddit and Twitter; this allowed for the analysis of language change at both user and community levels.

The results varied in the ability to detect new innovations across both the network abstractions and the metrics. The first two metrics, *variation in frequency* and *variation in form*, showed positive results, surfacing innovations such as *fleek* and *grexit* from the Twitter geo dataset, and detected more specialised innovations, such as *csgo* and *bruh*, when applied to Reddit. Additionally, the methods were applied to smaller subsets of the networks, representing the individual communities such as regional and subreddit subsets in the network, which allowed for community-specific language to be identified. Variations in regional language can be seen in locations such as South Wales with words such as *flook* and *depay* achieving prominent growth. Again, when applied to smaller communities within Reddit, such as *AMA*, innovations such as *commenter* and *sfw* can be identified. However, in measuring the *convergence in meaning*, the results were highly susceptible to underlying variations in a term's popularity caused by events in the news and other media sources. This was seen in the term *Ebola*, which saw spikes in popularity due to the global outbreak occurring during the data collection period.

In relation to the original research question, the results indicated that we can detect innovation growth and death through the application of simple and effective metrics. However, developed methods such as *convergence in meaning* are susceptible to sudden changes in popularity, which may not constitute language change but rather a population's reaction to an external event. Additionally, the methods used to answer the first question differ from that of previous research, initially through the use of *VERGT* [18] and *FUDGE* [144] as a grounding framework, which allowed for the development of simple and explainable methods.

9.1.2 Research Question 2 - Social Constructs on Language Change

Language that individuals use is influenced by the people and communities around them. This ultimately means that language can be defined through the selection of the users who use it and not the words and sounds themselves. Thus, the changes in a user's language are influenced by the users around them. This effect of the community influencing language change was identified in [45], although Croft proposed that the creation and subsequent propagation of innovations across populations is both intentional (normal replication) and accidental (altered replication), whereby the speaker generates the innovation and users and/or communities accept or reject it. Building on the ideas of information propagation, [148] proposed that users and communities have varying degrees of power over each other to influence which and what language they and others use. This effect of influence and control over the language individuals use can be seen in the act of users accommodating their language in relation to the person with whom they are communicating.

Domination and power within OSN has been modelled using computation methods before, aiming to mine the inter-user influence through past interactions: [83] showed that power and influence between users can be learnt as a function of past action cascades (the user performing the same actions as

people in their ego network). When looking at language and power, [188] showed that users were more likely on Twitter to accommodate their language dependent on centrality within the network structure. Additionally, [51] showed that, through the use of probabilistic models, we can distinguish between homophily and topic-based accommodation. However, [53] researched users' language accommodation over their lifetime within an online forum, showing that users are more likely to accommodate their language when first entering an OSN. However, as the user is about to leave the OSN, their language diverges from that of the community. These studies can be seen to model aspects of power and domination within users' language change and adoption, thus reinforcing that users base their language on the person with whom they are connected/communicating.

This research question focused on predicting users' language adoption from their exposure to innovations across their ego network. To model the process of users adopting an innovation based on their neighbours, we took influence from Granovetter's general threshold model [86], for which [83] proposed a scalable computation implementation. The aim of the model was to learn a global threshold (σ) at which users would adopt an innovation based on pressure from users within their ego networks. To implement this model, inter-user influence was quantified as a function of historic innovation diffusions, with the resulting influence values being used to quantify the pressure a user is exposed to upon a connection adopting a language innovation. A model was then trained and tested across both *macro* and *micro* representations of the Twitter and Reddit networks, thus showing that the processing through which a community and user adopt or reject an innovation is similar.

The aim was to identify a global threshold (σ) that, when breached, would mean a user would adopt an innovation. The threshold was computed using a ROC analysis, which meant that a resulting AUC of more than 0.5 signified that the model would be better than a random baseline. As seen in the results (Figure 7.6.2), varying levels of accuracy were seen across the four networks. However, the majority of networks achieved an accuracy greater than the 0.5 random baseline. Due to high network sparsity, the Twitter mention network achieved the worst result. For the remaining three networks, a models that quantified influence between users as static achieved AUC's as high as 0.91. The results also indicated that users follow the language of the more powerful users within the network, adopting terms and innovations after exposure from more influential users within their respective ego networks.

To further understand the extent to which a user's ego network influences their language adoption, and what proportion comes from sources that are external to the network, we shuffled the edges and re-ran the models. The results showed that the network structure contributed up to 25 % of the predictive accuracy, leading to the conclusion that the social structures within OSNs constitute 75 % of the pressures in influencing a user's language. However, when testing to see if community structure internal to the network was influential to a user's language adoption (Section 7.6.2), we could not see any statistically significant difference in the results. This potentially indicates that, even though, traditionally, community structures

have been highly influential in moderating the language individuals are exposed to and adopt [45], users now have the ability online to observe the whole network, thereby reducing the barriers to the number of innovations and the community-specific language to which they are exposed.

Not all innovations fulfil the same function within language. Therefore, to assess whether the context in which the innovation was used also affected its adoption, the given context of each usage of an innovation was determined by a POS Tagger, which associated each innovation with a function in language (e.g. noun or verb). The aim of this was to accesses whether the accuracy of the model varied depending on the function of the innovation. The results indicated that the context of the innovation appeared to influence the accuracy of the models, indicating that open class word adoption was easier to predict (as indicated through higher AUC). The increase in accuracy for open class words potentially comes from them being more expressive of the conversation in which the innovation is being used.

The original question in Chapter 7 focuses on using social constructs such as domination and power to model the process of language innovation adoption at the user level. The results indicated that we are able to model the influence between users (social domination and power) as a function of their past join actions; the results were then used to quantify the pressures applied to individuals from their ego networks to adopt an innovation based on a general threshold model [86]. Ultimately, we can infer that, within OSN, influence and domination pay a key role in the process of individual language adoption. However, explicit relationships and interactions account for only 25 % of the influence, with the remaining coming from external unmeasurable sources. Finally, we can model social constructs on OSN data and predict language adoption. However, there are many other external pressures that cannot be observed that affect the processes of individuals and communities.

9.1.3 Research Question 3 - Structural Influence on Language Innovations

The final question investigated how the size of an innovation's diffusion can be predicted across OSNs, and not when individual users will adopt an innovation (as answered in the previous question). This question ultimately looked at the network structure and its influence on the diffusion of innovations across its nodes. This drew on the grounded work of Milroy [148], who showed that the diffusion language innovations between communities are heavily influenced by their access to structural holes, as this gives the innovations access to a greater proportion of the social network. Milroy's work mirrors that of [27], who showed that structural holes mediate the flow of information through networks. This effect of structural holes has been shown to influence the diffusion of English in becoming a dominant global language, compared to other languages such as Icelandic. In addition, structural holes explain the merging of Northern Ireland dialects through points of contact (social clubs) between distinct rival communities.

However, the diffusion of language innovations bares similarities to meme and content diffusions

across OSNs, following a similar power laws. [135] identified that, when predicting the diffusion of memes, there is a positive correlation between network and temporal features with respect to the predictability of the diffusion. In this question, the model proposed by [135] was applied to predict innovation diffusion across OSN, as the model utilised the power law distribution of final diffusion sizes. This means that, instead of predicting the final size of a diffusion, the model asks the question, 'will the final size of a diffusion be above or below the median of all diffusions that have at least n observations?'. Using such a model results in a binary classifier that can be trained on balanced classes (will the diffusion double or not?), reducing over-fitting of the model to rare large diffusions. The model developed was trained on a number of feature sets (e.g. features drawn from the network topology, time metrics across the diffusion and the contextual usage of the innovation across communities). The results showed that, across the four networks, innovations diffuse in a predictable manner, achieving accuracies with highs of 0.9; the highest accuracies were achieved when using the network-based feature sets. These high accuracies in predicting innovations' diffusion processes based on network features show that it is the path the innovation takes that influences the final size of the path it takes.

To investigate the influence of the network structure on the diffusion of innovations, we looked at what effects structural holes have on the diffusion of innovations. The results indicated that structural holes have significant effects on the size of the final diffusion as innovations with increased access achieving larger diffusions on average. This was reaffirmed through the negative correlation between the number of communities (as detected through a community-detection algorithm) and the average constraint. The results also showed that, for an innovation to successfully diffuse, it must have access to users who have access to structure holes.

As seen in Chapter 7, temporal features appear to play a limited role in both user adoption and diffusion over a network. The bases of the model used for this question [135] showed that, for memes diffusion, predictability was dependent on temporal features, though this is not the case with language innovations. This can be attributed to innovation diffusions (and ultimately language change) accruing over both long- and short-term time periods, whereas memes online predominantly diffuse over short time windows. In relation to the original question, which focused on the influence of the social structures (the networks), the results showed that language innovations diffuse in a predictable manner, and larger diffusions that have access to structural holes will diffuse further as these users and communities have greater influence over a larger proportion of the population's language. Additionally, this question has ratified the work of [148], [149], whose original studies identified that structural holes influence language change and diffusion, allowing them to diffuse between communities. Ultimately, innovations across OSN diffuse in a predictable manner.

9.1.4 Research Question 4 - Implementation with Scalable Technology

Recently, there has been a movement from data-sparse to data-rich research, whereby the researcher has access to not only a sample of data but also the complete set. This raises questions about how we can process the data in a scalable, effective and replicable manner. Historically, researchers have developed custom code and libraries that they can run on their laptops or a large computer in order to process and model data for their research. These have been implemented using programming language such as *R*[1] or *matlab*[2]. These are specialised or proprietary languages; therefore, even if we have access to the code base, we either will have limited knowledge to run the code or need expensive licences to do so. Additionally, systems such as *R* and *matlab* were not designed to be easily scalable across multiple machines, with their functionality only existing in specialised libraries to allow for specific operations to be paralysed across multiple machines. For this reason, as we are dealing with big data, we focused the implementations on using open-source big data frameworks that have been designed to distribute workloads across clusters of machines.

Focusing on big data led this research to use three core technologies: Apache Spark, Apache Hadoop and Apache HBase. The first two are processing frameworks and the last is a storage and online query framework, and all three come from the Apache big data family. Each of the frameworks is open sourced, and each has an active community developing and supporting the multiple applications and frameworks on top of it. In each of the chapters, we detailed how each of the metrics and models is implemented, as well as identifying any constraints that we encountered in using them.

The technologies were used in different ways; for example, the social networks stored in HBase in Chapter 7 had to have their layout optimise in order to improve the query time. Whereas, in Chapter 6, we had to use Reddit as an intermediate data store for large joins across multiple dimensions. Additionally, the first two chapters were implemented in JAVA on top of Spark, whereas Chapter 8 used Python and Hadoop. By using Python and Hadoop over Spark we were able to use the popular network library NetorkX to mine features from across each of the four OSNs, and also paralyse and use `scikit-learn` to implement the logistic regression model.

In relation to the original question, which focused on whether we could implement the research questions in a scalable manner using big data technology, this thesis has shown that, with relative ease, we can implement systems in a scalable replicable manner, using highly popular big data technologies such as Spark and Hadoop. This meant that we could utilise not only a sample of the data but the whole dataset for each chapter. Ultimately, we have shown the benefits of using these given technologies, but also some of the caveats, such as memory and storage issues, that must be considered when using them in future technologies.

[1] https://www.r-project.org/
[2] https://www.mathworks.com/products/matlab.html

9.2 Overall

Aspects of each of the research questions can be seen across each of the three core research chapters. Thus, the results of the three research questions are now taken as one, with four key observations taken from them.

- **Language innovation is expressed through both micro and macro network structures** - By abstracting social networks with abstraction representing both user and group interactions, this thesis was able to show that language diffusion is not only an inter-user phenomenon but also an inter-community one. For question 1.1.1, the effects of both these micro and macro interactions can be seen in the growth of regional-specific words (representing the growth of community-specific language) and global words (representing the growth across communities). Whereas, when developing predictive models in questions 1.1.2 and 1.1.3, we could see that macro representation (inter-community interactions) achieved higher accuracies over inter-user representations. This may suggest that individuals adopt the language based not only on their neighbours, but also the communities (subreddit and geo-locations) in which they exist, though communities only adopt language based on their neighbours. This means that, for an individual, even though they may not have direct exposure to an innovation, they have indirect access to the innovation through the neighbouring communities to which they belong.

