TEACHERS IN SOCIAL MEDIA: A DATA SCIENCE PERSPECTIVE

By

Hamid Karimi

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computer Science – Doctor of Philosophy

2021

ABSTRACT

TEACHERS IN SOCIAL MEDIA: A DATA SCIENCE PERSPECTIVE

By

Hamid Karimi

Social media has become an integral part of human life in the 21st century. The number of social media users was estimated to be around 3.6 billion individuals in 2020. Social media platforms (e.g., Facebook) have facilitated interpersonal communication, diffusion of information, the creation of groups and communities, to name a few. As far as education systems are concerned, online social media has transformed and connected traditional social networks within the schoolhouse to a broader and expanded world outside. In such an expanded virtual space, teachers engage in various activities within their communities, e.g., exchanging instructional resources, seeking new teaching methods, engaging in online discussions. Therefore, given the importance of teachers in social media and its tremendous impact on PK-12 education, in this dissertation, we investigate teachers in social media from a data science perspective. Our investigation in this direction is essentially an interdisciplinary endeavor bridging modern data science and education. In particular, we have made three contributions, as briefly discussed in the following.

Current teachers in social media studies suffice to a small number of surveyed teachers while thousands of other teachers are on social media. This hinders us from conducting large-scale data-driven studies pertinent to teachers in social media. Aiming to overcome this challenge and further facilitate data-driven studies related to teachers in social media, we propose a novel method that automatically identifies teachers on Pinterest, an image-based social media popular among teachers. In this framework, we formulate the teacher identification problem as a positive unlabelled (PU) learning where positive samples are surveyed teachers, and unlabelled samples are their online friends. Using our framework, we build the largest dataset of teachers on Pinterest.

With this dataset at our disposal, we perform an exploratory analysis of teachers on Pinterest while considering their genders. Our analysis incorporates two crucial aspects of teachers in social

media. First, we investigate various online activities of male and female teachers, e.g., topics and sources of their curated resources, the professional language employed to describe their resources. Second, we investigate male and female teachers in the context of the social network (the graph) they belong to, e.g., investigating structural centrality, gender homophily. Our analysis and findings in this part of the dissertation can serve as a valuable reference for many entities concerned with teachers' gender, e.g., principals, state, and federal agencies.

Finally, in the third part of the dissertation, we shed light on the diffusion of teacher-curated resources on Pinterest. First, we introduce three measures to characterize the diffusion process. Then, we investigate these three measures while considering two crucial characteristics of a resource, e.g., the topic and the source. Ultimately, we investigate how teacher attributes (e.g., the number of friends) affect the diffusion of their resources. The conducted diffusion analysis is the first of its kind and offers a deeper understating of the complex mechanism driving the diffusion of resources curated by teachers on Pinterest.

To my wife, daughter, parents, siblings, and entire family for their love and support.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Dr. Jiliang Tang, my Ph.D. advisor, for his support and encouragement during my Ph.D. He helped me with numerous skills and provided me with countless professional and academic opportunities, e.g., writing a research paper, writing a grant proposal, polishing a novel idea, mentoring undergrad and grad students, job interview skills, conducting interdisciplinary research, managing a research lab to just a few. I feel honored and lucky to have been his Ph.D. student. He tirelessly helped me to be a better scholar and prepared me for my future job.

I want to extend my gratitude to my other Ph.D. committee members: Dr. Pang-Ning Tan, Dr. Arun Ross, Dr. Kenneth Frank, and Dr. Kaitlin Torphy, for their insightful comments and feedback. I had a chance to collaborate with Dr. Tan's lab in 2018 on a joint project about compromised account detection. This collaboration was through Dr. Courtland VanDam – the Ph.D. student of Dr. Tan at the time- whom I am also thankful for our fruitful collaboration. Dr. Tan's expertise in data mining has been a great source of help both in our joint project and this dissertation. Dr. Ross has had an instrumental role in improving the quality of this dissertation by providing me with great comments. His invaluable expertise has been constructive and inspiring. I met Dr. Kenneth Frank and Dr. Kaitlin Torphy through the Teachers in Social Media (TISM) project– an interdisciplinary project founded by Dr. Torphy wherein I have been leading computational efforts since Fall 2018. Dr. Frank has been an extraordinary mentor for me. He taught me how to think like a social scientist. He helped me hone my data science skills to answer educational research questions. I have continuously utilized the precious expertise of Dr. Frank on social network analysis and education science during my Ph.D. and especially in this dissertation. His endless support during my Ph.D. has had a significant impact on my professional and academic growth, for which I am always grateful. Finally, I am thankful for all the support I have received from Dr. Torphy. Throughout our close collaboration in the TISM project, I have learned a lot from her. Her support has prepared me for being better equipped as an interdisciplinary researcher, especially for

applying data science techniques to education.

I joined the Data Science and Engineering (DSE) Lab at the end of the Summer 2017 semester. During my Ph.D., I have had the pleasure and fortune of having supportive and encouraging friends and colleagues from the DSE lab. I thank all the fantastic members of the DSE lab. In particular, I want to thank my dear friend, Dr. Tyler Derr, who has been highly supportive during our collaborations in the DSE lab. Moreover, I want to thank several outstanding undergraduate students that I have had a chance to mentor, including Liyang Ye, Aaron Brookhouse, and Xochitl Weiss. Finally, I am thankful to all my collaborators from outside the DSE Lab. In particular, I want to express my special gratitude to the leading social scientist, Dr. H. Russell Bernard, who has been an incredible mentor for me.

What I am today owed to my parents: my dear and kind mom, Farast Hossein Panahi, and my wonderful late father, Esmaeil Karimi. I am eternally grateful for their unconditional love and support. Also, I am thankful for the love and support I have received from my incredible wife, Nazanin Donyapour, who is not only a wife but also a dear friend. Moreover, I am thankful to my older brother, Omid Karimi, for his support during the years of my education. Furthermore, I am also thankful for the support from my in-laws, especially my late mother-in-law, whose kindness was immeasurable. Finally, again, I would like to thank my entire family, especially my amazing and stunning siblings.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Social media has become an integral part of human life in the 21st century. The number of social media users in 2020 was estimated to be around 3.6 billion individuals [1]. Social media platforms (e.g., Facebook) have facilitated interpersonal communication, diffusion of information, the creation of groups and communities, to name a few. As far as education systems are concerned, online social media has transformed and connected traditional social networks within the schoolhouse to a broader and expanded world outside [2]. Thanks to advancements in communication, educators have access to ample online instructional resources curated and shared across social media platforms. In such an expanded virtual space, teachers engage in various activities within their community, e.g., exchanging instructional resources, seeking new teaching methods, and engaging in online discussions [3, 4, 5, 6, 7, 8, 9, 10, 11]. Students use social media as well–for example, to supplement educational materials and interact with others [12, 13, 14, 15, 16, 17]. Furthermore, educational policymakers take advantage of social media to infer public opinion about new policies [18]. In addition, parents seek out resources within social media to supplement their children with educational materials [19]. Hence, today's education, specially PK-12 education, is closely intertwined with online social media and entails various entities. Nevertheless, the essential entities who play a critical role in bridging education and social media are *teachers*. Next, we provide several reasons behind the importance of teachers in social media and why it deserves our investigation.

- While we might deem social media merely a communication tool, it is far beyond that. In a broader sense, social media is the reflection of who we are as humans, and it increasingly plays a critical role in shaping our identity [20]. Moreover, this *digital* identity is not limited to our personal interests and inherent beliefs; it encompasses the professional aspect of our life as well [21]. In other words, a significant portion of many people's professional life

is reflected on social media platforms. As far as teachers are concerned, this reflection involves performing various professional activities, e.g., curriculum development and lesson planning [22, 23, 24, 25, 4, 26, 27, 28]. Furthermore, as explained in [29], today's teachers conceptualize their professional identity through online social media. Thus, teaching as a career is no longer confined to the physical world, and its development has a huge presence in online social media. In summary, as much as we care about teachers and their profession (i.e., teaching), we need to care about their online social media presence as well.

- One of the primary motivations of teachers to turn to online social media is to supplement their instructional and educational resources. In the classroom, many teachers encounter needing additional educational resources to improve their students' learning. However, traditional educational resource curation (e.g., asking a colleague) is time-consuming and not scalable [30]. In contrast, seeking out educational resources from other teachers in online social media is easy. Specifically, the diffusion of online resources can be rapid within the same day, and teachers may integrate resources into their classroom practices quickly and conveniently. It is worth mentioning that during the COVID-19 pandemic, online educational resources curated by teachers have become specially essential [31]. Moreover, while teachers often find a minimal voice in school decisions, they are provided with tremendous flexibility and diversity in the portfolio of resources they can obtain from online social media. Hence, social media acts as a reliable support group for teachers and aids in the widespread diffusion of educational resources used in various classroom activities.

- Compared to the traditional data for studying teachers (i.e., interviews or surveys), the social media data offers several significant benefits. First, it is directly related to their educational classroom practices, while in surveys/interviews, we essentially create a proxy to tap into teachers' pedagogical efforts. Second, we can have real-time access to a large amount of teacher-related data in social media. Third, the data from social media is without the response bias (incorrect responses from participants in surveys/interviews [32]) and the observer

bias (inaccuracy or subjectivity in recording the responses [33]). Ultimately, compared to surveys/interviews, the social media data of teachers offer a wide variety of formats, e.g., images, videos, texts. Thus, the social media data of teachers offers a great potential to study teachers in the current digital age.

We refer the reader interested in further discussion about why we need to care about teachers in social media to the engaging article by Frank and Torphy [34].

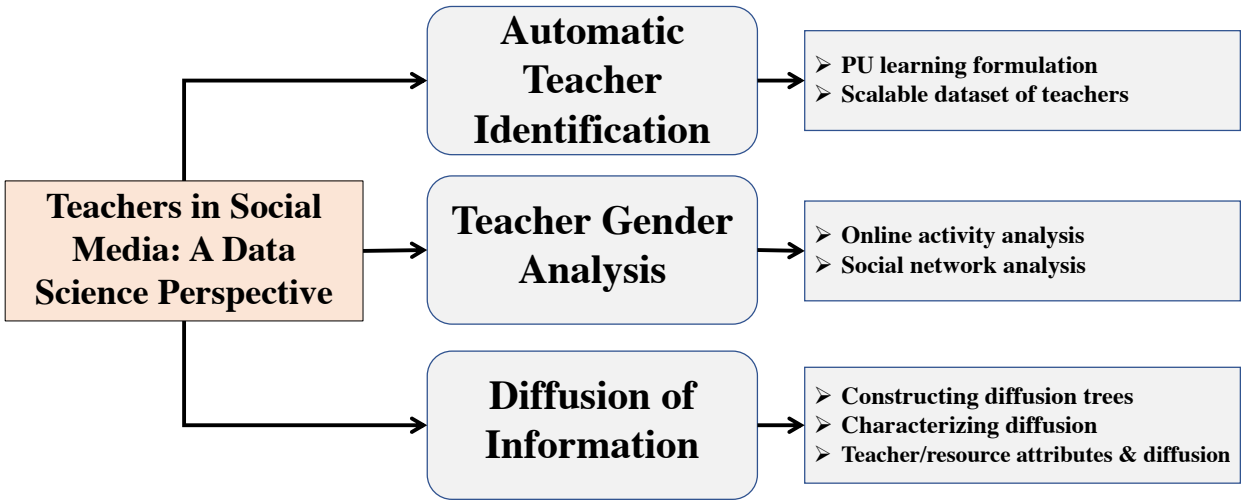


Figure 1.1: An overview of the research contributions presented in this dissertation.

Therefore, given the importance of teachers in social media and its tremendous impact on PK-12 education, in this dissertation, we investigate teachers in social media from a data science perspective. Our investigation in this direction is essentially an interdisciplinary endeavor bridging modern data science and education. A unique characteristic of this dissertation is that we have incorporated real social media data from thousands of teachers and their online friends in our investigations. Moreover, we have overcome significant technical challenges, proposed novel machine learning, and data mining algorithms, and performed various novel analyses about teachers in social media. An overview of the dissertation research contributions is summarized in Figure 1.1. Next, we introduce our major contributions and the addressed challenges.

1.1 Contributions

1.1.1 Automatic Teacher Identification

In response to the importance of teachers in social media, recent years have witnessed a rapid increase in studies assessing teachers in social media and its impact on the quality of education. These studies, however, suffice to a small number of surveyed teachers (at most a few hundred) while there are plenty of other teachers in social media. For instance, previous studies showed that more than 75% of American teachers use Pinterest to seek lessons and educational materials [35, 36]. Hence, using only a small number of teachers is not representative of the large population of teachers present in social media. This underrepresentation hinders us from conducting large-scale data-driven studies pertinent to teachers in social media. For instance, if we intend to study the diffusion of information among teachers, using only a small number of teachers, we cannot fully characterize this diffusion since we need to have a broader picture of the online network that embeds teachers. Aiming to overcome this challenge and further facilitate data-driven studies related to teachers in social media (including those presented in this dissertation), we propose a framework that automatically identifies teachers on Pinterest, an image-based social media popular among teachers. In this framework, we formulate the teacher identification problem as a positive unlabeled (PU) learning [37] where the positive samples are some surveyed (labeled) teachers, and unlabeled samples are some of their online friends. Using our framework, we build the largest dataset of teachers on Pinterest. We believe our proposed method has great potential in advancing research on teachers in social media.

1.1.2 Teacher Gender Analysis

For decades, there have been numerous studies in educational literature investigating the role of teacher's gender and how it affects the quality of education and particularly students' success [38, 39, 40, 41, 42, 42]. The primary motivation behind these studies is that the academic environment created by teachers largely influences the way students see themselves as learners/students. In

this regard, some argue that the behavioral differences in male vs. female teachers are what indeed matter, e.g., the way teachers manage the classroom or prepare materials [43]. However, while identifying the behavioral differences between male and female teachers has been investigated before [44, 41], no study to date juxtaposes male and female teachers in social media and investigates their behavioral differences (and similarities). Given the significance of social media in shaping teachers' professional lives, we believe analyzing teachers' online behavior while considering their gender is essential. Unfortunately, an obstacle to conducting such a study has been the availability of a large dataset of teachers in social media. However, thanks to our automatic teacher identification framework, we have overcome this obstacle where we have built a rich dataset of teachers in social media. Hence, given this dataset at our disposal, we perform an exploratory analysis of male and female teachers on Pinterest. Our study incorporates two crucial aspects of teachers. First, we investigate various online activities of male and female teachers, such as topics of their curated resources. The motivation for this type of analysis is to understand male and female teachers through the lens of their resource curation process. Second, we investigate male and female teachers in the context of their social network (the graph). The performed social network analysis complements online activity analysis by examining how teachers are connected in a social network. Notably, we look into the critical notion of homophily (i.e., the tendency of similar individuals to connect in a network) and substantiate gender homophily among teachers. Our analysis and findings in this part of the dissertation can serve as a valuable reference for many entities concerned with teachers' gender, e.g., educational scientists, policymakers, principals, state and federal governments.

1.1.3 Diffusion of Teacher-curated Resources

As mentioned before, previous studies have reported that the diffusion of information from social media to the classroom can be potentially very fast [4, 5, 34, 45]. This is encouraging as teachers can quickly implement these resources in their pedagogical practices. Nevertheless, our understanding of how resources curated by teachers diffuse across the network and what factors affect this diffusion is still slim. Again, a barrier to such understating has been the availability of a large

dataset of teachers in social media, including curated resources and their diffusion dynamics. We overcome this obstacle by constructing the diffusion trees for more than one million resources curated by teachers. Another major challenge in investigating the diffusion of information is how to characterize the diffusion. To address this challenge, we introduce three crucial measures which consider different aspects of the diffusion process. The first measure is *volume*, i.e., the number of users who have received and saved a resource. The second measure is *virality* which captures the structural virality of a resource based on its diffusion tree. Finally, the third metric includes *velocity* measured by calculating (a) the first time a resource is re-pinned and (b) the average time difference between consecutive re-pins.[1] Using the introduced diffusion measures, we investigate how different attributes of resources (e.g., their topics) affect the diffusion. Moreover, through several regression analyses, we determine how teacher attributes (e.g., the number of their followers) affect the diffusion. Our investigation in this part of the dissertation is the first of its kind and offers a deeper understating of the complex mechanism driving the diffusion of resources curated by teachers.

1.2 Organization

The remainder of this dissertation is organized as follows. First, in Chapter 2, we introduce the preliminaries, including current studies on teachers in social media and the data collection process. In Chapter 3, we present our proposed approach for automatic teacher identification. We verify the working of this approach on a large set of carefully constructed sets of teacher and non-teacher users on Pinterest. Chapter 4 is devoted to analyzing the online behavior of male and female teachers. We conduct two types of analysis, namely online activity analysis (Section 4.3) and social network analysis (Section 4.4). In Chapter 5, we present our study on the diffusion of resources curated by the teachers. Finally, Chapter 6 concludes the dissertation and offers promising future research directions.

[1] In Pinterest, *share* is called *re-pin*.

CHAPTER 2

FOUNDATIONS AND PRELIMINARIES

In this chapter, we present the foundations and preliminaries necessary for the rest of the dissertation. In Section 2.1, we review current studies related to teachers in social media, and in Section 2.2, we present the data collection process.

2.1 Current Studies on Teachers in Social Media

Social media platforms are ubiquitous and have transformed almost every aspect of our lives. As far as the education system is concerned, many teachers use online social media platforms (e.g., Facebook, Twitter, Pinterest) for educational engagement. Hence, during the past few years, there has been a growing number of studies focusing on why and how teachers use online social media. In this section, we review some of these studies. As for particular social media platforms, most studies have focused on Facebook, Twitter, and recently Pinterest since these platforms are the ones predominately used by teachers [46]. Therefore, we review the notable studies whose social media platform is Facebook, Twitter, or Pinterest.

2.1.1 Facebook-driven Studies

Rutherford [47] conducted one of the earliest studies on how teachers use Facebook for professional career development. They used a mixture of qualitative and quantitative analysis and investigated 384 users who actively participated in the *Ontario teacher Facebook group*. They found out that majority of discussions were practical and related to teacher professional development. Cinkara and Arslan [48] found similar results for EFL (English as a Foreign Language) teachers leveraging Facebook groups for professional career development. Aiming at determining the professional development practices, Ranieri et al. [49] investigated five Italian Facebook groups used by 1,1170 teachers. They thoroughly analyzed the dynamics of these groups and their members by taking advantage of social capital theory [50]. Moreover, they inspected membership motivations based

on the type of the professional group (i.e., generic and thematic) and how the dynamic of social capital differs in these groups. In general, they indicated that Facebook assists in improving professional development. Bicen and Uzunboylu [51] investigated the usefulness of Facebook in education. They set up an online learning environment for 71 teachers on Facebook where teachers could do various activities such as sharing pedagogical videos and participating in discussions with students and other teachers. Based on their results, participating teachers responded positively to the incorporation of Facebook in teaching, which ultimately helped improve students' learning. Authors in [52] investigated views of students and teachers on introducing information and communications technology outside the classroom, namely Google and Facebook discussion groups. They gathered survey and interview data from 283 teachers and students after the deployment of Google and Facebook groups. Their findings demonstrated students' positive attitude toward these technologies, while some teachers were reluctant to integrate them in their educational activities, mostly due to time affordability. Sumuer et al. [53] looked into the habits related to how teachers use Facebook (N=616). Similar to other Facebook-related studies, they showed professional usage of Facebook. However, interestingly, they recognized that many teachers have privacy concerns, such as sharing their personal information with students and parents. Similar privacy concerns were identified by [54]. In an interesting study, Forkosh-Baruch et al. [55] investigated *willing-to-connect* (via Facebook) for 160 Israeli teachers and 587 students based on various personal attributes e.g., age. They showed that *willing-to-connect* teachers were younger than *not-willing-to-connect* ones. However, it turned out that *willing-to-connect* students were older than *not-willing-to-connect* ones. In a similar study, Asterhan and Rosenberg [56] focused on student-teacher communication on Facebook (N=198) and pointed out the different challenges facing teachers on Facebook, e.g., privacy concerns. Additionally, their results showed that teachers utilize Facebook for instructional and psycho-pedagogical communications. Following the study in [56], Schwarz and Caduri [57] performed an in-depth analysis of the interaction logs of five teachers with students to characterize their communications. Although one teacher merely practiced the transmission of knowledge (which has been shown to be an ineffective teaching style [58, 59]), the other teachers help foster

positive educational practices, namely social learning, autonomy, and active engagement. In somewhat a unique study, Robson [29] shed light on the conceptualization of the professional identity of teachers by conducting interviews with 20 teachers who were using Facebook. They concluded that social media allows teachers to express themselves and form their ideal professional image. Blonder and Rap [60] investigated how Technological Pedagogical Content Knowledge (TPACK) and the self-efficacy beliefs of 12 high school chemistry teachers changed after introducing several chemistry-related Facebook groups. Their study showed that TPACK improved among teachers, and they developed TPACK skills geared more specifically toward teaching chemistry. Ab Rashid [61] analyzed Facebook timelines of 34 high school teachers using thematic, and discourse analysis approaches. Their investigation showed that teachers receive support from their peers through conversations reflected on their Facebook timeline, which eventually help them in their teaching. In conclusion, most studies have demonstrated that Facebook is primarily used for professional development by teachers, which consequently can improve the quality of education. Despite this, some teachers are reluctant to use Facebook due to privacy concerns.

2.1.2 Twitter-driven Studies

Cano [62] experimented with introducing Twitter as a learning and teaching tool in three Spanish high schools across three subjects: Spanish language, social science, and natural science (15 teachers and 280 students). Their results revealed improvement in students' grades after introducing Twitter. In a similar study, Van Vooren and Bess [63] investigated the relationship between the use of Twitter and the students' success (N=86 students). Their findings suggested that students who received support from their teacher via Twitter performed significantly better in standardized tests compared to their counterparts who received no Twitter-related support. Similarly, Noble et al. [64] indicated that teacher-student interactions on Twitter led to improved student learning and reinforced the trust between students and teachers. Wesely [65] investigated the role of Twitter in the professional development of 9 world language teachers. They monitored teachers participating in #edchat on Twitter for more than a year and tried to determined how these teachers develop

communities of practice which "[are] groups of people who share a concern or a passion for something they do and learn how to do it better as they interact regularly" [66]. They showed that Twitter helps improve teacher career development. Following on the work of [65], Britt and Paulus [67] performed a qualitative analysis of interviews with eight teachers who participated in #edchat on Twitter. Their findings indicated that #edchat is an effective community of practice reinforcing teacher professional development. Moreover, [68] found similar results for 30 identified influential educators (teachers) on Twitter. In their prominent studies Carpenter and Krutka [69, 70] conducted a thorough analysis of how and why teachers leverage Twitter for professional career development. Their findings indicated that Twitter plays an essential role in the professional development of teachers. In particular, the individualization offered by Twitter is a major advantage compared to traditional professional development approaches. Moreover, their results suggested that Twitter is of big help in combating teacher isolation where they received valuable support from their peers online. Visser et al. [71] surveyed 324 active educators on Twitter about their experience of using Twitter for professional purposes. Their findings indicated that teachers emphasized more on professional usage of Twitter than personal. Through a qualitative analysis, they showed that the surveyed teachers had created a positive socialized knowledge community accommodating interpersonal communication and fostering collaboration. A noteworthy aspect of this study is its teacher recruitment (sampling) process. Unlike most studies wherein teachers are pre-determined (e.g., through contacting school or district officials), they employed a snowball sampling strategy where they broadcast the link to their survey and asked teachers to participate themselves and further invite their friends to participate. Our proposed teacher identification method in Chapter 3 follows a similar strategy where we start from some initial teachers and attempt to expand (or snowball) the sample size, i.e., identify more teachers. However, since our purpose is to incorporate teachers' social media data, we focus on automating the teacher identification instead of directly asking for the teacher's participation. Trust et al. [72] performed a similar snowball sampling on multiple platforms (e.g., Twitter, Facebook, Google+) to survey teachers on the effectiveness of online platforms in professional development. Their results suggested that participating teachers have

found social media platforms to be supportive of their professional growth. Also, teachers reported improvement in their student learning after utilizing online social media. Studies in [73, 74] reported on the critical role of Twitter in teacher professional development as well as participating teachers' perception regarding Twitter. In particular, participants in [73] revealed that Twitter had offered them a sense of belonging and community, which is even stronger than what their physical workplace would deliver. Rosenberg et al. [75] conducted an interesting data-driven analysis of teachers on Twitter through the lens of the affinity space framework [76], a physical or virtual space revolving around a certain topic wherein people utilize a medium to interact with each other about that topic. The authors used state educational Twitter hashtags (SETHs) to define the affinity space on Twitter. SETHs are educational state-level Twitter hashtags developed by educators to participate in the educational discussions, e.g., *#miched* for Michigan or *#nebedchat* for Nebraska. They collected more than 500k tweets over six months, covering 68,552 unique Twitter users. Then, to answer who is participating in these affinity spaces, they manually identified 500 Twitter profiles belonging to educators (e.g., teacher, administrator). They also determined how active educators are in each state and further characterized their tweet timing behavior, e.g., percentage of tweets per day of the week. They found out that SETHs are effective spaces for teacher professional development. Rehm and Notten [6] leveraged social capital theory [50] and performed a qualitative study of 4,196 Twitter users participating in #EDchatDE, a hashtag developed for educational conversations in Germany. They found out that teachers' social capital would increase through participating in #EDchatDE. In conclusion, similar to Facebook, Twitter is used for professional development. Moreover, according to reviewed studies, Twitter offers a more interactive environment where teachers can participate in education-related discussions around certain topics. Compared to Facebook, the open and public nature of Twitter makes teachers less concerned with privacy issues. Eventually, the interested reader can refer to [77] to know more about how and why educators utilize Twitter.

2.1.3 Pinterest-driven Studies

Pinterest is an image-based personalized social media platform that draws 150 million active users per month. American teachers frequently use Pinterest as a common social media platform and virtual resource pool for professional purposes [4, 34, 78, 79, 80]. According to a national survey conducted by RAND Corporation, the majority of elementary and secondary teachers in the U.S. turn to Pinterest in response to recent national education reform (e.g., Common Core State Standards Reform [81]) or their instructional needs [82]. Frank et al. [25] analyzed the role of social networks in providing emerging beneficial opportunities for education. They argued that social networks outside schools, especially online ones like Pinterest, have a great potential to distribute knowledge and expertise among teachers equally. Hence, given the importance of Pinterest in education, there has been a growing number of studies investigating teachers on Pinterest.[1] Next, we review notable studies in this area.

Through an interview with eight teachers, Carpenter et al. [80] conducted a qualitative analysis on how teachers use Pinterest. They recruited teachers via a snowball sampling on Twitter, where they asked Twitter users to participate in an interview. Theoretically, they based their study on 1) Pinterest as an *affinity space* wherein teachers share common interests, and 2) teachers on Pinterest as *teacherpreneurs* who strive and spend time to impact beyond their classroom while not necessary to do so via traditional means, e.g., administrative roles. Although their sample size is small, their findings are interesting. They identified seven themes on how and why teachers use Pinterest. Notably, participants described Pinterest as a content curation tool. More specifically, teachers perceived Pinterest as an organizer or binder of the resources they encounter on the web or create themselves. This unique property of Pinterest has been demonstrated in previous (non-educational) studies [83, 84, 85, 86, 87, 88]. In fact, Pinterest as a social curation tool is what makes it very appealing among teachers [4]. Through a qualitative study of 117 teachers, Schroeder et al. [89] showed that teachers primarily utilize Pinterest to look for educational resources according to their classroom needs. They surveyed two types of teachers: preservice teachers (PSTs) and in-service

[1] The number of Pinterest-driven studies is smaller than Twitter and Facebook.

teachers (ISTs). Although both types sought specific instructional materials on Pinterest, the PTSs were more interested in "cute" and "fun" materials since, according to [89], PSTs have more time to implement these resources. Interestingly, our analysis shows that the word "fun" to be among the top words used by teachers to describe their pins– See Sections 3.5.4 and 4.3.4. Torphy et al. [79] performed a thorough analysis of teacherpreneurial behaviors of teachers on Pinterest. They characterized the source of 140,287 resources (pins) curated by 197 teachers on Pinterest. Their findings indicated that educational blogs were the predominant source of resources. Also, teacher-to-teacher market websites (notably *teacherspayteachers.com*) constituted a considerable portion of pins' sources. Our findings in Chapter 4 are in line with that of [79], where we also discovered that the predominant source of pins in our dataset is educational websites. Moreover, they found out that a significant portion of pins (82.8%) are monetized. This and the widespread diffusion of educational resources curated by teacherpreneurial educators signify that we face a new and decentralized open market of educational resources, which Torphy and Drake [4] named the "Fifth Estate within the digital age". Hu et al. [78] examined teachers' curation mechanism of mathematical resources on Pinterest. They characterized the mathematical pins and showed that they usually have low cognitive demand (difficulty). Sawyer et al. [90] found similar results regarding the cognitive demand levels of mathematical pins on Pinterest. Additionally, their work illustrated that socialized knowledge communities formed by mathematics teachers assist them in locating resources relevant to teaching mathematics. In our previous study [11], we also characterized mathematical resources shared by elementary school teachers on Pinterest and further proposed a method to predict the cognitive demand of resources. Hu et al. [26] performed an interesting analysis of how mathematical resources are curated. They identified three types of curation, namely self-directed, incidental, and socialized. One of the crucial features of this study is that they shed light on how online educational resources acquired from Pinterest are enacted in the classroom. Liu et al. [45] examined the diffusion of educational resources on Pinterest. They collected the Pinterest resource curation process for 34 early career teachers (ECTs) from three Midwestern states. They only studied the diffusion of resources for an ECT and their colleagues on Pinterest, i.e., those who

work with the ECT in the same school and had been nominated by them as close colleagues. Their results indicated that Pinterest act as a bridge between weakly connected teachers within the same school. Similarly, in Chapter 5, we analyze the diffusion of teacher-curated resources on Pinterest.

In conclusion, similar to Twitter and Facebook, teachers leverage Pinterest for professional purposes, which has improved their teaching according to the reviewed studies. However, unlike Twitter and Facebook, Pinterest is perceived as a social curation platform rather than a direct means for teacher-teacher or teacher-student communications. Moreover, perhaps the less politicized and polarized nature of Pinterest has contributed to its widespread usage by teachers. The interested reader can refer to [4, 28, 34, 46, 91, 92] for more details about other teachers in social media studies, especially those concerned with Pinterest. Eventually, it is worth mentioning our previous study [10] where we offered a roadmap on how to incorporate online social media in educational research, especially teachers in social media.

2.2 Data Collection

In this part, we explain the data collection process—first, describing the Pinterest data we acquired for each user. Then, we discuss the surveyed teachers and finally describe the constructed network of teachers.

2.2.1 Pinterest Data Description

Within Pinterest, users may encounter a personalized newsfeed of resources from various topics such as *education* and *sports*. Figure 2.1 illustrates an example Pinterest newsfeed. Each resource in Pinterest is called a *pin*. As demonstrated in Figure 2.2, each pin includes several pieces of information, including image, description, title, source, domain, comments, and board. By clicking on the source's URL, one will be redirected to the original website where the pin comes from. A user saves a pin in a *board* which is essentially a user-created directory holding pins with a similar topic (e.g., "My Math Pins" in Figure 2.2). Figure 2.3 shows an example of a Pinterest user's page, including the curated boards.

Figure 2.1: An example of Pinterest newsfeed.

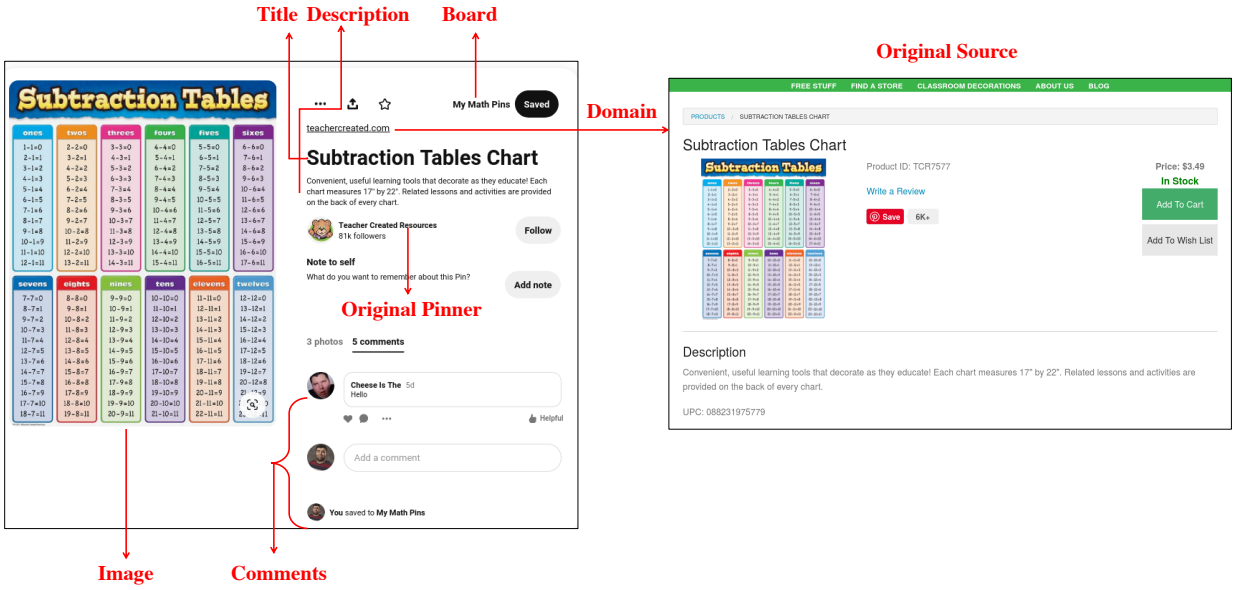


Figure 2.2: An example of a pin and its original source on the web.

We used API (application programming interface) provided by Pinterest and obtained data about Pinterest users, users' pins, and users' boards. Table 2.1 shows the pin-related information we retrieved for each pin. The API provided us with some crucial information. In particular, we have the topic of a pin which is a pre-defined topic (category) assigned by Pinterest. In our dataset,

Figure 2.3: An example of a Pinterest user's page.

there are 34 topics such as *education*, *sports*, *food_drink*. Moreover, similar to other social media platforms (e.g., Facebook) where a user's post is further shared, a pin can be a re-pin from another pin. Luckily, we also have access to the re-pin information, including the parent pin (i.e., the previous re-pin) and the original pin (i.e., the initial pin from which all re-pins have occurred). We can capture the diffusion process of the pins in the network from this information, which is used in our investigation in Chapter 5. Furthermore, Table 2.2 shows the fields related to a board. Through the board ID in Table 2.1 and ID in Table 2.2, one can find the corresponding board of a pin.

Table 2.1: Pin-related fields in our dataset.

Field	**Description**	**Example**
ID	A unique Pinterest-generated identifier of the pin	"715359134715179880"
title	User-generated title of the pin	*Subtraction Tables Chart*
description	User-generated description of the pin	Convenient, useful learning tools ..
domain	Domain of the original source	`www.teachercreated.com`
topic	A pre-defined topic assigned to a pin	*education*
created at	Time and date that the pin has been saved	*Sun, 23 Jul 2017 01:00:56*
image URL	URL of the image saved on Pinterest	*https://i.pinimg.com/...*
parent pin	ID of the successor pin	"320951910950109285"
original pin	ID of the original pin	"597430706814302354"
original pinner	ID of the original pinner of the pin	"144115394234286193"
board ID	A unique Pinterest-generated identifier of the board	"255227572571702183"

Table 2.2: Board-related fields in our dataset.

Field	Description	Example
ID	A unique Pinterest-generated identifier	"255227572571702183"
name	User-generated title of the board	*My Math Pins*
created at	Time and date of board creation	*Mon, 08 Aug 2016 07:22:18*
number of pins	Number of pins in the board	287

In addition to information about pins and boards in a user's account, we obtained information about each user, as shown in Table 2.3. In particular, we have the user-declared gender, which is used to distinguish male and female teachers in our investigation presented in Chapter 4. In addition, a Pinterest user can add a short description about themselves and a link to their website, both of which are available in our dataset and will be used in our analysis in subsequent Chapters.

Table 2.3: User-related fields in our dataset.

Field	Description	Example
username	A unique user-generated username	*karimihamid65*
name	First and last names of the user	Hamid Karimi
joined at	Time and date that the user has joined Pinterest	*Tue, 19 June 2016 23:17:11*
gender	User-declared gender *Male*, *Female*, or *Unspecified*	Male
self-description	A short self-introduction visible in a user's profile	*I am a 2nd grade teacher ...*
user website	User-declared website or blog	`www.hamidkarimi.com`
country	Country where the user has logged in during their last session	US

Remark. We used the Pinterest API to collect our data in November 2019. Unfortunately, since then, the Pinterest API has been the subject of a couple of changes, and consequently, some of the information we collected might not be provided any longer.

