RELATEDNESS AND COMPATIBILITY: SEMANTIC DIMENSIONS OF THE CONCEPT OF PRIVACY IN MANDARIN CHINESE AND AMERICAN ENGLISH CORPORA

Yuanye Ma

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the School of Information and Library Science

Chapel Hill
2022

Approved by:

Stephanie W. Haas

Rafael Capurro

James A. Danowski

Zeynep Tufekci

Yue (Ray) Wang

ABSTRACT

Yuanye Ma: Relatedness and Compatibility: Semantic Dimensions of the Concept of Privacy in Chinese and English Corpora

(Under the direction of Professor Stephanie W. Haas)

This dissertation is a study of how privacy as an ethical concept exists in two languages: Mandarin Chinese and American English. The assumption for this dissertation is that different languages have their own distinctive expressions and understandings when it comes to privacy. Specifically, this study is designed as a cross-genre and cross-language study that included two genres of language corpora for each of the languages: social media posts and news articles. In addition, the language corpora span from 2010 to 2019, which supported an observation of how privacy-related languages may have changed and evolved over the decade.

I took a mixed-methods approach, by using two computational methods: semantic network analysis (SNA) and structural topic modeling (STM), for processing the natural language corpora. When it comes to labeling and interpreting the results of topic modeling, I relied on external coders for labeling and my own in-depth reading of the topic words as well as original documents to make sense of the meaning of these topics. Last but not least, based on the interpretations of topics, I proposed four semantic dimensions and used these four dimensions to come back to code all the topics to have an overall depiction of the topics across these two languages and genres.

The four semantic dimensions, though were found in both languages, have revealed *unequal* presence in the two languages. Specifically, the *institution* dimension has much more presence in the English language; and in the Chinese language, it is the *individual* dimension that is frequently seen across topics in both genres. Apart from topics, this different emphasis on these two semantic dimensions (*institution* and *individual*) is also reflected through the semantic network analysis where the nodes with leading centrality scores over the years in these two languages differ. After considering the limitation of the data in this study, I argue that it is more cautious and appropriate to conclude that when it comes to privacy, the two languages *differ by their emphasis on different dimensions*.

This study is one of the first empirically-grounded intercultural explorations of *the concept of privacy* using language. It not only provides an examination of the concept as it is understood at the current time of writing, but also reveals that natural language is promising to operationalize intercultural privacy research and comparative privacy research.

To my advisor, friends and family, I couldn't have done this without you. Thank you for all of your support along the way.

ACKNOWLEDGEMENTS

This dissertation would not have been written, at least in its present scope and comprehensiveness, were it not for the support of my advisor Dr. Stephanie W. Haas, and my committee: Dr. Rafael Capurro, Dr. James Danowski, Dr. Zeynep Tufecki, and Dr. Yue (Ray) Wang.

First of all, my biggest critic and ally, Professor Stephanie W. Haas. Stephanie's patience and encouragement are key to my development as a junior scholar. I have always valued and desired to have a *coach* who could help train me by pushing me hard on the tiniest details. It is through Stephanie's numerous comments, questions on iterations of reviews of my crappy manuscripts, that I practiced and learned how to write and think with clarity as a scholar. Our weekly (during academic semesters) and bi-weekly (during summer breaks) meetings had kept me on track of reaching multiple goals including but not limited to my dissertation. Stephanie was there for me for many of my first-times as a junior scholar, including my first conference presentation, my first conference long paper, my first journal paper, and my first academic job interview. The lockdown of the COVID-19 pandemic in the United States began after the Spring Break in March 2020. Less than a month later, on April 2nd, 2020, I defended my comprehensive literature review on Zoom. This sudden change made many things then and thereafter more difficult for both of us. However, what's unchanged is Stephanie's unwavering support throughout this journey. I am beyond grateful for what Stephanie has done for me, she is and will be my role model as a scholar as a mentor in my own career.

The ideation of this dissertation received inspiration from classes, casual and formal discussions, and numerous readings. I started my doctoral study with only this vague interest of privacy. Not long into the program, I received Dr. Tufekci's "warning" that privacy is a well researched topic in the United States. I could recall vividly that it was at the same meeting with Dr. Tufekci that I first thought of the idea of a comparative study of privacy.

Later on, this idea of comparing privacy across cultures was further fleshed out by my taking the Comparative Ethics class led by Dr. David B. Wong at Duke University Department of Philosophy. The Comparative Ethics class and many materials included in this class, in particular the discussions on *the*

conception of the person, had inspired me to think more deeply about the *conceptualization* of concepts. In that class I learnt that comparing concepts across cultures is a *real thing* in academia that philosophers spent decades and longer enquiring about. In addition to this class, I would also like to acknowledge one paper from which I had received great inspirations quite early on. In the first semester of my doctoral study (Fall 2017), while I was buried in collections of papers in Zotero and worrying that I would never be able to finish reading them, I encountered: “Global Privacy in Flux: Illuminating Privacy Across Cultures in China and the U.S.” by Dr. Kenneth Neil Farrall. This paper’s elaboration on the concept of privacy by dissecting the vocabulary and characters that consist of this concept in the Chinese language is, in retrospect, the earliest hint of my decision to rely on language for my own research.

My continued readings on ethics eventually led me to discover Dr. Capurro and the research area of Intercultural Information Ethics (IIE). IIE immediately struck me as someone in Information Science, and it is where I situate the topic of this dissertation study. I have learnt so much from IIE scholars, including but not limited to: Dr.Pak-hang Wong, Dr. Charles Ess, Dr. Soraj Hongladorm, and I aspire to join and contribute to the future discussions of IIE.

Flashing back to my second year doctoral study: though a research topic was emerging, I was still clueless regarding the methodology part. This was when I encountered the paper: ‘“Privacy” in Semantic Networks on Chinese Social Media: The Case of Sina Weibo’, co-authored by Dr. Yuan, Dr. Feng, and Dr. Danowski. This paper ignited my interests in semantic network analysis. And I was absolutely thrilled when I received a positive reply from Dr. James Danowski who agreed to join my dissertation committee — my first committee member! Later on, Dr. Yue (Ray) Wang brought to my attention a few readings on digital humanities, topic modeling and word embeddings. Since then, I had spent months learning about topic modeling, word embedding, and other computational techniques for working with language. These readings eventually led me to discover a specific type of topic modeling (structural topic modeling, STM, developed by Dr. Margaret E. Roberts) that was one of the main computational techniques used in this study.

While I was learning to apply STM to my data in R, Dr. Roberts and Dr. Brandon Stewart generously provided consultancies through emails. While data were being gathered and processed, the following persons had provided help and consultancies: Yuzhu Jiang had assisted with collecting Weibo data; Dr. James Danowski very timely brought to my attention the availability of Twitter Academic API when I was worried about not being able to retrieve enough historical Twitter data. Dr. Yue (Ray) Wang

provided me with advice on computing. And Aaron T. Brubaker, SILS Director of Information Technology, helped set up a virtual machine for this study, which had expedited the computing process. Two UNC librarians: Lily C. Kirkhoff and Hsi-chu Bolick, helped with moderating my conversations with the vendor from which I ended up purchasing the Chinese news data for this dissertation.

I was also very fortunate in having worked with a group of colleagues on multiple projects throughout the process of this dissertation project. Including Dr. Mary Grace Flaherty, Dr. Mohammod Jarrahi, Dr. Francesca Tripodi, and Cami Goray who was a fellow student at SILS then and who is now a Ph.D. student at the University of Michigan School of Information. These projects and experiences may not be directly relevant to my dissertation project, but they provided spaces and opportunities for me to grow as a researcher. Even earlier, while I just started this journey, Dr. Arcot Rajasekar was my advisor for a couple of semesters. I would like to express my appreciation for Raja's encouragement and support early on.

During the writing of this dissertation, I was able to connect some ideas developed at a number of conferences. The most important of these is the idea of *relational privacy, which* was first published at ASIS&T 2019. The idea of understanding privacy via language was discussed in my presentation at the 2020 ASIS&T Asia-Pacific Regional Conference. Some early results from this dissertation study was presented at ASIS&T 2021 in the form of a poster. While I was preparing for the defense of this dissertation, I was able to submit parts of this dissertation to the 2022 ICA-Preconference where more discussions about comparative privacy are scheduled to happen this coming summer.

Various sections of the manuscript were read by my writing group members. In particular, Dr. Gigi Taylor and Dr. Warren Christian from the UNC Writing Center, and a number of my writing group members from around the world, offered comments and suggestions which had helped improve the clarity of my writing. At various stages of this dissertation, I have received support from my cohorts and fellow doctoral students, including: Cami Goray, Mengtian Guo, Heesoo Jang, Yuan Li, Kelsey Urgo, Wenyuan Wang, Austin Ward, and Hanlin Zhang.

I was supported through Research Assistantship by the School of Information and Library Science (SILS) 2017-2021. For four years, I have also received the Asheim Fellowship from SILS. I have received from SILS the Edward G. Holley Research Grant which supported my trip to ASIS&T 2019 in Australia, Melbourn. Earlier, the Graduate Student Transportation Grant supported my trip to the

iConference in Washington D.C. in 2019. My final year's study was supported through the Research Assistantship by the Center for Information, Technology, and Public Life (CITAP).

My study is done under the name of Information Science, but I would like to acknowledge the studies from other fields that have greatly informed me as a researcher. My initial interest in ethics can be traced back to my time in college, when I watched the videos of the open course "What is Justice" by Professor Michael Sandel for the sake of learning English. Other "leisurely" readings/podcasts/ documentary films also inspired me to think about having a scholarly life. Scholars who have particularly struck me, included but not limited to: translator Yan Fu, linguist Yuen Ren Chao, sociologist Anthony Giddens, sociologist Max Weber, sociologist Ulrich Beck, linguistic anthropologist Michael Silverstein. Academia may have disciplinary boundaries, but academic exploration is and should not be bounded. Learning about their work and career helped me see clearly my motivation for research and the questions a researcher can aim at.

I would also like to express my sincere appreciation to a few more people. First, to Dr. Don Blumenthal. Don taught me at University of Michigan School of Information between 2011 and 2013. Don passed away as I began at SILS in Fall 2017. In the years between 2014 and 2017, Don had written for me multiple recommendation letters while I was attempting multiple times with applications to various doctoral programs. In addition, Dr. Koji Yatani, my internship mentor at Microsoft Research Asia in Beijing, and Cynthia Wang, my manager when I was working at Thomson Reuters in Beijing — thank you for your letters, without which this journey may have never begun.

Finally, a special word of appreciation is owed to my friends and family, who have provided crucial support to me in the past few years. Their affection, empathy, and enthusiasm keep me a sane person and help me maintain a somewhat balanced life. They are: my cooking buddy Lu Xu, my running buddies Cami Goray and Chu-wen Hsieh, my friends from college Shujun Wang, Yuzhu Jiang; my roommates at Baity Hill at different times: Elizabeth Rosero-Pavon, and Yunwei Cao. Last but not least, my dearest partner Will Carroll, and my dearest parents Liyin Ma and Mingping Zhang: thank you for your love and I hope to pay back with more love.

Sincerely,

Yuanye Ma
Spring 2022, Chapel Hill

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

APEC	Asia-Pacific Economic Cooperation
CNNIC	China Internet Network Information Center
COPPA	Children's Online Privacy Protection Act
CCPA	California Consumer Privacy Act
CISA	Cybersecurity and Infrastructure Security Agency
CSF	Cybersecurity Framework
CISPA	Cyber Intelligence Sharing and Protection Act
DOD	Department of Defense
ECPA	Electronic Communications Privacy Act
FAA	Federal Aviation Administration
FCC	Federal Communication Commission
FISA	Foreign Intelligence Surveillance Act
FTC	Federal Trade Commission
HIPAA	Health Insurance Portability and Accountability Act
HUD	Department of Housing and Urban Development
IE	Information Ethics
IIE	Intercultural Information Ethics
ISP	Internet Service Provider
ISACA	Information Systems Audit and Control Association
LDA	Latent dirichlet allocation
GDPR	General Data Protection Regulation
MIIT	Ministry of Industry and Information Technology of China
NIST	CSF The NIST Cybersecurity Framework
NTIA	The National Telecommunications and Information Administration
NSA	National Security Agency
OSD	The Office of the Secretary of Defense
OMB	The Office of Management and Budget
OTA	Online Trust Alliance

SNA	Semantic Network Analysis
SOA	Semantic object of analysis
SORNs	System of Records Notices
SSA	Social Security Administration
STM	Structural Topic Modeling
TSA	Transportation Security Administration
USCIS	United States Citizenship and Immigration Services

CHAPTER 1: INTRODUCTION

"… comparing apparently similar or dissimilar concepts that were coined in different historical and cultural settings is dangerous in at least two ways. One danger is that we remain satisfied with merely juxtaposing such concepts; the second is that we thereby remain in such an early stage of an intercultural dialogue, defined by what may only look like a common ground or an incompatible view ..." (Capurro, 2005, p.45).

1.1 Studying the concept of privacy using language

In 1890, the seminal work of Warren and Brandeis introduced privacy as the "right to be let alone" (Warren & Brandeis, 1890). In retrospect, it became clearer that the time for the two authors was characterized by both the technological and the societal changes in America, including the wide use of cameras, telegraphs, and "newspaperization" (in other words, "unwanted newspaper publicity" (Glancy, 1979, p.8)). These changes gave the government, the press, and other institutions an increasing amount of capacity to invade previously inaccessible aspects of personal activity (Glancy, 1979; Solove, 2002). Afterwards, with more varied and advanced applications of information communication technologies (ICTs) collecting sensitive personal information by multiple parties, the understanding of privacy was reformulated to specify the aspect about access, or as "restricted access" (Allen, 1988; Moor, 1990) by appropriate entities/people. Before long, new understandings of privacy were again proposed, including privacy as a concept of "family of resemblance" (Solove, 2002), "categorical privacy" (Vedder, 1999), "contextual integrity" (Nissehbaum, 2004). The variety of understandings of privacy led scholars to argue explicitly that the meaning of privacy is "essentially contested" (Mulligan et al, 2016).

This very brief recount of the history of privacy since the end of the twentieth century reveals the conceptual variations of privacy as an ethical concept, which have been discussed mostly by privacy theorists. There are many influencing factors that can contribute to the conceptual variation, or, different understandings of privacy. Including technology, cultural norms, etc, as we will see more in the next section. Furthermore, an important factor that contributes to the various understandings of privacy is *language* itself. This dissertation project is a study of the *conceptual variation of privacy via language*.

To study privacy via language is distinctive from privacy experts' theoretical discussions in that language reflects how common people who are not privacy experts think of privacy. Language is not only an enumeration of the established meaning (Warglien & Gärdenfors, 2013), in that the language of privacy is not only an expression of existing meanings of privacy. In addition, being exposed to various language expressions of privacy is also where people who are not privacy experts keep themselves informed about privacy. In other words, language is where new meanings are learned and formed. People who are not privacy experts, for example, through scanning news articles, browsing on the internet, etc., keep themselves informed about privacy. And some people may choose to also voice their thoughts about privacy through natural language in various online platforms. This is why it's important to look at privacy through language. This *verbalization* process where individuals who are not privacy experts express the

meanings of privacy means not only expressing the pre-existing meanings of privacy; it also refers to a broader sense where implicit and potential understandings are transformed into language, or "putting into words that which is not language" (Hirschauer, 2007, p.415). In other words, an individual's verbalization of the concept of privacy is a process where the individual keeps learning about the meaning of privacy through interactions with available privacy expressions. It is a way to keep sustaining and informing oneself about the possible meanings of privacy.

How privacy is typically understood will depend on the culture an individual finds themselves in. For instance, privacy in American culture has strong associations with *individual autonomy* (Capurro, 2005) and *individual liberty* (Whitman, 2003). Over time, these understandings of privacy become implicit, at least when privacy is invoked in American culture. In that, these implicit understandings turn privacy into a *symbol* with associated meanings and connections. Similar processes could happen in other cultures when understanding privacy. The fact that the language of privacy could become a ready symbol is not inherently problematic since it helps the communication process by saving effort, "Using language to communicate about concepts therefore involves a process of discovering without an excess of mental effort a cognitive scenario where the implications of an expression are satisfactorily coherent." (McGregor et al., 2015, p.57). However, the understanding of privacy as a ready package could become *problematic* when it is taken for granted during communication between and among multiple languages, to the point that it risks being "masked" or "objectified" (Ng & Bradac, 1993, p.147). In other words, these implicit and associated understandings are taken for granted. This is perhaps part of the reasons why privacy researchers are increasingly motivated to examine the variations of the meaning of privacy in different social and language contexts (Abokhodair et al., 2016; González et al., 2019; Yuan et al., 2013).

The variety of language (including different languages, different genres, and different periods of collection of language) provides an opportunity to understand how different people understand privacy differently. In particular, such variations can be highlighted when one language is compared to another. For example, a comparative study design could either compare two completely different natural languages (like Chinese and English); or, it could compare two genres of language expression in one language. Existing studies that have used natural language to understand privacy have yet to fully leverage the contrastive design. Existing examples like Yuan et al., (2013) and Abokhodair et al., (2016) studied privacy primarily in one natural language (Mandarin Chinese and Arabic, respectively) in the genre of social media posts. Both studies only included one language and made discussions using

English as an implicit comparison. Their comparisons are hence considered one-way, because they only examineed one language.

This dissertation project is designed to be a two-way comparison of the concept of privacy using two languages, Mandarin Chinese and American English. More specifically, this dissertation is based on corpora that consist of two genres of language, social media and news articles, and span ten years (2010-2019). Previous studies have shown that leveraging a specific genre of language (for example, personal correspondence) can provide a unique perspective for understanding privacy (Zarrow, 2002; see also McDougall & Hansson, 2002). Hence, the selection of these two frequently used genres of natural language enables an examination of privacy as it has been expressed and enacted in these specific forms of expressions. Most importantly, the comparison across two languages enables me to approach privacy as an intercultural information ethics (IIE) (Capurro, 2005; 2008) subject. The comparison can help reveal distinctive assumptions and traditions underlying privacy in each of the languages. Based on these revelations I could approach more specific questions including if and what do these two languages share when it comes to understanding privacy.

1.2 Language and conceptual understanding

Humans are social animals who use language as a medium for shared meaning and understanding (Dove, 2014; Mercer, 2013). In a sense, language not only functions to help individuals understand their world; language, as it is practiced collectively, also creates meaning (Mercer, 2013; Rączaszek-Leonardi et al., 2017). Hence, language can be employed as a window *to understand how we understand*. Furthermore, the variety of languages suggests questions such as how the difference of language may impact people's understanding. Examples of how various languages are associated with different understandings include the work of Jackson et al., (2019) where different languages conceptualize emotions differently; the work of Gieck et al. (2016), where it was shown that historical events are understood differently in various languages; and Uz (2014), where it was shown how a single *concept* may be interpreted differently in different languages.

Gumperz (1964) developed this idea of *linguistic repertoire* to refer to "... all the accepted ways of formulating messages. It provides weapons of everyday communication. Speakers choose among this arsenal in accordance with the meanings they wish to convey" (p.138). Hence, language can be used to capture the concept of privacy via its *linguistic repertoire*. Furthermore, *linguistic repertoire* exists in the actual use of language and it is *a result* of the individual language practitioner's interaction with the society (Busch, 2017):

> "I do not understand the speaker as an (independently acting) individual but—in a poststructuralist move — as a subject formed through and in language and discourse, and I understand the repertoire not as something the individual possesses but as formed and deployed in intersubjective processes located on the border between the self and the other" (p.346).

That is to say, the meaning of a concept exists in the communication of *groups of people* through *shared language*. In other words, to understand a concept through language is *not* to grasp the standalone or ready meanings contained within the language as if one is collecting something contained in a bucket. Rather, to understand a concept through language is to shed light on how the meaning of the concept is being formed through the interaction between language and collective individuals that make use of that language.

> "Ultimately it is the individual who makes the decision, but his freedom to select is always subject both to grammatical and social restraints. …The power of selection is [therefore] limited by commonly agreed on conventions which serve to categorize speech forms as informal, technical, vulgar, literary, humorous, etc." (Gumperz, 1964, p.138)

Describing concepts by using language is particularly helpful for understanding particular kinds of concepts, in this case, *abstract* concepts (Borghi et al., 2019; Dove, 2014; Scorolli et al., 2011; Villani et

al., 2019). *Abstract concept* is proposed in contrast to *concrete concept*. Unlike concrete concepts that typically have a physically tangible and specific referent. For example, “apple” as a concrete concept can refer to something less disputable in terms of its meaning; as opposed to concepts like “love”, or “beauty”. These latter two concepts are abstract and could present much more challenges for understanding and are much more disputable depending on people who are practicing these words. In other words, abstract concepts, unlike concrete concepts, do not have ready and concrete external objects of reference. In a way, the meaning of abstract concepts depends more on the use and interpretation of language.

Beyond using language, there are other ways to study abstract concepts, including emotions and body gestures (Hill et al., 2014). However, language remains a crucial way to understand abstract concepts (Bowler, 1975; Zdrazilova et al, 2018). Because, unlike concrete concepts which can ground meaning primarily in concrete embodiments, abstract concepts need to rely on language as symbols to keep its meaning, in that “the contiguity of a sign with another sign offloads the necessity for grounding” (Louwerse, 2018, p.576).

Furthermore, relying on languages enables the study of abstract concepts in three more ways. First, language can be used for identifying the *substructure* (Evans, 2006) of meaning:

> “... conceptions are a function of language use. Lexical representations, or rather more technically, lexical concepts, represent the semantic pole of linguistic units, and are the mentally instantiated abstractions which language users derive from conceptions and the specific semantic contribution perceived to be associated with particular forms.” (p.499).

This way of understanding concepts recognizes that meaning is complex and multi-dimensional, and can reflect the complex social context in which language is used. In other words, to describe the meaning of a concept through language means to describe how a concept is invoked when it is in actual usage. For example, one could identify the metaphorical associations of a concept in language when used (Bundgaard, 2019; Xu et al., 2017).

Another reason that studying abstract concepts via language can be useful is that it enables us to better understand how meaning is *enacted* through language, in addition to how meaning could be *represented* by language. This is a difference that Lupyan & Lewis (2019) refers to as the two ways language and words function: they could function *as mapping,* and *as cues*. Language is not only capable of mapping or reflecting meaning, language is also capable of invoking and enacting meaning. Language enables meaning because the meaning of a concept also depends on the context in which this particular language operates, or “symbolic interdependency” (Louwerse, 2018). Partly because of the symbolic

interdependency, over time, the meaning of concepts could become “arbitrary symbols” (Corballis, 2002; Louwerse, 2018, p.583) in that the context concepts operate in also becomes implicit.

The third reason that mapping abstract concepts via language can be particularly helpful is that language can be preserved and recorded. Recorded language could reveal, for example, by using semantic distribution, a particular concept and its linguistic associations. Each specific genre or format of language, like news articles, novels, or legal reports, could provide a unique affordance and boundary for understanding. The genre itself determines the expression of a concept. For example, certain topics are most likely discussed in certain genres and could be framed variantly by genre, or with different sentiments (Ceron, 2015; Du et al., 2019; Jiang et al., 2018; Zhao et al., 2011). This way, language can be considered as data or evidence through which observation of the subject concept occurs. Moreover, since language can be preserved over time and space, a mapping of language can be informative for understanding concepts and conceptual changes in time and space.

Lastly, two important assumptions underlie the study of concepts. First, the assumption of the existence of abstract concepts, or concepts as *abstract objects*; and second, assumption of the coherence of concept over time despite its change and variation (Rosen, 2001; Margolis & Laurence, 2019). In that, though a concept could change dramatically compared to how it once was some time ago, the *essence* of the concept still holds so that it can still be considered as the *same* concept. Here, a similar analogy may be the perception of one's *self* or *identity*: in that it is in constant change, but a person, even after many years, still could be considered as the same person. With these understandings about language and concepts, we now move on to discuss more this specific concept that is the subject of this dissertation study, *privacy*.

CHAPTER 2: RELATED WORK

“Indeed, to get a handle on our transatlantic privacy conflicts, we must begin by recognizing that continental European and American sensibilities about privacy grow out of much larger and much older differences over basic legal values, rooted in much larger and much older differences in social and political traditions.” (Whitman, 2003, p.1160)

2.1 A review of privacy research

2.1.1 The social and cultural aspects of privacy

Despite that many still rely on this heavily criticized view of understanding privacy as *having control* over one's personal information (Radaelli et al., 2018; Vedder, 1999); increasingly, scholars have begun to propose richer conceptualizations of privacy that speak to the contextualized and networked nature of the social environment in which privacy as a concept exists, and recognizing that privacy is "essentially contested" (Mulligan et al., 2016).

The social aspect of privacy refers to how understandings of privacy derive from social norms where power and value are deeply contingent on the social context and traditions. For instance, in Chinese culture, it is more acceptable to disclose people's financial income information than in American culture (Acquisti et al., 2015). Or, cultures could demonstrate different expectations of privacy for different genders: one example is that females' privacy is of profound importance compared to males' in Arabic culture (Abokhodair et al., 2016). Studies have found that participants in India reported fewer concerns with providing personal health information online and lower levels of privacy concerns than Americans (Kumaraguru & Cranor, 2006). Even when compared across cultures that are neighbors, there can be drastic differences in terms of the expectations of privacy: when compared across Norwegian and Denmark, the information published online by the Norwegian Tax Authority is only considered highly sensitive by people from Denmark (Ess, 2019).

Moore (1985) demonstrated well how in human history, the concept of privacy could be meaningful in a diverse range of social situations, including the physiological (e.g., excretion), the intellectual (the space that scholars need to produce independent work), and the psychological (to temporarily get away from either the domestic or public space). Westin (2003) offers a three-level perspective to understand the social and cultural aspects of privacy. At the most macro level, it refers to *political* factors that demarcate the overarching division between public and private, and democratic vs. authoritarian societies. At the most micro level, it refers to how the individuals concern the *personal* and how each individual's specific needs and choices can vary regarding privacy. In between the macro and micro is the *socio-cultural* and *organizational* ones which refers to factors like race, wealth, all could have an impact on privacy.

Gao & O'Sullivan-Gavin (2015) put China's development of consumer privacy protection in the most recent decades into four periods (the 1980s, 1990s, 2000s, and since 2010), and recognized that

each period is situated in its economic, political, and social contexts. Some studies focus on specific components of society to examine their impact on privacy, for instance, gender (Abokhodair et al., 2016; Tifferet, 2019); or, socioeconomic status (Marwick et al., 2017).

Another even broader frame that has been leveraged to understand privacy from a cultural perspective is the Western and Non-Western culture contrast. Privacy theorists had traced the connection between privacy and the idea of *individual autonomy* in the West (Schwartz, 2000; Cohen, 2012). Meanwhile, studies have begun to discuss the conceptualizations of privacy that are not based on typical Western philosophical traditions in relatively recent years (Hongladarom, 2016; Ma, 2019a; Reviglio & Alunge, 2020).

In addition to examining each culture and society's distinctive understanding of privacy, scholars also showed interest in capturing a society's changing understanding of privacy in time, though with broad strokes. In particular, the change of privacy from a somewhat negative concept to a more positive and valuable one is discussed in both Western and Non-Western cultures. For example, Lü (2005) suggested that privacy is no longer being perceived with as strong a negative sense as it might have been in certain historical moments in Chinese history. This evolution of privacy also exists in the English language: privacy had been associated with a negative meaning in English until about the 1700s (Baldwin Lind, 2015, p.59).

Apart from these examinations of specific cultural components to understand privacy, another way to understand existing research of privacy is to divide work between seeing privacy as an *intrinsic* versus *extrinsic* value. Altman's (1981) Privacy Regulation Theory argues for understanding privacy as "an interpersonal process" that is valuable in itself. The seminal definition of privacy as "the right to be let alone" by Warren and Brandeis (1890) framed privacy as an intrinsic value of one's inviolate personality. In comparison, subsequent definitions of privacy started to lean towards understanding privacy as a tool to achieve some other goals. For example, privacy was considered as the "right to conceal discreditable facts about himself" (Posner, 1978), where privacy was considered as a tool to prevent potential harm being caused to the person.

2.1.2 The technical aspects of privacy

The studies of privacy have been increasingly situated within specific digital technological contexts in the most recent years, such as privacy for drones (Yao et al., 2017), or privacy in big data analytics applications (Tran & Hu, 2019). Today, digital technologies are the major focus for understanding privacy. However, the understanding of privacy in the context of technical or technological affordances extends beyond digital technologies.

Historians have sought to discuss that understandings of privacy might have interacted with the existence of private dwellings. For instance, it was observed that in England during the period between the ascension of Elizabeth and the English Civil War, rural homes were rebuilt and new ones tended to be designed and structured to have two stories (instead of just one); with more rooms, servants were able to move into separate rooms. These changes were afforded by not only cheaper materials, but also by the change in attitudes towards housing and the desire for privacy within those houses (Berg, 2018). In that, on the one hand, the technical capacities afforded more privacy. While on the other hand, the desire for privacy also precipitated the change in physical arrangement.

Private bedrooms (Reid, 2012) and private bathrooms (Stone, 1991) are both recent implementations if considered against human history. Both implementations have shaped people's understanding of what is *private*. Even today, private bathrooms that are taken for granted in many parts of the world could still be absent in other places. Shared bedrooms of *kang*[1], for example, though less seen, still exist in rural areas of China.

How physical living conditions could have a direct impact on the conceptualization of a concept like privacy was discussed in Mizutani et al., (2004): the bedroom and living conditions in the traditional Japanese society may have influenced the understanding of privacy in the Japanese society. More recently, it has been discussed how the urban living conditions in Chinese families might have interacted with the idea of children's privacy (Naftali, 2010). It is beyond the scope of this dissertation study to tease out the causal relationships between the concept of privacy and the affordance of physical conditions and technologies. However, undoubtedly, the establishment of the concept of privacy is highly influenced by the physical environment.

---

[1] "kang, a heated platform connected to the kitchen stove in northern Chinese houses. The bedrolls of all the members of the household are stacked on it because it is where they sleep" (McDougall, 2002, p.215).

Information communication technologies have been spurring discussion over privacy since the 1960s, it was argued that new challenges for privacy were not being raised by vague scientific developments, but by very concrete items embodying technological progress: for instance, battery-powered microphones, portable tape recorders, telephones that could be connected to a single mainline, etc. (González Fuster, 2014). One information technological advancement, the telephone, was a wild success in the early 20th century (Ferenstein, 2015). How understanding about privacy responds to the evolving information technological capabilities can be seen via a few court decisions in North America. For instance, in the 1928 Supreme Court decision over *Olmstead vs. the United States*, the court ruled that the Fourth Amendment did not apply to wiretapping a person's home phone. In another case, The Pennsylvania Supreme Court's holding in *Commonwealth v. Rekasie* states that one has no reasonable expectation of privacy during a telephone conversation conducted in one's own home (Tener, 2002). In contrast, the Ontario Court of Appeal in January 2003 acquitted Tessling of charges related to a large quantity of marijuana found in his home. The court ruled the police violated his privacy rights by not getting a search warrant before using Forward-Looking Infra-Red aerial cameras to detect heat coming from buildings on Tessling's property (Wageningen, 2004).

In more recent decades, the discussions of privacy have been dominated by the application of information communication technologies (ICTs); specifically, how technological applications play a role in continually challenging the understandings of privacy, including wiretapping, video surveillance, biometric identification, Global Positioning System (GPS), and Radio Frequency Identification (RFID) (Holvast, 2008).

More recent discussions of privacy have revealed that the applications of various ICTs demand a more fundamental re-conceptualization of privacy by challenging our assumptions regarding *the person* when thinking of privacy. For example, privacy scholars have begun to propose an alternative conceptualization of privacy that can better accommodate the information reality: *group privacy* (Mittelstadt, 2017), which fundamentally challenges the *individualistic* understanding of privacy. Being categorized as a member of such a group could drive a variety of automated decisions with harmful or beneficial effects on individuals. For example, being identified as a member of a group of "healthy people" could result in a preferential rate for health insurance. The existence of such groups is essentially informational, and these informational traits are only meaningful to algorithms on the group level when individuals' information is considered together. Literally, this information is beyond the scope of one's own

information and is not under the individual's control; in other words "... a recommendation system does not record an already delineated characteristic of the user but, rather, it is through the operations of the system that this 'characteristic' is constituted and brought to the attention of users" (Karakayali et al., 2017, p.8).

Privacy scholars went on to further clarify the meaning of *group* in *group privacy*, which is also termed as the "collective aspects" (Garcia et al., 2018; Sarigol et al., 2014). Loi & Christen (2019) explicitly discussed that there are two "types" of groups. The first type, or "type-a", is the more traditional way of understanding human groups. For example, a close circle of friends or family is a group. Intimate information is shared only within this close group. The second type of group, the "type-b" group, corresponds to groups created by online algorithms and inferences. Type-b groups' characteristics include: the member of a type-b group may not even be aware of themselves belonging to such groups; and such groups can be arbitrary/random and ephemeral. In addition, group membership may have a far-reaching effect on the person and persons being categorized (Cheney-Lippold, 2011; Haber, 2019). The examination of the conceptualizations and understandings of privacy offers an opportunity to update our *language* and working *vocabulary* of privacy.

Existing understandings of personal information and privacy rely on the idea of individual autonomy (Cohen, 2012; Capurro, 2005). Specifically, this way of conceiving the person as being individualistic, with self-government (Soares, 2018), serves not only as the conceptual and theoretical foundation for understanding privacy, but has been used to come up with specific designs and implementations of privacy in policy and technology; for instance, companies updating their privacy policies, or requesting user consent. For example, Internet users being asked to give their consent for the use of cookies in Google's Chrome browser (The Associated Press, 2022). The influence of this individualistic view extends to the understanding of genetic privacy. Following this individual control narrative, genetic privacy considers an individual's right of privacy to cover a new type of information that is genetic information, while the focus is still on informed consent, control, and voluntary disclosure (Lunshof et al., 2008; Erlich et al., 2014).

Though being criticized for not being able to encompass the current environment, the individualistic or individual autonomy-based conceptualization of privacy remains to be an important perspective for thinking of privacy. The individualistic view has been criticized for its narrow consideration of individuals in isolation (Vedder, 1999). "The limited guidance offered by informational self-determination

as a core conceptual component of privacy presents a challenge and an opportunity to expand the way we conceive of privacy, its risks and our strategies for protection" (Mulligan et al., 2016, p.2). Specifically, it has been pointed out that an individualistic-based privacy understanding is inadequate to understand privacy risks in networked societies (Radaelli et al., 2018). The shortcoming of individualistic-based privacy is highlighted when put in the context of recommender systems. Recommender systems are commonly known in the form of "micro-targeted ads" (Korolova, 2011), "personalized ads" (Tran, 2017), or "behavioral targeting" (Lee et al., 2018). Micro-targeted ads, taking Facebook as an example, may appear like the user's Facebook friend's updates (Korolova, 2011); whereas in reality, they are customized advertisement targeting groups generated by algorithm. Using the example from Mittelstadt (2017, p.477): a group can be "dog owners living in Wales aged 38–40 that exercise regularly." This information, being aged 38-40 and having a dog, etc. does not belong to any individual, it would be odd to say an individual owns or can control this information.

2.1.3 The sociotechnical aspects of privacy

The “sociotechnical” is a concept in Science and Technology Studies (STS), it denotes the interplay of technological infrastructures and practices of social and material agencies, involving both human and non-human actors. Sociotechnical was proposed because “a purely technological understanding of the complex, interactive systems fall short of their deep embeddedness and situatedness in, or entanglement, social (or cultural) contexts” (Ochs & Ilyes, 2014, p.75). Sociotechnical not only refers to the case that under many situations, the social and technical aspects are also intertwined and cannot be separated as they might be in theory; but also, sociotechnical systems are systems in which social and technical factors shape one another (Ellison & boyd, 2013, p.166).

From a sociotechnical view, studies have strived to illustrate how the understanding of privacy is subject to both social and technical influences; and most importantly, to the interplay of the two. Westin (2003)’s categorization of the contemporary development of privacy in the United States into three phases was a delineation of the socio-technical phases in which privacy was situated. The first phase, 1961 to 1979, was where data surveillance technologies were embraced by both the government and the private sector; this period also witnessed the “rise of advocacy journalism and television-age media competition” (p.437) as a result of the interplay of both the social and technical. The second phase from 1980 to 1989, was the period where enhanced computer and telecommunications performance and distributed computing and personal computing began to enter corporate and individual people’s lives. The third phase from 1990 to 2002, was a period that witnessed multiple major developments in technology, including the Internet, cellphone, large-scale data warehousing, and their implementations.

The seminal work of Warren and Brandeis (1890) introduced privacy as the “right to be let alone”. Though being criticized as being too narrow and vague (Solove, 2002), the two authors were responding to concerns triggered by both the technological and societal changes in American society at the time. For instance, the introduction of the Eastman Kodak Camera in 1884, which made photography portable and affordable to the general public. This coincided with the rise of sensationalistic journalism and circulation of the journalism and the public’s passion towards “newspaperization” (Solove & Schwartz, 2015). The interaction of the technical and the social together gave the increasing capacity to government, the press, and other institutions to invade previously inaccessible aspects of personal activity (Glancy, 1979; Solove, 2002).

The physiological examples discussed by Moore (1985) (see Section 2.1.1 The social and cultural aspects of privacy) may appear to be less relevant to today's privacy issues; however, with the application of biometrics information and the collection of real-time geolocation data (Jain et al., 2016; Vincent et al., 2019), the issue of physiological privacy can become meaningful again because of the interaction of the socio and the technical.

In summary, we have seen that the social, the technical and the socio-technical each offer a distinctive perspective for understanding the conceptualization of privacy. Moreover, it is increasingly the interaction of the social and the technical that provides the context in which understanding of privacy continues to evolve: from *right to be let alone* (Warren & Brandeis, 1890), to *contextual integrity* (Nissenbaum, 2004), and then to *group privacy* (Taylor et al., 2017; Mittelstadt, 2017). Lastly, I have also pointed out how despite the socio and technical changes over the years, the underpinning assumption of privacy (namely, the individualistic assumption of the person) can remain at play across varying socio and technical contexts.

2.2 Using language to research privacy as an intercultural information ethics concept

2.2.1 Privacy as an intercultural information ethics concept

Existing intercultural discussions of privacy devote their attention to differences when comparing the understanding of privacy in a Non-Western culture to a generalized Western understanding of privacy. To understand the story of privacy in recent Chinese society, Lü (2005) proposed that because of Chinese culture's collectivist nature, it could be quite difficult to have privacy, which is supposed to grow in individualistic cultures. This portrayal perhaps captures something true for understanding privacy in Chinese culture; however, it also risks oversimplifying not only Chinese culture[2], but also the multiplicity of Western societies, and privacy itself. Perhaps a more accurate and cautious way to put it is that the comparisons of the conceptualization of privacy across cultures can be considered as a spectrum, where there can be a range of similarities and dissimilarities. Focusing the attention on the comparison between the collectivist aspect of Chinese culture and the individualistic characteristic of privacy is really drawing two apparent two dissimilarities together, which is an interesting and valuable way to start a discussion. However, to move forward and produce more meaningful conversation to fully reveal the complexities of privacy in different cultures, perhaps more attention shall be given to the rest of the spectrum[3]. Lü (2005) did raise an interesting issue regarding on what grounds privacy might be supported in Chinese society. Whereas in Western cultures the way privacy was advocated has originated from a primarily individual autonomy perspective, it is likely that China, because of the mixture of its traditional culture and the influx of western ideas, will have to come up with a story of advocating privacy from "both individual and collective perspectives" (p.14). This recognition is essential in that it tries to clarify the working ground for the continuation of the conceptualization of privacy for Chinese culture.

Nakada & Tamura (2005) argued that privacy was an imported concept for Japan, by demonstrating that the Western conceptualization of privacy does not fit with the Japanese worldview *trichotomy*, which consists of *Ikai*, *Seken*, and *Shakai*. In Japanese culture, *Seken* refers to the aspect of the world that consists of traditional and indigenous worldviews or ways of thinking and feeling; *Shakai*

---

[2] As Wong (2009) described, "Chinese culture ... is mainly constituted by Confucianism, Daoism and Zen Buddhism, and each has their own moral systems." (p.55). Each tradition could have different implications for understanding privacy.

[3] The discussion in Ma (2019) is an early attempt in this direction, where the discussion was directed towards compatibilities between the conception of "relational person" that can be found in both the Confucinism tradition, and the feminist philosophy.

includes modernized worldviews and ways of thinking influenced in many respects by the thoughts and systems imported from 'Western' countries; whereas *Ikai* is the world of 'the other(s),' i.e., the hidden or forgotten meanings or values in *Seken* or *Shakai* as normal aspects of the world. *Ikai* is the aspect of the world from which evils, disasters, crimes, and impurity – along with freedom and the sources of energy-related to art and spiritual meanings – seem to emerge (Nakada & Tamura, 2005; p.27).

Because of the influence of Western culture, the Japanese find themselves grappling with two systems of understanding: one of traditional Japanese culture (which has been heavily influenced by Buddhism and Confucianism), and the other of the West, which was imported to Japan more recently. It was suggested by Nakada & Tamura (2005) that a typical Western understanding of privacy might be only applicable within the realm of *Shakai*; while *Seken* is related to the social relationships and the social community where an individual finds him/herself. And from the perspective of the larger social community, namely *Seken*, what might be considered as an invasion of privacy in Western societies, is considered as *necessary* in Japanese society. For example, victims' information needs to remain confidential in American societies. However, in Japan, "people need information about the victims' personalities and relationships to understand the meanings of this homicide …", because personal information is not just about individuals, it is entitled of broader social duties. "What may seem like a violation of privacy to Westerners is thus justified from the perspective of Seken" (p.30).

To further explain the two systems of understanding and their implications for understanding privacy, Nakada & Tamura (2005) discussed the concepts of *public (Ohyake)* and *private (Watakusi)* as an example. The traditional Japanese understanding would say "things related to Watakusi are less worthy than things related to Ohyake" (p.32); however, this understanding of *Watakusi* in Japan misaligns with how *private* is understood in the West. The downplay of *Watakusi* is perhaps one of the reasons that privacy may not be as important as it is in the West. But perhaps more importantly, it further suggests three things that are crucial for understanding privacy across cultures. First, an understanding of how the concept of privacy and its understanding hinge upon other related concepts. Second, how the partial correspondence of the concept between Japanese and American understandings is not unique to just privacy; the concept of *Ohyake* and *Watakusi* cannot be simply equated with the *public* and *private* in the West. And third, the coexistence of two systems, *Ohyake* and *Watakusi*, and the *public* and the *private*, could give rise to an internal conflict that perhaps each Japanese individual or any individual who lives in such an environment will need to resolve.

Based on the above illustrations, Nakada & Tamura (2005) suggested that Japan might have imported privacy only partially. And it was further suggested that not only is privacy a less valued concept by the Japanese people's view, other more dominant norms regarding one's social roles, for instance, the concept of *Bun (份)*, is what guides people's behavior in Japan. *Bun* in general refers to different roles depending on one's relationship with other people (Nakada & Tamura, 2005, p.31).

An alternative way of describing the incompatibilities of the concept of privacy across cultures comes from Mizutani et al., (2004). When trying to decipher the comparison of privacy across Japanese society and American society, they make this distinction between *descriptive privacy* and *normative privacy*. *Descriptive privacy* is understood as *the absence of privacy as a matter of fact (p.121)*, it describes the situation regarding privacy. For instance, generations of families live together and do not have each of their own separate rooms. In contrast, *normative privacy* refers to the situation that, although privacy is absent, expectations of it still exist; or, such normative rules about privacy still exist. Mizutani et al., (2004) argue that the lack of the former does not indicate the lack of the latter, and the existence of the former might not be a guarantee of the latter.

Indeed, Mizutani et al., (2004) consider the lack of privacy in Japanese society largely a result of practical constraints (e.g., close and limited physical living spaces), while privacy as a concept of inherent value still exists in Japanese society. Based on the distinction of *descriptive* and *normative* privacy, it seems possible for Japanese and American cultures to share a "minimal conception" (p.124) of privacy, while a full equivalence of the richness of privacy in the two cultures is less likely. Mizutani et al., (2004)'s concern resonates with Ess (2005) in that it would be a rush to conclude the absolute nonexistence of the concept of privacy in Japanese culture. In other words, Japanese society seems to have imported some aspects of the Western concept of privacy but not the ''individualistic'' perspective that ascribes privacy to the dignity of the person (Capurro, 2005, p.46).

To summarize the discussion thus far, Mizutani et al., (2004) is insightful in that it makes a distinction among multiple levels or dimensions of how privacy exists. In that, explicit normative elaborations only constitute a portion of how the concept itself exists. In real-world situations, the concept can present itself in people's behaviors, and specific design, and many more ways of expressions or enactment. The intimate and intricate connection between the concept of privacy and its expressions in various ways deserves more discussion. Among many ways of expression, language deserves particular attention. Mizutani et al., (2004), in particular, brings out that language can both guide and mislead, in that

the absence of certain language object does not indicate the complete absence of meaning, “the absence of a single word to describe a concept does not mean the concept is totally lacking, it does suggest that the contours of that concept and its discursive role may be different” (p.121).

2.2.2 Privacy in Mandarin Chinese and English

The meaning of privacy as a concept has been in constant evolution (Tavani, 2007), such changes can be manifested through the actual vocabulary used to refer to privacy. The word “privacy” per se in the English language has remained stable for centuries; in contrast, in Mandarin Chinese, the word “privacy” has gone through some significant changes in recent decades. The word “privacy” in the Chinese language as it is accepted currently is a compound word that consists of two Chinese characters, yin (隐) and si (私), each character with its meaning. The pre-modern Chinese written language, today known as classical Chinese, retained a strong monosyllabic character throughout its use well into the 19th century. Only a tiny percentage of the population, scholars and government officials, could understand classical Chinese, or manage the often ambiguous, highly contextual meanings clustered around each individual character (Rosemont, 1974). Compound words began to appear during the Han dynasty (206 BC-220 AD) but did not increase substantially until modern times, from roughly 20% percent of the written lexicon before the Qin dynasty to more than 80% today (Shi, 2002).

It turned out that privacy as the compound word yin3si1[4] was quite a recent adoption of practice (Gao & O’Sullivan-Gavin, 2015). Before yin3si1, it was a different compound word yin1si1 that was used in the context of the Chinese language. The transition from yin1si1 to yin3si1 appears to have occurred over the past three decades (McDougall, 2005). It was further suggested by McDougall that the last decade of the 20th century was when yin3si1 became “an independent concept that did not need to be contrasted with the ideal of public service” (McDougall, 2005, p.112). This transition that had occurred over several decades was also a period of time when Chinese society witnessed a growing appreciation of privacy (Gao & O’Sullivan-Gavin, 2015).

Specifically, the Criminal Procedure Law (National People’s Congress, NPC, 1979), which came into effect in 1980, used the term “dark secrets (yin1si1 阴私) when stipulating that trials involving personal secrets should not be open to the public” (Gao & O’Sullivan-Gavin, 2015, p.235). It was suggested by McDougall (2005) that the first bilingual dictionary appearance of yin3si1 might have occurred in A Chinese-English dictionary; the dictionary was compiled by the English Department of Peking Foreign Languages College in 1979 (p.113).

---

[4] I use numbers to indicate the tones: 1 indicates the first tone, and 3 indicates the third tone; yin1 is not the same character as yin3.

Looking back, researchers have suggested that the period after 1949 (the year the People's Republic of China was founded) was when the concept of privacy was harshly attacked in the Chinese culture when "private property was banned, and personal desires, including the desire for a space of one's own, were strictly abhorred among Chinese citizens, old and young alike" (Ong & Zhang, 2008, p.6; Naftali, 2010, p.301). Privacy in the Chinese language, in particular, its change from yin1si1 to yin3si1, correlates with the decades-long process of privacy adopting an increasingly positive tone. In a sense that, the transition in vocabulary may be both a result of, and a contributing factor to privacy adopting an increasingly positive tone.

The change of the vocabulary of privacy in the Chinese language has motivated researchers to map out the semantics associated with privacy in Mandarin Chinese one way or another. Specifically, several researchers have discussed that in the Chinese language context, the two characters that constitute privacy and especially the second character si1 (private) still could bear negative connotations; namely, the meaning of si1 inherently is part of the meaning of privacy so that privacy/yin3si1 bears a derogatory sense to some extent (Farrall, 2008; Huang, 2000; McDougall, 2005; Naftali, 2010; Zarrow, 2002). The negative connotation of si1 originates from its antonym to gong1 (public, 公); another antonym of si1 in the Chinese language is guan (official, 官). In either situation, the character si1 bears a negative connotation, with implications of "disreputable actions carried out in secret and/or from disreputable motives" (Farrall, 2008, p.2). McDougall (2004) further explained the possible cultural and philosophical sources of the negative sense of the character si1 (私) by tracing it back to Confucianism.

The other character, yin, can be another source of a negative connotation for yinsi. yin3 (literally means hidden) and yin1 (literally means shade, feminine, negative, or even sinister, but arguably it carries significantly more meaning and cultural connotations) indicate two different Chinese characters. At the same time, both have been used to form the compound word of privacy until yin3 has become the commonly accepted one at present in the most recent decades. Because yin1 could carry a more significant derogatory sense than yin3, the shift from yin1si1 to yin3si1 in the Chinese language could reflect how the understanding of privacy in the Chinese language has been an ongoing process and a process of getting rid of the negative connotations, especially within the most recent years, which can also be seen in the expansion of the meaning of privacy in court decisions.

In an opinion on the implementation of the Criminal Procedure Law in cases of rape, China Supreme People's Court (1982) used the phrase "隐私" (yin3si1) to describe cases of sexual crimes and emphasized that trials of such cases should not be open even to the internal staff of law enforcement agencies other than those directly involved in the trials. In Opinion on Several Issues Regarding the Implementation of the General Principle of Civil Law (Trial), China Supreme People's Court (1988) formally expanded the concept of privacy beyond cases of sexual crime and categorized it under the right to reputation, stipulating that any act, written or oral, that exposed to the public another person's "private secrets (yin1si)" and caused some damage to that person's good name must be deemed an infringement of that person's right to reputation. Even before the Supreme People's Court's 1988 interpretation was issued, lower Chinese courts were accepting civil cases of privacy infringement (Gu, 1988).

The word "privacy" seems to be relatively new to the English vocabulary as discussed in Huebert (1997, p.28), "the earliest example given by the Oxford English Dictionary appears in a mid-fifteenth-century text." Despite the fact that privacy in the English language nowadays has been widely accepted and understood as a value to be preserved and upheld, it has not always been considered this way.[5] The etymology of "private" reveals that the word comes from the Latin *privatus*, meaning 'to be deprived' or 'limited' (Baldwin Lind, 2015, p.51-52), which bears a derogatory sense. Moreover, and similar to the discussion in the Chinese language, the sense of negativity of "private" was revealed for its opposing relation with the public. The sense of dispossession from the public that the private space originally conveyed meant "withdrawing from the public body or restricted to one person or a few persons as opposed to the wider community; largely in opposition to public" (Baldwin Lind, 2015, p.51-2).

In addition to the opposing relation between *public* and *private*, it seems that the opposing relation between *private* and *official* (*guan*官 in Chinese) might have also existed in the English language context: "... the early modern public, often opposed to the private, was strongly linked to office-holding … an official persona was almost always a public figure with public responsibilities in a specific sphere. Within this defining context in which the public was understood, the private became the sphere of those who were subordinate or had to obey those exercising office" (Baldwin Lind, 2015, p.58; Condren, 2009).

[5] Historians of ancient Greece and Rome have argued that the concept of privacy is unknown to the ancient world (Berg, 2018).

This brief examination of the characters and words used to represent the concept of privacy in the Chinese and English languages has demonstrated how the concept can be revealed through language. We have seen that privacy being associated with a negative tone exists or have existed in both the Chinese language and English language. The negative sense in yinsi is probably only several decades away, while for the English word privacy, it could be centuries away, as was suggested by Baldwin Lind (2015) that “before 1700, private was essentially a negative term: whatever did not pertain to the nation or community” (p.59). It can be challenging to separate the meaning of language from its social environment, and it is not something that this dissertation aims to do. However, the focus on language looks promising and inspiring.

A question that has been under-explored in privacy related research is whether and how the expression of privacy in these two languages could come back to influence the understanding of privacy. This question aligns with the perspective of Whorf (1956) in that patterns of thought are under the influence of language. Though in translation practice, it is acceptable to equate privacy with yin3si. In intercultural discussion, the goal is to examine the ways privacy does not equal yin3si1; while doing so, also re-open the complexity of privacy to reveal how the concept can be colored with different cultural connotations.

2.3 Computational methods to work with language for conceptual understanding

This dissertation is a study of privacy by analyzing two languages in two genres, and I propose to work with the natural language corpora by using computational methods. There are many ways to work with natural language corpora. One example is qualitative content analysis, which is frequently used in fields such as sociology (Kozlowski, et al., 2019). Content analysis is a "systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding and categorizing" (Stemler, 2001, p.1). It relies on researchers' knowledge and expertise and their manual efforts of reading through the entire content. For example, researchers could read through interview transcripts and identify themes from the content. These themes could come from a qualitative coding process (Saldaña, 2016), and themes are usually represented as concrete words, phrases, or codes. However, content analysis has been criticized for its inability to cover a large amount of data and its overall generalizability. Although content analysis studies can be evaluated using inter-coder reliability[6], it has been questioned for relying on a *priori* categorizations by which the text under study can be put into categories of meaning or described statistically (Rice & Danowski, 1993, p.373).

In comparison, computational methods can process a larger scale of data that far exceeds human capacity, and can do so without presupposing any categorizations, unlike qualitative content analysis. In addition, another reason that social science researchers choose to use computational methods is that, computational methods such as topic modeling, have the capacity to capture *polysemy* (DiMaggio et al., 2013, p.587) and disambiguate different uses of a word based on its context. Some of the popular computational techniques include semantic network analysis (SNA) (Jiang et al., 2017; Doerfel & Connaughton, 2009; Doerfel, 1998) and topic modeling (Blei et al., 2003).

[6] Intercoder reliability aims to test the reliability of content analysis results; in actual operation, reliability can be measured and assessed using different statistics, including Holsti's method, Cohen's kappa, Scott's pi and Krippendorff's alpha each with pros and cons (Mouter & Vonk Noordegraaf, 2012).

2.3.1 Topic modeling

Topic modeling (Blei et al., 2003) is a computational content analysis technique that has been used to help researchers to gain a *thematic* understanding of the topics or themes of a corpus without having to read through each of the documents manually. Topic modeling draws on the notion of distributional semantics (Turney & Pantel, 2010) and makes use of the so-called *bag-of-words* assumption. That is, the meaning of a word comes from the context of the word (co-occurrence with other words), and the ordering of words within each document can be ignored. It is sufficient to describe the distribution of words in order to grasp the themes of a document (Grimmer & Stewart, 2013). Since topic modeling uses the *bag-of-words* approach, it essentially treats each document as a vector of word counts; each document is represented as a probability distribution over some topics, while each topic is represented as a probability distribution over several words.

The work of topic modeling relies on a few key assumptions about the corpus (Feng, 2019a, p.29). First, each document has multiple topics; second, the number of topics of the corpus is fixed; third, each document is assumed to be generated by a known process; and fourth, words are generated independently of other words (i.e. the *bag-of-words* assumption). These assumptions are why topic modeling, in general, may be criticized. For example, the *bag-of-words* assumption has been criticized because semantic information can get lost with the discarding of word order information (Lenci, 2008, p.21). Therefore, researchers have developed more customized topic models that either consider word order (Wallach, 2006), or the relations between topics (Blei & Lafferty, 2007), or the structure of topics (Blei et al., 2010).

Latent Dirichlet allocation (LDA) is an example of topic modeling (Blei, 2012). LDA starts with random initialization, hence, even when using the same text, coding, and preprocessing, generating exactly the same topic model would not be possible because it will still receive impact from “the software libraries used and their versions”, and the “random seed that may not be known” (Hecking & Leydesdorff, 2018, p.4). Hecking & Leydesdorff (2018) also discussed how results of LDA are irreproducible and suggested that one can use Gibbs sampling with a fixed seed to resolve the problem of the random initialization; however, updates of the hard and software used remain as factors that contribute to irreproducible results (p.265). There is continued research on dealing with the instability caused by the random initialization of topic models (Qiang et al., 2018; Sokolov & Bogolubsky, 2015). However, for customized topic modeling like Structural Topic Modeling (STM), this random initialization is no longer a

problem because the prior parameter is replaced by the spectral initialization so that randomness is removed (Roberts et al., 2014; Roberts et al., 2016).

There are three challenges for using topic modeling, some are posed by the characteristics of the natural language corpus, some are inherent characteristics of topic modeling. First, topic modeling may not work well with short text like Tweets (Cheng et al., 2014; Guo et al., 2016; Tang et al., 2014; Yan et al., 2013). In the case of working with a short text, researchers have to further augment the text by either aggregating the short text in some way; for example, to group the same Twitter handle into one bigger document (Hong & Davison, 2010); or to combine content generated in a certain time into one document (Zhao et al., 2011). Zhao et al., (2018) proposed another way to compensate for the short texts like Tweets, which is to make use of bi-gram instead of uni-gram for analysis.

The second challenge for working with topic modeling is the selection of the *hyperparameter K*, which depends on the step-by-step practical implementation of topic modeling. Topic modeling relies on two matrices to define the latent topical structure: the word-topic assignment matrix ϕ, and the document-topic assignment matrix θ; the computational core challenge is to estimate these two matrices (Maier et al., 2018). In Bayesian statistics, theoretically reasonable distributions are randomly assigned to the two unknown variables ϕ and θ. These distributions are called *prior distributions*, as they are assigned *before* data analysis. LDA uses probability distributions from the Dirichlet family of distributions, which is a continuous multivariate probability distribution frequently used in Bayesian statistics. Each of the two Dirichlet priors is governed by the *hyperparameter K* (equal for ϕ and θ) which indicates the number of topics.

There are different ways to decide on a value for *K*. First, the value of *K* can be dictated by the researcher based on domain knowledge (Grimmer & Stewart, 2013). Researchers can compare their manual coding results of the sample data with the results of topic modeling to see whether the model has meaningfully discriminated between topics. Second, the value of *K* can be selected based on the performance of the model in terms of perplexity, or coherence of topics. Third, researchers can also make conjectures of *K* based on results from other content analysis techniques like SNA and factor analysis (Leydesdorff & Nerghes, 2016). Also, experience running topic modeling might also be able to provide a hint in terms of deciding on the *hyperparameter*. For instance, for shorter and more focused corpora (i.e., those ranging from a few hundred to a few thousand documents in size), an initial choice between 5 and 50 topics is best; whereas for larger, and unfocused corpora (i.e., those ranging from tens of thousands to

hundreds of thousands of documents in size or larger), previous research has found that between 60 and 100 topics are the best (Roberts et al., 2018).

The third challenge for working with topic modeling is the *interpretation* of topics. The interpretations of topic modeling results depend on the purpose of research, and the assumptions of specific research questions. The meaning of “topic” in topic modeling may appear obvious, where a topic is understood as "what is being talked/written about" (Günther & Domahidi, 2017, p. 3057). Similarly, topic in topic modeling takes on an intuitive and quite "abstract notion" (Blei et al., 2003, p. 995), hence the meaning of a topic in topic modeling still must be interpreted substantially, “... what exactly topics represent, and if they represent different concepts given different input parameters in the model, is ultimately an empirical question” (Jacobi et al., 2016, p. 91). In other words, though substantive interpretability (meaning to provide an explanation of the meaning of topics) is not required by topic modeling, it is crucial for social and cultural research (DiMaggio et al., 2013, p.578). Many studies rely on the results from topic modeling as an indication of topics for the corpora analyzed. Specifically, interpretation of topics can be done by providing a summarization of the topics using a succinct phrase, namely “topic labeling” (Boyd-Graber et al., 2017, p.40). There are times that the human readers find topic modeling results hard to understand or not intuitive (Leydesdorff & Nerghes, 2015). However, even when topics are difficult for humans to interpret, they can still be very useful for predicting purposes (Resnik et al., 2013; Yun & Geum, 2020).

Operationally, to apply topic modeling properly to textual data, researchers need to take care of at least four major steps (Maier et al., 2017): first, the pre-processing of the corpora; second, the selection of model hyperparameters *K*; third, model evaluation; and fourth, interpreting the topic modeling results. The preprocessing of the corpus can impact the evaluation of the number of topics *K*, and the selection of the number of topics *K* will directly influence how the performance of the as well as how topics can be interpreted.