- **Influence comes from sources that are both internal and external to the network** - The language that individuals use is not only used within one community but across many, though with slight variations. This means that the language used within OSN is also used and observed offline. Thus, when developing the predictive models in Chapters 7 and 8, the reasons why a user may use an innovation may stem from sources that are external to the collected datasets. By randomising the networks in Chapter 7, the results showed that 25 % of the influence that causes a user to adopt an innovation comes from sources that are external to the network, representing interactions that cannot be mined or modelled. Even though we are not predicting user actions directly in Chapter 8, the external influence will still affect the results, possibly suggesting that 25 % of the underlying features are not representative of the 'true' diffusion process online. This large effect of external influences means that language change is inherently different than suggested in similar research that has focused on modelling the diffusion of explicit content across social networks.

- **Network structure influences which innovations a user has access to** - Chapters 7 and 8 focused on the effects of the social structures in which users and communities exist have on the diffusion of language innovations. The results indicated that users and communities who span structural holes influence the final size of an innovation diffusion, to a greater extent, as they allow the innovation greater access to far-reaching parts of the network. Ultimately, this means that, if

we were to artificially create an innovation, we should target users who have a low constraint in order to achieve the highest impact on language change.

- **Language change happens over both the short and long term** - Across each of the questions, time plays a mixed role in assessing language change. The results appeared to indicate that, as language change happens over both short and long periods of time, thus it is challenging to include temporal features in the developed models. For question 1.1.1, this can be seen in the inclusion of both the slow and fast growth of innovations being classified as significant changes, whereas, for 1.1.2 and 1.1.3, it can be seen in the variation (reduction in accuracies) in using time features within the models.

  These variations in speed of diffusion across both users and communities raise the question of what constitutes language change and evolution. Is it defined by the speed of the adoption, or is it the long-term usage of the innovation? The conflicts between time scales highlight the potential difference between language change and the diffusion of information (e.g. topics of interest). Even though both forms of diffusion are bound by the actions (e.g. generation of text) that can be performed in the network, they are fundamentally different. Therefore, when assessing language change, we should distinguish between the long- and short-term changes, and not aim to combine them into one, as has been done in this thesis.

- **Language change process can be modelled by the operationalisation of historic models** - Across this thesis, classical models and theories have been drawn upon to assess and model language change. In Chapter 6, this was the application of lexicographical models to quantify innovation growth. For Chapter 7, this was seen through the development of threshold models proposed by [86] and [196]. In Chapter 8, this was seen in the application of structural hole theory [27], and measuring network effects on innovation diffusion. In this way, the results of this thesis could be related to grounded research of the past, but also show that existing theories and hypotheses are still relevant within current contexts. Additionally, with the movement from data-sparse to data-rich research (section 4.2), there is a need to understand how we can translate traditional models in a scalable manner to deal with big data. This work has shown how we can use popular open-source big data technology, Apache Spark, HBase, Hadoop and Hive time mine metrics, and learn models in a scalable manner, distributing the computational power across clusters of computers and storing data in optimised queriable formats.

These four observations from across the three core research questions answer the one over-arching research question:

We *can* predict the diffusions of language innovation across OSN. However, as shown, there are numerous factors on which the diffusion of language innovations depends. For networks such as Reddit,

which are explicitly defined and follow common small world structures, we can say that the community heavily influences the language diffusions, as seen in the growth of community-specific words, and the large diffusions when there is increased access to structural holes. However, for geographic networks, we have to take into account the distribution of users and their movement across geographical locations, as this heavily influences the network construction and innovation diffusion as these movements influence who users come into contact with and thus what they are exposed to. Additionally, we need to distinguish between long- and short-term language change, as, within the current models, both are treated in the same way, thus introducing a large proportion of noise, and reducing the accuracy of the model.

9.3 Limitations

This work has a number of limitations, which can be divided into two: issues with the underlying data, and issues with the implementation/method.

9.3.1 Data

Social media data is inherently noisy, with the noise originating from the inclusion of Uniform Resource Locators (URLs) or hashtags in the text, to users being bots and corporate accounts; each of these adds a layer of complexity when dealing with the data. This meant that the first challenge was identifying what was considered to be an innovation. For this, we used the BNC as a baseline for English (filtering out words that appear within it), and also removed features such as emojis and emoticons, to name a few, along with some light text normalisation. However, this still left a significant amount of noise within the datasets, as can be seen in Chapter 6, where French words appeared within the data. This came from a combination of a large French-speaking population in the south of England, and the inclusion of the Channel Islands within the bounding box when collecting tweets.

Additionally, the influence of bot and corporate (not individual) accounts can be seen in Chapter 6, which influenced the number of usages of an innovation, thus affecting measures that relied on frequency as an indicator of acceptance. For *variation in frequency*, we could see that innovations such as *mph* were classed as 'significant' as automated weather stations tweeted out weather conditions. We could also see the influence of users re-tweeting content in the convergence of meaning through the growth of the word 'Ebola'. Some of the influence was rectified through the addition of metrics that counted the number of unique users per time period, thus quantifying language change as a function of the number of users of an innovation in each time period. Additionally, this effect of multiple usage of an innovation, such as re-tweeting, not only affected Chapter 6 but also 7 and 8, as the models based on the Twitter networks could be predicted if a user was going to re-tweet content based on their exposure, raising the question of whether users re-tweeting can be considered an adoption of an innovation.

The issue was not only with mining the innovations themselves, but also mining the interactions between users. Reddit user interactions can easily be mined as they are explicitly defined within the data and all users exist within the dataset. However, for Twitter, the data collected was a sample of all traffic, with constraints meaning that it contained many partial interactions. These partial interactions come in two forms: first, incomplete sets of users' tweets, as users do not always tweet with GPS; additionally, if a user mentions a fellow user, that user also needs to have tweeted with GPS coordinates turned on. This means that we cannot fully extract the true interactions between users, only identify a fraction of them. However, the data collected from each network does not contain all the interactions that a user has, but rather only explicit interactions in which the user generates content (tweets and posts). Thus, if a user observes content without generating content themselves in the reactions, it is not logged as an interaction. This effect of only using only explicit interactions has two effects on the research. First, the micro and macro network abstract may not be representative of how, at a fundamental level, users *observe* and travel through the network, but only represents what users react to. Second, models that use only explicit interactions have a number of implications. First, for question 1.1.2, when computing based on exposure, if a user is going to adopt an innovation, this means that we might not be capturing all the instances of an innovation to which they may have been exposed. For question 1.1.3, the topological and community-based measure may not represent the correct distance between nodes, which could mean that communities who observe each other, without generating content, may appear to be far apart in the network, but in fact should be closer to each other. This underrepresentation of user interaction and exposure to content could, in question 1.1.2, account for some of the 25 % of influence that cannot be modelled from the underlying data.

9.3.2 Methods

The methods used across the three research chapters varied. Chapter 6 focused on quantifying change, whereas Chapters 7 and 8 focused on predictive models. However, there are limitations in each of the methods used, originating both from limitations in the data but also limitations in the model and the analysis performed in answering the questions.

Chapter 6 focused on the detection and quantification of innovation growth and death. The methods used, however, were susceptible to sudden changes in popularity that may not be representative of language change but rather just a collective interest in an event. Additionally, assessing the prefix and suffix additions may have been over simplistic, and potentially better results may have been achieved using state-of-the-art stemmers and systematisers, as using the Levenshtein distance between small strings resulted in potentially large clusters of unrelated terms.

When looking at the networks and how language moves around them in Chapters 7 and 8, for the majority of time, they were treated as static objects (with the exception of edge addition when modelling

user language adoption). Thus, when predicting the size of diffusions in Chapter 8, each edge was presumed to always be there, even though it may not have been at the time of the diffusions. Treating the networks as static could have affected the results significantly, as social networks are inherently dynamic places, with structures that change as user interactions pattern with other users, as well as location changes.

However, when computing the global threshold in Question 1.1.2, each node/user had the same value. This breaks away from some of the work of Gravonetter [86], who believed that the activation threshold was unique to each user, with the global threshold then being the mean of all activation thresholds. Computing a global value meant that we could not distinguish the innovators (low threshold) or lagers (high threshold) in language innovation and adoption. If we had explored the local threshold, we could have explored the effect of thresholds in relation to a user's position in the network. As highlighted in section 2.3, the threshold of each user is relative to their local community and their position in the network [86], with users who bridge structural holes having a relatively high threshold as they 'manage' information diffusion across the network, and thus only a finite amount of resources in order to reproduce the innovations. However, one of the benefits of using a global threshold compared to an individual threshold means that, if the model is deployed, then there is no need to re-run the model every time a new user enters the network as they will user the global threshold.

For Chapter 8, the issue with the method came from the data. This is because the size of the diffusions across both Twitter networks did not follow a power law exponential of 2. This meant that, when training and testing the predictive models, the classes were unbalanced, resulting in a model that also learnt from the distribution of examples and not reservedly the features within the model. This could have been corrected by sampling only innovations that started in specific regions of the network, or defining the class boundaries differently and attempting to correct for the imbalance of the classes.

9.4 Conclusion

In this discussion chapter, we have brought together the results of the three research chapters. The aim of these chapters was to answer each of the four research questions, but, in the process, they aided in answering each other's questions. We first identified the three questions by breaking down the process of language change into three measurable components: detecting the innovation, predicting user adoption and predicting the adoption at the network level.

The results identified that we can quantify and predict language change. Breaking language change into its components, the results showed that individual innovations' growth/emergence can be modelled with relative ease through the application of time series measures that quantify the variations in popularity, the probability of morphological variations and convergences in context. For the usage of the

innovation (at both the user and community levels), the results indicated that the adoption of the innovation is dependent on the ego network applying pressure to the node to adopt the innovation. This drew on the concept of domination that exists between users within the social network, where users accommodate their language to each other dependent on structures of domination and power. Finally, we looked at the social structure and its influence on the diffusion of innovations across the network, showing that structural holes allow the innovation to gain access to further parts of the network.

However, the results indicated that we may need to have a better understanding of what language change is. This work focused on modelling all innovations in the networks; however, the evidence suggests that the time granularity of the innovation usage highly influences the definition of language change. Thus, it might have been better to distinguish between the long-term diffusion process and short-term innovation adoptions, then model these individually. This is not to say that what was captured in this work is not language change, but rather that what may have been captured is a combination of both language change and users reacting to events in the short term. Additionally, the results identified that, across this work, there were limitations in the methods and data that were used, the main limitation being the data sparsity within the Twitter mention network and the inability to log what users have seen when browsing the networks. Finally, we have shown through the implementation of the methods the relative ease with which we can utilise big data technologies within the context of academic research.

Chapter 10

Conclusion

This thesis has looked at language change and evolution through the medium of OSN. The result show how we can first detect growth, predict adoption and quantify diffusions of language innovations across OSNs through the application of grounded models (from linguistics) and machine learning. However, this work is not without its limitations, from the skewed sample of language from the given data sources, to the relatively simplistic models that do not quantify outside influences. Although the data was from biased sources, the results highlight the growing dominance of these OSN as playing an important role in culture, both online and offline as the language identified in chapter 6 includes terms mirrored offline in popular terminology and events of the time. Additionally, the diffusions of language (Chapters 7 and 8) could be modelled across both inter-user and inter-community iterations (Twitter geographical and mention networks), identifying that the language used in both abstractions is similar and interconnected. This shows that online sources are as important for studies of language as offline sources, and that, as online media becomes more dominant in daily life, the language used across them will mirror language used offline to a greater extent as the two converge into one form.

This conclusion of multiple influences in language formation can be seen in the results from chapter 7, which indicate that, firstly, the influence over a user's language does not come from one source but from all interactions in their life. However, as community structure has represented little influence, this gives rise to the idea that the language individuals are using is harmonising as users' communication with distant communities increases; thus, communities look for the same common language, which can also be seen in chapter 8, as larger diffusions spread in a more predictable manner.