2.2.2 The Surveyed Teachers

Our efforts in this dissertation are part of an interdisciplinary project named *Teachers in Social Media Project*.[2] Founded by Dr. Kaitlin Torphy,[3] this project considers the intersection of the cloud to class, the nature of resources within virtual resource pools, and the implications for equity as educational spaces grow. Much of the work coming out of the Teachers in Social Media project concerns instructional and educational resources shared on Pinterest. Hence, as a part of this project, we have surveyed various American teachers whose information is used as the basis of our data collection in this dissertation. More specifically, the surveyed teachers used in this dissertation are sampled from three sets described as follows.

Set 1: SEMI (Study of Elementary Mathematics Instruction). This sample includes 340 ECTs (early career teachers) from 75 schools in 31 districts across four mid-west states, including Ohio, Michigan, Illinois, and Indiana. ECT is defined as a teacher in the first four years of their teaching career.

Set 2: OER (Open Educational Resources). 100 Michigan teachers were sampled from two rural pilot districts utilizing open educational resources. Teachers were identified and sampled across K-12 grade levels and eight schools.

Set 3: Texas Teachers. Finally, we selected a random sample of 100 Texas teachers from a non-CCSS (common core state standard) state. Teachers are from 16 schools across 16 distinct districts.

In total, we have 540 teachers across five states, 48 districts, and 99 schools. Figure 2.5 shows the number of teachers in each of the five states. Among the surveyed teachers, 428 are females, 13 males, and 99 unspecified. Figure 2.4 shows the distribution of the grade levels for the surveyed teachers. For a teacher teaching multiple grades, we consider the highest grade they teach.[4] More than 84% of teachers are teaching grades K to 6.

[2] `https://www.teachersinsocialmedia.com/`

[3] `https://torphyka.wixsite.com/kaitlintorphy`

[4] Twelves teachers were teaching multiple grade levels.

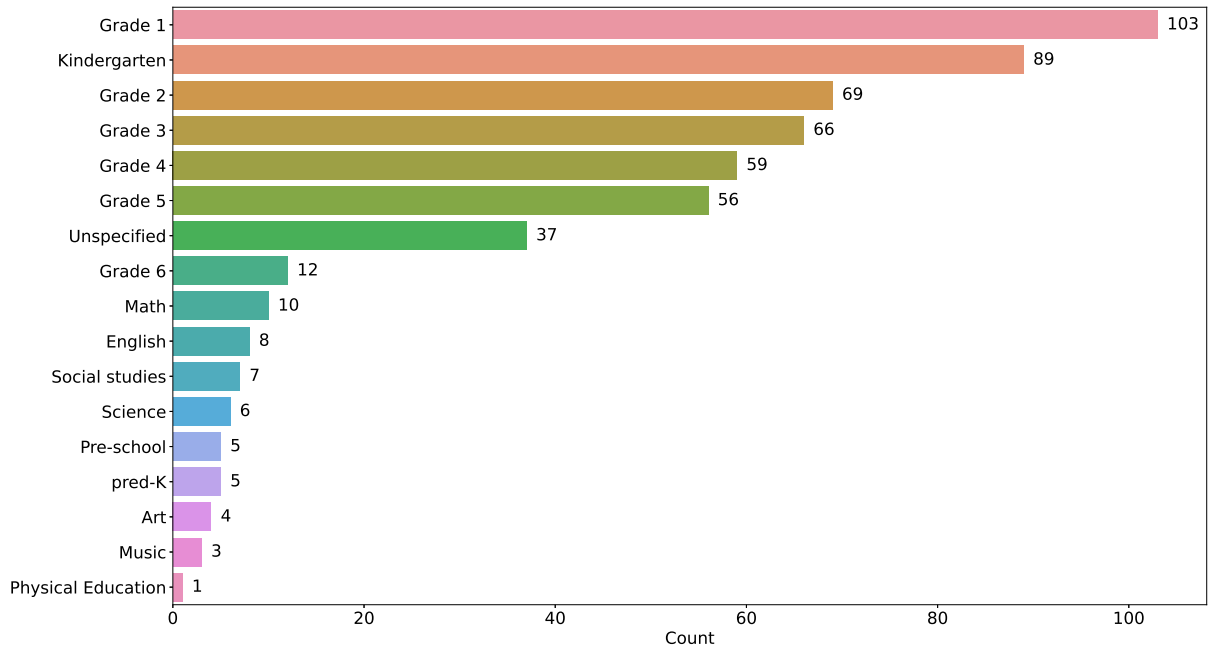


Figure 2.4: The distribution of grade levels of the surveyed teachers.

2.2.3 Network of Teachers

In addition to the data of the surveyed teachers, we acquired the data of all their online friends, i.e., their *followers* and *followees*. A user's follower is another user who follows that user, and similarly, a followee of a user is whom the user follows. Then, we constructed the network (graph) between all users. More formally, let $G = (V, E)$ represent our directed Pinterest network where V denotes the set of *nodes* (i.e., Pinterest users) and E denotes the set of *edges* (connections). Here, an edge $e : (u, v)$ indicates that user u follows user v. Some of the statistics of the network are shown in Table 2.4. Our network has 83,768 users and millions of edges. To the best of our knowledge, this is the largest network of teachers on social media. Furthermore, while the surveyed teachers reside in five U.S. states, by including their online friends, we have ended up with a single network (a connected graph).[5] This indicates the presence of small-world property where on average, two nodes (users) have a small distance from each other [93].

Figure 2.6 demonstrates the complementary cumulative distribution (CCDF) of degree distri-

[5]Technically, our network is weakly connected while its undirected version is strongly connected.

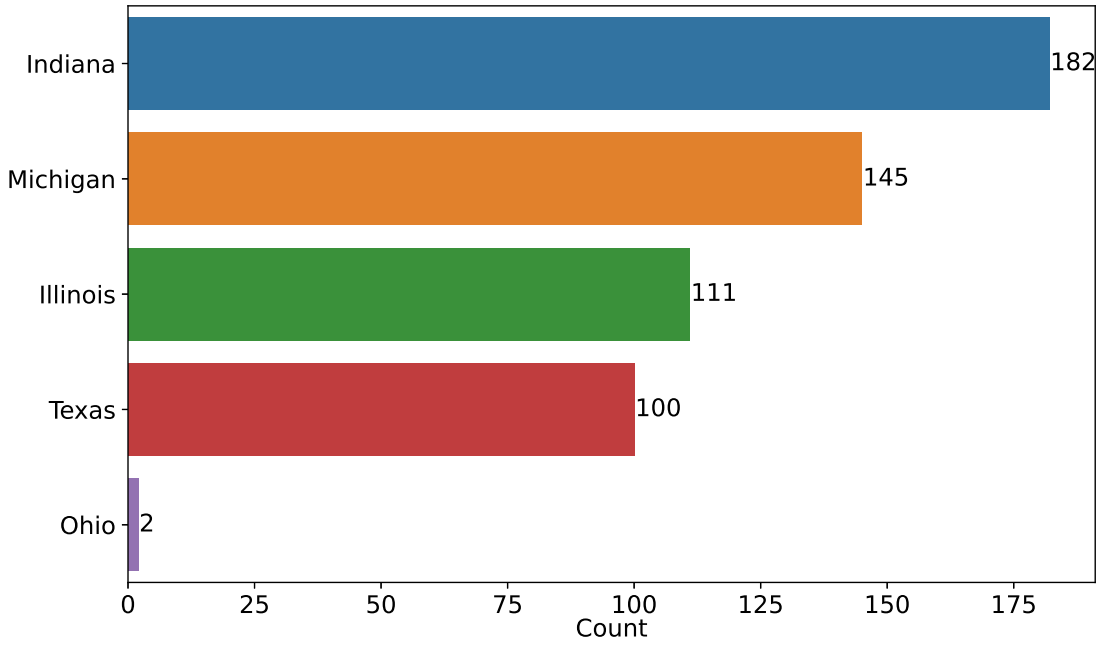


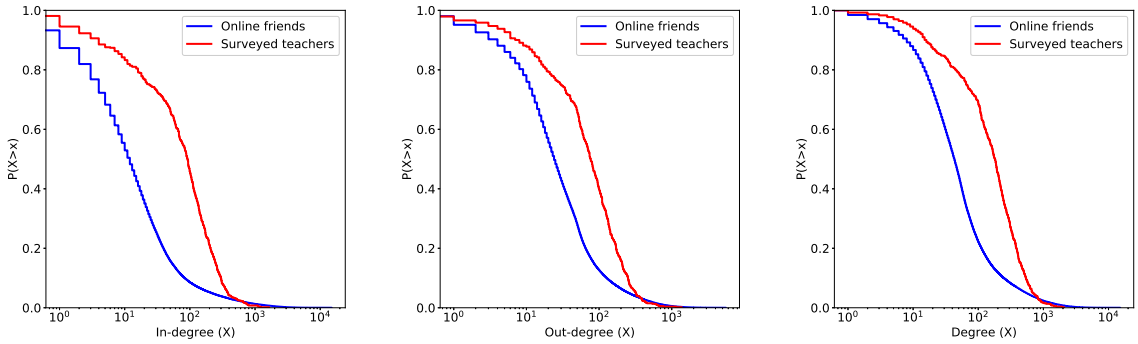
Figure 2.5: The number of surveyed teachers across five U.S. states.

Table 2.4: Some of the statistics of the Pinterest network.

The surveyed teachers	540
Followers+Followees (friends)	83,228
Total nodes (users)	83,768
Total edges (connections)	5,868,122
Average degree	131.58
Average in-degree	65.79
Average out-degree	65.79

butions for the surveyed teachers and all nodes where in-degree, out-degree, degree (in & out) distributions are shown in Figures 2.6a, 2.6b, 2.6c, respectively. Similar to other (online) social networks, the distributions follow a power-law distribution [94] where most of the nodes have a small (in/out)-degree and a tiny percentage have high degrees. Also, it seems the surveyed teachers have followed and are being followed more than their online friends. Figure 2.7 demonstrates the CCDF of the number of pins for the surveyed teachers and their online friends. The number of pins follows a power-law distribution.

Remark. Note that an online friend of a surveyed teacher can be a teacher or non-teacher.



(a) In-degree (b) Out-degree (c) Degree (in+out)

Figure 2.6: The CCDF of the degrees for the surveyed teachers and their online friends . x-axes are in log scale.

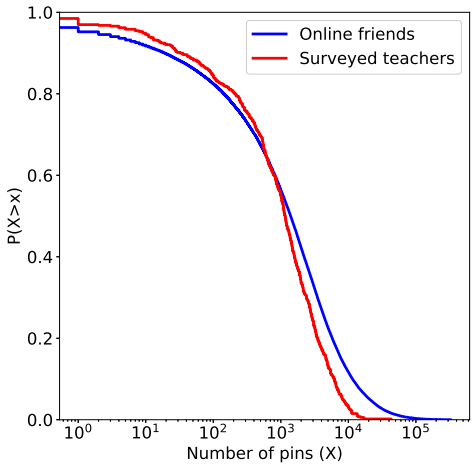


Figure 2.7: The CCDF of the number of pins for the surveyed teachers and their online friends. x-axis is in log scale.

We use this fact in the next section to identify more teachers among online friends of current the surveyed teachers.

CHAPTER 3

AUTOMATIC TEACHER IDENTIFICATION

3.1 Introduction

In Chapter 1, we discussed the importance of teachers in social media and its significant impact on education. In response to this importance, there have been many studies investigating teachers in social media in the past few years, a review of which was presented in Chapter 2. These studies have made significant progress in illuminating the potential of teachers in social media and the extra benefit it brings to education. Nevertheless, they suffer from a major limitation: they base their analysis on a limited number of surveyed/interviewed teachers. More specifically, they survey/interview a small number of teachers, and then they acquire their online social media data, i.e., a bottom-up data collection from surveyed/interviewed teachers to their online data. This causes two drawbacks. First, the study's outcome may not be statistically significant as the number of teachers is small. Second, they cannot harness the power of modern data-driven machine learning algorithms in their analysis since these algorithms usually require a sufficient amount of data. Hence, having an efficient mechanism to identify more teachers is crucial to advance the research on teachers in social media. An immediate option is to sample more teachers via surveys or interviews. However, surveying/interviewing is usually labor-intensive, costly, and non-scalable. Thus, we need a method that can identify teachers in social media *automatically*. Essentially, we need to have a binary *classifier* that for a given user reliably predicts whether they are a teacher or not.

While being very beneficial, the automatic identification of teachers in social media faces a significant technical challenge. For a binary classifier, we need to have training data from both teachers and non-teachers so we can train a supervised learning model classifying users to teachers and non-teachers. However, we do not have access to non-teacher users in practice since only a small set of surveyed/interviewed teachers are available. Given this, we cannot simply set up a

supervised model. To solve this challenge, we formulate the automatic teacher identification as a positive unlabelled (PU) learning task. As far as our dataset is concerned, positive samples are the surveyed teachers described in Section 2.2.2, and unlabelled samples are other users connected to the surveyed teachers[1], i.e., they follow or are followed by the surveyed teachers as described in Section 2.2.3. Note that an unlabelled user can be either a teacher or a non-teacher. PU learning has gained popularity in machine learning literature as its setting arises naturally in many applications such as automatic diagnosis [95], marketing [96], remote sensing [97]. Likewise, our conceptualization of the automatic teacher identification problem as a PU learning task reflects this problem's most practical and natural setting. Specifically, we train an efficient teacher identification classifier from a limited number of teachers and many readily available unlabelled users. Next, we briefly explain our proposed approach.

This chapter proposes a PU framework to identify teachers on Pinterest automatically. We call our framework `PUTeacher`, which entails three components. In the *Unsupervised Representation Learning* component, we develop a deep neural autoencoder to learn a salient and compact representation for both positive (teacher) and unlabelled users. This component's main advantage is that it can encode the underlying semantic of the entire training data (positive plus unlabelled) without requiring ground truth labels. In the second component, called *Automatic User Labeling*, we propose a method that utilizes the learned representations from the first component and then automatically marks unlabelled samples as potentially non-teachers or teachers, i.e., in the PU learning terminology, finding *reliable negative* samples and additional *reliable positive* samples. In the last component, *User Classification*, we utilize the marked samples as well as original positive samples and perform a binary classification to predict the class of unlabelled users (teacher or non-teacher). We conduct extensive experiments and show the effectiveness of our proposed framework. In summary, our contributions are as follow:

- We formulate the teacher identification problem as a PU learning task reflecting this problem's realistic and practical setting.

[1] In this chapter, we use terms *sample* and *user* interchangeably.

- We propose an effective PU learning method for teacher identification, which can reliably identify thousands of teachers on Pinterest.

The rest of this chapter is organized as follows. First, in Section 3.2, we formally define the problem, followed by presenting `PUTeacher` in Section 3.3. Next, Section 3.4 includes the experimental results and discussions. Finally, in Section 3.5, we perform an extensive resiliency analysis of `PUTeacher` to ensure its working.

3.2 Problem Statement

Let $\mathcal{X} = \{x_1, x_2, \cdots, x_n\}$ represent a dataset of n online social media users where $x_i \in \mathbb{R}^d$ and d is the dimension of feature inputs representing x_i. Suppose the random variable $y = \{+1, -1\}$ represents the label of a sample in $\mathcal{X}$ where $+1$ indicates the sample is a teacher and -1 otherwise. Further, let $\mathcal{X}$ consist of two distinct sets of l positively labelled users and $n-l$ unlabelled users, i.e., $\mathcal{X}^p = \{x_1, x_2, \cdots, x_l\}$ and $\mathcal{X}^u = \{x_{l+1}, x_{l+2}, \cdots, x_n\}$, respectively. For convenience, let $|\mathcal{X}^u| = m$ i.e., $m = n - l$. Following the PU learning setting, $\forall x_i \in \mathcal{X}^p$, $y_i = +1$ while y_j for $x_j \in \mathcal{X}^u$ is unknown.

Now, given the notations listed above, we seek to utilize $\mathcal{X}^p$ and $\mathcal{X}^u$ to learn a model $f_\theta(x)$ having parameters θ such that it can predict the label for an unseen user in $\mathcal{X}^u$.

Table 3.1: PU learning terminology and its equivalents in our automatic teacher identification task.

PU learning	**Automatic teacher identification task**
Positive samples	Teachers
Negative samples	Non-teachers
Unlabelled samples	Followers and followees of the survyed teachers
Reliable positive samples	Automatically marked teachers
Reliable negative samples	Automatically marked non-teachers
Original positive samples	The surveyed teachers

Table 3.1 demonstrates a mapping between PU learning terminology and its equivalent in our automatic teacher identification task.

3.3 The Proposed Framework (PUTeacher)

An overview of the proposed framework, PUTeacher, is demonstrated in Figure 3.1. Our framework falls into the category of two-stage PU learning, in which we first try to identify reliable negative and positive samples and then utilize them to train a supervised learning model. An important assumption in the two-stage PU learning is the *smoothness* property, which asserts that if two samples x_i and x_j are *similar*, the probabilities $P(y = +1|x_i)$ and $P(y = +1|x_j)$ are close [37]. The smoothness property has been leveraged in various two-stage PU learning algorithms [98, 99, 100]. Assuming this property, we can identify reliable negative samples as those far away from all labeled samples. The key to this assumption is to determine the similarity between the two samples. To this end, we need to have an effective method to encode the input data, which is what the *Unsupervised Representation Learning* component is tasked for. Then, in the *Automatic User Labeling* component, we utilize these representations, and based on the smoothness assumption, we propose a novel method to identify reliable negative and positive samples from the unlabelled data. In other words, *Automatic User Labeling* component automatically marks penitential non-teachers (reliable negative samples) and potential teachers (reliable positive samples). These two components will be presented in Sections 3.3.1 and 3.3.2, respectively.

Using the identified reliable negative and positive samples and the original positive labeled samples (i.e., the surveyed teachers), we can transform the PU learning into a supervised learning task, which is what we will perform in the *User Classification* component. In the two-stage PU learning, this supervised learning is predicated on another important assumption known as *separability*, under which it is assumed that two classes (i.e., teachers and non-teachers) are naturally separated [37]. In other words, theoretically, there should exist a 'perfect' classifier that distinguishes positive samples from negative ones. In Section 3.4, we empirically demonstrate that both the *smoothness* and *separability* assumptions hold, and hence our proposed framework is justified.

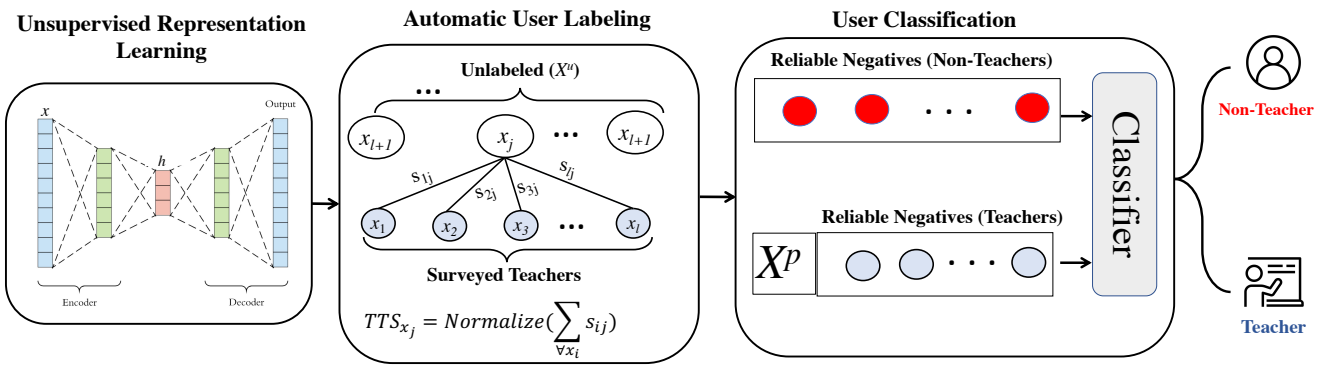


Figure 3.1: An illustration of the proposed method for automatic teacher identification (PUTeacher).

3.3.1 Unsupervised Representation Learning

We need to extract salient and semantically informative features of the input data. Such features are crucial for subsequent components of our framework. Unfortunately, we cannot use the labels to represent the input data since we only have labels for a single class (i.e., teachers), and most of the data is simply unlabelled. Hence, we train an autoencoder model to extract meaningful features from the input data without using the labels. The autoencoder takes an input sample, encodes it into a lower-dimensional hidden representation (embedding), and eventually decodes the hidden representation to an output, aiming to reconstruct the input. By doing so, we force the autoencoder to learn a condensed hidden representation that retains meaningful information about the input data. Moreover, another benefit of autoencoders is reducing the dimensionality while reconstructing the input sample. Autoencoders are widely used in representation learning and have shown tremendous performance in various applications [101, 102].

Let $E_\Theta(.)$ and $D_\Omega(.)$ denote the encoder and decoder with parameters Θ and Ω, respectively. Then, we optimize the following loss function:

$$\mathcal{L} = \underset{\Theta,\Omega}{\operatorname{argmin}} \frac{1}{m} \sum_{\forall x \in \mathcal{X}^u} \|x - D_\Omega(E_\Theta(x)\|_2^2 \tag{3.1}$$

where $\|.\|$ denotes L2 norm (euclidean distance) of a vector. For convenience, let $h_i = E_\Theta(x_i)$ denote the hidden representation for a sample x_i. Note that, we train the autoencoder on the

unlabelled samples.

3.3.2 Automatic User Labeling

Based on the smoothness assumption, *similar* samples have close probabilities of being positive. Hence, we leverage this assumption and attempt to mark samples in the unlabelled data. Let s_{ij} denote the similarity between user $x_j \in \mathcal{X}^u$ and $x_i \in \mathcal{X}^p$:

$$s_{ij} = e^{(-\left\|h_i - h_j\right\|_2^2 / 2\sigma^2)} \tag{3.2}$$

s_{ij} is essentially RBF (radial basis function) kernel determining the similarity between two feature vectors.[2] Then, we sum up the similarity between x_j and all positive samples (denoted as S_{x_j}):

$$S_{x_j} = \sum_{\forall x_i \in \mathcal{X}^p} s_{ij} \qquad \forall x_j \in \mathcal{X}^u \tag{3.3}$$

where σ is a free hyperparameter. We normalize the similarity score in Eq. 3.3 to range $[0, 1]$ and get *Teacher Tendency Score* (TTS_{x_j}) for each unlabelled user:

$$TTS_{x_j} = \frac{S_{x_j} - min(\{S_{x_k} : x_k \in \mathcal{X}^u\})}{max(\{S_{x_k} : x_k \in \mathcal{X}^u\}) - min(\{S_{x_k} : x_k \in \mathcal{X}^u\})} \quad \forall x_j \in \mathcal{X}^u \tag{3.4}$$

TTS_{x_j} encodes the similarity between an unlabelled user and all positive users i.e., the surveyed teachers. Now, based on the smoothness assumption, the closer TTS_{x_j} is to 1, the higher chance for x_j to be a positive user (a teacher). Similarly, the closer TTS_{x_j} is to 0, the higher chance for x_j to be a negative user (a non-teacher). Hence, let $\tilde{y}_{x_j}$ denote the automatic label (the pseudo-label) assigned to an unlabelled user:

$$\tilde{y}_{x_j} = \begin{cases} +1 & \text{if } TTS_{x_j} > \alpha \\ -1 & \text{if } TTS_{x_j} < \beta \end{cases} \tag{3.5}$$

[2] We found RBF kernel performing better empirically, but one can utilize other similarity measures, e.g., cosine similarity.

where α and β are two hyperparameters controlling the level of sensitivity for automatically marked teachers and non-teachers, respectively. Usually, α needs to be close to 1 (e.g., 0.9) and β close to 0 (e.g., 0.05). In addition, by adjusting α and β, we can control the number of newly identified teachers and non-teachers, respectively.[3] Also, one might wonder that instead of the summation in Eq. 3.3, we could take the average of similarities and consider that as TTS_{x_j}. However, we found that the average of similarities for RBF is usually small, and thus flexibly selecting proper values for α and β is difficult.

3.3.3 User Classification

Using the *Automatic User Labeling*, we create two sets of reliable negative (non-teacher) and reliable positive (teacher) samples denoted as $\mathcal{R}^n = \{x_i \in \mathcal{X}^u, \tilde{y}_i = -1\}$ and $\mathcal{R}^p = \{x_i \in \mathcal{X}^u, \tilde{y}_i = +1\}$, respectively. For the classification, we develop a deep feedforward neural network $f_\theta(.)$ trained on representations of samples in $\mathcal{R}^n$ and $\mathcal{R}^p$. The output of $f_\theta(.)$ is the probability distribution of being teacher and non-teacher i.e., $f_\theta(h_i) = [y_t^i, y_{nt}^i]$, where y_t^i (y_i^{nt}) is the probability of being a teacher and a non-teacher, respectively. Note that $0 \leq y_i^t, y_i^{nt} \leq 1$ and $y_i^t + y_i^{nt} = 1$. Then, we use the backpropagation and the cross entropy loss function (Eq. 3.6) to optimize the neural network.

$$\mathcal{L}_c = -\sum_{\forall x_i} \tilde{y}_i \times log(y_i^t) + (1 - \tilde{y}_i) \times log(y_i^{nt}) \tag{3.6}$$

Note that, the input to neural network is the representations learned in the first component of `PUTeacher` described in Section 3.3.1.

3.4 Experiments

To verify the effectiveness of the proposed method, we conduct some experiments. In Section 3.4.1, we explain the experimental settings. In Section 3.4.2, we verify the data assumptions

[3] One can set $\beta = 1 - \alpha$. However, we prefer to keep α and β independent for flexibility purposes. However, their ranges should not overlap.

discussed in Section 3.3 i.e., the smoothness and separability assumptions. Ultimately, we compare the performance of PUTeacher with several baselines in Section 3.4.5.

3.4.1 Experimental Settings

In this part, we present the experimental settings, including the dataset, input features, and hyperparameter tuning.

3.4.1.1 Dataset

Our users include 540 surveyed teachers and their followers and followees as described in Section 2.2.1. In addition to the surveyed teachers, we manually annotated 3,058 teachers and 2,079 non-teachers. The annotation procedure is described in the Appendix. Hence, the number of unlabelled users is 78,091 i.e., $|\mathcal{X}^n| = 78,091$. Table 3.2 demonstrates the specific data used to train, evaluate, and test PUTeacher's components. To train the *Unsupervised Representation Learning* component described in Section 3.3.1, we only used the unlabelled users. The validation set of this component, utilized for hyperparameter tuning, is 5-fold cross-validation on its training set. To train the *User Classification* component, we used 3038 teachers and 3038 non-teachers. The teachers consist of 1519 annotated teachers and 1519 automatically identified teachers acquired from the *Automatic User Labeling*. All 3038 non-teachers are automatically identified. To tune this component's hyperparameter, we used 3-fold cross-validation on its training set. Eventually, to evaluate the performance of PUTeacher, we created a test set for the *User Classification* component. This set consists of 1539 annotated teachers, 540 surveyed teachers, and 2079 annotated non-teachers, i.e., in total, 2079 teachers (1539 + 540) and 2079 non-teachers. Note that teacher and non-teacher samples used in the final evaluation of PUTeacher have ground truth labels. This helps reinforce the reliability of the test set. Moreover, using the entire 540 surveyed teachers as part of the test set further strengthens its reliability. Note that the *Automatic User Labeling* component utilizes the unlabelled users and the surveyed teachers to mark users, and thus, unlike the other two components, it does not entail any learning process.

Table 3.2: Samples used in training, evaluating, and testing PUTeacher's components. *Ann*: Annotated, *Auto*: Automatically identified, *Surv*: Surveyed

Component	Split	Samples
Unsupervised Representation Learning	Train	78,091 unlabelled users
Unsupervised Representation Learning	Validation	5-fold cross validation
Unsupervised Representation Learning	Test	–
User Classification	Train	3038 teachers: 1519 *Ann* + 1519 *Auto* 3038 *Auto* non-teachers
User Classification	Validation	3-fold cross validation
User Classification	Test	2079 teachers: 1539 *Ann* + 540 *Surv* 2079 *Ann* non-teachers

3.4.1.2 Input Features

We used the following input features to represent each user.

Topic. As mentioned in Section 2.2.1, each pin belongs to one of 34 general topics (categories) pre-defined by Pinterest, e.g., *food*, *fashion*, *education*. Hence, this feature vector has 34 values corresponding to 34 existing topics. Each element of the vector is the number of the user's pins in a topic divided by the total number of their pins (i.e., the input is normalized).

Domain. To represent the domain features, we extracted the top 200 domains in the entire dataset. Then for each user, we created a vector of size 200. Each element of the vector holds the number of pins whose domain is the corresponding domain in the top 200 domains. To normalize the vector, we divide it by the total number of the user's pins.

Description. We extracted the top 100 words used in the description of pins shared by a user. Then we represented each word using a pre-trained word embedding model known as fastText, which includes one million word vectors trained on Wikipedia 2017 [103] (a vector of size 300 represents each word). We took the average of word embeddings for the top 100 words. Moreover, we included an weighted average of the top 100 word embeddings where weights are frequencies of the words in the pin descriptions of the user. Hence, the dimension of this feature vector is 600. Note that, before acquiring word ebmeddings, we used NLTK package [104] and pre-processed pin

descriptions, e.g., removed punctuations and stopwords (e.g., '!', 'the'), stemmed the tokens (e.g., 'education' to 'educ').

3.4.1.3 Hyperparameter Tuning

The encoder and decoder of the *Unsupervised Representation Learning* are two-layer, fully connected neural networks. There is one hyperparameter associated with this autoencoder, namely the dimension of the hidden representation. To tune this hyperparameter, we performed 5-fold cross-validation on the unlabelled samples and evaluated the dimensions $\{10, 20, 30, 40, 50\}$. The dimension size 20 yielded the best performance based on the L2 loss in Eq. 3.1. The *Automatic User Labeling* has two crucial hyperparameters, namely α and β in Eq. 3.5. We did not tune these two hyperparameters since we treat them as flexible variables to be set by the practitioner of `PUTeacher`. Despite this, the selection criteria for these two hyperparameters are that α should be close to 1 and β close to 0. Hence, we set α to 0.9 and β to 0.05. The low value of β ensures identifying reliable non-teacher users, which are crucial for the *User Classification* component. In Section 3.5.3, we will explain how different values of α and β compete and affect the performance of `PUTeacher`. We also set $\sigma = 12$ in Eq. 3.3. Eventually, for the *User Classification* component, we developed a two-layer fully neural network connected. To tune the dimension of the hidden layer of this network, we performed 3-fold cross-validation on its training set.

3.4.2 Verification of Two-stage PU Learning Assumptions

As mentioned before, two crucial assumptions in the two-stage PU learning are separability and smoothness. To verify these assumptions, we trained a supervised multi-layer neural network on the input features of 3,598 labeled teachers (3,058 annotated + 540 surveyed) and 2,079 non-teachers. We randomly selected 70% of the data for the training and 30% for the test. We call this classifier, *data-assumption-verification-classifier*. Note that *data-assumption-verification-classifier* is merely used for the verification of the above assumptions and is distinct from our proposed automatic teacher identification approach, i.e., `PUTeacher`.

3.4.3 Separability

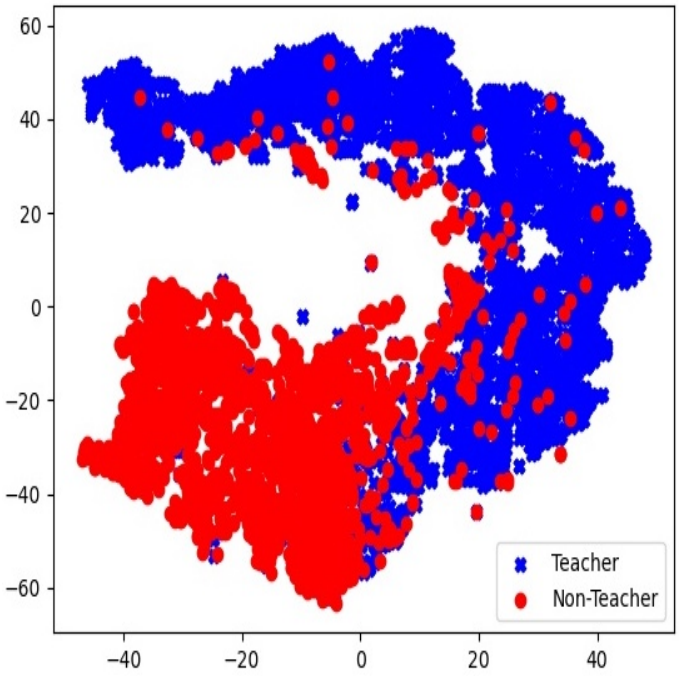


Figure 3.2: t-SNE visualization of teacher and non-teacher embeddings for the verification of the separability assumption.

Data-assumption-verification-classifier achieved a very high performance of 0.95 for AUC (Area Under Curve) and 0.92 for F1-score. Also, we used t-SNE (t-distributed stochastic neighbor embedding) [105] and visualized the learned representations for test teachers and non-teachers, as demonstrated in Figure 3.2. As it can be observed, the two classes are perfectly separated. Hence, we can conclude that the data separability assumption holds for our dataset.

The question is, what does this separability mean in the context of teachers in social media? It means that teachers are using online social media (here Pinterest) in such a way that we can distinguish them from other users. More specifically, as far as topics, domains, and descriptions of pins are concerned, teachers' online activity makes them *identifiable* from other Pinterest users. This is in line with previous studies [77, 23, 24, 4, 34] showing that teachers leverage social media for their specific professional needs. However, our findings corroborate this in a large-scale data-driven manner.

3.4.4 Smoothness

To verify the smoothness property, we calculate the Pearson correlation between two variables, namely $d(F_{x_i}, F_{x_j})$ and $|s^t_{x_i} - s^t_{x_j}|$ for all pairwise x_i and x_j in the test set. F_{x_i} denotes the final embedding from *data-assumption-verification-classifier* (i.e., the output before the last linear layer), $s^t_{x_i}$ denotes the score of being a teacher (i.e., the output of the last linear layer before the softmax), and d is the Euclidean distance. The correlation is 0.88 with a p-value of 1.0×10^{-8}. This high positive correlation between these two variables indicates a high degree of smoothness since samples that are mapped to the same regions (thus having a small distance in the embedding space) belong to the same class (thus a small difference in scores). Finally, the high correlation between embedding level distances and teacher score differences is visually demonstrated in Figure 3.3 by drawing a fitted linear regression between these two variables.

Based on the above analysis regarding the verification of the separability and smoothness assumptions, we can conclude that the two-stage PU learning is suitable for the automatic teacher identification problem, and thus, the design of `PUTeacher` is justified.

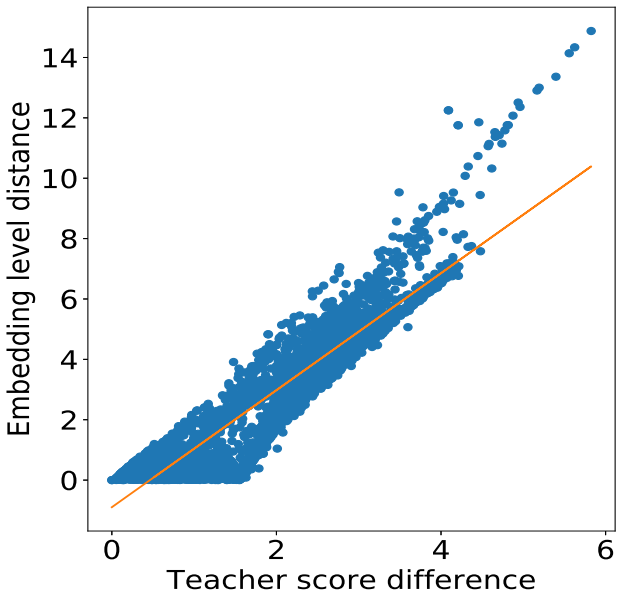


Figure 3.3: The fitted regression line between pairwise embedding distances and teacher scores differences for the verification of the smoothness assumption.

3.4.5 Baseline Comparison

We compare the performance of PUTeacher with the following baseline methods.

- ElkaNoto [97]. This method is based on training a non-traditional classifier to predict whether a sample is labeled. Then, it utilizes the SCAR (selected completely at random) assumption and adjusts the classifier to a traditional one, i.e., predicting the label of a sample.

- WElkaNoto [97]. It is similar to Elkanoto, except it assigns weights to training samples.

- BaggingPU [95]. This method is based on bootstrap aggregation. It repeatedly trains classifiers to identify positive examples in the unlabelled set and eventually takes the average of these classifiers to distinguish positive samples from negative ones.

- nPU [106]. This method proposes a convex formulation while canceling the bias introduced when one attempts to separate unlabelled data from positive data.

- nnPU [107]. The authors of this paper proposed a method to minimize the risk while reducing the bias and overfitting of flexible models in unbiased risk estimation.

- ProbTagging [108]. This method is a recent two-stage PU learning approach that identifies reliable negatives and positives and trains multiple ordinary supervised classifiers. Their tagging process is based on the k-Nearest Neighbor (kNN) in the input space. Due to the high-dimensionality of the input space in our dataset, however, kNN was not very effective. Hence, we trained this method based on representations learned in the first component of PUTeacher.