Preprocessing can vary by study. However, some commonly used steps include tokenization, lowercasing, punctuation and special characters removal, stop-words removal, stemming and/or lemmatization, and further filtering or pruning to strip words that are extremely rare or frequent. At the current stage, how the preprocessing impacts the research using topic modeling in terms of reliability, interpretability, and validity is largely an unexplored area (Maier et al., 2017). Research has found that the ratio between the document length and the vocabulary size of the document impacts the performance of

topic modeling, and studies have suggested that the model performs the best when the document lengths are at least four times the vocabulary size (Feng, 2019a).

There are, in general, two ways of evaluating the performance of the topic model on which the selection of the *hyperparameter K* depends. The first is *internal* or *intrinsic* measures, the second is *external* or *predictive* measures. Intrinsic measures perform the validation by relying on statistical measures like the *coherence* score and the *perplexity* score (Blei et al., 2003; Wallach et al., 2009; DiMaggio et al., 2013). More specifically, the *coherence* score indicates how frequently the top words of a topic co-occur. Mimno et al. (2011) provide a more detailed discussion on the calculation of semantic *coherence* of a given topic. Stevens et al. (2012) provide further evidence that the coherence score is adequate to compare the outcome of different topic modeling approaches.

The diagnosis of the hyperparameter *K* can be done by a joint consideration of the *coherence score* and *exclusivity* score of topics. *Exclusivity* is a measure of the probability for a word to fall primarily within the top rankings of a single topic. Those topic models that perform the best on both the *coherence* score and the *exclusivity* scores will be chosen. When plotted visually, the best-performing models will come from those that are placed near the upper right-hand quadrant of the coherence-exclusivity plot (see Figure 4.1 for a sample *coherence-exclusivity score* plot).

The second way of evaluating the performance of topic models is called *predictive*, or *external* (Grimmer & Stewart, 2013). The idea is to cross-check the results from topic modeling with the results from another source, such as human interpretation or modeling results from a different technique. For example, the topic modeling results could be compared with semantic network analysis (Leydesdorff & Nerghes, 2015), or principal component analysis (Hecking & Leydesdorff, 2018).

Despite the different measures to evaluate topic modeling, the evaluation of topic modeling still needs to balance between statistical diagnostics and human interpretability, because models that perform well by statistical measures may be less interpretable to humans, which is known as the “prediction-interpretability trade-off” (Lindstedt, 2019, p.310). So the evaluation of topic modeling results will depend on each study and its research questions, and will likely result from a trade-off and balance among multiple goals, including validity and interpretability.

Overall, compared to traditional qualitative content analysis, it is this capacity to reveal *latent* semantic relations between words that make topic modeling distinctive among computational linguistic methods. Other strengths of topic modeling include its ability to capture *polysemy* (DiMaggio et al., 2013,

p.587) and to disambiguate different uses of a word based on its context, in addition to processing large scales of data that is beyond manual capacity (Bohr & Dunlap, 2018; DiMaggio et al., 2013; Levy & Franklin, 2014; McFarland et al., 2013). In research applications, topic modeling has been used to understand a corpus over years (Lindstedt, 2019); or, to explore the temporal changes of a social issue or event (Hall et al., 2008; Jacobi et al., 2016; Lindstedt, 2019; Robinson, 2019).

2.3.2 Semantic Network Analysis

Semantic Network Analysis (SNA) is “a form of content analysis that identifies the network of associations between concepts expressed in a text” (Jiang et al., 2017, p.16; see also Doerfel, 1998; Carley & Palmquist, 1992). SNA has its origin in cognitive science and linguistics that argue human memory retains meaning via a system structure (Collins & Quillian, 1972), and relies on the structure of language (Doerfel & Connaughton, 2009) as its primary source of insight. SNA differs from more traditional content analysis common in social research, in that although it still relies on the calculation of co-occurrences of words, its structural perspective indicates that its emphasis is shifting towards a paradigmatic one where more emphasis is put on the structure and relationships that remain latent until revealed by the analysis. The meaning extracted from texts is not merely based on the presence of certain words or concepts, “... meaning is revealed by the relationships (networks) among the concepts” (Doerfel, 1998, p.17).

Similar to a social network, a semantic network can be presented visually through a graph consisting of two types of functional units: nodes, and edges. Typically, a node represents an n-gram, while an edge is a connection between n-grams in terms of their co-occurrence, which is usually visually represented by a line. A semantic network can be characterized by its components, including nodes, edges, and paths that connect nodes. See Drieger (2013) for a detailed description of a variety of measures for understanding nodes, edges, and paths.

Nodes can appear at the center or peripheral areas of the graph with various connections to other nodes. To describe the importance of nodes within the graph, *centrality* is commonly used as a quantitative measure. *Centrality* describes how nodes are “systematically linked to other nodes in a system” (Drieger, 2013, p.203). There are multiple types of centrality, including *degree centrality* and *betweenness centrality*. *Degree centrality* measures the number of nodes one node is connected with (i.e., popularity); and *betweenness centrality* measures the *extent* to which one node is connected to other groups of nodes (i.e., connectivity) (Drieger, 2013).

Moreover, the *joint consideration* of the two centralities can be used to construct a *structural space* of the corpus (See *Figure 2.1* for a sample of the structural space plot), where nodes within a network can be mapped onto four different areas in a quadrant, and each node acting a distinctive *structural role* depending on where they are mapped onto the network (Nerghes, 2016; Shim et al., 2015): “globally central”, “gatekeeper”, “locally central”, and “marginal”. *Globally central* nodes refer to nodes that

are high on both centrality scores, which indicates that they are closely related to the entire network (Shim et al., 2015). *Locally central* nodes, because they do not have as high betweenness centrality; hence, they can be understood as related to nodes in their own neighborhood. *Gatekeeper* nodes can be understood as indicating the possibility of intersection of different sub-networks. *Marginal* nodes are neither popular, nor connective, and hence can be considered the least important among the four roles. But they may have the potential to turn into the other three roles.

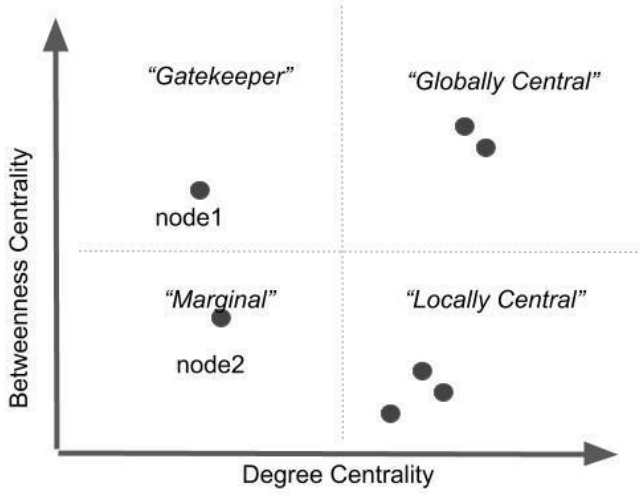


Figure 2.1 Four structural roles of nodes in a semantic Network.

Like topic modeling, building a semantic network typically involves a few steps, and details can vary by each study. SNA starts with the preparation and preprocessing (which include typical text pre-processing steps like cleaning, tokenization, removing stopwords, and filtering out extreme low/high-frequency words) of a language corpus, and typically concludes with a network graph visualization with the most important nodes (by centrality).

Overall, SNA offers an alternative perspective for understanding the text data that is not possible by reading text. SNA is not only able to process significantly larger amounts of data that can be less feasible to read manually but also offer additional perspectives to understand single words *in the context* of groups of words. Moreover, similar to arguments about the benefits of using topic modeling, some have argued that SNA could help address some of the shortcomings of traditional content analysis in that it can maintain the "richness of the data" and "multiplicity of meanings" while not having to reduce original content "to a few categories" (Doerfel, 1998, p.21). In other words, SNA can be considered more advantageous than traditional content analysis because the method does not employ *a priori categories*

based on theory, which might suppress unexpected emergent meanings (Rice & Danowski, 1993; Doerfel, 1998, p.23). However, it is important to note that the generation of a network is only the beginning of analysis, like topic modeling, researchers must further interpret the network analysis results.

Like topic modeling, SNA has been used to provide insights when the purpose of the research is to identify themes/frames when working with larger collections of documents in social sciences (Choi & Lecy, 2012; Jiang et al., 2018; Kwon et al., 2016; Smith & Parrott, 2012). Additionally, SNA has been used in conjunction[7] with other methods to enhance research exploration. In particular, researchers have used topic modeling and SNA together; findings from one method can be used to verify findings from the other (Leydesdorff & Nerghes, 2016).

[7] Other work where SNA has been used in conjunction with another computational methods include, for example, SNA and spatial modeling (Kwon et al., 2009), SNA and survey (Kim & Kim, 2015), SNA and thematic analysis (Veltri & Atanasova, 2017; Xiong et al.,2019), and SNA and social network analysis (Basov et al., 2017; Shen & Eliassi-Rad, 2006).

CHAPTER 3: RESEARCH DESIGN & QUESTIONS

“By way of analogy, if we reflect on our best efforts to read and teach classical Greek philosophy, most of us do not have an expert knowledge of the original language texts. But in developing a sophisticated understanding of an extended cluster of the most important Greek philosophical terms—logos, nomos, nous, phusis, kosmos, eidos, psyche, soma, arche, alethea, and so on—we are with imagination, able to get behind our own uncritical Cartesian assumptions and at least in degree, read these Greek texts on their own terms.” (Ames, 2020)

This dissertation proposes a study of privacy via language. Using language to study privacy is to focus on privacy as a concept itself, while temporarily suspending all the operational contexts in which privacy resides. Unlike other studies that aim at understanding how privacy may be impacted by a range of contextual factors, this study focuses on understanding how privacy is *expressed through language*, or the *conceptualization* of privacy. The goal of researching privacy using natural language is to reveal the complexities of privacy as an abstract concept.

In existing research of privacy, the internal complexity of privacy itself as a concept is suspended deliberately where a more simplified understanding of privacy is adopted for operations (Quinn et al., 2019). For instance, privacy equates to concrete decisions about whether and how to share one's information (Acquisti et al., 2015); or, privacy equates to the provision of related information so that people can make informed privacy decisions (Schaub et al., 2017). In other words, these above examples are where the research of privacy tends to attend more to the context of privacy. The focus is not quite on deciphering the meaning of privacy itself as a concept, but on how as a practical issue, privacy can be implemented or realized. This is where the current study differentiates itself from existing work.

Legal scholars and philosophers have contributed many discussions on the conceptualizations of privacy (Ess, 2019; Bannerman, 2018; Nissenbaum, 2004; Solove, 2002; Taylor et al., 2017). In contrast to the approach legal scholars and philosophers took, this research studies the conceptualization of privacy through the evidence of language. Working with language to understand privacy differentiates itself from the theoretical works in that language helps reveal how privacy as a concept is actually being enacted in everyday situations, rather than pointing out theoretical possibilities for conceptualizing privacy.

In addition, studying privacy via language is a potential way to enable the comparison of the conceptualization of privacy across cultures. Shared languages, if discovered, provide a foundation to approach the question of what is the "minimal conception" (Mizutani et al., 2004) of privacy across cultures. Specifically, through language analysis, this study strives to reveal *topics* that are typically invoked when privacy is used in that specific language. In other words, this study could tell when privacy is talked about in different languages, if it is expressed via similar topics. In addition to looking at topics, language provides various granular levels of observation that can be as small as single nodes/words, or as big as groups of topics.

3.1 Research Design

This study uses a corpus consisting of two genres in two languages across 10 years. The entire analysis space enabled by the design of this research project is depicted in *Figure 3.1.* The three coordinate axes represent language (Chinese and English), computational technique (SNA and STM), and genre (News articles and Social media posts). STM and SNA can supplement each other in that the former provides an observation at the topic level, while the latter provides an observation at the single node/word level. In addition, the results from STM and SNA can corroborate with each other as we will see soon.

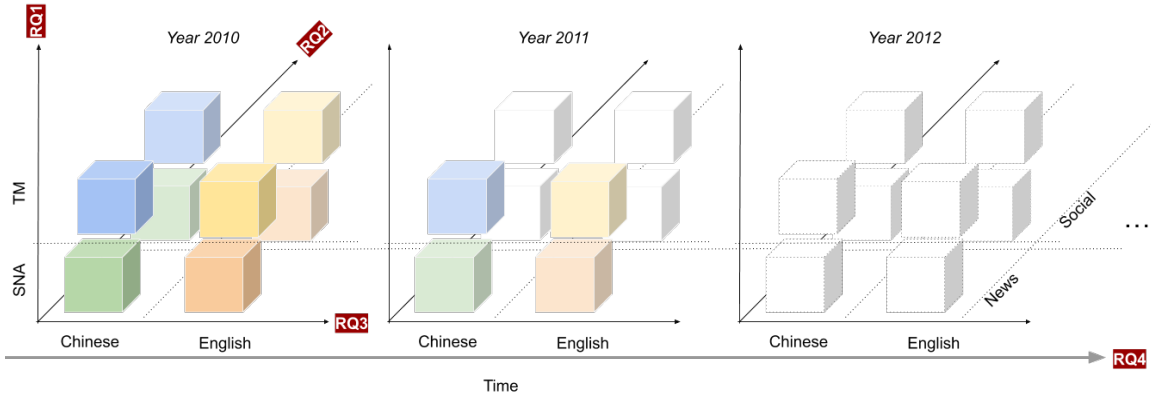


Figure 3.1. A visualization of the comparative analysis space

Next, I describe the major concepts used in this project, including *Semantic Object of Analysis (SOA), Semantic Feature*, and more. These concepts serve as analysis proxies for understanding privacy. In other words, I mainly engage with conceptual constructs to derive understandings of privacy. And I will only refer to the original natural language text as a potential double-check/validation which will be described more in Chapter 4. Method & Data.

Semantic Object of Analysis (SOA)

SOA refers to a section of the corpus of one year, one language, and one genre that serves as a basic unit of observation; for example, the 2011 English News corpus. Based on SOAs, observations can be made at multiple levels of granularity, for instance, across multiple SOAs, or at the single document or word level within one SOA.

Semantic Measure and Semantic Relatedness

Semantic measure refers to quantifiables that can be used to describe semantic objects of analysis as well as more or less granular units of language: words/nodes can be measured by their centrality scores; topics can be measured by their topic proportion. Semantic relatedness describes semantic objects of analysis when compared to each other, the comparison can be across genre, across language, and across time. If one semantic measure of two semantic objects increases as observed by a covariate, I may be able to suggest that these two objects of analysis are positively related to each other on the covariate. For example: for a topic *T*, if over time *T*'s prevalence increases in one language but decreases in another, then I may have an observation that *T*'s prevalence is inversely related by these two languages. Semantic measure and semantic relatedness are both considered semantic features.

Topics, Sub-themes/-topics, and Core Semantics

Topics refer to the results generated by structural topic modeling (STM). Sometimes, a topic can contain multiple privacy-related topic words that *cannot* be interpreted under one coherent theme, which is a situation I refer to by using sub-topics, or sub-themes. By sub-topics or sub-themes, they refer to the semantics within a topic that refer to a distinctive aspect of privacy that cannot be categorized under a higher-level concept. The *granularity of topics* refers to this phenomenon that topics generated by STM appear to have multiple sub-topics embedded within one topic. The difference between topics and sub-topics is that the latter is more granular.

Semantics is a general term that is used in this study to refer to all the topics and sub-topics that are meaningful for understanding privacy. Topics or sub-topics that are shared across genres in a language, are considered *core semantics*. Topics and sub-topics can overlap/repeat.

Note that the identification of sub-topics derives from the subjective interpretation of STM results by the researcher's interpretation and the external coders' labelings. See Section 4.1.2 Interpretation & Labeling of Topics for details.

Semantic Dimension

Semantic dimension[8] is used to refer to the higher level of categorizations of topics. Based on analysis, I propose four dimensional coding to categorize topics as a result of inductive interpretation, the four dimensional codes are: *technology*, *institution*, *individual*, and *public*. See Section 4.1.4 for more details.

Semantic (in)compatibility

Semantic (in)compatibility is used when topics, sub-topics, and semantic dimensions are compared across the two languages. When similar topics or subtopics are found across the two languages, I would argue for semantic compatibility. If topic categorizations of the two languages show similar patterns (for example, many topics in both the Chinese and the English language can be categorized under the *technology* dimension), then I would also argue for semantic compatibility. On the contrary, if subtopics, topics, or even dimensions in one language are missing in another language, I would consider these instances as semantic incompatibility.

[8] The word “dimension” here needs to be differentiated from the typical understanding of it in the context of vector space-based language models, where a word could be considered as a dimension (Sahlgren, 2006).

3.2 Research Questions

This dissertation aims to address four research questions. Research Question 1 focuses on understanding the corpus by language and by genre. Research Question 2 focuses on comparing the two genres in one language. Research Question 3 focuses on comparing across the two languages. Research Question 4 focuses on understanding the topics over time.

RQ1: What are the most notable semantic features that can be observed in the corpora regarding the concept of privacy?

This question lays the foundation for understanding privacy in these two languages. Relying on Semantic Network Analysis, and Topic Modeling, answers to this question describe what are some of the topics and subtopics in these two languages. STM will reveal what are the topics in these two genres and languages? What are they about? Which ones are leading in topic proportions? SNA will reveal what are the nodes/words with the highest centrality score? What are the *structural roles* of words; in other words, how are words positioned in their *structural space*.

RQ2: What are the trends or patterns of semantic features that can be observed across genres? Or, what are the (core) semantics of privacy?

This question guides the comparison between news and social media corpora; it explores how semantic features are similar or different by genre. The rationale for asking this question is that genre is a significant factor that impacts language expression: certain expressions and semantics of privacy are more likely (if at all) to occur in one genre than the other (see more details in Section 1.2). Specifically, this question will investigate whether certain topics are distinctive and/or more prevalent in one genre than another. The combination of findings from both genres provides a fuller semantic picture of the concept for one language, and those topics or subtopics that are shared across genres will be considered as *core* semantics for that language. For example, if a topic from topic modeling occurs in both Twitter and English News, then, this particular topic will be considered as a *core* semantics for privacy in the English language.

RQ3: What are the trends or patterns the semantic features observed across the two languages? On what dimensions are the two languages (in)compatible?

This question focuses on identifying topics that are shared or distinctive when compared across languages. The assumption is that privacy as an abstract concept, its expression can differ by language. Different languages may have similar or shared topics or subtopics of privacy; in addition, those topics or

subtopics may share similar or have different semantic dimensional coding. Shared topics across languages are considered as cases where the understandings of privacy are *compatible*; whereas those distinctive topics indicate where the understandings of privacy may be *incompatible*. Specifically, observations from the structural roles of words in these two languages can be also used to respond to this question. For example, if certain nodes only show up in one language, this suggests a distinctive understanding of privacy in one language that may be incompatible with the language.

RQ4: What are the trends or patterns semantic features and dimensions observed across time?

This question focuses on identifying patterns or trends across time. The assumption behind this question is that privacy is a dynamic concept and stays in constant change, and such changes can be observed in language. Specifically, such changes, when reflected through language, may be shown as at both the node/word and topic levels. At the node level, the changes may be revealed as the change of centrality score of the same node over time. At the topic level, it could be shown as how certain topics' proportion increases/decreases in time.

CHAPTER 4: METHODS & DATA

"... the world is presented in a kaleidoscopic flux of impressions which has to be organized by our minds--and this means largely by the linguistic systems in our minds." (Whorf, 1956, p. 213)

This chapter describes the two computational methods used in this study, they are: structural topic modeling (STM) and semantic network analysis (SNA), as well as ways used to interpret the results from these two computational methods. STM (Roberts et al., 2013 & 2019) is used to reveal the *topics* of privacy in the corpora, while SNA is used to reveal the most important words/nodes in each language leveraging *structural space* (Nerghes, 2016; Shim et al., 2015) in which words/nodes occur.

STM has been constructed using: first, the corpus in each genre and language; this is to establish a basic understanding of topics in each genre and language (in other words, the stand-alone by-genre and by-language analysis). Then, STM has been constructed using both languages and genres to enable cross-genre and cross-language analysis. For the stand-alone by-genre and by-language analysis, the original Chinese language corpus was used, the results are presented with my own translation of the topic words. While in the cross-genre and cross-language analysis, the translated Chinese corpus was used to build the model, and enable comparison with the English language.

For the structural space analysis, only the news corpus was selected. The social media corpus was not selected because of the short length of the social media posts. Similarly, for the Chinese language corpus, the structural space analysis was done using the translated corpus to enable comparison with the results from the English corpus.

4.1 Structural Topic Modeling

Structural topic modeling was built using the by-language and by-genre corpora first; and then, structural topic modeling was built using the cross-language and cross-genre corpora.

For by-language and by-genre analysis: once topics by each language and genre were identified, I put topics from the two genres in that one language together to create a complete list of topics for that language (which I refer to using the Union symbol). In addition, I put the topics from two genres in that one language together to compare and identify if any of the topics are shared by both genres (which I refer to using the Intersection symbol). Table 4.1 provides a conceptual representation of the joining and intersecting of STM topics from by language and by genre analysis.

	News	Social	Semantics in/across* language	Core semantics in / across* language
CN	A	B	A∪B	A∩B
EN	C	D	C∪D	C∩D
Semantics in News/Posts	A∪C	B∪D	(A∪B) ∩ (C∪D)*	(A∩B) ∩ (C∩D)*
Core Semantics in News/Posts	A∩C	B∩D		

Table 4.1. A conceptual display from topics, core semantics, to cross-language semantics

For cross-language and cross-genre analysis: two different topic modelings were constructed. First, a cross-language analysis of topics using the news corpora of both languages. Second, a cross-language and cross-genre analysis of topics using a sample from both languages and both genres. As we will see soon, topics generated from the by-language and by-genre analysis are the most semantically meaningful, which is what the subsequent interpretations and discussions are based on. For the same reason of generating semantically meaningful topics, a cross-language analysis of topics was done using only news corpora.

4.1.1 Modeling

To determine the hyperparameter *K* for topic modeling, different *K* values between 5 to 31 (incremented by 2, so *K* = 5, *K* = 7, all the way up to *K* = 31) were selected to run the estimation. This range appears to be reasonable considering the size of this corpus (Blei, 2012). Among models with different *K* values, the best performing model by the *coherence-exclusivity score* is selected (see Figure 4.1 for identifying *K* by the *coherence-exclusivity* score). In addition to the identification of topics, correlation analysis was done to understand the trends or patterns of topics as mediated by: language (CN, EN), genre (News, Social), and time (2010, 2011, … , 2019).

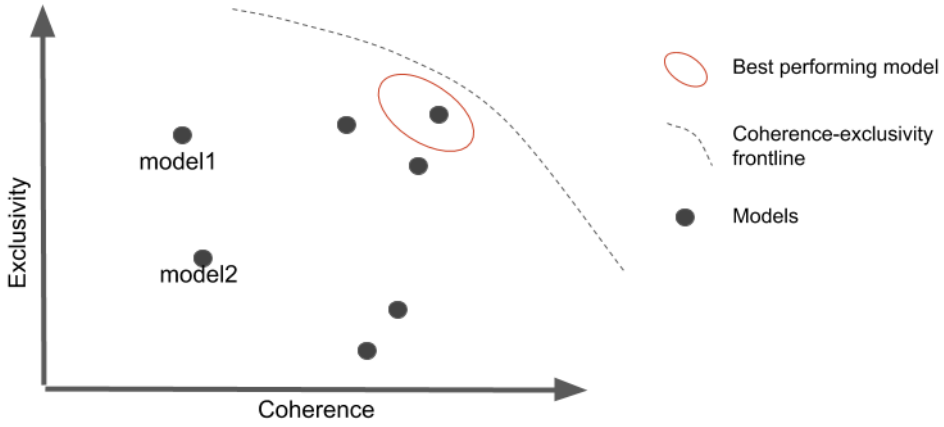


Figure 4.1 A sample plot of the joint measure of coherence and exclusivity

4.1.2 Interpretation & Labeling of Topics

The interpretation of topics relies first on a reading of the topic words returned from STM. Specifically, the interpretation of topics relied on topic words that have the highest probability score and the highest FREX score. Topic words with the highest FREX scores are those topic words that occur with high frequency in one topic and have high exclusivity to the topic (see Section 2.3.1). The decision to use both types of topic words is based on the possibility that the two could supplement each other in illustrating what a topic is likely about (probability score), and how this topic can be differentiated (FREX score) from the rest of the topics.

In addition, the researcher's interpretation of topics was supplemented and supported by a reading of the original documents (the original news articles and social media posts) and a further search in Google and Baidu to get the full spelling of acronyms and proper nouns. Each topic's exemplar documents were retrieved from the original corpus. Exemplar documents are those original documents used to produce the topics that have the highest proportion of the topic words of each of the topics. Reading the exemplar documents enables the researcher to see how topic words appear in their original context. Operationally, the exemplar documents can be retrieved using the *findThoughts* function in the *stm* r package (Roberts et al., 2019)[9].

Most of the topics are informative and *semantically meaningful* for understanding privacy. By *semantically meaningful*, it means that a native speaker of the language can understand intuitively how the topic words are related to privacy without additional clues. I identified four scenarios I use to classify each topic's level of *semantic meaningfulness*: the first two scenarios are considered *semantically meaningful*.

Scenario one_Obvious words: topic words that are immediately obvious how and why they could relate to privacy; for example, *GDPR*, *Facebook*.

Scenario two_Somewhat obvious words: words or terms that have a common meaning or use that is related to privacy to native speakers, for example, *artificial intelligence*, *cryptocurrencies*, *real-name registration*.

---

[9] More specifically, the identification of exemplar documents leverages the document-topic assignment matrix θ, which "Can be used to identify the documents that devote the highest or lowest proportion of words to a particular topic. Those with the highest proportion of words are often called "exemplar" documents and can be used to validate that the topic has the meaning the analyst assigns to it." (Roberts et al., 2014, p.6)

Scenario three_Words that require background knowledge to understand: words that are unclear in terms of how they could relate to privacy even for native speakers. Typically, these words require background knowledge on the part of the human interpreter, or require the use of contextual cues (from other topic words, or from additional searches). For example, the word “tissue box” appears in Topic 7 in Weibo K13 results. This word was successfully interpreted as related to privacy by only one of the three native speaker coders. The other coders, including the researcher herself, failed to interpret this word because of the lack of background knowledge or the inability to make use of other topic words (“shampoo”, “electronic clock”) that provide contextual information for understanding this particular word for this topic[10].

Scenario four_Words that are not meaningful: these include words that are too common or over generic to demonstrate a connection to privacy; for example, words like “Peopl” (people) is too generic to understand how it relates to privacy. Or, words that appear as mis-spellings or made-up words on the internet; for example: “xoxoxoxo”.

After reading through the topic words, an *initial label* is given to each of the topics by the researcher, preferably in the form of a short phrase that represents the gist of this topic. In the case where a topic contains multiple themes that cannot be summarized succinctly into one phrase, multiple representative keywords are used instead of a one-phrase label. This served as an *initial labeling* of the topics. The *finalized labeling* of topics incorporated both the researcher’s and the external coders’ interpretation of topics which is explained next.

One example of *initial labeling* is from CN K13 Topic 10 (with the researcher’s translation into English). This topic was labeled as “*Mobile phone”*, because the topic words point out several mobile phone related issues, each with a slightly different focus, but converge on the theme of “*Mobile phone”.*

- Highest Prob: 智能手机, 手机用户, 个人隐私, 通讯录, 运营商, 应用程序, 二维码
- Highest Prob: smartphone, mobile user, individual privacy, contact list, internet service provider, application, QR code
- FREX: 智能手机, 手机软件, 恶意软件, 恶意程序, 数据恢复, 下载安装, 二手手机
- smartphone, mobile software, malware, malicious program, data recovery, download and install, second-hand mobile phone

[10] There are news articles reporting how hidden cameras were installed in tissue boxes at hotel rooms (*Youth.cn,* 2019, January 25).

Another example is Topic 1 in CN K13. This topic was initially labeled “Privacy at work context”, which is indicated by topic words like “employee”, “mobile number”, “person in charge”, and “individual privacy”.The rest of the topic's words are examples of *Scenario four*, hence were ignored in the initial labeling.

- Highest Prob: 工作人员, 手机号, 手机号码, 电话号码, 负责人, 个人隐私, 李女士
- FREX: employee, mobile number, mobile number, person in charge, individual privacy, Ms. Li
- Highest Prob: 李女士, 航天员, 李亚鹏, 任志强, 丁嘉丽, 许先生, 王小姐
- FREX: Ms. Li, astronaut, Yapeng Li, Zhiqiang Ren, Jiali Ding, Mr. Xu, Ms. Wang

To summarize, the interpretation and labeling of topics is a process of understanding how the topic words may be understood as being *related to privacy* by summarizing them into a representative phrase when possible. The interpretation is an inherently subjective process that relies heavily on the researcher’s knowledge and judgment. To compensate for the potential bias in the researcher’s interpretation, this study also involved additional coders to help understand and label the topics.

4.1.3 Validation: external coders' interpreting and labeling

In addition to the researcher's own interpretation of topics, this study also leveraged additional coders' interpretations as a way to compensate for the subjectiveness and potential bias of the researcher's own understanding. The involvement of external coders serves two purposes. First, the external coders' interpretations *supplement* the researcher's understanding in case the researcher missed or misunderstood certain words of the topics. Second, the external coders' interpretation can help *validate* the researcher's understanding when the researcher and coders reach similar interpretations of the topics. See APPENDIX G for a detailed description of the recruitment of additional coders and the procedures of coders interpretation of topics.

After the researcher had finished summarizing the topics, the topics were sent to three additional coders who are native speakers of that language who also have a basic understanding of topic modeling. After collecting topic labeling results from the coders, a two-step process was used to integrate the topic labelings provided by the external coders with those assigned by the researcher.

Step one focused on comparing the interpretation of topics among the three coders. The purpose of this step was to see if the three coders achieved a convergence, or disagreed with each other on the interpretation of the topics. Convergence of interpretation was defined in two ways. First, leveraging the synonym relationship of words. This is when at least two out of the three coders used the same words or *synonymous* words (for example, "smartphone" and "mobile phone" are considered synonymous) when interpreting a topic. Second, leveraging the hypernym relationship of words. This is when two of the three coders used terms that fall into the narrower-broader term relation. For example, one coder may use "Facebook", the other coder may use "tech companies". Here, "tech companies" is the hypernym of "Facebook".

Step two is to compare the labeling results from the coders' to the researcher's. In situations where the researcher's interpretation did not align with the converged interpretation from the external coders: scenario one is when the converged interpretation from the external coders added something new that may have been missed by the researcher. In this case, the finalized coding would incorporate the converged coding as supplements to the researcher's interpretation. Scenario two is when the converged coding conflicted with the researcher's interpretation. In this case, the converged interpretation would override the researcher's interpretation. See Figure 4.2. for a flow chart of working with external

coders. See APPENDIX I for a complete documentation of all the topics as they fall under either the synonym relationship, the hypernym relationship, or the not semantically meaningful scenario.

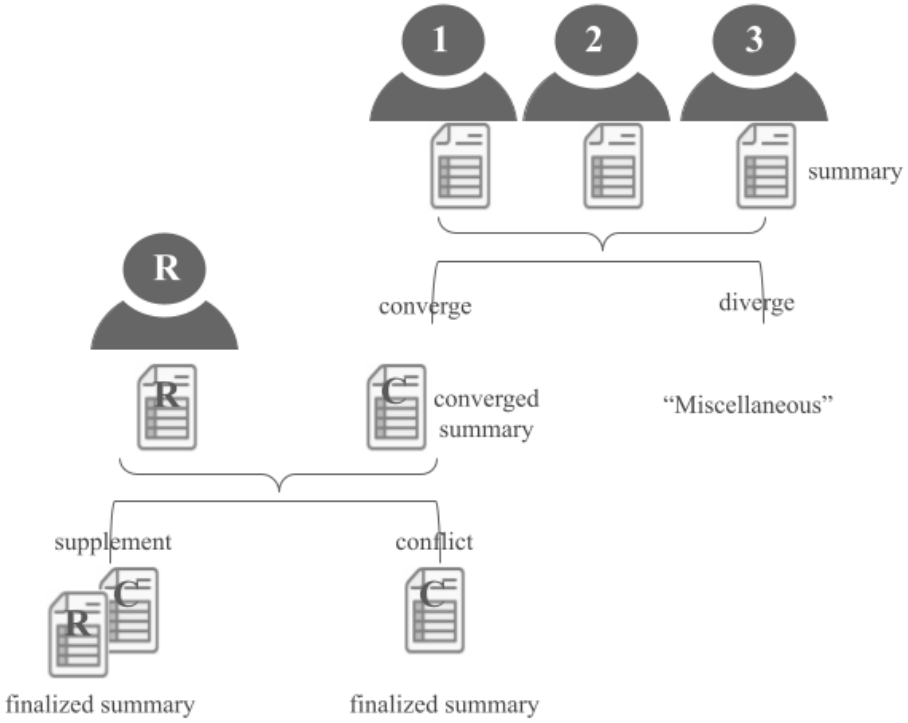


Figure 4.2. A flow chart of interpreting topics by external coders
(numbers 1 2 3 indicate coders, *R* indicates the researcher)

This example shows synonymous topic words from CN K13 Topic 10, with the researcher's translation into English. Here, the two coders used synonymous words (underlined), "smartphone", and "personal phone"; which is also synonymous with the researcher's initial labeling. Hence, the finaling labeling kept the researcher's initial labeling.

- Highest Prob: 智能手机, 手机用户, 个人隐私, 通讯录, 运营商, 应用程序, 二维码
- Highest Prob: smartphone, mobile user, individual privacy, contact list, internet service provider, application, QR code
- FREX: 智能手机, 手机软件, 恶意软件, 恶意程序, 数据恢复, 下载安装, 二手手机
- smartphone, mobile software, malware, malicious program, data recovery, download and install, second-hand mobile phone

- Summary 1: 恶意软件偷窃用户隐私 (malware spying on user privacy)
- Summary 2: 智能手机 (smartphone)，软件安全 (software safety)，个人隐私 (individual privacy)
- Summary 3: 个人手机用户数据 (personal phone user data)
- Initial labeling: Mobile phone
- Finalized labeling: Mobile phone

An example of the hypernym relationship between topic words comes from cross-genre cross-language K11 Topic 3. In Summary 1 (see below), the coder used “Cambridge Analytica”, whereas in Summary 2 the coder used “tech companies”, which is a hypernym of “Cambridge Analytica”. “tech companies” is also a hypernym of the initial label assigned by the researcher. Hence, the finalized labeling used the hypernym term.

- Highest Prob: user, googl, data, facebook, said, privaci, compani
- FREX: patent, analytica, abstract, stoddart[11], inventor, cambridg, trademark

- Summary 1: Cambridge Analytica leaked user information
- Summary 2: tech companies, data analysis
- Summary 3: personal internet footprint
- Initial labeling: Facebook and Google
- Finalized labeling: Technology companies

In situations where there was no convergence among the three coders, I labeled the topics as “Miscellaneous”; these are more often observed in social media topics rather than news topics.

For example, Topic 7 from Twitter K15:

- Highest Prob: privaci, protect, data, secur, free, american, real
- FREX: mellon, encyrypt, cisa, stealthcoin, idltweet, barbi, newpanda, stitm

- Summary 1: protection of financial communications
- Summary 2: Cybersecruity, cryptocurrency, and privacy
- Summary 3: blockchain, security

Initial labeling: Data security, cryptocurrency, data service
Finalized labeling: Miscellaneous

[11] Stoddart is a British-American chemist who is Board of Trustees Professor of Chemistry and head of the Stoddart Mechanostereochemistry Group.

4.1.4 Identification of the dimensions

Labeled topics are further categorized under the four semantic dimensions, they are: *technology*, *institution*, *individual*, and *public*. A further dimensional coding of topics was included to better compare the topics across these two languages. In that, whether a topic exists in one language matters for understanding privacy. Moreover, how much of certain topics exist in one language matters equally for understanding privacy. This dimensional coding, by grouping topics into four broad categorizations, could tell how many certain topics exist in one language and if certain topics exist more or less in one language compared to the other.

These four dimensions are proposed as a result of an inductive process of understanding all the topics. After interpreting all the topics, I have learnt that multiple topics appear to touch on digital technologies, or privacy related regulations. In other words, the same or similar topics are recurring across genres and even languages. Hence, I proposed these four codes to summarize at a high level some of the most commonly seen topics. Once the four dimensions are proposed, I go back to tag each of the topics from each of the language and genres by using either one or multiple dimensions out of the four. This inductive tagging process helps reveal if topics coming from one particular language or genre have more tags from any particular of the dimensions. Below I explain each of the four dimensions, and these four dimensions are visually presented in Figure 4.3.

Technology: This dimension is used to refer to various technologies and their applications, for example, social network site Facebook, search engine Google, various cryptocurrencies, drones, facial recognition, etc.

Institution: this dimension is used to refer to both tangible organizations like departments, companies, associations (Federal Trade Commission (FTC), Ministry of Industry and Information Technology (MIIT)). In addition, the institution dimension also refers to policies and laws like the General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), etc.

Individual: this dimension is used to refer to things that concern the person as individuals who traverse through various social situations and contexts concerning privacy. For example, individuals in their role as practitioners and users of digital applications and technologies can be reflected by topic words like “user”.

Public: this dimension is used to refer to the groups or communities that extend beyond the scope of the individual, including groups of populations, like “patients”, “students”, or the “state/country”.

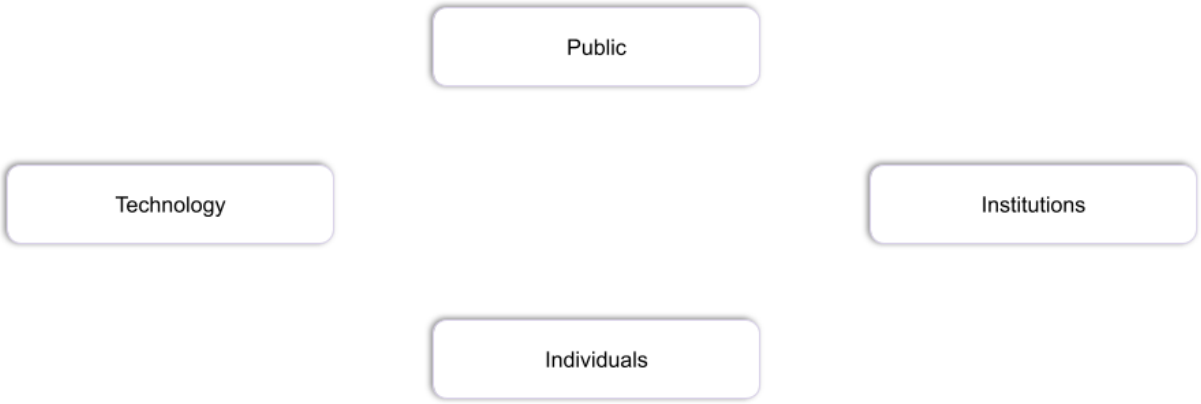


Figure 4.3 Four dimensions for topics categorization

4.2 Semantic Network Analysis

Operationally, *Wordij*[12] (Danowski, 2013) and *Gephi*[13] (Bastian et al., 2009) were used for the analysis of the structural space. The pre-processed corpus was the input file to Wordij to produce a .net file, which was then used to calculate the centrality scores in *Gephi*.

In the analysis using Wordij, word tokens that appeared fewer than 3 times were dropped. The window size was set to be the length of each of the preprocessed documents. Finally, both the degree and betweenness centrality scores were normalized to enable visualization of the structural role space, by dividing each token's score by the largest score of that corpus.

The structural space was plotted using the Python *matplotlib* library to represent the four distinctive structural role quadrants. Specifically, the top 50 nodes from each of the years' centrality lists were selected to produce this quadrant analysis. The quadrant mark was set as the mean score of the top 50 nodes' degree and betweenness centrality scores, respectively.

---

[12] Wordij Semantic Network Tools. (n.d.). Retrieved January 19, 2022, from https://www.wordij.net/index.html

[13] The *Open Graph Viz Platform*. Gephi. (n.d.). Retrieved January 19, 2022, from https://gephi.org/

4.3 Data

4.3.1 Collection

The keywords “privacy”(for English), and “隐私” for Mandarin Chinese, were used to retrieve the natural language corpora in the two languages and two genres. No additional keywords (related terms like data, information, etc) are included for two reasons. First, to make sure (to the extent possible) that semantic content included is indeed *about* privacy; and second, to avoid any potential bias due to the inclusion of other words. Some of the retrieved content may likely be only *tangentially* about privacy. However, the possibility of stray items should be of a limited impact considering the total size of data collection.

For news articles, the search was done using the keyword in the article *title* rather than in full text, to help improve the likelihood that articles discussing privacy are indeed *thematically* about privacy. For social media corpora, the search was done using “privacy” as a keyword in Twitter and Weibo posts.

Time range is another major criterion for data collection. A ten-year corpus, from January-1-2010 to December-31-2019, was gathered for multiple reasons. The first reason to include a multi-year corpus is that the researcher is interested in exploring temporal trends and patterns with the conceptualization and understanding of the concept of privacy. In addition, this particular ten-year period can provide adequate amounts of digital corpora for both genres and the two different social media sites, considering that Twitter was founded in 2006, and Weibo in 2009.

4.3.2 Data sources and their corresponding geographies

The English news corpora were searched and downloaded from Nexis Uni, an academic research database that contains news, business, and legal sources. The Chinese news was purchased from a data vendor Wisers (慧科讯业)[14]. Twitter posts were acquired via the Twitter Academic API, which is a new API service launched by Twitter in January 2021 that enables free access to the full history of public Twitter via the full-archive search endpoint. Weibo posts were acquired via a crawler.

All Mandarin Chinese news articles were from news sources in mainland China. Weibo is mainly used in mainland China. Therefore, it can be assumed that the Chinese corpora, in general, represents not just a simplified Mandarin Chinese language corpora, but it also corresponds to language communication in the mainland China geography (e.g., excluding special administrative regions like Hong

[14] https://www.wisers.com/about.html

Kong and Macau). Notably, the Chinese news articles were published by sources that are under the supervision of either central or local Chinese governments (see APPENDIX D.1 for a list of the top 30 sources from which I obtained the articles, and the number of the articles included from each, as well as a chart of the number of articles retrieved for each year). For example, Reference News (参考消息) is sponsored by Xinhua News Agency (a state-run news agency in China). The Southern Metropolis Daily (or, Nanfang Metropolis Daily, 南方都市报) which mainly covers cities in Southern China with a primary focus on Guangzhou and Shenzhen, is supervised by the Guangdong Provincial government. It is recognized that the governmental affiliation of news agencies, as well as the fact that media are under government censorship in China (Kuang, 2018), could have an impact on news content. However, this is beyond the scope of the current study to tease out.

For the English corpora, Nexis Uni supports a filter by publication geography (see the search interface of Nexis Uni in APPENDIX K), which was used when filtering for a collection of American English news articles that originated in the United States. For Twitter, the data collection was filtered by the location parameter (*place_country: US*). Therefore, tweets collected correspond to the tweets and Twitter users geographically based in the United States. The English corpora in total corresponds to language communication in the geography of the United States, which I assume is primarily American English.

The raw data were preprocessed using the Python *nltk* package (Bird et al, 2009). The preprocessed corpora were prepared following the same data frame template; namely, a data frame that contains five columns, including Number, Text, Year (2010/2011/… 2019), Genre (Social/ News), and Language (CN/EN). See below Table 4.2.

Two additional steps were performed for the preprocessing of the Mandarin Chinese corpora. The first step was the segmentation of the Chinese characters, which was performed using a package in Python called *Jieba*, an Open Source Chinese segmentation application. The second step was to transform all potentially traditional Chinese characters to simplified Chinese characters. This is done using a Python package *hanziconv*, an open-source tool that converts between simplified and traditional Chinese characters. To identify stopwords, the Chinese stopword list from the *Stopwords ISO* (a comprehensive collection of stopwords for multiple languages) was used. The English stopword list in *nltk* was used for the English corpora.

Text	Year	Genre	Lan
google paying record million fine settle …	2012	News	EN
市民 状告 苹果 ...	2015	Social	CN

Table 4.2 A framework of data preparation

Each news article in itself can be considered naturally a document for later analysis. However, for social media content, considering the short length of each post, Tweets and Weibo posts were each further aggregated to form a longer document. More specifically, 30 social media posts from the same year were randomly grouped to form a bigger document.

The processed Chinese corpora (both News and Weibo) were translated into English using Google Translation API (see APPENDIX M). The translation process added back some punctuation, stopwords, and special characters (See APPENDIX A.2 for a sample of originally processed corpus and their translated version). Therefore, the translated text went through a second round of preprocessing. In addition, words less than four letters were filtered out before pre-processing to clean up the text. Because the translation process had introduced back into the corpus garbled text like “'” which is ASCII code for the single quote apostrophe[15].

[15] https://www.ascii-code.com/

I use the type-token ratio (TTR) (McArthur et al., 2018) as an indicator to illustrate the vocabulary diversity of the corpus. Specifically, the TTR is calculated for each of the documents in each genre and language, and then the average TTR is calculated by adding up all the TTRs for each document divided by the number of documents. Table 4.3 shows the average TTR for each corpus. The average TTR of Twitter is higher than that for English news; and the average TTR of Weibo is very close to that of Chinese news. The overall number of unique types in social media is much higher than unique types in news in both languages. Moreover, the English language average TTR is higher than the Chinese language. Overall, these corpus statistics suggest that the vocabulary diversity of social media is higher than that of news; and the diversity of English is higher than that of Chinese. High diversity could be a factor that influences the number of topics.

	doc	**types**	**total tokens**	**avg TTR/doc**	**avg doc length**
CN News	15,905	16,561	3,883,383	0.6096	224
Weibo	23,336	37,125	21,049,072	0.6077	902
EN News	24,998	23,792	6,775,000	0.6788	271
Twitter	25,000	47,259	8,500,000	0.7694	340

Table 4.3 A summary of corpus statistics

CHAPTER 5: RESULTS

“As we bump up against the limits of informational self-determination, we must reflect on what gets lost when we reify privacy as just one thing—one principle, one formalization, one method of protection. We must engage with the whole tangled, ambiguous, and essentially contested terrain of privacy.” (Mulligan et. al., 2016, p.2)

This chapter presents results from structural topic modeling (STM), and the semantic network analysis (SNA). For STM, first, I present the by-language and by-genre modeling results; second, I present the results from the cross-genre analysis (including both news and social media results in one language); and third, I move on to introduce cross-genre and cross-language results.

For each of the modeling results, I describe: first, how the hyperparameter *K* is selected; second, I present the interpretations for each of the topics. All topics are listed in descending order by their *proportion*[16] scores. The explanations of topics were made by considering both the topic words of the *highest probability* and highest *FREX* score.

The SNA was done using the top 50 nodes as determined by centrality scores from the news corpus in the two languages. Comparisons were made regarding the presence or absence of nodes across the two languages. In addition, top nodes identified from the structural space analysis were selected to examine their structural roles changed over time.

[16] Topic *proportions* are used to describe how much of a corpus is devoted to a topic, it is also referred to as topic *prevalence* (Roberts et al., 2016).

5.1 Structural Topic Modeling

5.1.1 Chinese News Results

5.1.1.1 CN News STM K diagnostics

The STM was built using 15,905 original non-translated Chinese news documents which consist of 16,561 unique terms. The *hyperparameter K* was set to 13 as this was the optimal choice suggested by the joint consideration of the coherence and exclusivity scores (see Figure 5.1).

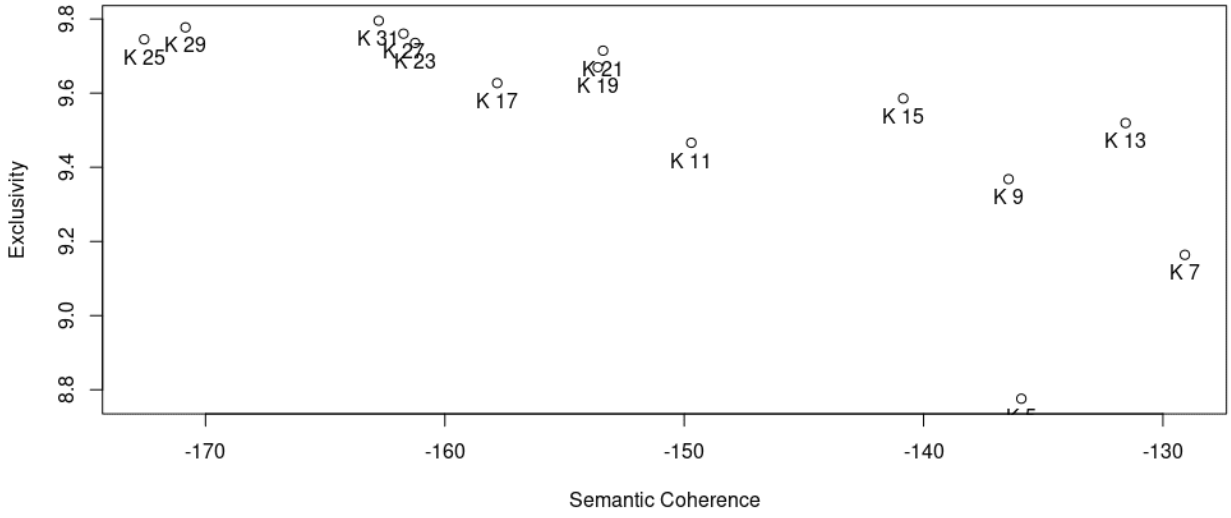


Figure 5.1 The coherence and exclusivity plot of CN analysis

5.1.1.2 Chinese News K13 results

For each topic in Chinese News K13, I present a labeling (in square brackets [ ]) after consolidating the labeling from external coders as described in Section 4.1.2 Interpretation and Labeling of Topics. I also add a brief summary of the meaning of a topic in a couple sentences.

Overall, the 13 topics appear semantically meaningful (see Section 4.1.2 for a detailed account of *semantically meaningful* topics) in that they are about some aspects of privacy. Table 5.1 contains all the CN K13 topics and their highest probability and FREX score topic words. Topics are distinctive with little overlapping or repetition of topic words across topics. However, I observed that one topic, Topic 8, is less straightforward to understand and can present a bit of a challenge when it comes to interpreting how they relate to privacy. In addition, it is also observed that there are topics that contain more than one privacy related sub-topic (for example, Topic 6 and Topic 7).

	Highest Prob	FREX
Topic 12	个人信息, 信息安全, 互联网, 网络安全, 个人隐私, 支付宝, 法律法规	携程网, 中消协, 刻不容缓, 个人主页, 宣传周, 杨建军, 网络安全
	Personal information, information security, Internet, network security, individual privacy, Alipay, law and regulation	Ctrip, China Consumer Association, urgent, personal homepage, Cybersecurity Week, Jianjun Yang, network security
Topic 10	智能手机, 手机用户, 个人隐私, 通讯录, 运营商, 应用程序, 二维码	智能手机, 手机软件, 恶意软件, 恶意程序, 数据恢复, 下载安装, 二手手机
	smartphone, mobile user, individual privacy, contact list, internet service provider, application, QR code	smartphone, mobile software, malware, malicious program, data recovery, download and install, second-hand mobile phone
Topic 6	互联网, 个人隐私, 数据保护, 浏览器, 保护器, 第三方, 委员会	保护器, 奥巴马, 美国政府, 执行官, 联邦贸易委员会, 安全局, 白皮书
	internet, individual privacy, data protection, browser, electric protector, third-party, the Administration	electric protector, Obama, American government, executive officer, Federal Trade Commission, Security Agency, whitepaper
Topic 8	微博上, 越来越, 没想到, 是不是, 王女士, 李先生, 幼儿园	心理咨询, 通知单, 前女友, 刘小姐, 日记本, 信报箱, 青春期
	Weibo, increasingly, unexpected, whether, Ms Wang, Mr Li, kindergarten	Psychological counseling, notice, ex-girlfriend, Ms. Liu, diary, mailbox, adolescence
Topic 2	隐私权, 个人隐私, 公共利益, 当事人, 名誉权, 人格权, 事务所	着作权, 闯红灯, 钱钟书, 男医生, 谢霆锋, 课堂教学, 贵州省
	privacy right, individual privacy, public interest, litigant, right of reputation, right of personality, firm	copyright, run a red light, Zhongshu Qian, male physician, Nicholas Tse, in class education, Guizhou Province
Topic 7	消费者, 互联网, 个人隐私, 公共安全, 征求意见, 信息系统, 人工智能	征求意见, 人工智能, 人脸识别, 保护性, 消保委, 十万元, 一千元
	consumer, internet, individual privacy, public security, request for comments, information system, artificial intelligence	request for comments, artificial intelligence, facial recognition, protective, Consumer Council, 100 thousand RMB, 1000 RMB
	个人隐私, 未成年人, 当事人, 嫌疑人, 公安机关, 年月日, 人民法院	被告人, 判决书, 检察院, 公开审理, 汉弗莱, 中介组织, 开庭审理

Topic 9	individual privacy, teenager/minor, litigant, suspect, public security organizations, year-month-date, People's court	defendant, court verdict, Procuratorate, public trial, Humphrey, intermediaries, court hearing
Topic 1	工作人员, 手机号, 手机号码, 电话号码, 负责人, 个人隐私, 李女士	李女士, 航天员, 李亚鹏, 任志强, 丁嘉丽, 许先生, 王小姐
	employee, mobile number, mobile number, person in charge, individual privacy, Ms Li	Mr Li, astronaut, Yapeng Li, Zhiqiang Ren, Jiali Ding, Mr. Xu, Ms. Wang
Topic 3	摄像头, 公共场所, 隐私权, 个人隐私, 摄像机, 网络平台, 工作人员	摄像头, 摄像机, 健身房, 巩义市, 急诊科, 淫秽物品, 云视通
	webcam, public space, privacy right, individual privacy, camera, network platform, staff	webcam, camera, gym, Gongyi city, emergency room, pornography, CloudSEE
Topic 4	艾滋病, 个人隐私, 用人单位, 医疗机构, 房产信息, 劳动者, 大学生	艾滋病, 房产信息, 劳动者, 感染者, 贫困生, 希拉里, 证明书
	HIV, individual privacy, employer, medical institute, property information, worker, college student	HIV, property information, worker, infected, impoverished/poor students, Hilary, proof
Topic 13	实名制, 身份证, 公务员, 信用卡, 银行卡, 寄件人, 个人信息	实名制, 寄件人, 人口普查, 火车票, 网约车, 一卡通, 垃圾袋
	real name registration, ID card, civil servant, credit card, bank card, sender, personal information	real name registration, sender, census, train ticket, ride hailing, all-purpose card, trash bag
Topic 5	出租车, 朋友圈, 个人隐私, 有限公司, 驾驶员, 工信部, 航空公司	出租车, 附加费, 出租汽车, 中奖者, 新闻节目, 起步价, 周杰伦
	taxi, wechat moments, individual privacy, corporate limited, driver, Ministry of Industry and Information Technology (MIIT), airline companies	taxi, additional fee, taxi, lottery winners, news program, base price, Jay Chow
Topic 11	泰迪熊, 通讯录, 安全卫士, 服务商, 电话号码, 数据安全, 不动产	泰迪熊, 信息源, 朱骏超, 不动产, 近些年, 智能家居, 较长时间
	Teddy Bear, contact list, Security Guard, service provider, phone number, data security, real estate	Teddy Bear, information source, Zhu junchao, real estate, recent years, smart home, extended time

Table 5.1 Translation of topic words of CN K13

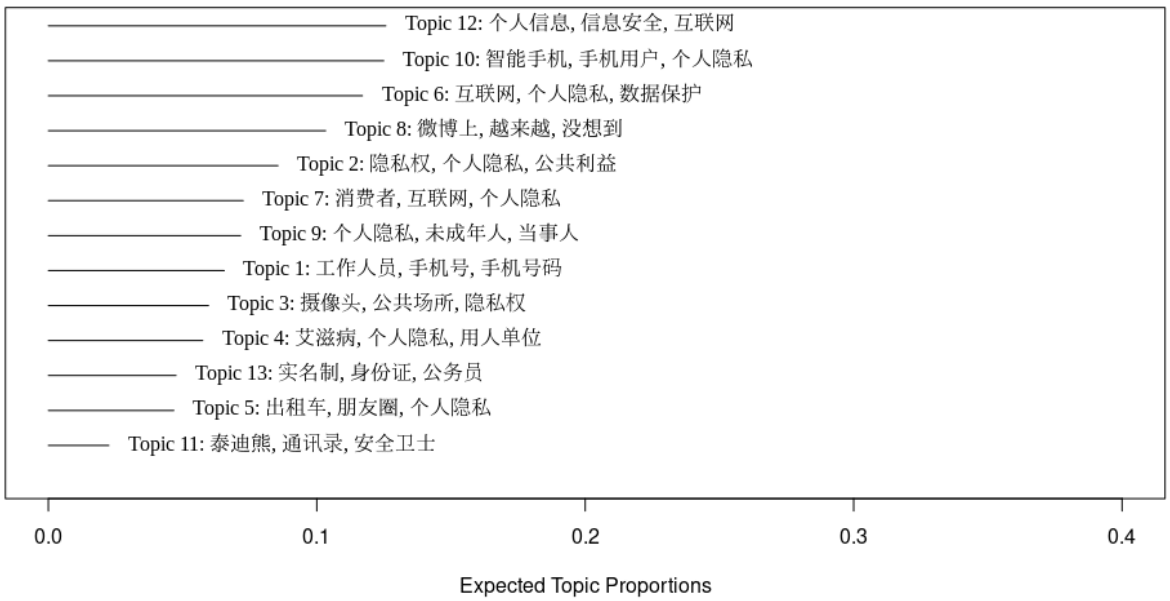


Figure 5.2 The plot of the topic proportions of CN K13
(including the top 3 FREX topic words)

[Network security] Topic 12 is about the Chinese government-led initiatives regarding network security, indicated by topic words like “information security”, “network security”, “Cybersecurity Week”, and “Jianjun Yang” (the associate director of the China Electronics Standardization Institute, “杨建军”).

[Mobile phone] Topic 10 is about a range of “smartphone” related privacy issues, which is indicated by topic words like “contact list”, “QR code”, “software on mobile phone”, “malware”, “data recovery”, and “second hand mobile phone”.

[America and online data protection] Topic 6 is about two privacy-related sub-topics: one is privacy in the United States indicated by topic words such as “American government”, “Obama”, “FTC”; the other concerns online data indicated by topics words like “third party”, “browser”, “internet”.

[Miscellaneous] Topic 8 appears to be in lack clear and coherent themes of privacy, though some topic words could indicate a few online and offline contexts where privacy is particularly at risk, including “psychology counseling”, “Weibo”, and “kindergarten”.

[Privacy related rights] Topic 2 is about the right of privacy and a few associated rights, including “right of reputation”, “right of personality”, and “copyright”.

[Consumer data protection, AI] Topic 7 contains two privacy-related sub-topics: one is about consumer data protection, indicated by topic words like “internet”, “individual privacy”, and “Consumer

Protection Commission"; the other is about AI, indicated by words like "artificial intelligence" and "facial recognition".

[Court trials of privacy] Topic 9 is about court trials of privacy-related cases, which is indicated by words like "litigant", "defendant", and "court".

[Medical information privacy] Topic 4 is about privacy concerns over medical information, indicated by topic words like "AIDS", "infected".

[Real name registration and personal ID information] Topic 13 is about privacy concerns over a few real-world everyday practices in China, which all could present privacy risks. These practices include the "real-name registration", "taxi-hailing", "express information of the sender", "ID card", and "bank card".

[Taxi] Topic 5 is about Taxi, indicated by words like "Taxi", and taxi-hailing service which is a feature embedded within the Wechat app, indicated by topic words like "wechat moments".

[Smart home appliance data] Topic 11 is about data security and services around smart home appliances, indicated by words like "smart home appliances", "data security".

[Celebrity and mobile contact privacy] Topic 1 contains two sub-topics: one is about privacy concerns of mobile contact, which is indicated by keywords like "mobile number", and "phone number"; and the other is about celebrities.

[Webcam privacy violation] Topic 3 is about privacy risks associated with "webcam" and pervasive "camera" installation at various public places, for example, at "gym".

The three topics with the highest topic proportion are Topic 12, Topic 10, and Topic 6. These leading topics generated from the Chinese news corpus suggest three most important aspects of concern when discussing privacy in the Chinese language today. Topic 12 suggests a sense of the government protecting online consumer information and safeguarding network security. Topic 10 suggests what may be concerning most people on an everyday basis, which is about their privacy and data security on their mobile phones. Topic 6. about America, reflects more about the political and strategic importance of the United States to China; this topic also reveals that the internet is another context (in addition to mobile phones) where privacy is concerned.