This work has implications for the future direction of research into language change and evolution, in that it has shown how we can use simplistic and grounded models on large scales to prove or disprove known classical theories. In addition, the thesis shows the power that big data brings to the fields of linguistics and IS in that it allows us to sample the whole of a population rather than a sample, though this change in paradigm means there may need to be a modification in the epistemological stance in the

discussion of the results as they may not be representative of a 'general' population. However, it also highlights the complexity (in developing the code) needed to learn models such as general threshold and circumventing scaling issues due to the JVM.

10.1 Future Direction

Taking this work forward, a number of research directions can be considered in both attempting to correct for some of the limitations in the data and representation but also applying more advanced and involved methods. As highlighted throughout, as we are only sampling language from online sources, thus we can only make the statement that we have been able to model language innovation diffusion as seen in the online world. In addition, we have limited knowledge of the authors of the texts taken from online sources, which means that it is challenging to comment in depth on the homophilc nature of user associations.

In order to achieve a more representative sample of language, we could use the methods developed in this thesis but apply them to data collected from offline sources. This could take the form of modelling the change of language and the diffusion of innovations across a corpus of academic writing, such as arxiv[1] or SSRN[2]. Even though the language within these datasets would be more formalised, it would be more representative of the diversity of topics and colloquial language within a population. Additionally, as there are in excess of 100 years' worth of published content, we can model language over not only small variations but major shifts across the century. One noticeable benefit of using such formalised data is that, unlike the data collected for this thesis, we could use the references between articles as the explicit relationship between papers, authors and academic disciplines. Alternatively, the methods could be applied to global language networks (samples of language of each country with the relationships between countries being the migration pattern) to identify, in a quantitative rather than qualitative manner, the diffusion patterns between languages across continents to understand how language is marginalising and why certain languages may be dying out.

The second direction that this research could be taken is to extend the models and methods developed to understand the interests and topics of a user, as, throughout this thesis, we have studied the interests and topics being defined in the network structure. In understanding the user, we can better understand homophily within the network, and extend the features within the prediction of a diffusion through the explicit cohesion of topics within the diffusion path. This could be achieved by applying LDA [143] to the text a user generates as this would extract the topics a user is writing about; thus, we would be able to quantify the user's topic cohesion with their ego network and the terms that they adopt.

[1] https://arxiv.org/
[2] https://www.ssrn.com/

10.2 Concluding Statement

In conclusion, as highlighted throughout this work, language is in constant flux, whether from the creation of new terms (innovations) to the subsequent adoption of new terms over time. We have shown that we can develop new scalable computation methods that can not only detect the growth and death of innovations (at both global and local levels) but also predict the innovation adoption and diffusion for both communities and individuals. This thesis has ultimately shown how, by utilising OSN and advances in big data and machine learning, we can measure and predict language change with relative ease, where before it was time-consuming and costly.

Glossary

amelioration the process of a word becoming positive over time.. 37

complex contagion within a social network multipul exposures are required for the node to adopt the innovation/phenomenon. 46

ego network A network consisting of one focal node (the "ego"), and the nodes with which it is directly connected (the "alters"). It additionally contains the connection between alters if they exist. Each alter will have their own ego network in addition. The sum of ego networks in a larger system represents the entire network.. 32, 39, 45, 188, 189, 197, 200, 204

emoji smiles and pictures used in electric messaging encoded within the ASICC characters set. 20, 194

emoticon a representation of facial expressions in text through the user of punctuation. . 20, 37, 194

hashtag A word or phrase preceded by a hash sign (#), used on social media websites and applications, especially Twitter, to identify messages on a specific topic. [178]. 45, 46, 194

homophily the tendency of induviduals assosiating themselves with people like themselves.. 44, 45, 51

meme An element of a culture or system of behaviour passed from one individual to another by imitation or other non-genetic means. [178]. 4, 46, 47, 49, 50, 52, 162, 189, 190

morpho-phonemic Of or relating to the written form of words. morphographemic rule noun (in transformational grammar) a rule for transforming phonetic units into written representations of their sounds; a rule for writing or spelling. [184]. 37

open class nouns, verbs and adjectives. 189

power law A relationship between two quantities such that one is proportional to a fixed power of the other. [178]. 26, 49–52, 190, 196

projection the process of a word becoming more negative over time.. 37

simple contagion within a social network a single exposure is required for the node to adopt the innovation/phenomenon. 46, 50, 52

social influence the process of a user changing their behaviour based on the users within their ego network.. 44, 45

spatio-temporal Belonging to both space and time or to space–time. [178]. 64

WordNet lexical database for the English language.. 115

Acronyms

AMA Ask me anything. 66

AMQP Advanced Message Queuing Protocol. 69

API Application Programming Interface. 62, 65–70, 77, 81, 110, 117, 118, 135

AUC Area under curve. 73, 143, 151, 154, 156, 158, 188, 189

BNC British National Corpus. 98, 101, 112, 137, 166, 194

CDF Cumulative distribution function. 31

CMC Computer Mediated Communication. 20, 60

CSV Comma Separated Values. 91, 98, 148

CV Cross Validation. 74

EPSRC Engineering and Physical Sciences Research Council. iii

FPR False Positive Rate. 142, 151

GPS Global Positioning System. 40, 64, 70, 87, 89–91, 110, 135, 158, 168, 195

HDFS Hadoop Distributed File System. 68, 76, 118, 175

HIV Human Immunodeficiency Virus. 27

HQL Hive Query Language. 84, 85, 90, 92, 94, 95, 97, 100

ICT Information Communication Technology. 35

IM Instant Messaging. 20

IO Principal Component Analysis. 118

IS Information Systems. 199

JS Jaccard similarity. 45

JSD Jensen–Shannon divergence. 38

JSON JavaScript Object Notation. 84–86, 93

JVM Java Virtual Machine. 91, 100

LGBT Lesbian Gay Bi-sexual and Trans(gender/sexual). 38

ML Machine Learning. 37

MSE Mean Square Error. 50

NER Names Entity Recognition. 37, 133

NLP Natural Language Processing. 6, 20, 37, 40, 99, 106, 110, 116, 133, 186

NoSQL Hadoop Distributed File System. 144

OED Oxford English Dictionary. 98, 114, 117

OOV Out of Vocabulary. 36–38, 43, 53, 129, 137

OSN Online Social Network. vi, vii, 1–5, 13, 20, 26, 34, 35, 39, 41–43, 45, 46, 51, 53, 57, 59, 60, 62, 66, 68, 73, 75, 79, 99, 105–107, 110, 112, 117, 126, 128, 131, 133, 135, 136, 162, 163, 166, 167, 185–193, 199, 201

PCA Principal Component Analysis. 40

PhD Doctor of Philosophy. 7, 8

PMI Point-wise mutual information. 37

POS Part of Speech. 6, 37, 38, 59, 98, 99, 101, 116, 132, 133, 137, 154, 169, 172, 181, 183, 189

REST Representational State Transfer. 67, 68

ROC Receiver operating characteristic. 44, 73, 132, 142, 148, 150, 151, 157, 158, 188

RSS Principal Component Analysis. 64

SQL Structured Query Language. 77, 84

SVM Support Vector Machines. 36, 37, 186

TPR True Positive Rate. 142, 151

TTA Time to Addoption. 29–33

U.K. United Kingdom. 70

URL Uniform Resource Locator. 37, 43, 101, 162, 194

USA United States of America. 70

WI Wiener Index. 48, 49

XML eXtensible Markup Language. 84

Bibliography

[1] X. W. A. C.-E. A McCallum, "Topic and role discovery in social networks with experiments on enron and academic email", pp. 1–24, Oct. 2007. [Online]. Available: https://www.aaai.org/Papers/JAIR/Vol30/JAIR-3007.pdf.

[2] L. A. Adamic, T. M. Lento, E. Adar, and P. C. Ng, "Information evolution in social networks", in *WSDM '16: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, Genome Institute of Singapore, New York, New York, USA: ACM, Feb. 2016, pp. 473–482, ISBN: 978-1-4503-3716-8. DOI: 10.1145/2835776.2835827. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2835776.2835827.

[3] B. Ahmed, "Lexical normalisation of twitter data", Sep. 2014. arXiv: 1409.4614 [cs.CL]. [Online]. Available: http://arxiv.org/abs/1409.4614v2.

[4] L. M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, and F. Menczer, "Friendship prediction and homophily in social media", English, *Acm Transactions on the Web*, vol. 6, no. 2, pp. –33, May 2012. DOI: 10.1145/2180861.2180866. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2180861.2180866.

[5] F. Al Zamal, W. Liu, and D. Ruths, "Homophily and latent attribute inference: inferring latent attributes of twitter users from neighbors.", *ICWSM*, 2012. [Online]. Available: http://www.networkdynamics.org/static/publication_files/ZamalLiuRuths_ICWSM2012.pdf.

[6] J. Algeo, "Where do all the new words come from?", *American Speech*, vol. 55, no. 4, pp. 264–277, 1980, ISSN: 00031283, 15272133. [Online]. Available: http://www.jstor.org/stable/454567.

[7] X. Amatriain, A. Jaimes*, N. Oliver, and J. M. Pujol, "Data mining methods for recommender systems", in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Boston, MA: Springer US, 2011, pp. 39–71, ISBN: 978-0-387-85820-3. DOI: 10.1007/978-0-387-85820-3_2. [Online]. Available: https://doi.org/10.1007/978-0-387-85820-3_2.

[8] C. Anderson, *The end of theory: The data deluge makes the scientific method obsolete*, Jun. 2008. [Online]. Available: https://www.wired.com/2008/06/pb-theory/.

[9] S. Aral, L. Muchnik, and A. Sundararajan, "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks", English, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 51, pp. 21 544–21 549, 2009. DOI: 10.1073/pnas.0908800106. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.0908800106.

[10] E. Aramaki, S. Maskawa, and M. Morita, "Twitter catches the flu: Detecting influenza epidemics using twitter", in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '11, Edinburgh, United Kingdom: Association for Computational Linguistics, 2011, pp. 1568–1576, ISBN: 978-1-937284-11-4. [Online]. Available: http://dl.acm.org/citation.cfm?id=2145432.2145600.

[11] G. Aston and L. Burnard, *The BNC Handbook: Exploring the British National Corpus with SARA*, ser. Edinburgh University Press Series. Edinburgh University Press, 1998, ISBN: 9780748610556. [Online]. Available: https://books.google.co.uk/books?id=s3iaoF1m8tAC.

[12] E. Bakshy, W. A. Mason, J. M. Hofman, and D. J. Watts, "Everyone's an influencer: quantifying influence on twitter", in *Proceedings of the 4th ACM International Conference on Web Search and Data Mining, WSDM 2011*, University of Michigan, Ann Arbor, United States, New York, New York, USA: ACM Press, Mar. 2011, pp. 65–74, ISBN: 9781450304931. DOI: 10.1145/1935826.1935845. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1935826.1935845.

[13] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The role of social networks in information diffusion", *ArXiv.org*, Jan. 2012. arXiv: 1201.4145v2 [cs.SI]. [Online]. Available: http://arxiv.org/abs/1201.4145v2.

[14] T. Baldwin, P. Cook, M. Lui, and A. MacKinlay, "How noisy social media text, how diffrnt social media sources?", in *Proceedings of the 6th ...*, 2013. [Online]. Available: http://www.aclweb.org/anthology/I/I13/I13-1041.pdf.

[15] D. Bamman, C. Dyer, and N. A. Smith, "Distributed representations of geographically situated language", 2014. [Online]. Available: http://repository.cmu.edu/lti/154/.

[16] A. L. Barabasi and R. Albert, "Emergence of scaling in random networks", vol. 286, no. 5439, pp. 509–512, 1999. DOI: 10.1126/science.286.5439.509. [Online]. Available: http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1999Sci...286..509B&link_type=EJOURNAL.

[17] A.-L. Barabási and E. Bonabeau, "Scale-free networks", *Scientific American*, vol. 288, no. 5, pp. 60–69, May 2003. DOI: 10.1038/scientificamerican0503-60. [Online]. Available: http://www.nature.com/doifinder/10.1038/scientificamerican0503-60.

[18] D. K. Barnhart, "A calculus for new words", English, vol. 28, no. 1, pp. 132–138, 2007. DOI: 10.1353/dic.2007.0009. [Online]. Available: http://muse.jhu.edu/content/crossref/journals/dictionaries/v028/28.barnhart.html.