- Supervised. For this method, we trained a supervised neural network classifier. Its test set is the same with PUTeacher as described in Section 3.4.1.1. For teachers in its training set, we used the same 1,519 annotated teachers used in the *User Classification* component of PUTeacher. Additionally, we selected 1,519 users from the unlabelled set whose unsupervised representations learned from the first component of PUTeacher is

closest to the representations of annotated non-teachers. Evidently, this method is not a PU learning approach as it uses both labeled teachers and non-teachers. However, it acts as a yardstick for other methods and informs us of the problem's upper bound performance.

For the baselines from literature, we used their publicly available codes. We tuned all methods' hyperparameters based on 3-fold cross-validation. Each method, including `PUTeacher`, was run five times, and the average performance on the test set is reported. The performance metrics are AUC and F1-score. We implemented our method using the PyTorch package [109]. Table 3.3 shows the results. In addition to reporting the performance on the entire test set, we exclusively report the performance against the surveyed teachers in terms of the recall. The reason to do this is the following. Our model uses data from social media (Pinterest) to perform automatic teacher identification. Given this, we need to ensure that the model is generalizable to other types of stratified data, i.e., the surveyed teachers. We make the following observations based on these results.

Table 3.3: Comparing *PUTeacher* with baseline methods.

	Entire Test Set		The Surveyed Teachers
Method	**AUC**	**F1-score**	**Recall**
`Supervised`	0.95 ± 0.008	0.92 ± 0.004	0.96 ± 0.001
`ElkaNoto`	0.82 ± 0.006	0.80 ± 0.002	0.83 ± 0.005
`WElkaNoto`	0.79 ± 0.008	0.79 ± 0.007	0.86 ± 0.002
`BaggingPU`	0.90 ± 0.008	0.87 ± 0.004	0.90 ± 0.008
`nPU`	0.88 ± 0.01	0.87 ± 0.01	0.87 ± 0.01
`nnPU`	0.90 ± 0.01	0.89 ± 0.01	0.91 ± 0.03
`ProbTagging`	0.91 ± 0.002	0.90 ± 0.007	0.93 ± 0.007
`PUTeacher`	0.93 ± 0.001	0.91 ± 0.001	0.96 ± 0.003

`ElkaNoto` and `WElkaNoto` achieved low performance. This is primarily due to their simplistic dependence on a non-traditional classifier and then adjusting it based on the probability of a positive sample being labeled, which is hard to estimate in practice. `BaggingPU` alleviates this problem by training a bag of classifiers and thus has outperformed `ElkaNoto` and `WElkaNoto`. `nPU` and `nnPU` achieved a relatively good performance. However, their dependence on the class prior has made it hard for them to obtain high performance. `ProbTagging` achieved a good performance.

However, in addition to being costly, their employed k-NN is not suitable for high-dimensional data. PUTeacher outperformed all baselines, and its performance is very close to the supervised classifier. Regarding the surveyed teachers, we can observe that PUTeacher has achieved a very high recall, has outperformed all baselines methods, and is on par with the Supervised method. This indicates that PUTeacher is generalizable to other types of stratified data.

3.5 Resiliency Analysis

In the previous section, we demonstrated that PUTeacher offers excellent performance in classifying users as teachers and non-teachers. To further ensure the robustness of PUTeacher, in this section, we perform several important resiliency experiments. These experiments look into different aspects of PUTeacher (e.g., its input and output) and enable us to ascertain its resiliency. The resiliency analysis presented in this section is particularly crucial since identified teachers from PUTeacher are the basis of the studies in the following two chapters of the dissertation.

3.5.1 Input Perturbation

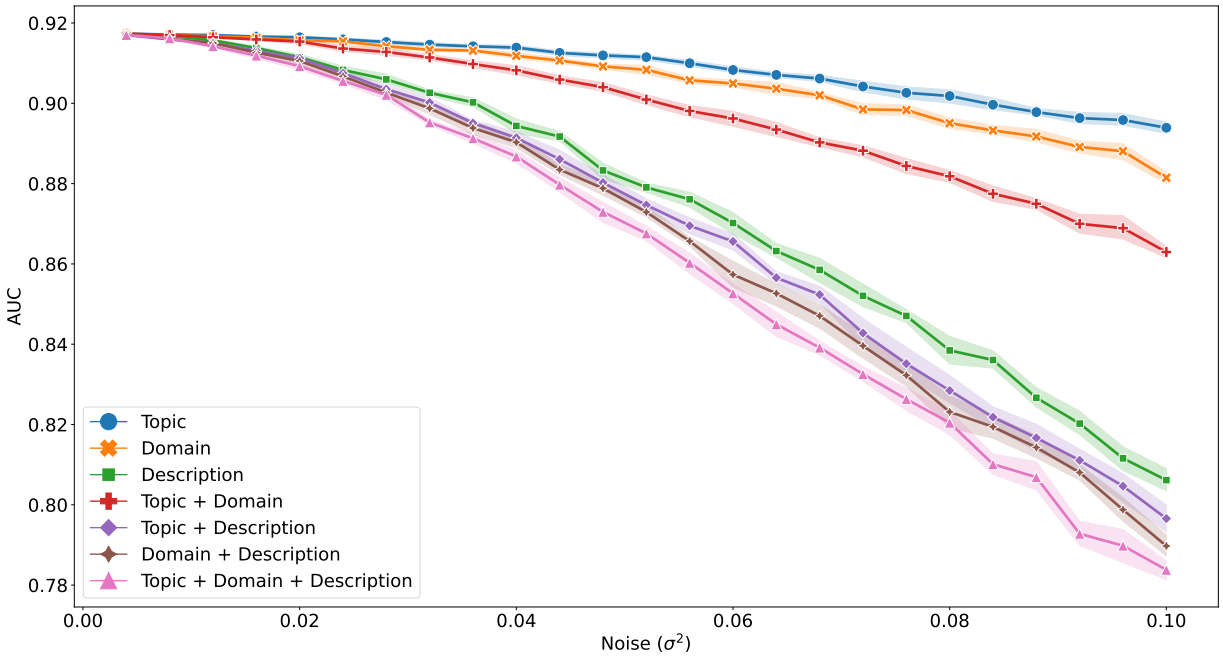


Figure 3.4: Perturbing the input features using the Gaussian noise.

The first analysis includes testing the robustness of PUTeacher against noise. To this end, we added noise to the input data and inspected the performance of the model. More specifically, we perturbed the three input features, i.e., topic, domain, and description, by adding the Gaussian noise from a normal distribution $\mathcal{N}(\mu, \sigma^2)$ where μ is the mean and σ^2 is the standard deviation of the distribution. For this analysis, we set $\mu = 0$ and considered σ^2 in range $[0.001, 0.1]$. That is, we changed the magnitude of the noise. Note that we only perturbed the input data of the test set since our goal is to determine the robustness of the trained PUTeacher when faced with predicting the class of unseen noisy instances. Given the uncontrolled and prone-to-noise nature of social media, encountering such instances is likely in practice. Figure 3.4 demonstrates the results where the x-axis is the amount of noise (σ^2), and the y-axis is the AUC on the perturbed test set. We ran each experiment ten times. Also, as shown in this figure, we considered all combinations of feature types. We make the following observations based on the results presented in Figure 3.4.

- Even when all features are perturbed and the noise is as high as 0.1, PUTeacher manages to deliver a good performance in terms of the AUC score. This indicates that our proposed model, to a large degree, is robust against the noise and can be used as a reliable model to identify teachers.

- Three cases wherein the descriptions have been kept intact are consistently above the other ones at each noise level. This shows that among feature types, the description is the most robust against the noise. The main reason is that while non-teachers can curate resources from similar domains and similar topics with teachers, the specific vocabulary used by teachers to describe their pins is somehow unique.

- The robustness of the topic and domain against noise exhibits a similar pattern. This means that these two feature types have some similarities. This seems logical since there are specific domains tied to particular topics, e.g., *teacherspayteachers.com* to *education*. Despite this, both feature types are essential since when both are perturbed, the performance has dropped significantly– See the red plot in Figure 3.4.

3.5.2 Imbalance in Number of Pins

As far as the number of curated pins is concerned, not all users on Pinterest are equally active. As was illustrated in Figure 2.7, the distribution of the number of pins follows a power-law distribution where most of the users have a small number of pins while a tiny portion of users has a massive number of pins. Similarly, the distribution of the number of pins for unlabelled users follows a power-law distribution as demonstrated in Figure 3.5. Furthermore, the first component of `PUTeacher` learns representations from this skewed set of users in terms of the number of pins. Given the importance of the first component of `PUTeacher`, we need to investigate whether the imbalance in the number of pins of unlabelled users influences the performance of `PUTeacher`?

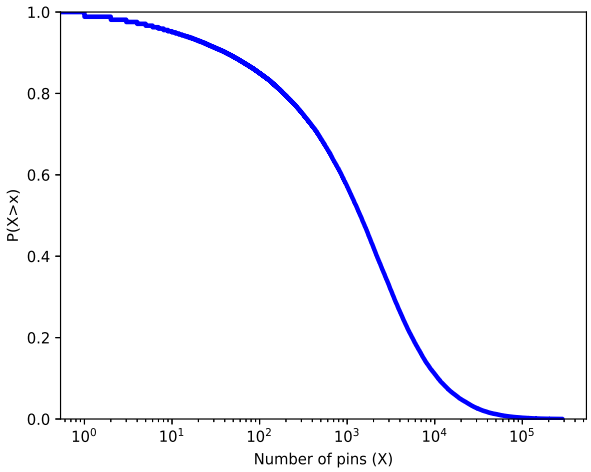


Figure 3.5: The CCDF of the number of pin for unlabelled users. x-axis is in log scale.

To answer this question, we trained four distinct versions of `PUTeacher`. In each version, we trained `PUTeacher` only on unlabelled users whose number of pins is in a certain range. These ranges include $[1, 500]$, $[501, 5000]$, $[5001, 20000]$, and $[20001, 287762]$[4], which cover 23319, 36749, 13253, and 3785 number of users, respectively. We call these ranges *low*, *medium*, *high*, and *very high*, respectively, signifying the number of pins they include. Note that the only difference in the four versions of `PUTeacher` is their unlabelled users. Other parts of the framework are kept intact. Figure 3.6 shows the ROC curves of these four models. According to this figure, we can

[4]The maximum number of pins for unlabelled users is 287,762.

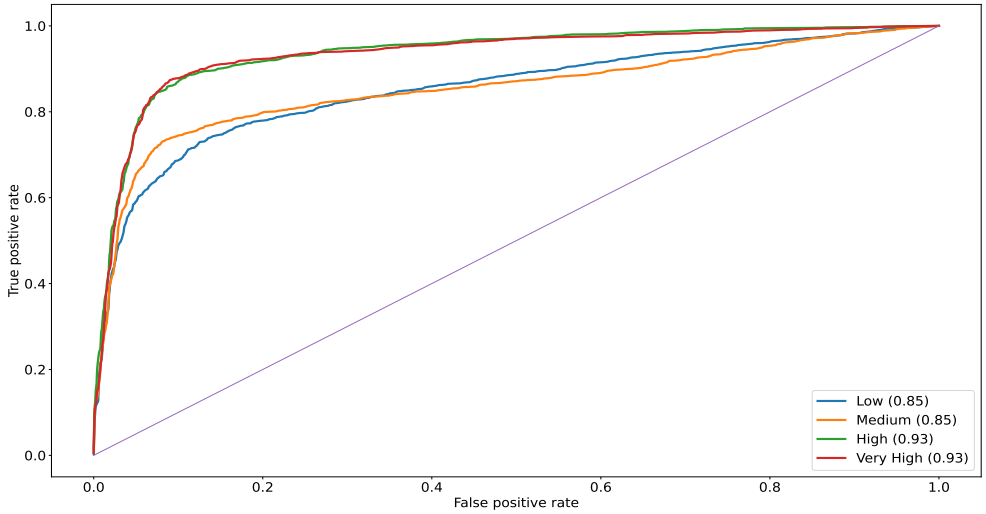
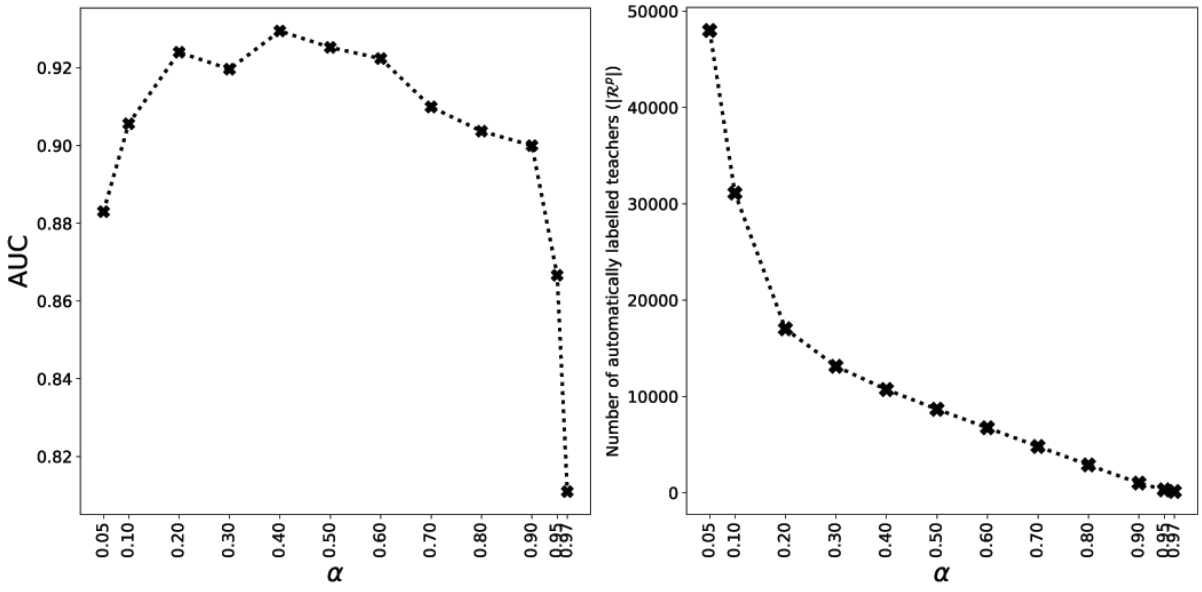


Figure 3.6: The ROC curves of training `PUTeacher` on four ranges of the number of pins. Numbers in the parentheses are AUC scores.

observe that training on unlabelled users in the *low* and *medium* ranges leads to decreasing the performance. In contracts, ranges *high* and *very high* deliver a better performance. Hence, we can state that users with a larger number of pins have a larger impact on the performance of `PUTeacher`.

Based on the above observation, we can assert that the only way that the number of pins causes a problem is when most unlabelled users have a low or medium number of pins. Nevertheless, this did not occur in our dataset since our collected unlabelled users consisted of a diverse set of users in terms of their number of pins. So the question is, what can one do in circumstances when most users have a relatively low number of pins, i.e., in ranges *low* and *medium* as described above? Here, we briefly mention two possible ways to address this and leave exploring more advanced approaches for the future. First, one can collect more unlabelled users. Note that unlabelled data from social media is significantly cheaper and easier to acquire than conducting surveys or annotating samples. For instance, using only 540 surveyed teachers, we easily collected thousands of their online friends as unlabelled users. Second, sample generation, e.g., using generative adversarial networks [110, 111], is another worthwhile direction. Using a generative method, for instance, one can synthesize pins for a user based on their pin attributes' distributions.

3.5.3 Teacher Filtering Parameter Analysis



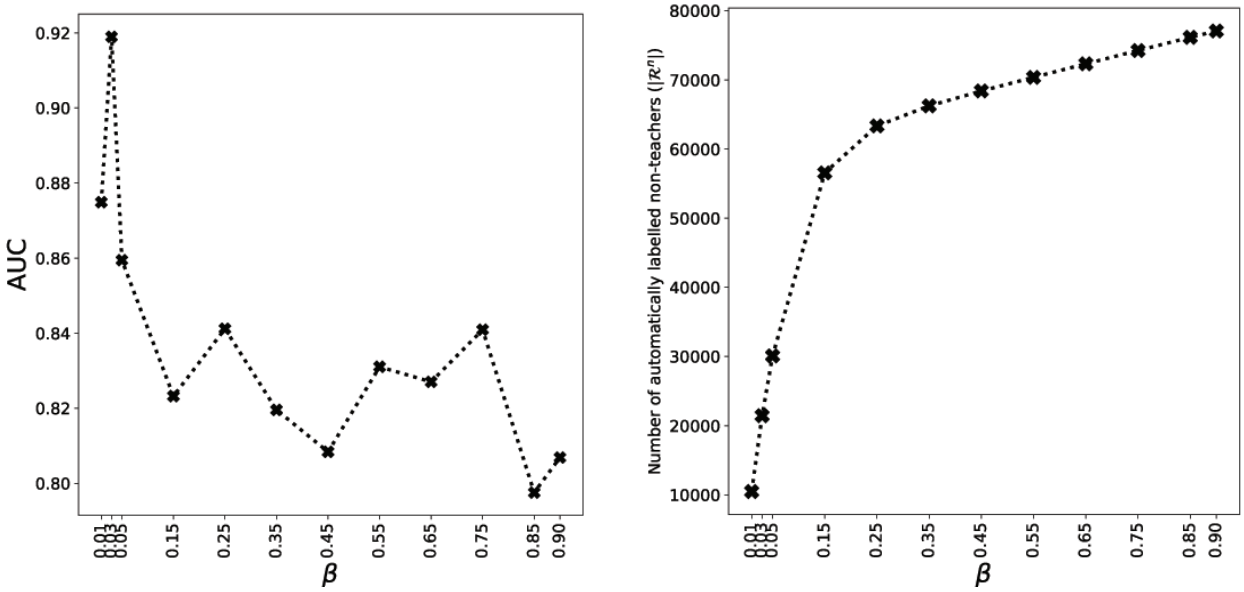
(a) (b)

Figure 3.7: Sensitivity analysis of the hyperparamter α.

As described in Section 3.3.2, there are two important hyperparameters in the *Automatic User Labeling* component of `PUTeacher`, namely α and β. In this part, we perform a sensitivity analysis of these two hyperparameters. Figures 3.7 and 3.8 show the sensitivity analysis for α and β, respectively. For each analysis, we keep one hyperparameter fixed and change the other one. While changing a hyperparameter, we report two measures: the AUC score of `PUTeacher` (Figures 3.7a and 3.8a) as well as the number of automatically labelled teachers and non-teachers (Figures 3.7b and 3.8b, respectively).

To assess the effect of α, we set $\beta = 0.03$. First, from Figure 3.7a, we can observe that for a proper α, our framework delivers a perfect performance, which indicates that the *Automatic User Labelling* component effectively identifies reliable teachers from unlabelled data. Moreover, when α is too small or too large, the performance drops. The reason for the former is that the *Automatic User Labeling* mistakenly marks many non-teachers as teachers. In other words, the teacher labeling filter is not restrictive enough. However, the latter case (i.e., when α is too large) makes the teacher

labeling filter too restrictive, and thus the framework marks only a small number of teachers– See Figure 3.7b. Consequently, this makes it hard for the final component of PUTeacher, *User Classification*, to learn an effective classifier. Finally, we can observe from Figure 3.7b that when α increases, the number of automatically labelled teachers decreases.

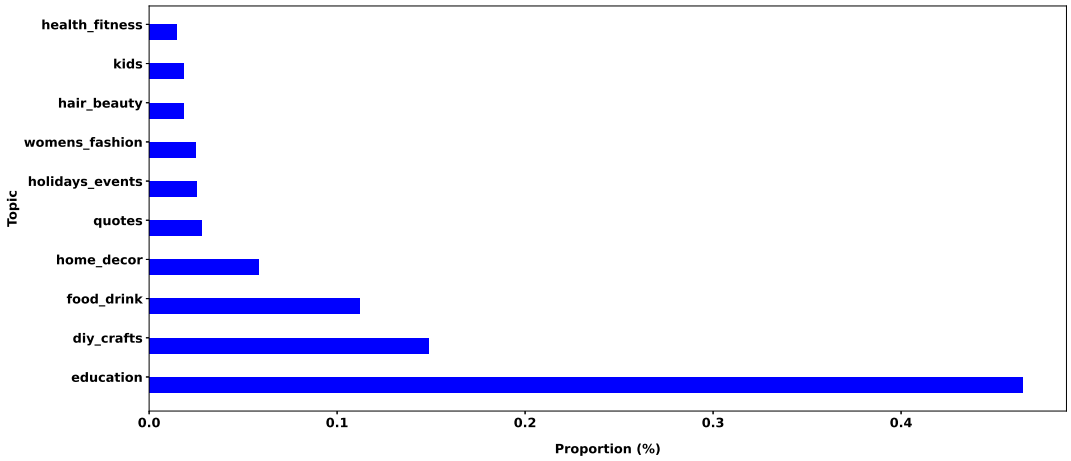


(a) (b)

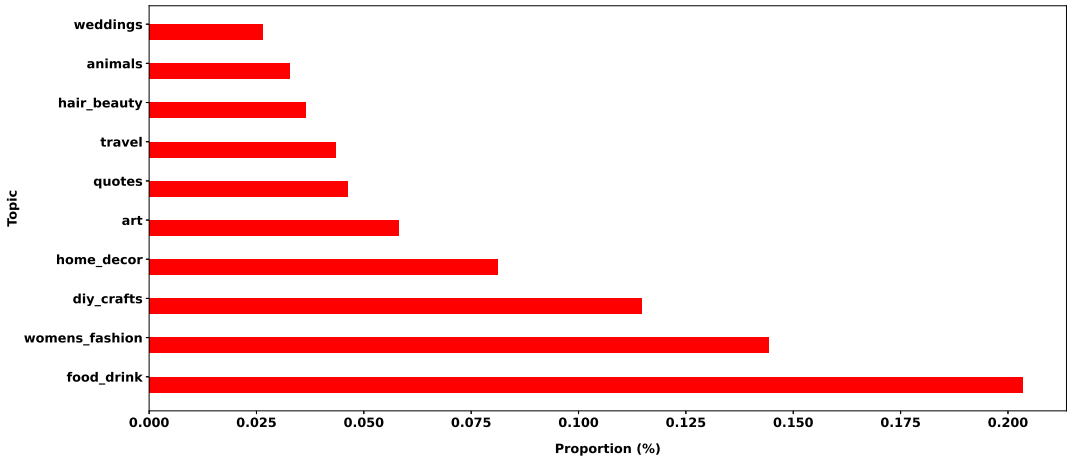
Figure 3.8: Sensitivity analysis of the hyperparamter β.

To assess the effect of β, we set $\alpha = 0.95$. We can observe from Figure 3.8b that when β is small, the performance is high since automatically labelled users enjoy high reliability. However, when β is becoming larger, the performance drops. Additionally, an interesting phenomenon occurs when β is set to a tiny number (less than 0.01): the performance drops significantly. The reason is that the number of identified non-teachers becomes very small, as shown in Figure 3.8b. Consequently, the *User Classification* component cannot properly learn to distinguish the two classes. Finally, one can observe in Figure 3.8b that by increasing β, the number of automatically labelled non-teachers increases.

3.5.4 Applying PUTeacher to Unlabelled Users

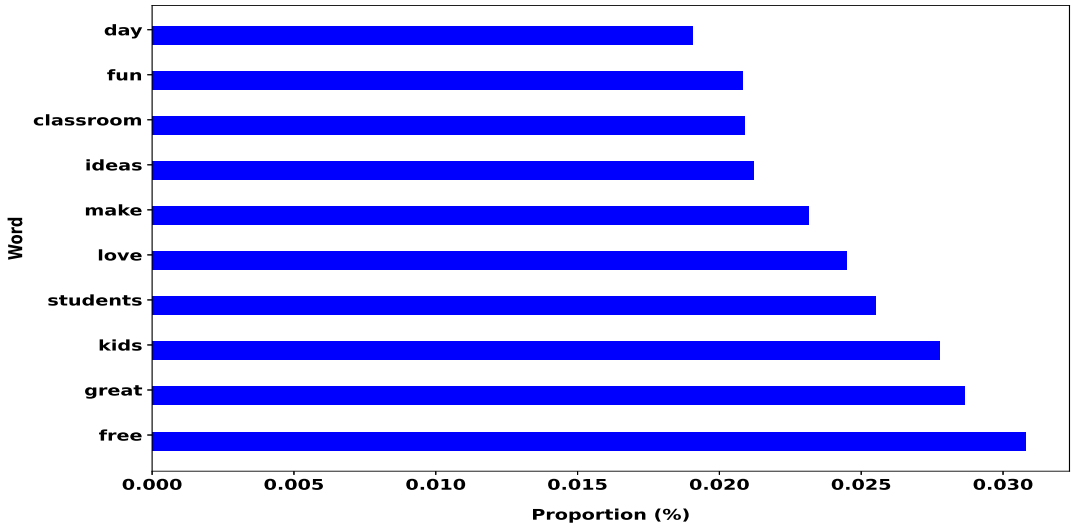


(a) *predicted-as-teachers*

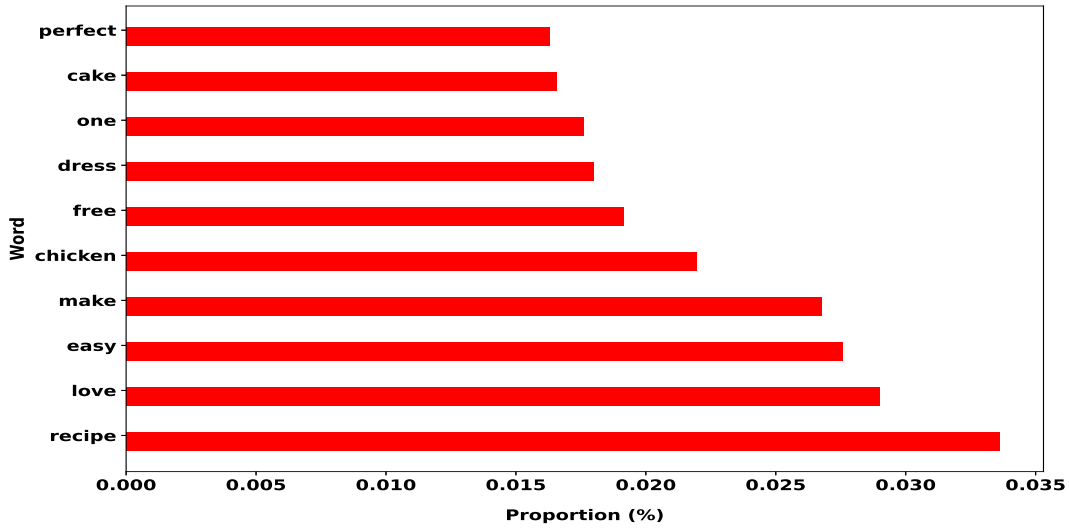


(b) *predicted-as-non-teachers*

Figure 3.9: The top 10 topics of pins of unlabelled users classified by PUTeacher.

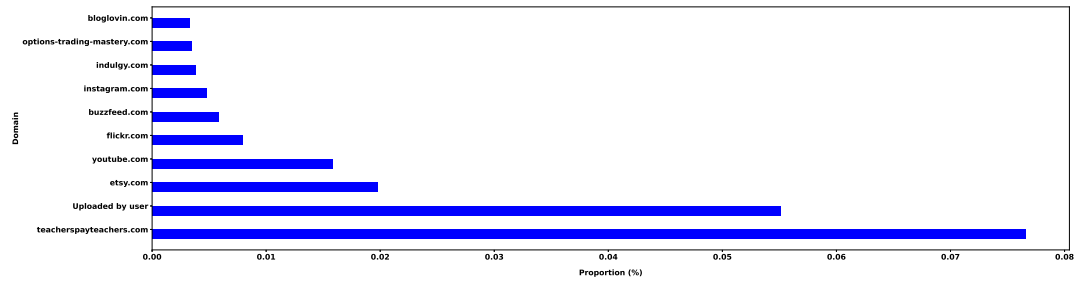


(a) *predicted-as-teachers*

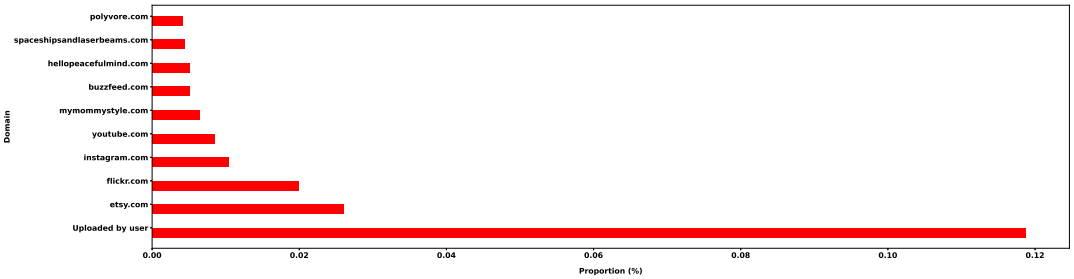


(b) *predicted-as-non-teachers*

Figure 3.10: The top 10 words of pin descriptions of unlabelled users classified by `PUTeacher`.



(a) *predicted-as-teachers*



(b) *predicted-as-non-teachers*

Figure 3.11: The top 10 domains of pins of unlabelled users classified by `PUTeacher`.

As mentioned before, the main goal of `PUTeacher` is to predict the class of unlabelled users, which reflects the practical scenario of how `PUTeacher` would be used. Hence, we used `PUTeacher` and predicted the class of unlabelled users (i.e., $\mathcal{X}^u$).[5] Now, we qualitatively assess two sets of predicted users: *predicted-as-teachers* and *predicted-as-non-teachers*. To make this assessment comparable, we shuffled predicted users and randomly select 5,000 *predicted-as-teachers* and 5,000 *predicted-as-non-teachers*. We retrieved the top 10 topics, 10 words (from pin descriptions), and 10 domains for *predicted-as-teachers* and *predicted-as-non-teachers*, as demonstrated in Figures 3.9, 3.10, and 3.11, respectively. We make the following observations based on these results.

- We can observe from Figure 3.9a that, as expected, *education* is the predominant topic of pins for *predicted-as-teachers* while it is not even in the list of top 10 topics for *predicted-as-non-teachers*.

[5] Recall that we do not have ground truth labels for these users.

- As far as the top 10 words are concerned, we can observe an interesting pattern. Almost all 10 words of pins belonging to *predicted-as-teachers* are somehow related to education and teaching, e.g., *kids*, *students*, *classroom*. In contrast, words associated with pins of *predicted-as-non-teachers* are related to other areas such as cooking– Note some words like *recipe*, *chicken*, and *cake* in Figure 3.10b.

- The word 'free' is the first and sixth frequently used word for *predicted-as-teachers* and *predicted-as-non-teachers*, respectively. We believe the main reason is as the following. First, Pinterest is widely used for business and marketing purposes. Given this, sometimes users mention the word *free* to indicate that their shared resources are free of charge explicitly, e.g., *free lessons on how to do photography in nature*. Hence, they probably mention the word *free* to help propagate their resources. In particular, teachers mention *free* in pin descriptions to attract the attention of their fellow teachers to their curated resources.

- The word 'ideas' is among the top 10 frequently used terms for pin descriptions of *predicted-as-teachers*. We believe such a significant emphasis of teachers on 'ideas' in their curated resources speaks to the distinct pattern employed by teachers in leveraging social media, in particular Pinterest, where they look for *novel* and *innovative* teaching *ideas* to supplement their educational resources. Such resources might be otherwise unavailable in their curriculum-based resources [4, 28, 34, 46, 91, 92].

- Another noteworthy word for *predicted-as-teachers* is 'fun'. We think the main reason for this word is that teachers tend to make their curated educational resources more engaging, e.g., *Fun counting coins games for first grade and second grade students*.[6]

- Ultimately, regrading the top 10 domains of pins, it is interesting to observe that *teachers-payteachers.com* is the top domain for *predicted-as-teachers*. *teacherspayteachers.com* is an online marketplace connecting millions of teachers and containing more than 5 million educational resources.

[6] `https://www.pinterest.com/pin/27443878969323415/`

From these observations, we can conclude that 1) PUTeacher can reliably identify teachers when it is applied to unlabelled (unseen) samples on Pinterest, and 2) the way that teachers leverage Pinterest for resource curation makes them outstanding among other Pinterest users, i.e., non-teachers.

3.5.5 State Representativeness

In this part, we shed light on the resiliency of PUTeacher from the perspective of the U.S. states of users in our dataset. First, we present the distribution of the U.S. states of users in our dataset. Then, we specify whether overrepresentation by certain states affects the generalization of PUTeacher to users from underrepresented states. Eventually, we present the distribution of the U.S. states for teachers automatically identified by PUTeacher.

3.5.5.1 The U.S. State Distribution

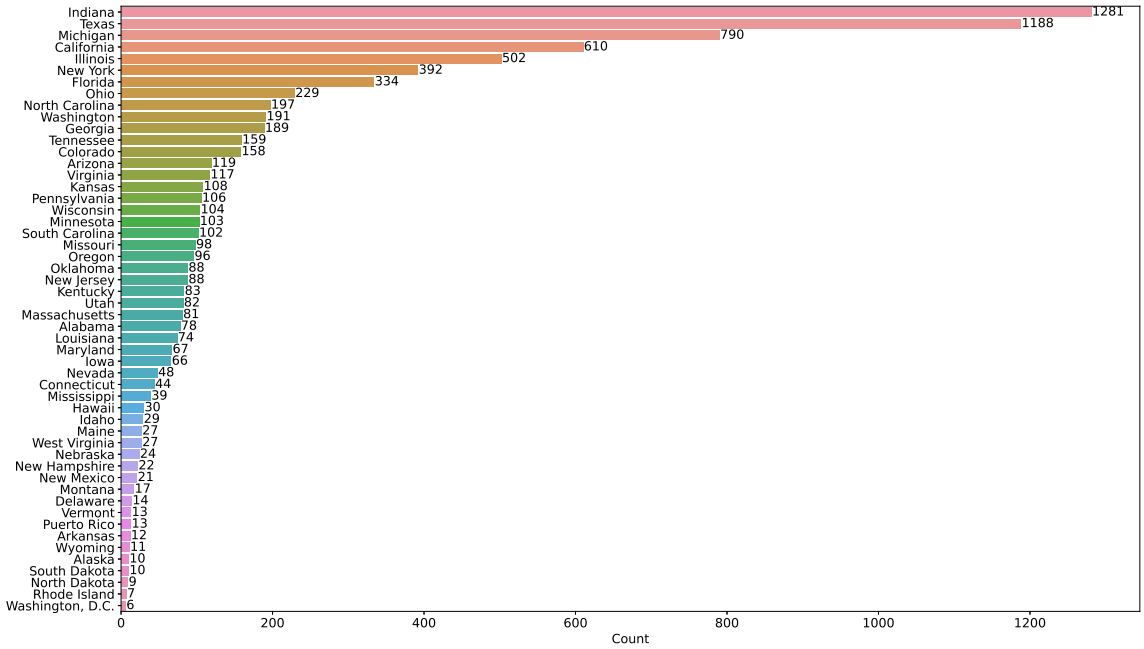


Figure 3.12: The distribution of the U.S. states for users in our dataset.

12,302 users in our dataset (around 14%) have shared their locations. We processed the strings of shared locations and managed to extract the U.S. states for 8,313 users. The state distribution for these users is shown in Figure 3.12. As can be seen in this Figure, the majority of teachers come from Indiana, Texas, and Michigan (around 40% combined). The main reason is that our surveyed teachers were sampled from four Midwest states (including Michigan, Indiana, Illinois, and Ohio) and Texas– See Figure 2.5. Consequently, those who are friends with these teachers probably come from the same states. Recall from Section 2.2 that other users in our dataset are Pinterest followers and followees (i.e., friends) of the surveyed teachers.

3.5.5.2 The U.S. State Generalization of PUTeacher

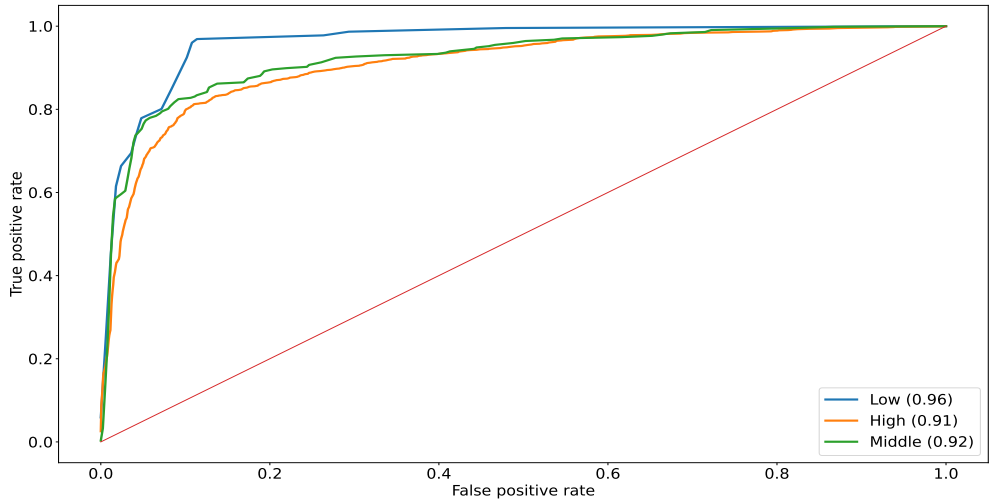


Figure 3.13: The ROC curves of PUTeacher's performance for three levels of state representativeness. Numbers in the parentheses are AUC scores.