5.1.2 Weibo Results

The STM was built using in total 23,336 original non-translated Chinese Weibo documents which consist of 37,125 unique terms. The coherence and exclusivity plot of Weibo results did not differentiate significantly for the *hyperparameter K* between the values of 11 and 13 (see Figure 5.3). Hence, a choice of *K*13 was made to better facilitate comparison between topic modeling results between Weibo and Chinese News (which also set the value of *K* as 13).

5.1.2.1 Weibo STM K diagnostics

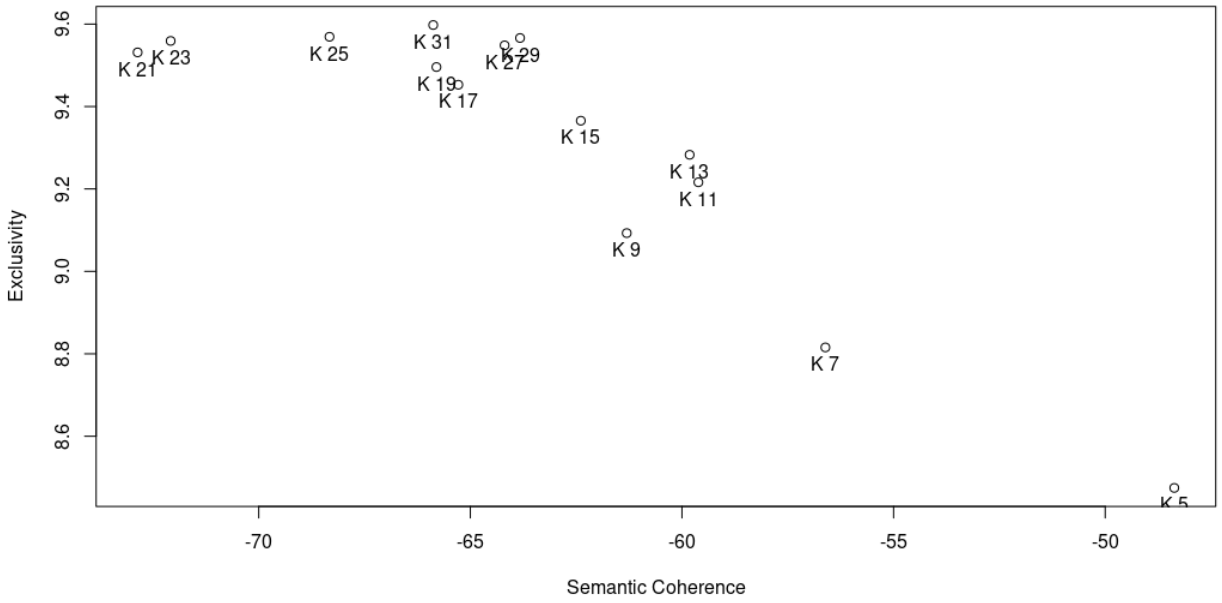


Figure 5.3 The coherence and exclusivity plot of Weibo analysis

5.1.2.2 Weibo K13 results

Overall, compared to CN news topics, Weibo topics are less semantically meaningful and interpretable. There is more repetition and overlapping across topic words in Weibo topics. Table 5.2 contains all the Weibo K13 topics and their highest probability and FREX score topic words. For example, words “individual privacy” and “personal information” appear as leading topic words in 6 out of the 13 topics. There are also topics with only 1 to 3 topic words that can be interpreted as being related to privacy, while the rest of the topic words may be less related to privacy. For example, among the topic

words of Topic 5, only “individual privacy”, “Meitu”[17] and “Tangdui”[18] appear to relate to privacy; and among topic words of Topic 6, only “individual privacy”, “QR code” and “personal information” appear to relate to privacy. For Topic 9, only the two topic words “individual privacy”, and “webcam” appear to be intuitively related to privacy.

In addition to the fact that Weibo documents were manually aggregated, the challenge with understanding Weibo topics may be partly attributed to the language of Weibo posts, which appear to have a much more diverse and much less formal vocabulary compared to news. The processed Weibo corpus resulted in 37,125 unique terms, whereas the Chinese news pre-processed resulted in a total of 16,561 unique terms (see also in Chapter 4: Methods & Data, for a table of corpus statistics).

Weibo topics consist of many idioms that were not seen in Chinese news topic words, such as “illegal court” (“私设公堂”) and “nepotism” (“裙带关系”)[19], which are common in everyday colloquial language (Jiao, 2016). In addition, Weibo topics also have many words that are from fictions and popular entertainment and TV shows, such as Monkey King (“孙悟空”) and Xuanyuan Sword (“轩辕剑”). Lastly, the Weibo corpus also contains buzzwords, celebrity names and product names, and onomatopoeia which are not seen in the topics generated from the Chinese news corpus, such as, “competent women” which is a buzzword in Chinese society; and “Yuehua”, which is a short name for an entertainment company.

	Highest Prob	**FREX**
Topic 11	个人隐私, 个人信息, 水瓶座, 朋友圈, 隐私权, 越来越, 手机号	马背上, 航天中心, 私设公堂, 磁力线, 券用券, 气质端庄, 妻葛黛瓦
	Individual privacy, personal information,aquarius, wechat moments, privacy right, increasingly, mobile number	horseback, space center, illegal court, coupon, magnetic line, elegant, wife Godiva

[17] Meitu, in short for Meitu Xiu Xiu, is an image editing software that is very popular in China.

[18] Tangdui is an image sharing website.

[19] “裙带关系” (as well as “私设公堂”) is a Chinese idiom which I managed to roughly translate as “nepotism”. Idioms are commonly used in daily conversations (Jiao, 2016) and can be difficult to translate for its embeddedness in cultural background (Zhong, 2015).

Topic 2	个人隐私, 办公室, 发生冲突, 善解人意, 表达意见, 人际关系, 水瓶座	黄金律, 吃眼前亏, 裙带关系, 祥林嫂, 绊脚石, 方舟子, 顺口溜
	Individual privacy, office, conflict, considerate, express pinion, interpersonal relationship, aquarius	Golden rule, suffer losses, nepotism, Aunt Xianglin, stumbling block, Fang Zhouzi, tongue twister
Topic 8	个人隐私, 双子座, 天蝎座, 巨蟹座, 金牛座, 白羊座, 射手座	能谋善, 第十一名, 第十二名, 变形金刚, 第六名, 第七名, 第八名
	Individual privacy, gemini, scorpio, cancer, taurus, aries, sagitarius	Resourceful, eleventh, twelfth, transformers, sixth, seventh, eighth
Topic 5	个人隐私, 脑子里, 一点点, 美图秀, 幼儿园, 下半生, 拦路虎	堆糖网, 蜘蛛精, 脑子里, 孙悟空, 美图秀, 乔布斯, 豆腐心
	Individual privacy, in one's mine, a bit, meitu, kindergarten, second half of life, obstacle	Tangdui net, Spider ghost, in one's mind, Monkey King, Meitu, Jobs, soft-hearted
Topic 4	个人隐私, 朱正廷, 个人信息, 互联网, 越来越, 朋友圈, 请乐华	朱正廷, 请乐华, 亚布力, 计算机硬件, 杜华乐华, 李宗伟, 交易商
	Individual privacy, Zhu Zhengting, personal information, internet, increasingly, wechat moments, ask yuehua	Zhu Zhengting, ask yuehua, Yabuli, computer hardware, Duhua yuehua, Li Zongwei, dealer
Topic 10	个人隐私, 明月光, 高风险, 客户端, 逃不了, 系统安全, 个性化	北京电影学院, 轩辕剑, 表演系, 天之痕, 文博会, 民族服装, 王老古
	Individual privacy, moonlight, high risk, customer end, inescapable, system security, customized	Beijing film academy, Xuanyuan Sword, performance department, Scar of Sky, culture exhibition, ethinic apparel, Wang Laogu
Topic 13	个人隐私, 自由空间, 个人信息, 是因为, 隐私权, 朋友圈, 发牢骚	智键护, 价元起, 全金属, 客客气气, 蜻蜓点水, 阴阳怪气, 君子之交淡如水
	Individual privacy, free space, personal information, because, privacy right, wechat moments, complain	Smart fingerprint protection, price, All metal, polite, touch on something, bad-tempered, A hedge between keeps friendship green
Topic 7	个人隐私, 个人信息, 摄像头, 朋友圈, 隐私权, 信息安全, 互联网	一分钟, 非必要, 沐浴液, 光大银行, 纸巾盒, 生活用品, 电子钟
	Individual privacy, personal privacy, webcam, wechat moments, privacy rights, information security, internet	One minute, unnecessary, shampoo, Everbright Bank, tissue box, daily necessities, electronic clock

Topic 6	个人隐私, 弄清楚, 保持中立, 尖酸刻薄, 二维码, 有没有, 个人信息	退避三舍, 衣衫褴褛, 骨子里, 衣冠楚楚, 冠冕堂皇, 趋之若鹜, 弄清楚
	Individual privacy, figure out, remain objective, mean, QR code, whether or not, personal information	Avoid, poorly dressed, in one's heart, dressed up, high-sounding, go after sth., figure out
Topic 3	个人隐私, 隐私权, 保护器, 互联网, 是不是, 越来越, 实名制	保护器, 人口普查, 打击报复, 郭德纲, 安全部队, 看人脸色, 抱佛脚
	Individual privacy, privacy right, protector, internet, whether or not, increasingly, real name registration	Protector, census, retaliate, Guo Degang, security troops, being given an attitude, last minute effort
Topic 1	隐私权, 当事人, 个人隐私, 哈哈哈, 名誉权, 肖像权, 未成年人	健康权, 姓名权, 荣誉权, 生命权, 专利权, 判决书, 诉讼法
	Privacy right, stakeholder, Individual privacy, hahaha, reputation right, portrait right, teenager	Health right, name right, reputation right, life right, patent right, verdict, Procedural law
Topic 12	个人隐私, 隐私权, 个人信息, 摄像头, 网络安全, 朋友圈, 互联网	致一人, 萨福克, 流离失所, 绝一人, 下议院, 英国议会, 妇女节
	Individual privacy, privacy right, personal privacy, webcam, network security, wechat moments, internet	To one, Suffolk, homeless, no one, House of Commons, British Parliament women's day
Topic 9	摄像头, 个人隐私, 最会学, 第二语言, 双子座, 女强人, 自我解嘲	最会学, 第二语言, 满满当当, 牢骚满腹, 自我解嘲, 怒气冲天, 女强人
	Webcam, individual privacy, best learning, second language, gemini, capable women, laughing at oneself	Best learning, second language, full, complain, laughing at oneself, angry, capable women

Table 5.2 Translation of topic words of Weibo K13

[Personal information and individual privacy] Topic 11 appears to be about privacy and personal information risks associated with "mobile number", and "wechat moments".

[Privacy at work context and interpersonal relationship] Topic 2 is about privacy concerns in the context of interpersonal relationships, as indicated by topic words like "interpersonal relationship", "considerate", and "office".

[Astrology] Topic 8 is about astrology.

[Miscellaneous] Topic 5 appears to lack coherent themes of privacy, though it does touch on a couple of apps that may have potential privacy issues, including photoshopping apps and websites, indicated by words like "Meitu", and "Tangdui".

[Celebrity] Topic 4 appears to be about privacy concerns over celebrities, indicated by topic words like “Zhu Zhengting”, “Yuehua”, and “Li Zongwei”.

[Information system security] Topic 10 is about system security, which is indicated by topic words like “customer end” and “system security”, “customize”.

[Wechat moments] Topic 13 appears to be about privacy concerns regarding “Wechat moments”, and how WeChat moments may be considered as a “free space” for “personal information” and even “complain”.

[Hidden webcam installed in daily items] Topic 7 is about privacy concerns over webcam and a general sense of concern over the security of personal information on the internet, which is indicated by topic words like “personal information”, “webcam”, “internet”, “information security”, and daily objects that have a camera embedded in them as have reported by journalists, including “tissue box”[20].

[Miscellaneous] Topic 6 appears to lack clear and coherent themes of privacy, though one topic word may be related to privacy, which is “QR code”.

[Real name registration and individual privacy on the internet] Topic 3 is about privacy risks related to the practices of “real name registration”, “individual privacy”, and the “internet”.

[Privacy related rights] Topic 1 is predominantly about several privacy-related rights, indicated by topic words like “reputation right”, “portraiture right”.

[Network security] Topic 12 appears to be about privacy associated with network security, which is indicated by topic words like “network security” and “Suffolk” who is the Global Cyber Security Officer at Huawei.

[Miscellaneous] Topic 9 does not have coherent and clear themes of privacy as the majority of the topic words are not associated with privacy, the only exception is the topic word “webcam”.

---

[20] Huanqiu.com 环球网 (2019, May 29) *Chazuo, matongshua, zhijinhe, zhexie dongxi douneng biancheng toupai nide shexiangtou……* [Socket, toilet brush, and tissue box: all things that can embed hidden cameras filming you in secret …]. Retrieved January 19, 2022, from https://baijiahao.baidu.com/s?id=1634844206467836252&wfr=spider&for=pc

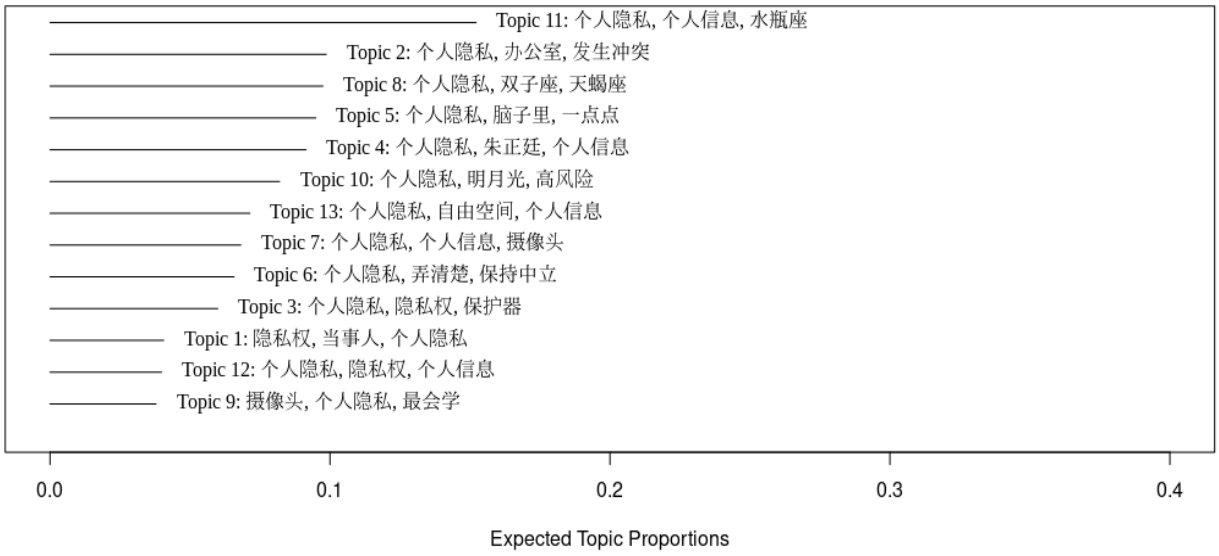


Figure 5.4 The plot of the topic proportions for Weibo K13
(displaying also the top 3 FREX topic words)

The top two Weibo topics in terms of topic proportion (see Figure 5.4), Topic 11 and Topic 2, reveal the aspects of concern that are related to most people's daily life and practices. Topic 11 is about the specific tech application that concerns privacy and personal information, which are "mobile number", and "wechat moments"; whereas Topic 2 reveals the context in which a range of privacy-related issues could arise.

In addition, though some of the topics and subtopics that touch on individuals' daily concern of privacy can be seen in the Chinese news results (for example, Topic 10 and Topic 13), they appear more prominent in Weibo results (for example, Topic 11 and Topic 2) both in terms of the frequency they appear in topics, and the ranking of topic proportion.

Weibo, as a social network platform, is where users post and discuss personal concerns and feelings, which may result in the casual nature of some of the Weibo topics. In addition, unlike news that are professionally edited content, Weibo topics perhaps reveal what everyday people truly concern about privacy.

5.1.3 A comparison of CN News K13 and Weibo K13 topics

To identify a complete set and a core set of semantics in the Chinese language, topics generated from Chinese news and Weibo were compared (see Table 5.3). Then, topics or subtopics shared across the two genres were identified as the *core semantics* (see Table 5.4 and 5.5).

CN K13	No.	Weibo K13
Celebrity and mobile contact privacy*	Topic 1	Privacy related rights
Privacy related rights	Topic 2	Privacy at work context and interpersonal relationship
Webcam privacy violation	Topic 3	Real name registration and individual privacy on the internet
Medical information privacy	Topic 4	Celebrity
Taxi	Topic 5	Miscellaneous*
America and online data protection*	Topic 6	Miscellaneous*
Consumer data protection, AI*	Topic 7	Hidden webcam installed in daily items
Miscellaneous*	Topic 8	Astrology
Court trials of privacy	Topic 9	Miscellaneous*
Mobile phone	Topic 10	Information system security
Smart home appliance data	Topic 11	Personal information and individual privacy
Information security on the internet	Topic 12	Network security
Real name registration and personal ID information	Topic 13	Wechat moments

Table 5.3 All topics from the Chinese corpus
(*indicates topics that have more than one subtopics)

Compared to CN News, more Weibo topics have more than one themes or subtopics within one topic. Hence, Weibo topics are higher in granularity, indicated by the higher presence of topics that were labeled as “Miscellaneous”. The higher granularity of Weibo topics can be attributed to two sources of influence. One is the nature of Weibo as a more casual genre, and the more casual and more diverse the language is, the more granular the topics are; the other possible reason is the manual aggregation of Weibo posts.

Some topics and sub-topics can be found in both genres. Hence, they are considered the core semantics for understanding privacy in the Chinese language (see Table 5.5). However, there are also topics and sub-topics that appear unique to either the news or social media genre. The four topics that only appeared in CN news topics include: Topic 4 Medical information privacy, Topic 6 America and online data protection, Topic 9 Court trials of privacy, and Topic 11 Smart home appliance data. One topic appears uniquely in Weibo results, which is Topic 8 Astrology.

Chinese news topics, compared to the casual style of Weibo topics, touch on more serious and formal issues like court trials (in Topic 9), and international issues indicated by the presence of the United States (in Topic 6). In contrast, Weibo topics' more casual style is revealed not only through topics that are entertaining in nature such as astrology (in Topic 8); but also through topics that appear to be more common in daily practices of personal information and information technologies, which are reflected through the presence of the topic words like “QR code” (Topic 6), and “Meitu”[21] (Topic 5).

No.	Complete CN Semantics
1	Celebrity and mobile contact privacy
2	Privacy related rights
3	Webcam privacy violation
4	Medical information privacy
5	Taxi
6	America and online data protection
7	Consumer data protection, AI
8	Court trials of privacy
9	Mobile phone
10	Smart home appliance data
11	Network security
12	Real name registration and personal ID information
13	Privacy at work context and interpersonal relationship
14	Astrology

[21] Meitu, in short for Meitu Xiu Xiu, is an image editing software that is very popular in China.

15	Wechat moments
16	Information system security
17	Personal information and individual privacy

Table 5.4 Combined semantics of the Chinese corpus

Four topics are shared across these two genres (see Table 5.5), hence they are considered the core semantics of privacy in the Chinese language. The majority of core semantics are associated with privacy from an individual's perspective, including privacy related rights, webcam violation of privacy that individuals could face in the everyday context, and real-name registration and personal ID information, which are indispensable everyday life elements for many people, from purchasing train tickets to posting information online. The only exception is network security that reflects more the perspective from institution and the public.

No.	**CN Core Semantics**
1	Privacy related rights
2	Webcam privacy violation
3	Network security
4	Real name registration and personal ID information

Table 5.5 Core semantics of the Chinese corpus

5.1.4 English News Results

The STM with the English news corpus was built using 24,998 documents and 23,792 terms. The *hyperparameter K* was set to 11 as this is the optimal choice suggested by the coherence and exclusivity plot (see Figure 5.5).

5.1.4.1 EN News STM K diagnostics

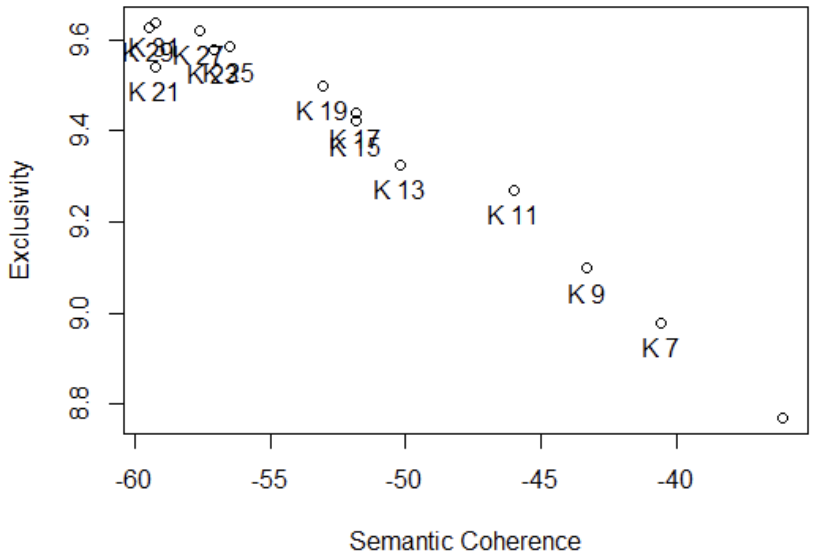


Figure 5.5 The coherence and exclusivity plot of EN analysis

5.1.4.2 English News K11 results

The vast majority of the 11 English news topics are semantically meaningful for how they relate to privacy; there is little overlapping across topics, and usually, the topic words can be interpreted under one coherent theme for each of the topics, rather than multiple subtopics. Further searches using Google were conducted to add back their full name for some of the acronyms when interpreting the topics (for example, “OTA” which is Online Trust Alliance, and “HUD” which is U.S. Department of Housing and Urban Development). Table 5.6 contains all the EN K11 topics and their highest probability and FREX score topic words

	Highest Prob	**FREX**
Topic 10	privaci, data, secur, inform, com, provid, manag	fairwarn, onetrust, hitrust, csf, patent, ota, isaca
Topic 1	user, facebook, privaci, compani, googl, said, data	zuckerberg, googl, facebook, app, appl, cambridg, analytica

Topic 9	privaci, consum, protect, data, bill, inform, senat	fcc, broadband, markey, rep, subcommitte, isp, ntia
Topic 8	record, system, inform, feder, act, offic, privaci	ssa, hud, sorn, usci, osd, dod, docket
Topic 5	inform, consum, data, privaci, person, provid, requir	coppa, ccpa, credit, hipaa, settlement, ftc, breach
Topic 2	said, student, health, state, school, use, patient	scanner, student, teacher, hospit, classroom, dna, patient
Topic 3	think, know, say, one, peopl, get, like	imus, malveaux, clip, tonight, velshi, cavuto, yeah
Topic 4	govern, secur, said, surveil, american, nation, agenc	nsa, drone, faa, snowden, terror, terrorist, aircraft
Topic 11	data, privaci, protect, european, compani, law, shield	schrem, shield, european, transatlant, apec, gdpr, europ
Topic 7	court, law, case, privaci, enforc, investig, search	suprem, circuit, warrant, judg, court, fourth, subpoena
Topic 6	person, violat, section, physic, imag, shall, record	subdivis, visual, shall, impress, compensatori, physic, sound

Table 5.6 Topic words of EN K11

[Privacy and data security services] Topic 10 is about privacy issues concerning privacy service and management and compliance solution providers. This includes companies that provide privacy compliance services like “fairwarning”, “onetrust”, “hitrust”. In addition, this topic also includes certificate providers, including OTA (Online Trust Alliance), and privacy certification providers, including ISACA (Information Systems Audit and Control Association), and the NIST Cybersecurity Framework (NIST CSF).

[Facebook, Google, Cambridge Analytica] Topic 1 is about privacy concerning three technology companies, “Facebook”, “Google”, and “cambridg” “analytica”.

[Regulations and regulators of privacy] Topic 9 concerns America’s national regulations of “isp” (Internet service providers), indicated by topic words like “bill”, “senate”, “fcc” (Federal Communications Commission), “broadband", “rep”, “markey”[22], subcommittee, and “NTIA” (The National Telecommunications and Information Administration).

[22] Edward Markey, U.S. representative from 1976 to 2013.

[U.S. governmental records] Topic 8 is about privacy issues concerning the “record” “system” of the “feder” governmental organizations of the United States, indicated by topic words including the Social security administration (“SSA”), U.S. Department of Housing and Urban Development (“HUD”), System of Records Notices (“SORNs”), United States Citizenship and Immigration Services (“USCI”), the Office of the Secretary of Defense (“OSD”), and The Department of Defense (“DOD”).

[Consumer data privacy protection] Topic 5 concerns privacy issues focusing on data protection enabled/supported by regulations, indicated by topic words “consum”, “settlement”, specific privacy regulations and regulatory institution, including “COPPA” (Children's Online Privacy Protection Act) and “CCPA” (California Consumer Privacy Act), “HIPAA” (Health Insurance Portability and Accountability Act), and “FTC” (Federal Trade Commission).

[Student and patient data] Topic 2 is about privacy concerns over specific types of data associated with specific population groups, indicated by topic words like “student”, and “patient”, “teacher”, “health”, “school”, “dna”, etc.

[TV show hosts and journalists] Topic 3 concerns the privacy of celebrities, indicated by topic words like “imus”, “malveaux”, “velshi”, “cavuto”, which are TV anchors/journalists’ names, and also TV show names like “tonight”.

[Government surveillance and national security] Topic 4 concerns privacy issues surrounding government surveillance and national security, indicated by topic words including “secur”, “nation”, “agenc”, and topic words like “faa” (Federal Aviation Administration) “NSA”, “snowden”; “drone”, “terror”, “aircraft”.

[International privacy laws] Topic 11 concerns privacy issues surrounding international privacy regulations and frameworks, indicated by topic words like “european”, “shield”[23], “transatlantic”, “gdpr” (General Data Protection Regulation), “apec” (Asia-Pacific Economic Cooperation), and “schrems”[24].

[Law enforcement and privacy] Topic 7 concerns “court” “subpoena” and privacy, and privacy issues during legal “search” and “investigation”, and the “fourth amendment”.

---

[23] *Welcome to the Privacy Shield*. Privacy Shield. (n.d.). Retrieved January 19, 2022, from https://www.privacyshield.gov/welcome

[24] Maximilian Schrems is an Austrian activist, lawyer, and author who was known for campaigns against Facebook for its privacy violations.

[Privacy violation of records] Topic 6 appears to be about privacy violation of records, indicated by topic words like “violat”, “record”, etc.

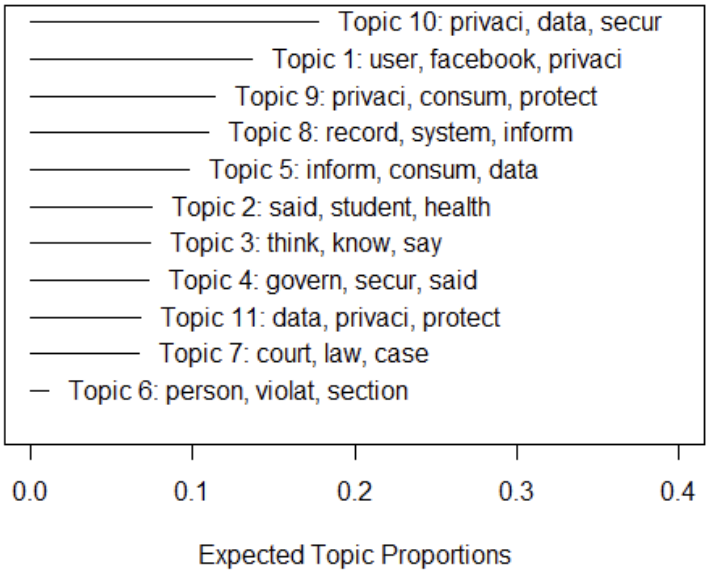


Figure 5.6 The plot of the topic proportions for EN K11
(displaying also the top 3 FREX topic words)

Topics with the highest proportions in EN K11 reflect primarily privacy concerns about two types of institutions. One is corporate organizations: Topic 10 Privacy and data security services, and Topic 1 Facebook, Google, Cambridge Analytica. The other one concerns governmental organizations: Topic 8 U.S. governmental records, and Topic 9 Regulations and regulators of privacy.

5.1.5 Twitter Results

The Twitter STM was built using a total of 25,000 documents and 47,259 terms. The exclusivity and coherence plot indicated that *hyperparameter K* is best set at 27 (see Figure 5.8). However, a quick review of K27 modeling results shows that K27 topics were difficult to interpret in that topics were full of words repeating across topics (see APPENDIX A.7 for a list of topics for K27). Therefore, K15 was used instead for the following reasons. First, compared to the next two options, K23 and K21, K15 significantly reduces the number of topics to interpret. Second, K15 is not far from K27 in terms of its coherence and exclusivity performance. Third, K15 is closer to the number of topics chosen for the analysis of the Chinese language and English news. Hence, this choice of K15 could better facilitate comparison of topics across language and genre.

5.1.5.1 Twitter STM K diagnostics

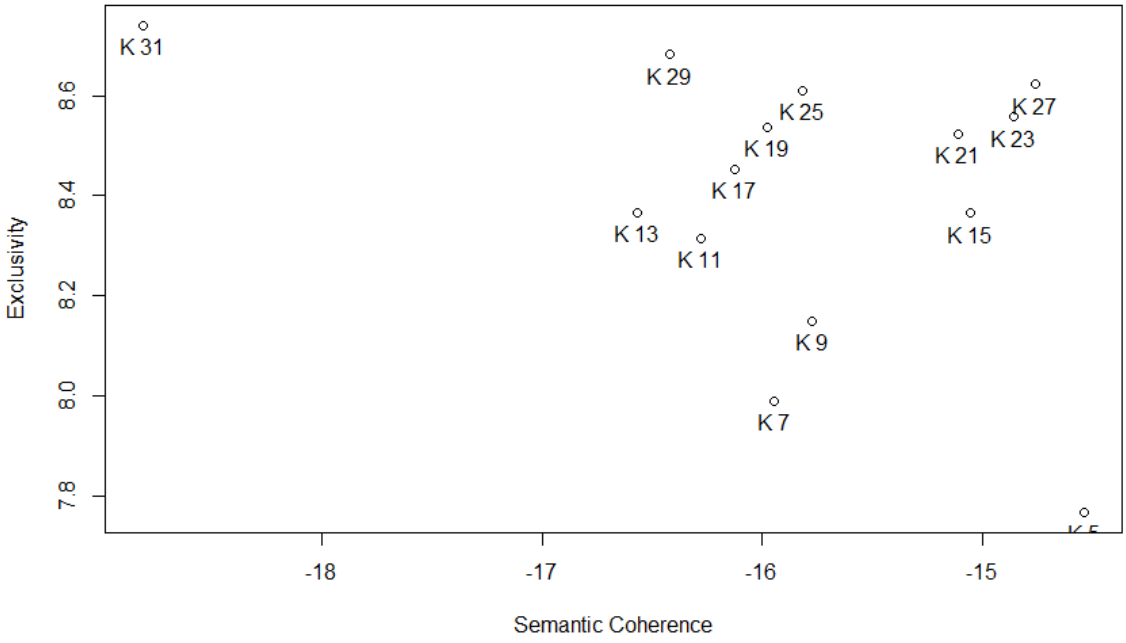


Figure 5.7 The coherence and exclusivity plot of Twitter analysis

Overall, Twitter topics are less intuitive than English news topics. Table 5.7 contains all the Twitter K15 topics and their highest probability and FREX score topic words. Twitter topics are high in the repetition of the words across all topics, those words are “facebook”, “google”, “data”, and “security” appear in multiple topics, making Twitter topics less distinct from each other. In addition, the majority of Twitter topics contain multiple sub-topics. For example, Topic 14 appears to have many distinctive topic

words related to privacy), which is quite similar to the observation for the Chinese Weibo topics (see Section 5.1.3 for a comparable discussion).

5.1.5.2 Twitter K15 results

Overall, leading Twitter topics centered around a couple of big tech companies, “Facebook” and “Google”. “Facebook” appears among leading topic words in 9 out of the total 15 topics. And “Google” appears in 5 out of 15 topics. In addition, Twitter topics also focused on very specific digital technology applications. As expected, Twitter topic words appeared more casual and less coherent than English news topics.

	Highest Prob	**FREX**
Topic 10	privaci, data, like, peopl, right, secur, need	xboxpaxaus, faceapp, leaderboard, ccpa, nica, doordash, securypto
Topic 6	privaci, facebook, googl, polici, need, set, user	demandprogress, plenti, petraeus, nich, randi, buyer, hysteria
Topic 5	privaci, data, secur, facebook, right, protect, like	datafund, cardi, deeponion, myhealthrecord, kavanaugh, ethereum, capitaltechnologiesresearch
Topic 7	privaci, protect, data, secur, free, american, real	encyrypt, cisa, stealthcoin, idltweet, barbi, newpanda, stitm
Topic 9	privaci, googl, facebook, like, protect, data, secur	cispaalert, endcispa, prism, obamacar, typewrit, bush, cispa
Topic 2	privaci, secur, data, real, free, show, absolut	glue, trumptransit, presidentelecttrump, static, infosecjob, maga, giveaway
Topic 11	privaci, data, real, free, secur, show, absolut	hat, foil, locker, evernot, pokmon, pokemon, fertil
Topic 1	privaci, facebook, googl, set, social, polici, onlin	icanstalku, buzz, tinyurl, nearbi, appspot, medal, proxypi
Topic 14	privaci, facebook, polici, set, onlin, like, protect	mellon, carnegi, onstar, hampton, geofenc, sacr, roethlisberg
Topic 13	privaci, data, secur, protect, facebook, onlin, internet	bonus, surfeasyinc, null, truecrypt, pear, papul, ncpol
Topic 4	privaci, data, secur, facebook, protect, like, need	granada, degrad, pension, nkdgvijlnn, classdojo, blackphon, glenn

Topic 3	privaci, facebook, like, need, set, peopl, want	blah, ittech, nut, jenni, shoe, vermont, medit
Topic 8	privaci, facebook, set, googl, polici, onlin, protect	geotag, error, kingston, carrier, timelin, datatravel, verdict
Topic 12	privaci, facebook, googl, set, polici, secur, onlin	doodl, czar, freelanc, lennon, creatur, invadin, buzz
Topic 15	privaci, data, secur, like, peopl, facebook, protect	nbsp, field, entri, npleas, soul, fill, pacif

Table 5.7 Topic words of Twitter K15

[Apps, privacy regulation] Topic 10 has two distinctive privacy-related sub-topics, first one is specific privacy and data protection regulation, indicated by the topic word “CCPA” (California Consumer Privacy Act); the other is digital currency “securypto”; and popular apps, including a food delivery service app “doordash”, and a photo editing app “faceapp”.

[Privacy policies of Facebook and Google] Topic 6 is mainly about “polici” (policy) of user privacy of two companies, “google” and “facebook”.

[Facebook, cryptocurrency] Topic 5 contains two privacy-related sub-topics; “facebook”; and a variety of online data services, and cryptocurrencies, including “myhealthrecord”, “deeponion”, “ethereum”, “datafund”.

[Miscellaneous] Topic 7 appears to have multiple sub-topics, lacking one clear and coherent theme of privacy. Topic words that are related to privacy include: “data”, “secur”, “encyrypt”, “CISA” (Cybersecurity and Infrastructure Security Agency), “stealthcoin”, and “newpanda”[25].

[Cyberintelligence] Topic 9 is about government online surveillance and cyberintelligence, indicated by topic words like “cispaalert”, “endcispa”, “prism”, “cispa” (Cyber Intelligence Sharing and Protection Act).

[Security and Trump] Topic 2 is broadly about security during the Trump administration, as indicated by topic words including “secur”, and “trumptransit”, “presidentelecttrump”, and “infosecjob”.

[Data security of apps] Topic 11 Appears to refer to data security and privacy concerns over two specific digital apps, one is “evernot”, the other one is “pokemon”.

[Privacy of social media] Topic 1 is mainly about privacy concerning social media, which is indicated by topic words like “facebook”, “social”

[25] newpanda. (n.d.). Retrieved January 19, 2022, from https://www.getnewpanda.com/

[Location data] Topic 14 appears to be mainly location data, indicated by topic words like "geofencing", and "OnStar"[26].

[Encryption service] Topic 13 is about data encryption services, indicated by topic words like "truecrypt", and "surfeasyinc".

[Tech companies and data security] Topic 4 is about privacy related to tech companies indicated by topic words like "facebook", "blackphon"; and concerns over data security in general, which is indicated by topic words like "data" and "secur".

[Facebook] Topic 3 appears to be lacking a clear theme, though the presence of the topic word "facebook" indicates that this topic could be about facebook.

[Flash memory] Topic 8 is about flash memory, which is indicated by topic words that are referring to data storage service providers: "kingston", "datatravel".

[Facebook, Google, and Russia] Topic 12 is about privacy concerns over Russia's involvement of two American tech companies, "facebook" and "google".

[Data security, facebook] Topic 15 is about privacy and data security and Facebook.

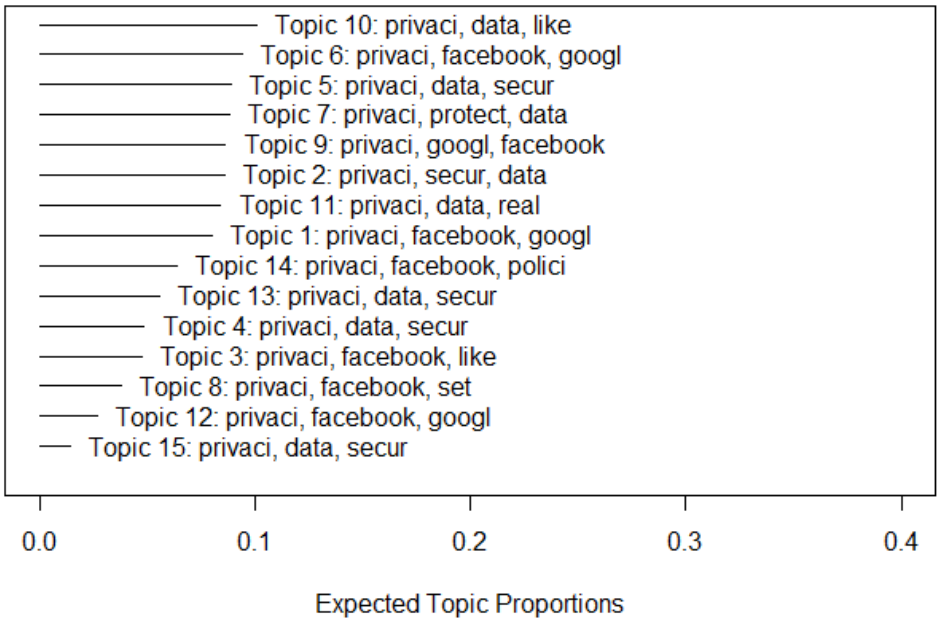


Figure 5.8 The plot of the topic proportions for Twitter K15
(displaying also the top 3 FREX topic words)

[26] OnStar is a subsidiary of General Motors that provides services including subscription-based communications, in-vehicle security.

Topics that are leading in proportion appear to be centering on first, the regulations and policies of privacy; and second, some of the most cutting edge digital technology applications, including cryptocurrencies (“securypto”, “ethereum”, “stealthcoin”). Topics that focus on governmental records and government practices appear in multiple topics. Including: Topic 10 Apps, privacy regulation; and Topic 9 Cyberintelligence, etc. Technology companies and applications are present among multiple topics, these technology companies and applications include: “Facebook”, “Google”, as well as some other popular apps like “doordash” and “faceapp”; they have appeared in topics including: Topic 6 Privacy policies of Facebook and Google, and Topic 11 Data security of apps.

5.1.6 A comparison of English News K11 and Twitter K15 results

Overall, English news topics concerned regulations and laws, and governmental practices; whereas, topics of Twitter appear to be more specific, focusing on very concrete applications and technologies. Topics that are unique to English news include Topic 2 Student and patient data, Topic 3 TV show hosts and journalists, and Topic 11 International privacy laws. Topics and subtopics that are unique to Twitter tend to focus on very concrete technological applications, cryptocurrency, and encryption, including: Topic 8 flash memory, Topic 13 encryption service, Topic 14 location data for topics. Apps also appear as subtopics in Topic 10 and Topic 11. In addition, the three shared topics/subtopics between the two genres are government surveillance, Facebook and Google, and Privacy law/regulation. These three subtopics together constitute the *core* semantics of privacy in the English corpus. In other words, the core semantics in the English language all concern the institutions of privacy, emphasizing big tech companies, governmental organizations, and regulations.

EN K11	**No.**	**Twitter K15**
Facebook, Google, Cambridge Analytica	Topic 1	Privacy of social media
Student and patient data	Topic 2	Security and trump
TV show hosts and journalists	Topic 3	Facebook
Government surveillance and national security	Topic 4	Tech companies and data security
Consumer data privacy protection	Topic 5	Facebook, Cryptocurrency
Privacy violation of records	Topic 6	Privacy policies of facebook and google
Law enforcement and privacy	Topic 7	Miscellaneous
US governmental records	Topic 8	Flash memory
Regulations and regulators of privacy	Topic 9	Cyberintelligence
Privacy and data security services	Topic 10	Apps, privacy regulation
International privacy laws	Topic 11	Data security of apps
-	Topic 12	Facebook, Google and Russia
-	Topic 13	Encryption service
-	Topic 14	Location data
-	Topic 15	Data security, facebook

Table 5.8 All topics from the English corpus

No.	Complete EN Semantics
1	Facebook, Google, Cambridge Analytica
2	Student and patient data
3	TV show hosts and journalists
4	Government surveillance and national security
5	Consumer data privacy protection
6	Privacy violation of records
7	Law enforcement and privacy
8	US governmental records
9	Regulations and regulators of privacy
10	Privacy and data security services
11	International privacy laws
12	Social media
13	Security and trump
14	Cryptocurrency
15	Privacy policies of Facebook and Google
16	Flash memory
17	Cyberintelligence
18	Apps
19	Facebook, Google and Russia
20	Encryption service
21	Location data

Table 5.9 Combined semantics in the English corpus

No.	EN Core Semantics
1	Facebook, Google, Cambridge Analytica
2	Government surveillance and national security
3	Regulations and regulators of privacy
4	Privacy and data security services

Table 5.10 Core semantics in the English corpus

5.1.7 Cross-Language News STM Results

This cross-language analysis used in total 20,000 news articles documents: 10,000 from the English news corpus, and 10,000 from the translated Chinese news corpus. 1,000 articles were drawn from each of the years from 2010 to 2019. As a result of preprocessing, 41,433 of the original 59,313 terms (56,168 of 3,469,902 tokens) were removed due to having a frequency lower than 3.

The social media corpus (Twitter and Weibo) was not considered for cross-language STM modeling because of the high repetition and overlapping across topics that were already seen when analyzing each genre in each language (see 5.1.2 Weibo Results and 5.1.5 Twitter Results).

5.1.7.1 Cross language News STM K diagnostics

The exclusivity and coherence plot indicated that 9, 11, and 13 were the three leading *hyperparameter K* options (see Figure 5.10), among which K11 was selected mainly because selecting the same *K* value that is used for the cross-genre and cross-language analysis could better facilitate the comparison of results (see 5.1.8 Cross-Genre and Cross-Language STM Results).

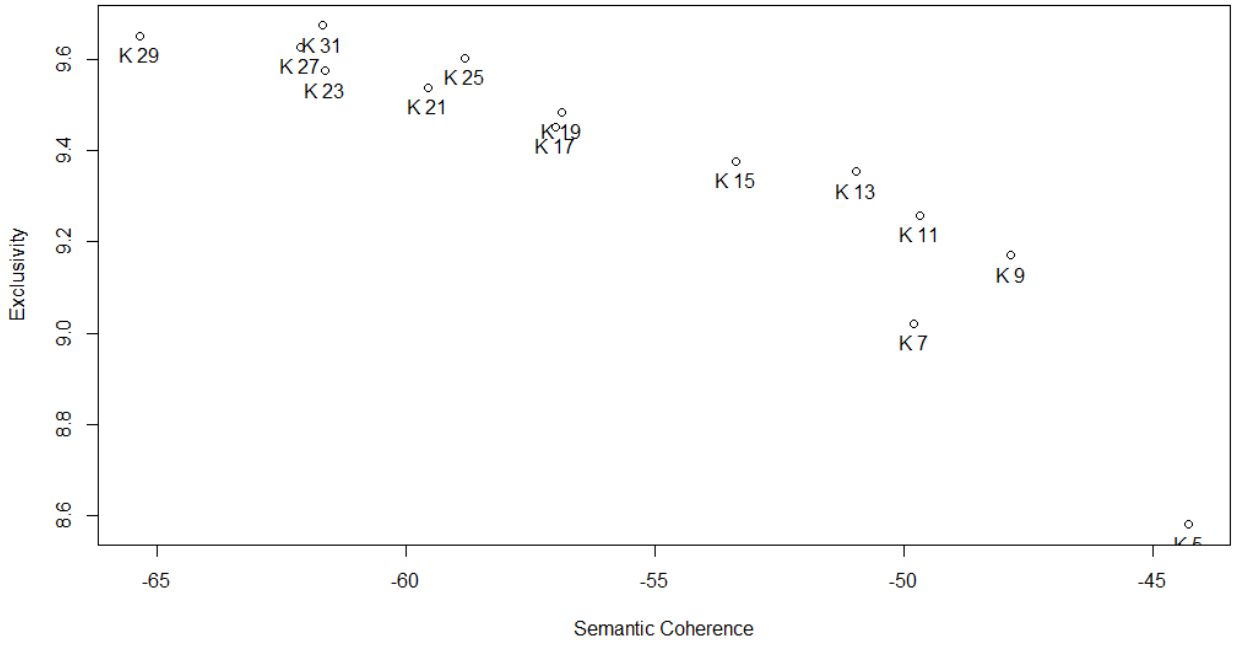


Figure 5.9 The coherence and exclusivity plot of cross-language analysis

Most of the topics appear semantically meaningful, and there is little overlapping or repetition across topics. Table 5.11 contains all the cross-language topics and their highest probability and FREX score topic words.

5.1.7.2 Cross language News STM K11 results

	Highest Prob	FREX
Topic 3	data, privaci, consum, protect, provid, secur, inform	ccpa, gdpr, framework, shield, subdivis, solut, complianc
Topic 7	inform, person, privaci, protect, right, public, regul	supervis, claus, strengthen, china, stipul, punish, ministri
Topic 2	said, privaci, govern, state, court, right, bill	snowden, ecpa, leahi, drone, warrant, fisa, liberti
Topic 8	privaci, compani, data, facebook, googl, user, inform	googl, analytica, zuckerberg, facebook, amazon, coppa, uber
Topic 9	record, system, inform, offic, privaci, feder, notic	docket, foia, citat, alexandria, sorn, patent, routin
Topic 6	user, mobil, phone, secur, data, privaci, softwar	teddi, softwar, tencent, mobil, virus, trojan, bear
Topic 11	privaci, photo, report, court, public, famili, case	divorc, husband, marriag, wife, celebr, girlfriend, kong
Topic 4	inform, report, phone, person, number, card, bank	bank, card, ticket, shop, merchant, crack, wechat
Topic 5	think, know, peopl, like, want, thing, time	clip, tonight, inaud, malveaux, yeah, somebodi, crosstalk
Topic 1	camera, student, school, live, parent, privaci, children	taxi, classroom, passeng, camera, teacher, student, broadcast
Topic 10	express, health, patient, medic, inform, hospit, deliveri	patient, medic, hospit, courier, deliveri, diseas, doctor

Table 5.11 Topic words of cross-language news analysis K11

[Consumer data protection and law] Topic 3 is about consumer data protection and law, indicated by topic words like “data”, “consum”, “protect”, “ccpa” (California Consumer Privacy Act), and “gdpr” (General Data Protection Regulation), “shield”[27], and “complianc”.

[27] “Shield” refers to the Privacy Shield, it is a framework designed by the U.S. Department of Commerce and the European Commission and Swiss Administration. Privacy Shield provides companies on both sides of the Atlantic with a mechanism to comply with data protection requirements when transferring personal data from the European Union and Switzerland to the United States in support of transatlantic commerce.

[Privacy regulation] Topic 7 is about privacy protection in general, indicated by words like “privaci”, “protect”, “right”, “public”, “regul”, “claus”.

[Regulation and government surveillance] Topic 2 is about two subtopics, one is about the regulation of privacy, indicated by words like “govern”, “right”, “court”, and “bill”. And the other sub-topic is about government surveillance, which is indicated by topic words like “snowden”, ecpa (Electronic Communications Privacy Act), and fisa (Foreign Intelligence Surveillance Act).

[Tech companies and children’s privacy] Topic 8 is about privacy issues related to several big tech companies, including “facebook” (including its related entities like “analytica”, “zuckerberg”), and “google”, “amazon”, and “uber”. This topic also touches on the privacy of children, indicated by “coppa” (Children's Online Privacy Protection Act).

[Government record privacy] Topic 9 is about privacy concerns over record and information in American government and federal agencies, indicated by topic words like “record”, “system”, “feder”, “foia” (Freedom of Information Act), and “sorn” (System of Records Notices).

[Mobile phone and privacy service] Topic 6 is about privacy regarding the mobile phone, indicated by topic words like “mobil”, “phone”, and related mobile phone data protection services like “teddi”[28].

[Interpersonal relationship] Topic 11 appears to be about interpersonal relationships, which is indicated by topic words, including “famili”, “husband”, “wife”, and “gossip”, etc.

[Phone, personal finance] Topic 4 is about two privacy-related subtopics, including phone, indicated by words including “phone”, “number”; and personal finance, indicated by words like “bank”, “card”, “shop”, “merchant”, “seller”.

[Mobile phone and privacy service] Topic 6 is about privacy regarding the mobile phone, indicated by topic words like “mobil”, “phone”, and related mobile phone data protection services like “teddi”.

[TV and celebrities] Topic 5 is about privacy issues related to American TV and radio celebrities, indicated by topic words like “malveaux”, “imus”, “tonight”.

[Camera, student, taxi] Topic 1 appears to concern three subtopics, one is about student, indicated by topic words like “student” and “teacher”, “school” and “classroom”; the second subtopic is about “camera”; and the third one about “taxi”, and “passeng”.

[28] Teddi is a name of a company that provides mobile phone data protection services.

[Healthcare and express] Topic 10 is primarily about privacy in two specific domains/industries, one is healthcare, indicated by topic words including “medic”, “patient”, “health”, “hospital”, “clinic” etc; the other is delivery, indicated by topic words like “express”, “courier”.

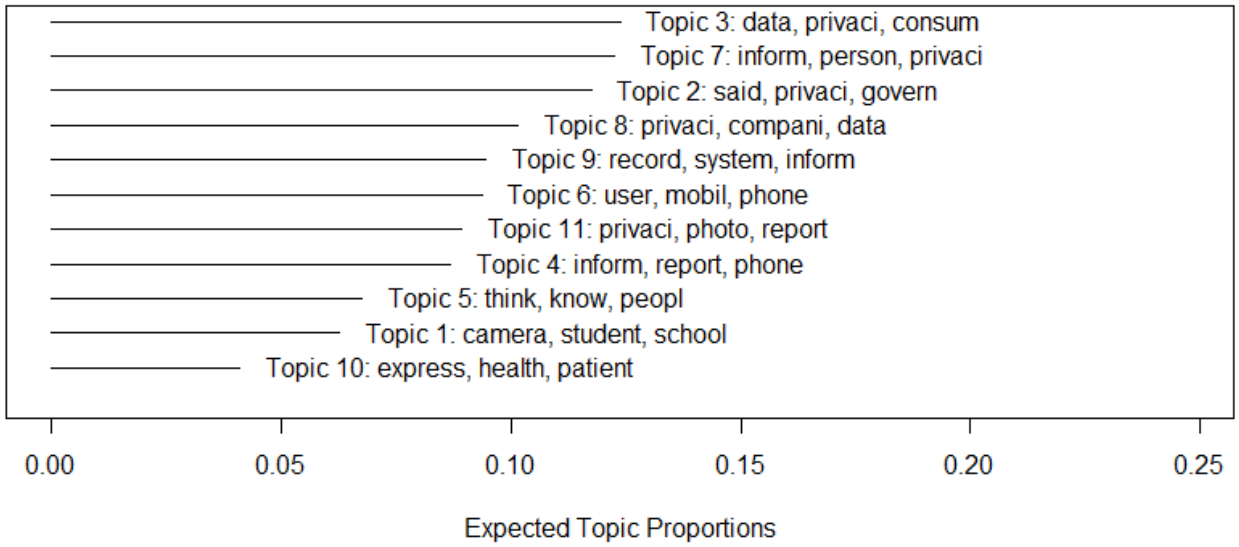


Figure 5.10 The plot of the topic proportions for cross-language analysis K11
(displaying also the top 3 FREX topic words)

The leading topics of this cross-language news analysis are concerned with issues about privacy regulation and law, and the role the government plays in the protection of privacy, as indicated by the top leading topics in news STM analysis results like Topic 3, Topic 7.

Topic 3 demonstrates an increase of proportion over time in the English language, while the increase in the Chinese language is less significant (see Figure 5.12 left panel). For Topic 7, the proportion increase in the Chinese language is greater than in the English language (see Figure 5.12 right panel).

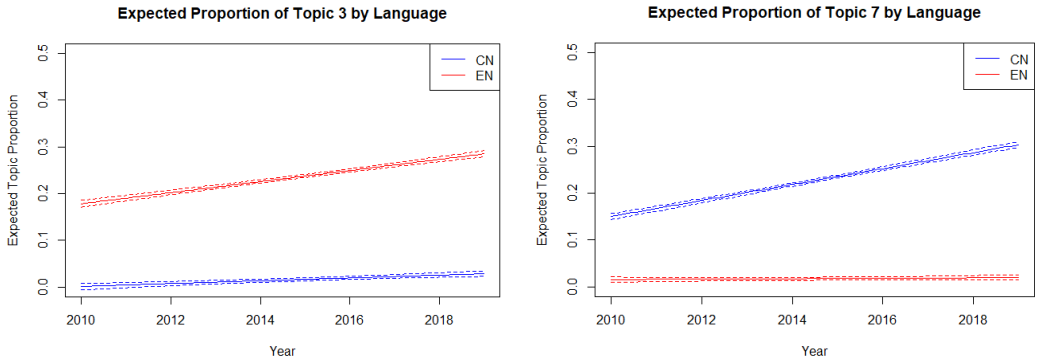


Figure 5.11 Topic 3 & 7: proportion over time by language (cross-languages analysis K11)

A distribution of topic words by language revealed that the topic words for Topic 3 for the Chinese language appear to focus more on personal finances (indicated by topic words like “bank”, “credit”); whereas the English language appears to be about consumer privacy in general (indicated by topic words like “consum”, “privaci”) (see Figure 5.13). Considering the increasing trend of topic proportion for the English language in Figure 5.12 (left panel), this could suggest that consumer data protection has become increasingly important for understanding privacy in the past few years in both languages; however, in the American English language, the focus is more on providing general protection, whereas the focus in the Chinese language is on protecting financial related consumer information.

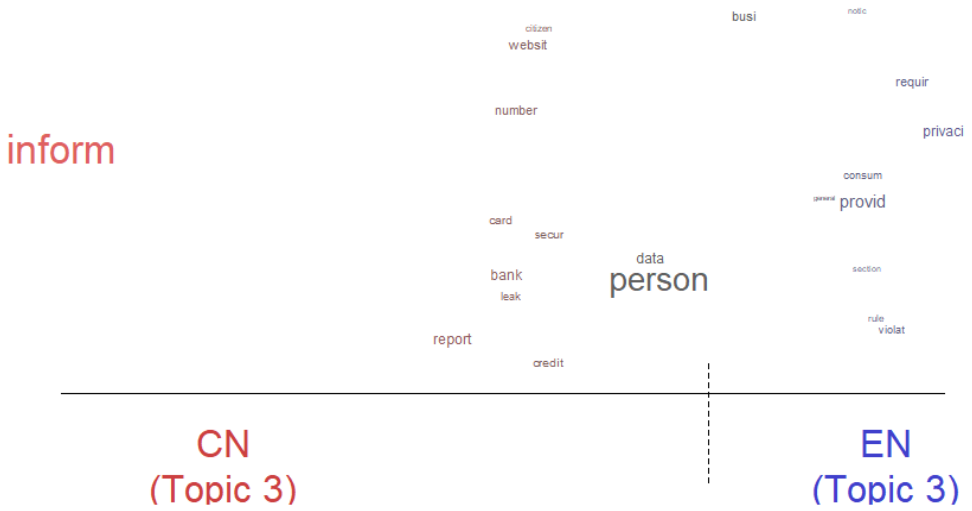


Figure 5.12 Topic 3: topic words variation by language (cross-languages analysis K11)

The topic words for Topic 7 for the Chinese language appear to focus more on the medical and healthcare (indicated by topic words like “medic”, “hospit”, etc), whereas the English language appears to

be more about student data (indicated by topic words like “student”, “school”, “children”, etc.) (see Figure 5.14). Considering the increasing trend of topic proportion for the Chinese language in Figure 5.12 (right panel), this distribution of topic words by language suggests that, although both languages concern privacy regulation, there is an increase in attention to health related data protection issues in the Chinese language. In contrast, the protection of students' education has remained an important issue in the English language.

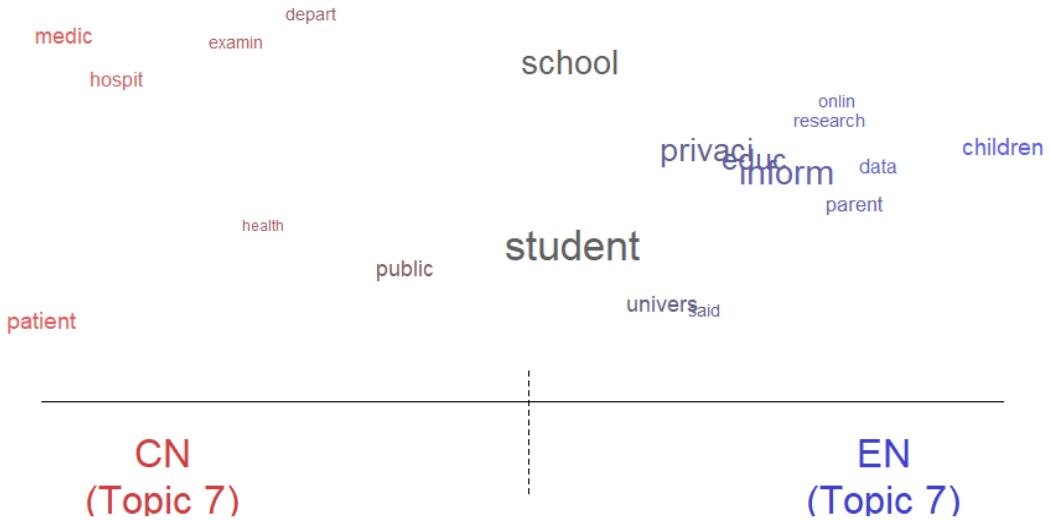


Figure 5.13 Topic 7: topic words variation by language (cross-languages analysis K11)

5.1.8 Cross-Genre and Cross-Language STM Results

The data used to conduct this cross-genre and cross-language analysis is a sample of all data, which consisted of 8,450 documents for social media corpus for both of the languages; and 10,000 documents for news articles for both of the languages. A topic model was built using a total of 36,900 documents, and 37,334 terms.

This decision to use only a sample of data rather than all the data is based on the goal of creating a balanced corpus across genre and language; i.e., to create a corpus that has the same or at least similar amount of documents across genre, across language, and across each of the ten years. Hence, the size of the smallest corpus (Weibo corpus for 2010 that has 8,450 documents) determined the amount of the other corpus .

5.1.8.1 Cross-language cross-genre News STM K diagnostics

The exclusivity and coherence plot (see Figure 5.15) indicated that for the *hyperparameter K*,11 and 13 were close in their performance. K11 was selected for it is slightly better in coherence. Compared to the previous analysis done using just one genre in one language where topics are mostly distinctive to each other and with little overlapping across topics, topics generated from this cross-genre and cross-language corpus were repetitive (for example, Topic 10 and Topic 2, and Topic 8 and Topic 3). In addition, several topics appear to be difficult to understand in terms of how they are semantically meaningful to privacy. In general, cross-genre and cross-language results are the least intuitive among all models. Table 5.12 contains all the cross-genre cross-language topics and their highest probability and FREX score topic words.