[19] M. G. Beiró, A. Panisson, M. Tizzoni, and C. Cattuto, "Predicting human mobility through the assimilation of social media traces into mobility models", Jan. 2016. [Online]. Available: http://arxiv.org/abs/1601.04560v1.

[20] M. S. Bernstein, A. Monroy-Hernández, D. Harry, and P. André, "4chan and/b: an analysis of anonymity and ephemerality in a large online community.", *ICWSM*, 2011. [Online]. Available: http://www.aaai.org/ocs/index/ICWSM/ICWSM11/paper/viewFile/2873/4398.

[21] J. T. Bertrand, "Diffusion of innovations and hiv/aids", *Journal of health communication*, 2004. DOI: 10.1214/10-aoas368suppb. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/10810730490271575.

[22] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks", English, *Journal of Statistical Mechanics: Theory and Experiment*, vol. physics.soc-ph, no. 10, Oct. 2008. DOI: 10.1088/1742-5468/2008/10/P10008. [Online]. Available: http://stacks.iop.org/1742-5468/2008/i=10/a=P10008?key=crossref.46968f6ec61eb8f907a760be1c5ace52.

[23] J. Borge-Holthoefer, R. A. Banos, S. González-Bailón, and Y. Moreno, "Cascading behaviour in complex socio-technical networks", English, *Journal of Complex Networks*, vol. 1, no. 1, pp. 3–24, May 2013. DOI: 10.1093/comnet/cnt006. [Online]. Available: http://comnet.oxfordjournals.org/cgi/doi/10.1093/comnet/cnt006.

[24] R. Brown and A. Gilman, "The pronouns of power and solidarity", in *Style in Language*, T. A. Sebeok, Ed., Cambridge, Mass: MIT Press, 1960, pp. 253–276.

[25] A. Bruns and J. E. Burgess, "The use of twitter hashtags in the formation of ad hoc publics", *ARC Centre of Excellence for Creative Industries and Innovation; Creative Industries Faculty; Institute for Creative Industries and Innovation*, Aug. 2011. [Online]. Available: http://eprints.qut.edu.au/46515.

[26] C. Buntain and J. Golbeck, "Identifying social roles in reddit using network structure", in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW '14 Companion, Seoul, Korea: ACM, 2014, pp. 615–620, ISBN: 978-1-4503-2745-9. DOI: 10.1145/2567948.2579231. [Online]. Available: http://doi.acm.org/10.1145/2567948.2579231.

[27] R. S. Burt, "The social capital of structural holes", in *The New Economic Sociology*, Russell Sage Foundation, May 2005, pp. 201–247, ISBN: 9780871543653.

[28] R. S. Burt, "The network structure of social capital", *Research in Organizational Behavior*, vol. 22, no. Supplement C, pp. 345–423, 2000, ISSN: 0191-3085. DOI: https://doi.org/10.1016/S0191-3085(00)22009-1. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0191308500220091.

[29] R. S. Burt, "Structural holes versus network closure as social capital", in *Social Capital: Theory and Research*, Aldine de Gruyter, 2001, pp. 1–30, ISBN: 9780202368948.

[30] C. Butterworth, D. Kershaw, J. Devine, J. Gallagher, and J. Croft, "Presenting the past: a case study of innovation opportunities for knowledge dissemination to the general public through pervasive technology", in *21st Annual Meeting of the European Association of Archaeologists*, 2015, pp. 391–392.

[31] E. Carmel and R. Agarwal, "Tactical approaches for alleviating distance in global software development", *IEEE Software*, vol. 18, no. 2, Mar. 2001. [Online]. Available: http://dl.acm.org/citation.cfm?id=626245.

[32] CBR. (Apr. 2013). Uk leads the world for twitter users - computer business review, [Online]. Available: http://www.cbronline.com/news/social/twitter-users-are-migrating-from-pcs-to-tablets-and-phones-report-120413.

[33] J. K. Chambers and N. Schilling-Estes, *The Handbook of Language Variation and Change*, English. John Wiley & Sons, Jun. 2013, ISBN: 1118335570. [Online]. Available: http://books.google.co.uk/books?id=UnXlipEpMAMC&printsec=frontcover&dq=intitle:The+handbook+of+language+variation+and+change&hl=&cd=1&source=gbs_api.

[34] P. Chang and K. Ahrens, "Towards a model for the prediction of chinese novel verbs.", in *PACLIC*, 2008, pp. 131–140. [Online]. Available: http://www.aclweb.org/anthology/Y08-1012.

[35] L. Chen, K. S. M. T. Hossain, P. Butler, N. Ramakrishnan, and B. A. Prakash, "Flu gone viral: Syndromic surveillance of flu on twitter using temporal topic models", in *2014 IEEE International Conference on Data Mining*, Dec. 2014, pp. 755–760. DOI: 10.1109/ICDM.2014.137. [Online]. Available: http://people.cs.vt.edu/~badityap/papers/fluviral-icdm14.pdf.

[36] J. Cheng, L. Adamic, L. A. Adamic, P. A. Dow, J. M. Kleinberg, J. Kleinberg, and J. Leskovec, "Can cascades be predicted?", *ArXiv.org*, pp. 925–936, Mar. 2014. DOI: 10.1145/2566486.2567997. arXiv: 1403.4608v1 [cs.SI]. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2566486.2567997.

[37] D. Choi, J. Han, T. Chung, Y.-Y. Ahn, B.-G. Chun, and T. T. Kwon, "Characterizing conversation patterns in reddit", in *The 2015 ACM*, New York, New York, USA: ACM Press, 2015, pp. 233–243, ISBN: 9781450339513. DOI: 10.1145/2817946.2817959. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2817946.2817959.

[38] N. Chomsky, "On certain formal properties of grammars", *Information and Control*, vol. 2, no. 2, pp. 137–167, 1959, ISSN: 0019-9958. DOI: https://doi.org/10.1016/S0019-9958(59)90362-6. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0019995859903626.

[39] C.-T. Chu, S. K. Kim, Y.-A. Lin, Y. Yu, G. Bradski, A. Y. Ng, and K. Olukotun, "Map-reduce for machine learning on multicore", in *Proceedings of the 19th International Conference on Neural Information Processing Systems*, ser. NIPS'06, Canada: MIT Press, 2006, pp. 281–288. [Online]. Available: http://dl.acm.org/citation.cfm?id=2976456.2976492.

[40] P. Cook, "Using social media to find english lexical blends", *Euralex.org*, 2012. [Online]. Available: http://www.euralex.org/elx_proceedings/Euralex2012/pp846-854%5C%20Cook.pdf.

[41] P. Cook, B. Han, and T. Baldwin, "Statistical methods for identifying local dialectal terms from gps-tagged documents", *... : Journal of the Dictionary Society of North ...*, 2014. [Online]. Available: http://muse.jhu.edu/journals/dictionaries/v035/35.cook.html.

[42] P. Cook and S. Stevenson, "Automagically inferring the source words of lexical blends", in *Proceedings of the Tenth Conference of the ...*, 2007. [Online]. Available: http://www.cs.utoronto.ca/~pcook/CookStevenson2007.pdf.

[43] ——, "Automatically identifying changes in the semantic orientation of words.", *LREC*, 2010. [Online]. Available: ftp://learning.cs.toronto.edu/public_html/public_html/cs/ftp/pub/gh/Cook+Stevenson-2010-LREC.pdf.

[44] P. Cook and S. Stevenson, "Automatically identifying the source words of lexical blends in english", English, *Computational Linguistics*, vol. 36, no. 1, pp. 129–149, Mar. 2010. DOI: 10.1162/coli.2010.36.1.36104. [Online]. Available: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2010.36.1.36104.

[45] W. Croft, *Explaining Language Change: An Evolutionary Approach*, ser. Explaining Language Change : an Evolutionary Approach. Longman, 2000, ISBN: 9780582356771. [Online]. Available: https://books.google.co.uk/books?id=5%5C_Ka7zLl9HQC.

[46] W. Croft, "Mixed languages and acts of identity: an evolutionary approach", in *The Mixed Language Debate*, Y. Matras and P. Bakker, Eds., Berlin, New York: Mouton de Gruyter, 2003, pp. 41–73, ISBN: 9783110197242. [Online]. Available: http://www.worldcat.org/title/mixed-language-debate-theoretical-and-empirical-advances/oclc/965696847.

[47] ——, "Evolution: language use and the evolution of languages", English, *The Language Phenomenon*, no. Chapter 5, pp. 93–120, Jan. 2013. DOI: 10.1007/978-3-642-36086-2_5. [Online]. Available: http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2013lphc.book...93C&link_type=EJOURNAL.

[48] D. Crystal, *Txtng: The Gr8 Db8*. OUP Oxford, 2009, ISBN: 9780191623400. [Online]. Available: https://books.google.co.uk/books?id=pKBnEGwmtZoC.

[49] ——, *Internet Linguistics: A Student Guide*. Routledge, 2011, ISBN: 9780415602716. [Online]. Available: https://books.google.co.uk/books?id=DMShSgAACAAJ.

[50] A. Culotta, "Detecting influenza outbreaks by analyzing twitter messages", *ArXiv.org*, Jul. 2010. arXiv: 1007.4748v1 [cs.IR]. [Online]. Available: http://arxiv.org/abs/1007.4748.

[51] C. Danescu-Niculescu-Mizil, M. Gamon, and S. Dumais, "Mark my words! linguistic style accommodation in social media", *ArXiv.org*, May 2011. DOI: 10.1145/1963405.1963509. arXiv: 1105.0673v1 [cs.CL]. [Online]. Available: http://arxiv.org/abs/1105.0673v1.

[52] C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts, "A computational approach to politeness with application to social factors", *ArXiv.org*, p. 6078, Jun. 2013. arXiv: 1306.6078 [cs.CL]. [Online]. Available: http://arxiv.org/abs/1306.6078v1.

[53] C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, and C. Potts, *No country for old members: user lifecycle and linguistic change in online communities*. International World Wide Web Conferences Steering Committee, May 2013, ISBN: 978-1-4503-2035-1. [Online]. Available: http://dl.acm.org/citation.cfm?id=2488388.2488416.

[54] Y. Darmaillac and S. Loustau, "Mcmc louvain for online community detection", *ArXiv.org*, Dec. 2016. arXiv: 1612.01489v1 [cs.SI]. [Online]. Available: http://arxiv.org/abs/1612.01489v1.

[55] J. A. Davis, "Clustering and hierarchy in interpersonal relations: testing two graph theoretical models on 742 sociomatrices", *American Sociological Review*, vol. 35, no. 5, p. 843, Oct. 1970. [Online]. Available: http://www.jstor.org/stable/2093295?origin=crossref.

[56] M. De Choudhury, *Tie Formation on Twitter: Homophily and Structure of Egocentric Networks*, English. IEEE, 2011, ISBN: 978-1-4577-1931-8. DOI: 10.1109/PASSAT/SocialCom.2011.177. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6113149.

[57] P. De Meo, E. Ferrara, G. FIUMARA, and A. PROVETTI, "On facebook, most ties are weak", *ArXiv.org*, Mar. 2012. DOI: 10.1145/2629438. arXiv: 1203.0535v2 [cs.SI]. [Online]. Available: http://arxiv.org/abs/1203.0535v2.

[58] D. Dediu, M. Cysouw, S. C. Levinson, A. Baronchelli, M. H. Christiansen, W. Croft, N. Evans, S. Garrod, R. D. Gray, A. Kandler, and E. Lieven, "Cultural evolution of language", English, in, 2013, ISBN: 978-0-262-01975-0. [Online]. Available: http://books.google.com/books?hl=en&lr=&id=xdqbAQAAQBAJ&oi=fnd&pg=PR5&dq=Cultural+Evolution+Society+Technology+Language+and+Religion&ots=Jo76j5mMSF&sig=Wmi7tWnJIcYFiVd1tlJ6ZgGj1ok.

[59] M. Duggan and A. Smith, "6% of online adults are reddit users", *Pew Internet & American Life Project*, 2013. [Online]. Available: http://www.pewinternet.org/files/old-media/Files/Reports/2013/PIP_reddit_usage_2013.pdf.