Now the question is, does this state overrepresentation affect the performance of PUTeacher when it is used to predict the class of teachers in underrepresented states? To answer this question, we assessed the performance of PUTeacher on the test set users for three levels of the U.S. state representativeness, namely **over**-representative, **middle**-representative, and **under**-representative states. Following the top-down order of states in the y-axis of Figure 3.12, over-representative states include Indiana, Texas, and Michigan. Middle-representative states are from California to

South Carolina. Ultimately, under-representative states include the rest of the states from Missouri to Washington D.C. Figure 3.13 demonstrates the results of this experiment. Interestingly, we can observe that despite the overrepresentation by several states, `PUTeacher` delivers an excellent AUC score for middle-representative and under-representative states. Surprisingly, the performance for under-representative states is even better than the other two cases. From this experiment, we can conclude that `PUTeacher` is entirely robust against the underrepresentation in terms of the U.S. states of users (including teachers).

3.5.5.3 The U.S. State Distribution of Automatically Identified Teachers

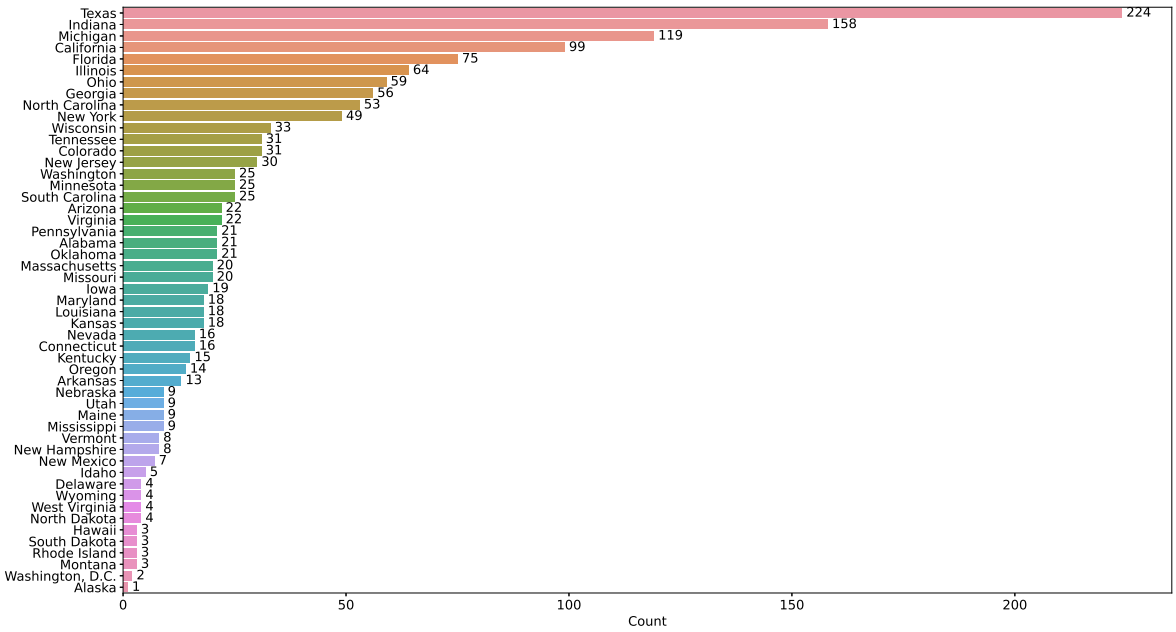


Figure 3.14: The distribution of the U.S. states of automatically identified teachers.

Finally, it is worthwhile to look into the U.S. state distribution of automatically identified teachers. To this end, similar to Section 3.5.4, we applied `PUTeacher` to all unlabelled users and considered a user as a teacher if their probability of being a teacher is larger than 0.9 i.e., according to the notation in Section 3.3.3, $y_i^t > 0.9$. From those predicted as teachers, we managed to extract the U.S. states for 1,769 teachers. Figure 3.14 shows the U.S. state distribution of these

1,769 automatically identified teachers. Similar to the U.S. state distribution of the entire dataset demonstrated in Figure 3.12, there are high numbers of teachers from Indiana, Michigan, and Texas. Again, this is because most of the surveyed teachers come from these three states, and consequently, their online teacher friends likely come from the same states. Despite this, we can observe that `PUTeacher` has been able to identify teachers from all states, including Washington D.C. This indicates the effectiveness of our proposed method and the widespread usage of Pinterest by teachers.

CHAPTER 4

GENDER ANALYSIS OF TEACHERS ON SOCIAL MEDIA

4.1 Introduction

Despite being a controversial topic, it is well known that there is a gap between male and female students when it comes to educational achievements [38]. Some studies attempting to determine this gap has focused on the role of teacher's gender. The general motivation behind this focus is that the academic environment created by teachers can significantly influence the way students see themselves as learners/students. Then, arguably, the teacher's gender has a significant role in the dynamic of this environment. For instance, students who have been discouraged from participation in classroom activities based on their gender have shown to be uniquely disadvantaged [112]. Moreover, some studies have explicitly investigated the difference in male and female teachers' treatment of boy and girl students [39, 40, 41, 42]. Although the evidence for the relationship between same-gender teachers and improvement in students' achievement is arguable [39], some still find same-gender teachers educationally relevant. The main reason is that same-gender teachers can affect engagement via perpetuating the role model effect and stereotype threat [42]. Regarding the latter, several studies have shown that male teachers face societal prejudice and judgment for violating gender stereotypes [113] such as the fear of being accused of inappropriate contact with students [114, 115] or being labeled as "weird", "gay", or "weak" [116]. Given this, perhaps it is no surprise to know that more than 75% of teachers in the U.S. are females (this number is around 82% for elementary school teachers) [117]. A similar trend persists in most other countries, especially in the Western nations [118]. In addition, it is worth mentioning that there have been efforts to recruit more male teachers due to the lack of male role models for boy students, otherwise knows as the decline of masculinity [119, 120]. Finally, the discussion about the impact of teachers' gender on the quality of education has extended beyond the academic literature, where teacher's gender among their other demographic attributes (e.g., race) have been the focus of many discussions

among policymakers, parents, students, and other educational stakeholders.

From the brief discussion above, we can infer that the current studies concerning the gender of teachers assume either: a) the teacher's gender has an innate value, e.g., the role model effect or b) the behavioral difference in male vs. female teachers is what indeed matters, e.g., the way teachers manage the classroom or prepare materials [43]. The former is beyond the scope of this dissertation, and thus this chapter focuses on the latter, where we attempt to determine differences and similarities in male and female teachers' behavior. However, unlike current studies, we focus on analyzing the behavior of male and female teachers through the lens of online social media where, as discussed in Chapter 1, nowadays plays a significant role in teachers' professional career development and is reshaping the entire teaching profession. Moreover, teacher gender analysis using online social media data compared to traditional educational data (e.g., surveys or interviews) have several advantages such as the fast and accessible data, less selection bias, and larger sample size (refer to Chapter 1).

Hence, in this study, we perform an exploratory analysis of male and females teachers on Pinterest. We focus only on teachers in the U.S. One of the main challenges to perform such a study is having a representative sample of male and female teachers whereby we mean the percentage of male and female teachers should be as close as possible to that of the general population of teachers in the U.S, i.e., 76.5% and 23.5% for female and male teachers (89% and 11% for elementary school teachers), respectively [117]. Moreover, to effectively perform a quantitative data-driven analysis, the sample size should be sufficiently large. Using only our surveyed teachers does not satisfy these two criteria. First, the percentage of male and female teachers in the surveyed teachers is 2.95% and 97.05%, respectively, which significantly differs from that of the teacher population in the U.S. Second, using only 540 teachers on Pinterest might bring into question the statistical significance of any quantitative analysis. Thanks to the automatic teacher identification framework proposed in Chapter 3, however, we can address these challenges. As will be presented in Section 4.2, using this framework, we can automatically identify thousands of teachers on Pinterest while their gender distribution matches that of the U.S. population of teachers. Using this dataset at our disposal, we

study male and female teachers on Pinterest from two crucial aspects, as briefly discussed in the following.

First, we investigate various online activities of male and female teachers, e.g., topics of pins, domains of pins, the number of boards. The motivation for this type of analysis is to understand male and female teachers through their resource curation process. Second, we investigate male and female teachers in the context of the social network (graph) they belong to, e.g., comparing the centrality of male and female teachers, determining gender homophily. The performed social network analysis complements online activity analysis by examining how male and female teachers are connected in a social network.

The novel analysis presented in this chapter sheds light on an unexplored area of research –male and female teachers on social media– which we believe has a great potential in fostering further research as it illuminates an important part of teachers' professional life in the information age (i.e., social media). From a practical perspective, our analysis and findings in this chapter can serve as a useful reference for many entities concerned with teachers' gender, e.g., educational scientists, policymakers, principals, state-level and national-level institutes. Moreover, as will be presented, some of the findings of this chapter are generalizable to all teachers regardless of their gender. In summary, our contributions in this chapter are as follow:

- Using our previously developed automatic teacher identification framework (Chapter 3), we build a large and representative sample of male and female teachers on Pinterest.
- To the best of our knowledge, this is the first data-driven study that analyses the teachers in social media concerning their gender.
- Our analysis considers two crucial aspects of teachers in social media and further informs us how teacher's gender plays a role in these aspects.

The rest of this chapter is organized as follows. First, in Section 4.2, we explain how we set up the dataset. Afterward, the online activity analysis is presented in Section 4.3, followed by the social network analysis in Section 4.4.

4.2 Dataset

In this section, we present how we constructed our dataset of male and female teachers. First, we discuss how we used our automatic teacher identification framework to incorporate more teachers in our teacher gender investigation. Afterward, we explain how we specified the gender of teachers, and finally, discuss the privacy concerns and taken measures to address these concerns.

4.2.1 Employing Automatic Teacher Identification

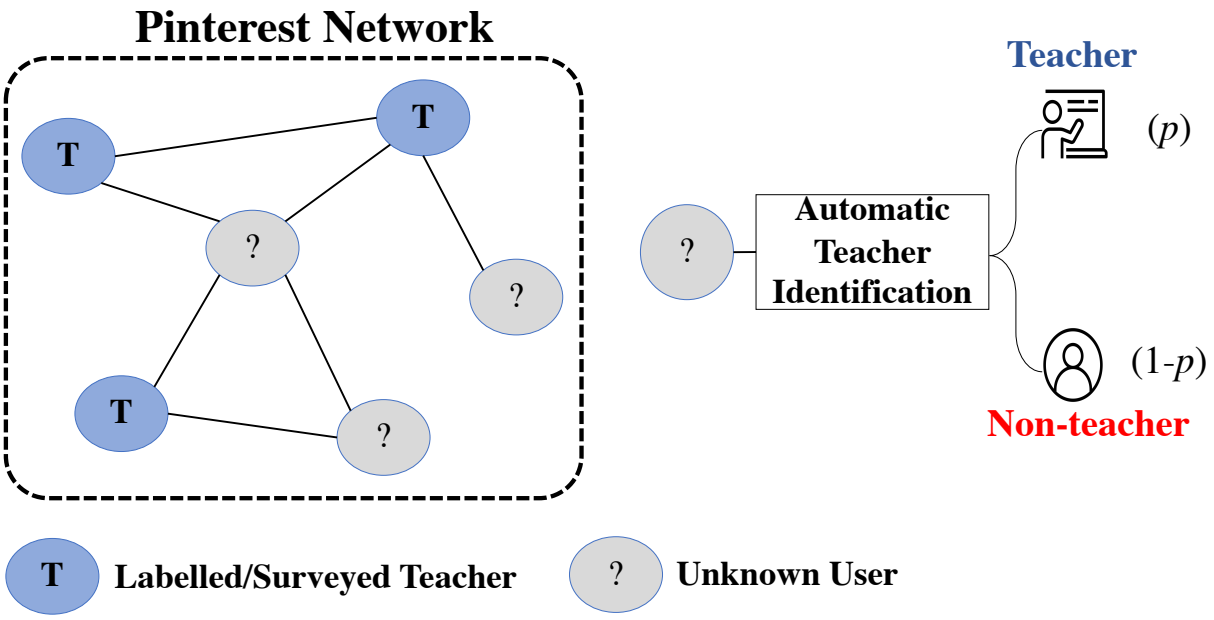


Figure 4.1: An overall illustration of our proposed automatic teacher identification approach (`PUTeacher`) presented in Chapter 3.

As mentioned in Section 4.1, two major data-related challenges facing our exploratory analysis of male and female teachers on Pinterest are 1) an insufficient number of teachers and 2) representativeness of teachers regarding their gender distribution. Using only our surveyed teachers does not resolve these challenges. While surveying more teachers seems like an immediate option, needless to say, that surveying is time-consuming, costly, and cumbersome. Hence, it is highly beneficial to devise a method that can automatically identify teachers on Pinterest. Fortunately, in Chapter 3, we proposed such an approach. For reference, we have included an overall illustration of our teacher

identification framework (PUTeacher) in Figure 4.1. To recall, as the input, PUTeacher takes the data of an unlabelled user, i.e., a Pinterest user connected to the surveyed teachers. As the output, it yields the probability that the user is a teacher, which is denoted as p in Figure 4.1. Obviously, $1 - p$ would be the probability of being a non-teacher.

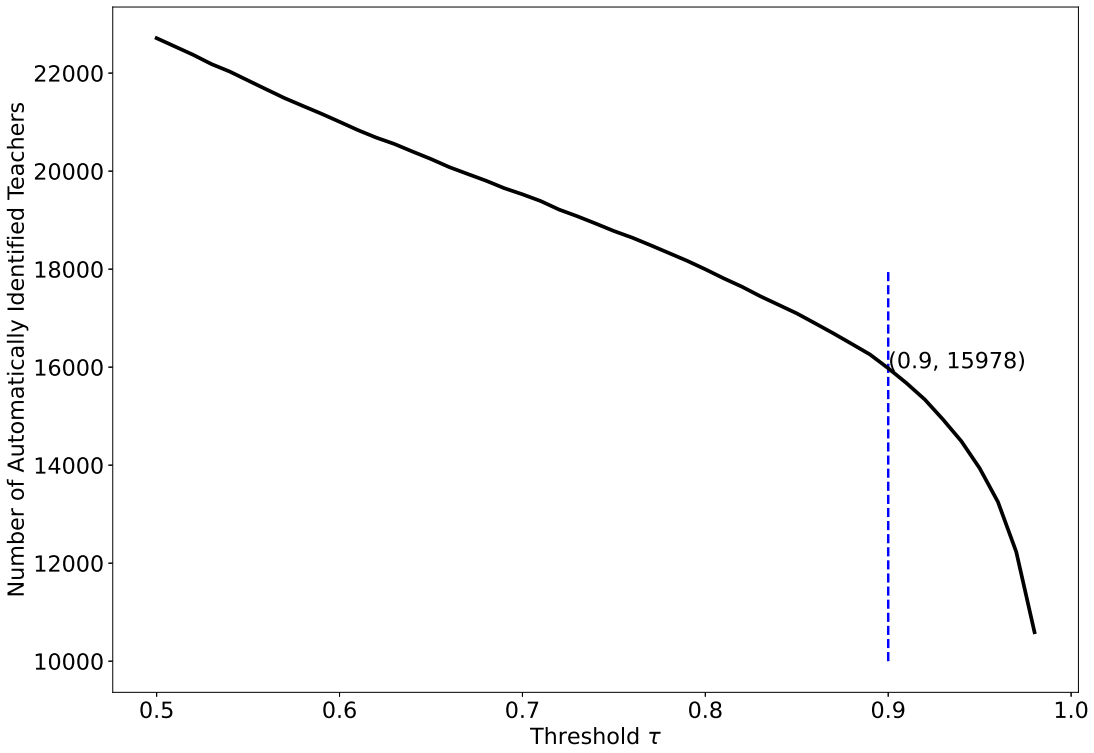


Figure 4.2: The number of identified teachers for different values of threshold τ.

Since the output of our automatic teacher identification approach is a probability distribution, we can set a threshold τ to specify the classification outcome, where if $p > \tau$, the user will be considered a teacher and otherwise a non-teacher. In Figure 4.2, we have plotted the number of automatically identified teachers when τ changes from 0.5 to 1. As shown in this figure, even being as conservative as $\tau = 0.9$, we can identify around 16,000 teachers on Pinterest, which significantly enlarges the sample size and thus addresses the first challenge mentioned above. In this study, we set the threshold to 0.9 to ensure high reliability in included users.

Remark. One might speculate that there might be some non-teachers among automatically identified teachers. While this can be possible, we believe it does not drastically affect the subsequent analyses in this chapter due to the following reasons. First, our rigorous evaluation in Chapter 3 revealed that the error in our method is very small. In particular, in Section 3.5, we performed a thorough resiliency analysis of `PUTeacher` and ensure it is a robust and reliable approach for automatic teacher identification on Pinterest. Second, the impact of a small number of incorrectly identified users will be "smoothed out" by a large number of correctly identified teachers and thus would not harm the *generalizability* of our results.

4.2.2 Gender Identification

Now, we have access to a large set of teachers on Pinterest; we need to a) identify their gender and b) keep only teachers who reside in the U.S. To do so, we perform the following three steps one by one.

Step 1. As mentioned in Chapter 2, the Pinterest API provided us with the self-declared gender of users– See Table 2.3. We only kept users whose recorded genders are specified (i.e., "Male" or "Female") and excluded "Unspeccified" ones.

Step 2. To further ensure that genders recorded in our dataset are correct, we utilized a secure and reliable commercial tool named Gender API.[1] For a given first name, Gender API determines its gender. It additionally provides an accuracy value in the range $[0, 1]$ specifying the certainty in the gender determination. Therefore, we passed all the first names of users from Step 1 to this program. Afterward, we applied two filtering operations. First, we excluded users whose corresponding accuracy acquired from Gender API is less than 0.8. Second, we kept only users whose genders from Pinterest API and Gender API match, i.e., both are male or female.

Step 3. As mentioned before, we need to restrict our analysis to teachers in the U.S. To this end, we included teachers whose field of *country* is "US" –See Table 2.3.

[1] `https://gender-api.com/`

We believe Step 1 and Step 2 combined offer a robust and reliable way of the gender specification of users. Moreover, Step 3 ensures including only teachers in the U.S. We also excluded users who had less than 20 pins in their accounts since they were very inactive on Pinterest. Table 4.1 demonstrates some basic statistics about our dataset after the above steps. Given this dataset of male and female teachers, we can assert that the second challenge (i.e., the gender representativeness) has been alleviated drastically. More specifically, compared to the surveyed teachers (2.95% male and 97.05% female), our new set of teachers (88% female and 12% male) is significantly more similar to the overall gender distribution of teachers in the U.S. (76.5% female and 23.5% male). Furthermore, the percentages of males and females in our dataset, i.e., 88% female and 12% male, perfectly match percentages of American elementary school teachers, i.e., 89% female and 11% male. This is particularly important since it has been shown that most teachers on Pinterest are elementary school teachers [117]. It is worth mentioning that the overrepresentation of female teachers on Pinterest is driven by two major factors: a) as mentioned before, the majority of teachers in the U.S. are female [117], and b) most of Pinterest users are female (around 77% [121]). In fact, Pinterest has been referred to as 'feminine' social media [122, 123, 124].

In addition to the number of users, Table 4.1 shows the statistics of several basic attributes about male and female teachers, e.g., the number of pins, the number of boards. In the remainder of this chapter, we will provide a rigorous analysis of these attributes. Finally, we emphasize that our constructed dataset of male and female teachers is the largest dataset of teachers in social media, with available gender information.

Table 4.1: Basic statistics of our constructed dataset of male and female teachers.

	Female Teachers	**Male Teachers**	**Total**
#Users	11,675 (88%)	1,592 (12%)	**13,267**
#Pins	67,705,475 (84%)	13,026,307 (16%)	**80,731,782**
#Boards	762,669 (88%)	102,986 (12%)	**865,655**
#Followees	975,775 (82%)	209,165 (18%)	**11,84,940**
#Followees (unique)	61,016 (70%)	25,940 (30%)	**86,956**
#Followers	738,326 (70%)	308,403 (30%)	**1,046,729**
#Followers (unique)	69,288 (68%)	33,036 (32%)	**102,324**
#Friends	1,714,101 (82%)	517,568 (18%)	**2,231,669**
#Friends (unique)	81,813 (67%)	40,120 (32%)	**121,933**

4.2.3 Privacy Concerns

Dealing with the demographic information of human subjects (acquired either through surveys or online social media) is not without privacy concerns. Nonetheless, we have been fully aware of these concerns and ensured they are appropriately addressed, as explained in the following.

- Pinterest data is publicly available, and we used authorized Pinterest API to acquire this data. Notwithstanding this, only authorized individuals have had access to this data.

- Regarding using the Gender API tool, we only submitted users' first names to this tool. In addition, we carefully reviewed the privacy statement of this tool[2] and ensured it is in line with our guidelines.

- Although we could have proceeded with using the Gender API tool to determine the gender of users whose value in our dataset is "Unspecified", we respected their decision in disclosing their gender and thus excluded those users in the analysis of this Chapter.

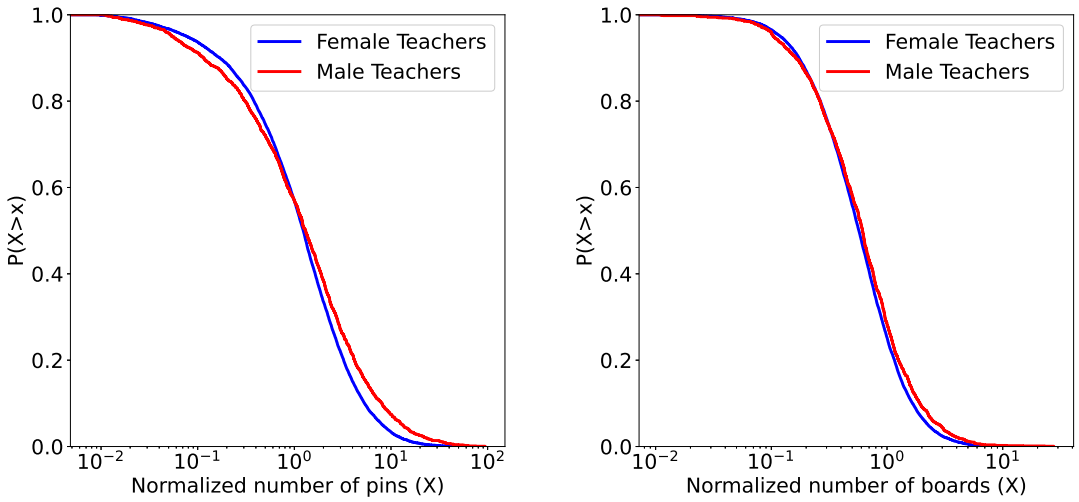
- Sharing our dataset with the scientific community can further advance the research in PK-12 education in general and teachers in social media in particular. Nevertheless, to protect the privacy of individuals, this sharing will only be possible upon proper communication and further institutional approval.

[2] `https://gender-api.com/en/privacy-policy`

4.3 Online Activity Analysis

In this section, we present a set of investigations related to the online activities of male and female teachers. These activities are primarily associated with pins and boards curated by teachers on Pinterest. We aim to delineate similarities and differences in how male and female teachers perform their online activities. In Section 4.3.1, we investigate the pin and board curation rate of male and female teachers. Sections 4.3.2 and 4.3.3 investigate the topics and domains of pins curated by male and female teachers, respectively. Afterward, in Section 4.3.4, look into the language used by male and female teachers to describe their pins and name their boards. Finally, in Section 4.3.5, we investigate how male and female teachers interact with Pinterest over the time.

4.3.1 Resource Curation Rate



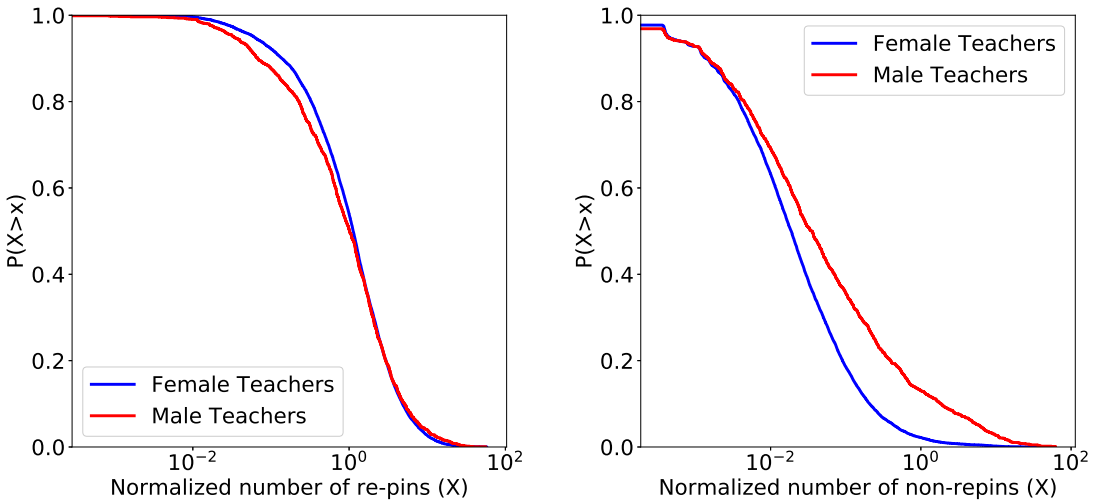
(a) Pinning rate

(b) Board curation rate

Figure 4.3: The CCDF of the number of pins and boards. x-axes are in log scale.

As its name suggests, **Pin**terest is all about pins. Also, as mentioned before, pins on Pinterest are organized in boards. Hence, our first online activity analysis is concerned with the number of pins and boards generated by male and female teachers. Figure 4.3a and Figure 4.3b show the cumulative distribution function (CCDF) of the number of pins and boards, respectively. Since not

all teachers have joined Pinterest at the same time, we need to consider the duration of an account. To this end, we divided the number of pins and boards by the number of *days* and *weeks* from the time a teacher has joined Pinterest to their last pin and board curation time, respectively. The reason for normalizing the number of pins and boards by different scales (days vs. weeks) is to account for the faster rate of pin curation than board curation. As shown in Figure 4.3a, the pinning rates for males and females are very similar, while for very active users (the tail of the distribution), male teachers tend to generate more pins. We also found almost identical distributions of the numbers of boards for male and female teachers (Figure 4.3b). This is in line with [125], which investigated the board creation rate of the general population of male and female Pinterest users and showed that their distributions are very similar.



(a) Rate of re-pins

(b) Rate of non-repins

Figure 4.4: The CCDF of re-pins and non-repins. x-axes are in log scale.

A pin on a user's account can be either a *re-pin* of someone else's pin or an original pin curated by the same user, which we call *non-repin*. Note that non-repin does not necessarily mean the user has created the pin's content (e.g., image). Instead, it simply means the user has not obtained it from *another Pinterest user*. To determine how male and female teachers behave regarding the re-pinning, Figure 4.4a and Figure 4.4b illustrate the CCDF of the number of re-pins and non-repins, respectively. In terms of re-pinning, we can observe from Figure 4.4a that both male and

female teachers behave very similarly. However, as far as the number of non-repins is concerned, we can observe from Figure 4.3 that male teachers tend to curate more non-repins (original pins) than female ones. To put it from a different perspective, female teachers on Pinterest are more "receptive" to sharing others' resources than male teachers. This is in line with [83] wherein the authors showed that female users tend to participate more in re-pinning than male users.

4.3.2 Topic of Pins

As discussed in Chapter 2, previous studies have approached teachers in social media from several different theoretical frameworks. One of these theories, which is pertinent to our investigation in this part of the dissertation, is *affinity space* [76] which has been the basis of several teachers in social media studies [69, 70, 75, 126]. In the context of teachers in social media, affinity space means teachers leverage social media to interact with each other about certain *topic(s)* relevant to their profession. Hence, motivated by the affinity space theory, in this part, we perform an in-depth investigation of teachers' *topics* of interest in their resource curation process. Since the main focus of this Chapter is the gender analysis of teachers, we conduct our investigation and present the findings while considering teachers' genders. As far as topics are concerned, luckily, Pinterest supports a set of pre-defined topics covering a wide range of categories such as *art*, *animal*, *travel*, *education*. A pin can belong to any of the existing 33 topics or 'others', i.e., in total, we have 34 topics.[3] Since these topics essentially encode the inherent users' preference, previous Pinterest-based studies have incorporated them into their user behavior analysis [86, 83, 125].

4.3.2.1 Top Topics

Figure 4.5 demonstrates the average percentage of each topic for male and female teachers. We make the following observations based on this figure.

[3]Both the supported topics and their numbers have slightly changed during the past few years. Hence, the topics in our dataset might be different from those in previous studies or future ones.

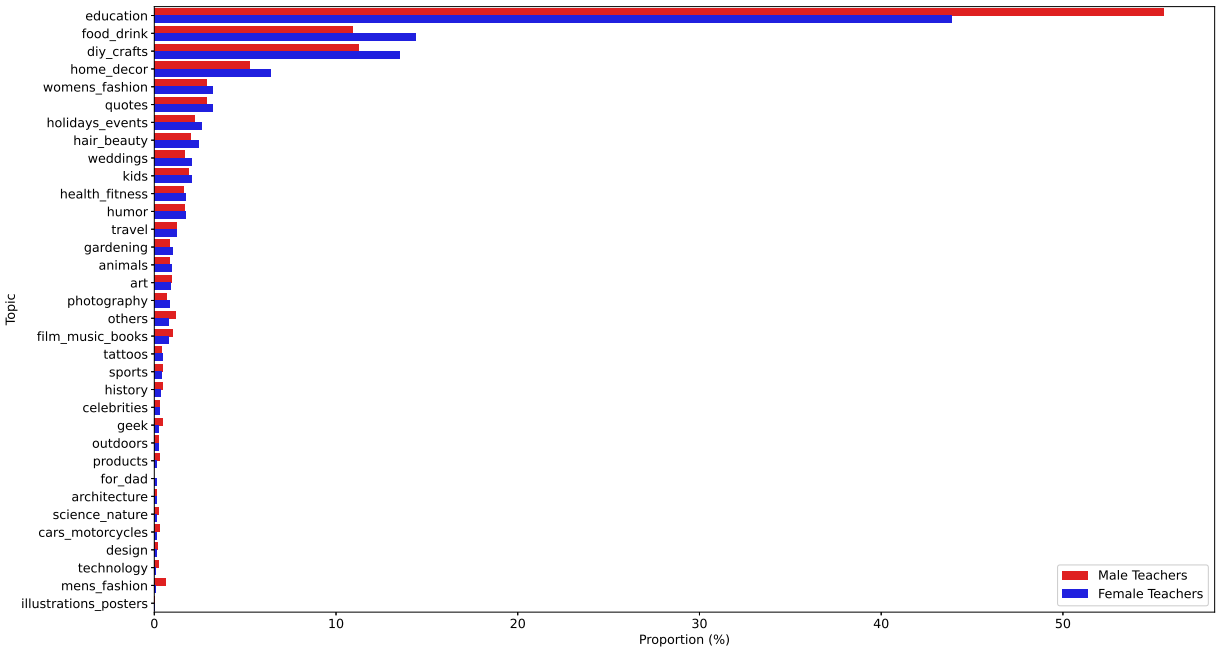


Figure 4.5: The average proportion of topics for male and female teachers.

- A crucial pattern that is immediately noticeable is that *education* is the predominant topic for both gender groups. Regarding our discussion about the affinity space, this signifies that Pinterest acts as a proper affinity space for teachers since they have leveraged Pinterest as a medium to focus on a specific topic relevant to their profession, i.e., *education*.

- Although the predominant topic for both gender groups is *education*, the percentage of *education* is higher for male teachers (55.57%) than female teachers (43.87%). This indicates that male teachers have *focused* more on educational resources while female teachers have *explored* other types of pins as well. This can be further inferred by looking at the other top topics, e.g., *food_drink*, *diy_crafts*, wherein female teachers have higher contributions than male teachers. In the subsequent analysis of this part, we dig deeper into this difference.

- The top four topics cover 78.21% and 83.02% of pins for female and male teachers, respectively. These skewed distributions indicate the existence of the power-law phenomenon in the topic distribution. Previous studies have observed a similar pattern regarding the topic distribution on Pinterest [83, 86, 125].

- Excluding *education*, male and female teachers' preferences in some other topics match what previous studies have identified. For instance, *food_drink* and *diy_crafts* are of the high interest for both male and female Pinterest users [86, 125].

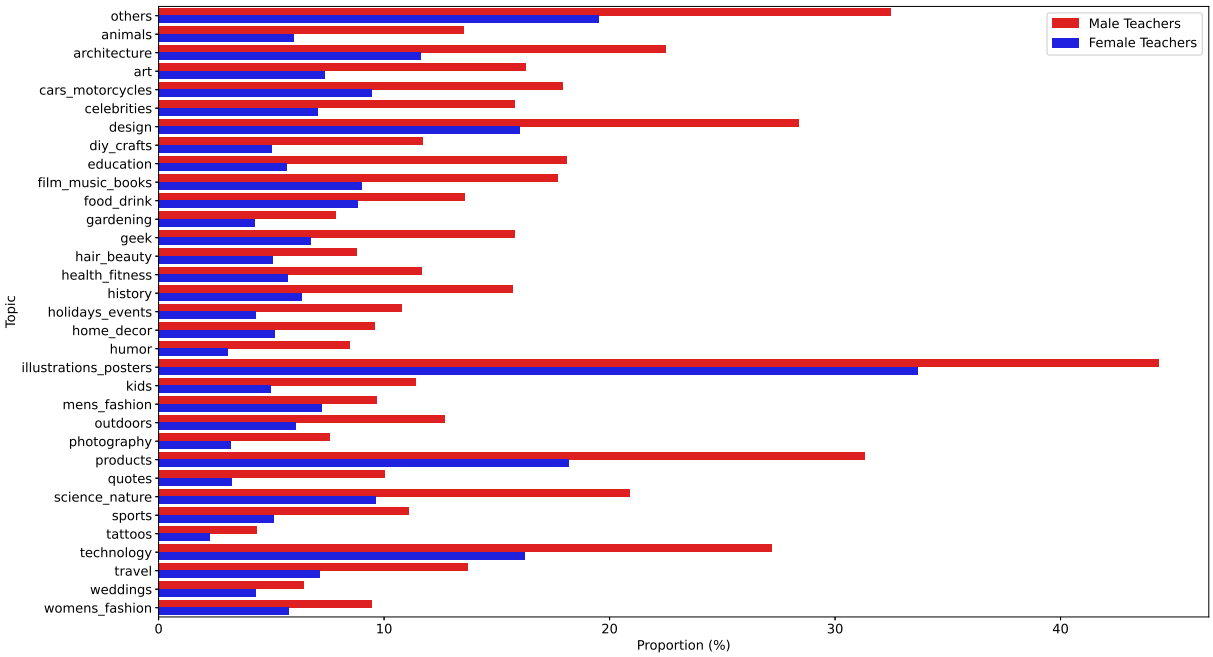


Figure 4.6: Average proportion of topics of non-repins for male and female teachers.

To deepen our understating of the distribution of topics, we computed the contribution of male and female teachers in each topic where we included only non-repins. The result is shown in Figure 4.6. Consistent with what was presented in Section 4.3 and the previous studies on the behavior of male and female Pinterest users [83, 84], for all topics, male teachers are more active in pinning than re-pinning. Moreover, both male and female teachers have made more non-repins in topics for which the overall number of pins is low, e.g., *illustrations_posters*. In other words, teachers curate the primary resources of their interest (i.e., educational pins) by taking them from other teachers and other Pinterest users. However, they locate pins related to their secondary interests (i.e., non-educational pins) on the web or directly upload them from their device.

4.3.2.2 Topic Entropy

For the chart demonstrated in Figure 4.5, we combined all pins curated by a gender group and then calculated the proportion of each topic. This reveals the overall behavior of male and female teachers regarding topics. Nevertheless, it does not inform us how an individual teacher behaves with respect to the topics of the pins they have curated. Therefore, to acquire such information, similar to previous studies [86, 83, 125], we utilize the notion of *topic specialization*. A completely topic-specialized user merely sticks to a single topic while a less topic-specialized user contributes to various topics. To quantify the topic specialization, we introduce the *topic entropy*. Suppose we have k topics $\mathcal{T} = [t_1, t_2 \cdots t_k]$ ($k = 34$ in our dataset). Further, for a given user, let p_{t_i} denote the fraction of their whose topic is t_i. Then, the *topic entropy* (TE) of a user u is defined as follows:

$$TE(u) = -\frac{m}{k} \times \sum_{\forall t_i \in \mathcal{T}} p_{t_i} \times ln(p_{t_i}) \tag{4.1}$$

where m is the number of topics wherein the user has at least one pin. The term $\frac{m}{k}$ smooths out the effect of the natural logarithm in the entropy formulation by accounting for the total number of topics. TE is in the range $[0, ln(k)]$ and the closer TE is to 0 ($ln(k)$) the more (less) topic-specialized a user is.