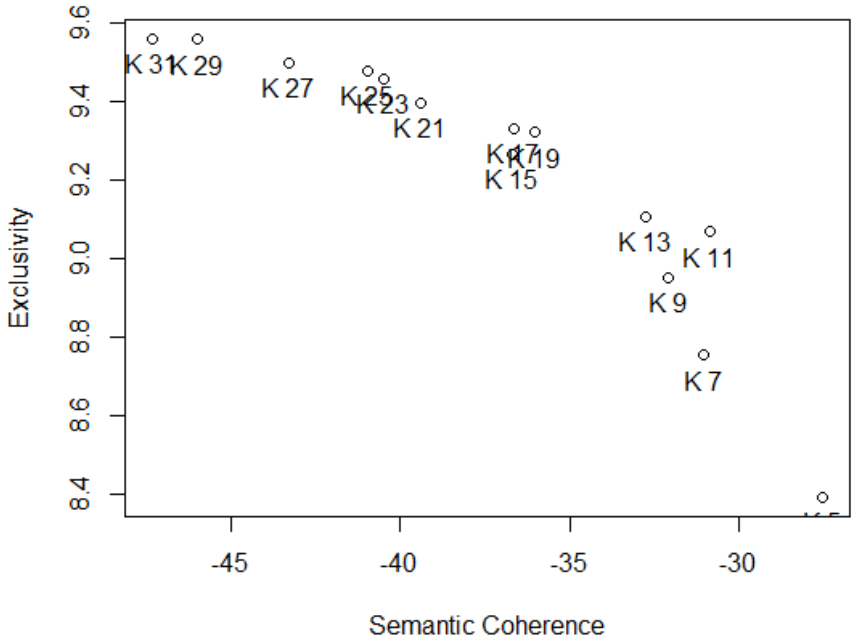


Figure 5.14 The coherence and exclusivity plot of cross-genre cross-language analysis

5.1.8.2 Cross-language cross-genre STM K11 results

	Highest Prob	FREX
Topic 8	privaci, facebook, data, secur, protect, like, googl	fami, socialmedia, probab, eeolshjjmv, dataprotect, presidenttrump, yall
Topic 10	privaci, link, person, talk, phone, love, mobil	villain, gemini, taurus, michao, ari, capricorn, disk
Topic 2	inform, user, mobil, phone, secur, person, data	teddi, mobil, softwar, bear, leakag, vulner, malici
Topic 4	data, privaci, consum, protect, inform, secur, provid	fairwarn, framework, harbor, ecpa, ccpa, stakehold, subcommitte
Topic 3	user, googl, data, facebook, said, privaci, compani	patent, analytica, abstract, stoddart, inventor, cambridg, trademark
Topic 6	privaci, protect, person, know, want, friend, inform	zhengt, lehua, gome, blockchain, ecard, onlook, artist
Topic 1	person, public, inform, right, privaci, student, protect	stipul, shall, subdivis, judgment, patient, municip, tort
Topic 5	camera, express, privaci, live, report, photo, children	taxi, courier, meow, pipe, meter, slip, passeng
Topic 7	said, peopl, think, know, like, year, want	malveaux, inaud, clip, clinton, videotap, gutfeld, unidentifi
Topic 11	privaci, internet, technolog, social, right, data, public	recognit, artifici, facial, census, genet, hong, scienc
Topic 9	record, system, inform, offic, feder, privaci, agenc	docket, sorn, routin, notic, supplementari, citat, submiss

Table 5.12 Topic words of cross-genre cross-language analysis K11

[Facebook and Google] Topic 8 is mainly about privacy over two American tech companies Facebook and Google. This topic appears to be overlapping with the other topic (Topic 3).

[Mobile phone] Topic 10 is broadly about mobile phone, which is indicated by topic words like “mobil” “phone”, similar to Topic 2.

[Mobile phone] Topic 2 is about privacy concerns over mobile phones users, and services that help protect mobile data, which are indicated by topic words like “user”, “mobile”, “phone”, “data”. “Teddi”, etc.

[Consumer protection laws] Topic 4 has two privacy-related subtopics: one is about consumer data and privacy protection, indicated by topic words like “data”, “privaci”, “consum”, “protect”. In addition, domestic and international privacy laws and regulations, indicated by words like “ecpa” (Electronic Communications Privacy Act), “ccpa” (California Consumer Privacy Act) and “harbour”.

[Facebook and Google] Topic 3 appears to be mainly concerning user privacy associated with big tech American companies, indicated by topic words like “user”, “data”, and “facebook” (and “cambridg” “analytica”), “google”, and “compani”. Another subtopic that can be seen in this topic is intellectual property, indicated by words like “patent” and “trademark”.

[Miscellaneous] Topic 6 appears to be lacking one coherent theme; however, it does touch on a couple privacy-related themes, including “blockchain”, and a Chinese entertainment company “lehua”.

[Specific population and regulation] Topic 1 is mainly about the privacy of specific populations, indicated by topic words like “student”, and “patient”. In addition, this topic also appears to touch on privacy regulation, indicated by topics words like “right”, “stipul”, “judgment”, “municip”, and “tort”.

[Privacy concern over daily applications of information] Topic 5 is about multiple daily digital technology applications that involve personal information, including “camera”, and “taxi”, and “express” and “courier”.

[America TV celebrity] Topic 7 is mainly about American TV anchor and journalist, indicated by topic words like “malveaux” and “gutfeld”.

[Artificial intelligence] Topic 11 is to be about privacy concerns over two novel “internet” and “data” applications driven by “artifici” intelligence, including “facial” “recognit”.

[US federal record system] Topic 9 is primarily about privacy concerns over US federal agencies, indicated by topic words like “record”, “system”, “offic”, “feder”, “agenc”, and “sorn” (System of Records Notices).

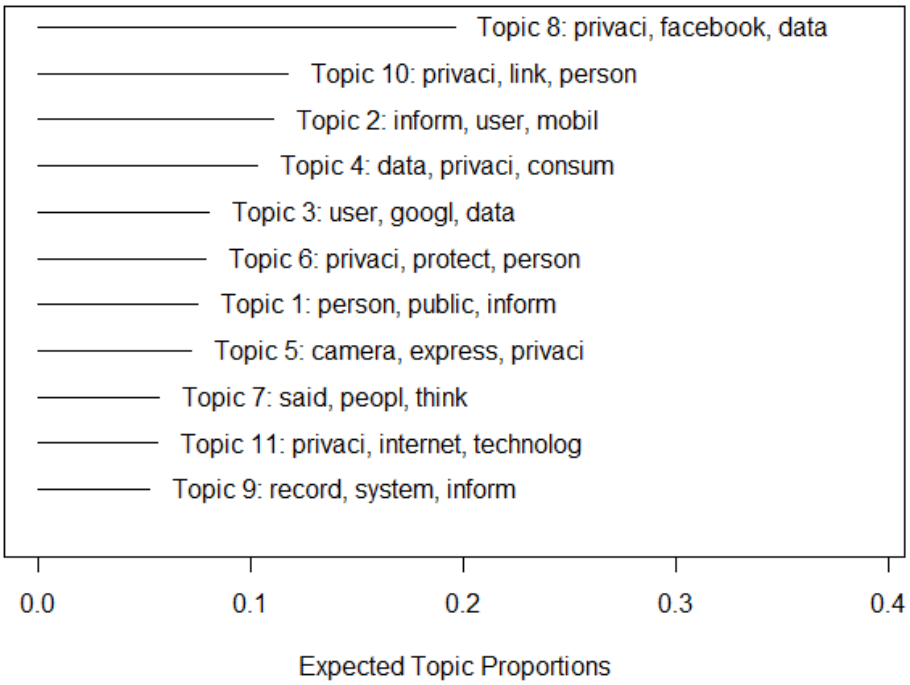


Figure 5.15 The plot of the topic proportions for cross-genre cross-language analysis K11 (displaying also the top 3 FREX topic words)

Topics that are leading in proportion among the 11 topics are Topic 8, Topic 10, Topic 2, and Topic 4. These four leading topics can be further divided into two groups: those that concern individuals' everyday activities (i.e., mobile technology); and those that concern organizations and regulations of privacy (i.e. the two big tech companies, and a few privacy regulations).

5.1.8.3 A correlational analysis cross-language, cross-genre, and cross-time

The correlational analysis aimed to understand how topics correlate with language, genre, and time (see Figure 5.17 and Figure 5.18).

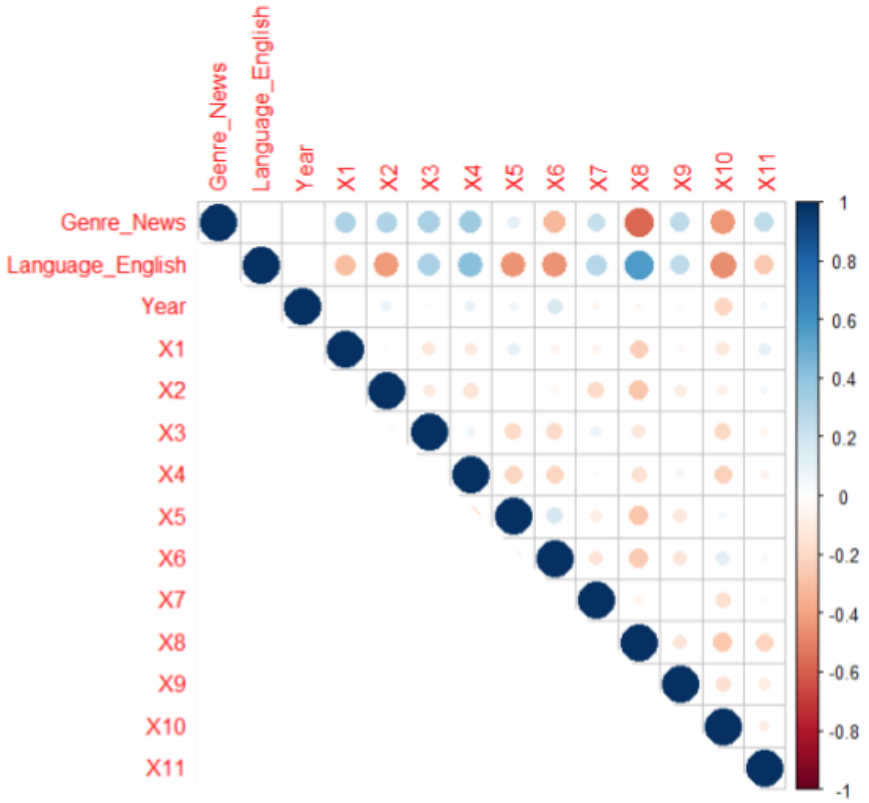


Figure 5.16 Pearson correlations between prevalence variables and proportions of topics (News_EN) (X1 indicates Topic 1, etc.)

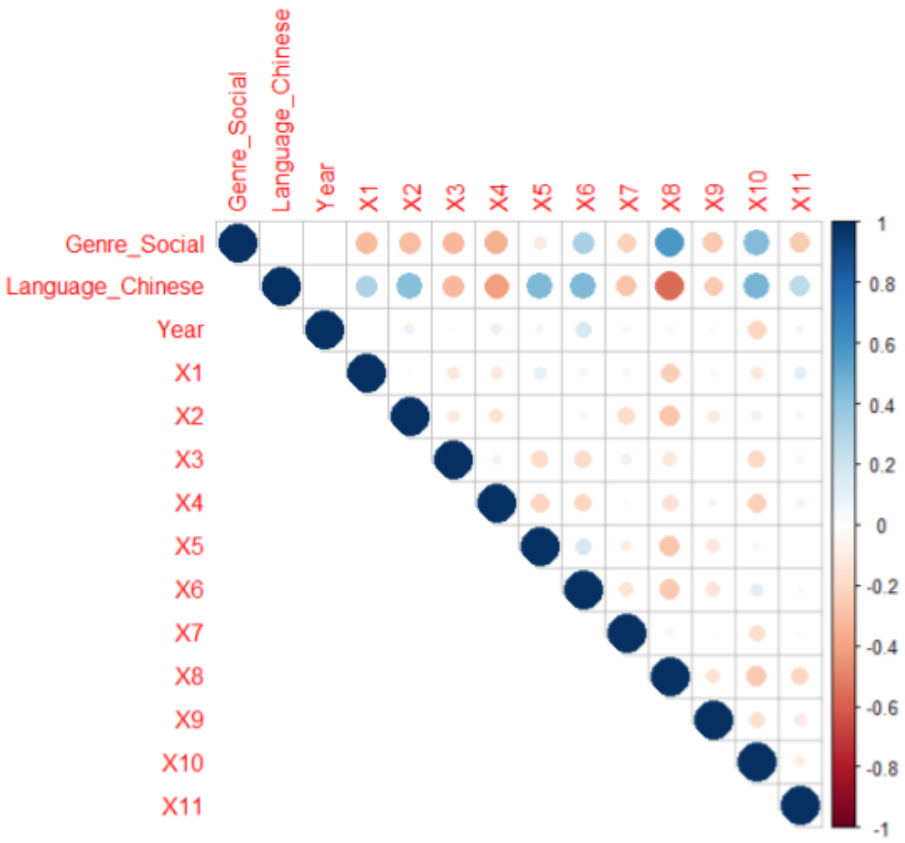


Figure 5.17 Pearson correlations between prevalence variables and proportions of topics (Social_CN) (X1 indicates Topic 1, etc.)

This correlation by language analysis did not reveal new results compared to results from the previous stand-alone analysis of each of the two languages. The leading three topics that were positively correlated with the English language were digital companies and privacy laws. Topic 8 and Topic 3 are both about Facebook and Google, Topic 4 is about privacy laws (for example, the California Consumer Privacy Act, CCPA).

In comparison, topics that lead in having a positive correlation with the Chinese language include topics 2, 5, 6, and 10. Topics 2 and 10 are both about mobile phones, which is one of the prevalent and frequent topics that have been found when analyzing the Chinese language corpus alone. Topic 5 and Topic 6 appear to be referring to some of the themes unique to the Chinese social context. Topic 5 concerns several very practical oriented privacy themes, including privacy concerns over camera, express and courier, children, and taxi and passenger, which are all themes that have shown up in Chinese

corpus analysis. Topic 6 contains some of the topic words that are unique to the Chinese context, in particular privacy concerns over Chinese celebrities.

This correlation by genre analysis did not reveal new results compared to results from the previous stand-alone analysis of each of the two languages for each genre. For example, Facebook and Google have been seen appearing as leading topics across both genres in English in previous standalone analysis; this is also seen in this correlation analysis.

More specifically, Topic 3 and Topic 4 show the strongest correlation with the news genre. Topic 3 refers to tech American companies Facebook and Google. Topic 4 refers to privacy regulations. Topics that are positively correlated with the social genre appear to be Topic 8 and Topic 10 which align with the earlier standalone analysis of both Weibo and Twitter, as many of the topic words in Topic 8 and Topic 10 are from social media. Topic 8 appears to be mainly about privacy over two American tech companies Facebook and Google. And Topic 10 appears to have one privacy-related theme, which is “mobil” “phone”.

Examination of the exemplar documents shows that Topic 2 is derived from the Chinese news corpus while Topic 10 was from the Chinese Weibo corpus, which explains why one is positively correlated with the news genre, and the other with social media genre.

The correlation with time analysis revealed weak correlations; it also suggested some interesting but conflicting trends. One topic that is positively correlated with time is Topic 6 Miscellaneous. The fact that Topic 6 contains words like “blockchain” and words that refer to the Chinese entertainment company “lehua”, could be part of the reason this topic is positively correlated with time, especially in the most recent couple of years.

The topic that appears to be negatively correlated with time is Topic 10 Mobile phone. The negative correlation could suggest that the topic is becoming less talked about over time. In that over time, people may have become used to mobile phones and related issues and hence no longer discuss it as a trendy issue on social media in particular. This observation adds one more layer of interpretation for understanding the presence of topics. In that, though overall mobile phones related topics remain important for understanding privacy, their presence in different corpus genres over time have been going through more subtle changes.

The declining trend of Topic 10 is revealed better through the change of topic proportion chart by language and by genre (see Figure 5.19). First, aligning with the previous standalone topic analysis by

language (See Chapter 5: Results), mobile phone as a topic indeed is more prevalent in the Chinese language than in English (see Figure 5.19 left panel). Moreover, mobile phone in the Chinese language show a clear decline (see Figure 5.19 left panel), especially in the social media (Weibo) genre (see Figure 5.20 right panel).

However, this decline of Topic 10 alone does not necessarily mean the overall decline of mobile-related themes when it comes to understanding privacy in the Chinese language context because an opposing trend has been observed in Topic 2 (see Figure 5.20).

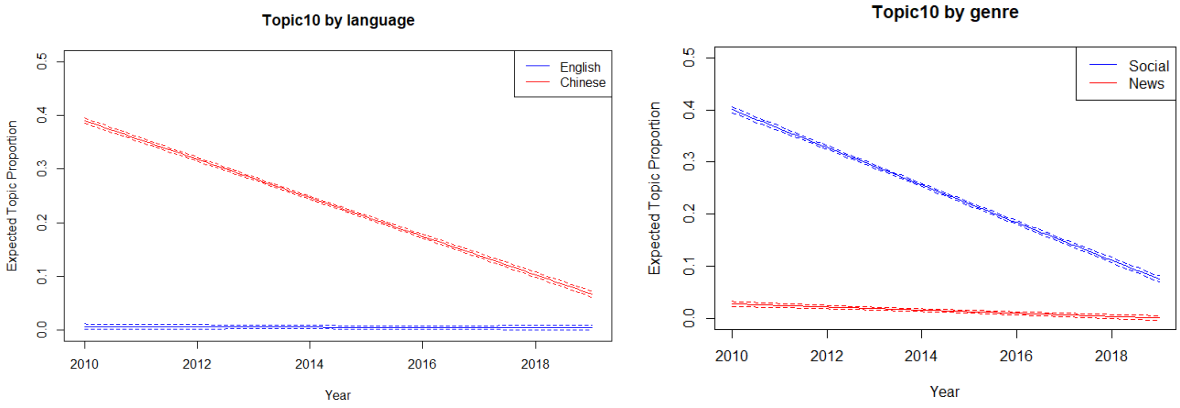


Figure 5.18 Topic 10: proportion over time by language and genre
(cross-genre cross-language analysis K11, with 99.9% confidence intervals)

Topic 2 over time, demonstrates an increase of prevalence in the Chinese language (see Figure 5.20 left panel). In addition, it shows an increase of proportion in specifically the news genre (see Figure 5.20 right panel).

The contradicting trends of Topic 10 and Topic 2 could be explained after an examination of their exemplar documents. Topic 10 appears to be primarily derived from the social media genre (namely, Weibo), which, when put together with the fact that Topic 2 is showing an increase in the News genre (see Figure 5.20 right panel), suggests a more complicated picture for understanding the trend of mobile phone related topics. Mobile phone related topics may have been evolving from being merely present in casual social media posts that were created by random users, to have become a theme that is increasingly recognized and reported by formal genres of language. In other words, mobile phone related topics are moving away from being covered by the social media genre to more being covered by the news genre.

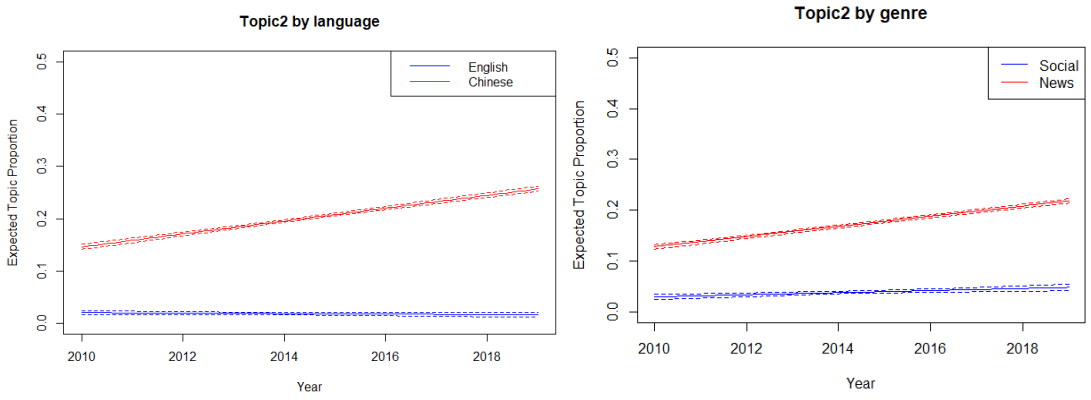


Figure 5.19 Topic 2: proportion over time by language and genre
(cross-genre cross-language analysis K11, with 99% and 99.9% confidence intervals respectively)

As the number one leading topic in terms of topic proportion, Topic 8 Facebook and Google does show a *slight* decline of topic proportion in the English and Social media genres over time (see Figure 5.21). This suggests that Facebook and Google, over a decade, remain as quite central concerns when it comes to understanding privacy in the English language context in general.

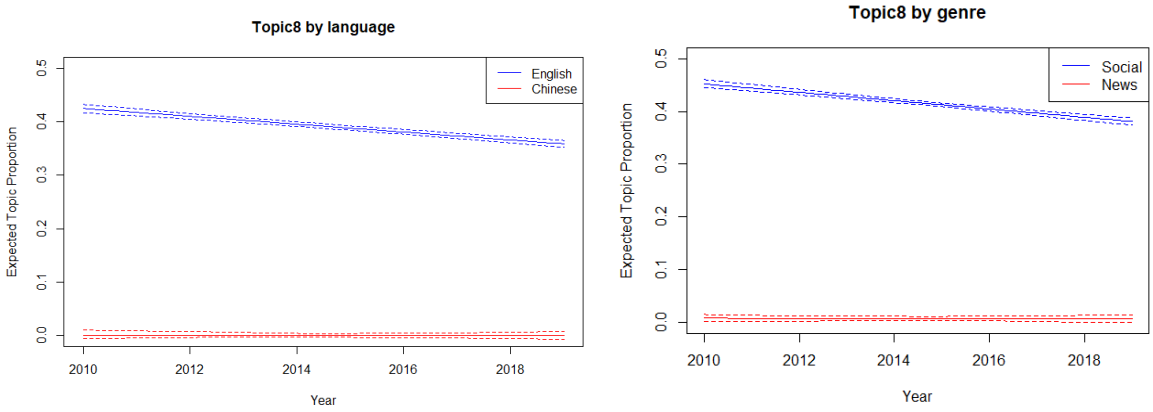


Figure 5.20 Topic 8: proportion over time by language and genre
(cross-genre cross-language analysis K11, with 99.9% confidence intervals)

Topic 6 Miscellaneous (see Figure 5.22) has shown an increase over time in terms of topic proportion for the Chinese language and the social media genre. There are multiple themes embedded in this topic, including some of the most cutting edge technology applications like “blockchain”, and the Chinese entertainment company name “lehua”. This presence of topic words that are associated with

Chinese entertainment company “lehua” suggests that this topic is primarily derived from Chinese social media corpus (Weibo), which is confirmed by a manual examination of the exemplar documents (see in APPENDIX C.2 for exemplar documents of Topic 6).

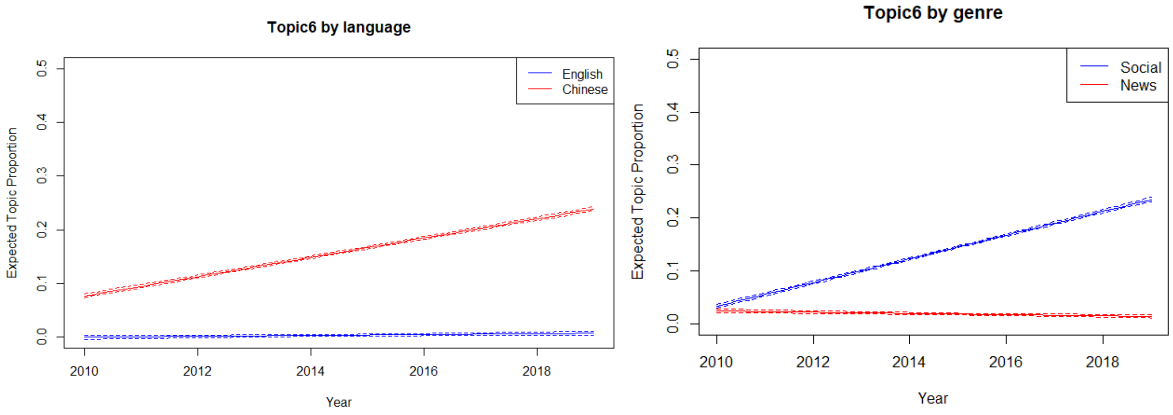


Figure 5.21 Topic 6: proportion over time by language and genre
(cross-genre cross-language analysis K11, with 99.9% confidence intervals)

Topic 5 Privacy concern over daily applications of information (see Figure 5.23) shows a mild increasing trend over the years in the Chinese language; while the topic proportion in the English language remained stable across the years. When examined by genre, the social media genre revealed an increase of proportion over time. When considered together, this by-genre and by-language topic change over time suggest that Topic 5 has been experiencing increasingly more exposure and presence in Weibo. The themes included in this topic touch on some of the most popular digital and technological applications that concern everyday life of a vast population, including “taxi” (possibly taxi-hailing), and “express” and “courier”, it is not surprising that these themes are increasingly discussed on Chinese Weibo.

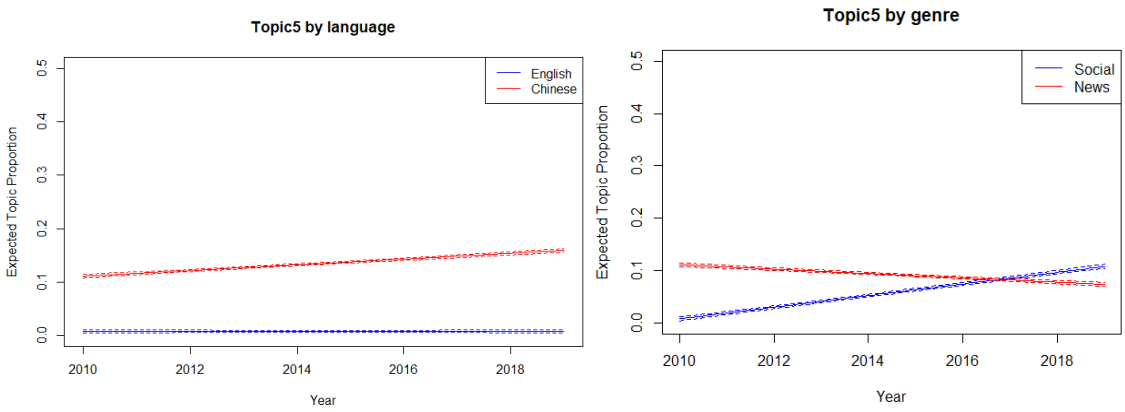


Figure 5.22 Topic 5: proportion over time by language and genre
(cross-genre cross-language analysis K11, with 99.9% confidence intervals)

5.2 Semantic Network Analysis

The semantic networks analysis (SNA) was conducted using a sample of 1,000 news articles in each language, from each year. Specifically, for the Chinese news, the translated corpus was used (instead of the original Chinese content). This decision was made for two reasons: first, to maintain consistency with the use of translated documents for the two methods of STM and SNA for enabling comparison. Second, to facilitate the comparison of SNA results between the English and Chinese languages. The SNA analysis was focused on the top 50 words of each language based on their centrality scores. A structural space analysis (see details in Section 4.2) was done by mapping the top 50 words into a quadrant structural space (see Figure 5.24 and Figure 5.25). For a better visualization (i.e. to properly show those words that are clustered together at the bottom left-hand side of the chart, an adjusted chart was also produced which excludes the top 5 words for each of the plots (see Figure 5.26 and Figure.27 for structural space plots that have excluded the top 5 words).

In addition, informed by topic modeling results, I focused on words that are leading topic words in previous topic modeling analysis (see Table 5.13). Specifically, for the Chinese language, “mobil”, “internet”, and “right” (each correspond to a leading topic that was identified from topic modeling results) were selected to understand their evolution of structural roles over the ten years. In addition, “public” and “technolog”, and “compani”, and “netizen”[29] were selected to correspond to the three semantic dimensions (namely, *public*, *technology* and *institution*). For the English language, words were selected by similar principles, including “state”, “google”, “facebook”, “technolog”, “feder”, and “govern”, which correspond to some of the leading topics’ topic words as well as the semantic dimensions (namely, *public*, *technology* and *institution*).

[29] 网民 in Chinese. “Netizen” refers to citizens on the internet. It describes a person who is actively involved in online communities or the Internet.

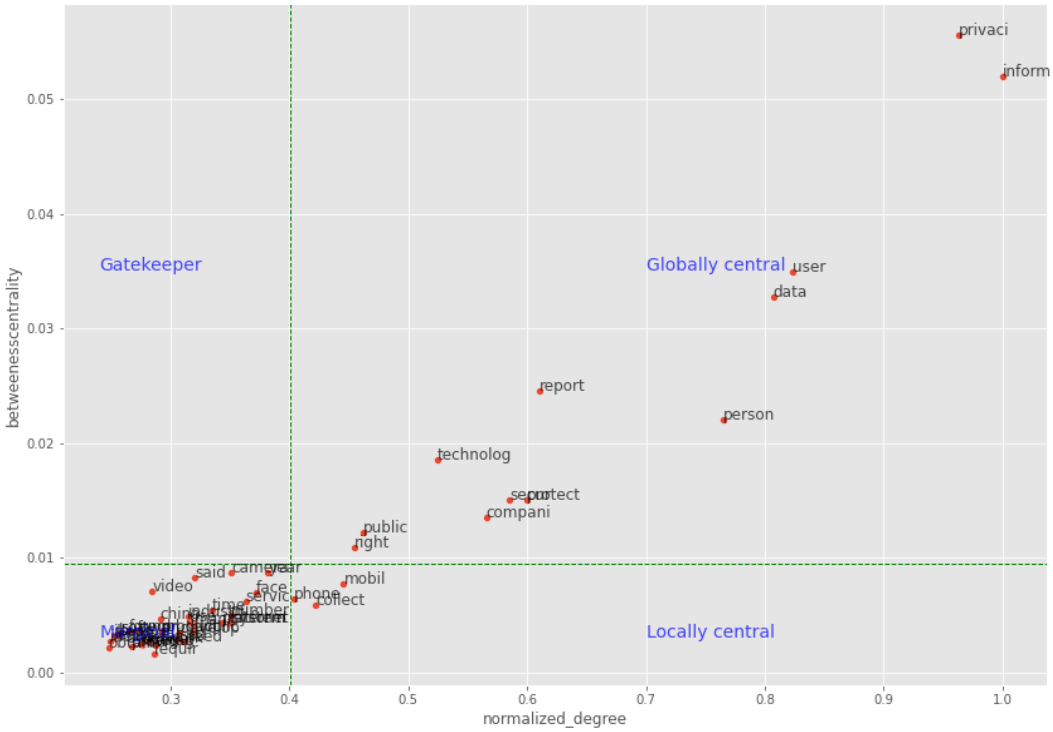


Figure 5.23 Structural space of 2019 Chinese news corpus top 50

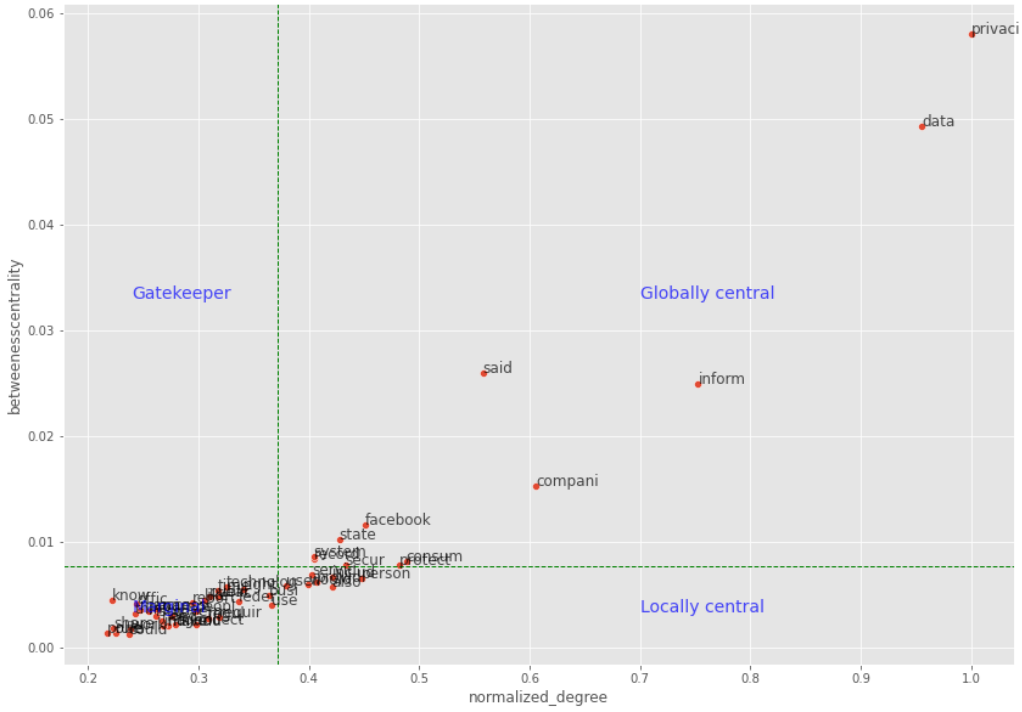


Figure 5.24. Structural space of 2019 English news corpus top 50

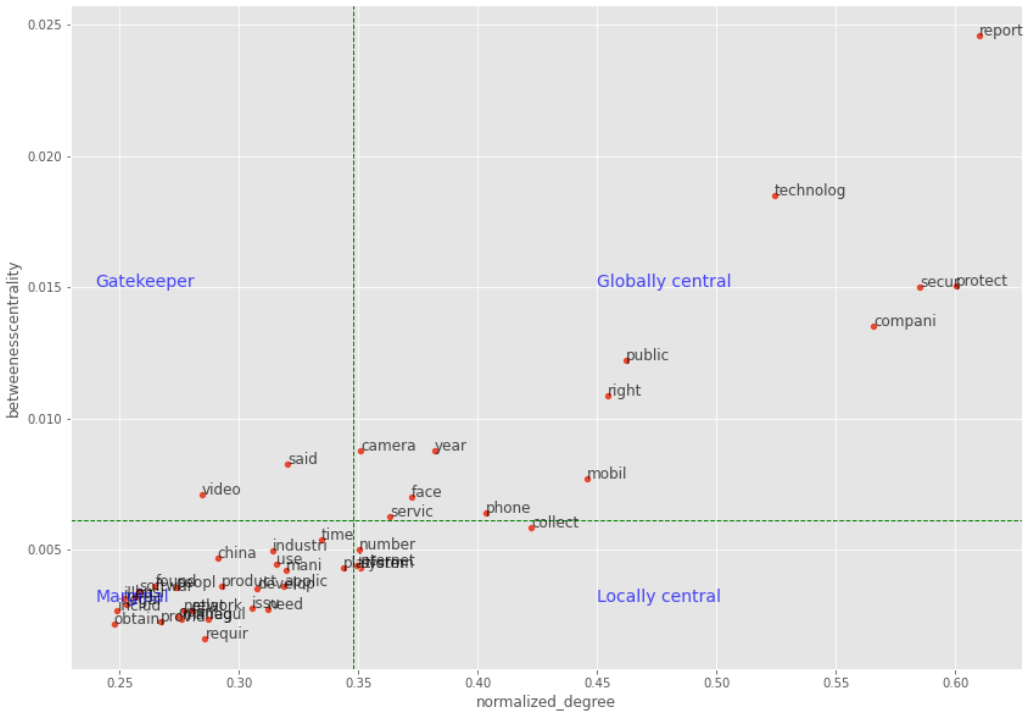


Figure 5.25 Structural space of 2019 Chinese news corpus top 45

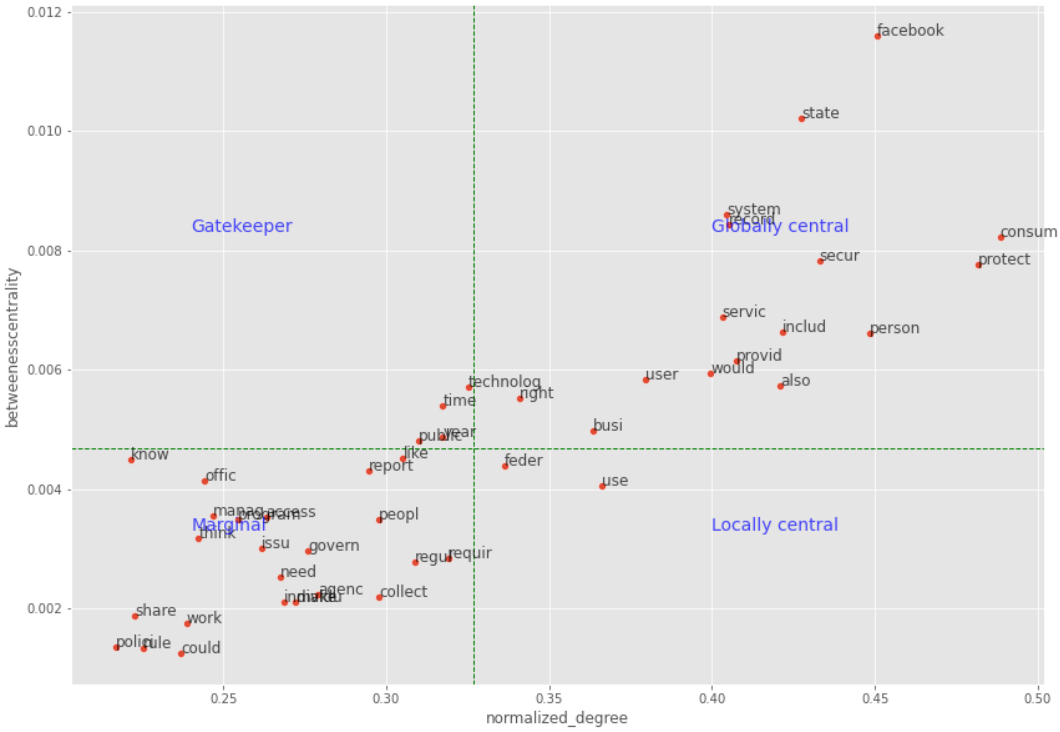


Figure 5.26 Structural space of 2019 English news corpus top 45

Unsurprisingly, the words that lead in both languages are “privaci”, “data”, and “inform”, which is a pattern seen across all years. The two languages across the years both have privacy (“privaci”) as the leading global central node. What appears to distinguish the top globally central words in the Chinese language from thos in English is that the Chinese nodes refer more to concrete technology related applications, including “mobile”, “phone”, and “camera”, and “express”. in contrast, the nodes in the English plot refer to institutions, indicated by nodes including “company”, “system”, “record”, and “service” (see Table 5.13 and Table 5.14 for the top 15 globally central nodes across 2010 to 2019 in the Chinese and English languages respectively). In particular, two digital companies, “facebook” “google” appeared in multiple years among the top English globally central nodes, which is another characteristic that distinguishes the leading nodes between these two languages. In comparison, the presumed Chinese language equivalent digital companies like “tencent”, “baidu” have no presence among the top globally central nodes. This observation on digital companies aligns with the overall characteristic that the English nodes emphasize the institutions of privacy, while the Chinese nodes put emphasis on the individual users.

The Chinese language analysis reveals a strong emphasis on users. In particular, the node “user” appears in 9 out of 10 years’ top 15 nodes. When considered together with the other two nodes that are also present across the 10 years, “mobile” and “phone”, it would be reasonable to suggest that the “users” are invoked in the sense of mobile phone users, which aligns with findings from the topic modeling. In comparison, this strong presence of mobile phone users is *not* seen in the English language. The English structural space analysis aligns with the previous results from topic modeling analysis as well, in that the leading nodes were also the topic words revealed by the topics from the STM, such as “data”, “service”, “facebook”, and “google”.

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
1	privaci	privaci	privaci	privaci	privaci	privaci	privaci	privaci	privaci	privaci
2	said	inform	inform	said	inform	said	data	data	data	data
3	inform	said	said	inform	data	data	said	inform	said	inform
4	data	data	data	data	said	inform	inform	said	inform	said
5	system	compani	compani	state	secur	secur	state	secur	facebook	compani
6	secur	secur	googl	secur	system	state	secur	protect	compani	facebook
7	servic	protect	consum	govern	record	system	protect	state	secur	state
8	record	consum	secur	includ	provid	compani	compani	system	protect	consum
9	compani	servic	protect	protect	state	protect	provid	compani	person	protect
10	right	system	user	right	protect	record	system	servic	state	system
11	facebook	record	system	system	compani	technology	govern	record	servic	record
12	peopl	provide	state	record	servic	would	would	provide	record	secur
13	would	person	would	googl	govern	includ	person	would	provid	person
14	protect	offic	person	compani	would	person	servic	person	user	includ
15	state	state	record	servic	person	govern	record	consum	consum	servic

Table 5.13 Top 15 globally central nodes in English news corpus across 2010-2019

(color coded to highlight the same node across the years)

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
1	privaci	privaci	inform	inform	inform	inform	inform	inform	inform	privaci
2	inform	report	privaci	privaci	privaci	privaci	privaci	privaci	privaci	inform
3	user	inform	person	person	report	report	person	person	user	user
4	report	person	report	report	phone	data	report	user	data	data
5	person	phone	public	public	mobil	person	public	report	person	person
6	internet	public	user	user	person	phone	data	express	protect	report
7	phone	mobil	compani	secur	user	public	secur	phone	report	protect
8	compani	user	phone	compani	public	secur	phone	protect	compani	secur
9	protect	year	secur	phone	secur	mobil	protect	public	public	technolog
10	secur	weibo	protect	protect	data	user	internet	secur	phone	compani
11	public	compani	year	year	year	compani	time	compani	mobil	public
12	softwar	time	mobil	data	protect	internet	peopl	mobil	internet	right
13	year	secur	time	time	time	protect	compani	data	express	mobil
14	time	photo	mani	internet	compani	system	right	camera	year	phone
15	mobile	netizen	system	mobil	right	right	mobil	live	servic	year

Table 5.14 Top 15 globally central nodes in Chinese news corpus across 2010-2019

(Color coded to highlight the same node across the years)

In the English language, some of the nodes that have remained globally central over the past ten years include: “state”, and “compani”. This suggests that the state government (rather than local governments) and various companies remain the most important themes for considering and understanding privacy in the English language context.

Nodes that have either demonstrated a decrease or increase in importance, include, “google” and “facebook”, which both witnessed a decrease of significance around 2013. However, “facebook” suddenly rejoined the globally central cluster since 2018, which may be related to Facebook’s implication with the Cambridge Analytica scandal[30].

In the English language, the “technolog” node has played Gatekeeper roles in multiple years. This gatekeeper role suggests that when it comes to understanding privacy in relation to technology, there are probably multiple different themes for understanding privacy. For example, how specific types of digital technologies may pose a threat to privacy, or technology companies and their accountability for privacy, etc.

[30] Meredith, S. (2018, April 10). *Facebook-cambridge analytica: A timeline of the data hijacking scandal*. CNBC. Retrieved February 16, 2022, from https://www.cnbc.com/2018/04/10/facebook-cambridge-analytica-a-timeline-of-the-data-hijacking-scandal.html

EN	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
state	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
google	G	M	GC	GC	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
facebook	GC	*NA*	G	*NA*	*NA*	*NA*	*NA*	*NA*	GC	GC
technolog	GC	GC	M	M	GC	GC	G	GC	G	G
feder	M	GC	GC	M	M	GC/LC	M	GC	M	LC
govern	LC	M	LC	GC	GC	GC	GC	GC	M	M
compani	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC

CN	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
internet	GC	LC	LC	GC	GC	GC	GC	M	LC	LC
mobil	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
compani	GC	GC	GC	GC	GC	GC	*NA*	GC	GC	GC
public	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
technolog	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	M	M	GC
right	GC	LC	GC	GC	GC	GC	GC	M	GC	GC
netizen	GC	GC	G	M	M	*NA*	*NA*	*NA*	*NA*	*NA*

Table 5.15 The changing structural roles of nodes

(GC: globally central, LC: locally central, G: gatekeeper, M: marginal, *NA*: non-applicable refers to when the node was not among the top 50)

CHAPTER 6: DISCUSSION

“We cannot simply start by asking ourselves whether privacy violations are intuitively horrible or nightmarish. The job is harder than that. We have to identify the fundamental values that are at stake in the "privacy" question as it is understood in a given society.” (Whitman, 2003, p.1220)

In this chapter, I consolidate findings from Chapter 5: Results, and compare and contrast findings from these two languages in the order of the four research questions. I will use the four semantic dimensions as a framework for understanding the major semantic characteristics regarding privacy when compared across these two languages.

6.1 Privacy in the corpora

6.1.1 Privacy in the Chinese corpora

Internet and mobile applications as the major contexts of privacy in the Chinese language

Topics with top proportions in Chinese corpora (Topics 12, 10, and 6 from Chinese news analysis, and Topics 4, 5, 11, and 13 from Weibo analysis) are all associated with some of the most popular internet, mobile applications, and mobile phones. This is a not surprising topic considering that China has the world's largest smartphone user population (Slotta, 2021). These leading topics suggest that the Internet, mobile phones, and mobile applications have become very important contexts for understanding privacy in the Chinese language, and what's driving discussions of privacy are problems of privacy in these mobile technological applications as they get adopted and used by an increasing number of users every day.

In Topic 12 of Chinese news (the topic with the highest proportion), we see words/phrases like "personal information", "internet" among the highest probability topic words (see 5.1.1 Chinese News Results). This suggests that in the Chinese news context, privacy is considered as "personal information" on the "internet", and online protection of personal information is one of the dominating ways of understanding privacy. In addition, privacy in the Chinese news appears to be associated with some of the most popular Internet/mobile applications in China, as indicated by the presence of "Alipay" (which is the one of the leading mobile payment apps), and "Ctrip" (which is one of the leading online booking platforms); both apps have hundreds of millions of monthly active users. The presence of these popular applications among the top topic words suggests that understanding of privacy is likely associated with the use of these apps.

Government leading the protection of consumer privacy in the language of Chinese news

Topic 12 from Chinese news, also includes the leading topic words "China Consumer Association" ("中消协"), "Cybersecurity Week" (which is a recurring annual event hosted by the Cyberspace administration office of China, "宣传周"), and a person's name, "Jianjun Yang" (the associate director of the China Electronics Standardization Institute, "杨建军"). It appears that the *government* and government-driven initiatives are another thread of ongoing privacy discourse in Chinese news. This sense of government-led privacy protection can also be seen in Topic 7, where we see topic words like "Consumer Council" (which is a generic name that refers to municipal level consumer protection

governmental agencies, “消保委”), “100 thousand RMB”, and “1000 RMB” (both refer to penalty fines for companies for breaching consumer privacy, “十万元”, “一千元”) (see APPENDIX C: EXEMPLAR DOCUMENTS FOR TOPIC 7 OF CN NEWS).

The picture that emerges from these two leading topics with increasing prevalence in the Chinese language corpora is that: first, understanding of privacy seems to be centered around major internet applications, and in particular mobile applications. And second, protection of privacy is what the government is pushing companies to do. Additionally, digital technology applications appear to be another emerging theme that's driving privacy discourse, as revealed through top vocabularies in Topic 7; in particular, topic words like “facial recognition”, and “artificial intelligence”.

Privacy at Offline Domains: Surveillance Camera, Health Information, and Court Information in the Chinese news

In addition to the digital or online realm (the Internet and mobile phone) as the major context for privacy, there are a few distinctive offline domains where privacy is concerned: Topic 3 Webcam privacy violation, Topic 4 Medical information privacy, and Topic 9 Court trials of privacy.

Topic 3 centers around “surveillance camera” in various offline “public spaces”, these public spaces include “public gym”, “swimming pool”, “hotel room”, “fitting room” at shopping malls, and “emergency room” at hospitals. Topic 3 forms a contrast with the previous discussion regarding privacy on the internet and mobile applications, in that Topic 3 predominantly concerns the misuse of cameras in various offline environments. In addition to the inappropriate use of cameras, the risk of inappropriate access to the camera recordings are also part of the privacy concern; topic words like “staff”, and “network platform” among those of highest probability. This suggests that discussions regarding privacy are concerned with the practice and operational details, as well as the accountability of the actual persons and organizations managing these devices and content.

Topic 4 is primarily about privacy issues concerning several specific domains. . First is health-related information, as indicated by words like “HIV”, “medical institute”, and “infected”. In addition to health information, Topic 4 appears to be related to employment and college/higher education information. The presence of topic words like “employer”, and “employee/worker” points to the possibility of a situation where a potential employee may have little choice or protection over their information they may be required to disclose to the potential employer. Similarly, “college students” may face disclosure of

personal information during the process of applying for college aid, as suggested by topic words like “impoverished/poor students”.

Topic 9 suggests one more offline domain where privacy is concerned; the information associated with court trials, as indicated by words like “litigant”, “the suspect”, “People’s Court”, “court verdict”, etc; This topic suggests that there are privacy concerns regarding the inappropriate access or disclosure of court information, in particular, court trial information regarding teenagers.

The pervasive theme: privacy of the individual in the Chinese news

In addition to these topics concerning privacy in various contexts and domains, what’s hard to ignore is the presence of the compound word/phrase “individual privacy” (个人隐私) across multiple topics. The presence of the phrase “individual privacy” as one of the topic words with the highest probability score, occuring in 10 out of all 13 topics suggests that there is a prevalent sense of framing privacy as associated with the *individual*s.

The fact that privacy in the Chinese language presents itself in the phrase of “individual privacy” is intriguing. The protection of the *individual* (Lü, 2005) and *personality* (Wang & Xiong, 2021) is relatively new in the Chinese context. In this context, referring to privacy by using the compound phrase *individual privacy* appears to be playing a role in emphasizing that the protection is *for the individuals*. This forms an interesting contrast with the English language where the protection of individual autonomy is inherent or implicit in the meaning of privacy.

Privacy as an interpersonal issue and privacy risks over specific practices and applications

One theme that stood out across topics generated from the Weibo corpus is very specific daily applications or practices that could cause privacy concerns for many people: the real-name registration (in Topic 3), the use of QR codes (in Topic 6), the use of webcams (in Topics 7 and 9), and discussing privacy issues that come up during daily interpersonal communication and interaction (Topic 2).

Ordinary Chinese people may face a dilemma between the old ways of interpersonal interactions that rely on *guanxi* (Hwang, 1987), and the *relatively* new advocacy of *individual privacy*. *Guanxi*, which can be roughly translated as *relationships*, depends on social contexts and many times implies making compromises by the individual in expectation of potential future returns during interpersonal interaction[31]. In contrast, privacy protection needs clear rules about what can and cannot be shared. The interaction of old and new values, namely *guanxi* and *individual privacy*, may present a unique challenge for understanding privacy in light of interpersonal interaction in the Chinese language. Because guanxi stands in contrast to "[c]lear property rights, an independent judiciary, and predictable impersonal enforcement of regulations [which] provide institutional protection that does not depend on the particularistic knowledge of others" (Xin & Pearce, 1996, p.1645).

Situating privacy in the pre-existing social background of *guanxi* helps with understanding the presence of privacy in a range of interpersonal contexts in the Chinese language (including the work context revealed in Weibo *Topic 2 Privacy at work context and interpersonal relationship*, and other interpersonal contexts discussed in Feng (2019b)). This suggests that for many people, despite the increasing advocacy for "individual privacy", when it comes to what to do regarding personal data in specific everyday situations, they still lack clear or established practices about sharing or disclosing personal information and data. People rely on their own sense of judgment and the specific social and/or contextual contingencies to make decisions and take actions when it comes to privacy.

[31] "The cultivation of guanxi involves more than the negotiation of a deal and the usage of customary forms to disguise what might otherwise be recognized as a corrupt and illegal exchange. The exchanges are also used to cultivate and strengthen relationships that are expected to continue. In the process, not only advantages and obligations are obtained, but also some degree of trust (Smart, 1993, p.400)."

6.1.2 Privacy in the English corpora

Governmental organizations and surveillance in the English news

Multiple topics in the English news involve governmental organizations, including government administrative organizations, specialized national security and intelligence agencies, and law enforcement authorities (court).

Specifically, Topic 8 in English news corpus enumerated multiple governmental organizations that maintain online record systems, including SSA (Social Security Administration), HUD (Department of Housing and Urban Development), System of Records Notices (SORNs), USCIS (United States Citizenship and Immigration Services), OSD (The Office of the Secretary of Defense), DOD (Department of Defense), and OMB (The Office of Management and Budget). Topic 4 revealed multiple government surveillance-related topic words, including NSA (National Security Agency), Snowden, and the Foreign Intelligence Surveillance Act (FISA). In addition, Topic 7 in English news is about caution towards privacy invasion by subpoena during court investigation.

Data and privacy-related services, and domestic and cross-Atlantic privacy regulations

There are also multiple topics in the English news analysis that are about privacy-related data and compliance services and frameworks. Topic 10 in English news topics enumerated several online privacy management and data security services, including “fairwarn”, “onetrust”, and “hitrust”; and certification providers, including OTA (Online Trust Alliance), ISACA (Information Systems Audit and Control Association), and the NIST Cybersecurity Framework (NIST CSF).

This presence of privacy and compliance services and frameworks can also be associated with the cross-border and cross-Atlantic privacy regulatory framework because regulations’ compliance requirements prompted the development of the various data and compliance services.

Privacy-related regulations and legislation in the English news spans across domestic and international domains. On the one hand, Topic 9 mainly concerned US domestic legislation and regulation, and related authorities including the Federal Communications Commission (FCC). Similarly, Topic 5 revealed several specialized and state-specific privacy legislations and law enforcement, including COPPA (Children's Online Privacy Protection Act), CCPA (California Consumer Privacy Act), and government organization Office for Civil Rights (OCR). On the other hand, Topic 11 specifically focused on international data privacy frameworks and regulations, including the GDPR, and the privacy shield.

The persisting presence of Facebook and Google in the English language context

Facebook and Google appeared in both the news and social media genres in the English language. For example, in English news topics, Topic 1 is about privacy concerns regarding user data of Facebook and Google. In Twitter topics, Facebook and Google appear in multiple topics (Topic 6, Topic 12, and more). The presence of these two companies suggests that in the English language context, privacy concerns revolve around the technology companies. Furthermore, users and law enforcement seek to hold these technology companies accountable for protection of privacy, data security, and national security.

What's more, Facebook and Google both have popular affiliating applications and platforms (for instance, Instagram of Facebook, and Youtube of Google). And it is these two parent company entities that showed up among topic words rather than any of their affiliating applications, which suggests that the discourse of privacy in the English language views Facebook and Google as accountable *institutions*, rather than just specific digital technology applications (social media, and search engines).

Technology-specific discussions of privacy

The final characteristic of privacy in the English language and in particular in the Twitter topics is about specific technology applications. For example, drones in Topic 4 of English news topics. For Twitter topics: cryptocurrencies in Topic 5, location data in Topic 14, and Encryption in Topic 13, etc.

Topic 4 in English news in addition to mentioning drones, also revealed topic words like "faa" (Federal Aviation Administration) that links the *technology* to the *institution* dimension*,* in particular with governmental organizations. In comparison, technologies as they appear inTwitter topics were reported from the perspective of *individual* users and the corporate organizations.

To summarize, the themes described in this section about privacy in the English language align with comments by Westin in that "... democratic societies value and institutionalize privacy… "(2003, p.432), as we have seen that topics revealed various institutions.

6.2 The core semantics of privacy

6.2.1 Core semantics of privacy in the Chinese language

Overall, four topics are shared across Chinese News and Weibo analysis (see Table 5.5): 1) Privacy related rights, 2) Webcam privacy violation, 3) Network security, and 4) Real-name registration and personal ID information. These four topics constitute the core semantics of privacy in the Chinese language.They suggest that privacy in the Chinese language is portrayed as an issue that concerns individuals' practices involving popular digital technologies, and the application of these technologies.

How the two core topics: Webcam privacy violation and real-name registration and personal ID information, are concerned with an individual's privacy may be easy to understand. To understand the appearance of privacy-related rights properly, we will need to see how privacy issues that arise from interpersonal interactions were dealt with given the existing privacy legal protection in China. This is where cases documented by Feng (2019b, p.70) are informative. Case one: "... the husband circulated the phones of his wife and her lover in their workplaces to humiliate the latter two". Case two: "a high school showed video clips of two students kissing each other at the back of the classroom". And case three: "a social media outlet published a lengthy article about a young woman's adultery activities with a movie star, which lead to the suicide of her father, who suffered financial and psychological pressure from his daughter's adultery".

Feng (2019b) emphasized that, though it appeared that in these cases, the wife, the two students, and the young woman, were all victims of privacy loss, none of the courts supported their claims after they had filed lawsuits in the local courts. What appears at stake in the above cases is a mixture of privacy with other related rights, including reputation and portraiture rights. In reality, there have been many privacy cases that were heard and judged on the basis of the right to reputation (Wang, 2012, p.148). The lack of protection for privacy by the courts can be attributed to the lack of protection by the Constitution. China's Constitution does not add clarity to the protection of privacy: Article 38 and Article 99 define the protection of a few personal rights, including personal dignity, portrait, and reputation (Feng, 2019; Wang, 2012, p.147-148); however, there is no protection of privacy.[32]

---

[32] The Personal Information Protection Law of the People's Republic of China (中华人民共和国个人信息保护法) came into effect in November 2021, which presumably offers stronger protection for privacy. However, as the data used in this study spanned from 2010 to 2019, this is also the period of time that discussions of this study are focused on.

Apart from the mixture of privacy rights and a few related rights, what stands out from the above cases is that privacy remains an issue that is heavily embedded in interpersonal interactions, which makes it important to discuss how the understanding of interpersonal relationships may impact the understanding of privacy in the Chinese context. Specifically, interpersonal relationships in the Chinese context are heavily influenced by the Confucian relational ethics (Ma, 2019a), which is still underlying many aspects of Chinese society. I will come back to this point for more discussion in the subsequent section (see Section 6.3).

To summarize, privacy in the Chinese corpora is revealed as an issue of an individual's daily practice, where individuals face privacy risks from technology use and interpersonal relationships. In addition, though core semantics are present across both genres, Weibo topics revealed more about individuals in interpersonal relationships (for example, Topic 11 Personal information and individual privacy, and Topic 2 Privacy at work context and interpersonal relationship). In comparison, Chinese news topics focus more on individuals as technology useds (for example, Topic 10 Mobile phone).

6.2.2 Core semantics of privacy in the English language

Overall, there are four topics of privacy that are shared across both genres in the English language (see Table 5.10 in Section 5.1.6), they are: 1) privacy concerns related to tech companies; in particular, Google and Facebook; 2) government surveillance and national security; 3) privacy regulations and regulators; and 4) privacy and data security services. Hence, these four topics constitute the *core semantics* of privacy in the English language.

The topics in English news regarding caution against governmental organizations can be further divided into two areas. First, those that caution against specialized intelligence governmental organizations like the National Security Agency (NSA) that are more associated with surveillance and national security. And second, those that concern more administrative governmental entities, like the Department of Housing and Urban Development (HUD), and the Social Security Administration (SSA), where there is less concern with government surveillance, but more concern about records security and protection.

In addition to domestic laws and regulations, topics from English also reveal a theme that concerns international, specifically cross-Atlantic, privacy regulations and frameworks. This finding aligns with a previous study (Zheng & Bashir, 2020) that also examined privacy-related news frames in these two languages.

Privacy in the English corpora, as understood by looking at the core semantics, are mainly associated with *institutions*. Despite that privacy research literature in the English language is often associated with an individualistic, or individual-control oriented conceptualization (Radaelli et al., 2018; Vedder, 1999). It turned out that *Individuals*, whether individuals as the users of technologies, or individuals as consumers of commercial products, do not present *explicitly* as a strong, or at least as strong a theme in the English language in neither genres.

Though core semantics can be seen across both genres, some appear to be more prominent in the English news than in Twitter. These topics/themes are about international privacy frameworks, and doubts and concerns towards U.S. governmental organizations. Caution against U.S. governmental organizations appears as the strongest topic in the English news; it appears in multiple topics, including Topic 4 (Government surveillance and national security) with a topic proportion of about 8%, Topic 8 (U.S. governmental records) with a topic proportion of about 12%, and Topic 9 (Regulations and regulators of privacy) with a topic proportion of about 12%. Similarly, international data protection also appears in

multiple topics in English news analysis, including in Topic 11 (International privacy laws) with a topic proportion of about 7%, and in Topic 5 (consumer data privacy protection) which has a topic proportion of about 10%. In comparison, there are two topics from Twitter that touch on governmental organizations and institutions, including Topic 9 Cyberintelligence (which has a topic proportion of about 9%), and Topic 10 Apps, privacy regulation (with a topic proportion of about 11%).