[60] M. Duggan and A. Smith, "Social media update 2013", *Pewinternet.org*, Dec. 2013. [Online]. Available: http://www.pewinternet.org/2013/12/30/social-media-update-2013/.

[61] J. Eisenstein, "Identifying regional dialects in online social media", 2014. DOI: 10.2307/50567?ref=no-x-route:999e2e100a68b5431a4dc6b948260b8b. [Online]. Available: https://smartech.gatech.edu/handle/1853/52405.

[62] J. Eisenstein, "What to do about bad language on the internet", in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia: Association for Computational Linguistics, Jun. 2013, pp. 359–369. [Online]. Available: http://www.aclweb.org/anthology/N13-1037.

[63] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, "A latent variable model for geographic lexical variation", in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, MA: Association for Computational Linguistics, 2010, pp. 1277–1287. [Online]. Available: http://www.aclweb.org/anthology/D10-1124.

[64] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, "Mapping the geographical diffusion of new words", *ArXiv.org*, p. 5268, Oct. 2012. arXiv: 1210.5268v3 [1210]. [Online]. Available: http://arxiv.org/abs/1210.5268v3.

[65] E. Ferrara, O. Varol, F. Menczer, and A. Flammini, "Traveling trends: social butterflies or frequent fliers?", *ArXiv.org*, Oct. 2013. DOI: 10.1145/2512938.2512956. arXiv: 1310.2671v1 [cs.SI]. [Online]. Available: http://arxiv.org/abs/1310.2671v1.

[66] M. Fischer, ""birds of a feather flock together" reloaded: Homophily in the context of web 2.0 in online social networking sites such as facebook", PhD thesis, College of Charleston, 2010.

[67] R. Fischer, *Lexical Change in Present-day English: A Corpus-based Study of the Motivation, Institutionalization, and Productivity of Creative Neologisms*, ser. Language in performance. G. Narr, 1998, ISBN: 9783823349402. [Online]. Available: https://books.google.co.uk/books?id=H93nAVbwZwwC.

[68] J. Frenzen and K. Nakamoto, "Structure, cooperation, and the flow of market information", *Journal of Consumer Research*, 1993. DOI: 10.2307/2489353. [Online]. Available: http://www.jstor.org/stable/2489353.

[69] S. Gao, J. Ma, and Z. Chen, "Modeling and predicting retweeting dynamics on microblogging platforms", in *The Eighth ACM International Conference*, New York, New York, USA: ACM Press, 2015, pp. 107–116, ISBN: 9781450333177. DOI: 10.1145/2684822.2685303. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2684822.2685303.

[70] A. Gazzard, M. Lochrie, A. Gradinar, P. Coulton, D. Burnett, and D. Kershaw, "From the board to the streets: a case study of local property trader", *Transactions of the Digital Games Research Association (ToDIGRA)*, vol. 1, no. 3, 2014.

[71] A. Ghaziani and M. J. Ventresca, "Keywords and cultural change: frame analysis of business model public talk, 1975-2000", *Sociological Forum*, vol. 20, no. 4, pp. 523–559, Dec. 2005. DOI: 10.1007/s11206-005-9057-0. [Online]. Available: http://www.scopus.com/inward/record.url?partnerID=HzOxMe3b&scp=29344451239&origin=inward.

[72] S. Ghemawat, H. Gobioff, and S.-T. Leung, "The google file system", *SIGOPS Oper. Syst. Rev.*, vol. 37, no. 5, pp. 29–43, Oct. 2003, ISSN: 0163-5980. DOI: 10.1145/1165389.945450. [Online]. Available: http://doi.acm.org/10.1145/1165389.945450.

[73] A. Giddens and C. Pierson, *Conversations with Anthony Giddens: Making Sense of Modernity*. Stanford University Press, 1998, ISBN: 9780804735681. [Online]. Available: https://books.google.co.uk/books?id=YXDsvo90iZsC.

[74] A. Giddens, *The Giddens Reader*, English. Stanford University Press, Jan. 1993, ISBN: 9780804722049. [Online]. Available: http://books.google.co.uk/books?id=hjGsAAAAIAAJ&printsec=frontcover&dq=intitle:The+Giddens+Reader&hl=&cd=1&source=gbs_api.

[75] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, "Part-of-speech tagging for twitter: Annotation, features, and experiments", in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, ser. HLT '11, Portland, Oregon: Association for Computational Linguistics, 2011, pp. 42–47, ISBN: 978-1-932432-88-6. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002736.2002747.

[76] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, "Part-of-speech tagging for twitter: annotation, features, and experiments", in *HLT '11: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short papers*, Association for Computational Linguistics, Jun. 2011. [Online]. Available: http://portal.acm.org/citation.cfm?id=2002736.2002747&coll=DL&dl=GUIDE&CFID=581530381&CFTOKEN=26152960.

[77] B. Gliwa, P. Bródka, and A. Zygmunt, "Predicting community evolution in social networks", English, *ArXiv.org*, no. 5, pp. 3053–3096, May 2015. DOI: 10.3390/e17053053. arXiv: 1505.01709v1 [cs.SI]. [Online]. Available: http://www.mdpi.com/1099-4300/17/5/3053/.

[78] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence", *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 5, no. 4, Feb. 2012. DOI: 10.1145/2086737.2086741. [Online]. Available: http://dl.acm.org/citation.cfm?id=2086741.

[79] B. Gonçalves and D. Sánchez, "Crowdsourcing dialect characterization through twitter", *ArXiv.org*, p. 7094, Jul. 2014. arXiv: 1407.7094 [physics.soc-ph]. [Online]. Available: http://arxiv.org/abs/1407.7094v1.

[80] ——, "Learning spanish dialects through twitter", *ArXiv.org*, p. 4970, Nov. 2015. arXiv: 1511.04970v1 [1511]. [Online]. Available: http://arxiv.org/abs/1511.04970v1.

[81] S. González-Bailón, N. Wang, and J. Borge-Holthoefer, "The emergence of roles in large-scale networks of communication", English, *EPJ Data Sci*, vol. 3, no. 1, p. 32, 2014. DOI: 10.1140/epjds/s13688-014-0032-y. [Online]. Available: http://www.epjdatascience.com/content/3/1/32.

[82] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, *Discovering leaders from community actions*. New York, New York, USA: ACM, Oct. 2008, ISBN: 978-1-59593-991-3. DOI: 10.1145/1458082.1458149. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1458082.1458149.

[83] ——, "Learning influence probabilities in social networks", in *WSDM '10: Proceedings of the third ACM international conference on Web search and data mining*, ACM Request Permissions, Feb. 2010. DOI: 10.1145/1718487.1718518. [Online]. Available: http://dl.acm.org/citation.cfm?id=1718518.

[84] A. Goyal, B.-W. On, F. Bonchi, and L. V. S. Lakshmanan, "Gurumine: a pattern mining system for discovering leaders and tribes", in *2009 IEEE 25th International Conference on Data Engineering (ICDE)*, IEEE, Mar. 2009, pp. 1471–1474, ISBN: 978-1-4244-3422-0. DOI: 10.1109/ICDE.2009.59. [Online]. Available: http://ieeexplore.ieee.org/document/4812550/.

[85] P. A. Grabowicz, J. J. Ramasco, E. Moro, J. Pujol, and V. M. Eguiluz, "Social features of online networks: the strength of intermediary ties in online social media", English, *ArXiv.org*, no. 1, e29358, Jul. 2011. DOI: 10.1371/journal.pone.0029358. arXiv: 1107.4009v2 [physics.soc-ph]. [Online]. Available: http://dx.plos.org/10.1371/journal.pone.0029358.

[86] M. Granovetter, "Threshold models of collective behavior", *American Journal of Sociology*, vol. 83, no. 6, pp. 1420–1443, 1978. DOI: 10.1086/226707. eprint: https://doi.org/10.1086/226707. [Online]. Available: https://doi.org/10.1086/226707.

[87] M. S. Granovetter, "The strength of weak ties", *American journal of sociology*, pp. 1360–1380, 1973. [Online]. Available: http://www.jstor.org/stable/10.2307/2776392.

[88] I. Gregory, P. Atkinson, A. Hardie, A. Joulain-Jay, D. Kershaw, C. Porter, P. Rayson, and C. J. Rupp, "From digital resources to historical scholarship with the british library 19th century newspaper collection", *Zurnal Sibirskogo federal'nogo universiteta. Seria: Gumanitarnye nauki*, vol. 9, no. 4, pp. 994–1006, 2016.

[89] J. GRIEVE, A. NINI, and D. GUO, "Analyzing lexical emergence in modern american english online", *English Language and Linguistics*, vol. 21, no. 1, pp. 99–127, 2017. DOI: 10.1017/S1360674316000113.

[90] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace", in *Proceedings of the 13th International Conference on World Wide Web*, ser. WWW '04, New York, NY, USA: ACM, 2004, pp. 491–501, ISBN: 1-58113-844-X. DOI: 10.1145/988672.988739. [Online]. Available: http://doi.acm.org/10.1145/988672.988739.

[91] M. Gupte, P. Shankar, J. Li, S. Muthukrishnan, and L. Iftode, "Finding hierarchy in directed online social networks", in *Proceedings of the 20th International Conference on World Wide Web, WWW 2011*, Rutgers, New York, New York, USA: ACM Press, Dec. 2011, pp. 557–566, ISBN: 9781450306324. DOI: 10.1145/1963405.1963484. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1963405.1963484.

[92] S. A. Hale, "Global connectivity and multilinguals in the twitter network.", *CHI*, pp. 833–842, 2014. DOI: 10.1145/2556288.2557203. [Online]. Available: http://doi.acm.org/10.1145/2556288.2557203.

[93] B. Han and T. Baldwin, "Lexical normalisation of short text messages: Makn sens a #twitter", in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, ser. HLT '11, Portland, Oregon: Association for Computational Linguistics, 2011, pp. 368–378, ISBN: 978-1-932432-87-9. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002472.2002520.

[94] B. Han, P. Cook, and T. Baldwin, "Automatically constructing a normalisation dictionary for microblogs", in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, ser. EMNLP-CoNLL '12, Jeju Island, Korea: Association for Computational Linguistics, 2012, pp. 421–432. [Online]. Available: http://dl.acm.org/citation.cfm?id=2390948.2391000.

[95] ——, "Geolocation prediction in social media data by finding location indicative words", in *24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers*, University of Melbourne, Parkville, Australia, Dec. 2012, pp. 1045–1062. [Online]. Available: http://anthology.aclweb.org/C/C12/C12-1064.pdf.

[96] ——, "Lexical normalization for social media text", English, *Acm Transactions on Intelligent Systems and Technology*, vol. 4, no. 1, pp. –27, Jan. 2013. DOI: 10.1145/2414425.2414430. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2414425.2414430.

[97] M. D. Hauser, N. Chomsky, and W. T. Fitch, "The faculty of language: What is it, who has it, and how did it evolve?", *Science*, vol. 298, no. 5598, pp. 1569–1579, 2002, ISSN: 0036-8075. DOI: 10.1126/science.298.5598.1569. eprint: http://science.sciencemag.org/content/298/5598/1569.full.pdf. [Online]. Available: http://science.sciencemag.org/content/298/5598/1569.

[98] T. Hey, S. Tansley, and K. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Oct. 2009, ISBN: 978-0-9825442-0-4. [Online]. Available: https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/.

[99] C. Honey and S. C. Herring, "Beyond microblogging: conversation and collaboration via twitter", English, pp. 1–10, Jan. 2009. DOI: 10.1109/HICSS.2009.89. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4755499.

[100] Y. Huang, D. Guo, A. Kasakoff, and J. Grieve, "Understanding us regional linguistic variation with twitter data analysis", *Computers*, 2015. DOI: 10.1016/j.compenvurbsys.2015.12.003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0198971515300399.

[101] Y. Huang, D. Guo, A. Kasakoff, and J. Grieve, "Understanding u.s. regional linguistic variation with twitter data analysis", *Computers, Environment and Urban Systems*, vol. 59, pp. 244–255, 2016, ISSN: 0198-9715. DOI: https://doi.org/10.1016/j.compenvurbsys.2015.12.003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0198971515300399.