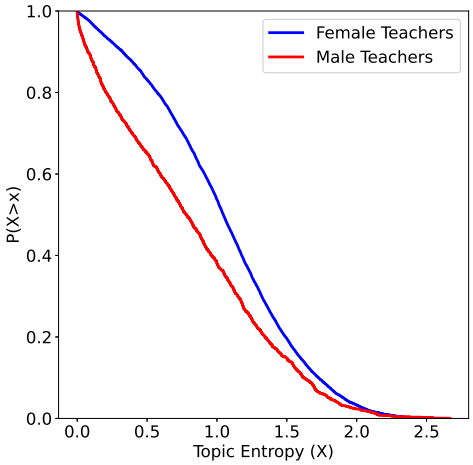
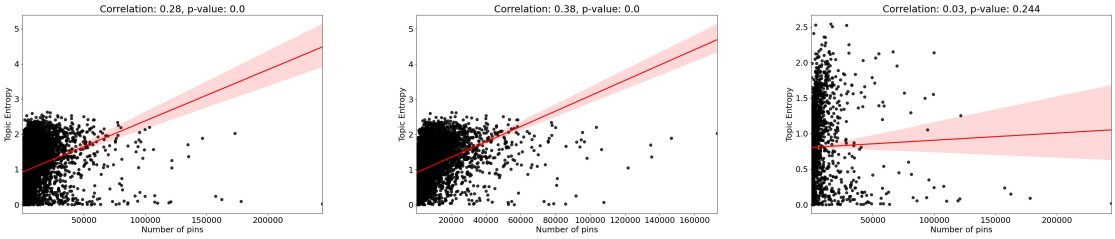


Figure 4.7: The CCDF of the topic entropy (Eq. 4.1).

Figure 4.7 shows the CCDF of the topic entropy for male and female teachers. As it can be observed from this figure, female teachers exhibit a smaller degree of topic specialization. Ottoni et al. [125] and Chang et al. [86] discovered a similar finding regarding female Pinterest users. Moreover, several psychological/medical studies have demonstrated that men are more focused than women while women are better at multi-tasking than men [127, 128]. Perhaps our finding regarding the difference in the topic specialization between male and female teachers (and even males and females on social media) can be linked to these psychological/medical studies. Nevertheless, this connection needs further rigorous analysis, and we leave it for future work.

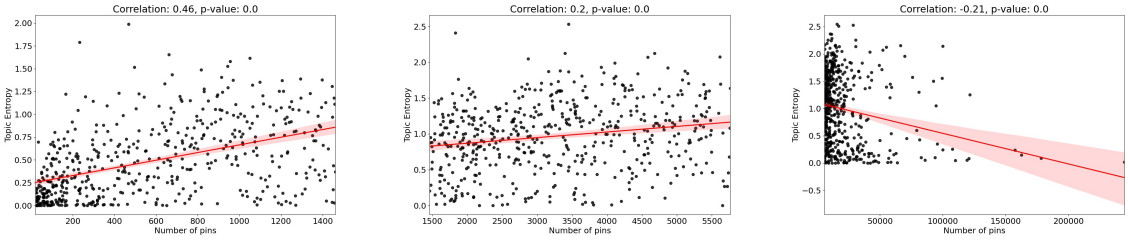


(a) Male & female teachers (b) Female teachers (c) Male teachers

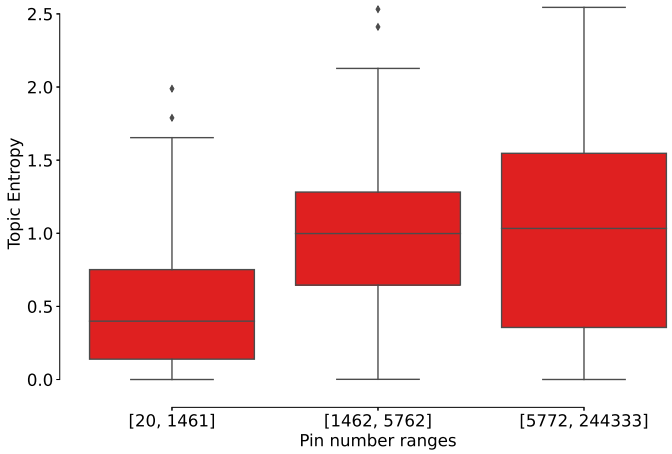
Figure 4.8: The topic entropy based on the number of pins.

One might wonder whether the topic specialization for a user is related to the number of their pins or not. For instance, we might expect a higher degree of topic specialization for a user with more pins since potentially more topics could be covered. Hence, we investigated the relationship between the topic entropy and the number of pins, as demonstrated in Figure 4.8. Figure 4.8a shows a case where we combined the data of male and female teachers. To exclude the gender from the relationship between the topic entropy and the number of pins, we investigated this relationship for each gender group separately. Specifically, Figures 4.8b and 4.8c show the topic entropy vs. the number of pins for female teachers and male teachers, respectively. For each chart, we also included a fitted regression line between the topic entropy and the number of pins and the Pearson correlation between these two variables. For the male and female teachers combined shown in Figure 4.8a, we can observe that as the number of pins increases, the topic entropy exhibits moderate growth, and the fitted line enjoys a positive slope. Also, the correlation is 0.28 with the p-value of almost zero,

which is statistically significant.[4] For female teachers, the trend is similar to that of Figure 4.8a, where we observe a positive correlation (c=0.38) between the topic entropy and the number of pins. Also, this correlation is statistically significant (p=0.0). Nevertheless, the story is different for male teachers. According to Figure 4.8c, there is no correlation between the topic entropy and the number of pins for male teachers. However, since the p-value is large (p=0.24), we cannot categorically assert no correlation. This motivated us to dig deeper into the relationship between the topic entropy and the number of pins. To this end, we discretized the number of pins into three distinct ranges for both male and female teachers. Afterward, we determined the relationship between the topic entropy and the number of pins in each range. Next, we present the results of these experiments.



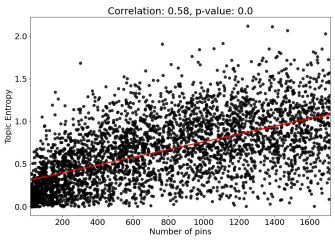
(a) Range $[20, 1461]$ (b) Range $[1462, 5762]$ (c) Range $[5762, 244333]$



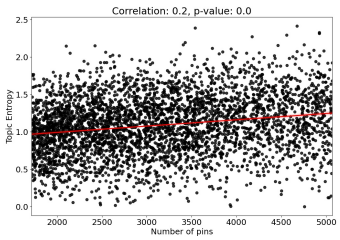
(d) Boxplots of the topic entropy for each range of the number of pins

Figure 4.9: The topic entropy for male teachers across three distinct ranges of the numbers of pins.

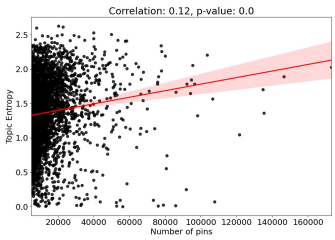
[4]Hereafter, we consider $p < 0.05$ as statistically significant.



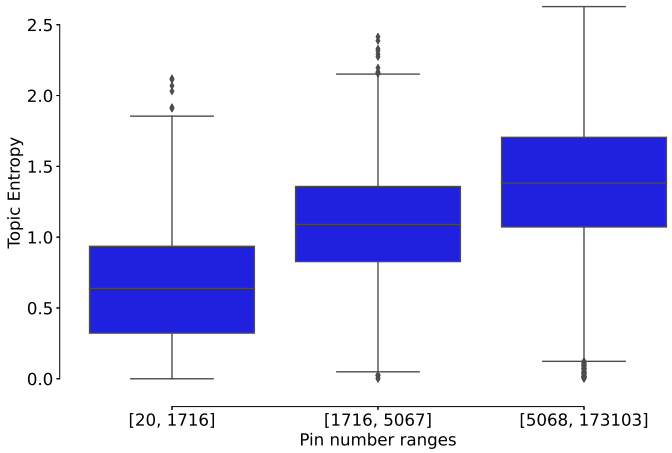
(a) Range $[20, 1716]$



(b) Range $[1716, 5067]$



(c) Range $[5068, 173103]$



(d) Boxplots of the topic entropy for each range of the number of pins

Figure 4.10: The topic entropy for female teachers across three distinct ranges of the numbers of pins.

Figure 4.9 illustrates the results for male teachers.[5] Here, we can observe an interesting pattern: in the low range of the number of pins, i.e., in Figure Figure 4.9a, when the number of pins increases, male teachers tend to try their hand in different topics, thus increasing the topic entropy. However, for the middle range, shown in Figure 4.9b, the correlation, while still being positive, drops significantly. Eventually, for the high range of the number of pins demonstrated in Figure 4.9c, the correlation becomes negative, which indicates that male teachers tend to become more topic-specialized. Hence, overall, we can conclude that the more prolific a male teacher becomes, the more he focuses on specific topics, i.e., becoming more topic-specialized. This can be observed from the boxplots in Figure 4.9d as well.

[5]The first ranges start from 20 since we filtered out those teachers with less than 20 pins in their accounts– See Section 4.2.

We conducted the same experiment for female teachers and the results are shown in Figure 4.10. Similar to male teachers, by moving from the lowest range of the number of pins (Figure 4.10a) to the highest range of the number of pins (Figure 4.10c), the correlation between the topic entropy and the number of pins drops significantly. However, unlike the male teachers' case, the correlation always remains positive regardless of the number of pins. This shows that prolific female teachers, similar to prolific male teachers, tend to focus on specific topics (being more topic-specialized) while, compared to males, the extent of this specialization is consistently smaller.

4.3.2.3 Topic Oscillation

Teachers might exhibit topic *variation* in the sequence of their pins over time. Capturing this variation can help us understand how much teachers stay *on-topic* while curating resources. However, the topic entropy (Eq. 4.1) cannot capture this variation since it calculates the entropy in a set of pins without considering the sequential order between them. More specifically, a user can be topic-specialized (i.e., focusing on certain topics) while still frequently drifting in these topics. To fix the idea, consider a simple example where a user has curated 8 pins having this sequence of topics $[c_1, c_2, c_1, c_2, c_1, c_2, c_1, c_2]$. Further, suppose the total number of topics is 10 ($k = 10$). TE for this user is 0.13. Compared to the maximum value of the topic entropy, i.e., 2.30,[6] this is a low value for the topic entropy of a user. However, while this user is topic-specialized to a large degree, they have varied from one topic to another in every two consecutive pins. Therefore, to capture the variation in topics of pins, we propose the *topic oscillation*. Let $S = [c_1, c_2 \cdots c_n]$ denote the chronologically ordered sequence of topics of pins for a given user where $c_i \in \mathcal{T}$. Then, the topic oscillation (TO) of a user is defined as follows:

$$TO(u) = \frac{1}{n-1} \times \sum_{i=1}^{n-1} \mathbb{1}(c_i = c_{i+1}) \tag{4.2}$$

[6] $ln(10) = 2.30$.

where $\mathbb{1}$ is the indicator function.[7] The minimum value of $TO(.)$ is 0 and it occurs when the user has pinned resources from a single topic. The maximum value of the $TO(.)$ is 1 and it occurs when the topic has changed for every two consecutive pins.

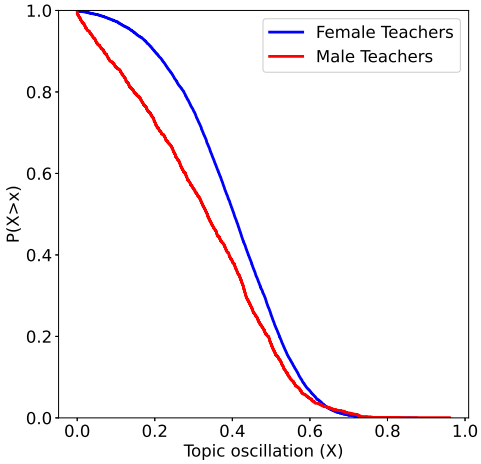
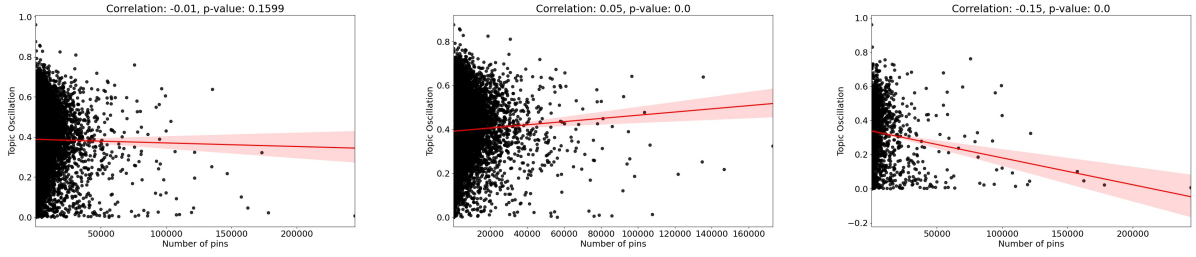


Figure 4.11: The CCDF of the topic oscillation (Eq. 4.2).

Figure 4.11 shows the CCDF of the topic oscillation for male and female teachers. As it can be observed from this figure, female teachers exhibit a smaller degree of the topic oscillation. In other words, overall, compared to female teachers, male teachers stay more on-topic while curating pins.

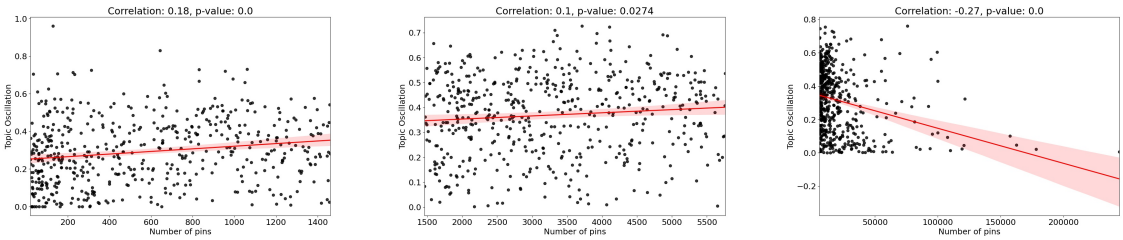


(a) Male & female teachers (b) Female teachers (c) Male teachers

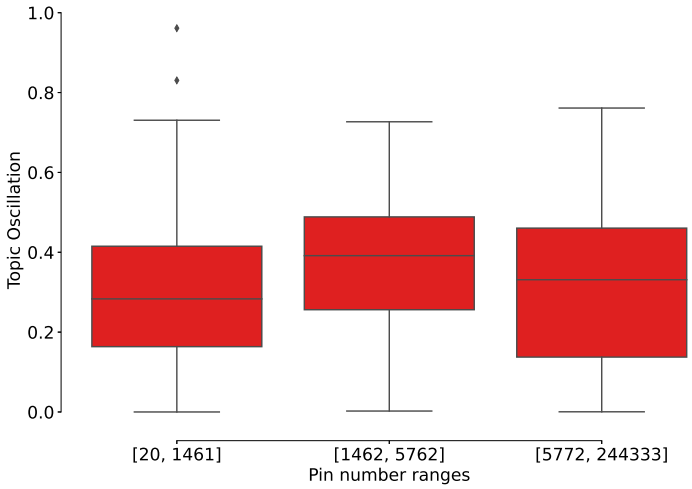
Figure 4.12: The topic oscillation based on the number of pins.

Does having a high or a low number of pins in a user's account is correlated with their topic oscillation? To answer this, similar to the topic entropy, we investigated the relationship between the topic oscillation and the number of pins, as shown in Figure 4.12. When we combined the data

[7] https://en.wikipedia.org/wiki/Indicator_function



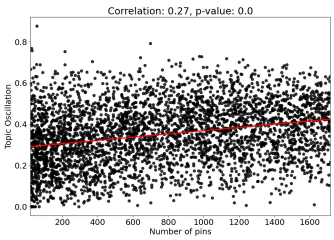
(a) Range $[20, 1461]$ (b) Range $[1462, 5762]$ (c) Range $[5762, 244333]$



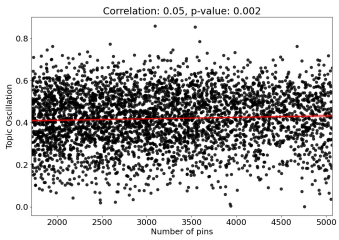
(d) Boxplots of the topic oscillation for each range of the number of pins

Figure 4.13: The topic oscillation for male teachers across three distinct ranges of the numbers of pins.

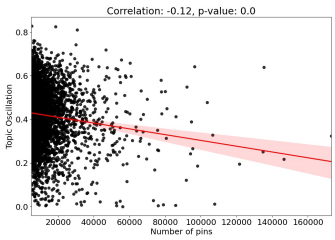
of male and female teachers (i.e., Figure 4.12a), there is almost no correlation between the topic oscillation and the number of pins. However, the p-value is relatively high, and thus the Pearson correlation in Figure 4.12a is not statistically significant. A closer look at Figures 4.12b and 4.12c reveals why the combined data of male and female teachers exhibit no correlation, and the p-value is high. The reason is that the data of female teachers and male teachers contains two distinct patterns in terms of the correlation between the topic oscillation and the number of pins where the former, in general, shows almost no correlation while the latter is associated with a negative correlation. Therefore, to understated the behavior of male and female teachers regarding the topic oscillation, we investigated the correlation between the topic oscillation and the number of pins in the three ranges similar to what we performed for the topic entropy in Section 4.3.2.2.



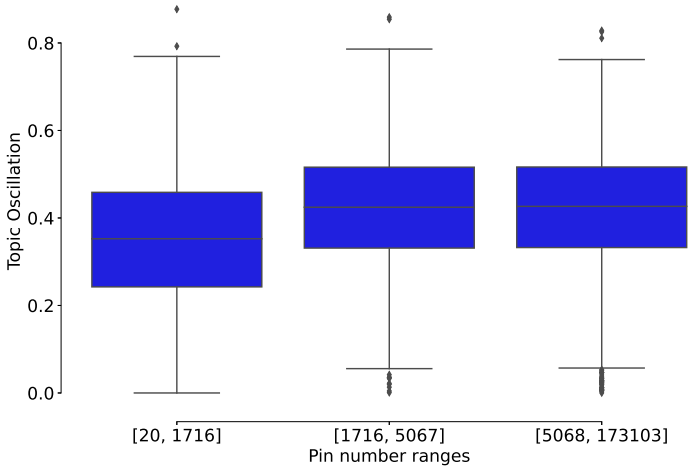
(a) Range $[20, 1716]$



(b) Range $[1716, 5067]$



(c) Range $[5068, 173103]$



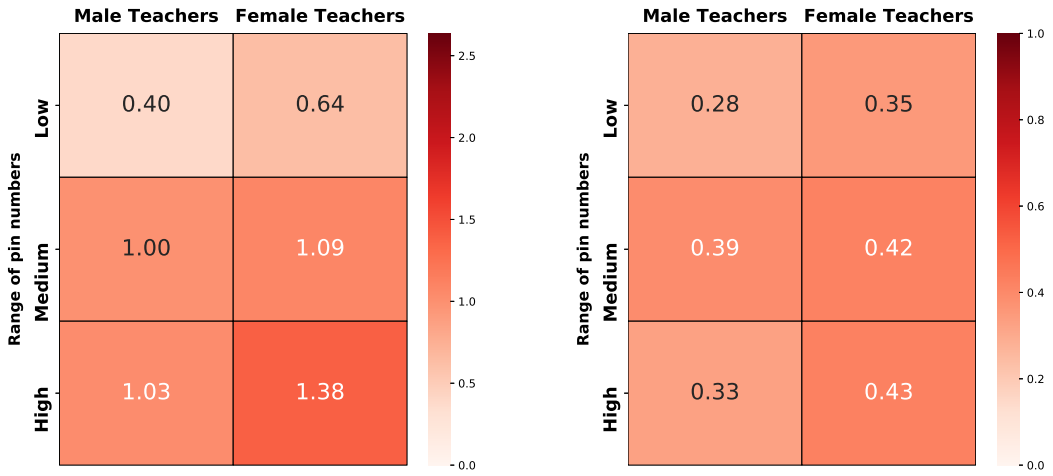
(d) Boxplots of the topic oscillation for each range of the number of pins

Figure 4.14: The topic oscillation for female teachers across three distinct ranges of the numbers of pins.

Figure 4.13 shows the correlation between the topic oscillation and the number of pins for male teachers in the three ranges of the number of pins. Here we observe a similar pattern with the topic entropy. For the low range of the number of pins illustrated in Figure 4.13a, the correlation is positive. For the next range (Figure 4.13b), the correlation decreases, yet it remains positive. However, the correlation becomes negative for the last range (Figure 4.13c). Hence, based on the results shown in Figure 4.13, we can conclude that industrious male teachers are more likely to stay on-topic in their pinning endeavor.

The correlation between the topic oscillation and the number of pins for female teachers in the three ranges is shown in Figure 4.14. The correlation is positive for the first two ranges, while for the last one, it becomes negative. Also, the correlation decreases as we move from a range to

the adjacent one. Thus, as far as female teachers are concerned, they exhibit two distinct patterns regarding the topic entropy and the topic oscillation. While they accumulate a larger number of pins, they tend to become less topic-specialized, whereas they manage to stay on-topic.



(a) The topic entropy in the three ranges (b) The topic oscillation in the three ranges

Figure 4.15: A summary of the topic entropy and the topic oscillation for male and female teachers (values are median in ranges).

Finally, Figure 4.15a summarizes the topic entropy and the topic oscillation values for the three defined ranges of the number of pins. We have labeled the three ranges as *low*, *medium*, and *high*, signifying their relative coverage of the number of pins. In summary, we can conclude that, compared to female teachers, male teachers are more topic-specialized and tend to stay more on-topic in their pinning endeavor.

4.3.3 Domain of Pins

As demonstrated in Section 2.2.1, a Pinterest user can pin (save) an image/video from virtually anywhere on the web in their account. Because of this feature, Pinterest has been referred to as a social curation website [83, 84, 85, 86, 87, 88]. The social curation nature of Pinterest has made it very appealing to teachers [4, 80]. Moreover, since the source of a resource can be practically any place on the web, it is very important in the context of teachers in social media. More specifically,

the source informs us whom teachers turn to for obtaining educational resources. Then, through this knowledge, we can characterize the educational resources curated by teachers and assess their quality which ultimately paves the way to determine the teacher's quality [79]. Additionally, not only for teachers but also for the general Pinterest users, the sources of pins play an essential role in understating the behavior of users [125, 129, 86]. This is because the sources of pins essentially embed crucial information about the user's preference and pinning behavior. Hence, in this part of the dissertation, we investigate the sources of pins. While doing so, we go one step further and incorporate the gender of teachers in our investigation. Pinterest records the source of a pin, a URL (Uniform Resource Locator) of the image/video– See Figure 2.2. This URL includes the *domain* of the source, e.g., *teacherspayteachers.com*. Therefore, we investigate the sources of pins via their domains.

4.3.3.1 Top Domains

Figures 4.16 and 4.18 demonstrate the percentage of each of the top 20 domains of pins for male and female teachers, respectively. In addition, we calculated the distribution of the top domains across topics, as shown in Figure 4.17 for male teachers and Figure 4.19 for female teachers. More specifically, we created the domain-topic matrices DT^m and DT^f for male teachers and female teachers, respectively. Then, an entry $DT^m_{(d,t)}$ or $DT^f_{(d,t)}$ represents the percentage of pins whose domain is d and topic is t where d is a domain from the 20 top domains and t is a topic from 34 existing topics i.e., $\mathcal{T}$. For instance, $DT^m_{(etsy.com,diy\_craft)}$ in Figure 4.17 is 75.2, which means the topic of 75.2% of male teachers' pins coming from *etsy.com* is *diy_craft*. We make the following observations from these figures.

- First of all, we can observe that the predominant domain for male and female teachers is *teachers-payteachers.com*. This seems reasonable since, as mentioned before, *teacherspayteachers.com* is the largest marketplace of online educational resources and is very popular among American educators. In fact, *teacherspayteachers.com* is colloquially considered *ebay.com* for educational resources.

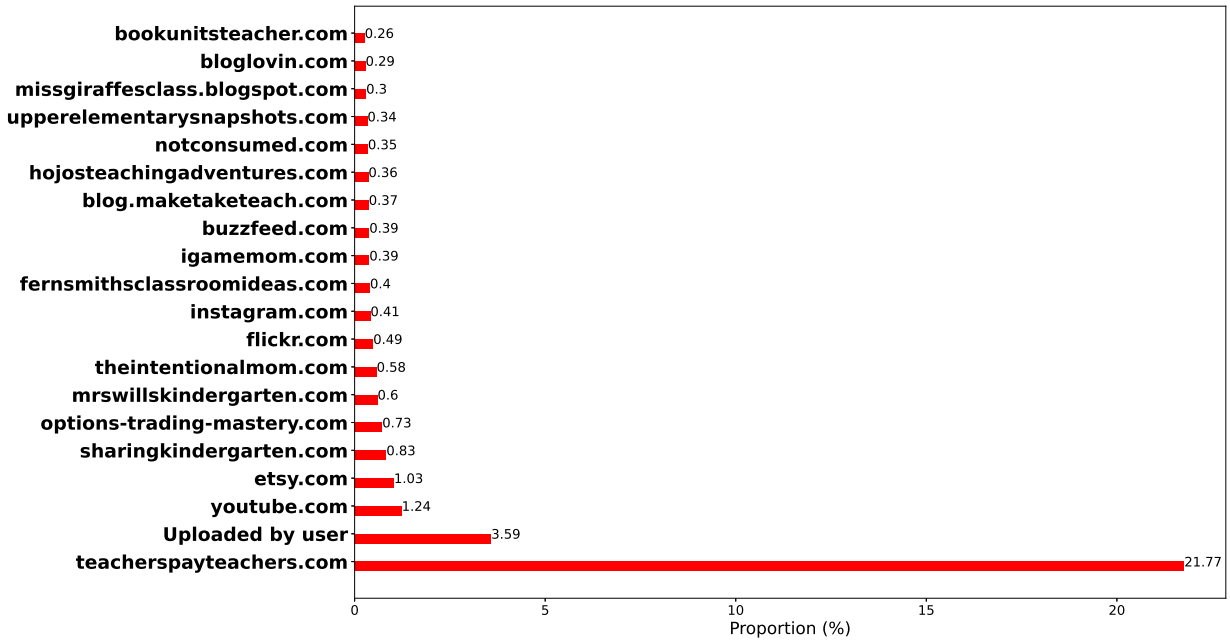


Figure 4.16: The top 20 domains of pins for male teachers.

- Although the predominant domain for both gender groups is *teacherspayteachers.com*, its percentage is significantly higher for male teachers (21.77%) than for female teachers (7.88%). Similar to what we observed for topics of pins in Section 4.3.2, this indicates that male teachers have *focused* more on certain domains while female teachers have *diversified* their attention on different domains.

- Following on the previous point, we can observe from Figure 4.19 that although male teachers have curated resources from various domains, their attention has been fairly concentrated on educational content from these domains. On the contrary, female teachers have explored curating other types of resources from their domains of interest. For instance, 81.7% of pins from *polyvore.com* curated by female teachers are related to women's fashion. In other words, the domain-topic matrix DT^m shown in Figure 4.17 is more *sparse* than DT^f in Figure 4.19. As a simple measure of sparsity, we calculated the number of zero elements in each matrix divided by 680, the number of entries in a matrix, since 20 domains $\times$ 34 topics = 680. The sparsity for DT^m is 0.62, while its value for DT^f is 0.47.

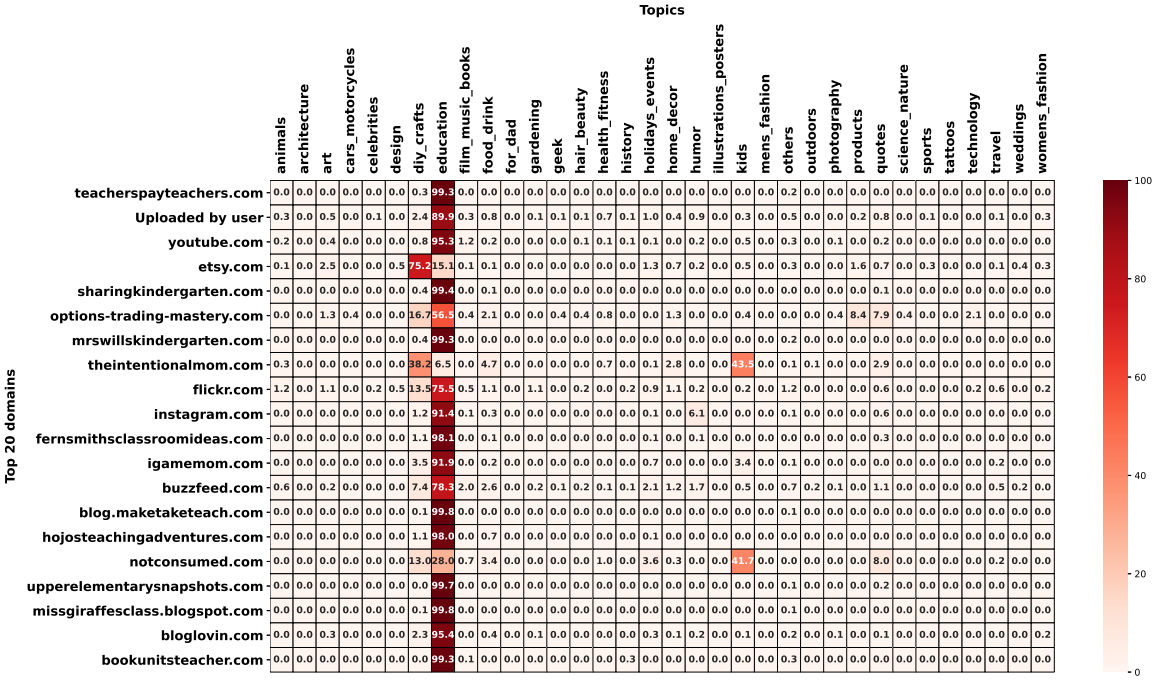


Figure 4.17: The distribution of the top 20 domains for male teachers across topics (DT^m).

- Despite the difference in sparsity between DT^m and DT^f, for both gender groups, the predominant topic of pins coming from the top 20 domains is still *education*– See the column corresponding to *education* in DT^m (Figure 4.17) and DT^f (Figure 4.19). While some of these domains are evidently educational websites such as *missgiraffesclass.blogspot.com*, both male and female teachers have curated a significant number of educational resources from the general-purpose sources as well e.g., *youtube.com*, *amazon.com*. This indicates that, overall, both gender groups seek educational materials from most sources they encounter. Figure 4.20 illustrates an example of an educational pin curated from *youtube.com*.

- A relatively large portion of pins for both gender groups does not have a domain from the web as the user has directly uploaded them, namely 3.59% for male teachers and 5.72% for female teachers.[8] To put these numbers in perspective, we processed the data of non-teachers in our dataset[9] and discovered that there is a tiny percentage (0.23%) of pins whose domain is "Uploaded

[8]Technically, the *domain* for these pins in our dataset is "Uploaded by user"–See Table 2.1.

[9]For non-teachers, we used a similar procedure described in Section 4.2 and considered a user as non-teacher if $p < 0.05$.

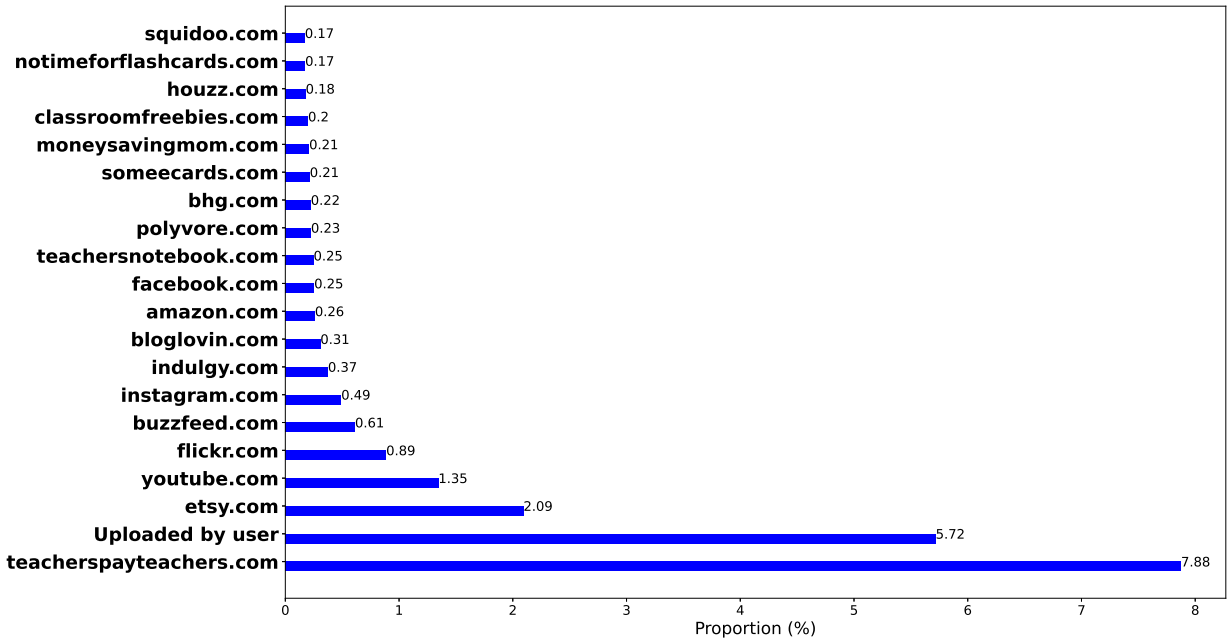


Figure 4.18: The top 20 domains of pins for female teachers.

by user". Furthermore, results presented in Figures 4.17 and 4.19 reveal that the topic of most of these pins is *education* (89.9% and 80.4% for male and female teachers, respectively). Given this, we can state that both male and female teachers actively curate educational resources not only by acquiring them from online sources but by directly creating and sharing them with their peers.

Consistent with Torphy et al. [79], our findings indicate that teachers frequently turn to their fellow teachers online for educational resources and professional materials. We mentioned "fellow teachers" because online educational resources are mainly prepared by other teachers/educators. For example, resources from *teacherspayteachers.com* are primarily curated by teachers themselves. However, compared to Torphy et al. [79], our investigation has three distinct differences. First, we performed the domain analysis of teachers' online resources in a significantly larger scale fashion: ours includes 80,731,782 pins from 13,267 teachers (male and female teachers combined) while [79] included 140,287 pins from 197 teachers. Second, they found out that educator blogs were the predominant sources of pins, followed by "Teacher-to-Teacher Consumption Markets (TTM)" websites. However, according to our findings, TTMs, especially *teacherspayteachers.com* are the

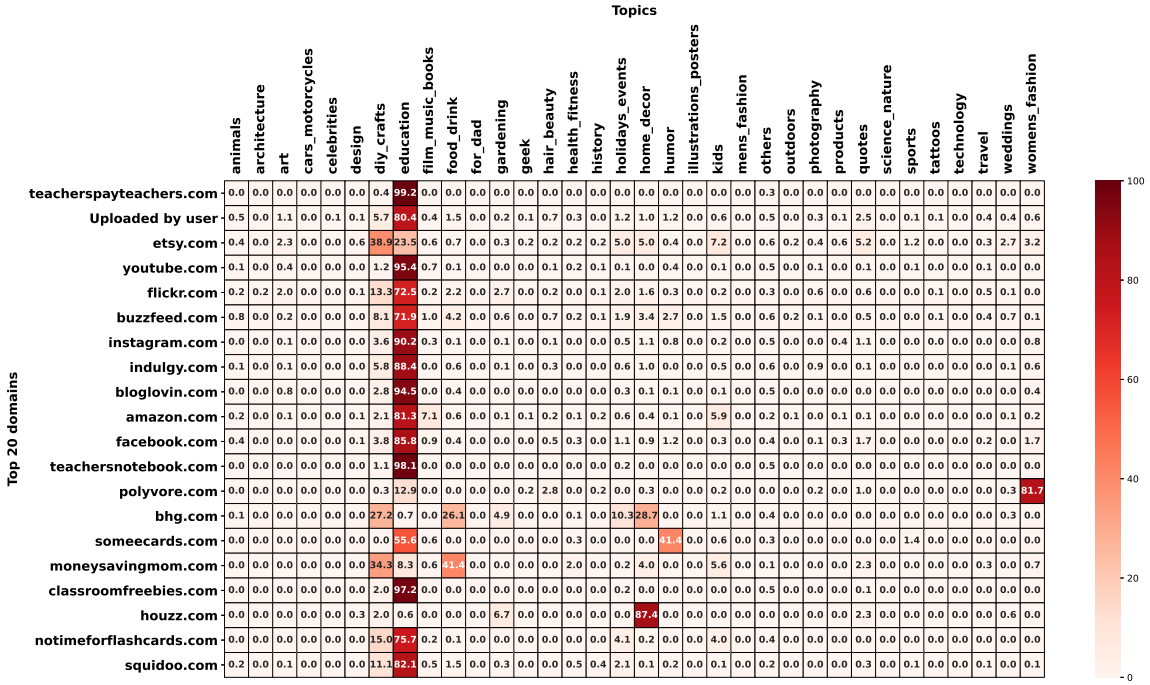


Figure 4.19: The distribution of the top 20 domains for female teachers across topics (DT^f).

predominant sources of pins. According to [79], "educator blogs include independent websites created by individuals or groups of teachers who openly reflect and share their professional values," e.g., *missgiraffesclass.blogspot.com*. Finally, we incorporated the gender of teachers in our domain analysis and demonstrated male and female differences and similarities regarding the sources of their pins.