6.3 Dimensions where the two languages are (in)compatible

Before diving into more details, I will first revisit the framework that I use to interpret the dimensions across these two languages. The four dimensions were inductively proposed to better compare the topics across these two languages. Once the dimensions were finalized (see Section 4.1.4 Identification of the dimensions for more details), I went back to code all the topics. These dimensions are *individual*, *institution*, *public*, and *technology*. See Tables 6.1, 6.2, 6.3, and 6.4 for the coded topics in the two languages.

Individual: citizens or persons in the societies as single individuals. This dimension can be seen in topics that concern individual daily practices and interests, for example, individuals as digital technology users, and individuals in interpersonal relationships.

Institution: formal societal regulations and organizations, including law, corporations, governmental organizations, for example, privacy and related rights and regulations (e.g., GDPR), companies (e.g., Facebook), and governmental organizations (e.g., MIIT).

Public: this refers to the broad sense of social good that extends beyond even the scope of institutions, for example, national security and network security are associated with the broad idea of public interests.

Technology: refers to concrete technological designs and implementations that could impact privacy, for example, mobile phones, smart/wearable technologies, drones, etc.

No.	CN Semantics	Dimension
1	Celebrity and mobile contact privacy	tech | individual
2	Privacy related rights	institution
3	Webcam privacy violation	tech | individual
4	Medical information privacy	individual | public
5	Taxi	tech | individual
6	America and online data protection	institution | tech
7	Consumer data protection, AI	institution | tech
8	Court trials of privacy	institution
9	Mobile phone	tech
10	Smart home appliance data	tech

11	Network security	public | institution
12	Real name registration and personal ID information	individual | institution
13	Privacy at work context and interpersonal relationship	individual
14	Astrology	NA
15	Wechat moments	tech | individual
16	System security	institution | tech
17	Personal information and individual privacy	individual

Table 6.1 Coded semantics of CN corpora

No.	CN Core Semantics	Dimension
1	Privacy related rights	individual | institution
2	Webcam privacy violation	tech | individual
3	Network security	public | institution
4	Real name registration and personal ID information	individual | institution

Table 6.2 Coded core semantics of CN corpora

No.	EN Semantics	Dimension
1	Facebook, Google, Cambridge Analytica	tech | institution
2	Student and patient data	individual | public
3	TV show hosts and journalists	individual
4	Government surveillance and national security	institution | public
5	Consumer data privacy protection	institution
6	Privacy violation of records	individual | institution
7	Law enforcement and privacy	institution
8	US governmental records	institution
9	Regulations and regulators of privacy	institution
10	Privacy and data security services	institution
11	International privacy laws	institution
12	Social media	tech
13	Security and trump	institution

14	Cryptocurrency	tech
15	Privacy policies of Facebook and Google	institution
16	Flash memory	tech
17	Cyberintelligence	institution
18	Apps	tech
19	Facebook, Google and Russia	institution
20	Encryption service	tech | institution
21	Location data	tech

Table 6.3 Coded semantics of EN corpora

No.	EN Core Semantics	Dimension
1	Facebook, Google, Cambridge Analytica	tech | institution
2	Government surveillance and national security	institution | public
3	Regulations and regulators of privacy	institution
4	Privacy and data security services	institution

Table 6.4 Coded core semantics of EN corpora

By using this dimensional coding, I have defined three different levels of compatibility for understanding the semantic dimensions of privacy across these two languages: 1) low compatibility, 2) medium compatibility, and 3) high compatibility. The three levels of compatibility provides a more elaborated comparison across these two languages, so that more detailed discussions of privacy across the two languages become feasible than when just relying on a simplified comparison across the two languages.

Low compatibility: where the topic or subtopic in one language is completely missing in the other language. For example, the “Privacy at work context and interpersonal relationship” topic from the Chinese corpora is completely missing in the English corpora.

Medium compatibility: where the topic or subtopic and their codings partially match in two languages. For example “Wechat moments” is coded as the *tech* and the *individual*, whereas “Facebook, google, Cambridge Analytica” is coded as *tech* and *institution*.

High compatibility: where the topic or subtopics exists in both languages and the topics coding match exactly. For example, “national security” in the Chinese language and “network security” in the English language, both are coded as *public* and *institution*.

Based on these comparison rules, this cross-language comparison of privacy reveals a mixed picture: there are certainly shared topics across these two languages; however, there are also topics that appear central for understanding privacy in one language that are missing in the other. More importantly, even for shared topics, these two languages show different emphasis for understanding these topics.

6.3.1 Low compatibility: individual and institution

Government protecting? Or, government to be protected against?

One of the least compatible areas when it comes to understanding privacy is the government. In the English language, privacy is considered as safeguarding individuals against organizations, and *especially* against government organizations. In comparison, in the Chinese language, the government is acting as a protecting role to safeguard individuals from other ill-behaving companies. Furthermore, this sense of safeguarding individuals against the government is missing in the Chinese language.

There is a strong sense in the English corpora which is about privacy risks associated with the federal governmental organizations, including both general governmental organizations (for example, the Department of Housing and Urban Development (HUD)), and specialized intelligence agency (for instance, National Security Agency (NSA)). For the general governmental organizations, there are concerns over potential data records disclosure. For the special intelligence organizations, there are concerns about state surveillance. When compared, neither of these concerns are seen in the Chinese language, especially the second aspect with regards to state surveillance (Feng (2019b) made a similar observation).

Moreover, the presence of governmental organizations in the English topics mostly center around the federal level government organizations. Topic modeling results did not quite capture topics in the Chinese language that caution against central governmental organizations. However, there has been news coverage in recent years regarding the (mis)practices in and by local Chinese governmental organizations and associated privacy risks (Chen, 2017). None of these local governments related concerns were captured by topic modeling in the Chinese language, which does suggest that privacy concerns towards governmental organizations (even with local governments) are quite peripheral among all topics of privacy in the Chinese language.

The weak presence of caution against governmental institutions as reflected by topics can be traced to the legal and legislation reality in China, in that China's constitution does not protect individuals from potential invasions by the government authorities (Feng; 2019; Wang, 2012).

There are two more related topics that are only present in the English language, they are: privacy and data compliance services (Topic 10 Privacy and data security services, in EN K11), and international privacy protection framework (Topic 11 International privacy laws, in EN K11). In that, the cross-border privacy regulations demand companies to prove their compliance with the regulations. In comparison, the

Chinese topics do not reveal any privacy compliance services, or any international or cross-border privacy regulations.

Individuals navigating on their own: interpersonal relationships, mobile phones, and surveillance cameras

Privacy is considered as an interpersonal issue in the Chinese language context. Under this light, privacy is seen as an individual-to-individual, or person-to-person issue to be dealt with as an interpersonal interaction issue that is highly context-dependent[33]; rather than a concept that is valued because it protects individuals from authorities. In contrast, in the English language context, privacy appears more within the framework of individuals-to-organizations, where organizations can include both commercial companies and governmental entities.

Seeing privacy as an interpersonal matter can be understood under the light of the ethical traditions of Chinese culture, specifically the Confucian role ethics (Ma, 2019a), which: "... originates in and radiates from the concrete family feelings that constitute the relations between children and their elders and the interdependent roles they live" (Rosemont & Ames, 2008, p.1). The Confucian roles ethics expects the person to behave differently depending on the specific context and the person's reflections, recognizing the entitlement of the person is different by their social contexts.

Understanding privacy as an interpersonal relationship issue can be thought of in the English language. For example, privacy can be understood as something that can be used to manage interpersonal interactions. Altman (1981) suggests: "Privacy is conceived of as an interpersonal boundary process by which a person or group regulates interaction with others (p.6)". In addition, Quinn et al. (2019) revealed understandings of privacy in the context of interpersonal interaction and relationships, where people consider interpersonal relationships when thinking about sharing personal information on digital social media platforms. Here, the key to interpreting the presence of understanding privacy as an interpersonal relationship in the Chinese language is that: first, it is revealed as one of the *leading* topics for understanding privacy; second, the interpersonal contexts found in the Chinese language refer to a range of offline interpersonal contexts (see in Section 5.1.2 for Weibo Topic 2 Privacy at work context; see also in Section 6.2.1 for real-world interpersonal interaction examples from Feng (2019b)).

As discussed in Section 6.1, when it comes to what to do regarding personal data in everyday specific situations, many still rely on their own sense of judgment and the specific social and/or contextual contingencies to make decisions and take actions when it comes to privacy. And the unique presence of understanding privacy as interpersonal relationships in the Chinese corpora suggests that the conflict

[33] Similar discussions can be seen in Yuan et al. (2013). In contrast, the value of privacy is to safeguard individual autonomy and is context-independent in the US (Zheng & Bashir, 2020).

between privacy (as a relatively new value) and more traditional social values (like *guanxi*), is seen in the Chinese language, but *not* in the American English.

There are two specific practices that are associated with privacy that are seen only in the Chinese language context. One is the real-name registration, and the other is about surveillance webcams. Both present challenges to privacy that individuals have to navigate on their own because of the lack of institutional protection. The key to interpreting the privacy concerns around these practices is that neither is associated with criticisms or questions explicitly towards the governmental organizations who designed and implemented these specific practices.

The real name registration initially was aimed at exposing only online bloggers' and micro-bloggers' real ID information (Jiang, 2016). However, with the popularization of smartphones and apps like Wechat, and the implementation of the second generation ID cards (Brown, 2008; Tiejun et al., 2010) that are required to be used in various social interactions (for example, when purchasing train tickets) (Wildau, 2017), the real name registration has, in fact, expanded its implementation beyond just people who do online blogging or micro-blogging, and has become a common practice in Chinese society. Though the implementation of real-name registration was complimented for bringing convenience to everyday life, research has revealed a spectrum of users' attitudes and concerns over the real name registration practice (Jiang, 2016) between positive and negative.

Similar to the wide adoption of smartphones and mobile apps, the number of surveillance cameras in China has been growing tremendously. Many are installed in urban areas which are further equipped with facial recognition algorithms (Ricker, 2019). There have been an increasing number of privacy concerns over the implementation of these surveillance cameras. However, the concern is mainly cast towards commercial entities and mal-practices of companies, while little criticism has been heard when it comes to the governments' use of cameras in general, which is an observation that is in alignment with findings from Su et al., (2021).

6.3.2 Medium compatibility: technology

In both languages, we can see topics that refer to digital technologies and their applications that concern many people on a daily basis, such as social media. For example, the topic “Wechat moments” appears in the Chinese corpora, and the topic “social media” in theEnglish corpora.;In both languages, there are concerns over some of the most cutting edge digital technologies, including AI, facial recognition, and cryptocurrencies.

However, in the English topics, Google and Facebook (rather than their popular affiliated apps like Instagram and Youtube) are the two tech companies that are present across multiple topics in both genres. In the Chinese language topics, though Wechat appears multiple times across topics, its parent company, Tencent, is not present among leading words of topics at all. Neither are other digital companies in the Chinese market that could be considered as the Chinese equivalent of the US companies, for example, Baidu,(which is the leading Chinese language search engine). Hence, although the technology dimension is present in both languages, it is present in the English language also in the sense of institution (hence the coding of *technology* and *institution*), whereas in the Chinese language, it is present mostly in the sense of individual technology users (hence the dalo coding of *technology* and *individual*).

In addition to understanding the presence of the technology dimension by the topics that contain digital technology related topic words, this dimension can also be illustrated by looking at the presence of the top *globally central* nodes in these two languages’ structural space analysis. The two digital companies that are frequently seen among topic words, “facebook” and “google”, appear also among the top English globally central nodes. In comparison, the Chinese language equivalents “tencent”, “baidu” did not appear among the top globally central nodes (see Section 5.2 Semantic Network Analysis). The weaker presence of digital technology companies suggests that, in the Chinese language, technology related privacy discussion has less intention to hold these companies accountable for their privacy and data protection practices than in the English language context. The weak presence of the technology companies is in alignment with the overall weak presence of the institution dimension in the Chinese language.

6.3.3 High compatibility: public

The *public* dimension is where the two languages are the most compatible. In the English language. In the Chinese news, Topic 12 refers to cybersecurity (Topic 12); and in the English news, Topic 4 refers to national security; both are coded as *public*.

It is worth mentioning that in both languages, there are topics that refer to privacy concerns over specific population groups (for instance, students and patients); topics that concern privacy with a scope that is broader than the individual. However, the key difference across the two languages is that, in the English language the presence of these population groups was accompanied by specialized privacy protection laws and regulations; for example, the Children's Online Privacy Protection Act (COPPA), and the Health Insurance Portability and Accountability Act (HIPAA). In comparison, privacy concerns over these specific population groups in the Chinese context were present more as a result of concerns people have: current Chinese legislation does not have dedicated protection for different types of information, or different population groups' privacy.

6.4 Semantic features across time

Some fluctuations of topic proportion appear to be associated with specific events (announcement of privacy regulation, etc). In addition, greater peaks of topic proportions are seen more in social media topics than in news topics (see Figures 6.2 - 6.5 for topic proportions 2010-2019 for the two genres and two languages). This is likely because social media are responding to real-world events in real time. Tweets can be used to detect bursts of events (Atefeh & Khreich, 2015). For example, Topic 9 Cyberintelligence in Twitter refers to the Cyber Intelligence Sharing and Protection Act of the year 2011.

The topic prevalence plot over time of the Chinese news corpus revealed that there were peaks of topic proportion between the years 2018 and 2019 in multiple topics (Figure 6.2), including Topic 5 Taxi and Topic 6 America and online data protection; and in particular in Topic 7 Consumer data protection, AI. After reviewing some of the exemplar documents associated with these topics, it appears that the peak of topic proportion was likely associated with some domestic regulation and legislation progress, for example, the Draft for Solicitation of comments for the Regulation on the management of the public security video imaging information system (公共安全视频图像信息系统管理条例征求意见稿) (see APPENDIX C). In addition, 2018 was the year China saw multiple regulations and legislation related to privacy (Luo & Wang, 2021), made by multiple governmental organizations, including the Ministry of Public Security of China's announcement of a new regulation of online information security (translated as Measures of Internet Security Supervision and Inspection by the Public Security Organs) (Zhang, 2018), and the Personal Information Security Specification that went into effect in May 2018 (Han & Munir, 2018). These could have contributed to the peaks between 2018 and 2019.

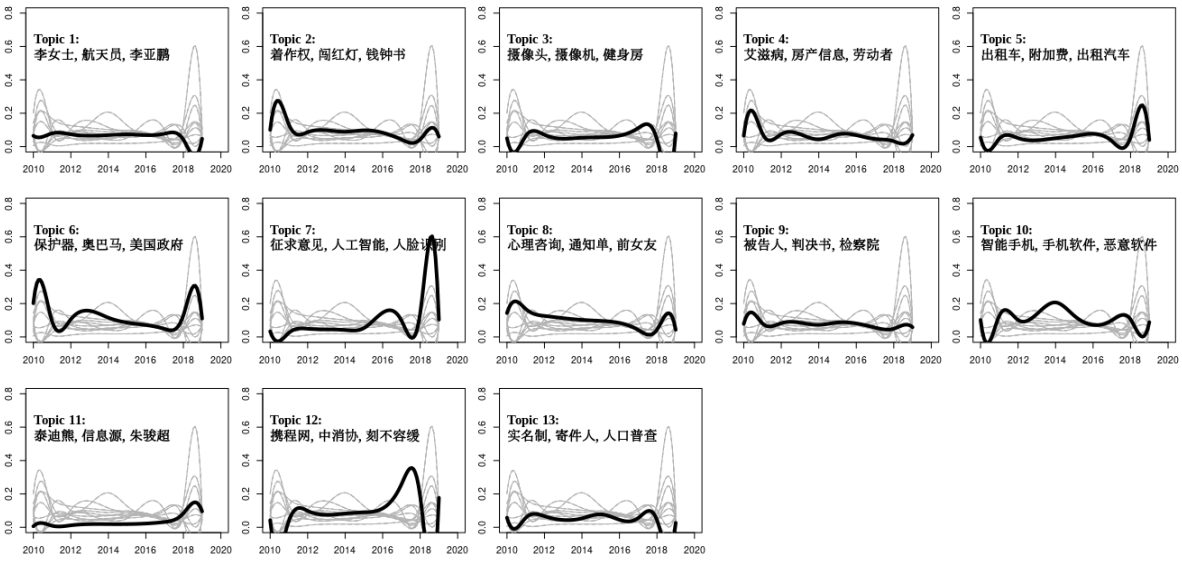


Figure 6.1 A plot of CN K13 topic proportions for 2010-2019

Similarly, for Weibo topics, the most noticeable peak is of Topic 4 Celebrity between 2018 and 2019, which turned out to reflect quite accurately some of the topic words' (including celebrity name and entertainment company) popularity during that period[34].

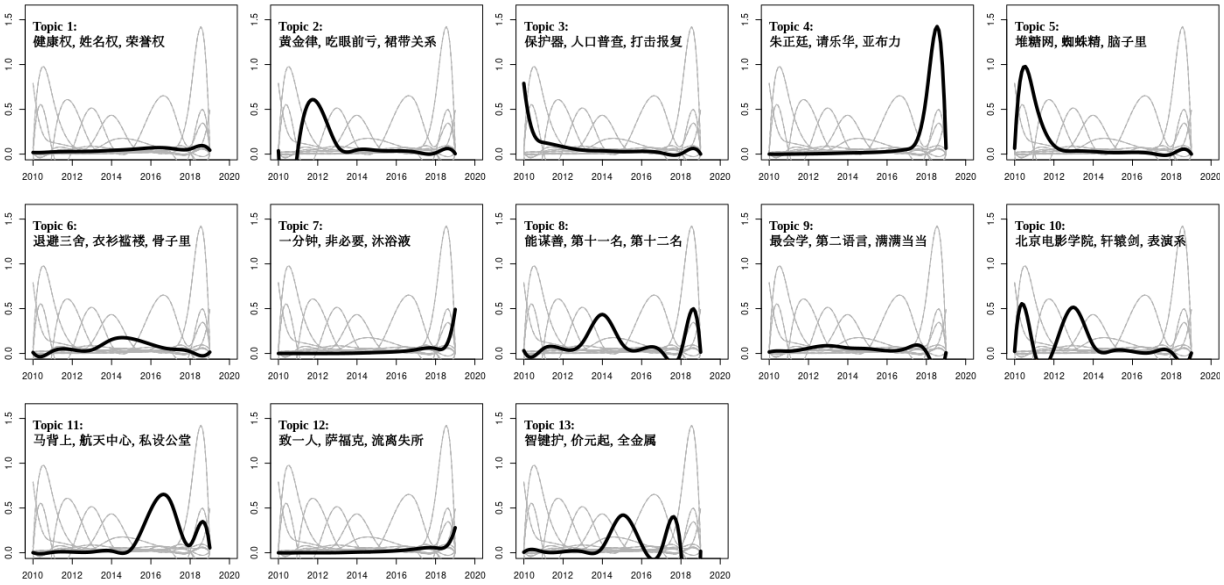


Figure 6.2 A plot of Weibo K13 topic proportions for 2010-2019

[34] Yulelajie (2018, March 24). *Renhong shifeiduo! Zhuzhengting zao renshen gongji, yuehua zhongyu zhenchulaile.* [Fame comes at a price! Yuehua finally came to defend Zhuzhengting, a much assaulted idol trainee]. Retrieved from https://baijiahao.baidu.com/s?id=1595781862369933011&wfr=spider&for=pc

For English news topics, the peak around 2011 in Topic 4 Government surveillance and national security appears to coincide with the updates of the U.S. PATRIOT Act around that time (Bureau of Justice Assistance, n.d.). In addition, the peak of Topic 11 International privacy laws around 2018 and 2019 was a period where discussions were still ongoing leading up to the final invalidation of the privacy framework between the US and EU in 2020[35]; 2018 was also the year when the GDPR took effect.

Lastly, the peak of Topic 1 Facebook, Google, Cambridge Analytica around 2018 and 2019, reveals correspondence with the timeline of the Cambridge Analytica incident of Facebook which was exposed in 2018: as the Cambridge Analytica scandal was exposed in March 2018[36], so 2018 was a period where there was a lot of media coverage regarding this scandal.

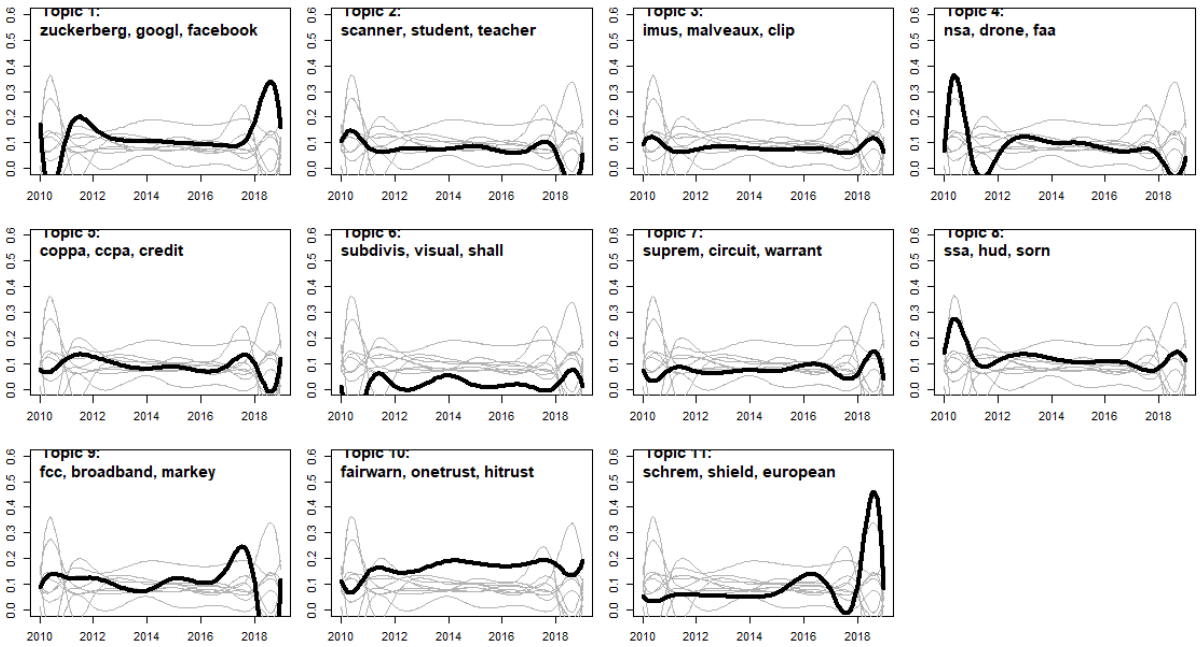


Figure 6.3 A plot of EN K11 topic proportions for 2010-2019

For Twitter topics, there are two significant peaks in Topic 9 Cyberintelligence and Topic 11 Data security of apps that have the possibility of being related to specific events at that time. Since Topic 9

[35] *FAQs – EU-U.S. privacy shield program updatefaqs – EU-U.S. privacy shield program*. Privacy Shield. (2021, March 21). Retrieved January 19, 2022, from https://www.privacyshield.gov/article?id=EU-U-S-Privacy-Shield-Program-Update#:~:text=On%20July%2016%2C%202020%2C%20the,the%20EU%2DU.S.%20Privacy%20Shield

[36] Meredith, S. (2018, April 10). *Facebook-cambridge analytica: A timeline of the data hijacking scandal*. CNBC. Retrieved January 19, 2022, from https://www.cnbc.com/2018/04/10/facebook-cambridge-analytica-a-timeline-of-the-data-hijacking-scandal.html

peaked around 2011, this peak is likely associated with “cispa” (Cyber Intelligence Sharing and Protection Act), which was introduced in 2011[37]. Another peak was for Topic 11 peak around 2019, it turned out that both Evernote (“evernot”) and Pokemon go (“pokmon”) were been criticized for their privacy practices in the most recent years (Bergen & Kulwin, 2016; Vincent, 2016).

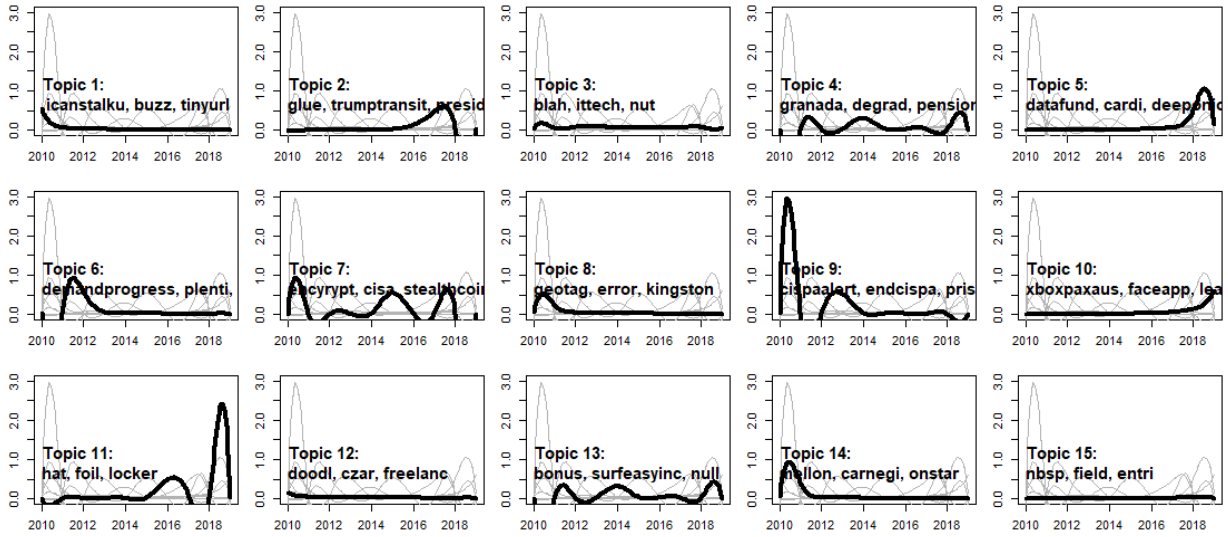


Figure 6.4 A plot of Twitter K15 topic proportions for 2010-2019

[37] *Text - H.R.3523 - 112th Congress (2011-2012): Cyber ...* Congress.gov . (n.d.). Retrieved January 19, 2022, from https://www.congress.gov/bill/112th-congress/house-bill/3523/text

CHAPTER 7: CONCLUSION

"Events of language use mediate human sociality. Such semiotic occasions develop, sustain, or transform at least part—some have argued the greater part—of people's conceptualizations of their universe." (Silverstein, 2004, p.621)

7.1 Summary of findings

The first research question in this study seeks to identify the major topics and semantics for understanding privacy in two languages: Mandarin Chinese and American English. My analysis shows that the semantics in the Chinese corpora include: first, Internet and mobile applications as the major context of privacy concerns; second, government leading the protection of consumer privacy; third, privacy at offline domains (Surveillance Camera, Health Information, and Court Information); fourth, the emphasis of privacy of the individual; and fifth, privacy as an interpersonal issue and privacy risks over specific practices and applications. Major semantics in the English corpora include: first, governmental organizations and surveillance; second, data and privacy-related services and cross-Atlantic privacy regulation; third, the persisting presence of Facebook and Google; and fourth, cutting edge technology-specific discussions of privacy.

The second research question seeks to identify patterns of topics across the two genres within one language. For each of the two languages, I found and discussed how topics vary by genre and identified topics that are shared across genres. The topics that are shared across the two genres in the Chinese language, i.e., the *core semantics* of privacy in the Chinese language (see Table 5.5), include privacy-related rights, webcam privacy violation, network security, and real-name registration and personal ID information. In other words, privacy in the Chinese language is frequently portrayed as an issue that concerns *individuals'* practices: as individual interpersonal interactions and individuals' use of popular digital technologies. For the English language, the *core semantics* of privacy (see Table 5.10) include digital technology companies (in particular, Google and Facebook), government surveillance and national security, privacy regulations and regulators; and privacy and data security services. Hence, I suggest that the *core semantics* of privacy in the English language mainly concerns *institutions* that are technological, corporate, and regulatory.

The third research question seeks to compare topics across these two languages. In the Chinese language, privacy is understood to concern interpersonal interactions, emphasizing the *individuals* in contexts. In comparison, this semantic is not present in the English language. This conclusion is reached after I have considered cases where privacy is discussed or understood in the context of interpersonal interactions in the English language (see Section 6.3 for discussions of Altman (1981) and Quinn et al. (2019)). However, the interpersonal interactions appear distinctive in the Chinese language context in that: first, they are deeply embedded in Confucian relational ethics; second, interpersonal interactions in

which privacy resides may need to resolve the conflict between the more traditional social value of *guanxi* and *privacy*.

In the English language, the concern of *institutions* focuses on governmental surveillance. In contrast, this semantic is not present in the Chinese language analysis. I reached this conclusion after considering the potential for media censorship in the Chinese language context (Song et al., 2013). I ruled out the possibility regarding the lack of government surveillance as a result of active media censorship because other studies have reported evidence that suggests media censorship may not actually occur. Specifically, Qin et al., (2017) revealed that much more sensitive words and topics (for example, “suppression”, “demonstration”, “strike”, “corruption”) were not censored on the Chinese social media (Weibo), which renders the likelihood of privacy targeted media censorship less plausible[38]. In other words, caution against government mal-practice and government surveillance may *indeed be missing* in the Chinese language, at least non-present in the two genres studied in this dissertation.

After coding all the topics in both languages using the four semantic dimensions, it became clearer how the two languages differ when compared to each other. The four semantic dimensions, though present in both languages, have an unequal presence in the two languages. Specifically, the *institution* dimension (either as governmental or commercial organizations) has much more presence in the English language than in the Chinese language. In other words, in the English language, *institutions* are held accountable for the protection of privacy as revealed by the fact that many topics revolve around the *institution* dimension. Whereas in the Chinese language, it is the *individual* dimension (either as individual technology users, or consumers in general) that is frequently seen across genres and topics. In the Chinese language, privacy currently remains an individual’s problem in their daily interactions with digital technology applications and interpersonal relationships.

This variation in emphasis can be seen in multiple topics in these two languages. For example, topics that concern student and patient information are seen in both languages. In the English results, we can see specialized privacy regulations for protecting these specific populations and their information; in contrast, there is no specialized regulation or legislation in the Chinese context for the protection of these specific population groups and their data or information.

---

[38] King et al., (2012) found that the purpose of media censorship by the Chinese government is *not* to remove criticisms, but to prevent collective actions. In other words, keywords targeted censorship is less likely unless it is associated with collective action.

This different emphasis on these two semantic dimensions (*institution* and *individual*) is also reflected through the structural space analysis of nodes where the nodes with leading centrality scores over the years in these two languages differ. Many leading nodes in the English language refer to institutions, which include “system”, “compani”, “state”, “feder”, etc. In comparison, in the Chinese language, it is “person”, “mobil”, “phone”, “user” that are leading nodes, focusing on the individuals.

I summarize these observations at multiple levels of granularity and respond to the third research question by suggesting that when it comes to privacy, there are incompatibilities across these two languages in two different ways. First, certain semantics exist in one language while missing completely in the other. Second, semantics exist in both these two languages but they emphasize different dimensions. Given the fact that the data included in this study is by no means exhaustive of all language (so that the missing of semantics could be a limitation as a result of the data used in this study), I conclude by arguing that overall, it is more cautious and appropriate to understand the incompatibilities by saying the two languages *differ by their emphasis on different dimensions*. The *individual* is emphasized in the Chinese language and the *institution* in the English language.

The last research question concerns topics over time in these two languages. There are two findings when comparing topics over time. First, fluctuations in topic proportion over time have been identified as potentially associated with specific events, after examining the exemplar documents of topics. Second, a richer analysis of time can be achieved when also considering the genre factor. For example, we have seen that for mobile phone-related topics, specifically Topic 10 and Topic 2 from the Chinese language, have revealed opposing trends: Topic 10 with an overall decline of topic proportion, while Topic 2 an overall increase. In addition, Topic 10 appears to be primarily derived from the social media corpus (Weibo) and Topic 2 shows an increase for the News genre. These observations, when considered together, suggest a more complicated picture for understanding the trend of topics in time. In that, mobile phone-related topics, despite their overall importance, are less talked about in social media corpus, and more in news. In other words, certain topics may migrate from one genre to another over time.

7.2 Additional thoughts on the missing of semantics and language in its environment

Some of the findings seen in this study should be read with these cautionary notes below in mind. First, I will briefly discuss the seemingly counterintuitive finding of the institution dimension in the English language, and the individual dimension in the Chinese language. Next, I will illustrate that the absence of a particular topic (for example, the lack of caution against government in the Chinese language) may be a result of multiple factors interacting with each other, on top of which also exists the possibility of media censorship. Third, I will discuss more what written language corpora do and can represent when put back into the broader language and information environment. These extra cautionary notes highlight how findings from this dissertation study may be read within the broader scholarly context.

Privacy has been evolving in the American English context for over a century. Its conceptualization and value has been associated with protecting individuals and individual autonomy against organizations and in particular government organizations. The protection revolving around privacy has been institutionalized so that the discussion of privacy focuses on the institutions: organizations (governmental and commercial) and laws and regulations. In comparison, the recognition of privacy explicitly is relatively recent in the Chinese context (Lü, 2005; Wang & Xiong, 2021). The fact that privacy is relatively new in the Chinese context explains why on the one hand, privacy-related language focuses on the individuals, while on the other hand, the society as a whole seems still to lack institutional establishment for the protection of privacy.

When it comes to interpreting the lack of *caution against government surveillance*, journalistic reports have revealed some opposing evidence. Undoubtedly, people's aversion towards government surveillance in the greater Chinese culture context was captured. For example, during the Anti-Extradition Law Amendment Bill Movement (which is also known as the 2019 Hong Kong protests), young people in Hong Kong clearly voiced their negative attitude toward government surveillance (Mozur & Lin, 2019). One possible way to explain the difference between the presence of government surveillance in journalistic evidence, and the absence of government surveillance in this study's Chinese language data, lies in the specific form of language that captured these semantics; in other words, it may be a difference between verbal expression and written language.

Whether a topic such as surveillance would get a *written* expression in society is a result of complicated social processes. The presence of a governmental organization actively deciding what to keep or delete is only one of the factors. Even if immediate and observable censorship is not present,

people may still choose to not say or write certain things as a result of historical censorship. In other words, people may be practicing self-censorship (Robinson & Tannenberg, 2019; Shen & Truex, 2021), with or without being explicitly aware of this. This possibility certainly deserves future research, especially work that focuses on understanding the long-term impact of censorship on the information environment, and various forms of expressions (including, but not limited to written forms) in that environment.

These considerations, while highlighting the limitations of this dissertation study's reliance on the specific written language corpora, also underline the *urgency* of researching concepts via language and the significance of understanding specific languages as they are situated within their broader social and information environment. Studying the concept of privacy by using language points to the relation between concepts and language, and it is a relation that has two directions. The presence of semantics in written language could indicate the presence of the concepts. However, if certain semantics, that are present in other languages or other forms of languages, are unseen in the written corpora, it does not necessarily mean that the semantics and the associated concepts are non-existent. Both the presence and non-presence of semantics in written language will impact the continued understanding of the concepts. After all, as stated at the beginning of this study, the meaning of privacy or any abstract concepts exists in the communication of groups of people through shared language. If certain expressions remain missing or silent in the written or spoken language, the meaning creation capacity of that language itself could be impacted.

7.3 Significance and contribution

Semantics and semantic dimensions observed in this study make it tangible what people are talking and even thinking about regarding privacy in these two languages, which is the first contribution of this dissertation to privacy research. In other words, natural language is a way to operationalize privacy research, and natural language can be especially useful for revealing the complexity of privacy as an abstract concept as we have seen in preceding chapters.

This study has revealed that natural language is also promising to operationalize intercultural privacy research and comparative privacy research (Masur et al., 2021). This study is one of the first empirically-grounded intercultural explorations of *the concept of privacy*. It provides an examination of the concept as it is understood at the current time of writing. Existing work that touches on the conceptualization of privacy with an intercultural perspective is based on historical literature and philosophical studies (Baldwin Lind, 2015; Farrall, 2008; McDougall et al., 2002; Whitman, 1985). This study looks at how the concept currently exists and is expressed in different languages. My analysis reveals that this concept is multidimensional and is situated within social and cultural traditions, and these findings can be used to support further cross-cultural discussions of privacy and related issues and concepts.

This study advances the use of natural language both as a *material* and a *method* to work with/on the topic of privacy in an intercultural and comparative setting. Approaching philosophical analysis through language is not new; however, applying computational textual analysis in cross-language settings appears to be an area that has just started to draw more attention (see Chapter 2 Related Work). This dissertation connects questions from intercultural information ethical discussions (for example, Hongladarom, 2016; Ess, 2005 & 2019) with the computational textual methods, and provides a multi-level description of how privacy as an abstract concept exists in the two natural languages. The multi-dimensional meaning of privacy as a concept that is shown in this study can be used to further investigate conceptualizations of privacy, especially when in cross-linguistic and intercultural conversations.

Making observations of a corpus at multiple levels of granularity is a method that can be used to cross-check findings from computational textual studies. In this dissertation, the inclusion of multiple computational methods has enabled the researcher to observe the corpus from levels of multiple granularities: first, at the node/word level; second, at the topis level; and third, at the dimensional level. Observing the language corpus at different granularity levels not only helps reveal the complexity of

privacy as an abstract concept, but observations from different granularity levels also help validate the interpretation for each other. For example, the emphasis on the *institution* dimension in the English language is first captured through the topics from STM (as there are multiple topics that concern government, corporate organizations as well as privacy-related regulations and laws). In addition, this emphasis on the institution is seen in the structural space analysis, as leading nodes in centrality scores are also about institutions, including “govern”, “feder”, etc. Finally, dimensional coding of topics revealed that the *institution* dimension is indeed present more in the English corpus than in the Chinese corpus.

7.4 Limitations and future directions

The main limitation in this dissertation's intercultural information ethics exploration is that it has presupposed privacy as a concept in both languages, particularly in the Chinese language. Given this constraint, what remains yet to be fully discussed is the possibility that privacy as a concept may *not* exist in the Chinese language *at all*. Or, vice versa, the possibility that yin3si1 (隐私) as a concept does not exist in the English language.

In addition, researching privacy using language has its limitations, which primarily come from the selection and design of the language corpus. In this study, the *caution against the government* is missing in the Chinese corpora of this study. However, this observation is preliminary mainly because the design of the corpora only included two genres: news, and social media. In other words, data used in this study are non-exhaustive and the corpus may have been subject to censorship. Hence, it will take more investigation to properly understand the missing of *caution against the government surveillance* for understanding privacy, rather than concluding for good here that it does not exist in the Chinese language for understanding privacy. One possibility for future exploration is that, by including more language corpus (for example, personal correspondence, novels, etc.), this topic may likely show up. Even if or when this topic does appear, it may not have the same emphasis on government surveillance that was seen in the English language.

The news articles and social media posts in this study were retrieved using only the term "privacy". Using a combination of both privacy and its synonyms or related words (for example, "personal data" or "data") could probably retrieve a bigger corpus that is not only greater in scale, but potentially also richer in semantics. Also, the news articles and social media posts corpora, should not be considered as unquestionably fair or adequate representation of what ordinary people conceptualize privacy, for the following reasons. First, when it comes to news articles, whatever the media reports does not necessarily represent what the people actually think[39]. Second, when it comes to social media posts, whatever individuals express on their public media accounts (which could be considered as one's online presentation of self), does not necessarily equate to what concerns them or what they truly believe (Hogan, 2010). In other words, it is likely that once situated in a different or more private or secure environment, individuals are more likely to voice certain aspects of their thoughts. Thirdly, the individuals

[39] Scholarly discussions of media misrepresentation are plenty; for example, media misrepresentation of minorities (Dixon et al., 2019).

who post on social media platforms like Twitter and Weibo, still only constitute a proportion of the entire population[40].

I refer to the language corpora used in this study as the Mandarin Chinese language, and the American English language, respectively. What has been ignored is the complexity and diversity within the languages themselves. In addition, by recognizing that the two genres studied are limited, I acknowledge that there are many more genres of language that can be examined. Also, how these two specific genres are situated within a broader written language expression, is also left unexplored in this study. In other words, even if it is through these two genres of language that many people read and learn about privacy, these specific two genres' impacts still exist in the broader language and information environment, mediated by other language expressions and information. The situatedness of these specific two genres and how they may actually interact with other genres of language and information remains unexplored.

One potential way to probe if there are any topics of caution against government surveillance when understanding privacy in the Mandarin Chinese language, is to compare privacy-related news articles that mention and/or originate from state/central governments versus those that concern local governments. This division is worth exploring because censorship can be selective, in that criticisms towards the local governments in the Chinese context are more likely to be allowed (Kuang, 2018). Local governments are already the target of various criticisms (Chen, 2017), including those concerning their mal-practices regarding privacy protection (Lin, 2020). This demarcation between local and central governments (Tai, 2014) is crucial in terms of revealing the complexity of attitudes towards governmental organizations regarding privacy in the Chinese language context. This distinction between local and central governments is also proposed in the hope that it could help with constructing a corpus for *challenging topics*, which for various reasons are less- or un-spoken in the language. Moreover, another reason that makes a distinction between local and central government necessary is that many data practices are primarily led by central governments. For example, the real-name registration system that is present in multiple topics in this study relies on collections of personal information and is led by central governmental organizations (including the State Council of China) rather than local governments (Ma, 2019b).

---

[40] The problems of social media data bias are discussed more extensively in Olteanu et al., (2019).

To build on the findings from this dissertation for the American English language, future studies could probe the difference between the understanding of state surveillance and corporate surveillance (Connor & Doan, 2021). Given the spread of corporate surveillance, a further question to ask is if government surveillance semantics remain as the central anchoring for understanding privacy in the English language; and if so, how. In addition, if and how the semantics regarding corporate surveillance has modified and changed the understanding of privacy in any way.

To build on the findings from this dissertation regarding the interpersonal relationships in the Chinese language, future studies could further tease out how more traditional values like *guanxi* may come into conflict, or compete with, privacy as a more recent value. The study of such potential conflicts may be highlighted when also taking into consideration the variation of privacy conceptualization among different populations in China (for example by economic status, age, etc.)

Another intriguing phenomenon revealed in this study, though preliminary, also invites future studies. The observation that one topic can display different or opposing trends over the years in different genres. For example, mobile phone-related topics show decreasing topic proportion in the social media corpus, and increasing topic proportion in the news corpus (see Section 5.1.7.2 Cross-language News STM K11 results). This observation suggests the possibility that new issues related to privacy may start with a discussion on social media platforms, and gradually move on to being discussed and reported in more formal language genres like news.

Lastly, this dissertation can also be tied to the ongoing discussions in linguistic relativity, where questions of how languages differ and how differences of language may have an impact on understanding and thinking are explored (Everett, 2013). Existing linguistic relativity studies have studied how different concepts that are fundamental to humans (including time and space) are expressed in different languages. Ethical concepts can be another area for linguistic relativity studies to investigate, for which this dissertation could be considered as a case study of a particular concept in two specific languages.

APPENDIX A.1: TOPIC LIST OF CN NEWS K13

A topic model with 13 topics, 15905 documents and a 16561 word dictionary.

Topic 1 Top Words:
Highest Prob: 工作人员, 手机号, 手机号码, 电话号码, 负责人, 个人隐私, 李女士
FREX: 李女士, 航天员, 李亚鹏, 任志强, 丁嘉丽, 许先生, 王小姐
Lift: 周筱赟, 度春宵, 抢红包, 条胡钢, 杨浦区, 水资源, 派件员
Score: 航天员, 李女士, 成绩单, 丁嘉丽, 李亚鹏, 私家侦探, 手机号
Topic 2 Top Words:
Highest Prob: 隐私权, 个人隐私, 公共利益, 当事人, 名誉权, 人格权, 事务所
FREX: 着作权, 闯红灯, 钱钟书, 男医生, 谢霆锋, 课堂教学, 贵州省
Lift: 戴平华, 素材库, 继承权, 黄海波, 关之琳, 凯宾斯基, 助人为乐
Score: 隐私权, 钱钟书, 名誉权, 人格权, 着作权, 闯红灯, 肖像权
Topic 3 Top Words:
Highest Prob: 摄像头, 公共场所, 隐私权, 个人隐私, 摄像机, 网络平台, 工作人员
FREX: 摄像头, 摄像机, 健身房, 巩义市, 急诊科, 淫秽物品, 云视通
Lift: 游泳馆, 监控室, 酒店客房, 三四条, 三圣乡, 下课铃, 不幸遇难
Score: 摄像头, 公共场所, 巩义市, 副镇长, 摄像机, 电子眼, 试衣间
Topic 4 Top Words:
Highest Prob: 艾滋病, 个人隐私, 用人单位, 医疗机构, 房产信息, 劳动者, 大学生
FREX: 艾滋病, 房产信息, 劳动者, 感染者, 贫困生, 希拉里, 证明书
Lift: 卫生部, 妇幼保健, 一平方米, 一汽大众, 三明市, 上台演讲, 上小新
Score: 艾滋病, 感染者, 贫困生, 希拉里, 用人单位, 房产信息, 劳动者
Topic 5 Top Words:
Highest Prob: 出租车, 朋友圈, 个人隐私, 有限公司, 驾驶员, 工信部, 航空公司
FREX: 出租车, 附加费, 出租汽车, 中奖者, 新闻节目, 起步价, 周杰伦
Lift: 一组组, 一门心思, 万人迷, 万多辆, 上辈子, 不知其可, 个人成长
Score: 出租车, 朋友圈, 附加费, 起步价, 出租汽车, 中奖者, 新闻节目
Topic 6 Top Words:
Highest Prob: 互联网, 个人隐私, 数据保护, 浏览器, 保护器, 第三方, 委员会
FREX: 保护器, 奥巴马, 美国政府, 执行官, 联邦贸易委员会, 安全局, 白皮书
Lift: 中所装, 华尔街日报, 参议员, 叙利亚, 方舟子, 欧洲法院, 称谷歌
Score: 保护器, 互联网, 数据保护, 奥巴马, 浏览器, 亿美元, 联邦贸易委员会
Topic 7 Top Words:
Highest Prob: 消费者, 互联网, 个人隐私, 公共安全, 征求意见, 信息系统, 人工智能
FREX: 征求意见, 人工智能, 人脸识别, 保护性, 消保委, 十万元, 一千元
Lift: 一针一线, 京津冀, 剩余次数, 原始记录, 张亚勤, 强磁场, 必做题
Score: 消费者, 人工智能, 公共安全, 征求意见, 人脸识别, 信息系统, 互联网
Topic 8 Top Words:
Highest Prob: 微博上, 越来越, 没想到, 是不是, 王女士, 李先生, 幼儿园
FREX: 心理咨询, 通知单, 前女友, 刘小姐, 日记本, 信报箱, 青春期
Lift: 信报箱, 心理专家, 收视率, 滕文超, 潘玉芳, 爱丽舍宫, 一个劲地
Score: 真人秀, 心理咨询, 幼儿园, 李先生, 信报箱, 微博上, 王女士
Topic 9 Top Words:
Highest Prob: 个人隐私, 未成年人, 当事人, 嫌疑人, 公安机关, 年月日, 人民法院
FREX: 被告人, 判决书, 检察院, 公开审理, 汉弗莱, 中介组织, 开庭审理

Lift: 健康检查, 八达通, 开庭审理, 韩德云, 一中院, 一法条, 丁二醇
Score: 未成年人, 被告人, 汉弗莱, 人民法院, 嫌疑人, 判决书, 当事人
Topic 10 Top Words:
Highest Prob: 智能手机, 手机用户, 个人隐私, 通讯录, 运营商, 应用程序, 二维码
FREX: 智能手机, 手机软件, 恶意软件, 恶意程序, 数据恢复, 下载安装, 二手手机
Lift: 渠道商, 计算机病毒, 一两分钟, 一角钱, 丁志刚, 万万不可, 三星电子
Score: 手机用户, 智能手机, 通讯录, 应用程序, 恶意软件, 无人机, 安全软件
Topic 11 Top Words:
Highest Prob: 泰迪熊, 通讯录, 安全卫士, 服务商, 电话号码, 数据安全, 不动产
FREX: 泰迪熊, 信息源, 朱骏超, 不动产, 近些年, 智能家居, 较长时间
Lift: 令人惊叹, 全频段, 厂商会, 屏蔽器, 手持式, 打差评, 程国斌
Score: 泰迪熊, 通讯录, 朱骏超, 安全卫士, 信息源, 如拨出, 机信息
Topic 12 Top Words:
Highest Prob: 个人信息, 信息安全, 互联网, 网络安全, 个人隐私, 支付宝, 法律法规
FREX: 携程网, 中消协, 刻不容缓, 个人主页, 宣传周, 杨建军, 网络安全
Lift: 中消协, 张家口市, 扩大范围, 一千余, 丁晓东, 上美图, 下抖音
Score: 个人信息, 网络安全, 信息安全, 支付宝, 互联网, 运营者, 携程网
Topic 13 Top Words:
Highest Prob: 实名制, 身份证, 公务员, 信用卡, 银行卡, 寄件人, 个人信息
FREX: 实名制, 寄件人, 人口普查, 火车票, 网约车, 一卡通, 垃圾袋
Lift: 三点一线, 交通卡, 人口总数, 免年费, 华龙网, 可买代, 咱不急
Score: 实名制, 身份证, 寄件人, 公务员, 人口普查, 普查员, 一卡通

APPENDIX A.2: TRANSLATION OF TOPIC WORDS OF CHINESE NEWS

	Highest Prob	FREX
Topic 12	个人信息, 信息安全, 互联网, 网络安全, 个人隐私, 支付宝, 法律法规	携程网, 中消协, 刻不容缓, 个人主页, 宣传周, 杨建军, 网络安全
	Personal information, information security, Internet, network security, individual privacy, Alipay, law and regulation	Ctrip, China Consumer Association, urgent, personal homepage, Cybersecurity Week, Jianjun Yang, network security
Topic 10	智能手机, 手机用户, 个人隐私, 通讯录, 运营商, 应用程序, 二维码	智能手机, 手机软件, 恶意软件, 恶意程序, 数据恢复, 下载安装, 二手手机
	smartphone, mobile user, individual privacy, contact list, internet service provider, application, QR code	smartphone, mobile software, malware, malicious program, data recovery, download and install, second-hand mobile phone
Topic 6	互联网, 个人隐私, 数据保护, 浏览器, 保护器, 第三方, 委员会	保护器, 奥巴马, 美国政府, 执行官, 联邦贸易委员会, 安全局, 白皮书
	internet, individual privacy, data protection, browser, electric protector, third-party, the Administration	electric protector, Obama, American government, executive officer, Federal Trade Commission, National Security Agency, whitepaper
Topic 8	微博上, 越来越, 没想到, 是不是, 王女士, 李先生, 幼儿园	心理咨询, 通知单, 前女友, 刘小姐, 日记本, 信报箱, 青春期
	Weibo, increasingly, unexpected, whether, Ms Wang, Mr Li, Kindergarten	Psychological counseling, notice, ex-girlfriend, Ms. Liu, diary, mailbox, adolescence
Topic 2	隐私权, 个人隐私, 公共利益, 当事人, 名誉权, 人格权, 事务所	着作权, 闯红灯, 钱钟书, 男医生, 谢霆锋, 课堂教学, 贵州省
	privacy right, individual privacy, public interest, litigant, right of reputation, right of personality, law firm	copyright, run a red light, Zhongshu Qian, male physician, Nicholas Tse, in class education, Guizhou Province
Topic 7	消费者, 互联网, 个人隐私, 公共安全, 征求意见, 信息系统, 人工智能	征求意见, 人工智能, 人脸识别, 保护性, 消保委, 十万元, 一千元
	consumer, internet, individual privacy, public security, request for comments, information system, artificial intelligence	request for comments, artificial intelligence, facial recognition, protective, Consumer Council, 100 thousand RMB, 1000 RMB
	个人隐私, 未成年人, 当事人, 嫌疑人, 公安机关, 年月日, 人民法院	被告人, 判决书, 检察院, 公开审理, 汉弗莱, 中介组织, 开庭审理

Topic 9	individual privacy, teenager/minor, litigant, suspect, public security organizations, year-month-date, people's court	defendant, court verdict, Procuratorate, public trial, Humphrey, intermediaries, court hearing
Topic 1	工作人员, 手机号, 手机号码, 电话号码, 负责人, 个人隐私, 李女士	李女士, 航天员, 李亚鹏, 任志强, 丁嘉丽, 许先生, 王小姐
	employee, mobile number, mobile number, individual privacy, Ms Li	Mr Li, astronaut, Yapeng Li, Zhiqiang Ren, Jiali Ding, Mr. Xu, Ms. Wang
Topic 3	摄像头, 公共场所, 隐私权, 个人隐私, 摄像机, 网络平台, 工作人员	摄像头, 摄像机, 健身房, 巩义市, 急诊科, 淫秽物品, 云视通
	webcam, public space, privacy right, individual privacy, camera, network platform, staff	webcam, camera, gym, Gongyi city, emergency room, pornography, CloudSEE
Topic 4	艾滋病, 个人隐私, 用人单位, 医疗机构, 房产信息, 劳动者, 大学生	艾滋病, 房产信息, 劳动者, 感染者, 贫困生, 希拉里, 证明书
	HIV, individual privacy, employer, medical institute, property information, worker, college student	HIV, property information, worker, infected, impoverished/poor students, Hilary, proof
Topic 13	实名制, 身份证, 公务员, 信用卡, 银行卡, 寄件人, 个人信息	实名制, 寄件人, 人口普查, 火车票, 网约车, 一卡通, 垃圾袋
	real name registration, ID card, civil servant, credit card, bank card, sender, personal information	real name registration, sender, census, train ticket, ride hailing, all-purpose card, trash bag
Topic 5	出租车, 朋友圈, 个人隐私, 有限公司, 驾驶员, 工信部, 航空公司	出租车, 附加费, 出租汽车, 中奖者, 新闻节目, 起步价, 周杰伦
	taxi, wechat moments, individual privacy, corporate limited, driver, Ministry of Industry and Information Technology (MIIT), airline companies	taxi, additional fee, taxi, lottery winners, news program, base price, Jay Chow
Topic 11	泰迪熊, 通讯录, 安全卫士, 服务商, 电话号码, 数据安全, 不动产	泰迪熊, 信息源, 朱骏超, 不动产, 近些年, 智能家居, 较长时间
	Teddy Bear, contact list, security guard, service provider, phone number, data security, real estate	Teddy Bear, information source, Zhu junchao, real estate, recent years, smart home, extended time

APPENDIX A.3: TRANSLATION OF TOPIC WORDS OF WEIBO

	Highest Prob	FREX
Topic 11	个人隐私, 个人信息, 水瓶座, 朋友圈, 隐私权, 越来越, 手机号	马背上, 航天中心, 私设公堂, 磁力线, 券用券, 气质端庄, 妻葛黛瓦
	Personal privacy, personal information, Aquarius, wechat moment, privacy right, increasingly, mobile number	horseback, space center, illegal court, coupon, magnetic line, elegant, wife Godiva
Topic 2	个人隐私, 办公室, 发生冲突, 善解人意, 表达意见, 人际关系, 水瓶座	黄金律, 吃眼前亏, 裙带关系, 祥林嫂, 绊脚石, 方舟子, 顺口溜
	Personal privacy, office, conflict occurred, considerate, express opinion, interpersonal relationship, Aquarius	Golden rule, suffer losses, nepotism, Aunt Xianglin, stumbling block, Fang Zhouzi, tongue twister
Topic 8	个人隐私, 双子座, 天蝎座, 巨蟹座, 金牛座, 白羊座, 射手座	能谋善, 第十一名, 第十二名, 变形金刚, 第六名, 第七名, 第八名
	Personal privacy, gemini, scorpia, cancer taurus, aries, sagittarius,	Resourceful, eleventh, twelfth, transformers, sixth, seventh, eighth
Topic 5	个人隐私, 脑子里, 一点点, 美图秀, 幼儿园, 下半生, 拦路虎	堆糖网, 蜘蛛精, 脑子里, 孙悟空, 美图秀, 乔布斯, 豆腐心
	Personal privacy, in mind, a bit, Meitu, kindergarten, second half of life, obstacle on the road	Tangdui, Spider monster, in one's mind, Monkey King, Meitu, Jobs, gentle mind
Topic 4	个人隐私, 朱正廷, 个人信息, 互联网, 越来越, 朋友圈, 请乐华	朱正廷, 请乐华, 亚布力, 计算机硬件, 杜华乐华, 李宗伟, 交易商
	Personal privacy, Zhengting Zhu, personal information, internet, increasingly, wechat moments, demand Yuehua	Zhu Zhengting, demand Yuehua, computer hardware, Duhua Yuehua, Li Zingwei, dealer
Topic 10	个人隐私, 明月光, 高风险, 客户端, 逃不了, 系统安全, 个性化	北京电影学院, 轩辕剑, 表演系, 天之痕, 文博会, 民族服装, 王老古
	Personal privacy, moon light, high risk, customer end, inescapable, system security, customization	Beijing Film Academy, department of performing arts, The sword of Xuan Yuan,
	个人隐私, 自由空间, 个人信息, 是因为, 隐私权, 朋友圈, 发牢骚	智键护, 价元起, 全金属, 客客气气, 蜻蜓点水, 阴阳怪气, 君子之交淡如水

Topic 13	Personal privacy,, free space, personal information, because, privacy right, wechat moment, complain	Smart fingerprint protection, price, all metal, polite, touch on something, bad-tempered, A hedge between keeps friendship green
Topic 7	个人隐私, 个人信息, 摄像头, 朋友圈, 隐私权, 信息安全, 互联网	一分钟, 非必要, 沐浴液, 光大银行, 纸巾盒, 生活用品, 电子钟
	Personal privacy, personal information, webcam, wechat moment, privacy right, information security, internet	One minute, unnecessary, body shampoo, Everbright Bank, tissue box, daily necessities electronic clock,
Topic 6	个人隐私, 弄清楚, 保持中立, 尖酸刻薄, 二维码, 有没有, 个人信息	退避三舍, 衣衫褴褛, 骨子里, 衣冠楚楚, 冠冕堂皇, 趋之若鹜, 弄清楚
	Personal privacy, figure out, stay neutral, mean, QR code, whether or, personal information	avoid, poorly dressed, in one’s nature, dapper, high-sounding, chasing after sth, figure out,
Topic 3	个人隐私, 隐私权, 保护器, 互联网, 是不是, 越来越, 实名制	保护器, 人口普查, 打击报复, 郭德纲, 安全部队, 看人脸色, 抱佛脚
	Personal privacy, privacy right, protector, internet, increasingly, real name registration	Protector, census, retaliate, Guo Degang, security troops, subservient, last minute work
Topic 1	隐私权, 当事人, 个人隐私, 哈哈哈, 名誉权, 肖像权, 未成年人	健康权, 姓名权, 荣誉权, 生命权, 专利权, 判决书, 诉讼法
	Privacy right, litigant, personal privacy, hahaha, reputation right, portraiture right, teenager	Health right, Naming right, Reputation right, Life right, Patent right, court verdict, litigation law
Topic 12	个人隐私, 隐私权, 个人信息, 摄像头, 网络安全, 朋友圈, 互联网	致一人, 萨福克, 流离失所, 绝一人, 下议院, 英国议会, 妇女节
	Personal privacy, privacy right, Personal information, webcam, network security, wechat moment, internet	To one, Suffolk, Homeless, only one, lower house, British parliament, Women’s day
Topic 9	摄像头, 个人隐私, 最会学, 第二语言, 双子座, 女强人, 自我解嘲	最会学, 第二语言, 满满当当, 牢骚满腹, 自我解嘲, 怒气冲天, 女强人
	Webcam, personal privacy, best learning, second language, gemini, competent women, self-mockery	best learning, second language, full, whining, self-mockery, angry, competent women

APPENDIX A.4: TOPIC LIST OF EN NEWS K11

A topic model with 11 topics, 24998 documents and a 23792 word dictionary.