[102] B. A. Huberman, D. M. Romero, and F. Wu, "Social networks that matter twitter under the microscope", *First Monday*, vol. 14, no. 1, Jan. 2009. [Online]. Available: https://www.scopus.com/inward/record.uri?partnerID=HzOxMe3b&scp=58449115775&origin=inward.

[103] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities", pp. 56–65, 2007. [Online]. Available: http://dl.acm.org/citation.cfm?id=1348556.

[104] E. Katz and P. Lazarsfeld, *Personal Influence, the Part Played by People in the Flow of Mass Communications*, ser. A Report of the bureau of applied social research Columbia university. Free Press, 1966, ISBN: 9781412830706. [Online]. Available: https://books.google.co.uk/books?id=rElW8D0D8gYC.

[105] P. Kazienko, P. Bródka, K. Musial, and J. Gaworecki, "Multi-layered social network creation based on bibliographic data", English, *Audio, Transactions of the IRE Professional Group on*, pp. 407–412, Aug. 2010. DOI: 10.1109/SocialCom.2010.65. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5591277.

[106] T. Kenter, M. Wevers, P. Huijnen, and M. de Rijke, "Ad hoc monitoring of vocabulary shifts over time", in *The 24th ACM International*, New York, New York, USA: ACM Press, 2015, pp. 1191–1200, ISBN: 9781450337946. DOI: 10.1145/2806416.2806474. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2806416.2806474.

[107] D. Kershaw, *Towards modelling language innovation acceptance in online social networks - dataset*, http://dx.doi.org/10.17635/lancaster/researchdata/46, 2016. DOI: 10.17635/lancaster/researchdata/46.

[108] ——, *Danjamker/networkdiffusionprediction-mr 0.0.1*, May 2018. DOI: 10.5281/zenodo.1253675. [Online]. Available: https://doi.org/10.5281/zenodo.1253675.

[109] D. Kershaw, M. Rowe, A. Noulas, and P. Stacey, "Birds of a feather talk together: user influence on language adoption", in *Hawaii International Conference on System Sciences*, Hawaii International Conference on System Sciences, Jan. 2017, ISBN: 9780998133102. DOI: 10.24251/HICSS.2017.225. [Online]. Available: http://hdl.handle.net/10125/41379.

[110] ——, *Birds of a feather talk together: user influence on language adoption - data set*, http://dx.doi.org/10.17635/lancaster/researchdata/99, 2017. DOI: 10.17635/lancaster/researchdata/99.

[111] D. Kershaw, M. Rowe, and P. Stacey, "Towards tracking and analysing regional alcohol consumption patterns in the uk through the use of social media.", *WebSci*, pp. 220–228, 2014. DOI: 10.1145/2615569.2615678. [Online]. Available: http://doi.acm.org/10.1145/2615569.2615678.

[112] ——, "Language innovation and change in on-line social networks", in *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, ser. HT '15, Guzelyurt, Northern Cyprus: ACM, 2015, pp. 311–314, ISBN: 978-1-4503-3395-5. DOI: 10.1145/2700171.2804449. [Online]. Available: http://doi.acm.org/10.1145/2700171.2804449.

[113] ——, "Towards modelling language innovation acceptance in online social networks", in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, ser. WSDM '16, San Francisco, California, USA: ACM, 2016, pp. 553–562, ISBN: 978-1-4503-3716-8. DOI: 10.1145/2835776.2835784. [Online]. Available: http://doi.acm.org/10.1145/2835776.2835784.

[114] D. Kershaw, M. Rowe, P. K. Stacey, and A. Noulas, *Birds of a feather talk together: user influence on language adoption - code*. DOI: 10.5281/zenodo.1216050. [Online]. Available: https://doi.org/10.5281/zenodo.1216050.

[115] S. Kim, S. Park, S. A. Hale, S. Kim, J. Byun, and A. Oh, "Understanding editing behaviors in multilingual wikipedia", *ArXiv.org*, p. 7266, Aug. 2015. arXiv: 1508.07266 [cs.SI]. [Online]. Available: http://arxiv.org/abs/1508.07266.

[116] S. Kim, I. Weber, L. Wei, and A. Oh, *Sociolinguistic analysis of Twitter in multilingual societies*. New York, New York, USA: ACM, Sep. 2014, ISBN: 978-1-4503-2954-5. DOI: 10.1145/2631775.2631824. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2631775.2631824.

[117] R. Kitchin, "Big data and human geography opportunities, challenges and risks", English, *Dialogues in Human Geography*, vol. 3, no. 3, pp. 262–267, Nov. 2013. DOI: 10.1177/2043820613513388. [Online]. Available: http://dhg.sagepub.com/content/3/3/262.short.

[118] R. Kitchin, "Big data, new epistemologies and paradigm shifts", English, *Big Data & Society*, vol. 1, no. 1, p. 2 053 951 714 528 481, Apr. 2014. DOI: 10.1177/2053951714528481. [Online]. Available: http://bds.sagepub.com/content/1/1/2053951714528481.full.

[119] E. M. Knorr and R. T. Ng, "Algorithms for mining distance-based outliers in large datasets", in *Proceedings of the 24rd International Conference on Very Large Data Bases*, ser. VLDB '98, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, pp. 392–403, ISBN: 1-55860-566-5. [Online]. Available: http://dl.acm.org/citation.cfm?id=645924.671334.

[120] A. Kramer and C. K. Chung, "Dimensions of self-expression in facebook status updates.", *ICWSM*, 2011. [Online]. Available: http://www.geekebook.com/file/hYGUOC/dimensions-of-self-expression-in-facebook-status-updates.pdf.

[121] S. Krishnan, P. Butler, R. Tandon, J. Leskovec, and N. Ramakrishnan, "Seeing the forest for the trees", in *The 8th ACM Conference*, New York, New York, USA: ACM Press, 2016, pp. 249–258, ISBN: 9781450342087. DOI: 10.1145/2908131.2908155. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2908131.2908155.

[122] V. Kulkarni, B. Perozzi, and S. Skiena, "Freshman or fresher? quantifying the geographic variation of internet language", *ArXiv.org*, Oct. 2015. arXiv: 1510.06786v1 [cs.CL]. [Online]. Available: http://arxiv.org/abs/1510.06786v1.

[123] V. Kulkarni, R. Al-Rfou, B. Perozzi, and S. Skiena, "Statistically significant detection of linguistic change", English, in *The 24th International Conference*, New York, New York, USA: ACM Press, 2015, pp. 625–635, ISBN: 9781450334693. DOI: 10.1145/2736277.2741627. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2736277.2741627.

[124] ——, "Statistically significant detection of linguistic change", English, in *The 24th International Conference*, New York, New York, USA: ACM Press, 2015, pp. 625–635, ISBN: 9781450334693. DOI: 10.1145/2736277.2741627. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2736277.2741627.

[125] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?", pp. 591–600, 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1772751.

[126] T. La Fond and J. Neville, "Randomization tests for distinguishing social influence and homophily effects", in *Proceedings of the 19th international conference ...*, 2010. DOI: 10.1145/1772690.1772752. [Online]. Available: http://dl.acm.org/citation.cfm?id=1772752.

[127] W. Labov, *Principles of Linguistic Change, Social Factors*, ser. Principles of Linguistic Change. Wiley, 2001, ISBN: 9780631179160. [Online]. Available: https://books.google.co.uk/books?id=LS%5C_Ux3CEI5QC.

[128] ——, *The Social Stratification of English in New York City*. Cambridge University Press, 2006, ISBN: 9780521821223. [Online]. Available: https://books.google.es/books?id=bJdKY0mZWzwC.

[129] ——, *Principles of Linguistic Change, Volume 3: Cognitive and Cultural Factors*, ser. Language in Society. Wiley, 2011, ISBN: 9781444351460. [Online]. Available: https://books.google.co.uk/books?id=uwMTUk4g2IoC.

[130] S. L. Lai and V. T. Ng, "Collaborative discovery of chinese neologisms in social media", in *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*, IEEE, 2014, pp. 4107–4112, ISBN: 978-1-4799-3840-7. DOI: 10.1109/SMC.2014.6974578. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6974578.

[131] A. Lamb, M. J. Paul, and M. Dredze, "Separating fact from fear: tracking flu infections on twitter", in *Proceedings of NAACL-HLT*, 2013. [Online]. Available: http://www.aclweb.org/anthology/N/N13/N13-1097.pdf.

[132] R. Lambiotte, "Rich gets simpler", English, *Proceedings of the National Academy of Sciences*, pp. 201 612 364–2, Aug. 2016. DOI: 10.1073/pnas.1612364113. [Online]. Available: http://www.pnas.org/lookup/doi/10.1073/pnas.1612364113.

[133] V. Lampos, T. De Bie, and N. Cristianini, "Flu detector-tracking epidemics on twitter", pp. 599–602, 2010. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-15939-8_42.

[134] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle", in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '09, Paris, France: ACM, 2009, pp. 497–506, ISBN: 978-1-60558-495-9. DOI: 10.1145/1557019.1557077. [Online]. Available: http://doi.acm.org/10.1145/1557019.1557077.

[135] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst, "Cascading behavior in large blog graphs", *ArXiv.org*, Apr. 2007. arXiv: 0704.2803v1 [physics.soc-ph]. [Online]. Available: http://arxiv.org/abs/0704.2803v1.

[136] F. Liu, F. Weng, and X. Jiang, "A broad-coverage normalization system for social media language", in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ser. ACL '12, Jeju Island, Korea: Association for Computational Linguistics, 2012, pp. 1035–1044. [Online]. Available: http://dl.acm.org/citation.cfm?id=2390524.2390662.

[137] I. Lunden. (Jul. 2012). Analyst: twitter passed 500m users in june 2012, 140m of them in us; jakarta 'biggest tweeting' city — techcrunch, [Online]. Available: http://techcrunch.com/2012/07/30/analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/.

[138] S. Maity, A. Chaudhary, S. Kumar, A. Mukherjee, C. Sarda, A. Patil, and A. Mondal, "Wassup? lol : characterizing out-of-vocabulary words in twitter", in *CSCW '16 Companion: Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, Indian Institute of Technology, Kharagpur, ACM, Feb. 2016. DOI: 10.1145/2818052.2869110. [Online]. Available: http://dl.acm.org/citation.cfm?id=2869110.

[139] A. Markham, E. Buchanan, and t. A. E. Committee, "Ethical decision making in internet research: version 2.0 ", Tech. Rep., Mar. 2013. [Online]. Available: http://aoir.org/reports/ethics2.pdf.

[140] I. Mavridis and H. Karatza, "Performance evaluation of cloud-based log file analysis with apache hadoop and apache spark", *Journal of Systems and Software*, vol. 125, pp. 133–151, 2017, ISSN: 0164-1212. DOI: https://doi.org/10.1016/j.jss.2016.11.037. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0164121216302370.

[141] R. Maybaum, "Innovation diffusion and language change:a review and some new directions", Tech. Rep., Mar. 2014.

[142] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks", *Annual Review of Sociology*, vol. 27, no. 1, pp. 415–444, 2001. DOI: 10.1146/annurev.soc.27.1.415. eprint: https://doi.org/10.1146/annurev.soc.27.1.415. [Online]. Available: https://doi.org/10.1146/annurev.soc.27.1.415.

[143] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "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*, ser. SIGIR '13, Dublin, Ireland: ACM, 2013, pp. 889–892, ISBN: 978-1-4503-2034-4. DOI: 10.1145/2484028.2484166. [Online]. Available: http://doi.acm.org/10.1145/2484028.2484166.

[144] A. A. Metcalf, *Predicting New Words: The Secrets of Their Success*. Houghton Mifflin, 2004, ISBN: 9780618130085. [Online]. Available: https://books.google.co.uk/books?id=ACsetPyuv8YC.

[145] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space", *ArXiv.org*, p. 3781, Jan. 2013. arXiv: 1301.3781 [cs.CL]. [Online]. Available: http://arxiv.org/abs/1301.3781v3.

[146] S. Milgram, "The small world problem", 1967. [Online]. Available: http://measure.igpp.ucla.edu/GK12-SEE-LA/Lesson_Files_09/Tina_Wey/TW_social_networks_Milgram_1967_small_world_problem.pdf.