4.3.3.2 Domain Entropy

Similar to the topic entropy in Section 4.3.2, we also investigated the domain specialization. A completely domain-specialized user gets their pins from a single domain while a less domain-specialized user tries out different domains.[10] Let $\mathcal{D} = [d_1, d_2 \cdots d_k]$ denote the set of k domains. In the dataset used in this chapter, combining the domains of pins of male and female teachers, in total, we have 46,377 unique domains (i.e., k= 46,377). Further, for a given user, let p_{d_i} denote

[10]The term "domain-specialized" is used in the context of this dissertation and the domains of teacher-curated resources on Pinterest. Therefore, it should not be confused with the term *domain expert* (`https://en.wikipedia.org/wiki/Subject-matter_expert`).

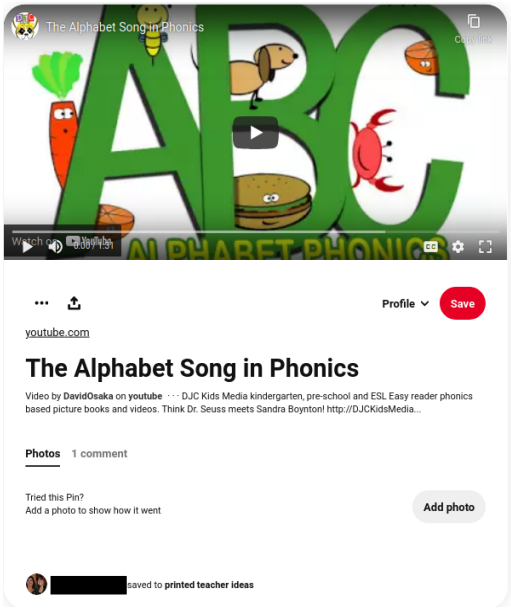


Figure 4.20: An example of an educational pin curated from *youtube.com*.

the fraction of pins whose domain is d_i. Then, we define the *domain entropy* (DE) as follows:

$$DE(u) = -\sum_{\forall d_i \in \mathcal{D}} p_{d_i} \times ln(p_{d_i}) \tag{4.3}$$

Compared to the topic entropy in Eq. 4.1, here we do not account for the number of domains from which a user has at least one pin, i.e., an equivalent of the term $\frac{m}{k}$ in Eq. 4.1 is omitted from the definition of the domain entropy. The reason is we have many unique domains (specifically 46, 377), and such a term would be almost zero and consequently would zero out the domain entropy.

Figure 4.21 shows the CCDF of the domain entropy for male and female teachers. As it can be observed from this figure, female teachers exhibit a smaller degree of domain specialization. Combined with what we found out in Section 4.3.2, male teachers are more specialized in both topic and domain. Since the range of existing domains is significantly larger than existing topics (46, 377 versus 34), the domain entropy has a larger scale than the topic entropy. More specifically, the maximum value for the domain entropy is $ln(46377) = 10.74$ while for the topic entropy the maximum value is $ln(34) = 3.52$.

Similar to Section 4.3.2, here we attempt to determine the relationship between the number of pins and the domain entropy. Figure 4.22 demonstrates the plots of the domain entropy versus the

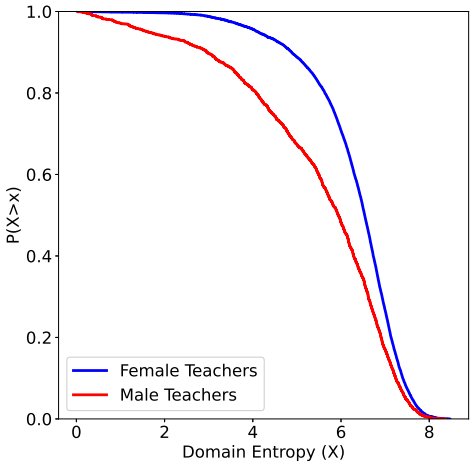
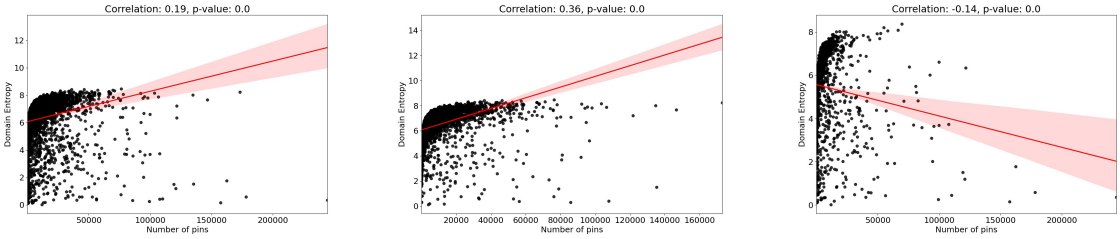


Figure 4.21: The CCDF of the domain entropy (Eq. 4.3).

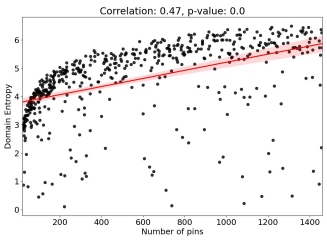


(a) Male & female teachers (b) Female teachers (c) Male teachers

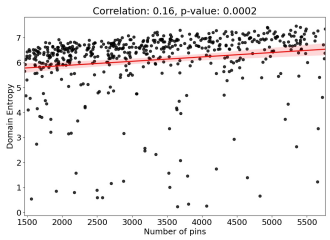
Figure 4.22: The domain entropy based on the number of pins.

number of pins with fitted regression lines for three cases: male and female teachers combined (Figure 4.22a), female teachers only (Figure 4.22b), and male teachers only (Figure 4.22c). The overall trend shown in Figure 4.22a shows that the more pins a user has, the higher the domain entropy is. Nevertheless, this trend is driven by two opposite forces from the data of male and female teachers. More specifically, for female teachers, the higher number of pins results in the higher domain entropy, whereas for the male teachers, the domain entropy decreases when the number of pins increases. This can be further confirmed by the positive (negative) correlation as well as the positive (negative) slope for the fitted regression lines in Figures 4.22b and Figures 4.22c, respectively.

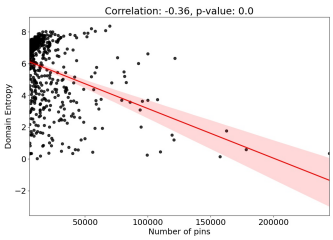
The distinct patterns that emerged from the data of male and female teachers regarding the relationship between the domain entropy and the number of pins motivated us to dig deeper into



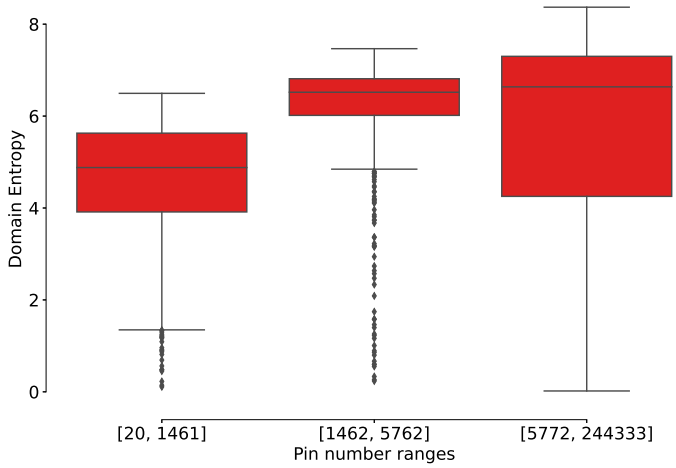
(a) Range $[20, 1461]$



(b) Range $[1462, 5762]$



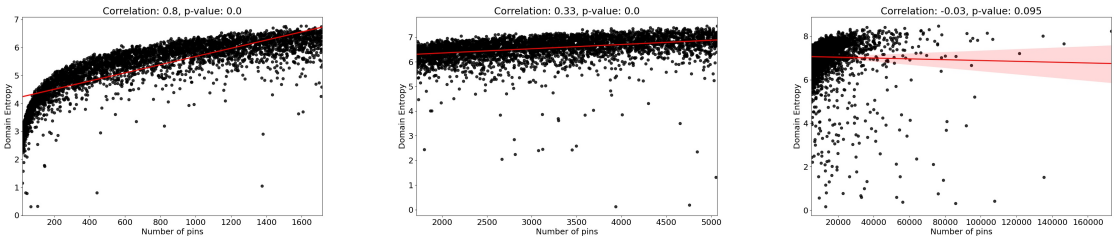
(c) Range $[5762, 244333]$



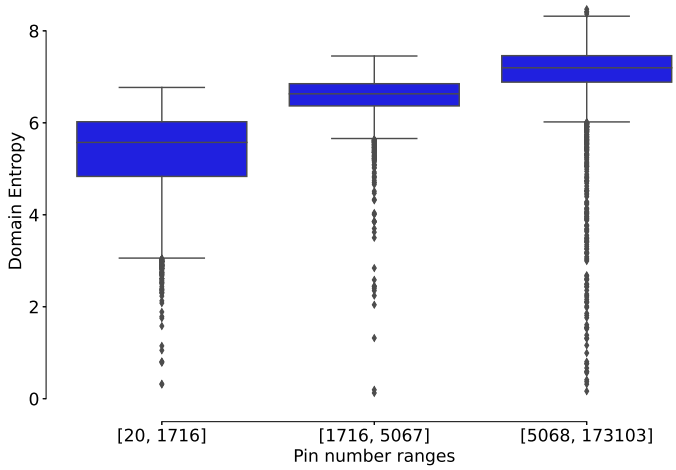
(d) Boxplots of the domain entropy for each range of the number of pins

Figure 4.23: The domain entropy for male teachers across three distinct ranges of the numbers of pins.

this relationship, similar to what we performed in Section 4.3.2. To this end, we investigated this relationship in three ranges of the number of pins. Figure 4.23 and 4.24 show the results for male and female teachers, respectively. For male teachers, we can observe a similar pattern with what we presented in Section 4.3.2. At the first range, demonstrated in Figure 4.23a, the correlation between the number of pins and the domain entropy is positive and relatively large. This correlation remains positive yet drops significantly in the middle range (Figure 4.23b) and eventually becomes negative in the last range (Figure 4.23c). Hence, the more prolific male teachers become, the more they tend not to explore more domains, i.e., they prefer to focus on a smaller set of domains. Regarding the changes from a range to the next one, female teachers exhibit a similar behaviour where the correlation from the first range (Figure 4.24a) to the second range (Figure 4.24b) and eventually



(a) Range $[20, 1716]$ (b) Range $[1716, 5067]$ (c) Range $[5068, 173103]$



(d) Boxplots of the domain entropy for each range of the number of pins

Figure 4.24: The domain entropy for female teachers across three distinct ranges of the numbers of pins.

to the third one (Figure 4.24c) decreases monotonically. Nevertheless, compared to male teachers, there exists a major difference: at each range, the magnitude of the correlation is significantly higher for female teachers, i.e., 0.8 vs. 0.47, 0.33 vs. 0.16, and -0.03 vs. -0.36, respectively.

4.3.3.3 Domain Oscillation

Although teachers might focus on certain domains (i.e., being domain-specialized), they might frequently switch from one domain to another. The domain entropy cannot capture this variation based on the same reasoning discussed for the topic entropy. Hence, similar to the topic oscillation defined in Eq. 4.2, we define the *domain oscillation*. Suppose $S = [a_1, a_2 \cdots a_n]$ denote the chronologically ordered sequence of domains of pins for a given user where $a_i \in \mathcal{D}$. Then the

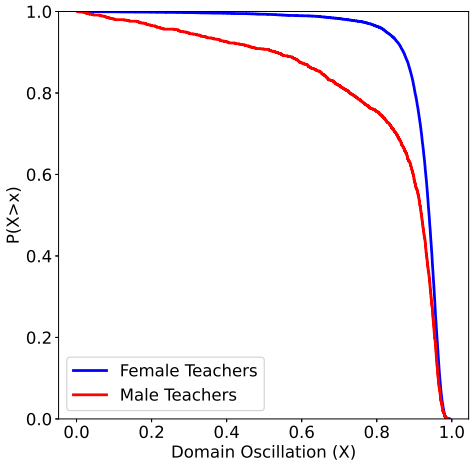


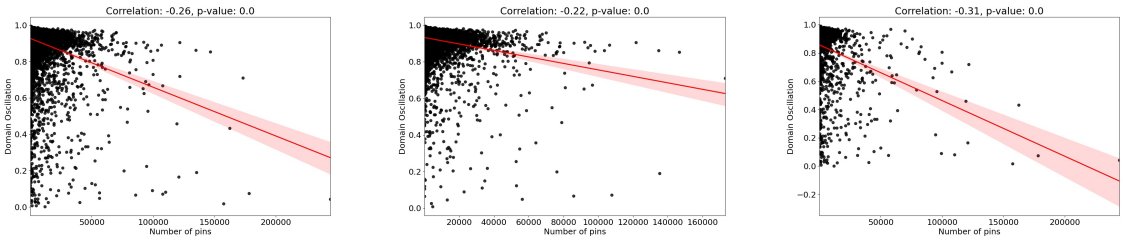
Figure 4.25: The CCDF of the domain oscillation (Eq. 4.4).

domain oscillation (DO) is defined as follows:

$$DO(u) = \frac{1}{n-1} \times \sum_{i=1}^{n-1} \mathbb{1}(a_i = a_{i+1}) \tag{4.4}$$

where $\mathbb{1}$ is the indicator function. The minimum value of $DO(.)$ is 0 and it occurs when the user has pinned resources from a single domain. The maximum value of the $DO(.)$ is 1 and it occurs when the domain has changed for every two consecutive pins.

Figure 4.25 demonstrates the CCDF of the domain oscillation for male and female teachers. As shown in this figure, female teachers exhibit a smaller degree of domain specialization, i.e., overall, they are less domain-specialized.

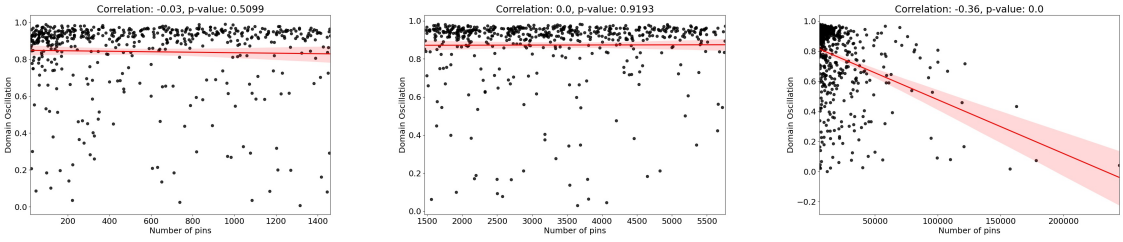


(a) Male & female teachers (b) Female teachers (c) Male teachers

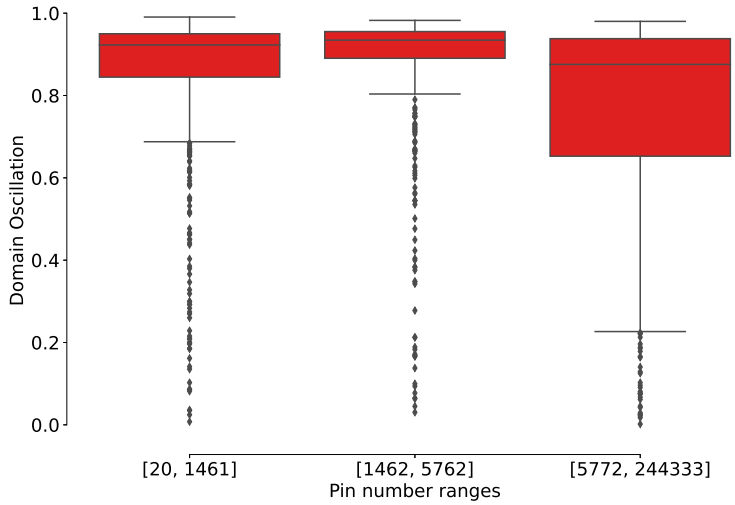
Figure 4.26: The domain oscillation based on the number of pins.

Again, we attempt to look into the domain oscillation for male and female teachers while con-

sidering the number of pins. In other words, does the number of pins is related to the domain oscillation? The same as before, we performed this investigation for three cases as shown in Figure 4.26: male and female teachers combined (Figure 4.26a), female teachers only (Figure 4.26b), male teachers only (Figure 4.26c). Interestingly, for all three cases, the domain oscillation negatively correlates with the number of pins. In other words, the more pins a teacher has, the less they oscillate from a domain to another. Here, unlike the domain entropy, male and female teachers show similar behavior.



(a) Range $[20, 1461]$ (b) Range $[1462, 5762]$ (c) Range $[5762, 244333]$

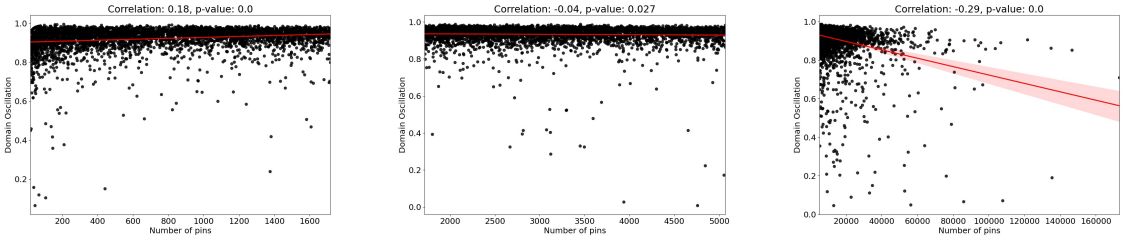


(d) Boxplots of the domain oscillation for each range of the number of pins

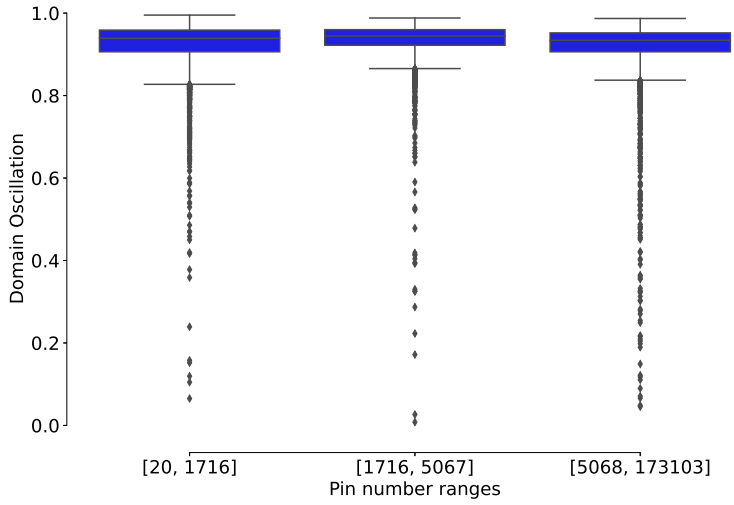
Figure 4.27: The domain oscillation for male teachers across three distinct ranges of the numbers of pins.

To deepen our understating of the relationship between the domain oscillation and the number of pins, again, we discretized the number of pins into three equal-size ranges for male and female teachers and then inspected the relationship. Figure 4.27 shows the results for male teachers. There

is no correlation between the domain oscillation and the number of pins for the first two ranges (i.e., Figures 4.27a and 4.27b). Nevertheless, the p-values for these two ranges are large, and thus they are not statistically significant. The main reason that the correlations in these two ranges are not statistically significant is the presence of a relatively large number of outliers. The outliers are visually visible in Figures 4.27a and 4.27b, which are the data points whose domain oscillation is less than 0.8. Later, we will discuss more about the observed phenomenon in these two ranges for male teachers (i.e., the results shown in Figures 4.27a and 4.27b) and how to make sense of it. However, the story for the third range, shown in Figure 4.27c, is different, where we observe a moderate negative correlation (-0.36), which is also statistically significant. Hence, for prolific male teachers, the more they pin, the less they tend to vary in the domains of the pins.



(a) Range $[20, 1716]$ (b) Range $[1716, 5067]$ (c) Range $[5068, 173103]$



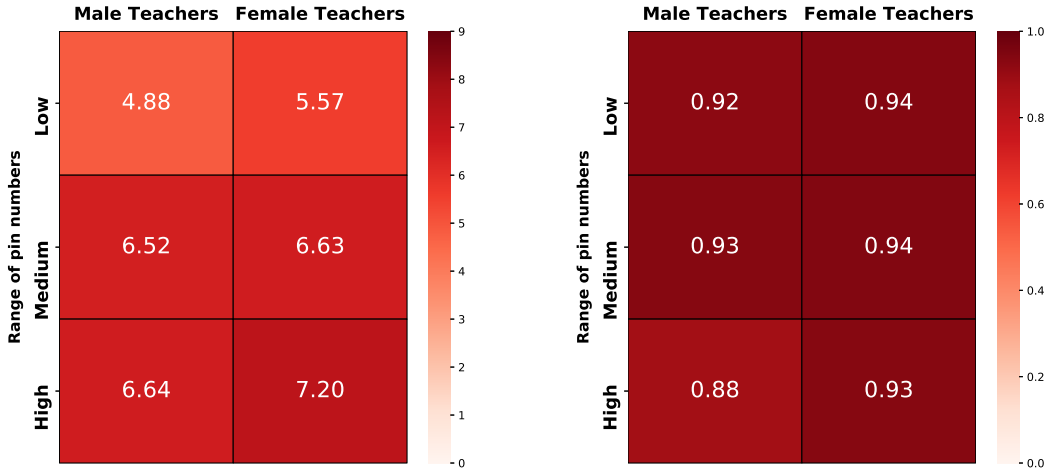
(d) Boxplots of the domain oscillation for each range of the number of pins

Figure 4.28: The domain oscillation for female teachers across three distinct ranges of the numbers of pins.

Figure 4.28 shows the results of the correlation between the number of pins and the domains oscillation across the three defined ranges of the numbers of pins for female teachers. The first range (Figure 4.28a) shows a small positive correlation between the domain oscillation and the number of pins. This correlation becomes almost zero in the second range (Figure 4.28b). Hence, if the number of pins for female teachers is not very large (i.e., the first two ranges in Figure 4.28), the more they pin does not have an impact on their behavior regarding the domain oscillation. For the last range, we observe a similar pattern with that of males teachers: a moderate negative correlation between the domain oscillation and the number of pins. Hence, similarly, for prolific female teachers, the more they pin resources, the less they tend to vary in the domains of pins.

Referring back to the results demonstrated in Figure 4.27, we can discern an interesting similarity between the plots in the first two ranges for male teachers and female teachers i.e., Figure 4.27a with Figure 4.28a and Figure 4.27b with Figure 4.28b. It seems the data in male-related ranges are the sparse versions of female-related ranges, where for male teachers, we only have fewer data points while its pattern is similar to female teachers. This fewer number of data points consequently made the p-value high. Furthermore, in the first two ranges for female teachers, more data points helped obtain statistically significant correlations. Hence, for the first two ranges, male and female teacher data distributions are very similar, and we can generalize our findings for female teachers to male teachers. Based on the above discussion, it is safe to state that when the number of pins for male teachers is not very large (i.e., the first two ranges in Figure 4.27), the more they pin does not have an impact on their behavior regarding the domain oscillation.

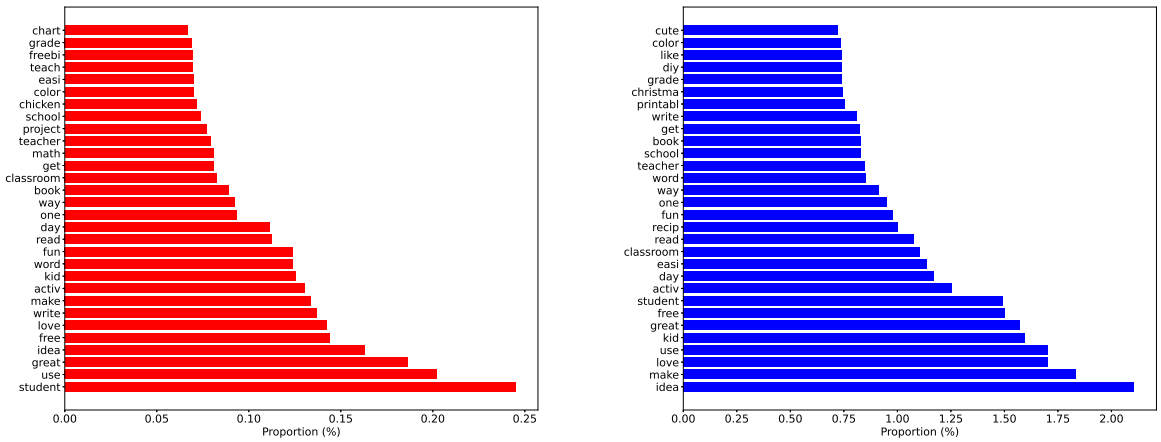
Figure 4.29 summarizes the domain entropy and the domain oscillation values for the three defined ranges of the number of pins. We have labeled the three ranges as *low*, *medium*, and *high*, signifying their relative coverage of the number of pins. In summary, we can conclude that male teachers are more domain-specialized than female teachers and tend to vary less in sources of pins they curate.



(a) The domain entropy in three ranges

(b) The domain oscillation in three ranges

Figure 4.29: A summary of the domain entropy and the domain oscillation for male and female teachers (values are median in ranges).



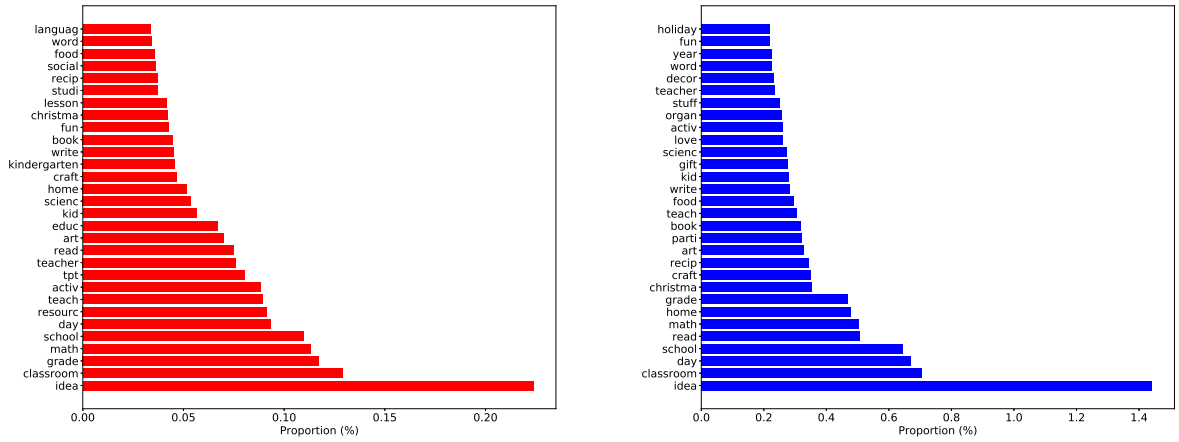
(a) Male teachers

(b) Female teachers

Figure 4.30: The top 30 words of pin descriptions for male and female teachers.

4.3.4 Language of Resources

To further investigate pins and boards curated by male and female teachers, we now look into the top words associated with pins and boards. For pins, we acquired the top words from pin descriptions and for boards from board names. For both of these textual inputs, we used the NLTK package [104]



(a) Male teachers

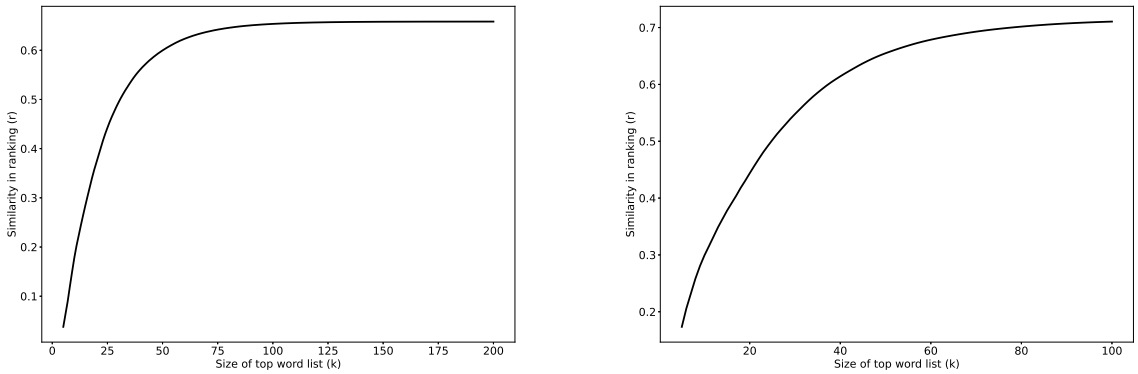
(b) Female teachers

Figure 4.31: The top 30 words of board names for male and female teachers.

and performed appropriate pre-processing, e.g., removing punctuations and stop words (e.g., '!', 'the'), stemming the tokens (e.g., 'education' to 'educ'). As a result, figure 4.30 demonstrates the top 30 words associated with pins curated by male and female teachers. Similarly, Figure 4.31 shows the top 30 words associated with board names for male and female teachers.

According to Figures 4.30 and 4.31, almost all the top words associated with pin descriptions and board names curated by both male and female teachers are related to education and teaching e.g., *student*, *classroom*, *grade*, *school*, *math*. This signifies that both male and female teachers predominately leverage Pinterest to curate resources related to their teaching profession. In other words, in line with previous studies [28, 5, 34], teachers utilize Pinterest for professional purposes. However, our study in this chapter is the first one corroborating this fact for both male and female teachers. Furthermore, similar to our findings in Section 3.5.4, some words like 'fun' and 'idea' are among the top words in the board names and pin descriptions of both gender groups. Again, this demonstrates the distinct way that teachers leverage social media for education-related purposes.

A closer look at Figure 4.30 reveals that 25 out of the top 30 pin-related words are common between males and females. Regarding board-related words, 22 out of the top 30 words are common between the two gender groups. These two observations indicate that male and female teachers



(a) Pin-related words

(b) Board-related words

Figure 4.32: Similarity of the top-k pin-related word lists using Rank-biased Overlap (RBO).

employ a very similar professional vocabulary in their pin and board curation activities. However, the order of the top pin-related (and board-related) words for male teachers is different from that of female teachers. Therefore, to more rigorously compare the ranked list of the top words for male and female teachers, we used Rank-biased Overlap (RBO) [130]. RBO takes two ranked lists as the input and returns a numeric value ($0 \leq r \leq 1$), indicating the similarity between the two lists. The closer r is to 1, the more similar the two ranked lists are. Compared to traditional ranked list comparison methods like Kendall tau[11], the main advantage of RBO is that it can handle non-conjointness, i.e., the items in the two ranked lists do not necessarily need to overlap (for more detail about RBO, refer to [130]). We calculated the similarity between the two ranked lists of words while varying the length of lists. Figures 4.32a and 4.32b demonstrate the result of this experiment for the top words of pins and boards, respectively. When the size of lists is small, the similarity between the male-related ranked list and the female-related ranked list is small. This seems reasonable since for a small list, the two list are very different as can be seen in Figures 4.30 and 4.31. Nevertheless, once we expand the ranked lists, the similarity increases. This indicates that, overall, male and female teachers act similarly in terms of the language of their curated resources.

[11] `https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient`

4.3.5 Resource Curation Over Time

Another aspect of teachers in social media is concerned with *when* teachers participate in online activities. Studying *time* helps us understand how teachers *interact* with space outside of their classroom. This is particularly important since teachers spend a considerable amount of time seeking out educational materials for their pedagogical needs, e.g., up to 12 hours, according to [131]. Given this, characterizing when teachers interact with social media has been investigated in the literature. For instance, Rosenberg et al. [75] considered the conversations around state-level education-related hashtags on Twitter as affinity spaces and further answered when teachers/educators participate in these spaces. They analyzed, for example, the percentage of participants per day of the week, which helped understand the unique ways that their sampled teachers are using Twitter for professional purposes. We have performed similar analyses, as will be explained in this part. In another relevant study, Greenhalgh and Koehler [126] characterized the timing around #educattentats, a hashtag about Paris terrorist attacks that occurred in November 2015. #educattentats attempted to organize teachers on how to discuss the attacks with their students.

Our investigation in this part has several unique characteristics. First, the social media data in the previous studies are associated with a specific scenario or project that gathers teachers, e.g., specific hashtags [126, 75]. Nevertheless, our data reflects the entire timeline of thousands of teachers on Pinterest from the day they joined until November 2019. Given this, we have a better picture of teachers' interaction with social media. Second, most previous studies looked into the interaction of teachers with social media through inter-teacher online conversations, e.g., tweeting about a specific topic or posting in a specific teacher-related Facebook group. This has its own merits e.g., helping to understand teacher professional development process [47, 46, 48, 49]. However, we believe approaching the temporal analysis of the interaction of teachers with social media through the lens of their curated resources is closely related to teachers' classroom pedagogical activities and, thus, educationally carries more weight. Third, for the first time, we have incorporated the gender of teachers in the temporal analysis of teachers' interaction with social media.

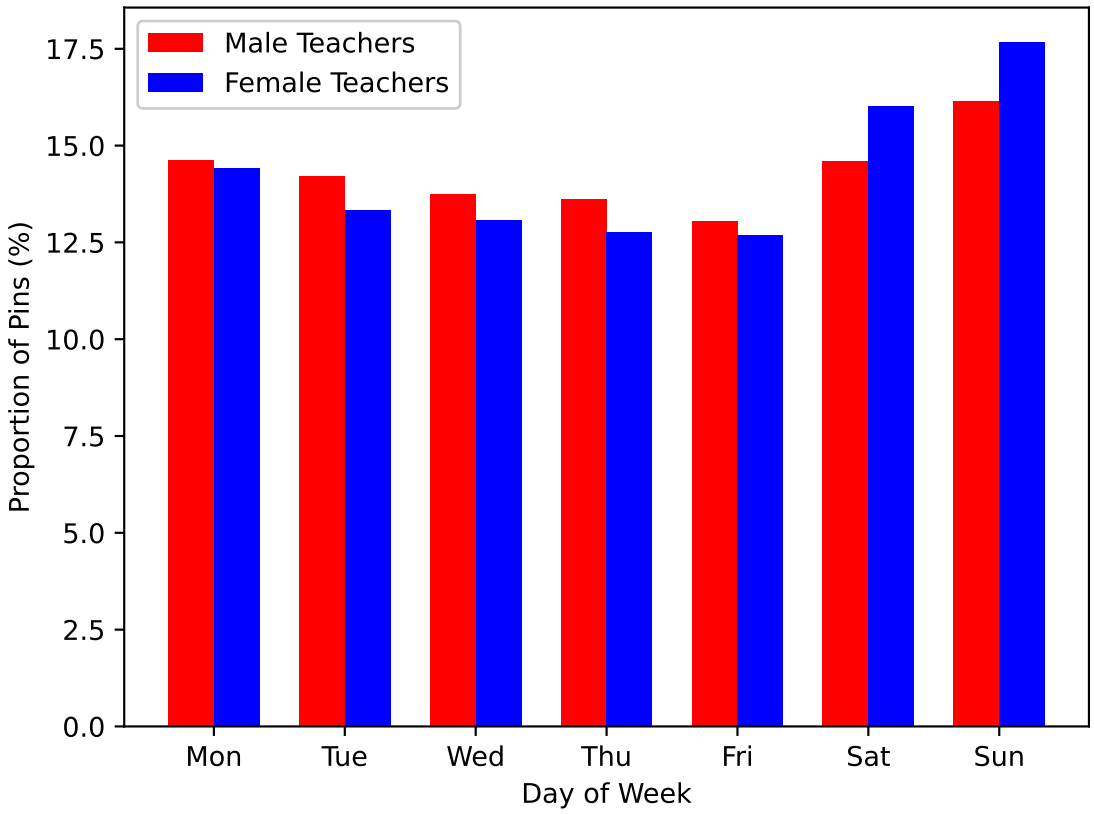


Figure 4.33: The average percentage of pin curations on each day of the week for male and female teachers.