Topic 1 Top Words:
Highest Prob: user, facebook, privaci, compani, googl, said, data
FREX: zuckerberg, googl, facebook, app, appl, cambridg, analytica
Lift: ananda, aran, arcad, blippi, bogost, britteni, chappl
Score: facebook, googl, user, app, appl, zuckerberg, said
Topic 2 Top Words:
Highest Prob: said, student, health, state, school, use, patient
FREX: scanner, student, teacher, hospit, classroom, dna, patient
Lift: antarctica, aston, biehl, derr, dioxid, docken, gedmatch
Score: student, patient, said, school, health, educ, tsa
Topic 3 Top Words:
Highest Prob: think, know, say, one, peopl, get, like
FREX: imus, malveaux, clip, tonight, velshi, cavuto, yeah
Lift: aleppo, aorta, arborvita, azuz, babe, bankrol, bashar
Score: imus, think, say, realli, get, talk, clip
Topic 4 Top Words:
Highest Prob: govern, secur, said, surveil, american, nation, agenc
FREX: nsa, drone, faa, snowden, terror, terrorist, aircraft
Lift: aopa, bieseck, binney, boghosian, bosh, cse, csec
Score: drone, nsa, snowden, surveil, said, intellig, terrorist
Topic 5 Top Words:
Highest Prob: inform, consum, data, privaci, person, provid, requir
FREX: coppa, ccpa, credit, hipaa, settlement, ftc, breach
Lift: andrequir, annualcreditreport, aorzehoski, arant, boult, casher, cjame
Score: consum, ftc, ccpa, coppa, data, hipaa, inform
Topic 6 Top Words:
Highest Prob: person, violat, section, physic, imag, shall, record
FREX: subdivis, visual, shall, impress, compensatori, physic, sound
Lift: disgorg, zenovich, auditori, ingress, subdivis, compensatori, enjoin
Score: subdivis, plaintiff, visual, shall, impress, compensatori, section
Topic 7 Top Words:
Highest Prob: court, law, case, privaci, enforc, investig, search
FREX: suprem, circuit, warrant, judg, court, fourth, subpoena
Lift: amarosa, asbl, ascia, ayala, beeler, bopp, cfca
Score: court, suprem, warrant, ecpa, judg, justic, plaintiff
Topic 8 Top Words:
Highest Prob: record, system, inform, feder, act, offic, privaci
FREX: ssa, hud, sorn, usci, osd, dod, docket
Lift: ahrc, altmey, apss, bryman, cncs, dpfpa, dtra
Score: system, dhs, dod, record, docket, ssa, hud
Topic 9 Top Words:
Highest Prob: privaci, consum, protect, data, bill, inform, senat
FREX: fcc, broadband, markey, rep, subcommitte, isp, ntia
Lift: atsc, sohn, ajit, alce, alertsbillhid, amarasingham, appright
Score: consum, fcc, senat, rep, bill, legisl, ftc
Topic 10 Top Words:
Highest Prob: privaci, data, secur, inform, com, provid, manag
FREX: fairwarn, onetrust, hitrust, csf, patent, ota, isaca
Lift: onetrust, abenant, abloy, accesspr, accordingth, acegroup, activa
Score: patent, data, fairwarn, trademark, solut, onetrust, mobil
Topic 11 Top Words:
Highest Prob: data, privaci, protect, european, compani, law, shield
FREX: schrem, shield, european, transatlant, apec, gdpr, europ

Lift: jourova, acirc, ald, allason, andrus, ansip, antal
Score: european, shield, data, gdpr, schrem, europ, compani

APPENDIX A.5: TOPIC LIST OF EN NEWS K13

A topic model with 13 topics, 24998 documents and a 23792 word dictionary.

Topic 1 Top Words:
- Highest Prob: use, data, privaci, inform, peopl, onlin, like
- FREX: survey, percent, wearabl, shop, toy, alexa, pew
- Lift: acquisti, ananda, anjana, blippi, britteni, cylindr, dopplr
- Score: think, consum, data, say, advertis, onlin, thank

Topic 2 Top Words:
- Highest Prob: health, said, student, state, patient, school, educ
- FREX: patient, student, classroom, hospit, genet, teacher, physician
- Lift: burzichelli, delano, dysphoria, gresham, greul, holcomb, kia
- Score: patient, student, health, said, school, educ, medic

Topic 3 Top Words:
- Highest Prob: say, know, peopl, one, think, right, get
- FREX: imus, malveaux, clip, abc, tonight, velshi, cavuto
- Lift: acosta, aleppo, alfresco, aorta, arborvita, aristocrat, armani
- Score: imus, say, think, clip, cnn, malveaux, realli

Topic 4 Top Words:
- Highest Prob: govern, secur, said, surveil, nation, american, agenc
- FREX: nsa, snowden, terror, patriot, fisa, surveil, metadata
- Lift: binney, dakwar, delong, emmerson, enrico, fountainhead, gchq
- Score: nsa, snowden, surveil, intellig, fbi, said, terrorist

Topic 5 Top Words:
- Highest Prob: inform, consum, privaci, data, person, provid, requir
- FREX: coppa, ccpa, credit, breach, ftc, ocr, url
- Lift: andrequir, annualcreditreport, aorzehoski, arant, boult, casher, cjame
- Score: consum, ftc, ccpa, coppa, data, inform, hipaa

Topic 6 Top Words:
- Highest Prob: person, violat, section, physic, imag, shall, subdivis
- FREX: subdivis, visual, shall, compensatori, impress, physic, sound
- Lift: disgorg, zenovich, auditori, compensatori, enjoin, ingress, subdivis
- Score: subdivis, plaintiff, visual, shall, impress, compensatori, section

Topic 7 Top Words:
- Highest Prob: court, law, case, privaci, enforc, investig, search
- FREX: suprem, circuit, warrant, judg, court, subpoena, fourth
- Lift: afifi, alsup, amarosa, aronberg, asbl, ascia, ayala
- Score: court, suprem, warrant, ecpa, justic, plaintiff, judg

Topic 8 Top Words:
- Highest Prob: record, system, inform, feder, act, offic, privaci
- FREX: ssa, hud, sorn, usci, osd, dod, omb
- Lift: ahrc, altmey, andnecessari, anyrecord, appropriateag, apss, arecord
- Score: system, dhs, dod, record, docket, ssa, hud

Topic 9 Top Words:
- Highest Prob: privaci, protect, bill, consum, senat, legisl, inform
- FREX: fcc, broadband, rep, markey, sen, isp, subcommitte
- Lift: amarasingham, arl, billhid, copra, dier, dstone, hudgin
- Score: consum, fcc, senat, rep, bill, legisl, broadband

Topic 10 Top Words:
- Highest Prob: privaci, secur, data, inform, com, manag, servic
- FREX: fairwarn, onetrust, hitrust, csf, isaca, patent, pct
- Lift: abenant, accesspr, accordingth, acegroup, activa, adt, agentless
- Score: patent, fairwarn, data, trademark, onetrust, solut, user

Topic 11 Top Words:
- Highest Prob: data, privaci, protect, european, compani, law, shield
- FREX: shield, european, schrem, gdpr, framework, transfer, apec

Lift: antal, brunei, buttarelli, chapnick, docx, edp, eeck
Score: european, data, shield, gdpr, schrem, framework, europ
Topic 12 Top Words:
Highest Prob: facebook, user, googl, said, compani, privaci, data
FREX: zuckerberg, facebook, googl, analytica, cambridg, whatsapp, appl
Lift: ashli, barley, bering, etonian, goncharov, hanspet, heyward
Score: facebook, googl, user, appl, zuckerberg, app, said
Topic 13 Top Words:
Highest Prob: use, said, privaci, canada, secur, drone, canadian
FREX: tsa, faa, drone, scanner, airspac, canadian, unman
Lift: abowd, airspac, kamloop, peevey, pirker, rideau, vonn
Score: drone, canadian, canada, tsa, airport, faa, aircraft

APPENDIX A.6: TOPIC LIST OF TWITTER K15

A topic model with 15 topics, 25000 documents and a 47259 word dictionary.

Topic 1 Top Words:
Highest Prob: privaci, facebook, googl, set, social, polici, onlin
FREX: icanstalku, buzz, tinyurl, nearbi, appspot, medal, proxypi
Lift: audioo, azzmdi, benefitsma, bgvfan, biochemist, bzizck, cfuzao
Score: tinyurl, buzz, facebook, ping, icanstalku, appspot, dlvr
Topic 2 Top Words:
Highest Prob: privaci, secur, data, real, free, show, absolut
FREX: glue, trumptransit, presidentelecttrump, static, infosecjob, maga, giveaway
Lift: airplaneapril, dysphoriajo, honeybadgerbit, infbloodcom, notmynigel, johngallj, antisnp
Score: webcam, guarante, presidenttrump, eeolshjjmv, maga, absolut, infosecjob
Topic 3 Top Words:
Highest Prob: privaci, facebook, like, need, set, peopl, want
FREX: blah, ittech, nut, jenni, shoe, vermont, medit
Lift: ittech, webbizceo, nbreach, absoltuley, anthea, itsjust, lolss
Score: ittech, facebook, privaci, like, set, blah, need
Topic 4 Top Words:
Highest Prob: privaci, data, secur, facebook, protect, like, need
FREX: granada, degrad, pension, nkdgvijlnn, classdojo, blackphon, glenn
Lift: gayformagcon, privacysolv, amendmentprotect, teamyamita, thefourth, widescre, ownprivaci
Score: snowden, granada, atomsoffic, blackphon, nkdgvijlnn, privaci, nkdgvirczf
Topic 5 Top Words:
Highest Prob: privaci, data, secur, facebook, right, protect, like
FREX: datafund, cardi, deeponion, myhealthrecord, kavanaugh, ethereum, capitaltechnologiesresearch
Lift: dacifac, raindov, sukiblueberri, caramelleacid, kataplixi, heltongreen, holidaycurr
Score: gdpr, blockchain, cryptocurr, cardi, coin, datafund, anatica
Topic 6 Top Words:
Highest Prob: privaci, facebook, googl, polici, need, set, user
FREX: demandprogress, plenti, petraeus, nich, randi, buyer, hysteria
Lift: callm, castroasesino, theburgerman, themarkhenri, toddkincannon, armynew, canadianarmi
Score: cispa, nich, demandprogress, privaci, facebook, plenti, googl
Topic 7 Top Words:
Highest Prob: privaci, protect, data, secur, free, american, real
FREX: encyrypt, cisa, stealthcoin, idltweet, barbi, newpanda, stitm
Lift: isao, aldubswitch, dontbeagooglesheep, pearsonto, tellpearson, lsectweet, cyberfeminist
Score: cisa, webcam, atomsoffic, guarante, snowden, stealthcoin, stitm
Topic 8 Top Words:
Highest Prob: privaci, facebook, set, googl, polici, onlin, protect
FREX: geotag, error, kingston, carrier, timelin, datatravel, verdict
Lift: arcadeweb, corporateupd, searchdainew, raysman, eyscfl, bppdcmmq, qwsysn
Score: ping, facebook, privaci, dlvr, tinyurl, geotag, arcadeweb
Topic 9 Top Words:
Highest Prob: privaci, googl, facebook, like, protect, data, secur
FREX: cispaalert, endcispa, prism, obamacar, typewrit, bush, cispa
Lift: plent, beepic, glueck, phreegal, pyrop, plattner, prettyperuvian
Score: cispa, cispaalert, prism, nich, privaci, endcispa, snowden
Topic 10 Top Words:
Highest Prob: privaci, data, like, peopl, right, secur, need
FREX: xboxpaxaus, faceapp, leaderboard, ccpa, nica, doordash, securypto

Lift: cerda, imperosoftwar, jennyabamu, matthewjshow, mrahmedserougi, myhealthchamp, pkathrani
Score: ccpa, gdpr, blockchain, xboxpaxaus, doordash, faceapp, artificialintellig
Topic 11 Top Words:
Highest Prob: privaci, data, real, free, secur, show, absolut
FREX: hat, foil, locker, evernot, pokmon, pokemon, fertil
Lift: didayimebn, glennon, euref, homeworkhelp, btrenchard, ashtonbthink, stilllook
Score: webcam, guarante, presidenttrump, absolut, eeolshjjmv, trumpcar, foil
Topic 12 Top Words:
Highest Prob: privaci, facebook, googl, set, polici, secur, onlin
FREX: doodl, czar, freelanc, lennon, creatur, invadin, buzz
Lift: nawti, toge, allsopp, doodl, wemissmj, nerney, aguirr
Score: tinyurl, buzz, facebook, googl, privaci, nawti, ping
Topic 13 Top Words:
Highest Prob: privaci, data, secur, protect, facebook, onlin, internet
FREX: bonus, surfeasyinc, null, truecrypt, pear, papul, ncpol
Lift: yrozlhpvkg, zzsfxkjgjx, ayrsyvsjgx, facebookexperi, sidecar, jmhattem, vsmfcyfuon
Score: surfeasyinc, bonus, atomsoffic, snowden, blackphon, nkdgvijlnn, papul
Topic 14 Top Words:
Highest Prob: privaci, facebook, polici, set, onlin, like, protect
FREX: mellon, carnegi, onstar, hampton, geofenc, sacr, roethlisberg
Lift: damnitstru, deadeyefr, dimntf, hmcipr, markgr, bluntspeakin, molcavil
Score: ping, tinyurl, dlvr, privaci, facebook, mellon, carnegi
Topic 15 Top Words:
Highest Prob: privaci, data, secur, like, peopl, facebook, protect
FREX: nbsp, field, entri, npleas, soul, fill, pacif
Lift: behr, nyccoffeeshop, postworkout, kilz, haemorrhag, fitnessphotograph, batzographi
Score: gdpr, behr, nbsp, sweepstak, blockchain, kilz, entri

APPENDIX A.7: TOPIC LIST OF TWITTER K27

A topic model with 27 topics, 25000 documents and a 47259 word dictionary.

Topic 1 Top Words:
Highest Prob: privaci, facebook, googl, polici, set, social, onlin
FREX: geotag, error, icanstalku, nearbi, hour, scan, stalk
Lift: factoidz, duckworth, bmxwzo, cwfa, twiddl, undat, participationy
Score: tinyurl, ping, geotag, facebook, icanstalku, dlvr, error
Topic 2 Top Words:
Highest Prob: privaci, real, free, absolut, show, guarante, webcam
FREX: eeolshjjmv, guarante, webcam, absolut, infosecjob, maga, trumptransit
Lift: themfor, shelburn, pmcg, johngallj, antisnp, awifefirst, ayrshirebog
Score: webcam, guarante, absolut, eeolshjjmv, presidenttrump, maga, show
Topic 3 Top Words:
Highest Prob: privaci, facebook, set, like, googl, polici, social
FREX: ittech, shaw, leach, gramm, resel, stewart, kristen
Lift: ittech, blogsit, adconion, bliley, gramm, babycent, tisk
Score: ittech, facebook, ping, privaci, tinyurl, social, set
Topic 4 Top Words:
Highest Prob: privaci, data, facebook, secur, protect, like, need
FREX: granada, null, nkdgvijlnn, blackphon, classdojo, pension, nkdgvirczf
Lift: amendmentprotect, thefourth, trabajo, teamyamita, widescre, ccours, ewaemi
Score: atomsoffic, snowden, granada, blackphon, nkdgvijlnn, nkdgvirczf, papul
Topic 5 Top Words:
Highest Prob: privaci, data, secur, right, facebook, protect, like
FREX: ethereum, chatbot, cryptocurr, datasci, blockchain, capitaltechnologiesresearch, fintech
Lift: ecomi, dacifac, raindov, sukiblueberri, realajbenza, wtfimontwitr, ciadaught
Score: gdpr, blockchain, cryptocurr, bitcoin, coin, ethereum, fintech
Topic 6 Top Words:
FREX: addr, hysteria, stumbler, petraeus, ric, demandprogress, statist
Lift: callm, karthika, muthukumaraswami, butterfield, gvcbel, inquirera, promicrosoft
Score: cispa, stumbler, demandprogress, googl, facebook, petraeus, privaci
Topic 7 Top Words:
Highest Prob: privaci, protect, data, secur, free, real, show
FREX: cisa, icit, stealthcoin, stitm, idltweet, barbi, lcevcsjdyu
Lift: aldubswitch, dpintens, brunet, bune, futut, pizd, functionaleleg
Score: webcam, cisa, guarante, atomsoffic, absolut, snowden, stealthcoin
Topic 8 Top Words:
Highest Prob: privaci, facebook, set, polici, googl, onlin, protect
FREX: foothold, carrier, cun, ubermedia, costolo, vodafon, arcadeweb
Lift: arcadeweb, trupanion, londonlipgloss, corporateupd, irealhatewhen, mediakil, pawedcard
Score: ping, facebook, dlvr, tinyurl, foothold, privaci, arcadeweb
Topic 9 Top Words:
Highest Prob: privaci, product, huge, plenti, facebook, nich, googl
FREX: plenti, buyer, nich, huge, product, plent, prism
Lift: plent, rajasa, tahanan, kota, bola, buyer, plenti
Score: nich, plenti, buyer, product, huge, cispa, plent
Topic 10 Top Words:
Highest Prob: privaci, data, like, peopl, right, secur, need
FREX: xboxpaxaus, faceapp, leaderboard, ccpa, nica, doordash, securypto
Lift: matthewjshow, pkathrani, ameensol, cerda, snolancollin, mrahmedserougi, engrapp
Score: ccpa, gdpr, xboxpaxaus, doordash, blockchain, faceapp, artificialintellig
Topic 11 Top Words:
Highest Prob: privaci, data, secur, protect, internet, free, real

FREX: hat, centen, foil, locker, privacyshield, ndata, broadband
Lift: centen, zscbjxpyuk, southernsej, featuresaf, coreysdavi, substanceabusej, shivakumar
Score: webcam, guarante, absolut, eeolshjjmv, presidenttrump, hat, foil
Topic 12 Top Words:
Highest Prob: privaci, facebook, googl, set, polici, concern, secur
FREX: doodl, creatur, raleigh, invadin, lennon, nawti, buzz
Lift: nawti, asbkxo, raleigh, invadin, techblogstoday, doodl, progra
Score: tinyurl, buzz, facebook, googl, nawti, ping, dlvr
Topic 13 Top Words:
Highest Prob: privaci, data, secur, protect, facebook, onlin, right
FREX: surfeasyinc, bonus, truecrypt, blackphon, ncpol, ncga, ncgop
Lift: zzsfxkjgjx, facebookexperi, ioreep, webrtcdigest, marekciesla, vsmfcyfuon, ilovemyjob
Score: surfeasyinc, atomsoffic, bonus, snowden, blackphon, truecrypt, nkdgvijlnn
Topic 14 Top Words:
Highest Prob: privaci, facebook, polici, set, like, onlin, protect
FREX: mellon, carnegi, sacr, hulu, telemarket, hampton, dropbox
Lift: deadeyefr, privacywherev, neuromarket, promul, pronlinenew, ftmvrw, internetpawn
Score: ping, dlvr, tinyurl, mellon, carnegi, privaci, facebook
Topic 15 Top Words:
Highest Prob: privaci, data, facebook, secur, like, protect, peopl
FREX: anatica, datafund, cardi, cambridg, iamcardib, ubymi, gdpr
Lift: behr, lapphund, qldlabor, fuatmcnhej, fionneorland, gdprcountdown, clintonviceb
Score: gdpr, blockchain, cardi, anatica, datafund, coin, ccpa
Topic 16 Top Words:
Highest Prob: privaci, protect, data, secur, right, browser, firefox
FREX: rival, experiment, firefox, credenti, allinternet, addon, unparallel
Lift: allinternet, vineet, experiment, adikamdar, domainkey, makena, ldtgihclac
Score: allinternet, experiment, firefox, rival, browser, snowden, mynam
Topic 17 Top Words:
Highest Prob: privaci, facebook, googl, social, polici, need, set
FREX: calendar, fortress, zoom, mccain, diff, drunken, randi
Lift: readdl, yopro, apr, dragani, cvqvvcth, pote, crextacom
Score: cispa, nich, facebook, readdl, plenti, privaci, instagram
Topic 18 Top Words:
Highest Prob: privaci, facebook, googl, proxi, onlin, protect, polici
FREX: appspot, proxypi, fastest, proxi, stabl, taemin, incept
Lift: tutorials, femenil, whereveryou, palabrasquedanmiedo, youkilledthemood, thestopadivorc, usethisproxi
Score: appspot, proxypi, tinyurl, proxi, fastest, ping, buzz
Topic 19 Top Words:
Highest Prob: privaci, facebook, screen, like, googl, protect, set
FREX: widescreen, notebook, filter, accessori, protector, glare, bold
Lift: gautham, nagesh, obtai, miafreedman, desbloqueada, jig, antistat
Score: screen, gautham, nagesh, filter, notebook, ping, protector
Topic 20 Top Words:
Highest Prob: privaci, secur, data, protect, internet, right, like
FREX: static, glue, giveaway, trumptransit, presidentelecttrump, decor, infosecjob
Lift: davidoro, whyidontusefacebook, electi, khuja, nmunich, hile, factsoup
Score: presidenttrump, maga, infosecjob, trumpcar, glue, giveaway, skyrocketad
Topic 21 Top Words:
Highest Prob: privaci, googl, facebook, data, like, right, need
FREX: alto, palo, retai, intim, alma, gatsbi, spain
Lift: retai, emord, rens, hagen, electronicfrontierfound, matthewbarbi, santaclaralaw
Score: retai, cispa, prism, palo, snowden, alto, drone
Topic 22 Top Words:
Highest Prob: privaci, protect, secur, data, right, facebook, onlin

FREX: pear, papul, penil, bobbi, westpac, marshal, atlas
Lift: stetson, bitcoinsoper, aciaicyla, pcgtw, hiit, gpodagrosi, subtledevi
Score: papul, pear, atomsoffic, snowden, penil, cisa, stealthcoin
Topic 23 Topic Words:
Highest Prob: privaci, facebook, googl, set, polici, onlin, like
FREX: insert, lever, chain, slat, aluminum, wreck, creeper
Lift: oyenamit, elkowiri, porteg, tecra, intelligencecommun, gastown, stargam
Score: cispa, facebook, demandprogress, oyenamit, petraeus, sopa, privaci
Topic 24 Top Words:
Highest Prob: privaci, facebook, googl, set, social, polici, onlin
FREX: buzz, whini, dipshit, tinyurl, myspac, medal, nsfw
Lift: fcvw, sanfranciscoinfonewspap, gcjuw, bsujto, digireport, cjnlff, idzjba
Score: tinyurl, facebook, buzz, ping, icanstalku, dlvr, googl
Topic 25 Top Words:
Highest Prob: privaci, real, free, show, data, absolut, webcam
FREX: hat, fertil, glow, foil, locker, evernot, pokemon
Lift: notprivaci, kinsman, fertilit, iacipp, btrenchard, fakeheadlinebot, makeatwitterbot
Score: webcam, guarante, presidenttrump, absolut, trumpcar, skyrocketad, eeolshjjmv
Topic 26 Top Words:
Highest Prob: privaci, googl, facebook, like, secur, protect, data
FREX: cispaalert, stopcispa, endcispa, prism, typewrit, lavabit, tunnelbear
Lift: blippex, mistasucr, mobileguard, plattner, gaggingbil, temityperkin, stiennon
Score: cispa, cispaalert, prism, endcispa, snowden, cispablackout, stopcispa
Topic 27 Top Words:
Highest Prob: privaci, facebook, polici, googl, set, mobil, secur
FREX: upsurg, applicat, nutrit, applic, respons, mobil, comment
Lift: upsurg, applicat, amalgam, fling, peen, contactless, chachigonzal
Score: upsurg, applicat, privaci, mobil, facebook, polici, nutrit

APPENDIX A.8: TOPIC LIST OF WEIBO K13

A topic model with 13 topics, 23336 documents and a 37125 word dictionary.

Topic 1 Top Words:
Highest Prob: 隐私权, 当事人, 个人隐私, 哈哈哈, 名誉权, 肖像权, 未成年人
FREX: 健康权, 姓名权, 荣誉权, 生命权, 专利权, 判决书, 诉讼法
Lift: 益物权, 博淑芬姐, 荣誉权, 专用权, 专利权, 名称权, 健康权
Score: 哈哈哈, 荣誉权, 健康权, 家庭暴力, 姓名权, 名誉权, 证明书
Topic 2 Top Words:
Highest Prob: 个人隐私, 办公室, 发生冲突, 善解人意, 表达意见, 人际关系, 水瓶座
FREX: 黄金律, 吃眼前亏, 裙带关系, 祥林嫂, 绊脚石, 方舟子, 顺口溜
Lift: 之多动哥, 之大戒, 偶发性, 儿吹得, 免疫针, 入围奖, 出血量
Score: 表达意见, 黄金律, 善解人意, 明月光, 吃眼前亏, 发生冲突, 逃不了
Topic 3 Top Words:
Highest Prob: 个人隐私, 隐私权, 保护器, 互联网, 是不是, 越来越, 实名制
FREX: 保护器, 人口普查, 打击报复, 郭德纲, 安全部队, 看人脸色, 抱佛脚
Lift: 三言二拍, 两不一, 他会出, 伯博爵, 傅盛称, 反三俗, 夏云涛
Score: 保护器, 安全卫士, 抱佛脚, 人口普查, 安全部队, 窥私门, 拿得起
Topic 4 Top Words:
Highest Prob: 个人隐私, 朱正廷, 个人信息, 互联网, 越来越, 朋友圈, 请乐华
FREX: 朱正廷, 请乐华, 亚布力, 计算机硬件, 杜华乐华, 李宗伟, 交易商
Lift: 一万名, 七八张, 三百六十五个, 下元券, 下愿用, 中国台北队, 之涂磊
Score: 朱正廷, 请乐华, 中心化, 密码学, 亚布力, 计算机硬件, 杜华乐华
Topic 5 Top Words:
Highest Prob: 个人隐私, 脑子里, 一点点, 美图秀, 幼儿园, 下半生, 拦路虎
FREX: 堆糖网, 蜘蛛精, 脑子里, 孙悟空, 美图秀, 乔布斯, 豆腐心
Lift: 一来一去, 一逼夫, 万人继, 不世出, 不雅门, 专家级, 东方日报
Score: 脑子里, 堆糖网, 美图秀, 蜘蛛精, 表达意见, 下半生, 豆腐心
Topic 6 Top Words:
Highest Prob: 个人隐私, 弄清楚, 保持中立, 尖酸刻薄, 二维码, 有没有, 个人信息
FREX: 退避三舍, 衣衫褴褛, 骨子里, 衣冠楚楚, 冠冕堂皇, 趋之若鹜, 弄清楚
Lift: 不蹭白, 共元别, 张清华, 微博剑网, 郭采洁, 点半起, 副部级
Score: 弄清楚, 衣冠楚楚, 衣衫褴褛, 保持中立, 退避三舍, 趋之若鹜, 衡量标准
Topic 7 Top Words:
Highest Prob: 个人隐私, 个人信息, 摄像头, 朋友圈, 隐私权, 信息安全, 互联网
FREX: 一分钟, 非必要, 沐浴液, 光大银行, 纸巾盒, 生活用品, 电子钟
Lift: 一肖请, 丁聪之子, 万余户, 上位法, 不计前嫌, 中数度, 交通图
Score: 摄像头, 王一博, 迪士尼, 杨洋方, 超范围, 朱一龙, 爆安宰贤
Topic 8 Top Words:
Highest Prob: 个人隐私, 双子座, 天蝎座, 巨蟹座, 金牛座, 白羊座, 射手座
FREX: 能谋善, 第十一名, 第十二名, 变形金刚, 第六名, 第七名, 第八名
Lift: 一没见, 刘泽刚, 可发往, 庄稼汉, 徐德明, 检查点, 热线专
Score: 狮子座, 金牛座, 白羊座, 魔羯座, 双鱼座, 巨蟹座, 管太多
Topic 9 Top Words:
Highest Prob: 摄像头, 个人隐私, 最会学, 第二语言, 双子座, 女强人, 自我解嘲
FREX: 最会学, 第二语言, 满满当当, 牢骚满腹, 自我解嘲, 怒气冲天, 女强人

Lift: 曝二怕, 拿手好戏, 米德兰, 配钥匙, 满满当当, 有三怕, 嵴柱炎
Score: 防窥贴, 摄像头, 第二语言, 最会学, 自我解嘲, 离婚率, 女强人
Topic 10 Top Words:
Highest Prob: 个人隐私, 明月光, 高风险, 客户端, 逃不了, 系统安全, 个性化
FREX: 北京电影学院, 轩辕剑, 表演系, 天之痕, 文博会, 民族服装, 王老古
Lift: 北京电影学院, 文博会, 民族服装, 一十八, 一场虚惊, 三十二篇, 上不应
Score: 明月光, 逃不了, 高风险, 系统安全, 密码锁, 文博会, 民族服装
Topic 11 Top Words:
Highest Prob: 个人隐私, 个人信息, 水瓶座, 朋友圈, 隐私权, 越来越, 手机号
FREX: 马背上, 航天中心, 私设公堂, 磁力线, 券用券, 气质端庄, 妻葛黛瓦
Lift: 不开森, 不雅床, 中井柏然, 中鹿晗, 丰密面, 之人伤, 乌当区
Score: 实际行动, 马背上, 网约车, 朋友圈, 妻葛黛瓦, 气质端庄, 葛黛瓦为
Topic 12 Top Words:
Highest Prob: 个人隐私, 隐私权, 个人信息, 摄像头, 网络安全, 朋友圈, 互联网
FREX: 致一人, 萨福克, 流离失所, 绝一人, 下议院, 英国议会, 妇女节
Lift: 佛言人于, 同剧素, 孙宇晨, 致一人, 萨福克, 一博谢, 互相冲突
Score: 致一人, 迪士尼, 萨福克, 杨洋方, 王一博, 绝一人, 英国议会
Topic 13 Top Words:
Highest Prob: 个人隐私, 自由空间, 个人信息, 是因为, 隐私权, 朋友圈, 发牢骚
FREX: 智键护, 价元起, 全金属, 客客气气, 蜻蜓点水, 阴阳怪气, 君子之交淡如水
Lift: 二三月, 何安下, 作收纳, 卷帘拉下, 吊挂在, 外其顶, 店庆大
Score: 蜻蜓点水, 客客气气, 阴阳怪气, 智键护, 君子之交淡如水, 自由空间, 价元起

APPENDIX A.9: TOPIC LIST OF CROSS-LANGUAGE NEWS ANALYSIS

A topic model with 11 topics, 19999 documents and a 17880 word dictionary.

Topic 1 Top Words:
Highest Prob: camera, student, school, live, parent, privaci, children
FREX: taxi, classroom, passeng, camera, teacher, student, broadcast
Lift: bayonet, cellophan, circumcis, erectil, foreskin, groin, invigil
Score: camera, student, taxi, passeng, school, teacher, parent
Topic 2 Top Words:
Highest Prob: said, privaci, govern, state, court, right, bill
FREX: snowden, ecpa, leahi, drone, warrant, fisa, liberti
Lift: clapper, dragnet, fisc, lieberman, litt, plouff, rousseff
Score: senat, warrant, liberti, court, feder, snowden, drone
Topic 3 Top Words:
Highest Prob: data, privaci, consum, protect, provid, secur, inform
FREX: ccpa, gdpr, framework, shield, subdivis, solut, complianc
Lift: ahima, akingump, auditori, availableher, bigid, bisgaard, bissel
Score: data, gdpr, ccpa, consum, healthcar, california, fairwarn
Topic 4 Top Words:
Highest Prob: inform, report, phone, person, number, card, bank
FREX: bank, card, ticket, shop, merchant, crack, wechat
Lift: akai, baimoutu, boje, bozang, cheyipai, dadao, danzhou
Score: wechat, netizen, yuan, leakag, phone, inform, leak
Topic 5 Top Words:
Highest Prob: think, know, peopl, like, want, thing, time
FREX: clip, tonight, inaud, malveaux, yeah, somebodi, crosstalk
Lift: abba, agonis, alright, aniston, armani, ashleymadison, backsplash
Score: think, malveaux, thank, clip, talk, mccord, realli
Topic 6 Top Words:
Highest Prob: user, mobil, phone, secur, data, privaci, softwar
FREX: teddi, softwar, tencent, mobil, virus, trojan, bear
Lift: antivirus, aoyou, ciphertext, daquan, dunjun, firmwar, ganji
Score: user, mobil, teddi, phone, softwar, tencent, yanbei
Topic 7 Top Words:
Highest Prob: inform, person, privaci, protect, right, public, regul
FREX: supervis, claus, strengthen, china, stipul, punish, ministri
Lift: foyu, lengshuijiang, zhaoqun, chunyao, daokui, expropri, hesheng
Score: inform, infring, china, leakag, supervis, stipul, netizen
Topic 8 Top Words:
Highest Prob: privaci, compani, data, facebook, googl, user, inform
FREX: googl, analytica, zuckerberg, facebook, amazon, coppa, uber
Lift: kalanick, libra, pichai, abin, accc, acton, adweek
Score: facebook, googl, user, zuckerberg, appl, data, app
Topic 9 Top Words:
Highest Prob: record, system, inform, offic, privaci, feder, notic
FREX: docket, foia, citat, alexandria, sorn, patent, routin
Lift: usac, abalo, adjudicatori, afpc, aggarw, alexandria, altmey
Score: feder, docket, pursuant, system, washington, sorn, foia
Topic 10 Top Words:
Highest Prob: express, health, patient, medic, inform, hospit, deliveri
FREX: patient, medic, hospit, courier, deliveri, diseas, doctor
Lift: waybil, alkaloid, angong, antigen, antiretrovir, antivir, apnea
Score: patient, hospit, medic, deliveri, health, courier, express
Topic 11 Top Words:
Highest Prob: privaci, photo, report, court, public, famili, case
FREX: divorc, husband, marriag, wife, celebr, girlfriend, kong

Lift: alin, authorship, cuili, huangyan, liyang, shengjia, yuntian
Score: weibo, netizen, zhang, wang, wife, husband, daughter

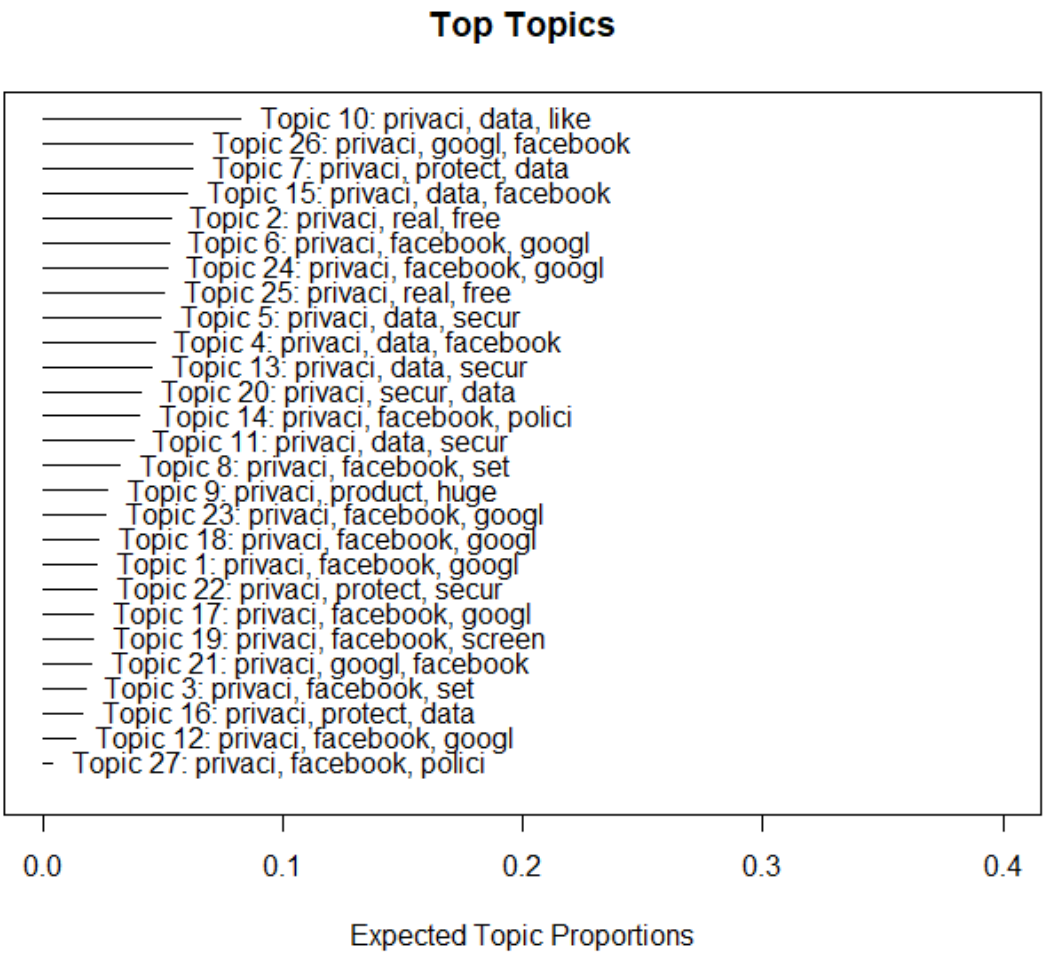
APPENDIX A.10: TOPIC LIST OF CROSS-GENRE AND CROSS-LANGUAGE ANALYSIS

A topic model with 11 topics, 36900 documents and a 37334 word dictionary.

Topic 1 Top Words:
Highest Prob: person, public, inform, right, privaci, student, protect
FREX: stipul, shall, subdivis, judgment, patient, municip, tort
Lift: beishankou, blackthorn, etiolog, maoheng, zenovich, benhui, canjun
Score: infring, patient, stipul, student, yuan, public, subdivis
Topic 2 Top Words:
Highest Prob: inform, user, mobil, phone, secur, person, data
FREX: teddi, mobil, softwar, bear, leakag, vulner, malici
Lift: daocui, dunjun, iimedia, kantou, minrui, modian, yuechuan
Score: mobil, user, phone, leakag, teddi, inform, wechat
Topic 3 Top Words:
Highest Prob: user, googl, data, facebook, said, privaci, compani
FREX: patent, analytica, abstract, stoddart, inventor, cambridg, trademark
Lift: arpi, asiekierska, baldera, biswa, cama, chowdhri, debmalya
Score: facebook, googl, user, said, appl, data, warrant
Topic 4 Top Words:
Highest Prob: data, privaci, consum, protect, inform, secur, provid
FREX: fairwarn, framework, harbor, ecpa, ccpa, stakehold, subcommitte
Lift: advogado, availableher, ayer, ballon, bissel, brisboi, cbpr
Score: data, gdpr, consum, ccpa, hipaa, senat, ecpa
Topic 5 Top Words:
Highest Prob: camera, express, privaci, live, report, photo, children
FREX: taxi, courier, meow, pipe, meter, slip, passeng
Lift: baoxi, changzhouren, congxin, jinguang, tianmuhu, xigong, xiuna
Score: yuan, wang, netizen, zhang, wechat, camera, weibo
Topic 6 Top Words:
Highest Prob: privaci, protect, person, know, want, friend, inform
FREX: zhengt, lehua, gome, blockchain, ecard, onlook, artist
Lift: baohong, cipherpunk, gongjin, huakan, jihan, jingpengcheng, lehua
Score: weibo, privaci, tencent, blockchain, wechat, realli, zhengt
Topic 7 Top Words:
Highest Prob: said, peopl, think, know, like, year, want
FREX: malveaux, inaud, clip, clinton, videotap, gutfeld, unidentifi
Lift: malveaux, abdin, baier, bartlett, blagojevich, blitzer, bollea
Score: said, republican, malveaux, think, realli, obama, trump
Topic 8 Top Words:
Highest Prob: privaci, facebook, data, secur, protect, like, googl
FREX: fami, socialmedia, probab, eeolshjjmv, dataprotect, presidenttrump, yall
Lift: bule, cdnpoli, cybersec, gamedev, lukewilliamss, stealthcoin, fami
Score: privaci, facebook, dont, googl, fami, infosec, info
Topic 9 Top Words:
Highest Prob: record, system, inform, offic, feder, privaci, agenc
FREX: docket, sorn, routin, notic, supplementari, citat, submiss
Lift: arsf, cigi, cmppa, dhap, dsca, facsimil, fdms
Score: docket, feder, record, pursuant, system, sorn, amend
Topic 10 Top Words:
Highest Prob: privaci, link, person, talk, phone, love, mobil
FREX: villain, gemini, taurus, michao, ari, capricorn, disk
Lift: absentmind, aihuo, cornet, croxin, dingbao, duantangwang, duitangwang
Score: villain, weibo, love, talk, privaci, mobil, gemini
Topic 11 Top Words:
Highest Prob: privaci, internet, technolog, social, right, data, public
FREX: recognit, artifici, facial, census, genet, hong, scienc

Lift: jihe, cortex, dorsolater, prefront, tirol, schnberg, zewei
Score: china, technolog, chines, recognit, facial, artifici, internet

APPENDIX B: EXPECTED TOPIC PROPORTION PLOT OF TWITTER K27



APPENDIX C.1: EXEMPLAR DOCUMENT FOR TOPIC 7 OF CN NEWS

公安部 日前 会同 部门 研究 起草 公共安全 视频 图像 信息系统 管理条例 征求意见 稿 征求意见 稿 指出 禁止 可能 泄露 隐私 场所 部位 安装 视频 图像 采集 设备 违法者 单位 安装 单位 处 一万元 十万元 以下 罚款 个人 安装 个人 处 一千元 五千元 以下 罚款 公安部 发布 公共安全 视频 图像 信息系统 管理条例 征求意见 稿 征求意见 稿 指出 旅馆 客房 集体 宿舍 公共 浴室 更衣室 卫生间 可能 泄露 隐私 场所 部位 禁止 安装 视频 图像 采集 设备 征求意见 稿 指出 单位 个人 利用 公共安全 视频 图像 信息系统 非法 获取 国家 秘密 工作 秘密 商业秘密 侵犯 公民 个人隐私 合法权益 公共安全 视频 图像 信息系统 建设 使用 单位 系统 设计 方案 设备 类型 安装 位置 地址码 基础 信息 获取 涉及 国家 秘密 工作 秘密 商业秘密 视频 图像 信息 负有 保密 义务 获取 涉及 公民 个人隐私 视频 图像 信息 非法 泄露 征求意见 稿 指出 社会 公共 区域 视频 图像 采集 设备 安装 位置 应当 居民 住宅 保持 合理 距离 旅馆 客房 集体 宿舍 公共 浴室 更衣室 卫生间 可能 泄露 隐私 场所 部位 禁止 安装 视频 图像 采集 设备 征求意见 稿 指出 视频 图像 信息 用于 公共 传播 时除 法律 另有 规定 外 应当 涉及 当事人 个体 特征 机动车 号牌 隐私 信息 采取 保护性 措施 征求意见 稿 指出 单位 个人 下列 行为 盗窃 损坏 擅自 拆除 公共安全 视频 图像 信息系统 设施 设备 破坏 擅自 删改 公共安全 视频 图像 信息系统 运行 程序 运行 记录 删改 隐匿 毁弃 留存 期内 公共安全 视频 图像 信息系统 采集 原始 视频 图像 信息 买卖 非法 使用 复制 传播 公共安全 视频 图像 信息系统 基础 信息 采集 视频 图像 信息 影响 公共安全 视频 图像 信息系统 正常 使用 情形 征求意见 稿 表示 国家机关 工作人员 违法 强制 要求 企事业 单位 组织 个人 建设 公共安全 视频 图像 信息系统 指定 变相 指定 公共安全 视频 图像 信息系统 设计 施工 维护 单位 设备 品牌 销售 单位 单位 个人 有权 公安机关 部门 履行 公共安全 视频 图像 信息系统 监督管理 职责 违法行为 进行 检举 控告 收到 检举 控告 机关 应当 职责 及时 查处 征求意见 稿 提出 违反 条例 规定 可能 泄露 隐私 场所 部位 安装 视频 图像 采集 设备 县级 地方 人民政府 公安机关 责令 立即 拆除 拒 拆除 依法 申请 人民法院 强制 拆除 单位 安装 单位 处 一万元 十万元 以下 罚款 个人 安装 个人 处 一千元 五千元 以下 罚款 征求意见 稿 指出 国家机关 工作人员 履行 监督管理 职责 工作 滥用职权 玩忽职守 徇私舞弊 依法 给予 处分 构成犯罪 依法追究 刑事责任 综合 新华社 中新网

APPENDIX C.2: EXEMPLAR DOCUMENT FOR TOPIC 6

Topic 6:

alias weijun wednesday please follow wechat public account questions weijun advice seeing issue weijun suddenly feels balance enough mind want bear strange questions anonymous little girl secretariat work recently inadvertently knowing privacy worrying about knowing long time knowing knowing knowing dismissed leaking talking sleep careful hearing deaf dumb first calm every story spreading world girlfriend loves world feel taste getting farther farther love xudu weijun thank pull girlfriend together knowing women silly cute fried eggs simple weijun bangs stay middle continue keep long hair short hair tangled choice phobia late prestige classmate rolls dice edition manuscript written reporter jiecai

minute tell protection personal information personal privacy leakage serious security risks browser social platforms financial software privacy leakage prevent personal information leakage look small coup hubei internet police patrol enforcement shoot video bank collection student funds disclose student information jiangxi bank collects student funds discloses names students addresses colleges universities jiangxi pingxiang rural commercial bank recently issued reminder notice cause disputes reminders urge overdue student loans announce name university graduates school overdue amount minimum amount overdue amount multiple deemed infringement individuals hidden youyou postgraduate entrance exam english exams vocabulary baby unfamiliar means equivalent today sharing usage means case especially suitable writing situation afraid writing mistakes best write king true questions unreasonable demands respect children privacy rights regrets trust children achievements represent early sleep lazy children concentrate reading distracted housework lack humor affection matter coquettish unreasonable excuses remember sentence betray friends privacy please unfamiliar friends phone number guizhou weibo liupanshui city henan province puyang henan province playing profile picture addresses live public whole country privacy concealment hide current curse find current curse turtle dare face current curse wrong fact interesting relationship high school classmates girls affirming interests hobbies always similarities personality kind independent self world usually play other completely connected other know roots know bottom least three four years friends familiar usual contact little holiday water bottle aquarius personality analysis aquarius smart biggest feature innovation pursuit unique life individualism strong constellation friendly people attention privacy aquarius definitely considered star friendship like make kind friends difficult make heart takes long time family members seem cold alienated many years boundaries things inherently sensitive privacy relatives friends never actively asks want embarrassed situation affects relatives friends care worry housework including asset allocation privacy industry urges legislation regulate restrict sina data protect privacy industry urges legislation regulate restrict cold knowledge hotel room must turn bathroom lights attention privacy bring knowledge vagrant weibo video huazhu hotel suspected infringing privacy chicken huawei translate english melon english internet buzzword word melon means news gossip events privacy melon share gossip meaning english buzzwords similar meanings world diapers convenient children deep sleep comfortable protect children privacy brothers sisters lezhen foshan come recruit wechat assist unblocking address time limit salary successful unblocking establishment process friend blocking take long open account group simple operation protection privacy shielding friends friends circle friends circle friends block colleagues fines meta urinary regulations understand wechat work software want block block last piece privacy news today news hear husband spokesperson major media respects privacy longer respond matter love many years welcomed first child secretly married year true week weibo fish sauce rumored kill family three haidian district reviews death wechat meng team swiped screen fraud overturned deleted fish sauce punishment detention article questioned many parties satirical meng statement protect privacy information meng meng apologized jiangsu police patriotic stamp contest getting started novices consult technical issues related stamp competition write article friends reference stamp competition important team event organized dynamic circle sports circle organizes major events every week competition period registration week competition week requires team members seven days every distance stars idol encounters rational fans wang jiaer besieged begged fans kneel down taeyeon male fans forcibly pulled away scared members suffer illegitimate childbirth crying waking middle night stranger stands bedside illegitimate child best life threatening making call best delete oxygen shooting video information from stage stage easy wechat privacy completely leaked website link chenyu today theme belt chenyu wants love world coolest nameplate hands protect personal privacy careful share mistakes seems ritual affected song left them secret together unreserved woman overly inquiring privacy along wisely give certain

amount private space likes respect privacy blessings look zhao liying feng shaofeng wechat settings prohibit adding friends wechat cannot prohibiting adding friends friends verification settings click open wechat program wechat homepage click settings click privacy address book friends need verify back button open choice privacy terrible thoughts terrible resource sharing circles weibo video fans suiyue first knows going facing professor charming kisses heartbeats turn cost much reverie dare think much think cute child bears heart think term want silly remind privacy violations feel test privacy comment reporter asks shing personal privacy issues public boss domineering wants know crocodile swallowing gold awareness always early children especially girls important church protect family education part daily life fathers mothers helps children understand body correctly cultivate children awareness gender privacy learn protect themselves professor wang dawei people public security university china explained huateng officially announced wechat behavior toleration tell wechat liked consumers reason wechat privacy good links things suning received express information sent things send places email legend mobile phone billing really privacy lost love breaks saves boyfriend girlfriends love dating secrets breaking finding relatives friends instead redemption feel good choice send relatives friends help solve emotional problems emotions always people privacy break find outsider replace communication mediation hope write something want know city watched film documentary read book experience feelings know like exposing people want privacy want talk life want philosophy buddhism nature mobile payment battle territory never stopped experts believe personal perspective personal data privacy needs strengthen public security many subways announced launch lightning network alternatives temporarily applicable bitcoin according news privacy blockchain company announced launch lightning network payment solution payment processor verifier chain generation batch payment validity proof verification verification verification digital signature verification payer sufficient funds transaction chain sent micro reminder casually like minute tell protection personal information security every personal information leakage disgusting life network closely related know browser social platform financial software privacy leakage prevent personal information leakage quickly look small coup prevent leakage personal information link remembers schedule apple privacy advertisements announced last year apple senior president global privacy participated consumer privacy roundtable last time apple participated exhibition recent years exhibitor status launch products apple always indirectly participating online dramas exclusive memory foggy roommates really disgusting move things privacy

days chengcheng coughed helped cover quilt knowing afraid heat cover shoulders feet satisfied feet covered whole body covered strangely usually cover quilt head chengcheng mother

follow dynamics blacklist weibo open trumpet follow classmates broke words play online games wear vest deliberately close continue sneak snipes online always alert network offensive defensive battle name love threw floor last year blushed yelled concerned fact classmates classmates concerned privacy smoke mother child came little suddenly scene mother zhang never imagined scene before invisible online fighting wisdom bravery sometimes front weibo message appears sometimes signs small account follow trend express disgust with want space zhang always feels relieved wants master dynamics days mother changes longer limited real life study daily life break online world start follow blacklist weibo zhang photo diary internet computing fashion person space often updated sina weibo multiple fans wechat circle friends activist commented that zhang mother stand playing multiple weibos high school pursues called freedom usually likes mother beginning year zhang relatives reposted weibo found opened weibo paid attention seeing sent weibo message comment expect days later protested classmates weibo parent attention parent comments concerned dare post smoking photo weibo fact scared death taking pictures mother sweet worried things things happy recently results easier grasp communication understanding weibo good thing change quiet follow mode soon learns smart software knows weibo mother browses look time weibo want follow mode weibo quietly knowing soon name weibo name blacklisted unceremoniously blacklisted following classmates time delete traces fans registered weibo account hardly posted weibo called trumpet zhang weibo trumpet following weibo listed special attention trumpet thinking anyway knows follow classmates classmates fans follow public wechat account blog number star weibo account miss weibo trumpet total attention fans never posted weibo strange existence frequent logins kind similar voyeurism feels zhang enjoys time knowing whereabouts thinking classmate relationship mastered classmates nearly half year never posted weibo account saturday night zhang woke room brushed weibo sofa living room suddenly snatch homework tired movie late quit weibo account late delete traces minutes angrily away beginning article smoke mother child internet filled real life zhang figure parents post weibo follow peeping privacy question answer parent group several parents said method imitating small unit want child know wearing vest organizes online game opens accounts deal mother fact

campaign zhang stop invisible footsteps internet weibo trumpet plan failed found playing online game league legends popular game requires team battles zhang registered league legends account young people inquire game skills attention account deliberately close form group attention game time feels headache deal mother accounts mother friend good account open cousins cousins turn active weibo wechat knows account registered become wechat friend zhang zhou believes method actually simple truly understands child preferences leader child makes friends classmates like interact zhou this weibo food picture daughter dormitory experience zhou experience children comments weibo actually indicate want communicate children possible ways arouse disgust

wood feel exchanging specific privacy secrets friendship love feeling closer like reciprocity idol selling boyfriends setting idols boyfriends fall love lower requirements climb wall climbing wall boyfriend sweetly asks idols really know talk love idol private life choose true crying easy know comes care year life career bring drastic changes future heartache thinks half year police media privacy review mass entertainment discuss little criminal suspicion search innocent knows pride sadness danger consumers worry leakage bank card privacy quick reminder remember protect privacy mobile phones privacy protection mobile phones store large amount personal privacy information protect data loss mobile phones dealt defraud cctv chinese police online weibo videos changli share changli guide preventing fraud pits daily life involves personal privacy property little partners must keep source public account changsha polytechnic remembered time broke someone distressed second envy actor country next door idol different actors idol different idols kind private problem passers help scold actors kind problem passerby privacy xiao zhan mobile phone number company long time artists fans held opinions this company must respond violation privacy artists basic protection company artists successful company always criticized leakage privacy management protection artists solve immediately declare privacy leaked endless ways leak personal information many people leak privacy earn benefits legal sanctions inevitable personal privacy handled picture answer someone collects sells faces seven uncles brains moon private romantic things best arms four pointed moon night arms stars take look moon take peek uncle brain present fiat currency payment processors appearing market processor integrates variety utility tools payment methods cryptocurrency payments payment portfolio provide unparalleled benefits consumers businesses benefits include simplicity security privacy lower overall costs improved personal fund control link supervision exposure privacy violations curious sohu entertainment every time open honest mosaic shows legal behavior sohu embarks must sensitive words cautious front talk endlessly seem show fact naked exposed eyes weight silence self protection lose privacy heart nowhere hide must think twice things tough things regret afterwards keep solid footsteps conscientiousness less regrets like always like privacy matter female encounters salty elbow crotch take pictures call police hangzhou crotch left seat quietly took picture obscene called police finally administratively detained japanese pear video first hand video foreign netizens satisfied design student dormitory make full space protect personal privacy victims pick privacy messages weibo turn weibo service function every parent want personal space want leak privacy leaks save mobile phone privacy security protection data every privacy protected mobile phone stores personal privacy information protect data lost mobile phones dealt videos taken minute tell protection personal information security ministry public security organizes collection personal information violation laws regulations centralized rectification rectification illegal illegal collection personal information take look prevent hubei internet police patrolling enforcing meet frankly unlock phones respect trust privacy real problem believe many couples troubled said boyfriend would take initiative give tian liang received girl text message said happen prove something couples look mobile phones popular dirtiest countries dirty country roadside dirty country roadside dirtiest country roadside dirty country roadside movie calls maddening easy remember last time feel like seeing easy remember last time felt like friends around want open wanted find place send yitu year holiday house recommended luxury vineyard holiday house yitu zealand largest chinese vacation rental platform vacations want escape hustle bustle city experience beautiful scenery rural life come decorate vineyard house located vineyards northern wine region canterbury provides privacy luxury accommodation experience french decoration style facebook suspected spying user privacy apple downloads immediately link privacy guard browser marriott fined million yuan leaking room opening information hundreds millions guests relevant booking system suspended kunming knowledge network security told protect personal information minute life network closely related know browser social platforms privacy financial software leaks minor receiving harassing calls severely suffering property damage threatening personal safety preventing personal information leakage take quick look prevent leakage personal information hubei internet police patrol enforcement take video wechat switch quickly turn privacy exposed thank watching video sharing beautiful world page link resume steal privacy headlines address book personal privacy minute tells protect safety personal

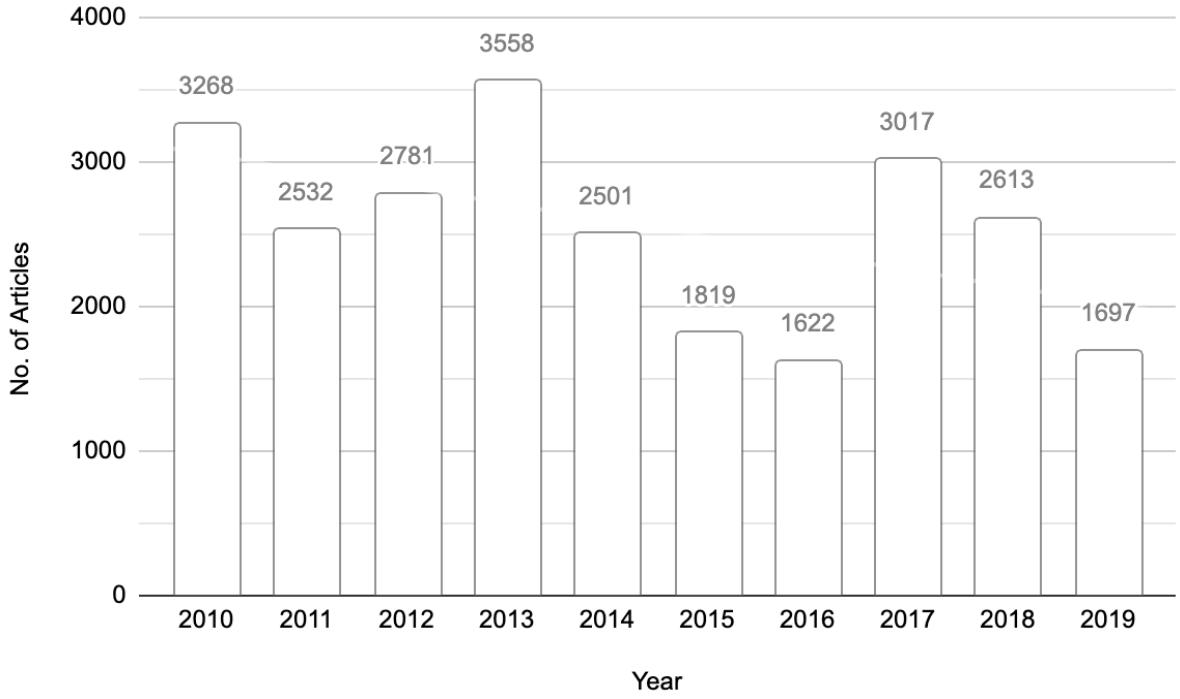
information personal privacy leaks serious security risks browser social platform financial software privacy leaks receive harassment calls personal safety prevent personal information leaking take quick look prevent small coups police report uniqlo sneak shots tear disguise careful pinhole cameras cameras transaction chaos discussion protect privacy pinhole cameras faked look like daily necessities hooked clock sockets places easy sneak shots sneak recordings check pinhole cameras poke videos shoot videos reveal leadership take care help gossip colleagues behind back gossip work methods tell colleagues salary tell colleagues leadership pictures mosaic follow protect privacy submitted tianxian baby warm reminder qunli tianxian baby recommends leaving group protect talk heart looking forward legal process baby deceived baby cheated wounded anything tongue anything attack people shortcomings expose people scars expose people scars invite hatred harm others harm oneself dignity world stands world face skin every dignity face life receiving shortcomings speaking people privacy praise word mouth boasting speaks speaks itself social psychologists found process falling love marriage surprisingly similar always going fixed steps always easy along other increasingly frequent longer time relationships objects falling love step gradually seek companionship time second step disclose private personal information third step open desire greatest scourge talking privacy greatest evil knowing negligence greatest illness aristotle zhang yixing mother posts zhang yixing thousands chinese youth love motherland artists kept mouth shut zhang yixing stood clear stand support hong kong police supporting state leak privacy forged donated organs kind person little life bottom line heart must severely punish kind villain defend justice patriots least respect support privacy leaks damage mobile phone privacy security protection leading heated discussions data every mobile phone pays great attention privacy protection stores large amount personal privacy information protect data loss mobile phone dealt cctv shouting opening door husband yelling toilet over talk management rules main function over talking amway encourages actively produce publish high quality originals hope abide rules mutual supervision reminders carefully read kicks points mutual fans experience kicks points guide dili lengba forbidden words excessive catharsis rhythm custom live action alarm clock choose service weather getting colder getting difficult manual wake services quietly emerging internet need yuan enjoy private customized stores provide monthly services people difficulty getting think type service quite good lawyers risk personal privacy leakage service modern express quick quick video

APPENDIX D.1: A LIST OF THE TOP 30 CHINESE NEWS SOURCES

Source	Count of articles
中国新闻社	1194
21世界经济报道（爱淘稿）	425
南方都市报	290
法制晚报	265
新京报	226
金陵晚报	187
齐鲁晚报(数字报)	185
北京晨报	181
法制日报(电子报)	181
广州日报	172
华商报	169
新快报	139
通信信息报	138
羊城晚报(全国版)	138
新闻晨报	133
南国早报	129
北京青年报	128
新民晚报	126
人民邮电报	121
汕头特区晚报	115
新晚报(数字报)	114
京华时报	113
深圳特区报	111
北方新报	108

京江晚报	107
参考消息	107
西宁晚报	106
电脑报	106
经济晚报	103
福州晚报(数字报)	102

APPENDIX D.2: NEWS ARTICLE COUNT BY YEAR



APPENDIX E: SAMPLE OF CHINESE CORPUS TRANSLATED INTO ENGLISH

Processed CN News

互联网安全 厂商 公司 近日 宣布 推出 一款 名为 扣扣 保镖 安全 工具 称该 工具 全面 保护 用户 安全 包括 阻止 查看 用户 隐私 文件 防止 木马 盗取 帐号 加速 功能 继 推出 隐私 保护器 再次 推出 一款 产品 该款 产品 目标 锁定 据介绍 扣扣 保镖 提供 阻止 查看 用户 隐私 文件 功能 用户 开启 隐私 保护 功能 自动 阻止 聊天 程序 电脑 硬盘 隐私 文件 强制 扫描 查看 据介绍 扣扣 保镖 重要 功能 加速 提供 禁用 开启 插件 功能 启动 程序 变小 聊天 加速 业内人士 认为 推出 产品 意味着 腾讯 之间 纠纷 更加 激烈 公司总裁 齐向东 此前 腾讯 口水战 表态 同行 吵架 一定 程度 曝光 业内 不足 推动 行业 发展 口水 代替 不了 网民 提供 优质服务 表示 希望 停止 腾讯 公司 争吵 目前 腾讯 暂未 此事 做出 回应 数日 前 金山 百度 腾讯 傲游 可牛 公司 联合 发布 反对 正当竞争 加强 行业 自律 联合声明 希望 表达 坚决 反对 正当竞争 行径 呼吁 加强 互联网 行业 自律 中国 互联网 健康 发展 创造 良好环境 腾讯 弹窗 形式 内容 用户 进行 告知 上述 行为 卫士 同样 桌面 弹窗 形势 进行 回应 腾讯 全网 弹窗 报复 公布 长期 超级 黑名单 方式 偷偷 扫描 用户 硬盘 获取 巨额 利益 一份 声明 中称 腾讯 软件 弹窗 官方网站 专题 博客 微博 方式 进行 恶意 传播 扩大 影响 每日 经济 新闻报道 指出 两家 公司 之间 口水仗 引来 众多 网民 关注 一切都是 客户端 强势 推送 促成 意义 讲 数亿 网民 被迫 围观 腾讯 互不相让 弹窗 举动 更是 掀起 全民 娱乐 高潮

Translated CN News

An Internet security company recently announced the launch of a security tool called buckle bodyguard, saying that the tool comprehensively protects user security, including preventing viewing of user privacy files, preventing Trojan horses from stealing accounts, and accelerating functions. According to the introduction, the buckle bodyguard provides the function of preventing the viewing of user privacy files. The user turns on the privacy protection function to automatically block the chat program. The computer hard disk privacy file is forced to scan and view. According to the introduction, the buckle bodyguard is important to accelerate the function. Disable the plug-in function to start the program to become smaller. He believes that the launch of the product means that the dispute between Tencent has become more intense. Qi Xiangdong, the president of the company, has previously expressed that Tencent's verbal warfare has stated that peer quarrels have been exposed to a certain extent. The industry's insufficiency has promoted the development of the industry. Saliva can not replace Internet users. A few days ago, Jinshan Baidu, Tencent, Maxthon Keniu jointly issued a joint statement against fair competition and strengthening industry self-discipline, hoping to express its firm opposition to fair competition and call for strengthening the self-discipline of the Internet industry. The healthy development of the Internet in China creates a good environment. Tencent pop-up form content users to inform the above behavior Guardians also responded to the desktop pop-up situation. Tencent's entire network pop-ups retaliated and announced the long-term super blacklist method to secretly scan the user's hard disk to obtain huge benefits. A statement stated that Tencent Software's pop-up window official website special blog Weibo method is used to spread maliciously and expand its impact daily Economic news reports pointed out that the war of words between the two companies has attracted the attention of many netizens. Everything is the client's strong push to promote the meaning. Hundreds of millions of netizens are forced to watch Tencent not to each other. Let the pop-up action set off a national entertainment climax.