[147] G. A. Miller, "Wordnet: a lexical database for english", *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995. DOI: 10.1145/219717.219748. [Online]. Available: http://portal.acm.org/citation.cfm?doid=219717.219748.

[148] J. R. D. Milroy, A. L. Milroy, and S. S. R. C. Great Britain, *Sociolinguistic Variation and Linguistic Change in Belfast*. Social Science Research Council, 1982. [Online]. Available: https://books.google.co.uk/books?id=mkAHHQAACAAJ.

[149] J. Milroy and L. Milroy, "Linguistic change, social network and speaker innovation", *Journal of Linguistics*, vol. 21, no. 2, pp. 339–384, 1985, ISSN: 00222267, 14697742. DOI: 10.2307/4175792. [Online]. Available: http://www.jstor.org/stable/4175792.

[150] J. L. Moreno, "Who shall survive?: a new approach to the problem of human interrelations.", English, 1934. DOI: 10.1037/10648-000. [Online]. Available: http://psycnet.apa.org/books/10648.html.

[151] M. E. J. Newman, "Detecting community structure in networks", *The European Physical Journal B*, vol. 38, no. 2, pp. 321–330, Mar. 2004. DOI: 10.1140/epjb/e2004-00124-y. [Online]. Available: http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2004EPJB...38..321N&link_type=EJOURNAL.

[152] M. Newman, "Power laws, pareto distributions and zipf's law", English, *Contemporary physics*, vol. 46, no. 5, pp. 323–351, 2005. DOI: 10.1080/00107510500052444. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/00107510500052444.

[153] D. Nguyen, D. Trieschnigg, and L. Cornips, "Audience and the use of minority languages on twitter", *Ninth International AAAI ...*, 2015. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10512.

[154] D. Nguyen and J. Eisenstein, "A kernel independence test for geographical language variation", *ArXiv.org*, Jan. 2016. arXiv: 1601.06579v1 [cs.CL]. [Online]. Available: http://arxiv.org/abs/1601.06579.

[155] B. O'Connor, J. Eisenstein, E. P. Xing, and N. A. Smith, "Discovering demographic language variation", 2010. [Online]. Available: http://repository.cmu.edu/lti/219/.

[156] R. S. Olson and Z. P. Neal, "Navigating the massive world of reddit: using backbone networks to map user interests in social media", *ArXiv.org*, p. 3387, Dec. 2013. arXiv: 1312.3387 [cs.SI]. [Online]. Available: http://arxiv.org/abs/1312.3387.

[157] O. Owoputi, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith, "Improved part-of-speech tagging for online conversational text with word clusters", in *In Proceedings of NAACL*, 2013. [Online]. Available: http://ttic.uchicago.edu/~kgimpel/papers/owoputi+etal.naacl13.pdf.

[158] U. Pavalanathan and J. Eisenstein, "Audience-modulated variation in online social media", English, *American Speech*, vol. 90, no. 2, pp. 187–213, Jun. 2015. DOI: 10.1215/00031283-3130324. [Online]. Available: http://americanspeech.dukejournals.org/content/90/2/187.short.

[159] U. Pavalanathan and J. Eisenstein, "Emoticons vs. emojis on twitter: a causal inference approach", *ArXiv.org*, Oct. 2015. arXiv: 1510.08480v1 [cs.CL]. [Online]. Available: http://arxiv.org/abs/1510.08480v1.

[160] M. Pennacchiotti and A.-M. Popescu, "A machine learning approach to twitter user classification.", *ICWSM 2011*, 2011. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2886.

[161] T. Plotkowiak and K. Stanoevska-Slabeva, "Information diffusion in twitter communities : Theory driven approach towards explaining topic based information diffusion", in *ASNA 2011 Proceedings*, Applications of Social Network Analysis, Sep. 2011. [Online]. Available: https://www.alexandria.unisg.ch/208579/.

[162] M. Prensky, "H. sapiens digital: From digital immigrants and digital natives to digital wisdom", *Innovate: Journal of online education*, vol. 5, no. 3, p. 1, 2009.

[163] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in twitter", in *Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents*, ser. SMUC '10, Toronto, ON, Canada: ACM, 2010, pp. 37–44, ISBN: 978-1-4503-0386-6. DOI: 10.1145/1871985.1871993. [Online]. Available: http://doi.acm.org/10.1145/1871985.1871993.

[164] A. Rapoport, "Spread of information through a population with socio-structural bias: i. assumption of transitivity", English, *The bulletin of mathematical biophysics*, vol. 15, no. 4, pp. 523–533, 1953. DOI: 10.1007/BF02476440. [Online]. Available: http://link.springer.com/article/10.1007/BF02476440.

[165] F. Riquelme, "Measuring user influence on twitter: a survey", *ArXiv.org*, p. 7951, Aug. 2015. arXiv: 1508.07951 [cs.SI]. [Online]. Available: http://arxiv.org/abs/1508.07951v1.

[166] F. Riquelme and P. González-Cantergiani, "Measuring user influence on twitter", *Inf. Process. Manage.*, vol. 52, no. 5, pp. 949–975, Sep. 2016, ISSN: 0306-4573. DOI: 10.1016/j.ipm.2016.04.003. [Online]. Available: https://doi.org/10.1016/j.ipm.2016.04.003.

[167] E. M. Rogers, *Diffusion of Innovations, 5th Edition*. Free Press, 2003, ISBN: 9780743258234. [Online]. Available: https://books.google.co.uk/books?id=9U1K5LjUOwEC.

[168] K. Rogers, "Why reddit banned some racist subreddits but kept others", Aug. 2015. [Online]. Available: http://motherboard.vice.com/read/why-reddit-banned-some-racist-subreddits-but-kept-others.

[169] M. Rosvall, A. V. Esquivel, A. Lancichinetti, and J. D. West, "Memory in network flows and its effects on spreading dynamics and community detection", English, *Nature*, vol. 5, p. 4630, 2014. DOI: 10.1038/ncomms5630. [Online]. Available: http://www.nature.com/doifinder/10.1038/ncomms5630.

[170] B. Ryan, "2012-04-04 online data collection and privacy", pp. 1–9, Apr. 2012. [Online]. Available: https://www.mrs.org.uk/pdf/2012-04-04%5C%20Online%5C%20data%5C%20collection%5C%20and%5C%20privacy.pdf.

[171] D. Sánchez and D. Isern, "Automatic extraction of acronym definitions from the web", *Applied Intelligence*, vol. 34, no. 2, Apr. 2011. DOI: 10.1007/s10489-009-0197-4. [Online]. Available: http://link.springer.com/article/10.1007/s10489-009-0197-4.

[172] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. P. Seligman, and L. H. Ungar, "Personality, gender, and age in the language of social media: The open-vocabulary approach", *PLOS ONE*, vol. 8, no. 9, T. Preis, Ed., e73791, Sep. 2013. DOI: 10.1371/journal.pone.0073791. [Online]. Available: https://doi.org/10.1371/journal.pone.0073791.

[173] E. Semino, Z. Demjén, J. Demmen, V. Koller, S. Payne, A. Hardie, and P. Rayson, "The online use of violence and journey metaphors by patients with cancer, as compared with health professionals: a mixed methods study", *BMJ Supportive & Palliative Care*, vol. 7, no. 1, pp. 60–66, 2017. DOI: 10.1136/bmjspcare-2014-000785. [Online]. Available: http://spcare.bmj.com/content/7/1/60.

[174] M. A. Serrano, M. Boguna, and A. Vespignani, "Extracting the multiscale backbone of complex weighted networks", English, *ArXiv.org*, no. 16, pp. 6483–6488, Apr. 2009. DOI: 10.1073/pnas.0808904106. arXiv: 0904.2389v1 [physics.soc-ph]. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.0808904106.

[175] C. R. Shalizi and A. C. Thomas, "Homophily and contagion are generically confounded in observational social network studies", English, *ArXiv.org*, no. 2, pp. 211–239, Apr. 2010. DOI: 10.1177/0049124111404820. arXiv: 1004.4704v3 [stat.AP]. [Online]. Available: http://smr.sagepub.com/cgi/doi/10.1177/0049124111404820.

[176] S. Sharoff, "Know thy corpus! exploring frequency distributions in large corpora",

[177] W.-Y. Shin, B. C. Singh, J. Cho, and A. M. Everett, "A new understanding of friendships in space: complex networks meet twitter", *ArXiv.org*, Jul. 2015. arXiv: 1507.02206v4 [cs.SI]. [Online]. Available: http://arxiv.org/abs/1507.02206v4.

[178] J. A. Simpson and E. Weiner, *The oxford english dictionary*, 1989. [Online]. Available: http://www.cinqueterreliguria.net/boat/aluminium/Oxford%5C%20English%5C%20Dictionary.htm.

[179] P. Singer, F. Flöck, C. Meinhart, E. Zeitfogel, and M. Strohmaier, "Evolution of reddit: From the front page of the internet to a self-referential community?", in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW '14 Companion, Seoul, Korea: ACM, 2014, pp. 517–522, ISBN: 978-1-4503-2745-9. DOI: 10.1145/2567948.2576943. [Online]. Available: http://doi.acm.org/10.1145/2567948.2576943.

[180] T. Snijders, "Statistical models for social networks", English, *Annual Review of Sociology*, vol. 37, no. 1, pp. 131–153, 2011. DOI: 10.1146/annurev.soc.012809.102709. [Online]. Available: http://www.annualreviews.org/doi/abs/10.1146/annurev.soc.012809.102709.

[181] R. V. Solé and S. Valverde, "Information theory of complex networks: onevolution and architectural constraints", English, in *Planetary-Scale Views on an Instant-Messaging Network*, Berlin, Heidelberg: Springer Berlin Heidelberg, Aug. 2004, pp. 189–207, ISBN: 978-3-540-22354-2. DOI: 10.1007/978-3-540-44485-5_9. [Online]. Available: http://link.springer.com/10.1007/978-3-540-44485-5_9.

[182] P. K. Stacey, D. Kershaw, N. Puntambekar, E. C. Egginton Draysey, G. Giering, Z. Anastasiou, J. Millington, C. Daramilas, J. Easton, and J. Mehlem, *Managing big data analytics projects*, 1st ed. Achamore Books., 2017, p. 110. [Online]. Available: https://dspace.lboro.ac.uk/2134/24465.

[183] I. Steadman, *Big data and the death of the theorist*. WIRED UK, Jan. 2013. [Online]. Available: http://www.wired.co.uk/article/big-data-end-of-theory.

[184] A. Stevenson, *Oxford Dictionary of English*, English. Oxford University Press, Aug. 2010, ISBN: 0199571120. DOI: 10.1093/acref/9780199571123.001.0001/acref-9780199571123. [Online]. Available: http://books.google.co.uk/books?id=anecAQAAQBAJ&printsec=frontcover&dq=Dictionary+Oxford+English&hl=&cd=2&source=gbs_api.

[185] G. Szabo and B. A. Huberman, "Predicting the popularity of online content", *Communications of the ACM*, vol. 53, no. 8, Aug. 2010. DOI: 10.1145/1787234.1787254. [Online]. Available: http://dl.acm.org/citation.cfm?id=1787254.

[186] A. Tagarelli and R. Interdonato, "Lurking in social networks: topology-based analysis and ranking methods", *ArXiv.org*, p. 4695, Sep. 2014. arXiv: 1409.4695 [1409]. [Online]. Available: http://arxiv.org/abs/1409.4695.

[187] H. Tajifel, *Differentiation Between Social Groups*. Academic Press, Inc, Jan. 1979, ISBN: 0126825505.

[188] N. Tamburrini, M. Cinnirella, V. A. Jansen, and J. Bryden, "Twitter users change word usage according to conversation-partner social identity", *Social Networks*, vol. 40, no. Supplement C, pp. 84–89, 2015, ISSN: 0378-8733. DOI: https://doi.org/10.1016/j.socnet.2014.07.004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S037887331400046X.

[189] C. Tan, L. Lee, and B. Pang, "The effect of wording on message propagation: topic- and author-controlled natural experiments on twitter", *ArXiv.org*, May 2014. arXiv: 1405.1438v1 [cs.SI]. [Online]. Available: http://arxiv.org/abs/1405.1438v1.