4.3.5.1 Days of the Week

Figure 4.33 demonstrates the average percentage of pins curated on each day of the week for male and female teachers. Similarly, Figure 4.34 demonstrates the average percentage of boards curated on each day of the week for male and female teachers. Regarding the gender of teachers, there is no significant difference between male and female teachers, neither for pins nor for boards. However, during the weekends, female teachers tend to become slightly more active in pin and board curation than male teachers. Moreover, interestingly, distributions of pin and board curations are very similar. This suggests that teachers have a specific weekly schedule to utilize social media for resource curation. Another crucial observation from Figures 4.33 and 4.34 is that during weekends, teachers (both males and females) are more active. Considering being busy at

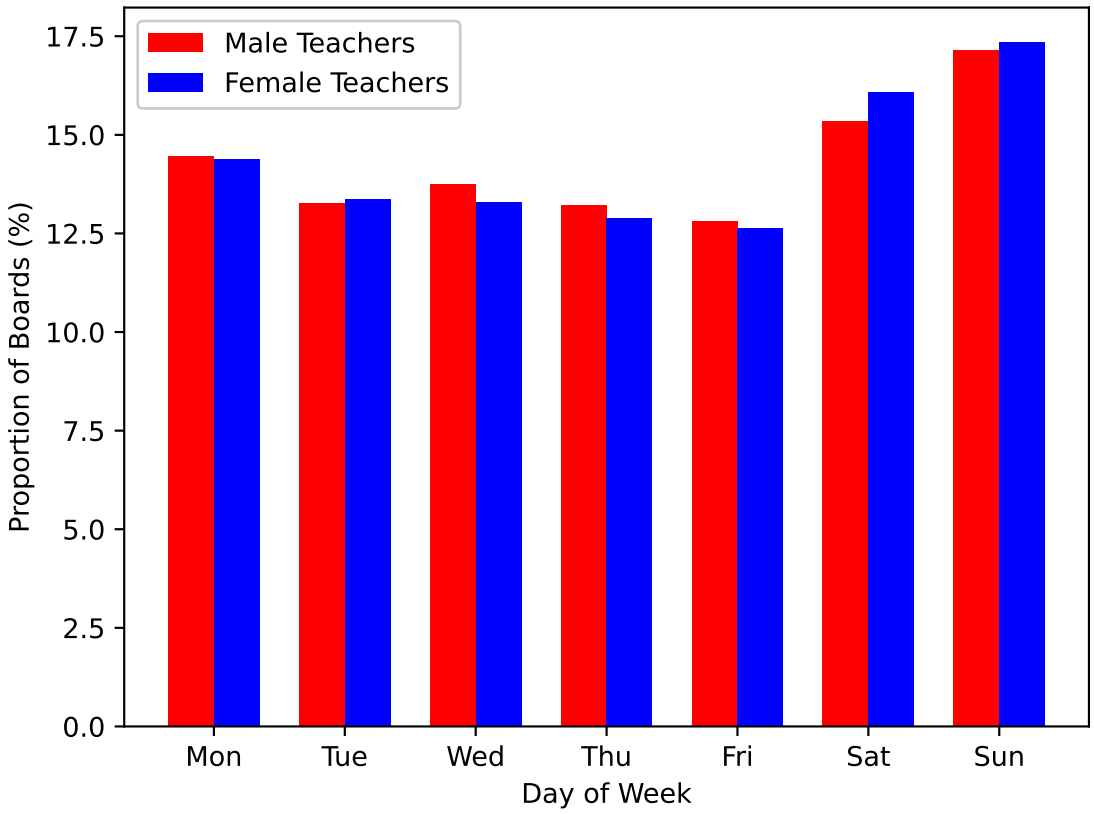


Figure 4.34: The average percentage of board curations on each day of the week for male and female teachers.

school during weekdays, it seems logical that teachers allocate more time on their weekends to use Pinterest. This finding is especially outstanding since the higher activity of teachers on Pinterest during the weekends is in contrast with the overall usage pattern of Pinterest, where weekends have the least amount of traffic [132]. Similarly, Rosenberg et al. [75] found a distinct weekly pattern of tweeting for teachers on Twitter, which happened to be at odds with the overall Twitter usage pattern.

4.3.5.2 Months of the Year

Continuing the temporal analysis of pin and board curations, in this part, we look into months of the year (i.e., January, February, $\cdots$, December). Figure 4.35 demonstrates the average percentage of

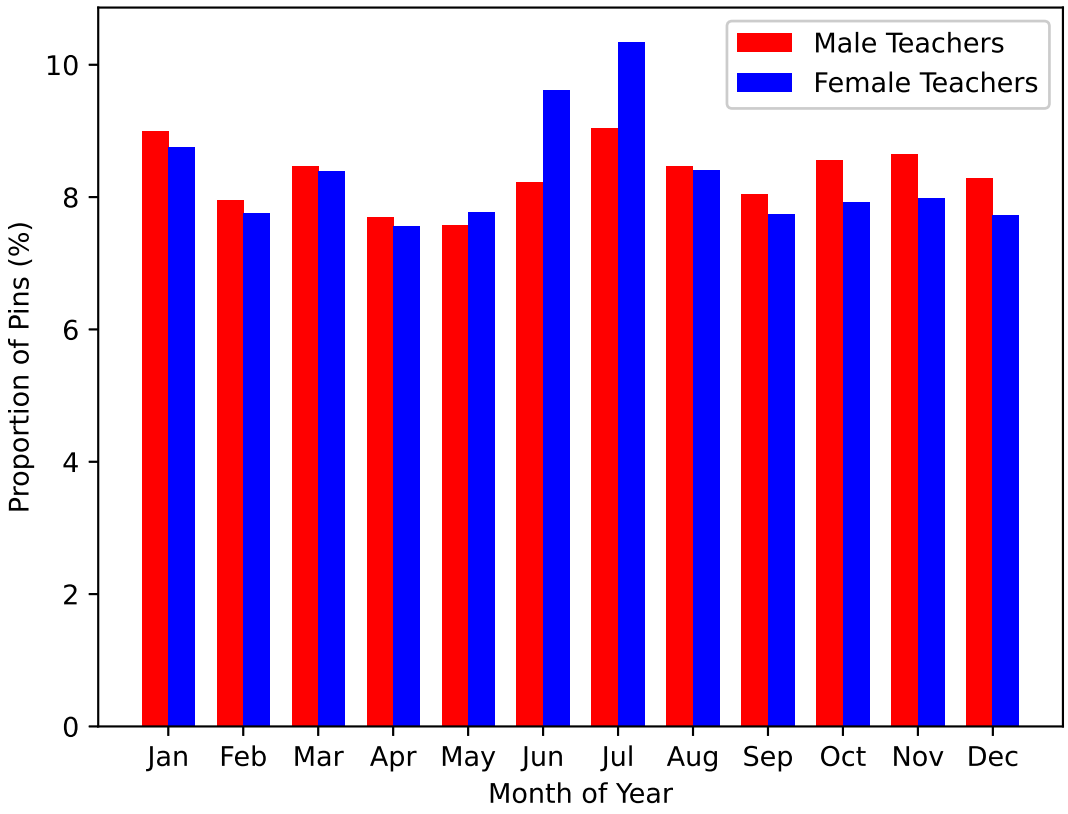


Figure 4.35: The average percentage of pin curations in each month of the year for male and female teachers.

pin curations in each month of the year for male and female teachers. Similarly, Figure 4.36 shows this for board curations. According to these figures, there is no significant difference between male and female teachers concerning the number of curations in months of the year. An interesting observation is the higher curation activity in the summer months, namely June, July, and August. In this period of the year, teachers have more free time, and they can curate more educational materials and thus prepare themselves for the next teaching semester in August/September. Another pattern is a relatively high degree of board curations in January. Note that boards are essentially organizational folders representing teachers' professional perspectives regarding pins worth saving and sharing [79]. Given this, perhaps at the beginning of the year, teachers start to create more boards to collect resources for the rest of the year.

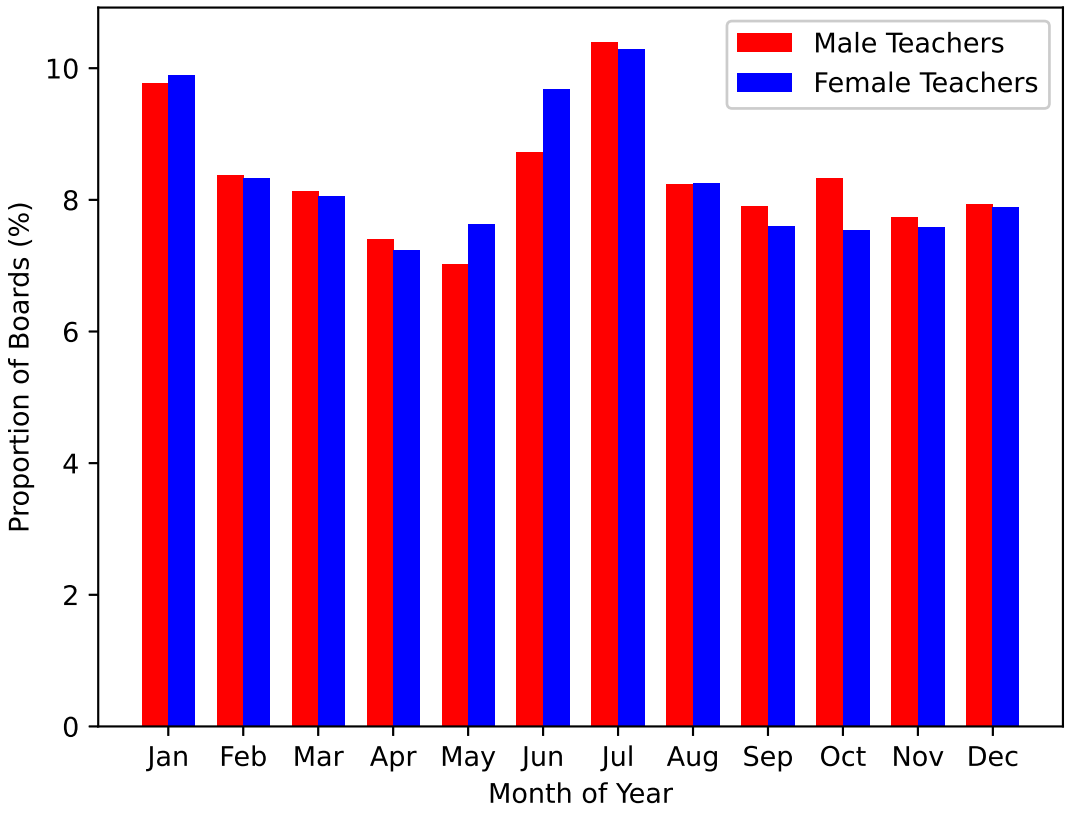


Figure 4.36: The average percentage of board curations in each month of the year for male and female teachers.

4.3.5.3 Days of the Month

Finally, in this part, we investigate the curation pattern for each day of the month, i.e., 1, 2, $\cdots$ 31. Figure 4.37 demonstrates the average percentage of pin curations on each day of the month for male and female teachers. Similarly, Figure 4.38 shows this for board curations. Since in the Gregorian calendar, the months are either 28, 29, 30, or 31 days long, there is less amount of activity for days 29, 30, and 31. For the rest of the days, we can observe that, on average, teachers perform resource curation across all days without any significant variation. However, the numbers for male teachers exhibit more variation than for female teachers. We believe the reason is the artifact of the underrepresentation of male teachers, where we might not have enough data to populate each day as much as we do for female teachers.

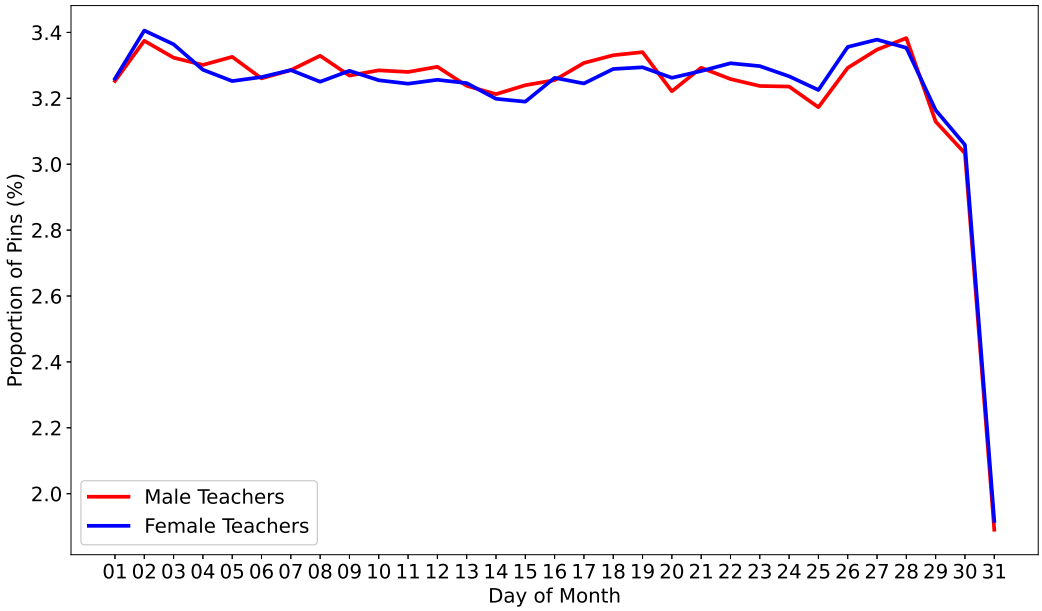


Figure 4.37: The average percentage of pin curations in each day of the month for male and female teachers.

In conclusion, our large-scale temporal investigations in Sections 4.3.5.1, 4.3.5.2, and 4.3.5.3 indicates three crucial findings:

- Teachers are committed to using social media for educational purposes. This commitment is especially notable since our analysis showed that they spend time from their leisure to interact with social media (here Pinterest), e.g., over the weekends.
- Teachers' usage of social media is persistent and perpetual, as demonstrated in previous studies as well [4, 92].
- Teachers shape their unique pattern of using social media compatible with their teaching schedule and the school-year calendar.

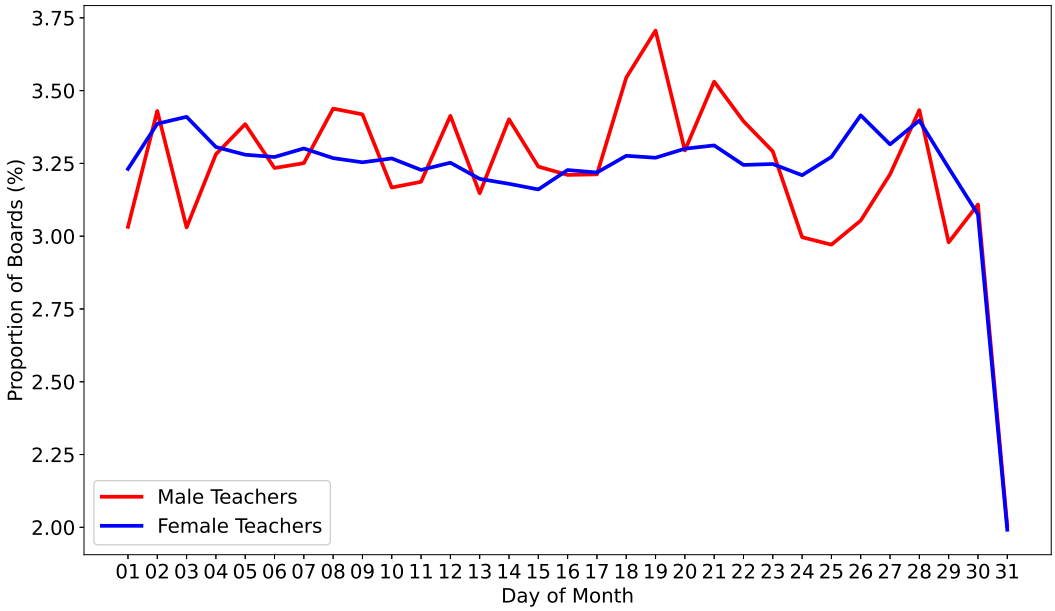


Figure 4.38: The average percentage of board curations in each day of the month for male and female teachers.

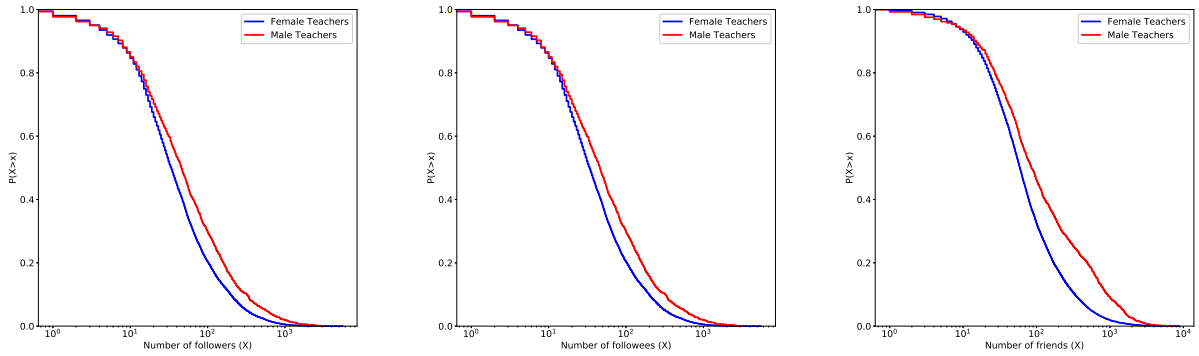
4.4 Social Network Analysis

In Chapter 1, we discussed the importance of teachers in social media and how they improve education. In particular, online social media helps teachers share information and resources. These resources can be diffused rapidly to them classroom. What has made this rapid diffusion possible is the *network* formed among teachers. Within this network, teachers form ties (connections) and communities (or socialized knowledge communities [78, 133]) to exchange knowledge. Not only online networks but also traditional school-level teacher networks have been shown to provide educational opportunities [25, 134]. Hence, to better understand teachers in social media, we need to investigate the network that embeds teachers. To this end, social network analysis offers us analytical approaches to study the network. Social network analysis is widely used in various areas [102, 135, 136, 137, 138]. In particular, it has been used an analytical framework to study teachers in social media [6, 139, 140, 141, 8, 142].

In this section, we utilize social network analysis and study our Pinterest network from multiple

perspectives. While performing the network/graph analysis, we consider the gender of teachers as well. In Section 4.4.1, we look into the distribution of online connections for male and female teachers. Then, in Section 4.4.1, we investigate centrality measures. Eventually, In Section 4.4.3, we investigate the gender homophily among male and female teachers.

4.4.1 Distribution of Connections

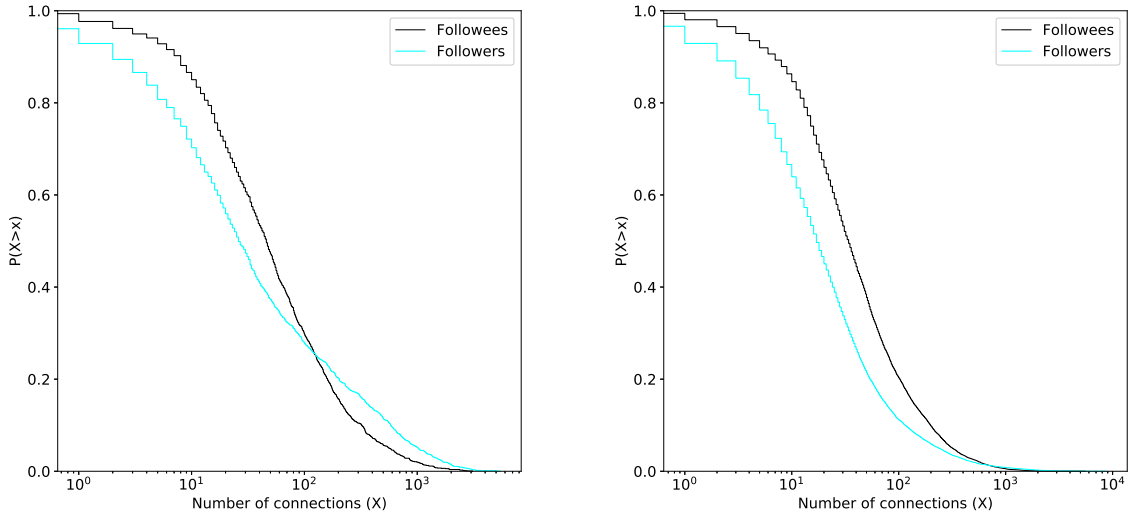


(a) #Followers (b) #Followees (c) #Followers + #Followees

Figure 4.39: The CCDF of number of connections for male and female teachers. x-axes are in log-scale.

Figures 4.39a, 4.39b, 4.39c demonstrate the CCDF of the number of followers, the number of followees, and the total number of friends (i.e., the followers and followees combined), respectively. As it can be observed, when the number of followers and the number of followees are small, the distributions are almost identical for male and female teachers. Nevertheless, the probability of having many followers/followees is slightly more for male teachers than females. Also, in general, male teachers tend to have more friends.

To go deeper into the distribution of the number of connections, Figure 4.40 shows the CCDF of the number of followers and the number of followees for each gender group separately. First, regarding the distributions of the number of followers vs. the number of followees, we can observe from Figure 4.40b that, in general, female teachers have more followees than followers. According to the CCDF curve, it means: $\forall\, n \quad P(\#followers > n) \leq P(\#followees > n)$. We face a different



(a) Male teachers

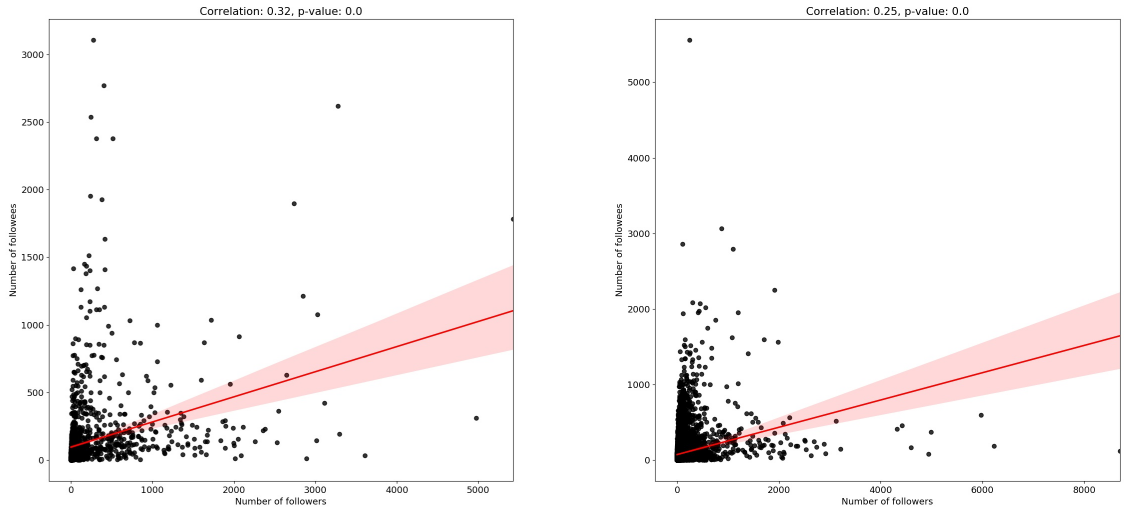
(b) Female teachers

Figure 4.40: The CCDF of number of connections based on their types for each gender group separately. x-axes are in log-scale.

scenario for male teachers demonstrated in Figure 4.40a. We see the same pattern when the numbers of followers and followeess are low: male teachers have more followees than followers. This technically means $\forall\ n \in [1, n_1] \quad P(\#followers > n) < P(\#followees > n)$ where n_1 in our data is 126. However, when the number of followers and followees increases, the chance of having more followers is higher than followees, i.e., $\forall\ n > n_1 \quad P(\#followers > n) > P(\#followees > n)$. Hence, we can conclude that when male teachers expand their networking, they can have more people follow them.

Is there any correlation between the number of followers and followees? Furthermore, how do male teachers and female teachers differ regarding that correlation? To answer these questions, Figure 4.41 illustrates the regression plot between the number of followers and the number of followees for male and female teachers. For both gender groups, the correlation is positive, not very large, however. This means that more number of followees leads to more number followers.[12] This correlation is higher for male teachers than for female ones. Another question is, does the

[12]The reason we mentioned the causality from followee to follower stems from the usual online user behavior where they follow more people to get more followers, not the other way around.



(a) Male teachers (b) Female teachers

Figure 4.41: Regression plots of the number and the number of followees.

increase in the number of followers come from those who have already been followed? More specifically, are friendships reciprocal in a way that if a user (teacher) follows someone, that person follows the user back? To answer this question, we define *reciprocity* for a user as follows:

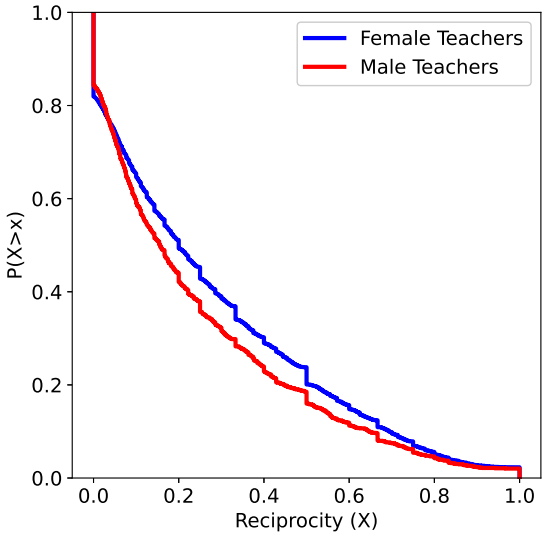


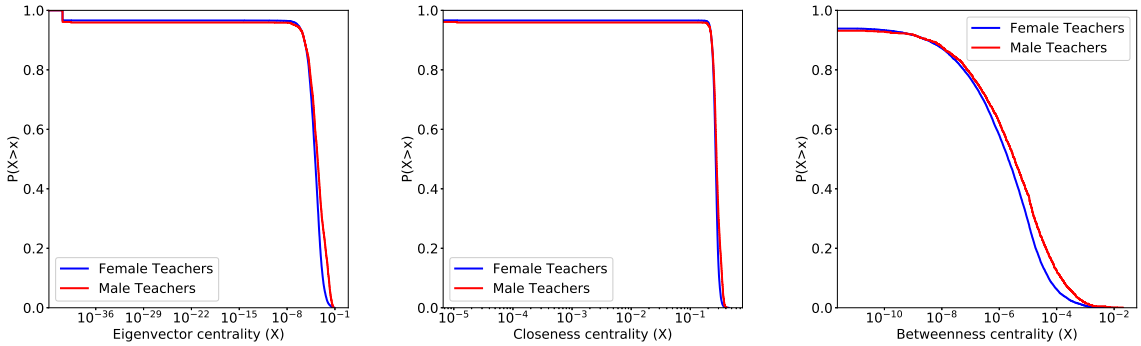
Figure 4.42: The CCDF of the reciprocity for male and female teachers.

$$reciprocity(u) = \frac{|FL(u) \cap FO(u)|}{|FO(u)|} \tag{4.5}$$

where $FL(u)$ denotes the list of followers for user u and $FO(u)$ denotes the list of followees. Essentially, $reciprocity(u)$ determines out of those whom u follows what fraction have followed them back. Note that $reciprocity(u) \in [0, 1]$. Figure 4.42 demonstrates the CCDF of reciprocity index for male and female teachers. We can observe that, in general, male and female teachers have a close reciprocity value. The reciprocity of female teachers is slightly higher, though (the average values of reciprocity for male and female teachers are 0.24 and 0.27, respectively). A previous study on some general users on Pinterest found similar results regarding the reciprocity of males and females [125]. Furthermore, to put the reciprocity of male and female teachers in perspective, the average value of reciprocity for non-teachers in our dataset is 0.35, which is not very high either. Hence, overall, both male and female teachers do not have a high reciprocity value. We speculate this is related to the social curation nature of Pinterest, where users might not feel "obligated" to follow back since the formation of friendship in order to receive information is not as crucial as other social media platforms such as Facebook and Twitter. We will leave further investigation about the exact reasons behind the relatively low reciprocity on Pinterest, especially for teachers, in the future.

4.4.2 Centrality

Centrality is one of the most important notions in social network analysis. It essentially assigns a node a number based on their position in the network. Centrality has many applications in various networks [143, 144, 145]. As will be discussed in the next chapter, centrality significantly influences the diffusion of information. Hence, as part of our investigation of male and female teachers on Pinterest, it is essential to compare their centrality. To this end, we computed three notable measures of centrality, namely eigenvector centrality, closeness centrality, and betweenness centrality. Eigenvector centrality assigns a high score to a node if connected to other high central nodes. Closeness centrality is the average length of the shortest path between the node and all other



(a) Eigenvector centrality (b) Closeness centrality (c) Betweenness centrality

Figure 4.43: The CCDF of centrality measures for male and female teachers. x-axes are in log scale.

nodes in the graph. Thus the more central a node is, the closer it is to all other nodes. Finally, betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two nodes. Figure 4.43 demonstrates the CCDF of these three centrality measures for male and female teachers. As it can be observed from this figure, both gender groups enjoy very similar values of centrality. However, the betweenness centrality is slightly higher for male teachers. Regardless of the gender, we can observe that centrality measures follow a power-law distribution where most nodes have low centrality while a small percentage has massive scores. In conclusion, based on our findings, the positional importance of teachers on Pinterest is not driven by their gender.

4.4.3 Gender Homophily

Homophily is a notion that individuals with similar personal or social traits tend to have a relationship with each other [146], as eloquently put in the famous proverb "birds of a feather flock together". Homophily-driven relationships and interactions based on different attributes such as race, gender, religion, and education level have long been identified and studies in the sociology literature [147]. With the advent of online social media platforms and the formation of friendships/ties on these platforms, homophilic behaviors have been identified and measures on these networks [148, 149, 150]. Some studies have pointed out positive impact of homophily e.g., improving coordination [151],

enhancing tolerance and cooperation [152], formation of social norms [153], better diffusion of information [154, 155]. However, it has been shown that homophily causes negative effects, e.g., political polarization [150], reducing diversity and negatively impacting minorities [156, 157].

Regarding teachers in social media, as explained by Frank et al. [25], social media networks provide a great potential to reduce differences among teachers by diffusion of information and exchange of social capital. Nevertheless, because of the homophilic behavior of teachers (like other human beings), such potential can be disrupted [25]. To devise effective measures to prevent or at least mitigate this disruption, the first and essential task is to understand and characterize homophily among teachers in social media. Therefore, in this part of the dissertation, we analyze gender homophily among our identified teachers on Pinterest .

We perform our homophily analysis through the lens of dyads and triads in the network. A dyad is a pairwise relationship between two individuals, which is the basic structure within a network and the core of any "intersubjective relationship" [158]. A triad or triangle is the relationship between three individuals, which acts as the building block of social order and society [159]. Dyadic and triadic relationships play an essential role in classic sociology [160]. Hence, studying gender homophily through dyadic and triadic relationships offers a better picture of homophilic structure in the network.

To characterize dyads, we recognize seven types of such relationships as demonstrated in Figure 4.44. Each circle denotes either a male or female teacher. We have three types of male-female relationships: a male follows a female, a female follows a male, both follow each other (reciprocal), which, respectively, are denoted as Type 1, Type 2, and Type 3 in Figure 4.44. These three types are non-homophilic relationships since connections have been established between the opposite genders. Type 4 denotes a female teacher follows another female teacher and Type 5 represents a bidirectional relationship between two female teachers. These two types are homophilic relationships since they involve only the same gender. Similarly, we define Type 6 and Type 7 for male-only relationships as illustrated in Figure 4.44.

Figure 4.45a demonstrates the proportion of each type in our dataset. Since the number of

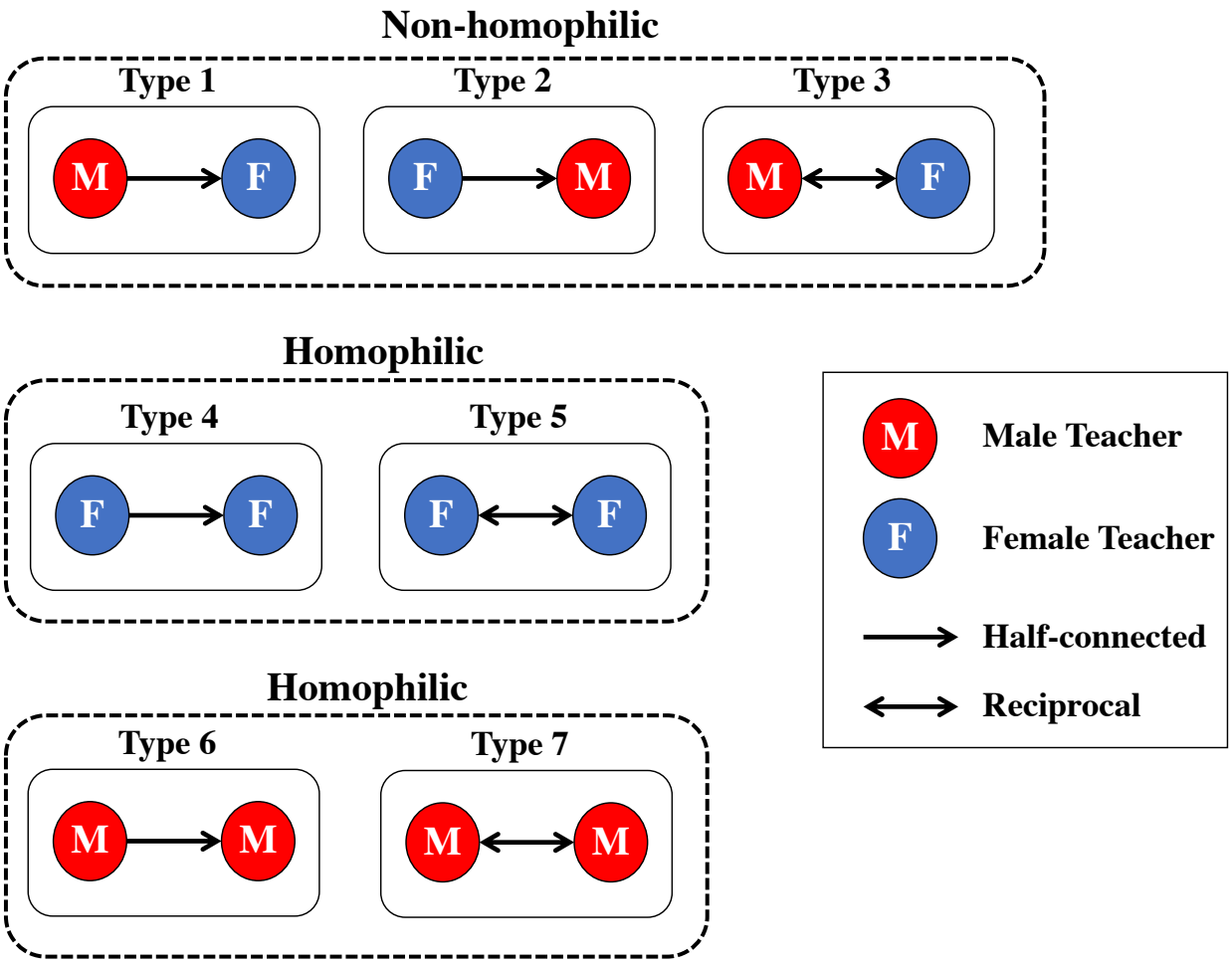
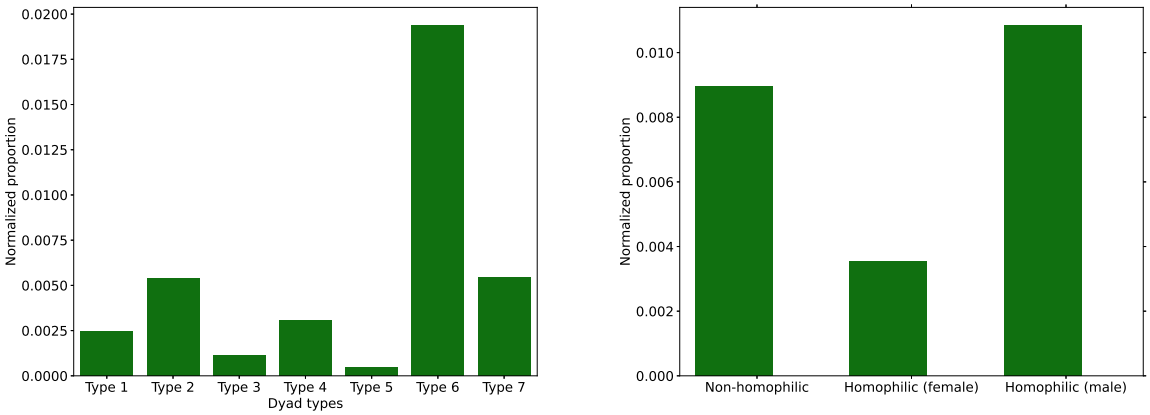


Figure 4.44: Dyad types.

male teachers and female teachers is not equal, we normalized each type and divided it by the total number of such types that could potentially exist. More specifically, the number of relationships of Type 1, Type 2, and Type 3 are divided by $\binom{11675}{1} \times \binom{1592}{1}$ since we have 11675 female teachers and 1592 male teachers–See Table 4.1. The number of relationships of Type 4 and Type 5 are divided by $\binom{11675}{2}$. Finally, for Type 6 and Type 7, the normalization factor is $\binom{1592}{2}$. As it can be observed from Figure 4.45a, Type 6 has the highest value. One might wonder that this is due to the artifact of the lower normalization factor for Type 6 since the number of male teachers is significantly smaller than female teachers. Nevertheless, Type 7 has the same factor while its proportion is smaller than Type 6. The high value for the percentage of dyads of Type 6 indicates a high degree of homophily between male teachers. We speculate this behavior is due to the low representation of male teachers



(a) Proportion of different types of dyads

(b) Homophilic and non-homophilic dyads

Figure 4.45: Gender homophily in dyadic relationships.

on Pinterest, which encourages them to seek each other actively on this platform. Figure 4.45b aggregates the results in Figure 4.45a based on homophilic and non-homophilic relationships. We can observe that homophilic relationships prevail over non-homophilic ones—nonetheless, there is a considerable percentage of non-homophilic relationships between male and female teachers.