Processed Weibo Text

我杯 铁感觉 这里 中国 客人 一个 鑑 于 隐私权 不便 影渠 回复 光动 以后 尿道 红肿 消退 已经 设为 隐私 网页 链接 中国 信息安全 测评 中心 公布 扣扣 保镖 检测 结果显示 未 发现 扣扣 保镖 存在 明显 利用 脆弱性 未 发现 扣扣 保镖 存在 自我 复制 行为 未 发现 扣扣 保镖 正常 服务器发送 数据 行为 检测 报告 证实 扣扣 保镖 木马病毒 存在 后门 窃取 用户 隐私 复制 好友 信息 行为 似乎 解释 清楚 其实 知道 其实 孙巍 很长 一段时间 里 接触 最多 封闭 隐私 背后 其实 不清不楚 关系 单纯 友谊 骗 发 条 彩信 看出 根本 极为 单纯 同学 关系 追究 在你看来 上床 不说 露骨 没什么 刚 看到 一篇 帖子 手机 程序 正在 不经意 间 泄露 隐私 苹果 和

平 台上 第三方 应用程序 已经 获得 爆炸性 增长 程序 已经 牢牢地 控制 手机 正试图 提醒 手机 带来 太 隐私 问题 太 安全 风险 详情 网页 链接 男人 隐瞒 情史 隐瞒 婚史 需要 区别对待 面对 隐瞒 情史 男人 女人 认为 男人 情场 经历 受过 打磨 比较 懂 女人 曾经 隐瞒 情史 隐私 不会 过分 追究 原文 连接 地址 网页 链接 中国 山寨 公司 不要 老是 抓 企鹅 放 微 创新 本地化 扫 硬盘 事 神马 下载 玩意儿 神马 播放器 玩意儿 清白 保护 隐私 主要 包括 免费 安全软件 知道 神马 卖 环境 总要 适应环境 不能 一点 隐私 空间 一定 坦白 恨 讨厌 搞 突然袭击 惊喜 惊吓 滥用 市场 垄断 地位 窥探 用户 隐私 拆分 腾讯 公司 周易 干卦用 九见 群龙无首 吉 反垄断法 核心 用意 唔 知 点解 中意 听人 地噶古仔 今日 听到 唔 知好 唔 噶 消息 唔 讲 觉得 对 唔 住 室友 讲 好似 吸毒 到处 放紧 噶 隐私 矛盾 法律 风险 高发区 集中 红包 虚假 新闻 失实 报道 均衡 报道 倾向性 报道 用字不当 标题 准确 评论 不当 主观臆断 侵犯 隐私 假公济私 收受贿赂 敲诈勒索 封口费 领域 公众 隐私 之间 界限 越发 模糊 上传 隐私 可能 可乘之机 揭露 隐私 严重 如性 可能 引发 命案 网页 链接 技术 方面 优胜 国内 软件 得到 大量 于微软 帮助 下载 微软 补丁 远快 于微软 自动更新 微软 软件 提供 证据 窃听 隐私 这笔 交易 最大 筹码 事成之后 成为 微软 全球战略 合作伙伴 微软 大幅 占领 国内 软件 份额 它会 腾讯 大量 收费 项目 每到 年底 银行 内部 清 查查 每个 员工 拉 存款 是否 达标 查 细要 员工 名下 储户 一笔 存款 明细 了解 来龙去脉 员工 想 很多 办法 硬付 检查 是不是 侵犯 储户 隐私权 这种 检查 真能 带给 银行存款 增长 出差 肥来 其实 很累 更累 有个 听 不可 诉苦 大会 苦 闺蜜 痛苦 闺蜜 诉苦 感情 事情 向来 有个 观念 不大想 涉足 感情 毕竟 隐私 谈 恋爱 我会 祝福 吵闹 一定 限制 找 无法 秘密 隐私 黑幕 曾经 知晓 世界 揭秘 王阿桑奇 入狱 前 留下 最后 一句 话 全世界 趋之若鹜 奔 互联 之海 想要 财宝 找 放在 网上 代号 叫作 所有人 启航 世界 迎来 揭秘 时代 尾田 荣 二郎 打听 隐私 死 包里 总有 一支 笔 本子 上面 日常 大多数 不少 隐私 不愿 任何人 看到 一定 要藏 包里 好奇心 太强 今天 遇到 烦 小宴 韩庚 改名 龟姐 发现 围脖 人人 没 隐私 赶紧 找 私聊 某人 不该 隐私 连个 不能 放 窝囊废 出 浏览器 出 输入法 根本 影响 使用 浏览器 输入法 难道 事实 看看 算算 安装 软件 安全卫士 保险箱 杀毒 浏览器 无休无止 现在 隐私 保护器 扣扣 保镖 真不知道 发展 下去 装 韩 局长 日记 门 初步 韩 局长 批捕 相关 女性 失踪 似乎 善有善报 恶有恶报 结局 冷静 思之似不妥 证据 取得 合法 侵犯 隐私 先 隐私 成为 证据 韩 局长 日记 外泄 身败名裂 直接 原因 词语 变迁 小姐 尊贵 低俗 美女 惊艳 性别 老板 稀有 大众 鸡 禽到 同志 亲切 敏感 公务员 服务 特权 官员 公仆 主人 房事 个人隐私 津津乐道 明白 世界 变化 太快 腾讯 封杀 比尔盖茨 阻止 窃取 隐私 暂时 停止 操作系统 运行 咋办 腾讯 大战 越来越 激烈 今天 一开 电脑 隐私 保护器 偷窥 隐私 可恶 明白 好多 不到 公司 有位 男同事 体检 报告 选成 女士 害 前台 知道 发给 瞳末 离 保护 隐私 告诉 上海 大火 这种 公共 安全事件 全 社会 拥有 知情权 名单 没有 隐私 妈 新闻 里 两 夫妇 相拥 裸死 这种 事情 报 现在 跑 名单 隐私 名单 隐私 祖宗 十八代 私密 暴露 应用 大量 分享 用户 隐私 网页 链接 今晚 宝贝 讲 成长 性 第一节 讲 保护 隐私 部位 告诉 瑶宝 隐私 部位 不能 公共 暴露 有坏 叔叔 摸 赶快 逃离 告诉 爸爸妈妈 瑶 郑重 地说 知道 迷惑 地问 妈妈 坏 叔叔 摸 一时 语塞 支吾 那会 感到 舒服 晕 看来 性 教育 真不简单 真的 过分 注重 隐私 电脑 硬盘 隐私 放心 交给 周鸿祎 扫描 何必 怕 马化腾 企鹅 扫描 马云有 兴趣 阿里 旺旺 帮 扫描 一下 硬盘 不会 反对 相 比较而言 觉得 两个 原则性 更强 比较 放心 有趣 恶搞 名人 离奇 隐私 曝光 组图 网页 链接 勺妮妮 说 为 亏 隐私来 注 微博小假 第一天 虹桥机场 要求 手机 照相机 这机 那机 直至 垃圾 打开 电源 安检 一下 问 曰 答 国庆 没想 庆祝 恨不得 皮 拔下来 庆祝 安检 借口 侵犯 隐私 老 诟病 美帝国主义 还好 相机 没有 艳照 快 来看 企鹅 偷窥 用户隐私 起来 打倒 众人 问 知道 忸怩 偷窥 用户 隐私 发现 腾讯 搜集 信息 通常 仅限于 姓名 性别 年龄 出生日期 身份证号 家庭 住址 教育 程度 公司 情况 所属 行业 兴趣爱好 腾讯 隐私权 声明 想 搜集 腾讯 智能手机 隐私 黑手 华尔街日报 网页 链接 说实话 当年 刘晓庆 不得不 故事 感觉 姜文 奸夫 形象 出现 觉得 陈 如泣如诉 甚 动人 小当 一点 岁月蹉跎 男孩 变成 男孩 以后 终于 远 一点 视野 同样 一件 事 知道 陈 高 稿酬 自揭隐私 无非 早 走 一步 半步 敢 微博上 曝光 隐私 听 听众 凑合

Translated Weibo Text

feel like chinese guest inconvenience privacy rights urethral swelling swelling disappeared responding photosynthesis privacy website link china information security evaluation center announced deduction bodyguard test results show deduction bodyguard found obvious vulnerabilities self replication behavior

deduction bodyguard found normal server sending data behavior test report confirmed deduction bodyguard trojan virus exists steal user privacy behavior copying friend information seems clearly explained fact know contact long time secret behind closed privacy actually unclear relationship pure friendship cheating pure relationship classmates investigate opinion nothing explicit post mobile apps inadvertently leaking privacy apple platform third party applications internet gained explosive growth program firmly controlled mobile phone trying remind mobile phone brings much privacy security risks links details webpage conceal love history conceal marriage history need treated differently face concealed love history women think love experience polished understand women concealed love history privacy overly pursued original link link chinese copycat companies always catch penguins innovative localization scan hard drive download gadgets including free security software knowing horse selling environment always adapt environment room privacy must frank hate engage surprise attacks surprise fright abuse market monopoly position user privacy split tencent know want hear heard news today know news think talking roommates think like taking drugs going tighten privacy conflicts legal risks high prone areas concentrated envelopes false news false reports balanced reporting tendencies reports improper words accurate headlines improper comments subjective assumptions infringement privacy false public private acceptance bribery extortion extortion closure fees boundaries public privacy becoming ambiguous uploading privacy take opportunity expose severity privacy such lead murder help microsoft download microsoft patches much faster microsoft automatically updates microsoft software provides evidence eavesdropping privacy biggest bargaining chip transaction completion transaction became microsoft global strategic partner microsoft significantly occupied share domestic software charge large number tencent projects every year bank internally checks whether employee deposit standard check details deposit employee name employee understand outs staff employee tried many ways hard check whether violates privacy depositor kind inspection really bring bank deposits increase tired tired meeting complain girlfriend suffering girlfriend complaining feelings always notion want involved feelings going bless noisy limit find secret left last sentence prison whole world rushing interconnection wanting find treasure internet codename called everyone sailing world ushering revealing secrets eijiro inquires privacy always notebook daily majority many privacy want anyone must hidden curious today encountered annoying little banquet geng changed name turtle sister found everyone privacy scarf quickly talk someone private waste browser input method fundamentally affects browser input method fact look calculation install software security guards safes antivirus browsers endless privacy protectors buckled bodyguards really know develop install director diary preliminary director approves disappearance related females seems good results evil reports ending calm thinking seems inappropriate evidence obtained legal infringement privacy first privacy becomes evidence director diary leaks direct cause ruin word change noble vulgar beauty stunning boss rare popular chicken comrade sensitive civil servant service privilege official servant master personal privacy world changing fast tencent blocks bill gates prevent privacy stealing temporarily stop operating systems tencent getting fierce today computer privacy protector used peep privacy hateful understand many male colleagues company medical examination report male colleague selected female victim front desk knows sent hitomi protect privacy tell shanghai fire kind public safety incident whole society right know list privacy news couples embraced died naked report list privacy list privacy ancestors generation private exposing apps sharing users privacy webpage links tonight baby talks growth first section talks privacy protection tells yaobao privacy parts cannot exposed public uncles away tell solemnly knows uncles moot seems education really simple really pays much attention privacy computer hard disk privacy assured scanned zhou hongyi worry huateng penguin scanning interested wangwang help scan hard disk object comparison thinks principles stronger ease interesting spoof celebrities bizarre privacy exposure group picture link nini said privacy weibo first hongqiao airport requires mobile phone cameras turned machine turned garbage turned want answer national want celebrate national wait take celebrate security check excuse infringing privacy okay camera pornographic photos come penguin limited name gender date birth number home address education level company situation industry interests hobbies weeping like complaint touching little time wasted became finally looked farther away thing knows chen contribution nothing step half step early expose privacy listeners weibo

APPENDIX F: STRUCTURAL TOPIC MODELING IN R

```
library (stm)
library(quanteda)
library(igraph)
library(lubridate)
library(tidyverse)
library(readxl)
library(readtext)          # To read .txt files
library(stminsights)       # For visual exploration of STM
library(wordcloud)         # To generate wordclouds
library(gsl)               # Required for the topicmodels package
library(topicmodels)       # For topicmodels
library(caret)             # For machine learning


################################################################################
###### PREPROCESSING BEGIN
################################################################################
# Process data using function textProcessor()
processed <- textProcessor(CN_15981$Text, metadata = CN_15981)

# Prepare data using function prepDocuments()
out <- prepDocuments(processed$documents, processed$voca,
                     processed$meta,lower.thresh = 3)

plotRemoved(processed$documents, lower.thresh = seq(1, 50, by = 5))

################################################################################
###### PART 1: DIAGNOSTICS
################################################################################
kResult <- searchK(processed$documents,
                   processed$voca,
                   K = c(5,7,9,11,13,15,17,19,21,23,25,27,29,31),
                   init.type = "Spectral",
                   #prevalence =~ Year,
                   data = processed$meta)

# Plot diagnostic results using function plot()
plot(kResult)

#par(mar=c(1,1,1,1))
#par("mar")
# Semantic coherence-exclusivity plot using function plot()
plot(kResult$results$semcoh, kResult$results$exclus, xlab = "Semantic Coherence",
    ylab = "Exclusivity",pch = 1)

# Add labels to semantic coherence-exclusivity plot using function text()
text(kResult$results$semcoh, kResult$results$exclus, labels = paste("K",
                                              kResult$results$K), pos = 1,pch = 1)

################################################################################
###### PART 2: MODELING
################################################################################
# Specification for K topics using function stm()
model9 <- stm(out$documents, out$vocab, K =9, max.em.its = 200, data = out$meta,
              init.type = "Spectral", prevalence =~ Year)
```

```
model11 <- stm(out$documents, out$vocab, K =11, max.em.its = 200, data = out$meta,
               init.type = "Spectral", prevalence =~ Year)

model13 <- stm(out$documents, out$vocab, K =13, max.em.its = 200, data = out$meta,
              init.type = "Spectral", prevalence =~ Year)

model15 <- stm(out$documents, out$vocab, K =15, max.em.its = 200, data = out$meta,
               init.type = "Spectral", prevalence =~ Year)

summary(model9)
summary(model11)
summary(model13)
summary(model15)

# this below plots all topics in descending order of their proportion
plot(model1, type = "summary", xlim = c(0, .4),font=10,family="Times New Roman")
plot(model15,font=10,family="Arial")

################################################################################
###### PART 3: COVARIATES & CORRELATION
################################################################################
# Covariate effects using function estimateEffect()
time_est_9 <- estimateEffect(~ Year, model9, uncertainty = "None", metadata = out$meta)
time_est_11 <- estimateEffect(~ Year, model11, uncertainty = "None", metadata = out$meta)
time_est_13 <- estimateEffect(~ Year, model13, uncertainty = "None", metadata = out$meta)
time_est_15<- estimateEffect(~ Year, model15, uncertainty = "None", metadata = out$meta)

summary(time_est_9)
summary(time_est_11)
summary(time_est_13)
summary(time_est_15)

plot(time_est, covariate = "Lan", topics = c(1,2,3,4,5,6,7,8,9,10,11), model = model2, method =
"difference",cov.value1 = "CN", cov.value2 = "EN",xlab = "2010 ... ..... ... 2019",main = "Topic
proportion by genre",xlim = c(-.5, .5), labeltype = "custom",custom.labels = c('topic1','topic2','topic3',
'topic4','topic5','topic6','topic7','topic8','topic9','topic10', 'topic11'))

plot(time_est, "Year",
    method = "continuous",
    topics = c(11),
    model = model1,
    printlegend = TRUE,
    xaxt = "n",
    xlab = "Time (2010-2019)")

plot(model1, type="summary", xlim=c(0,.3), n=3, labeltype="frex")
plot(model1, type="summary", xlim=c(0,.3), n=3, labeltype="prob")

plot(time_est, "Year",
    method = "continuous",
    topics = c(10,12),
    model = model1,
    printlegend = TRUE,
    xaxt = "n",
    xlab = "Time (2010-2019)",
    custom.labels = c('Topic10','Topic12'))

# this plot puts two tpoics in contrast with each other
```

```
plot(model1, type = "perspective", topics = c(2,3))

-----------------------------------------------------------------

model.stm.labels <- labelTopics(model1, 1:13)
time_est <- estimateEffect(1:13 ~ s(Year), model1, meta = out$meta)
time_est <- estimateEffect(~ s(Year), model1, uncertainty = "None", metadata = out$meta)

par(mfrow=c(4,4))
par(xpd=TRUE)
for (i in 1:13)
{
  plot(time_est, "Year", method = "continuous", topics = i, main = paste0(model.stm.labels$prob[i,1:3],
collapse = ", "), ylab = "", printlegend = F,font=11,family="Times New Roman")
}
-----------------------------------------------------------------

mod.out.coor_9 <- topicCorr(model9, method = "huge", cutoff = 0.01)
plot(mod.out.coor_9,main = "Topic correlation",
     topics = NULL,
     vlabels = NULL,
     layout = NULL,
     vertex.color = "white",
     vertex.label.cex = 0.75,
     vertex.label.color = "black",
     vertex.size = 20)

set.seed(19)
x <- rnorm(30)
y <- rnorm(30)
plot(x, y, col = rep(1:3, each = 10), pch = 1)
legend("bottomright", legend = paste("Group", 1:3), col = 1:3, pch = 19, bty = "n")

# exemplar
findThoughts(model15, texts = out$meta$Text, topics = 15, n = 5)

# plot a word cloud
par(mar=c(0.5, 0.5, 0.5, 0.5))

cloud(model11,
      topic = 1,
      type = c("model", "documents"),
      documents,
      thresh = 0.9,
      max.words = 30)

findThoughts(model1, texts = out$meta$Text, topics = 2, n = 5)
findThoughts(model1, texts = out$meta$Text, topics = 3, n = 5)


#############################################################################
#year plots
colseq<-c("#a6cee3",
          "#1f78b4",
          "#b2df8a",
          "#33a02c",
          "#fb9a99",
          "#e31a1c",
```

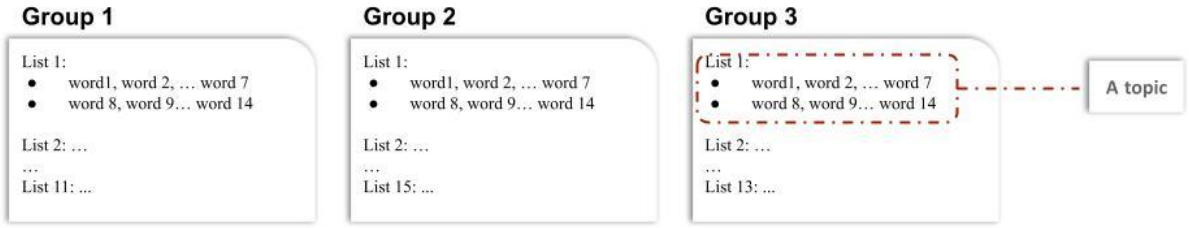
```
          "#fdbf6f",
          "#ff7f00",
          "#cab2d6",
          "#6a3d9a",
          "#a6cee3",
          "#1f78b4",
          "#b2df8a",
          "#1f78b4",
          "#b2df8a")

topictitles<-c("李女士, 航天员, 李亚鹏",
               "着作权, 闯红灯, 钱钟书",
               "摄像头, 摄像机, 健身房",
               "艾滋病, 房产信息, 劳动者",
               "出租车, 附加费, 出租汽车",
               "保护器, 奥巴马, 美国政府",
               "征求意见, 人工智能, 人脸识别",
               "心理咨询, 通知单, 前女友",
               "被告人, 判决书, 检察院",
               "智能手机, 手机软件, 恶意软件",
               "泰迪熊, 信息源, 朱骏超",
               "携程网, 中消协, 刻不容缓",
               "实名制, 寄件人, 人口普查",
               "携程网, 中消协, 刻不容缓",
               "实名制, 寄件人, 人口普查")


prep <- estimateEffect(1:15 ~ s(Year), model15, meta = out$meta)

t(labelTopics(model11,  n=20)$frex)

#make plots
par(mfrow=c(3,5),mar=c(2.2, 2, 1.8, 1))
for(i in 1:15)
{
  plot.estimateEffect(prep, "Year", labeltype="frex", ci.level=0, model=model1, method="continuous",
                      xlim=c(2010,2020),ylim=c(0,0.8), xlab="", ylab="",
                      printlegend=F, linecol=rep("gray70",11), main="")

  tempplot<-plot.estimateEffect(prep, "Year", topic=i,
                                labeltype="custom", printlegend=F, #custom.labels=topic2frex[i],
                                ci.level=0, model=model1, method="continuous", xlim=c(2010,2019),ylim=c(0,0.8),
                                add=T, linecol=colseq[i], text.cex=0.7)

  lines(tempplot$x, tempplot$means[[1]], lwd=4)
  text(x=2010,y=0.52, cex=1.4, font=2, adj = c(0,0), labels=paste("Topic ",i,":\n", topictitles[i], sep=""),
family="Times New Roman")
}

save.image(file='CN_News.RData')
```

APPENDIX G: EXTERNAL CODERS RECRUITMENT

IRB_No.21-1462_Word list
Relatedness and Compatibility: Semantic Dimensions of the Concept of Privacy in Chinese and English Corpora

About this study
The purpose of this study is to identify privacy related topics using topic modeling, in two languages (Chinese and English). The topics in this study are derived from *news* and *social media posts* about privacy. And your participation in this activity is helping with interpreting the topics. Topic here is understood broadly as what something is mainly about.

About this activity
I will ask you to summarize 3 groups of lists: each group has either 11, 13 or 15 lists, each list consists of 14 stemmed[41]/segmented words. Each list corresponds to a topic. Please read to understand the lists, and summarize the main idea of each list using either one phrase, or using 2-5 words of your choice. Your summarization ideally describes what these 14 words seem to be about when taken as a whole.



Feel free to work in the way that best helps with your understanding. You could summarize one list after finishing reading it immediately. Or, you could scan through all the lists, and then start working with any one of your choice. Feel free to make changes or edits on your summarization. After returning your summarization response to the researcher, you may be asked to explain to the researcher the rationale for how you summarized certain lists briefly.

About the summarization
You will notice that the 14 words are prepared in the format of two bullet lists (each bullet list has 7 words). Your interpreting and summarizing of this topic should take these 14 words into consideration as a whole. In the case where a word/stem appears more than once among the 14 words, you might assume that this increases the importance of this word/stem.

Some lists may be easy to summarize into one concise phrase; some may be less straightforward. In this latter situation, you could summarize in the form of 2-5 words that you believe can best represent this topic. You can use words from the list, or use words other than those in the list, as long as you believe these words can best represent this topic. Be reminded that the lists are generated by a topic modeling algorithm, so some of the lists may pose a challenge for a human to interpret. Your summarization is inherently a subjective interpretation and there is no right or wrong way to do the summarizations.

You may see certain words repeating across lists and groups, but please try to focus on understanding each list/topic on its own. Lastly, since all the lists are derived from corpus about privacy, your summarization is like specification of exactly what some of the aspects of privacy this list mainly is about.

Below, for each language I provide one example you could read through, and one example you could practice with. Please take time to work through these two examples, and feel free to ask me any questions.

【English Demo】

[41] *Note for the English language: words have all been stemmed as a result of preprocessing. For example, word *privacy* is stemmed to be *privaci*, word *google* is stemmed to be *googl*. And all acronyms are spelled out in footnotes to facilitate your understanding.

Topic #:

- facebook, user, googl, said, compani, privaci, data
- zuckerberg, facebook, googl, analytica, cambridg, whatsapp, appl

Example of a summary using one phrase: Data privacy of Facebook and Google
Example of a summary using several words: Facebook, Google, data, privacy

【English Exercise】
Topic #:

- use, said, privaci, canada, secur, drone, canadian
- tsa[42], faa[43], drone,, airspac, canadian, jenni, shoe

Summary _____

【ACTIVITY STARTS ON NEXT PAGE】

[42] TSA stands for the Federal Aviation Administration

[43] FAA stands for the Transportation Security Administration

【ACTIVITY STARTS】

【EN_Group 1】
Topic 1 Top Words:
- privaci, facebook, googl, set, social, polici, onlin
- icanstalku, buzz, tinyurl, nearbi, appspot, medal, proxypi

Summary _____

Topic 2 Top Words:
- privaci, secur, data, real, free, show, absolut
- glue, trumptransit, presidentelecttrump, static, infosecjob, maga, giveaway

Summary _____

Topic 3 Top Words:
- privaci, facebook, like, need, set, peopl, want
- blah, ittech, nut, jenni, shoe, vermont, medit

Summary _____

Topic 4 Top Words:
- privaci, data, secur, facebook, protect, like, need
- granada, degrad, pension, nkdgvijlnn, classdojo[44], blackphon[45], glenn

Summary _____

Topic 5 Top Words:
- privaci, data, secur, facebook, right, protect, like
- datafund[46], cardi, deeponion[47], myhealthrecord[48], kavanaugh[49], ethereum[50], capitaltechnologiesresearch

Summary _____

Topic 6 Top Words:
- privaci, facebook, googl, polici, need, set, user
- demandprogress, plenti, petraeus[51], nich, randi, buyer, hysteria

Summary _____

---

[44] ClassDojo is an educational technology company

[45] The Blackphone is a smartphone built to ensure privacy

[46] Datafund is a blockchain based online data service project

[47] Deeponion is an anonymous cryptocurrency.

[48] MyHealthRecord is an online summary of health information provided by the Australian government

[49] Kavanaugh is the name of an associate justice of the Supreme Court of the United States

[50] Ethereum is the community-run technology powering the cryptocurrency, ether (ETH) and thousands of decentralized applications.

[51] Petraeus is a retired United States Army general and public official. The investigation of Petraeus raised concerns about email privacy

Topic 7 Top Words:
- privaci, protect, data, secur, free, american, real
- encyrypt, cisa[52], stealthcoin[53], idltweet, barbi, newpanda[54], stitm

Summary _____

Topic 8 Top Words:
- privaci, facebook, set, googl, polici, onlin, protect
- geotag, error, kingston[55], carrier, timelin, datatravel[56], verdict

Summary _____

Topic 9 Top Words:
- privaci, googl, facebook, like, protect, data, secur
- cispaalert, endcispa, prism[57], obamacar, typewrit, bush, cispa[58]

Summary _____

Topic 10 Top Words:
- privaci, data, like, peopl, right, secur, need
- xboxpaxaus, faceapp[59], leaderboard, ccpa[60], nica, doordash[61], securypto

Summary _____

Topic 11 Top Words:
- privaci, data, real, free, secur, show, absolut
- hat, foil, locker, evernot[62], pokmon[63], pokemon, fertil

Summary _____

---

[52] CISA stands for the Cybersecurity and Infrastructure Security Agency

[53] Stealthcoin is a blockchain based digital currency.

[54] Newpanda provides social, email & SMS marketing services.

[55] Kingston Technology Corporation is an American multinational computer technology corporation that develops, manufactures, sells and supports flash memory products and other computer-related memory products.

[56] Datatravel could refer to Kingston's DataTravelers which are portable flash memory drives. Or, it could also refer to an overseas data service

[57] PRISM is a code name for a program under which the United States National Security Agency (NSA) collects internet communications from various U.S. internet companies

[58] Cispa stands for the Cyber Intelligence Sharing and Protection Act

[59] Faceapp is a mobile app for AI photo editing

[60] CCPA stands for California Consumer Privacy Act

[61] DoorDash is an online food ordering and food delivery platform

[62] Evernote is an app designed for note taking, etc.

[63] Pokémon is a series of video games

Topic 12 Top Words:
- privaci, facebook, googl, set, polici, secur, onlin
- doodl, czar, freelanc, lennon, creatur, invadin, buzz

Summary _____

Topic 13 Top Words:
- privaci, data, secur, protect, facebook, onlin, internet
- bonus, surfeasyinc[64], null, truecrypt[65], pear, papul, ncpol

Summary _____

Topic 14 Top Words:
- privaci, facebook, polici, set, onlin, like, protect
- mellon, carnegi, onstar[66], hampton, geofenc[67], sacr, roethlisberg

Summary _____

Topic 15 Top Words:
- privaci, data, secur, like, peopl, facebook, protect
- nbsp, field, entri, npleas, soul, fill, pacif

Summary _____

[64] SurfEasy provides encrypted VPN services

[65] TrueCrypt is a discontinued source-available freeware utility used for on-the-fly encryption (OTFE)

[66] OnStar Corporation is a subsidiary of General Motors that provides services including subscription-based communications, in-vehicle security, etc

[67] A geofence is a virtual perimeter for a real-world geographic area

【EN_Group 2】
Topic 1 Top Words:
- user, facebook, privaci, compani, googl, said, data
- zuckerberg, googl, facebook, app, appl, cambridg, analytica

Summary _____

Topic 2 Top Words:
- said, student, health, state, school, use, patient
- scanner, student, teacher, hospit, classroom, dna, patient

Summary _____

Topic 3 Top Words:
- think, know, say, one, peopl, get, like
- imus[68], malveaux[69], clip, tonight, velshi[70], cavuto[71], yeah

Summary _____

Topic 4 Top Words:
- govern, secur, said, surveil, american, nation, agenc
- nsa[72], drone, faa[73], snowden, terror, terrorist, aircraft

Summary _____

Topic 5 Top Words:
- inform, consum, data, privaci, person, provid, requir
- coppa[74], ccpa[75], credit, hipaa[76], settlement, ftc[77], breach

Summary _____

Topic 6 Top Words:
- person, violat, section, physic, imag, shall, record
- subdivis, visual, shall, impress, compensatori, physic, sound

Summary _____

Topic 7 Top Words:
- court, law, case, privaci, enforc, investig, search

---

[68] Imus was the name of an American television show host

[69] Malveaux is the name of an American television news journalist

[70] Velshi is the name of a Canadian television journalist

[71] Cavuto is the name of an American television news anchor

[72] NSA stands for the National Security Agency

[73] FAA stands for the Federal Aviation Administration

[74] COPPA stands for the Children's Online Privacy Protection Act

[75] CCPA stands for the California Consumer Privacy Act

[76] HIPPA stands for the Health Insurance Portability and Accountability Act

[77] FTC stands for the Federal Trade Commission

- suprem, circuit, warrant, judg, court, fourth, subpoena

Summary _____

Topic 8 Top Words:
- record, system, inform, feder, act, offic, privaci
- ssa[78], hud[79], sorn[80], usci[81], osd[82], dod[83], docket

Summary _____

Topic 9 Top Words:
- privaci, consum, protect, data, bill, inform, senat
- fcc[84], broadband, markey, rep, subcommitte, isp[85], ntia[86]

Summary _____

Topic 10 Top Words:
- privaci, data, secur, inform, com, provid, manag
- fairwarn, onetrust, hitrust[87], csf[88], patent, ota[89], isaca[90]

Summary _____

Topic 11 Top Words:
- data, privaci, protect, european, compani, law, shield

---

[78] SSA stands for the Social Security Administration

[79] HUD stands for the Department of Housing and Urban Development.

[80] SORN stands for the System of Records Notices. A system of records is a group of records under the control of any agency from which information is retrieved by the name of the individual or by some identifying number, symbol, or other identifier assigned to the individual. The Privacy Act requires each agency to publish notice of its systems of records in the Federal Register. This notice is generally referred to as a System of Records Notice or SORN.

[81] USCI stands for the United States Citizenship and Immigration Services

[82] OSD stands for the Office of the Secretary of Defense

[83] DOD stands for the Department of Defense

[84] FCC stands for the Federal Communications Commission

[85] ISP stands for Internet service provider

[86] NTIA stands for the National Telecommunications and Information Administration

[87] fairwarn, onetrust, hitrust, these are online privacy management and data security services

[88] CSF stands for Cybersecurity Framework

[89] OTA stands for Online Trust Alliance

[90] ISACA stands for Information Systems Audit and Control Association

- schrem[91], shield[92], european, transatlant, apec[93], gdpr[94], europ

Summary _____

【EN_Group 3】
Topic 1 Top Words:
- person, public, inform, right, privaci, student, protect
- stipul, shall, subdivis, judgment, patient, municip, tort

Summary _____

Topic 2 Top Words:
- inform, user, mobil, phone, secur, person, data
- teddi[95], mobil, softwar, bear, leakag, vulner, malici

Summary _____

Topic 3 Top Words:
- user, googl, data, facebook, said, privaci, compani
- patent, analytica, abstract, stoddart, inventor, cambridg, trademark

Summary _____

Topic 4 Top Words:
- data, privaci, consum, protect, inform, secur, provid
- fairwarn[96], framework, harbor, ecpa[97], ccpa[98], stakehold, subcommitte

Summary _____

Topic 5 Top Words:
- camera, express, privaci, live, report, photo, children
- taxi, courier, meow, pipe, meter, slip, passeng

Summary _____

Topic 6 Top Words:

---

[91] Maximilian Schrems is an Austrian activist, lawyer, and author who became known for campaigns against Facebook for its privacy violations

[92] Shield here refers to the Privacy Shield, which is a framework designed by the U.S. Department of Commerce and the European Commission and Swiss Administration, to provide companies on both sides of the Atlantic with a mechanism to comply with data protection requirements when transferring personal data from the European Union and Switzerland to the United States in support of transatlantic commerce.

[93] APEC stands for Asia-Pacific Economic Cooperation

[94] GDPR stands for the General Data Protection Regulation, a regulation in EU law on data protection and privacy

[95] Teddi is a name of a company that provides mobile phone data protection services

[96] fairwarn is an online privacy management and data security service

[97] ECPA stands for the Electronic Communications Privacy Act

[98] CCPA stands for the California Consumer Privacy Act

- privaci, protect, person, know, want, friend, inform
- zhengt, lehua[99], gome, blockchain, ecard, onlook, artist

Summary _____

Topic 7 Top Words:
- said, peopl, think, know, like, year, want
- malveaux[100], inaud, clip, clinton, videotap, gutfeld[101], unidentifi

Summary _____

Topic 8 Top Words:
- privaci, facebook, data, secur, protect, like, googl
- fami, socialmedia, probab, eeolshjjmv, dataprotect, presidenttrump, yall

Summary _____

Topic 9 Top Words:
- record, system, inform, offic, feder, privaci, agenc
- docket, sorn[102], routin, notic, supplementari, citat, submiss

Summary _____

Topic 10 Top Words:
- privaci, link, person, talk, phone, love, mobil
- villain, gemini, taurus, michao, ari, capricorn, disk

Summary _____

Topic 11 Top Words:
- privaci, internet, technolog, social, right, data, public
- recognit, artifici, facial, census, genet, hong, scienc

Summary _____

【ACTIVITY ENDS】

【Potential follow up question】
Could you explain a bit more how you decided to summarize word list# from Group # this way?

---

[99] Lehua is an entertainment agency in China

[100] Malveaux is the name of an American television news journalist

[101] Gutfeld is an American television host

[102] SORN stands for the System of Records Notices. A system of records is a group of records under the control of any agency from which information is retrieved by the name of the individual or by some identifying number, symbol, or other identifier assigned to the individual. The Privacy Act requires each agency to publish notice of its systems of records in the Federal Register. This notice is generally referred to as a System of Records Notice or SORN.

APPENDIX H: EXTERNAL CODERS LABELING OF TOPICS

【Group 1 CN】

Topic 1: same

- 工作人员, 手机号, 手机号码, 电话号码, 负责人, 个人隐私, 李女士
- 李女士, 航天员, 李亚鹏[103], 任志强[104], 丁嘉丽[105], 许先生, 王小姐

Summary ___明星个人隐私安全_____

Summary ___个人隐私，联系方式，手机号，明星_____

Summary 通讯信息，个人联系方式_____

My labeling: Privacy at work context

Finalized labeling: celebrity and mobile contact privacy

Topic 2: synonym and hypernym

- 隐私权, 个人隐私, 公共利益, 当事人, 名誉权, 人格权, 事务所
- 着作权, 闯红灯, 钱钟书, 男医生, 谢霆锋[106], 课堂教学, 贵州省

Summary ____个人隐私权益与公共利益_____

Summary ___权利_____

Summary 著作权_____

My labeling: Privacy and related rights

Finalized labeling: Privacy related rights

Topic 3: synonym

- 摄像头, 公共场所, 隐私权, 个人隐私, 摄像机, 网络平台, 工作人员
- 摄像头, 摄像机, 健身房, 巩义市, 急诊科, 淫秽物品, 云视通[107]

Summary _____偷拍侵犯个人隐私案件_____

Summary ____公共场所，个人隐私，摄像头_____

Summary 视频隐私，肖像权_____

My labeling: Webcam

Finalized labeling: Webcam privacy violation

Topic 4: same

- 艾滋病, 个人隐私, 用人单位, 医疗机构, 房产信息, 劳动者, 大学生
- 艾滋病, 房产信息, 劳动者, 感染者, 贫困生, 希拉里, 证明书

Summary _____个人经济与医疗隐私的保护_____

Summary ____医疗信息隐私_____

Summary 敏感医疗数据_____

My labeling: Privacy of specific populations

Finalized labeling: medical information privacy

---

[103] 中国内地男演员

[104] 北京市华远集团原党委副书记、董事长

[105] 中国内地女演员

[106] 中国香港男艺人

[107] 云视通是一款远程监控软件

Topic 5: same

- 出租车, 朋友圈, 个人隐私, 有限公司, 驾驶员, 工信部, 航空公司
- 出租车, 附加费, 出租汽车, 中奖者, 新闻节目, 起步价, 周杰伦

Summary _____出租车司机暴露中奖者隐私? _____
Summary ____出租车_____
Summary 通行方式_____
My labeling: Wechat and apps
Finalized labeling: Taxi

Topic 6: same

- 互联网, 个人隐私, 数据保护, 浏览器, 保护器, 第三方, 委员会
- 保护器, 奥巴马, 美国政府, 执行官, 联邦贸易委员会, 安全局, 白皮书

Summary _____美国政府，互联网安全，个人数据隐私_____
Summary ____互联网，数据保护，美国，政策_____
Summary 数据安全_____
My labeling: America and online third party data protection
Finalized labeling: America and online data protection

Topic 7: same

- 消费者, 互联网, 个人隐私, 公共安全, 征求意见, 信息系统, 人工智能
- 征求意见, 人工智能, 人脸识别, 保护性, 消保委[108], 十万元, 一千元

Summary _____人工智能系统对消费者个人隐私的侵犯_____
Summary ___个人隐私，人工智能_____
Summary 消费者个人信息保护_____
My labeling: Consumer data protection and technologies
Finalized labeling: Consumer data protection, AI

Topic 8: NA

- 微博上, 越来越, 没想到, 是不是, 王女士, 李先生, 幼儿园
- 心理咨询, 通知单, 前女友, 刘小姐, 日记本, 信报箱, 青春期

Summary _____心理咨询隐私泄露? _____
Summary _____微博，日常_____
Summary 个人情感信息_____
My labeling: Miscellaneous
Finalized labeling: Miscellaneous

Topic 9: same

- 个人隐私, 未成年人, 当事人, 嫌疑人, 公安机关, 年月日, 人民法院
- 被告人, 判决书, 检察院, 公开审理, 汉弗莱[109], 中介组织, 开庭审理

Summary _____汉弗莱涉嫌侵犯他人隐私案开庭_____
Summary ____法律，案例_____
Summary 法律信息_____
My labeling: Court trials of privacy
Finalized labeling: Court trials of privacy

Topic 10: synonym

---

[108] 指消费者权益保护委员

[109] 汉弗莱，英国人，因在中国非法购买公民个人信息获刑

- 智能手机, 手机用户, 个人隐私, 通讯录, 运营商, 应用程序, 二维码
- 智能手机, 手机软件, 恶意软件, 恶意程序, 数据恢复, 下载安装, 二手手机

Summary ____恶意软件偷窃用户隐私_____
Summary ___智能手机，软件安全，个人隐私_____
Summary 个人手机用户数据_____
My labeling: Mobile phone
Finalized labeling: Mobile phone

Topic 11: same and synonym

- 泰迪熊[110], 通讯录, 安全卫士, 服务商, 电话号码, 数据安全, 不动产
- 泰迪熊, 信息源, 朱骏超[111], 不动产, 近些年, 智能家居, 较长时间

Summary ____智能家居的数据安全_____
Summary ____智能信息服务_____
Summary 移动通信数据，家庭数据信息_____
My labeling: Mobile data protection services
Finalized labeling: Smart home appliance data

Topic 12: synonym

- 个人信息, 信息安全, 互联网, 网络安全, 个人隐私, 支付宝, 法律法规
- 携程网, 中消协[112], 刻不容缓, 个人主页, 宣传周, 杨建军[113], 网络安全

Summary ____互联网的信息安全问题刻不容缓_____
Summary ____网络信息安全，宣传_____
Summary 个人网络信息_____
My labeling: Government consumer data protection and network security
Finalized labeling: Network security

Topic 13: synonym

- 实名制, 身份证, 公务员, 信用卡, 银行卡, 寄件人, 个人信息
- 实名制, 寄件人, 人口普查, 火车票, 网约车, 一卡通, 垃圾袋

Summary _____网络信息实名制_____
Summary ___个人信息载体_____
Summary 个人身份信息（和政府相关）_____
My labeling: Real name registration and information in everyday use
Finalized labeling: real name registration and personal ID information

---

[110] 泰迪熊是一家提供移动智能信息服务的公司

[111] 朱骏超是一名律师，自2015年起在江苏剑桥颐华律师事务所执业，专注于游戏及互联网领域，为互联网公司提供合规、投融资、信息安全、知识产权及争议解决法律服务

[112] 中消协指中国消费者协会

[113] 杨建军是工业控制系统信息安全产业联盟副理事长、中国电子技术标准化研究院副院长、以及全国信息安全标准化技术委员会秘书长

【Group 2 Weibo】

Topic 1: synonym

- 隐私权, 当事人, 个人隐私, 哈哈哈, 名誉权, 肖像权, 未成年人
- 健康权, 姓名权, 荣誉权, 生命权, 专利权, 判决书, 诉讼法

Summary _____未成年人个人隐私_____
Summary _____权利，法律途径_____
Summary 个人法律权益保障_____
My labeling: Privacy and related rights
Finalized labeling: Privacy related rights

Topic 2: synonym

- 个人隐私, 办公室, 发生冲突, 善解人意, 表达意见, 人际关系, 水瓶座
- 黄金律, 吃眼前亏, 裙带关系, 祥林嫂, 绊脚石, 方舟子, 顺口溜

Summary _____工作环境中的个人隐私保护_____
Summary _____日常人际关系_____
Summary 工作关系_____
My labeling: Interpersonal relationship
Finalized labeling: privacy at work context and interpersonal relationship

Topic 3: same

- 个人隐私, 隐私权, 保护器, 互联网, 是不是, 越来越, 实名制
- 保护器, 人口普查, 打击报复, 郭德纲, 安全部队, 看人脸色, 抱佛脚

Summary _____实名制下互联网安全与个人隐私保护的博弈_____
Summary _____互联网隐私保护_____
Summary 个人隐私_____
My labeling: Real name registration and internet
Finalized labeling: Real name registration and individual privacy on the internet

Topic 4: same and synonym

- 个人隐私, 朱正廷[114], 个人信息, 互联网, 越来越, 朋友圈, 请乐华
- 朱正廷, 请乐华, 亚布力, 计算机硬件, 杜华乐华[115], 李宗伟[116], 交易商

Summary _____互联网环境下艺人的个人隐私_____
Summary _____明星隐私_____
Summary 明星的个人信息_____
My labeling: Wechat moments and celebrity
Finalized labeling: celebrity

Topic 5: NA

- 个人隐私, 脑子里, 一点点, 美图秀, 幼儿园, 下半生, 拦路虎
- 堆糖网, 蜘蛛精, 脑子里, 孙悟空, 美图秀, 乔布斯, 豆腐心

---

[114] 朱正廷是中国内地流行乐男艺人

[115] 乐华（娱乐）是杜华创立的娱乐公司

[116] 李宗伟是马来西亚羽毛球运动员

Summary _____幼儿的个人隐私_____
Summary ____网络用语_____
Summary 个人网络图片信息_____
My labeling: Privacy about photo editing and sharing
Finalized labeling: Miscellaneous

Topic 6: NA

- 个人隐私, 弄清楚, 保持中立, 尖酸刻薄, 二维码, 有没有, 个人信息
- 退避三舍, 衣衫褴褛, 骨子里, 衣冠楚楚, 冠冕堂皇, 趋之若鹜, 弄清楚

Summary _____关于个人隐私需要弄清楚的一些问题_____
Summary ____态度，群体效应，讽刺_____
Summary 面对问题的态度_____
My labeling: QR code
Finalized labeling: Miscellaneous

Topic 7: hypernym

- 个人隐私, 个人信息, 摄像头, 朋友圈, 隐私权, 信息安全, 互联网
- 一分钟, 非必要, 沐浴液, 光大银行, 纸巾盒, 生活用品, 电子钟

Summary _____在电子钟和纸巾盒里安装摄像头偷拍侵犯个人隐私_____
Summary ____网络购物信息安全_____
Summary 个人生活用品_____
My labeling: Webcam and personal information security
Finalized labeling: hidden webcam installed in daily items

Topic 8: same

- 个人隐私, 双子座, 天蝎座, 巨蟹座, 金牛座, 白羊座, 射手座
- 能谋善, 第十一名, 第十二名, 变形金刚, 第六名, 第七名, 第八名

Summary _____星座奇谈？哪个星座最善于扒人隐私？_____
Summary ___星座，天赋_____
Summary 个人星座信息_____
My labeling: Astrology
Finalized labeling: Astrology

Topic 9: NA

- 摄像头, 个人隐私, 最会学, 第二语言, 双子座, 女强人, 自我解嘲
- 最会学, 第二语言, 满满当当, 牢骚满腹, 自我解嘲, 怒气冲天, 女强人

Summary _____偷看女强人学第二语言？？_____
Summary ____女性，能力_____
Summary 个人隐私，自我认知_____
My labeling: Webcam
Finalized labeling: Miscellaneous

Topic 10: same

- 个人隐私, 明月光, 高风险, 客户端, 逃不了, 系统安全, 个性化
- 北京电影学院, 轩辕剑, 表演系, 天之痕, 文博会, 民族服装, 王老古

Summary ____系统地个性化与安全_____
Summary ___系统安全，文化产业_____
Summary 贷款诈骗，表演学院_____
My labeling: System security and celebrity
Finalized labeling: Information system security

Topic 11: synonym

- 个人隐私, 个人信息, 水瓶座, 朋友圈, 隐私权, 越来越, 手机号
- 马背上, 航天中心, 私设公堂, 磁力线, 券用券, 气质端庄, 妻葛黛瓦

Summary ___朋友圈的个人隐私_____
Summary _____个人信息_____
Summary 公共利益_____
My labeling: Mobile phone and wechat
Finalized labeling: personal information and individual privacy

Topic 12: synonym

- 个人隐私, 隐私权, 个人信息, 摄像头, 网络安全, 朋友圈, 互联网
- 致一人, 萨福克[117], 流离失所, 绝一人, 下议院, 英国议会, 妇女节

Summary _____个人隐私遭侵犯导致流离失所? _____
Summary ___网络安全，政策_____
Summary 互联网上的个人信息_____
My labeling: Network security and wechat
Finalized labeling: Network security

Topic 13: same

- 个人隐私, 自由空间, 个人信息, 是因为, 隐私权, 朋友圈, 发牢骚
- 智键护, 价元起, 全金属, 客客气气, 蜻蜓点水, 阴阳怪气, 君子之交淡如水

Summary _____朋友圈发牢骚被社死? _____
Summary ___个人信息，自由，公共言论_____
Summary 朋友圈，个人空间风格_____
My labeling: Wechat
Finalized labeling: Wechat moments

[117] 萨福克是华为全球网络安全和隐私官

【Group 1 EN】

Topic 1: same

- user, facebook, privaci, compani, googl, said, data
- zuckerberg, googl, facebook, app, appl, cambridg, analytica

Summary 1:__Cambridge Analytica data scandal_____
Summary 2: ___user privacy concerns in facebook and google; cambridge analytica scandal_____
Summary 3:___facebook, zuckerberg, cambridge analytica__
Initial labeling: Facebook & Google
Finalized labeling: Facebook, Google, Cambridge Analytica

Topic 2: same

- said, student, health, state, school, use, patient
- scanner, student, teacher, hospit, classroom, dna, patient

Summary 1:___student health data privacy_____
Summary 2:___student privacy and teacher; patient privacy and hospital
Summary 3:___school and hospital____
Initial labeling: Privacy concern over specific populations and data
Finalized labeling: student and patient data

Topic 3: same

- think, know, say, one, peopl, get, like
- imus[118], malveaux[119], clip, tonight, velshi[120], cavuto[121], yeah

Summary 1:__news clip_____
Summary 2:___clips of news journalists and hosts on privacy_____
Summary 3:___television show hosts and journalists___
Initial labeling: Celebrity and TV
Finalized labeling: tv show hosts and journalists

Topic 4: synonym and same

- govern, secur, said, surveil, american, nation, agenc
- nsa[122], drone, faa[123], snowden, terror, terrorist, aircraft

Summary 1:___government surveillance whistleblower_____
Summary 2:___privacy fears about US government's use of drones to monitor for terrorists_____
Summary 3:___national security and terrorism_____
Initial labeling: Government surveillance and national security
Finalized labeling: Government surveillance and national security

Topic 5: same

- inform, consum, data, privaci, person, provid, requir

[118] Imus was the name of an American television show host

[119] Malveaux is the name of an American television news journalist

[120] Velshi is the name of a Canadian television journalist

[121] Cavuto is the name of an American television news anchor

[122] NSA stands for the National Security Agency

[123] FAA stands for the Federal Aviation Administration

- coppa[124], ccpa[125], credit, hipaa[126], settlement, ftc[127], breach

Summary 1:__consumer data legal protection_____
Summary 2:___data privacy laws_____
Summary 3:____national policies around consumer privacy_____
Initial labeling: Consumer privacy protection law and regulations
Finalized labeling: consumer data privacy protection

Topic 6: same
- person, violat, section, physic, imag, shall, record
- subdivis, visual, shall, impress, compensatori, physic, sound

Summary 1:__violation of personal image records_____
Summary 2:___privacy violations on records_____
Summary 3:__various forms of data___
Initial labeling: Physical privacy
Finalized labeling: privacy violation of records

Topic 7: synonym
- court, law, case, privaci, enforc, investig, search
- suprem, circuit, warrant, judg, court, fourth, subpoena

Summary 1:___legal decisions on privacy_____
Summary 2:____supreme court rules privacy and enforcing law through searchs_____
Summary 3:___law, legal____
Initial labeling: Court subpoena
Finalized labeling: law enforcement and privacy

Topic 8: synonym
- record, system, inform, feder, act, offic, privaci
- ssa[128], hud[129], sorn[130], usci[131], osd[132], dod[133], docket

Summary 1:__privacy of governmental department records_____
Summary 2:____US departments that have information on citizens_____
Summary 3:____immigration____
Initial labeling: Privacy about federal organizations
Finalized labeling: US governmental records

---

[124] COPPA stands for the Children's Online Privacy Protection Act

[125] CCPA stands for the California Consumer Privacy Act

[126] HIPPA stands for the Health Insurance Portability and Accountability Act

[127] FTC stands for the Federal Trade Commission

[128] SSA stands for the Social Security Administration

[129] HUD stands for the Department of Housing and Urban Development.

[130] SORN stands for the System of Records Notices. A system of records is a group of records under the control of any agency from which information is retrieved by the name of the individual or by some identifying number, symbol, or other identifier assigned to the individual. The Privacy Act requires each agency to publish notice of its systems of records in the Federal Register. This notice is generally referred to as a System of Records Notice or SORN.

[131] USCI stands for the United States Citizenship and Immigration Services

[132] OSD stands for the Office of the Secretary of Defense

[133] DOD stands for the Department of Defense

Topic 9: synonym

- privaci, consum, protect, data, bill, inform, senat
- fcc[134], broadband, markey, rep, subcommitte, isp[135], ntia[136]

Summary 1:___government bill on privacy of communication data_____
Summary 2:__laws about consumer privacy and data protection with ISPs_____
Summary 3: __the lawmakers and regulators_____
Initial labeling: Broadband and ISP regulation
Finalized labeling: regulations and regulators of privacy

Topic 10: same

- privaci, data, secur, inform, com, provid, manag
- fairwarn, onetrust, hitrust[137], csf[138], patent, ota[139], isaca[140]

Summary 1:___online data security managers_____
Summary 2:____privacy management and security servcies_____
Summary 3:_associations related to privacy management and data security____
Initial labeling: Privacy management services
Finalized labeling: privacy and data security services

Topic 11: same and hypernym

- data, privaci, protect, european, compani, law, shield
- schrem[141], shield[142], european, transatlant, apec[143], gdpr[144], europ

Summary 1:___international online data protections_____
Summary 2:___Europeaon data protection and privacy laws_____
Summary 3:__European data/privacy regulations_____
Initial labeling: International privacy law and framework
Finalized labeling: International privacy laws

---

[134] FCC stands for the Federal Communications Commission

[135] ISP stands for Internet service provider

[136] NTIA stands for The National Telecommunications and Information Administration

[137] fairwarn, onetrust, hitrust, these are online privacy management and data security services

[138] CSF stands for Cybersecurity Framework

[139] OTA stands for Online Trust Alliance

[140] ISACA stands for Information Systems Audit and Control Association

[141] Maximilian Schrems is an Austrian activist, lawyer, and author who became known for campaigns against Facebook for its privacy violations

[142] Shield here refers to the Privacy Shield, which is a framework designed by the U.S. Department of Commerce and the European Commission and Swiss Administration, to provide companies on both sides of the Atlantic with a mechanism to comply with data protection requirements when transferring personal data from the European Union and Switzerland to the United States in support of transatlantic commerce.

[143] APEC stands for Asia-Pacific Economic Cooperation

[144] GDPR stands for the General Data Protection Regulation, a regulation in EU law on data protection and privacy

【Group 2 Twitter】

Topic 1: same
- privaci, facebook, googl, set, social, polici, onlin
- icanstalku, buzz, tinyurl, nearbi, appspot, medal, proxypi

Summary 1: __social media privacy_____
Summary 2: _____social media and data privacy_____
Summary 3:___privacy invasion due to online info____
Initial labeling: Facebook, Google and geolocation
Finalized labeling: Privacy of social media

Topic 2: same and hypernym
- privaci, secur, data, real, free, show, absolut
- glue, trumptransit, presidentelecttrump, static, infosecjob, maga, giveaway

Summary 1:__trump era security_____
Summary 2:_____president trump, data privacy_____
Summary 3:___trump, job security____
Initial labeling: Data security and Trump
Finalized labeling: security and trump

Topic 3: same
- privaci, facebook, like, need, set, peopl, want
- blah, ittech, nut, jenni, shoe, vermont, medit

Summary 1:__location-based privacy on facebook_____
Summary 2:_____facebook and privacy_____
Summary 3:_____people's need/want toward privacy
Initial labeling: Facebook
Finalized labeling: Facebook

Topic 4: same and hypernym
- privaci, data, secur, facebook, protect, like, need
- granada, degrad, pension, nkdgvijlnn, classdojo[145], blackphon[146], glenn

Summary 1:___facebook data security_____
Summary 2:____phone privacy and data security; facebook_____
Summary 3:__tech companies, data security___
Initial labeling: Facebook, data security, classdojo and blackphone
Finalized labeling: tech companies and data security

Topic 5: same
- privaci, data, secur, facebook, right, protect, like
- datafund[147], cardi, deeponion[148], myhealthrecord[149], kavanaugh[150], ethereum[151], capitaltechnologiesresearch

Summary 1:___facebook privacy sensitive information_____
Summary 2:__cryptocurrency, privacy, and facebook; brett kavanaugh_____

---

[145] ClassDojo is an educational technology company

[146] The Blackphone is a smartphone built to ensure privacy

[147] Datafund is a blockchain based online data service project

[148] Deeponion is an anonymous cryptocurrency

[149] MyHealthRecord is an online summary of health information provided by the Australian government

[150] Kavanaugh is the name of an associate justice of the Supreme Court of the United States

[151] Ethereum is the community-run technology powering the cryptocurrency, ether (ETH) and thousands of decentralized applications.

Summary 3:_blockchain/cryptocurrency-related__
Initial labeling: Facebook, online data services and cryptocurrencies
Finalized labeling: Facebook, Cryptocurrency

Topic 6: same
- privaci, facebook, googl, polici, need, set, user
- demandprogress, plenti, petraeus[152], nich, randi, buyer, hysteria

Summary 1:__facebook privacy policy concerns_____
Summary 2:____privacy policies with google and facebook_____
Summary 3:_demanding privacy from consumers/users___
Initial labeling: Facebook, google, and government surveillance
Finalized labeling: privacy policies of facebook and google

Topic 7: NA
- privaci, protect, data, secur, free, american, real
- encyrypt, cisa[153], stealthcoin[154], idltweet, barbi, newpanda[155], stitm

Summary 1:___protection of financial communications_____
Summary 2:_____Cybersecruity, cryptocurrency, and privacy_____
Summary 3:____blockchain, security, _____
Initial labeling: Data security, cryptocurrency, data service
Finalized labeling: Miscellaneous (no convergence)

Topic 8: same
- privaci, facebook, set, googl, polici, onlin, protect
- geotag, error, kingston[156], carrier, timelin, datatravel[157], verdict

Summary 1:__online location-based privacy concerns_____
Summary 2:___flash memory and privacy at google and facebook_____
Summary 3:_____memory/data-related products_____
Initial labeling: Facebook, Google, and data service
Finalized labeling: flash memory

Topic 9: synonym
- privaci, googl, facebook, like, protect, data, secur
- cispaalert, endcispa, prism[158], obamacar, typewrit, bush, cispa[159]

Summary 1:__government-based online privacy issues_____
Summary 2:_____cyberintellengce and privacy on google and facebook_____
Summary 3:___national policies related to online security___

---

[152] Petraeus is a retired United States Army general and public official. The investigation of Petraeus raised concerns about email privacy

[153] CISA stands for the Cybersecurity and Infrastructure Security Agency

[154] Stealthcoin is a blockchain based digital currency.