[190] S. E. Tchokni, D. O. Séaghdha, and D. Quercia, "Emoticons and phrases: status symbols in social media.", *ICWSM*, 2014. [Online]. Available: http://www.cl.cam.ac.uk/~do242/Papers/icwsm14_status.pdf.

[191] L. Trask, *Language Change*, ser. Language Workbooks. Taylor & Francis, 2013, ISBN: 9781134885671. [Online]. Available: https://books.google.co.uk/books?id=uzOfllBb3GYC.

[192] C. Troutman, B. Clark, and M. Goldrick, "Social networks and intraspeaker variation during periods of language change", *University of Pennsylvania Working Papers in Linguistics*, vol. 14, no. 1, p. 25, 2008. [Online]. Available: http://repository.upenn.edu/cgi/viewcontent.cgi?article=1041&context=pwpl.

[193] P. Trudgill, *A Glossary of Sociolinguistics*. Oxford University Press, 2003, ISBN: 9780195219432. [Online]. Available: https://books.google.co.uk/books?id=F5ffKI4qrEwC.

[194] P. Trudgill, "Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography", *Language in Society*, vol. 3, no. 2, pp. 215–246, 1974, ISSN: 00474045, 14698013. [Online]. Available: http://www.jstor.org/stable/4166764.

[195] O. Tsur and A. Rappoport, *What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities*, ser. content based prediction of the spread of ideas in microblogging communities. New York, New York, USA: ACM, Feb. 2012, ISBN: 978-1-4503-0747-5. DOI: 10.1145/2124295.2124320. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2124295.2124320.

[196] T. W. Valente, "Social network thresholds in the diffusion of innovations", *Social Networks*, vol. 18, no. 1, pp. 69–89, 1996, ISSN: 0378-8733. DOI: https://doi.org/10.1016/0378-8733(95)00256-1. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0378873395002561.

[197] F. Viger and M. Latapy, "Efficient and simple generation of random simple connected graphs with prescribed degree sequence", English, *Journal of Complex Networks*, vol. 4, no. 1, pp. 15–37, Feb. 2016. DOI: 10.1093/comnet/cnv013. [Online]. Available: http://comnet.oxfordjournals.org/lookup/doi/10.1093/comnet/cnv013.

[198] F.-Y. Wang, D. Zeng, K. M. Carley, and W. Mao, "Social computing: from social informatics to social intelligence", *IEEE Intelligent Systems*, vol. 22, no. 2, pp. 79–83, Mar. 2007. DOI: 10.1109/MIS.2007.41. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4136863.

[199] X. F. Wang and G. Chen, "Complex networks: small-world, scale-free and beyond", English, *Circuits and Systems Magazine, IEEE*, vol. 3, no. 1, pp. 6–20, 2003. DOI: 10.1109/MCAS.2003.1228503. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1228503.

[200] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks", *Nature*, vol. 393, no. 6684, pp. 440–442, 1998. DOI: 10.1038/30918. [Online]. Available: http://www.nature.com/doifinder/10.1038/30918.

[201] U. Weinreich, W. Labov, and M. I. Herzog, *Empirical Foundations for a Theory of Language Change*. University of Texas Press, 1968. [Online]. Available: https://books.google.co.uk/books?id=Wr1CAAAAIAAJ.

[202] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers", in *WSDM '10: Proceedings of the third ACM international conference on Web search and data mining*, Pennsylvania State University, New York, New York, USA: ACM, Feb. 2010, pp. 261–270, ISBN: 978-1-60558-889-6. DOI: 10.1145/1718487.1718520. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1718487.1718520.

[203] L. Weng, A. Flammini, A. Vespignani, and F. Menczer, "Competition among memes in a world with limited attention", *Scientific Reports*, vol. 2, pp. 1–8, Mar. 2012. DOI: 10.1038/srep00335. [Online]. Available: http://www.nature.com/articles/srep00335.

[204] L. Weng, M. Karsai, N. Perra, F. Menczer, and A. Flammini, "Attention on weak ties in social and communication networks", *ArXiv.org*, May 2015. arXiv: 1505.02399v1 [physics.soc-ph]. [Online]. Available: http://arxiv.org/abs/1505.02399.

[205] L. Weng and F. Menczer, "Topicality and social impact: diverse messages but focused messengers", English, *ArXiv.org*, no. 2, e0118410, Feb. 2014. DOI: 10.1371/journal.pone.0118410. arXiv: 1402.5443v1 [cs.SI]. [Online]. Available: http://dx.plos.org/10.1371/journal.pone.0118410.

[206] ——, "Topicality and impact in social media: Diverse messages, focused messengers", *PLOS ONE*, vol. 10, no. 2, R. Lambiotte, Ed., e0118410, Feb. 2015. DOI: 10.1371/journal.pone.0118410. [Online]. Available: https://doi.org/10.1371/journal.pone.0118410.

[207] L. Weng, F. Menczer, and Y.-Y. Ahn, "Virality prediction and community structure in social networks", *ArXiv.org*, Jun. 2013. DOI: 10.1038/srep02522. arXiv: 1306.0158v2 [cs.SI]. [Online]. Available: http://www.nature.com/articles/srep02522.

[208] ——, "Predicting successful memes using network and community structure.", in *ICWSM*, 2014. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/download/8081/8154.

[209] L. Weng, J. Ratkiewicz, N. Perra, B. Gonçalves, C. Castillo, F. Bonchi, R. Schifanella, F. Menczer, and A. Flammini, "The role of information diffusion in the evolution of social networks", *ArXiv.org*, pp. 356–364, Feb. 2013. DOI: 10.1145/2487575.2487607. arXiv: 1302.6276v2 [cs.SI]. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2487575.2487607.

[210] T. White, *Hadoop: The Definitive Guide*, ser. Oreilly and Associate Series. O'Reilly, 2012, ISBN: 9781449311520. [Online]. Available: https://books.google.co.uk/books?id=drbI%5C_aro20oC.

[211] World Book, Inc, *The World Book Dictionary*, ser. The World Book Dictionary. World Book, 2003, ISBN: 9780716602996. [Online]. Available: https://books.google.co.uk/books?id=oPW%5C_pTjpeCQC.

[212] S. Yang and G. M. Allenby, "Modeling interdependent consumer preferences", *Journal of Marketing Research*, 2003. [Online]. Available: http://journals.ama.org/doi/abs/10.1509/jmkr.40.3.282.19240.

[213] Y. Yang, "Rain: social role-aware information diffusion", in *Proceedings of the National Conference on Artificial Intelligence*, Tsinghua University, Beijing, China, Jun. 2015, pp. 367–373, ISBN: 9781577356998. [Online]. Available: https://dl.acm.org/citation.cfm?id=2887059.

[214] Y. Yang, J. Tang, Y. Dong, Q. Mei, R. A. Johnson, and N. V. Chawla, "Modeling the interplay between individual behavior and network distributions", *ArXiv.org*, Nov. 2015. arXiv: 1511.02562v1 [cs.SI]. [Online]. Available: http://arxiv.org/abs/1511.02562.

[215] S.-H. Yook, H. Jeong, and A.-L. Barabási, "Modeling the internet's large-scale topology", English, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 21, pp. 13 382–13 386, Oct. 2002. DOI: 10.2307/3073414?ref=no-x-route:9a1df0784036b87269d8743da88c7bfe.

[216] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets", in *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*, ser. HotCloud'10, Boston, MA: USENIX Association, 2010, pp. 10–10. [Online]. Available: http://dl.acm.org/citation.cfm?id=1863103.1863113.

[217] X. Zhang and Y. LeCun, "Text understanding from scratch", *ArXiv.org*, p. 1710, Feb. 2015. arXiv: 1502.01710 [cs.LG]. [Online]. Available: http://arxiv.org/abs/1502.01710v1.

[218] M. Zimmer, ""but the data is already public": on the ethics of research in facebook", *Ethics and Information Technology*, vol. 12, no. 4, Dec. 2010. DOI: 10.1007/s10676-010-9227-5. [Online]. Available: http://link.springer.com/article/10.1007/s10676-010-9227-5.

Appendix

Listing 1: Example Tweet JSON

```
{
  "coordinates": null,
  "created_at": "Thu Oct 21 16:02:46 +0000 2010",
  "favorited": false,
  "truncated": false,
  "id_str": "28039652140",
  "entities": {
    "urls": [
      {
        "expanded_url": null,
        "url": "http://gnip.com/success_stories",
        "indices": [
          69,
          100
        ]
      }
    ],
    "hashtags": [

    ],
    "user_mentions": [
      {
        "name": "Gnip, Inc.",
        "id_str": "16958875",
        "id": 16958875,
        "indices": [
```

```
          25,
          30
        ],
        "screen_name": "gnip"
      }
    ]
},
"in_reply_to_user_id_str": null,
"text": "what we've been up to at @gnip — delivering data to happy
    ↪ customers http://gnip.com/success_stories",
"contributors": null,
"id": 28039652140,
"retweet_count": null,
"in_reply_to_status_id_str": null,
"geo": null,
"retweeted": false,
"in_reply_to_user_id": null,
"user": {
  "profile_sidebar_border_color": "C0DEED",
  "name": "Gnip, Inc.",
  "profile_sidebar_fill_color": "DDEEF6",
  "profile_background_tile": false,
  "profile_image_url": "http://a3.twimg.com/profile_images/62803643/
      ↪ icon_normal.png",
  "location": "Boulder, CO",
  "created_at": "Fri Oct 24 23:22:09 +0000 2008",
  "id_str": "16958875",
  "follow_request_sent": false,
  "profile_link_color": "0084B4",
  "favourites_count": 1,
  "url": "http://blog.gnip.com",
  "contributors_enabled": false,
  "utc_offset": -25200,
  "id": 16958875,
  "profile_use_background_image": true,
```

```
    "listed_count": 23,
    "protected": false,
    "lang": "en",
    "profile_text_color": "333333",
    "followers_count": 260,
    "time_zone": "Mountain Time (US & Canada)",
    "verified": false,
    "geo_enabled": true,
    "profile_background_color": "C0DEED",
    "notifications": false,
    "description": "Gnip makes it really easy for you to collect social
      ↪ data for your business.",
    "friends_count": 71,
    "profile_background_image_url": "http://s.twimg.com/a/1287010001/images
      ↪ /themes/theme1/bg.png",
    "statuses_count": 302,
    "screen_name": "gnip",
    "following": false,
    "show_all_inline_media": false
  },
  "in_reply_to_screen_name": null,
  "source": "web",
  "place": null,
  "in_reply_to_status_id": null
}
```

Listing 2: Example Reddit JSON

```
{
   "name":"t1_ch5js6v",
   "author":"I_DO_C--C--COCAINE",
   "archived":true,
   "parent_id":"t1_ch5j1ri",
   "author_flair_css_class":null,
   "author_flair_text":null,
   "gilded":0,
```

```
    "ups":3,
    "subreddit_id":"t5_2s3qj",
    "retrieved_on":1433581667,
    "distinguished":null,
    "link_id":"t3_24awjc",
    "removal_reason":null,
    "score":3,
    "body":"FUCKIN NEO AND SHIT",
    "controversiality":0,
    "subreddit":"Bitcoin",
    "created_utc":"1398818816",
    "score_hidden":false,
    "edited":false,
    "id":"ch5js6v",
    "downs":0
}
```

Listing 3: Twitter Mention Graph SQL

```
CREATE TABLE twitter.mentions_5 as
SELECT edges.source as source, edges.target as target, min(edges.created_at
    ↪ ) as created_at, count(1) as weight
FROM(
select get_json_object(tweets.json, '$.user.id') as source, ms as target,
    ↪ CAST(UNIX_TIMESTAMP(get_json_object(tweets.json, '$.created_at'),'EEE
    ↪  MMM dd HH:mm:ss z yyyy')*1000 as timestamp) as created_at
from twitter.tweets
LATERAL VIEW explode(split(regexp_replace(get_json_object(tweets.json, '$.
    ↪ entities.user_mentions[*].id'), "\\[|\\]", ""),',')) x AS ms
) as edges
LEFT SEMI JOIN (
  SELECT distinct get_json_object(tweets.json, '$.user.id') as id
  FROM twitter.tweets
  ) AS users ON users.id = edges.target
GROUP BY edges.source, edges.target
```