[155] Newpanda provides social, email & SMS marketing services.

[156] Kingston Technology Corporation is an American multinational computer technology corporation that develops, manufactures, sells and supports flash memory products and other computer-related memory products.

[157] Datatravel could refer to Kingston's DataTravelers which are portable flash memory drives. Or, it could also refer to an overseas data service

[158] PRISM is a code name for a program under which the United States National Security Agency (NSA) collects internet communications from various U.S. internet companies

[159] Cispa stands for the Cyber Intelligence Sharing and Protection Act

Initial labeling: Facebook, google, and government surveillance
Finalized labeling: cyberintelligence

Topic 10: hypernym and synonym
- privaci, data, like, peopl, right, secur, need
- xboxpaxaus, faceapp[160], leaderboard, ccpa[161], nica, doordash[162], securypto

Summary 1:__facial recognition privacy policies_____
Summary 2:_____privacy and faceapp, CCPA_____
Summary 3:_online apps that collect people's data___
Initial labeling: Privacy law, cryptocurrency, apps
Finalized labeling: apps, privacy regulation

Topic 11: same and hypernym
- privaci, data, real, free, secur, show, absolut
- hat, foil, locker, evernot[163], pokmon[164], pokemon, fertil

Summary 1:___data security of popular apps_____
Summary 2:__privacy in evernote; pokemon_____
Summary 3:_free apps that might use your data___
Initial labeling: Evernote and Pokemon go
Finalized labeling: data security of apps

Topic 12: same and synonym
- privaci, facebook, googl, set, polici, secur, onlin
- doodl, czar, freelanc, lennon, creatur, invadin, buzz

Summary 1:___Russian involvement American online data_____
Summary 2:_____privacy in facebook and google, Russia_____
Summary 3: ___something related to what people do online___
Initial labeling: Facebook and Google
Finalized labeling: Facebook, Google and Russia

Topic 13: same
- privaci, data, secur, protect, facebook, onlin, internet
- bonus, surfeasyinc[165], null, truecrypt[166], pear, papul, ncpol

Summary 1:___encryption online and social media data_____
Summary 2:___data privacy using VPNs and encryption_____
Summary 3: _____ encrypted service_____
Initial labeling: Data security, Facebook and encryption service
Finalized labeling: encryption service

Topic 14: synonym
- privaci, facebook, polici, set, onlin, like, protect

---

[160] Faceapp is a mobile app for AI photo editing

[161] CCPA stands for California Consumer Privacy Act

[162] DoorDash is an online food ordering and food delivery platform

[163] Evernote is an app designed for note taking, etc.

[164] Pokémon is a series of video games

[165] SurfEasy provides encrypted VPN services

[166] TrueCrypt is a discontinued source-available freeware utility used for on-the-fly encryption (OTFE)

- mellon, carnegi, onstar[167], hampton, geofenc[168], sacr, roethlisberg

Summary 1: __policy security of location data_____
Summary 2:___privacy at Carnegie Mellon; geofencing, facebook_____
Summary 3:___geological areas____
Initial labeling: Miscellaneous
Finalized labeling: location data

Topic 15: same
- privaci, data, secur, like, peopl, facebook, protect
- nbsp, field, entri, npleas, soul, fill, pacif

Summary 1:___facebook personal data security_____
Summary 2:_____data security and privacy; facebook_____
Summary 3:_data security/protection of people's information_____
Initial labeling: Data security and Facebook
Finalized labeling: data security, facebook

[167] OnStar Corporation is a subsidiary of General Motors that provides services including subscription-based communications, in-vehicle security, etc

[168] A geofence is a virtual perimeter for a real-world geographic area

APPENDIX I: CODERS' LABELING COMPARISON

Corpus	Topics	synonym	hypernym	same	dissimilar
CN K13	topic 1			✓	
	topic 2	✓	✓		
	topic 3	✓			
	topic 4			✓	
	topic 5			✓	
	topic 6			✓	
	topic 7			✓	
	topic 8				✓
	topic 9			✓	
	topic 10	✓			
	topic 11	✓		✓	
	topic 12	✓			
	topic 13	✓			
Weibo K13	topic 1	✓			
	topic 2	✓			
	topic 3			✓	
	topic 4	✓		✓	
	topic 5				✓
	topic 6				✓
	topic 7		✓		
	topic 8			✓	
	topic 9				✓

	topic 10			✓	
	topic 11	✓			
	topic 12	✓			
	topic 13			✓	
EN K11	topic 1			✓	
	topic 2			✓	
	topic 3			✓	
	topic 4	✓		✓	
	topic 5			✓	
	topic 6			✓	
	topic 7	✓			
	topic 8	✓			
	topic 9	✓			
	topic 10			✓	
	topic 11		✓	✓	
Twitter K15	topic 1			✓	
	topic 2		✓	✓	
	topic 3			✓	
	topic 4		✓	✓	
	topic 5			✓	
	topic 6			✓	
	topic 7				✓
	topic 8			✓	

	topic 9	✓		
	topic 10	✓	✓	
	topic 11		✓	✓
	topic 12	✓		✓
	topic 13			✓
	topic 14	✓		
	topic 15			✓

APPENDIX J: ESTIMATION OF THE EFFECT OF LANGUAGE AND GENRE

```
Call: estimateEffect(formula = ~Lan, stmobj = model_K11_lan, metadata = out$meta,
   uncertainty = "None")
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Topic 1:
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.114337   0.001198   95.41   <2e-16 ***
LanEN       -0.082834   0.001631  -50.78   <2e-16 ***
---
Topic 2:
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.082648   0.001213  68.147  < 2e-16 ***
LanEN       -0.004358   0.001676  -2.601  0.00931 **
---
Topic 3:
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.087996   0.001323  66.514  < 2e-16 ***
LanEN       0.012635   0.001947   6.489 8.77e-11 ***
---
Topic 4:
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.080336   0.001281   62.73   <2e-16 ***
LanEN       0.024017   0.001788   13.44   <2e-16 ***
---
Topic 5:
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.109962   0.001078  102.00   <2e-16 ***
LanEN       -0.077120   0.001494  -51.62   <2e-16 ***
---
Topic 6:
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.041043   0.000947   43.34   <2e-16 ***
LanEN       0.017928   0.001351   13.27   <2e-16 ***
---
Topic 7:
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.118780   0.001182  100.47   <2e-16 ***
LanEN       -0.052647   0.001770  -29.74   <2e-16 ***
---
Topic 8:
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.059128   0.002205   26.81   <2e-16 ***
LanEN       0.312770   0.003110  100.57   <2e-16 ***
---
Topic 9:
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.127944   0.001491   85.80   <2e-16 ***
```

```
LanEN       -0.052432   0.002290  -22.89   <2e-16 ***
---
Topic 10:
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.112284   0.001139   98.61   <2e-16 ***
LanEN       -0.080392   0.001497  -53.71   <2e-16 ***
---
Topic 11:
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0657299  0.0007886   83.35   <2e-16 ***
LanEN       -0.0171693  0.0011074  -15.51   <2e-16 ***
---

Call: estimateEffect(formula = ~Genre, stmobj = model_K11_genre, metadata = out$meta,
    uncertainty = "None")
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Topic 1:
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0419877  0.0006239   67.30   <2e-16 ***
GenreSocial -0.0121813  0.0009504  -12.82   <2e-16 ***
---
Topic 2:
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.099396   0.001317   75.47   <2e-16 ***
GenreSocial -0.045211   0.001941  -23.29   <2e-16 ***
---
Topic 3:
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.105559   0.001361   77.57   <2e-16 ***
GenreSocial -0.052127   0.002025  -25.74   <2e-16 ***
---
Topic 4:
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.117306   0.001319   88.96   <2e-16 ***
GenreSocial -0.037923   0.001979  -19.16   <2e-16 ***
---
Topic 5:
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.090577   0.001173   77.25   <2e-16 ***
GenreSocial -0.033954   0.001667  -20.37   <2e-16 ***
---
Topic 6:
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.116267   0.001210   96.08   <2e-16 ***
GenreSocial -0.053477   0.001837  -29.11   <2e-16 ***
---
Topic 7:
Coefficients:
           Estimate Std. Error t value Pr(>|t|)
```

(Intercept) 0.063092 0.001146 55.043 < 2e-16 ***
GenreSocial 0.005564 0.001707 3.259 0.00112 **

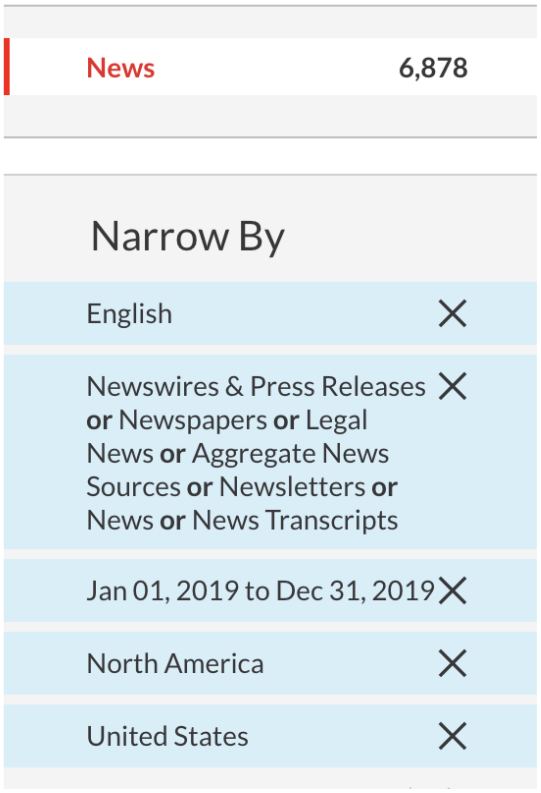
Topic 8:
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.118970 0.002318 51.33 <2e-16 ***
GenreSocial 0.287877 0.003401 84.65 <2e-16 ***

Topic 9:
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.073456 0.001220 60.20 <2e-16 ***
GenreSocial -0.029876 0.001821 -16.41 <2e-16 ***

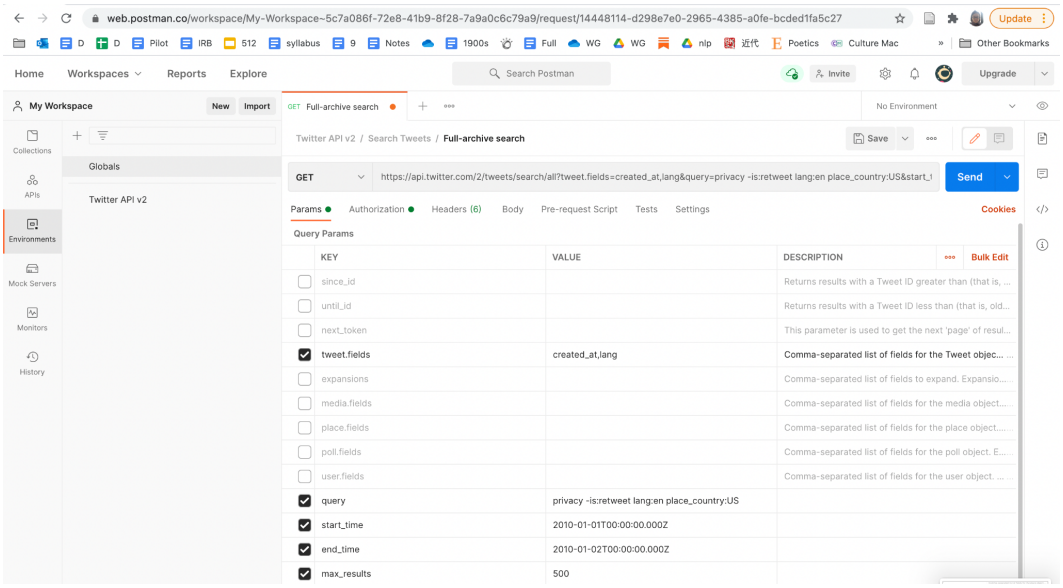
Topic 10:
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.104038 0.001400 74.290 <2e-16 ***
GenreSocial -0.018998 0.002118 -8.968 <2e-16 ***

Topic 11:
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.069224 0.000841 82.310 < 2e-16 ***
GenreSocial -0.009516 0.001210 -7.862 3.89e-15 ***

APPENDIX K: SCREENSHOT OF NEXIS UNI SEARCH INTERFACE



APPENDIX L: A SCREENSHOT OF TWEETS RETRIEVAL CRITERIA



APPENDIX M: GOOGLE CLOUD TRANSLATE

```
def explicit():
    from google.cloud import storage

    # Explicitly use service account credentials by specifying the private key
    # file.
    storage_client = storage.Client.from_service_account_json(
        'airy-ceremony-307203-1dbd14706eab.json')

    # Make an authenticated API request
    buckets = list(storage_client.list_buckets())
    print(buckets)
```

```
import os
from google.cloud import translate

os.environ['GOOGLE_APPLICATION_CREDENTIALS'] = '/Users/aliceyuanye_ma/Desktop/airy-ceremony-307203-1dbd14706eab.json'
```

```
translate_client = translate.Client()
```

```
import six
from google.cloud import translate_v2 as translate
translate_client = translate.Client()
```

```
text= '天气'
target = 'en'
output = translate_client.translate(text,target_language = target)
```

```
print(output)
```

```
{'translatedText': 'the weather', 'detectedSourceLanguage': 'zh-CN', 'input': '天气'}
```

```
import os
os.chdir('/Users/aliceyuanye_ma/Desktop/SMALL SAMPLE/')
cwd = os.getcwd()
import pandas as pd
df_1= pd.read_csv('Weibo_2019_845_clean.csv',index_col=0,encoding='utf-8')
```

```
translated = []
for item in list_one:
    translated_item = translate_client.translate(item,target_language = target)
    translated.append(translated_item)
```

```
translated_list = []
i = 0
while i < 845:
    translated_list.append(translated[i]['translatedText'])
    i=i+1
```

```
import pandas as pd
d1 = {'Text':translated_list}
df1 = pd.DataFrame(d1)
df1['Year']='2019'
df1['Genre']='Social'
df1['Lan']='CN'
```

```
df1.to_csv('Weibo_2019_translated_845.csv')
```

```
df_1.head()
```

	Unnamed: 0.1	Text	Year	Genre	Lan
1585	1585	每日 路报 丨 星期三 热点 告 台湾同胞 书 发表 周年 两岸 政治 交往 不断 创造 ...	2019	Social	CN
2021	2021	用户 诉 腾讯 浏览器 违法 收集 个人隐私 法院 立即 停止 互金 新闻网 贷 天眼 网...	2019	Social	CN
1684	1684	一分钟 告诉 保护 个人 信息安全 个人隐私 泄露 有着 严重 安全隐患 浏览器 社交 平...	2019	Social	CN
1300	1300	看看 微 博大 数据 有点 可怕 啊啊啊 不能 人点 隐私 一位 跨界 人士 隐私 问题 ...	2019	Social	CN
1382	1382	我刚 更新 新浪 微盘 客户端 方便 安全 贴心 照片 备份 新增 个性化 设置 密码锁 ...	2019	Social	CN

```
list_one  = df_1["Text"].tolist()
```

REFERENCES

Abokhodair, N., Abbar, S., Vieweg, S., & Mejova, Y. (2016). Privacy and Twitter in Qatar: Traditional Values in the Digital World. *Proceedings of the 8th ACM Conference on Web Science*, 66–77. https://doi.org/10.1145/2908131.2908146

Acquisti, A., Brandimarte, L., & Loewenstein, G. (2015). Privacy and human behavior in *the* age of information. *Science*, *347*(6221), 509–514. https://doi.org/10.1126/science.aaa1465

Altman, I. (1981). The environment and social behavior: Privacy, personal space, territory, crowding (1st Irvington ed. (1981)). Irving Publishers.

Ames, R. T. (2020). Preparing a New SOURCEBOOK IN CLASSICAL CONFUCIAN PHILOSOPHY. *Journal of Chinese Humanities*. http://www.journalofchinesehumanities.com/uncategorized/preparing-a-new-sourcebook-in-classical-confucian-philosophy%ef%bb%bf/

Atefeh, F., & Khreich, W. (2015). A Survey of Techniques for Event Detection in Twitter. *Computational Intelligence*, *31*(1), 132–164. https://doi.org/10.1111/coin.12017

Baldwin Lind, P. (2015). Looking for privacy in Shakespeare: Woman's place and space in a selection of plays and early modern texts [D_ph, University of Birmingham]. https://etheses.bham.ac.uk/id/eprint/5848/

Bannerman, S. (2018). Relational privacy and the networked governance of the self. Information, Communication & Society, 0(0), 1–16. https://doi.org/10.1080/1369118X.2018.1478982

Basov, N., Lee, J.-S., & Antoniuk, A. (2017). Social Networks and Construction of Culture: A Socio-Semantic Analysis of Art Groups. In H. Cherifi, S. Gaito, W. Quattrociocchi, & A. Sala (Eds.), *Complex Networks & Their Applications V* (pp. 785–796). Springer International Publishing.

Bastian M, Heymann S, Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. ICWSM; 8:361–362.

Berg, C. (2018). The Origins of Modern Privacy. In C. Berg (Ed.), *The Classical Liberal Case for Privacy in a World of Surveillance and Technological Change* (pp. 75–96). Springer International Publishing. https://doi.org/10.1007/978-3-319-96583-3_5

Bergen, M., & Kulwin, N. (2016, July 12). *Pokémon go's creators say they didn't mean to spy on google accounts*. Vox. Retrieved December 11, 2021, from https://www.vox.com/2016/7/11/12154354/pokemon-go-niantic-google-permissions.

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc."

Blei, D. M. (2012). Probabilistic Topic Models. Communications of the ACM 55(4): 77–84.

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies. *J. ACM*, *57*(2), 7:1-7:30. https://doi.org/10.1145/1667053.1667056

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. Annals of Applied Statistics, 1(1), 17–35. https://doi.org/10.1214/07-AOAS114

Blei, D. M., Ng, A.Y. and Jordan, M.I. (2003), "Latent Dirichlet allocation". Journal of Machine Learning Research, Vol. 3 Nos 4-5, pp. 993-1022.

Bohr, J., & Dunlap, R. E. (2018). Key Topics in environmental sociology, 1990–2014: Results from a computational text analysis. Environmental Sociology, 4(2), 181–195. https://doi.org/10.1080/23251042.2017.1393863

Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of Topic Models. Foundations and Trends® in Information Retrieval, 11(2–3), 143–296. https://doi.org/10.1561/1500000030

Borghi, A. M., Barca, L., Binkofski, F., Castelfranchi, C., Pezzulo, G., & Tummolini, L. (2019). Words as social tools: Language, sociality and inner grounding in abstract concepts. Physics of Life Reviews, 29, 120–153. https://doi.org/10.1016/j.plrev.2018.12.001

Bowler, P. J. (1975). The Changing Meaning of "Evolution." Journal of the History of Ideas, 36(1), 95–114. JSTOR. https://doi.org/10.2307/2709013

Brown, C. L. (2008). China's second-generation national identity card: Merging culture, industry and technology, In Bennett, C. J., & In Lyon, D. (2008). *Playing the identity card: Surveillance, security and identification in global perspective*. London ; New York : Routledge, 2008.

Bundgaard, P. F. (2019). The structure of our concepts: A critical assessment of Conceptual Metaphor Theory as a theory of concepts. Cognitive Semiotics; Berlin, 12(1). http://dx.doi.org/10.1515/cogsem-2019-2010

Bureau of Justice Assistance. (n.d.). *USA PATRIOT Act*. Retrieved November 16, 2021, from https://bja.ojp.gov/program/it/privacy-civil-liberties/authorities/statutes/1281.

Busch, B. (2017). Expanding the Notion of the Linguistic Repertoire: On the Concept of Spracherleben—The Lived Experience of Language. Applied Linguistics, 38(3), 340–358. https://doi.org/10.1093/applin/amv030

Capurro, R. (2005). Privacy. An Intercultural Perspective. Ethics and Information Technology, 7(1), 37–47. https://doi.org/10.1007/s10676-005-4407-4

Capurro, R. (2008). Intercultural information ethics: Foundations and applications. Journal of Information, Communication & Ethics in Society; Bingley, 6(2), 116–126. http://dx.doi.org/10.1108/14779960810888347

Carley, K., & Palmquist, M. (1992). Extracting, representing, and analyzing mental models. Social Forces, 70(3), 601-636.

Ceron, A. (2015). Internet, News, and Political Trust: The Difference Between Social Media and Online Media Outlets. Journal of Computer-Mediated Communication, 20(5), 487–503. https://doi.org/10.1111/jcc4.12129

Chen, D. (2017). "Supervision by Public Opinion" or by Government Officials? Media Criticism and Central-Local Government Relations in China. *Modern China*, *43*(6), 620–645. https://doi.org/10.1177/0097700417706704

Chen, C. C., Chen, X.-P., & Huang, S. (2013). Chinese Guanxi: An Integrative Review and New Directions for Future Research. *Management and Organization Review*, *9*(1), 167–207. https://doi.org/10.1111/more.12010

Cheney-Lippold, J. (2011). A New Algorithmic Identity: Soft Biopolitics and the Modulation of Control. Theory, Culture & Society, 28(6), 164–181. https://doi.org/10.1177/0263276411424420

Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). BTM: Topic Modeling over Short Texts. IEEE Transactions on Knowledge and Data Engineering, 26(12), 2928–2941.

https://doi.org/10.1109/TKDE.2014.2313872

Choi, C., & Lecy, J. D. (2012). A Semantic Network Analysis of Changes in North Korea's Economic Policy. Governance, 25(4), 589–616. https://doi.org/10.1111/j.1468-0491.2012.01597.x

Cohen, J. E. (2012). What Privacy Is For (SSRN Scholarly Paper ID 2175406). Social Science Research Network. https://papers.ssrn.com/abstract=2175406

Collins, A. M., & Quillian, M. R. (1972). Experiments on semantic memory and language comprehension. In L. W. Gregg (Ed.), Cognition in learning and memory (pp. 117–138). New York, NY: Wiley.

Condren, C. (2009). Public, Private and the Idea of the 'Public Sphere' in Early–modern England. *Intellectual History Review*, *19*(1), 15–28. https://doi.org/10.1080/17496970902722866

Connor, B. T., & Doan, L. (2021). Government and corporate surveillance: Moral discourse on privacy in the civil sphere. *Information, Communication & Society*, *24*(1), 52–68. https://doi.org/10.1080/1369118X.2019.1629693

Corballis, M. C. (2002). *From Hand to Mouth: The Origins of Language.* Princeton NJ: Princeton University Press.

Danowski, J. A. (2013). WORDij version 3.0: Semantic network analysis software. Chicago: University of Illinois at Chicago.

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. Poetics, 41(6), 570–606. https://doi.org/10.1016/j.poetic.2013.08.004

Dixon, T. L., Weeks, K. R., & Smith, M. A. (2019). *Media Constructions of Culture, Race, and Ethnicity*. Oxford Research Encyclopedia of Communication. https://doi.org/10.1093/acrefore/9780190228613.013.502

Doerfel, M. L. (1998). What constitutes semantic network analysis? A comparison of research and methodologies. Connections, 21, 16–26.

Doerfel, M. L., & Connaughton, S. L. (2009). Semantic networks and competition: Election year winners and losers in U.S. televised presidential debates, 1960–2004. Journal of the American Society for Information Science and Technology, 60(1), 201–218. https://doi.org/10.1002/asi.20950

Dove, G. (2014). Thinking in Words: Language as an Embodied Medium of Thought. Topics in Cognitive Science, 6(3), 371–389. https://doi.org/10.1111/tops.12102

Drieger, P. (2013). Semantic Network Analysis as a Method for Visual Text Analytics. Procedia - Social and Behavioral Sciences, 79, 4–17. https://doi.org/10.1016/j.sbspro.2013.05.053

Du, H., Nguyen, L., Yang, Z., Abu-Gellban, H., Zhou, X., Xing, W., Cao, G., & Jin, F. (2019). Twitter vs News: Concern Analysis of the 2018 California Wildfire Event. 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), 2, 207–212. https://doi.org/10.1109/COMPSAC.2019.10208

Ellison, N. B., & Boyd, D. M. (2013). Sociality Through Social Network Sites. The Oxford Handbook of Internet Studies. https://doi.org/10.1093/oxfordhb/9780199589074.013.0008

Erlich, Y., Williams, J. B., Glazer, D., Yocum, K., Farahany, N., Olson, M., Narayanan, A., Stein, L. D., Witkowski, J. A., & Kain, R. C. (2014). Redefining Genomic Privacy: Trust and Empowerment. PLOS Biology, 12(11), e1001983. https://doi.org/10.1371/journal.pbio.1001983

Ess, C. (2005). "Lost in Translation"?: Intercultural Dialogues on Privacy and Information Ethics (Introduction to Special Issue on Privacy and Data Privacy Protection in Asia). Ethics and Information Technology, 7(1), 1–6. https://doi.org/10.1007/s10676-005-0454-0

Ess, C. M. (2019). Intercultural Privacy: A Nordic Perspective. In H. Behrendt, W. Loh, T. Matzner, & C. Misselhorn (Eds.), Privatsphäre 4.0: Eine Neuverortung des Privaten im Zeitalter der Digitalisierung (pp. 73–88). J.B. Metzler. https://doi.org/10.1007/978-3-476-04860-8_5

Evans, V. (2006). Lexical concepts, cognitive models and meaning-construction. 17(4), 491–534. https://doi.org/10.1515/COG.2006.016

Everett, C. (2013). Linguistic Relativity: Evidence Across Languages and Cognitive Domains. In *Linguistic Relativity*. De Gruyter Mouton. https://doi.org/10.1515/9783110308143

Farrall, K. N. (2008). Global Privacy in Flux: Illuminating Privacy Across Cultures in China and the U.S. International Journal of Communication, 2(0), 38.

Feng, H. (2019a). *Performance of Latent Dirichlet Allocation with Different Topic and Document Structures* [Ph.D.]. https://search.proquest.com/docview/2300629819/abstract/96FB946EE6434301PQ/1

Feng, Y. (2019b). The future of China's personal data protection law: Challenges and prospects. Asia Pacific Law Review, 27(1), 62–82. https://doi.org/10.1080/10192557.2019.1646015

Ferenstein, G. (2015, November 25). The Birth And Death Of Privacy: 3,000 Years of History Told Through 46 Images. *Medium*. https://medium.com/the-ferenstein-wire/the-birth-and-death-of-privacy-3-000-years-of-history-in-50-images-614c26059e

Fu, X., Li, J., Yang, K., Cui, L., & Yang, L. (2016). Dynamic Online HDP model for discovering evolutionary topics from Chinese social texts. Neurocomputing, 171, 412–424. https://doi.org/10.1016/j.neucom.2015.06.047

Gao, Z., & O'Sullivan-Gavin, S. (2015). The development of consumer privacy protection policy in China: A historical review. *Journal of Historical Research in Marketing; Bingley*, *7*(2), 232–255. http://dx.doi.org/10.1108/JHRM-08-2014-0022

Garcia, D., Goel, M., Agrawal, A. K., & Kumaraguru, P. (2018). Collective aspects of privacy in the Twitter social network. EPJ Data Science, 7(1), 1–13. https://doi.org/10.1140/epjds/s13688-018-0130-3

Gieck, R., Kinnunen, H.-M., Li, Y., Moghaddam, M., Pradel, F., Gloor, P. A., Paasivaara, M., & Zylka, M. P. (2016). Cultural Differences in the Understanding of History on Wikipedia. In M. P. Zylka, H. Fuehres, A. Fronzetti Colladon, & P. A. Gloor (Eds.), Designing Networks for Innovation and Improvisation (pp. 3–12). Springer International Publishing. https://doi.org/10.1007/978-3-319-42697-6_1

Glancy, D. (1979). The Invention of the Right to Privacy. Faculty Publications. http://digitalcommons.law.scu.edu/facpubs/317

González, F., Yu, Y., Figueroa, A., López, C., & Aragon, C. (2019). Global Reactions to the Cambridge Analytica Scandal: A Cross-Language Social Media Study. *Companion*

Proceedings of The 2019 World Wide Web Conference, 799–806. https://doi.org/10.1145/3308560.3316456

González Fuster, G. (2014). *The Emergence of Personal Data Protection as a Fundamental Right of the EU* (Vol. 16). Springer International Publishing. https://doi.org/10.1007/978-3-319-05023-2

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Political Analysis, 21(3), 267–297. https://doi.org/10.1093/pan/mps028

Gu, J. (1988), "Ruling on a case that involved illegal disclosure of others' privacy", People's Justice, No. 8, pp. 27-28.

Gumperz, J. J. (1964). Linguistic and Social Interaction in Two Communities1. American Anthropologist, 66(6_PART2), 137–153. https://doi.org/10.1525/aa.1964.66.suppl_3.02a00100

Günther, E., & Domahidi, E. (2017). What communication scholars write about: An analysis of 80 years of research in high-impact journals. International Journal of Communication, 11, 3051–3071.

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the History of Ideas Using Topic Models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 363–371. http://dl.acm.org/citation.cfm?id=1613715.1613763

Han, S., & Munir, A. B. (2018). Information Security Technology - Personal Information Security Specification: China's Version of the GDPR Reports: Practitioner's Corner. European Data Protection Law Review (EDPL), 4(4), 535–541.

Haber, B. (2019). The digital ephemeral turn: Queer theory, privacy, and the temporality of risk. *Media, Culture & Society*, *41*(8), 1069–1087. https://doi.org/10.1177/0163443719831600

Hecking, T., & Leydesdorff, L. (2018). Topic Modeling of Empirical Text Corpora: Validity, Reliability, and Reproducibility in Comparison to Semantic Maps. ArXiv:1806.01045 [Cs]. http://arxiv.org/abs/1806.01045

Hill, F., Reichart, R., & Korhonen, A. (2014). Multi-Modal Models for Concrete and Abstract Concept Meaning. Transactions of the Association for Computational Linguistics. https://doi.org/10.1162/tacl_a_00183

Hirschauer, S. (2007). Puttings things into words. Ethnographic description and the silence of the social. *Human Studies*, *29*(4), 413. https://doi.org/10.1007/s10746-007-9041-1

Holvast, J. (2008). History of Privacy. *The Future of Identity in the Information Society*, 13–42. https://doi.org/10.1007/978-3-642-03315-5_2

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in Twitter. Proceedings of the First Workshop on Social Media Analytics, 80–88. https://doi.org/10.1145/1964858.1964870

Hongladarom, S. (2016). A Buddhist Theory of Privacy. Springer Singapore. https://www.springer.com/us/book/9789811003165

Hwang, K. (1987). Face and Favor: The Chinese Power Game. *American Journal of Sociology*, *92*(4), 944–974. https://doi.org/10.1086/228588

Jackson, J. C., Watts, J., Henry, T. R., List, J.-M., Forkel, R., Mucha, P. J., Greenhill, S. J.,

Gray, R. D., & Lindquist, K. A. (2019). Emotion semantics show both cultural variation and universal structure. Science, 366(6472), 1517–1522. https://doi.org/10.1126/science.aaw8160

Jacobi, C., Atteveldt, W. van, & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modeling. Digital Journalism, 4(1), 89–106. https://doi.org/10.1080/21670811.2015.1093271

Jain, A. K., Nandakumar, K., & Ross, A. (2016). 50 years of biometric research: Accomplishments, challenges, and opportunities. Pattern Recognition Letters, 79, 80–105. https://doi.org/10.1016/j.patrec.2015.12.013

Jiang, K., Anderton, B. N., Ronald, P. C., & Barnett, G. A. (2018). Semantic Network Analysis Reveals Opposing Online Representations of the Search Term "GMO." Global Challenges, 2(1), 1700082. https://doi.org/10.1002/gch2.201700082

Jiang, K., Benefield, G. A., Yang, J., & Barnett, G. A. (2017, January 4). Mapping Articles on China in Wikipedia: An Inter-Language Semantic Network Analysis. https://doi.org/10.24251/HICSS.2017.270

Jiang, M. (2016). Managing the micro-self: The governmentality of real name registration policy in Chinese microblogosphere. Information, Communication & Society, 19(2), 203–220. https://doi.org/10.1080/1369118X.2015.1060723

Jiao, L. (2016). *Chinese Idioms*. Routledge Handbooks Online. https://doi.org/10.4324/9781315675541.ch5

Karakayali, N., Kostem, B., & Galip, I. (2017). Recommendation Systems as Technologies of the Self: Algorithmic Control and the Formation of Music Taste: Theory, Culture & Society. https://doi.org/10.1177/0263276417722391

Kim, L., & Kim, N. (2015). Connecting opinion, belief and value: Semantic network analysis of a UK public survey on embryonic stem cell research. Journal of Science Communication, 14(1), A01. https://doi.org/10.22323/2.14010201

King, G., Pan, J., & Roberts, M. (2012). *How Censorship in China Allows Government Criticism But Silences Collective Expression* (SSRN Scholarly Paper ID 2104894). Social Science Research Network. https://papers.ssrn.com/abstract=2104894

Korolova, A. (2011). Privacy Violations Using Microtargeted Ads: A Case Study. *Journal of Privacy and Confidentiality*, *3*(1), Article 1. https://doi.org/10.29012/jpc.v3i1.594

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. American Sociological Review, 84(5), 905–949. https://doi.org/10.1177/0003122419877135

Kuang, X. (2018). Central State vs. Local Levels of Government: Understanding News Media Censorship in China. *Chinese Political Science Review*, *3*(2), 154–171. https://doi.org/10.1007/s41111-018-0091-5

Kwon, K. H., Bang, C. C., Egnoto, M., & Raghav Rao, H. (2016). Social media rumors as improvised public opinion: Semantic network analyses of twitter discourses during Korean saber rattling 2013. Asian Journal of Communication, 26(3), 201–222. https://doi.org/10.1080/01292986.2015.1130157

Kwon, K., Barnett, G. A., & Chen, H. (2009). Assessing Cultural Differences in Translations: A Semantic Network Analysis of the Universal Declaration of Human Rights. Journal of International and Intercultural Communication, 2(2), 107–138. https://doi.org/10.1080/17513050902759488

Kumaraguru, P., & Cranor, L. (2006). Privacy in India: Attitudes and Awareness. In G. Danezis & D. Martin (Eds.), Privacy Enhancing Technologies (pp. 243–258). Springer. https://doi.org/10.1007/11767831_16

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. Rivista Di Linguistica, 20(1), 1–31.

Levy, K. E. C., & Franklin, M. (2014). Driving Regulation: Using Topic Models to Examine Political Contention in the U.S. Trucking Industry. Social Science Computer Review, 32(2), 182–194. https://doi.org/10.1177/0894439313506847

Leydesdorff, L., & Nerghes, A. (2015). Co-word Maps and Topic Modeling: A Comparison Using Small and Medium-Sized Corpora (n < 1000). ArXiv:1511.03020 [Cs]. http://arxiv.org/abs/1511.03020

Lin, F. (2020, September 22). *Renmin lailun: zhengfu guanwang xielou geren xinxi, shuohaode yinsi baohu ne?* [People's forum: government official websites leaking personal information, what about privacy protection?]. Retrieved November 18, 2021, from http://opinion.people.com.cn/n1/2020/0922/c431649-31871150.html.

Lindstedt, N. C. (2019). Structural Topic Modeling For Social Scientists: A Brief Case Study with Social Movement Studies Literature, 2005–2017. Social Currents, 6(4), 307–318. https://doi.org/10.1177/2329496519846505

Loi, M., & Christen, M. (2019). Two Concepts of Group Privacy. Philosophy & Technology. https://doi.org/10.1007/s13347-019-00351-0

Louwerse, M. M. (2018). Knowing the Meaning of a Word by the Linguistic and Perceptual Company It Keeps. Topics in Cognitive Science, 10(3), 573–589. https://doi.org/10.1111/tops.12349

Luo, D., & Wang, Y. (2021, November). *China - Data Protection Overview*. DataGuidance. Retrieved January 10, 2022, from https://www.dataguidance.com/notes/china-data-protection-overview

Lü, Y.-H. (2005). Privacy and Data Privacy Issues in Contemporary China. Ethics and Information Technology, 7(1), 7–15. https://doi.org/10.1007/s10676-005-0456-y

Lunshof, J. E., Chadwick, R., Vorhaus, D. B., & Church, G. M. (2008). From genetic privacy to open consent. Nature Reviews Genetics, 9(5), 406–411. https://doi.org/10.1038/nrg2360

Lupyan, G., & Lewis, M. (2019). From words-as-mappings to words-as-cues: The role of language in semantic knowledge. Language, Cognition and Neuroscience, 34(10), 1319–1337. https://doi.org/10.1080/23273798.2017.1404114

Ma, Y. (2019a). Relational privacy: Where the East and the West could meet. Proceedings of the Association for Information Science and Technology, 56(1), 196–205. https://doi.org/10.1002/pra2.65

Ma, Y. (2019b). Unmapped Privacy Expectations in China: Discussions Based on the Proposed Social Credit System. In N. G. Taylor, C. Christian-Lamb, M. H. Martin, & B. Nardi (Eds.), *Information in Contemporary Society* (pp. 799–805). Springer International Publishing. https://doi.org/10.1007/978-3-030-15742-5_75

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Communication Methods and Measures*, *12*(2–3), 93–118. https://doi.org/10.1080/19312458.2018.1430754

Margolis, E., & Laurence, S. (2019). Concepts. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Summer 2019). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2019/entries/concepts/

Marwick, A., Fontaine, C., & boyd, danah. (2017). “Nobody Sees It, Nobody Gets Mad”: Social Media, Privacy, and Personal Responsibility Among Low-SES Youth. Social Media + Society, 3(2), 2056305117710455. https://doi.org/10.1177/2056305117710455

Masur, P. K., Epstein, D., Quinn, K., Wilhelm, C., Baruh, L., & Lutz, C. (2021). *A Comparative Privacy Research Framework*. SocArXiv. https://doi.org/10.31235/osf.io/fjqhs

McArthur, T. M., Lam-McArthur, J. L.-M., & Fontaine, L. F. (2018). Type Token Ratio. In T. McArthur, J. Lam-McArthur, & L. Fontaine (Eds.), The Oxford Companion to the English Language. Oxford University Press. https://www.oxfordreference.com/view/10.1093/acref/9780199661282.001.0001/acref-9780199661282-e-1462

McDougall, B. S. (2005). Discourse on Privacy by Women Writers in Late Twentieth-Century China. *China Information*, *19*(1), 97–119. https://doi.org/10.1177/0920203X05051022

McDougall, B. S. (2004). Privacy in Modern China. History Compass, 2(1), **-**. https://doi.org/10.1111/j.1478-0542.2004.00097.x

McDougall, B. S., & Hansson, A. (2002). Chinese concepts of privacy. BRILL. http://ebookcentral.proquest.com/lib/unc/detail.action?docID=253522

McFarland, D. A., Ramage, D., Chuang, J., Heer, J., Manning, C. D., & Jurafsky, D. (2013). Differentiating language usage through topic models. Poetics, 41(6), 607–625. https://doi.org/10.1016/j.poetic.2013.06.004

McGregor, S., Agres, K., Purver, M., & Wiggins, G. A. (2015). From Distributional Semantics to Conceptual Spaces: A Novel Computational Method for Concept Creation. *Journal of Artificial General Intelligence; Vienna*, *6*(1), 55–86. http://dx.doi.org/10.1515/jagi-2015-0004

Mercer, N. (2013). The Social Brain, Language, and Goal-Directed Collective Thinking: A Social Conception of Cognition and Its Implications for Understanding How We Think, Teach, and Learn. Educational Psychologist, 48(3), 148–168. https://doi.org/10.1080/00461520.2013.804394

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 262–272.

Mittelstadt, B. (2017). From Individual to Group Privacy in Big Data Analytics. Philosophy & Technology, 30(4), 475–494. https://doi.org/10.1007/s13347-017-0253-7

Mizutani, M., Dorsey, J., & Moor, J. H. (2004). The internet and Japanese conception of privacy. Ethics and Information Technology, 6(2), 121–128. https://doi.org/10.1023/B:ETIN.0000047479.12986.42

Moore, B. (1985). Privacy. Society, 22(4), 17–27. https://doi.org/10.1007/BF02701909

Mouter, N., & Vonk Noordegraaf, D. M. (2012). Intercoder reliability for qualitative research: You win some, but do you lose some as well? Proceedings of the 12th TRAIL Congress, 30-31 Oktober 2012, Rotterdam, Nederland. Retrieved from https://repository.tudelft.nl/islandora/object/uuid%3A905f391d-4b25-40cf-9292-e253b7e55db2

Mozur, P., & Lin, Q. (2019, October 3). Hong Kong Takes Symbolic Stand Against China's High-Tech Controls. *The New York Times*. https://www.nytimes.com/2019/10/03/technology/hong-kong-china-tech-surveillance.html

Mulligan, D. K., Koopman, C., & Doty, N. (2016). Privacy is an essentially contested concept: A multi-dimensional analytic for mapping privacy. Phil. Trans. R. Soc. A, 374(2083), 20160118. https://doi.org/10.1098/rsta.2016.0118

Nakada, M., & Tamura, T. (2005). Japanese Conceptions of Privacy: An Intercultural Perspective. Ethics and Information Technology, 7(1), 27–36. https://doi.org/10.1007/s10676-005-0453-1

Naftali, O. (2010). Caged golden canaries: Childhood, privacy and subjectivity in contemporary urban China. Childhood, 17(3), 297–311. https://doi.org/10.1177/0907568209345612

Nerghes, A. (2016). Words in Crisis: A relational perspective of emergent meanings and roles in text. https://research.vu.nl/en/publications/words-in-crisis-a-relational-perspective-of-emergent-meanings-and

Ng, S. H., & Bradac, J. J. (1993). Power in language: Verbal communication and social influence. Sage Publications, Inc.

Nissenbaum, H. (2004). Privacy as Contextual Integrity Symposium—Technology, Values, and the Justice System. Washington Law Review, 79, 119–158.

Ochs, C., & Ilyes, P. (2014). Sociotechnical Privacy. Mapping the Research Landscape.

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*, *2*. https://www.frontiersin.org/article/10.3389/fdata.2019.00013

Posner, Richard A. (1978). "The Right of Privacy". Sibley Lecture Series. Paper 22. http://digitalcommons.law.uga.edu/lectures_pre_arch_lectures_sibley/22

Qiang, J., Li, Y., Yuan, Y., & Liu, W. (2018). Snapshot ensembles of non-negative matrix factorization for stability of topic modeling. Applied Intelligence, 48(11), 3963–3975. https://doi.org/10.1007/s10489-018-1192-4

Qin, B., Strömberg, D., & Wu, Y. (2017). *Why Does China Allow Freer Social Media? Protests versus Surveillance and Propaganda* (SSRN Scholarly Paper ID 2910223). Social Science Research Network. https://doi.org/10.2139/ssrn.2910223

Quinn, K., Epstein, D., & Moon, B. (2019). We Care About Different Things: Non-Elite Conceptualizations of Social Media Privacy. *Social Media + Society*, *5*(3), 2056305119866008. https://doi.org/10.1177/2056305119866008

Radaelli, L., Sapiezynski, P., Houssiau, F., Shmueli, E., & de Montjoye, Y.-A. (2018). Quantifying Surveillance in the Networked Age: Node-based Intrusions and Group Privacy. ArXiv:1803.09007 [Cs]. http://arxiv.org/abs/1803.09007

Rączaszek-Leonardi, J., Fusaroli, R., & Caramelli, N. (2017). Rethinking Meaning: An Ecological Perspective on Language. Psychology of Language and Communication, 20(2), 92–97. https://doi.org/10.1515/plc-2016-0005

Resnik, P., Garron, A., & Resnik, R. (2013). Using Topic Modeling to Improve Prediction of Neuroticism and Depression in College Students. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1348–1353. https://aclanthology.org/D13-1133

Reid, J. (2012). “My Room! Private! Keep Out! This Means You!”: A Brief Overview of the Emergence of the Autonomous Teen Bedroom in Post-World War Ii America. Journal of the History of Childhood and Youth; Baltimore, 5(3), 419-443,500-501.

Reviglio, U., & Alunge, R. (2020). “I Am Datafied Because We Are Datafied”: An Ubuntu Perspective on (Relational) Privacy. Philosophy & Technology, 33. https://doi.org/10.1007/s13347-020-00407-6

Rice, R . E ., & Danowski, J . A . (1993). Is it really just a like a fancy answering machine? Comparing semantic networks of different types of voicemail users. Journal of Business Communication, 30, 369-397

Ricker, T. (2019). The US, like China, has about one surveillance camera for every four people, says report. The Verge. https://www.theverge.com/2019/12/9/21002515/surveillance-cameras-globally-us-china-amount-citizens

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: An R Package for Structural Topic Models. *Journal of Statistical Software*, *91*, 1–40. https://doi.org/10.18637/jss.v091.i02

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A Model of Text for Experimentation in the Social Sciences. Journal of the American Statistical Association, 111(515), 988–1003. https://doi.org/10.1080/01621459.2016.1141684

Roberts, M. E., Stewart, B. M., Tingley, D., & Airoldi, E. M. (2013). The structural topic model and applied social science. ICONIP 2013.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. American Journal of Political Science, 58(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Roberts M., Stewart B., & Airoldi E. (2018). A model of text for experimentation in the social sciences [Data set]. Harvard Dataverse. https://doi.org/10.7910/DVN/SIGIAU

Robinson, D., & Tannenberg, M. (2019). Self-censorship of regime support in authoritarian states: Evidence from list experiments in China. *Research & Politics*, *6*(3), 2053168019856449. https://doi.org/10.1177/2053168019856449

Robinson, S. D. (2019). Temporal topic modeling applied to aviation safety reports: A subject matter expert review. Safety Science, 116, 275–286. https://doi.org/10.1016/j.ssci.2019.03.014

Rosemont, H. (1974). *On Representing Abstractions in Archaic Chinese*. https://doi.org/10.2307/1397604

Rosemont, H., & Ames, R. T. (2008). The Chinese classic of family reverence: A philosophical translation of the Xiaojing. Honolulu, HI: University of Hawaii Press.

Rosen, G. (2001). Nominalism, Naturalism, Epistemic Relativism. Noûs, 35(s15), 69–91. https://doi.org/10.1111/0029-4624.35.s15.4

Sahlgren, M. (2006). The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Undefined. /paper/The-Word-Space-Model-%3A-Using-distributional-to-and-Sahlgren/1521ddb27860cc8834f8a82e62665bf983c8ad2c

Saldaña, J. (2016). The coding manual for qualitative researchers (Third edition.). Retrieved from

https://catalog.lib.unc.edu/catalog/UNCb8697471

Sarigol, E., Garcia, D., & Schweitzer, F. (2014). Online Privacy as a Collective Phenomenon. ArXiv:1409.6197 [Cs]. http://arxiv.org/abs/1409.6197

Shi, Y. (2002). *The Establishment of Modern Chinese Grammar: The formation of the resultative construction and its effects*. John Benjamins. https://doi.org/10.1075/slcs.59

Schwartz, P. M. (2000). Internet privacy and the state. Connecticut Law Review, 32(3), 815-860.

Scorolli, C., Binkofski, F., Buccino, G., Nicoletti, R., Riggio, L., & Borghi, A. (2011). Abstract and Concrete Sentences, Embodiment, and Languages. Frontiers in Psychology, 2, 227. https://doi.org/10.3389/fpsyg.2011.00227

Schaub, F., Balebako, R., & Cranor, L. F. (2017). Designing Effective Privacy Notices and Controls. IEEE INTERNET COMPUTING, 21(3), 70–77.

Shen, Z., & Eliassi-Rad, T. (2006). Visual Analysis of Large Heterogeneous Social Networks by Semantic and Structural Abstraction. IEEE Transactions on Visualization and Computer Graphics, 12(6), 1427–1439. https://doi.org/10.1109/TVCG.2006.107

Shen, X., & Truex, R. (2021). In Search of Self-Censorship. *British Journal of Political Science*, *51*(4), 1672–1684. https://doi.org/10.1017/S0007123419000735

Shim, J., Park, C., & Wilding, M. (2015). Identifying policy frames through semantic network analysis: An examination of nuclear energy policy across six countries. Policy Sciences, 48(1), 51–83. https://doi.org/10.1007/s11077-015-9211-3

Silverstein, M. (2004). "Cultural" Concepts and the Language-Culture Nexus. Current Anthropology, 45(5), 621–652. https://doi.org/10.1086/423971

Slotta, D. (2021, May 6). *Topic: Smartphone market in China*. Statista. Retrieved December 10, 2021, from https://www.statista.com/topics/1416/smartphone-market-in-china/.

Smart, A. (1993). Gifts, Bribes, and Guanxi: A Reconsideration of Bourdieu's Social Capital. *Cultural Anthropology*, *8*(3), 388–408.

Smith, R. A., & Parrott, R. L. (2012). Mental representations of HPV in Appalachia: Gender, semantic network analysis, and knowledge gaps. Journal of Health Psychology, 17(6), 917–928. https://doi.org/10.1177/1359105311428534

Soares, C. (2018). The Philosophy of Individualism: A Critical Perspective. International Journal of Philosophy and Social Values, 1. https://doi.org/10.34632/philosophyandsocialvalues.2018.2664

Sokolov, E., & Bogolubsky, L. (2015). Topic Models Regularization and Initialization for Regression Problems. Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications, 21–27. https://doi.org/10.1145/2809936.2809940

Solove, D. (2002). Conceptualizing Privacy. California Law Review, 90(4), 1087. https://doi.org/10.15779/Z382H8Q

Solove, D., & Schwartz, P. (2015). An Overview of Privacy Law (SSRN Scholarly Paper ID 2669879). Social Science Research Network. https://papers.ssrn.com/abstract=2669879

Song, S. Y., Shen, F., Yao, M., & Wildman, S. S. (2013). *Unmasking News in Cyberspace: Examining Censorship Patterns of News Portal Sites in China* (SSRN Scholarly Paper ID 2275437). Social Science Research Network. https://doi.org/10.2139/ssrn.2275437

Stemler, S. (2001). An Overview of Content Analysis. *Practical Assessment, Research & Evaluation*, *7*(17), 1–6.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring Topic Coherence over Many Models and Many Topics. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 952–961. https://www.aclweb.org/anthology/D12-1087

Stone, L. (1991). The Public and the Private in the Stately Homes of England, 1500-1990. *Social Research; Camden, N. J.*, *58*(1). https://search.proquest.com/docview/1297197441/citation/A9660A62017541E6PQ/1

Su, Z., Xu, X., & Cao, X. (2021). What explains popular support for government monitoring in China? *Journal of Information Technology & Politics*, *0*(0), 1–16. https://doi.org/10.1080/19331681.2021.1997868

Tai, Q. (2014). China's Media Censorship: A Dynamic and Diversified Regime. *Journal of East Asian Studies*, *14*(2), 185–210. https://doi.org/10.1017/S1598240800008900

Tang, J., Meng, Z., Nguyen, X., Mei, Q., & Zhang, M. (2014). Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis. *Proceedings of the 31st International Conference on Machine Learning*, 190–198. https://proceedings.mlr.press/v32/tang14.html

Tavani, H. T. (2007). Philosophical Theories of Privacy: Implications for an Adequate Online Privacy Policy. Metaphilosophy, 38(1), 1–22. https://doi.org/10.1111/j.1467-9973.2006.00474.x

Taylor, L., Floridi, L., & Sloot, B. van der (Eds.). (2017). Group Privacy: New Challenges of Data Technologies. Springer International Publishing. //www.springer.com/us/book/9783319466064

Tener, C. A. (2002). Hold the Phone: PA High Court Says No Reasonable Expectation of Privacy during Phone Call Note. *University of Pittsburgh Law Review*, *64*(4), 855–878.

Tifferet, S. (2019). Gender differences in privacy tendencies on social network sites: A meta-analysis. Computers in Human Behavior, 93, 1–12. https://doi.org/10.1016/j.chb.2018.11.046

Tiejun, P., Chunlei, X., Leina, Z., Yufeng, H., & Lingbin, B. (2010). ONE-CARD System Based on the Second Generation ID Card in China. 2010 International Conference on E-Business and E-Government, 108–111. https://doi.org/10.1109/ICEE.2010.35

The Associated Press. (2022, January 6). *France fines Google, facebook millions over tracking consent*. NBCNews.com. Retrieved January 31, 2022, from https://www.nbcnews.com/tech/tech-news/france-fines-google-facebook-millions-tracking-consent-rcna11220

Tran, T. P. (2017). Personalized ads on Facebook: An effective marketing tool for online marketers. *Journal of Retailing and Consumer Services*, *39*(C), 230–242.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research, 37, 141–188.

Uz, I. (2014). Individualism and First Person Pronoun Use in Written Texts Across Languages. Journal of Cross-Cultural Psychology, 45(10), 1671–1678. https://doi.org/10.1177/0022022114550481

Vedder, A. (1999). KDD: The challenge to individualism. Ethics and Information Technology, 1(4), 275–281. https://doi.org/10.1023/A:1010016102284

Veltri, G. A., & Atanasova, D. (2017). Climate change on Twitter: Content, media ecology and information sharing behaviour. Public Understanding of Science, 26(6), 721–737. https://doi.org/10.1177/0963662515613702

Villani, C., Lugli, L., Liuzza, M. T., & Borghi, A. M. (2019). Varieties of abstract concepts and their multiple dimensions. Language and Cognition, 1–28. https://doi.org/10.1017/langcog.2019.23

Vincent, J. (2016, December 16). *Evernote backtracks on controversial privacy policy*. The Verge. Retrieved December 11, 2021, from https://www.theverge.com/2016/12/16/13979778/evernote-privacy-policy-opt-out.

Vincent, P., Antoine, B., Sonia, B. M., & Lionel, B. (2019). The Long Road to Computational Location Privacy: A Survey. IEEE Communications Surveys & Tutorials, 21(3), 2772–2793. https://doi.org/10.1109/COMST.2018.2873950

Wallach, H. M. (2006). Topic modeling: Beyond bag-of-words. Paper presented at the 23rd International Conference on Machine Learning, Pittsburgh, PA.

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. In Proceedings of the 26th Annual International Conference on Machine Learning (pp. 1105–1112). Association for Computing Machinery. https://doi.org/10.1145/1553374.1553515

Wang, H. (2012). The Conceptual Basis of Privacy Standards in China and its Implications for China's Privacy Law. Frontiers of Law in China, 7(1), 134–160. https://doi.org/10.3868/s050-001-012-0007-4

Wang, L., & Xiong, B. (2021). Personality Rights in China's New Civil Code: A Response to Increasing Awareness of Rights in an Era of Evolving Technology. *Modern China*, *47*(6), 703–739. https://doi.org/10.1177/0097700420977826

Warglien, M., & Gärdenfors, P. (2013). Semantics, conceptual spaces, and the meeting of minds. *Synthese*, *190*(12), 2165–2193. https://doi.org/10.1007/s11229-011-9963-z

Warren, S. D., & Brandeis, L. D. (1890). The Right to Privacy. Harvard Law Review, 4(5), 193–220. JSTOR. https://doi.org/10.2307/1321160

Westin, A. F. (2003). Social and Political Dimensions of Privacy. Journal of Social Issues, 59(2), 431–453. https://doi.org/10.1111/1540-4560.00072

Whitman, J. Q. (2003). *The Two Western Cultures of Privacy: Dignity Versus Liberty* (SSRN Scholarly Paper ID 476041). Social Science Research Network. https://doi.org/10.2139/ssrn.476041

Whitman, C. (1985). Privacy in Confucian and Taoist Thought. Book Chapters. https://repository.law.umich.edu/book_chapters/21

Whorf, B. L. (1956). Language, thought, and reality: Selected writings. Cambridge, Mass.

Wildau, G. (2017, December 27). China unveils digital ID card linked to Tencent's WeChat. Financial Times. https://www.ft.com/content/3e1f00e2-eac8-11e7-bd17-521324c81e23

Xiong, Y., Cho, M., & Boatwright, B. (2019). Hashtag activism and message frames among social movement organizations: Semantic network analysis and thematic analysis of Twitter during the #MeToo movement. Public Relations Review, 45(1), 10–23. https://doi.org/10.1016/j.pubrev.2018.10.014

Wageningen, E. van. (2004). Top court to ponder high-tech snooping; Marijuana case involving Kingsville man tests privacy: [Final Edition]. *The Windsor Star*, A1 Front.

Wong, P.-H. (2009). What Should We Share?: Understanding the Aim of Intercultural Information Ethics. SIGCAS Comput. Soc., 39(3), 50–58. https://doi.org/10.1145/1713066.1713070

Xin, K. R., & Pearce, J. L. (1996). Guanxi: Connections as Substitutes for Formal Institutional Support. *The Academy of Management Journal*, *39*(6), 1641–1658. https://doi.org/10.2307/257072

Xu, Y., Malt, B. C., & Srinivasan, M. (2017). Evolution of word meanings through metaphorical mapping: Systematicity over the past millennium. Cognitive Psychology, 96, 41–53. https://doi.org/10.1016/j.cogpsych.2017.05.005

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. WWW. https://doi.org/10.1145/2488388.2488514

Yao, Y., Xia, H., Huang, Y., & Wang, Y. (2017). Privacy Mechanisms for Drones: Perceptions of Drone Controllers and Bystanders. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 6777–6788. https://doi.org/10.1145/3025453.3025907

Youth.cn 中国青年网 (2019, January 25). *Kaifang santianhou shipin chuxianzai seqing wangzhan, jiudian zhenkong shexiangji toulu fangbushengfang* [Hotel guests were videotaped and uploaded to pornography websites: miniature cameras are everywhere. China Youth Official Website]. Retrieved December 10, 2021, from https://baijiahao.baidu.com/s?id=1623597377823280003&wfr=spider&for =pc.

Yuan, E. J., Feng, M., & Danowski, J. A. (2013). “Privacy” in Semantic Networks on Chinese Social Media: The Case of Sina Weibo. Journal of Communication, 63(6), 1011–1031. https://doi.org/10.1111/jcom.12058

Yun, J., & Geum, Y. (2020). Automated classification of patents: A topic modeling approach. *Computers & Industrial Engineering*, *147*, 106636. https://doi.org/10.1016/j.cie.2020.106636

Zarrow, P. (2002) ‘The Origins of Modern Chinese Concepts of Privacy: Notes on Social Structure and Moral Discourse’, in B.S. McDougall and A. Hansson (eds) Chinese Concepts of Privacy, pp. 21–46. Leiden: Brill

Zdrazilova, L., Sidhu, D. M., & Pexman, P. M. (2018). Communicating abstract meaning: Concepts revealed in words and gestures. Philosophical Transactions of the Royal Society B: Biological Sciences, 373(1752), 20170138. https://doi.org/10.1098/rstb.2017.0138

Zhang, L. (2018). *China: New regulation on police cybersecurity supervision and inspection powers issued*. The Library of Congress. Retrieved January 10, 2022, from https://www.loc.gov/item/global-legal-monitor/2018-11-13/china-new-regulation-on-police-cybersecurity-supervision-and-inspection-powers-issued/

Zhao, X., Zhan, M., & Jie, C. (2018). Examining multiplicity and dynamics of publics’ crisis narratives with large-scale Twitter data. Public Relations Review, 44(4), 619–632. https://doi.org/10.1016/j.pubrev.2018.07.004

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., & Li, X. (2011). Comparing

Twitter and Traditional Media Using Topic Models. In P. Clough, C. Foley, C. Gurrin, G. J. F. Jones, W. Kraaij, H. Lee, & V. Mudoch (Eds.), Advances in Information Retrieval (pp. 338–349). Springer. https://doi.org/10.1007/978-3-642-20161-5_34

Zheng, Q., & Bashir, M. (2020). Investigating the Differences in Privacy News Based on Grounded Theory. In W. Karwowski, R. S. Goonetilleke, S. Xiong, R. H. M. Goossens, & A. Murata (Eds.), Advances in Physical, Social & Occupational Ergonomics (pp. 528–535). Springer International Publishing. https://doi.org/10.1007/978-3-030-51549-2_70

Zhong, J. (2015). On Translatability and Untranslatability in Translation of Chinese Idioms. Overseas English, *22*, 166–167.