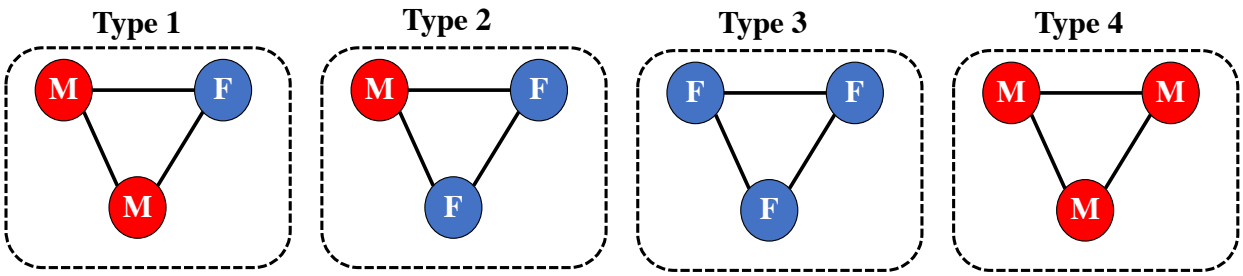
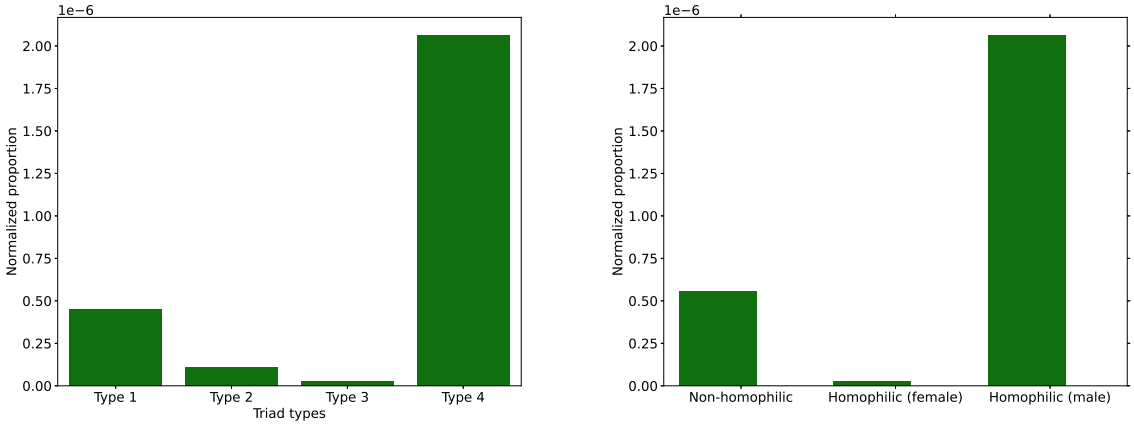


Figure 4.46: Triad types.

In addition to dyads, we also investigated triads. Similarly, we identified four types of triads as shown in Figure 4.46. Type 1 denotes a relationship between two male teachers and a female one. Type 2 denotes a relationship between two female teachers and a male teacher. Type 3 and Type 4 are all-female and all-male relationships, respectively. We normalized the number of occurrences of Type 1, Type 2, Type 3, and Type 4 through dividing them by $\binom{11675}{1} \times \binom{1592}{2}$, $\binom{11675}{2} \times \binom{1592}{1}$, $\binom{11675}{3}$, and $\binom{1592}{3}$, respectively. Figure 4.47a demonstrates the proportion of different types of



(a) Proportion of different types of triads.

(b) Homophilic vs. homophilic in triads.

Figure 4.47: Gender homophily in triadic relationships.

triads in our network. As can be observed from this figure, all-male triadic relationships are the most common type of relationship. Interestingly, the proportion of Type 1 is higher than Type 2, which means while two males establish homophilic relationships, their higher-order structure is less homophilic. Figure 4.47b shows the proportion of homophilic vs non-homophilic triadic relationships. Based on this figure, homophilic relationships prevail over non-homophilic ones.

In conclusion, our empirical evaluation shows that both in dyadic and triadic relationships, homophily between teachers exists. Nevertheless, non-homophilic relationships have also been established. Knowing the fact that more male teachers are being introduced in K-12 education [119, 120], it would be interesting to investigate whether the same homophilic patterns persist or not in the future.

CHAPTER 5

DIFFUSION OF TEACHER-CURATED RESOURCES ON SOCIAL MEDIA

As mentioned before, teachers' primary motivation to join online social media, especially Pinterest, is to curate resources for their pedagogical activities. These resources come from other teachers who have joined social media and established ties with their peers. These ties promote *diffusion* of information on social media and help teachers access an extensive collection of educational resources [45]. Given this, what makes social media very efficient for educational resource curation, unlike, say, asking a colleague, is the fast and widespread diffusion of resources across the network. Moreover, the fast and widespread diffusion of resources allow teachers to cross the traditional school-level boundaries and facilitate large-scale collaboration. Hence, it is of great importance to study the diffusion of teacher-curated resources on social media. Nevertheless, two challenges need to be addressed. The first challenge is concerned with the data. We need to construct the entire diffusion process of a sufficiently large sample of resources so we can perform an effective data-driven study. This process should entail several key elements about the diffusion, including the teacher who has initially curated a resource (i.e., the *producer*), other users *who* have further re-pinned (adopted) the resource, and the *time* when the resource has been re-pinned by a user. Essentially, we need to construct the diffusion tree of a resource, as will be explained in Section 5.1. Figure 5.1 illustrates an example of diffusion tree. Unfortunately, a large-scale dataset of diffusion trees is missing in current studies. Hence, to fill this gap, we construct the diffusion trees of more than 1 million teacher-curated resources on Pinterest. The second challenge is how we can characterize the diffusion process. More specifically, we need to have some measures quantifying the dynamics of diffusion. To this end, we introduce three crucial measures about the diffusion of resources on Pinterest. These measures consider several key aspects of the diffusion of information on Pinterest, namely the number of users who have received a teacher-curated resource, the popularity of a resource, and how fast a resource has been diffused. Using these measures, we perform a large-scale analysis of the diffusion of teacher-curated resources and answer two crucial

research questions: a) do different resources (e.g., in terms of their topics) affect the diffusion?, and b) how teacher attributes (e.g., the number of followers) affect the diffusion?

The remainder of this chapter is organized as follows. First, in Section 5.1, we discuss the constructed dataset of diffusion trees. Then, in Section 5.2, we introduce three measures to characterize the diffusion. Finally, in Section 5.3, we present the results of our diffusion analysis and answer the two research questions.

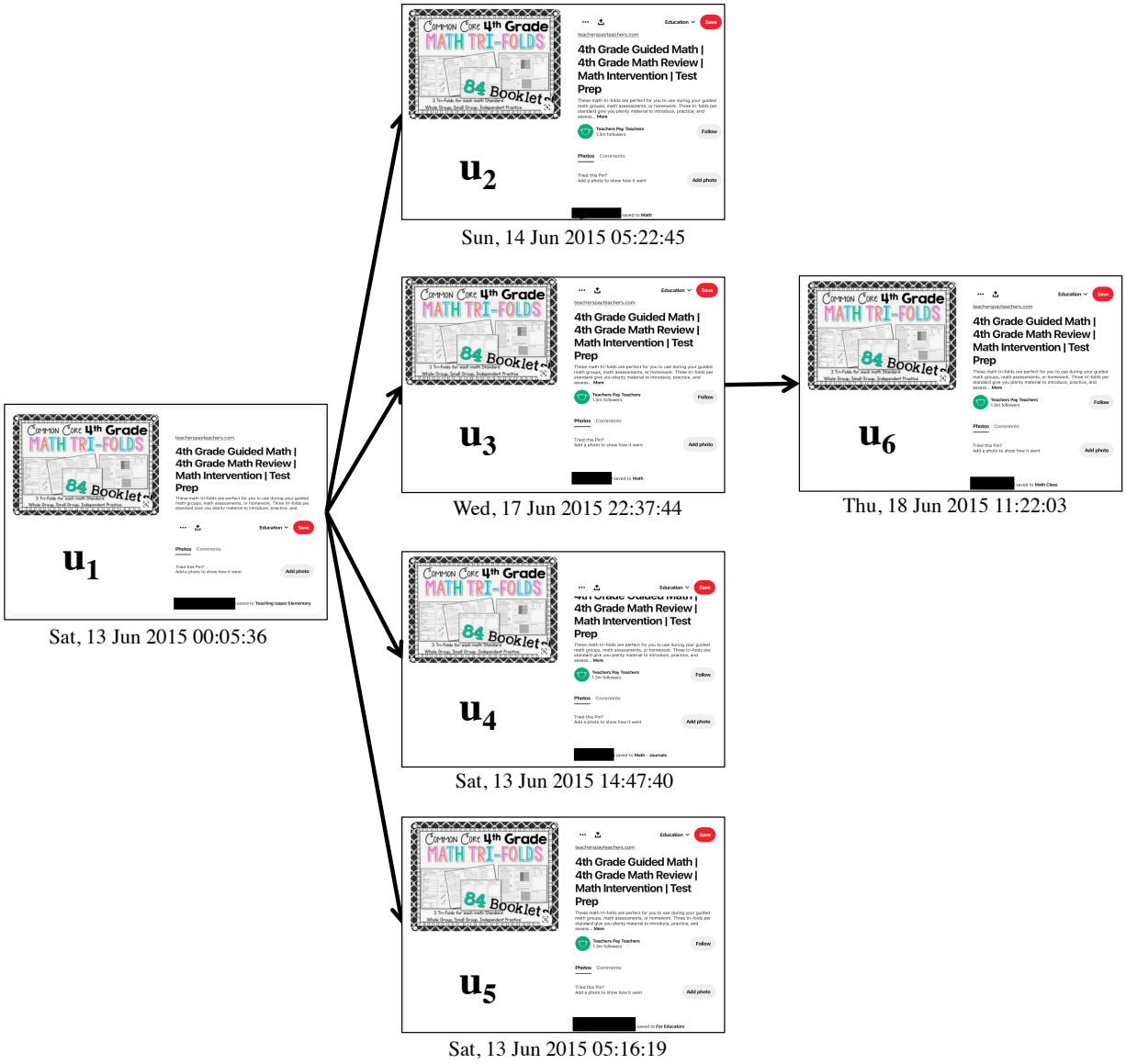


Figure 5.1: An example of diffusion tree.

5.1 Dataset: Diffusion Trees

As mentioned above, to investigate the diffusion of curated resources on Pinterest, we need to construct diffusion trees of resources. A diffusion tree is a tree representing the cascade of information among users. Formally, we define a diffusion tree as a directed graph $T = (U, E, p, r)$, where U is the set of users participating in the diffusion, E is the set of directed edges between users U, p is the pin being diffused among U, and r is the root of the tree– a teacher who has initially curated pin p. Each edge $e = (u_i, u_j) \in E$ indicates that user $u_i \in U$ has re-pinned (received) pin p from user $u_j \in U$. Figure 5.1 demonstrates an example of diffusion tree. In this Figure, user u_1 (the root) has curated a resource, which has been diffused in the network and further re-pinned by users $u_2, u_3, u_4, u_5,$ and u_6.[1] In this example, we have also shown the curation time below each node. As it can be seen, this pin has been diffused very fast, which signifies the power of social media in the rapid diffusion of (educational) resources.

We created diffusion trees for 1,162,983 unique pins curated by our identified teachers.[2] To create an edge, we used the *parent pin* field in our dataset, which holds the previous pin (refer to Table 2.1). If a pin does not have the parent pin, it means it is in the root of the tree, e.g., the one curated by user u_1 in Figure 5.1. Note that the content of a pin (i.e., image/video, description, etc.) does not change when it gets diffused. However, Pinterest assigns a unique identifier to each pin once a user re-pins it. Using these identifiers, we could trace back pins and construct diffusion trees. Moreover, we created trees for all types of resources curated by teachers, either educational or non-educational. We did this for two reasons. First, through incorporating non-educational pins, we can effectively contextualize how educational resources, compared to non-educational ones, are diffused. Second, in addition to investigating the diffusion of teacher-curated resources on Pinterest, the overarching goal of this dissertation is the behavior analysis of teachers in social media. Therefore, we believe investigating the diffusion of all types of teacher-curated resources is contributing to this goal. Finally, it is worth mentioning that our dataset of diffusion trees is the

[1] Names of the users have been redacted for privacy purposes.

[2] Identified teachers are the same teachers used in Chapter 4.

largest dataset of teacher-curated resources on Pinterest, which can foster future research on the diffusion of resources on social media.

5.2 Characterizing Diffusion

Now we have diffusion trees; we need to characterize them based on some measures. These measures should signify what previous studies have emphasized about the diffusion of educational resources on social media, particularly Pinterest, namely a) educational resources are diffused in a large scale manner among teachers, and b) the diffusion of educational resources is fast [4, 133, 45, 78]. We adopted these measures from Han et al. [161]. They introduced them to study the diffusion of information on Pinterest.

5.2.1 Volume

The first measure is the *volume* (VL), which is defined as the number of nodes in a diffusion tree:

$$VL(T) = |U| \tag{5.1}$$

For instance, the the volume of the tree in Figure 5.1 is 6. Despite its simplicity, the volume has a significant implication since it informs us how widely a piece of information has been diffused. In particular, the number of users who have received a piece of information is used in the popularity prediction/assessment of information on social media, e.g., the number of retweets on Twitter [162, 163]. Pertinent to this chapter, by determining the volume, we can ascertain how much other users/teachers are interested in a teacher-curated resource.

5.2.2 Virality

The volume, while still being important, reports the number of individuals who have adopted a resource. Nevertheless, depending on the structure of a diffusion tree, this adoption can take different forms. To fix the idea, Figure 5.2 demonstrates three distinct diffusion trees, all having volume 8. In T_1, there is a broadcast from the root to other nodes where only the root has participated

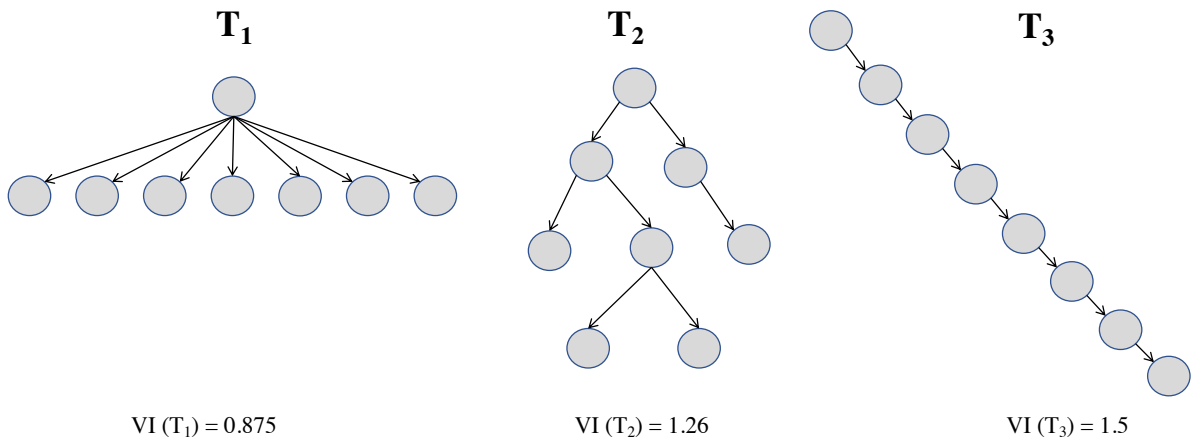


Figure 5.2: Three structurally different trees with the same volume but different virality values.

in the information propagation. However, in T_2, the resource has been relayed by different nodes where more nodes have participated in the diffusion. T_3 is an extreme scenario where we see a chain-wise 'deep' tree, and the message has been passed on consecutively. Distinguishing between diffusion scenarios based on their tree structure informs us about the virality and penetration of a message across the network [161]. Furthermore, such distinction is important in the context of teachers in social media as we can specify how others have responded to a teacher-curated resource. To this end, we define the *virality* (VI) of a diffusion tree:

$$VI(T) = \frac{2}{(|U|) \times (|U| - 1)} \sum_{\forall u_i, u_j \in U} d(u_i, u_j) \tag{5.2}$$

where $d(u_i, u_j)$ is the shortest distance between two users u_i and u_j in the diffusion tree T. The sum of shortest distances between nodes in a graph is known as Wiener Index [164, 165]. Han et al. [161] used a similar metric for the virality. The term $\frac{2}{(|U|)\times(|U|-1)}$ normalizes the Wiener Index.

5.2.3 Velocity

In addition to the number of people who have received a resource and how viral the resource has become, the speed of the diffusion is also important. In particular, previous studies have pointed out that the fast diffusion of educational resources on social media and then to the classroom is what makes online social media very appealing to teachers [4, 91, 166]. Hence, the third diffusion

measure is about the velocity (or speed) of diffusion. To this end, we introduce two metrics. The first metric is the *average re-pin time*, which calculates the average time between two re-pins in the diffusion tree. The average re-pin time (ART) for a diffusion tree is defined as follows:

$$ART(T) = \frac{1}{|U| - 1} \sum_{\forall e \in T} u_j(t) - u_i(t) \tag{5.3}$$

where (u_i, u_j) is an edge in the diffusion tree and $u_i(t)$ ($u_j(t)$) denotes the re-pin time by user u_i (u_j). Note that, in Eq. 5.3, we have subtracted $u_i(t)$ from $u_j(t)$ since user u_i has received the pin earlier. Given the fast diffusion of information on social media, we use an hour as the scale of time. ART for the example tree demonstrated in Figure 5.1 is 46.2 hours. Sometimes a resource can continue to get further diffused for a long time (say months), and thus makes $ART(T)$ large. Therefore, to better capture the velocity of diffusion, we additionally define the *first re-pin time* (FRT). It is the amount of time from the initial curation of a pin to when someone re-pins it for the *first* time:

$$FRT(T) = min\{u_i(t) - r(t)\} \;\; s.t. \;\; (r, u_i) \in E \tag{5.4}$$

where $r(t)$ denotes the time that the root has curated the pin. FRT for the example tree in Figure 5.1 is 5.16 hours.

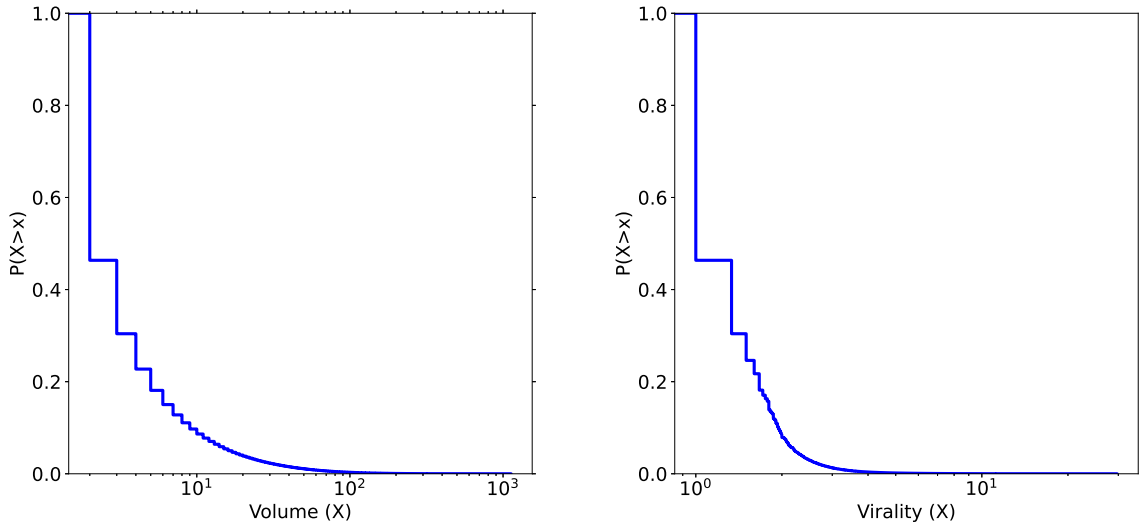
5.3 Diffusion Analysis

In this section, we analyze the constructed diffusion trees. First, in Section 5.3.1, we present some statistics about diffusion measures. Then, in Section 5.3.2, we present the results of our investigation regarding how different types of resources are diffused. Finally, in Section 5.3.3, we analyze the relationship between the introduced diffusion measures and some teacher attributes.

5.3.1 Distribution of Diffusion Measures

In this part, we look into the statistics and distributions of the three diffusion measures. Figure 5.3 demonstrates the CCDF of the volume and vitality for all resources. We can observe that both the

volume and virality follow a power-law distribution where most resources have low volume and vitality, and a small percentage has very high values for these two measures. In addition, Table 5.1 shows some statistics about the virality, volume, and velocity measures. According to this table, only the top 0.1% of diffused resources have the volume and virality larger than 174 hours and 5.99 hours, respectively. This means that there are a handful of resources curated by teachers that have become extremely popular. This is in line with previous studies on the virality and popularity of information on social media, where they have shown that some information becomes significantly viral across the network [167, 168, 161]. Finally, on average, around five people have adopted each teacher-curated resource on Pinterest.



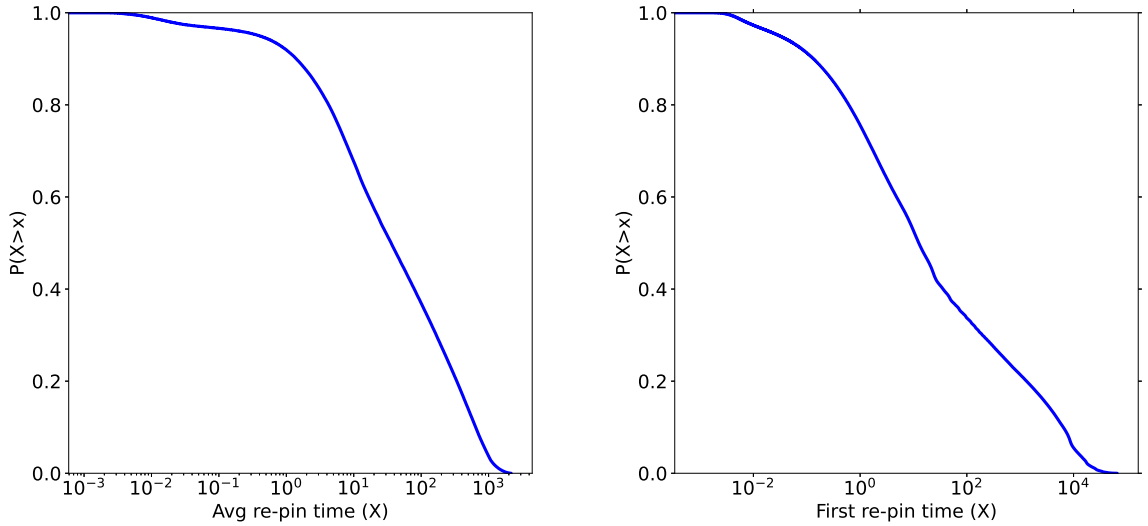
(a) Volume (Eq. 5.1)

(b) Virality (Eq. 5.2)

Figure 5.3: The CCDF of the volume and virality. x-axes are in log scale.

Table 5.1: Some statistics of the introduced diffusion measures of the constructed diffusion trees.

Measure	Min	Max	Mean	Median	Std	top 0.1%	top 0.01%
Volume	2	1,129	5.4	2	13.58	> 174	> 434
Virality	1	29.72	1.33	1	0.54	> 5.99	> 11.45
ART	0.0012	2,159.4	192.4	35.8	317.3	> 1,950.7	> 2,113.2
FRT	0.0008	65,655.0	1,814.4	12.5	4,975.0	> 45,020.7	> 56,960.9



(a) Average re-pin time (Eq. 5.3)

(b) Fist re-pin time (Eq. 5.4)

Figure 5.4: The CCDF of the velocity measures. x-axes are in log scale.

Figure 5.4 shows the CCDF of the velocity measures: the average re-pin time (ART) in Figure 5.4a and the first re-pin time (FRT) in Figure 5.4b. Unlike the volume and virality, neither of the velocity measures follows a power-law distribution. Moreover, there is a significant difference between the mean and median for ART and FRT. Specifically, while pins have a small median of the average re-pin and the first re-pin times (35.56 and 12.56 hours, respectively), their means have been skewed because of some outliers.

From the results presented in Figures 5.3 and 5.4 and Table 5.1, we can conclude that teacher-curated resources diffuse rapidly and are received by a significant number of other users on Pinterest, including other teachers. While this finding has been reported before [4, 133, 45, 78], this is the first study that confirms it through a large-scale data-driven assessment. In the next part, we perform a more detailed evaluation of the diffusion of teacher-curated resources based on the resources' attributes.

5.3.2 Resource Attributes and Diffusion Measures

In this part, we analyse the diffusion measures based on two crucial attributes of pins, namely the topic and the domain.

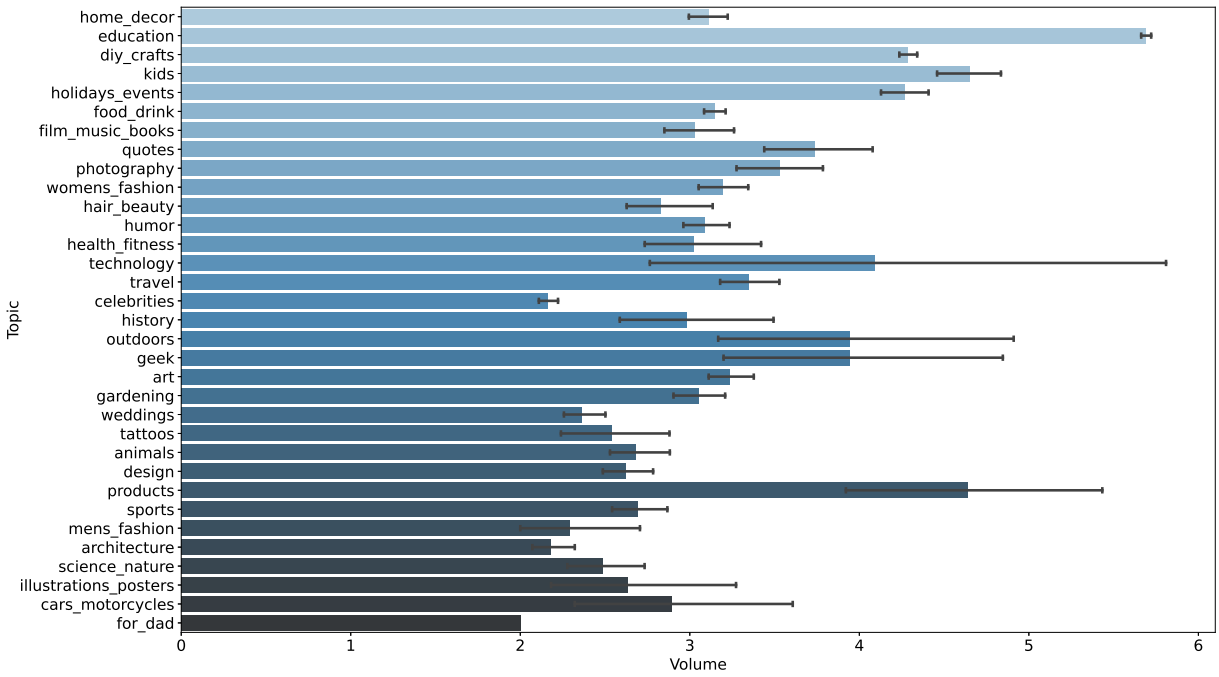


Figure 5.5: The average volume per topic.

5.3.2.1 Topic

Figure 5.5 shows the average value of the volume per topic. As shown in this figure, the pins whose topic is *education* have the highest volume where on average, each of such pins is adopted by six users on Pinterest. Interestingly, *kids* is the second topic in terms of volume. This can partially be explained by the similarity of this topic with *education* and being attractive to teachers, especially for pre-kindergarten or homeschooling-specific materials. Regarding the other topics, they all have low volumes, mostly below 4. Also, since the predominant topic is *education* and there is a small amount of data for the other topics, they exhibit high standard errors.

Figure 5.6 shows the average value of virality per topic. Similar to the volume, *education* has the highest virality. This signifies the high penetration of teacher-curated educational resources across

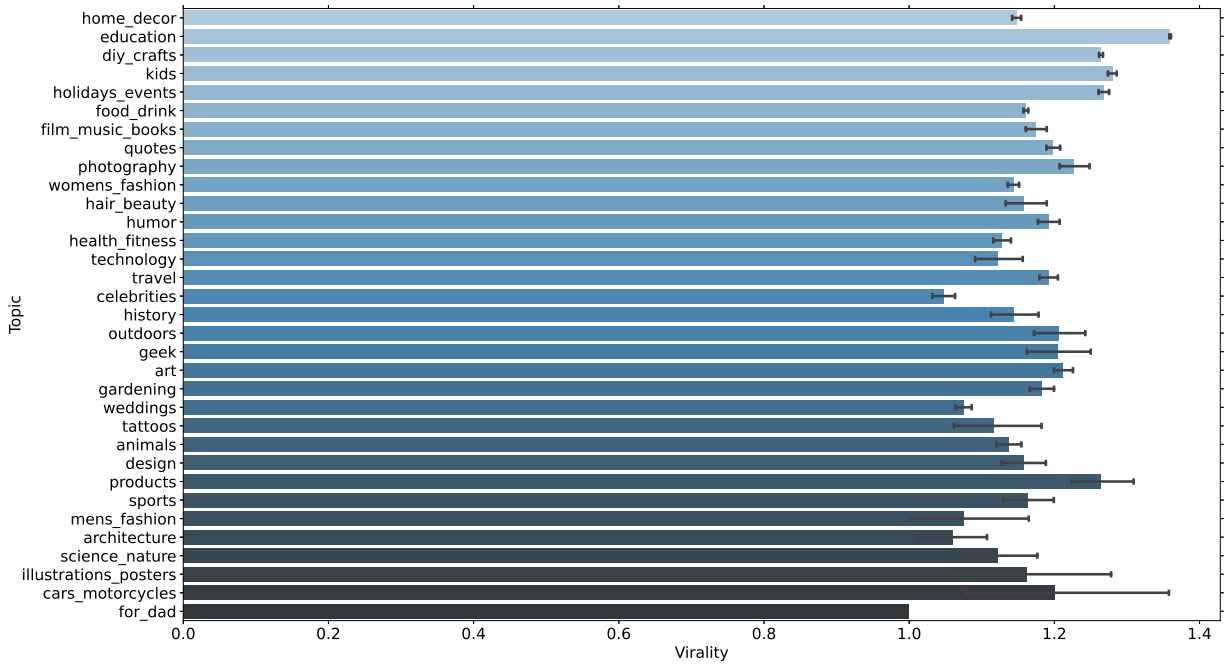


Figure 5.6: The average virality per topic.

Pinterest. The average value of virality for the topic *kids* is relatively high. Moreover, comparing the volume and virality values in Figure 5.6 and Figure 5.5, we can infer that the high volume does not necessarily mean high virality. For instance, the pins whose topic is *quotes*, on overage, have relatively high virality while their volume is not very high.

Figures 5.7 and 5.8 demonstrate the median of the average re-pin time and the first re-pin time, respectively. Here, we used the median to plot the charts since, as mentioned in Section 5.3.1, the values of ART and FRT of our constructed diffusion trees have some outliers. Furthermore, there is a small amount of data for certain topics, and consequently, their velocity measures are very skewed, as can be observed from Figures 5.7 and 5.8. Hence, for clarity purposes, in Figure 5.9, we have included the median of the velocity measures for topics whose pin proportion is at least 10%. We make two major observations based on the results shown for the velocity measures. First, *education* has a short average re-pin time and first re-pin time. In particular, the median of the first re-pin time for *education* is only 12 hours. This means a teacher-curated resource takes approximately only half a day to be adopted by another user on Pinterest, which signifies the fast

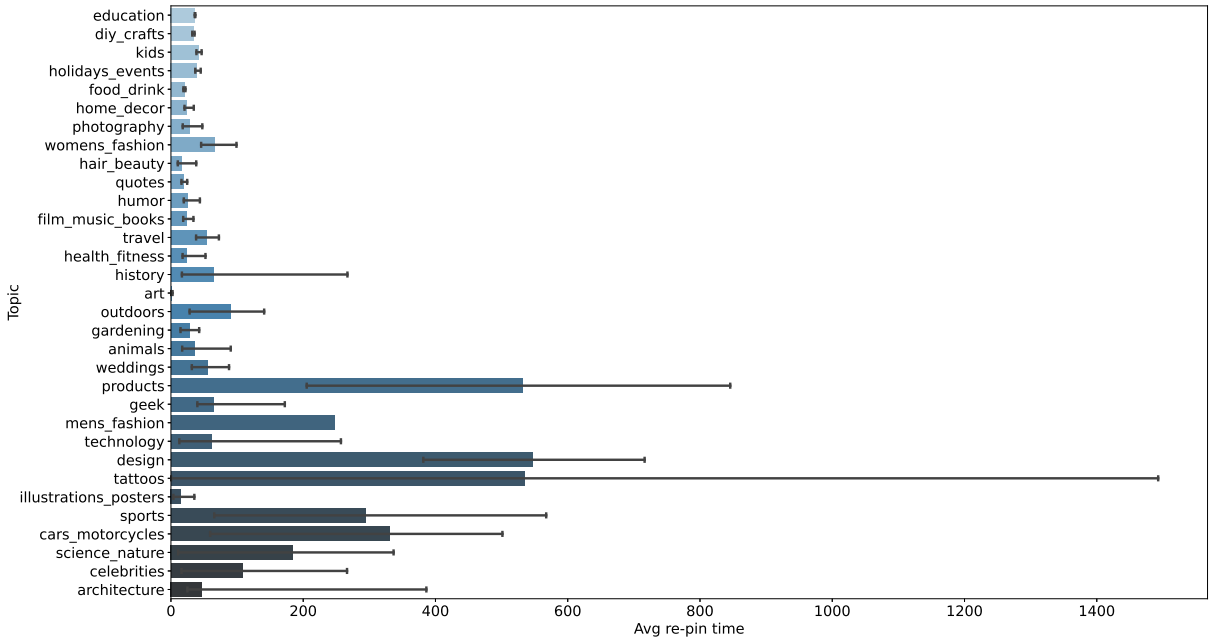


Figure 5.7: The median of the average re-pein time per topic.

diffusion of educational materials across Pinterest. Second, compared to the first re-pin time, the average re-pin time is generally longer. We believe this occurs because a user quickly saves a pin curated by the root, and then the pin is spread across the network at a lesser pace. There are a few exceptions, however, e.g., *travel*, *art*. We speculate this is due to the special nature of these topics to teachers, where their pins might take some time to attract someone's attention, while once they do, they eventually get diffused.

5.3.2.2 Domain

In this part, we investigate the diffusion of teacher-curated resources based on their domains. For this analysis, we only included the top 10 domains utilized by teachers. Figure 5.10 shows the average volume and virality values for the top 10 domains of teacher-curated resources. Figure 5.11 shows the median of the average re-pin time and first re-pin time for the same top 10 domains. We make the following observations based on these results.

Except for *youtube.com* and *Uploaded by User* (directly uploaded from the user's device), the rest of the domains are specifically related to education, e.g., *moffattgirls.blogspot.com*. Interestingly,

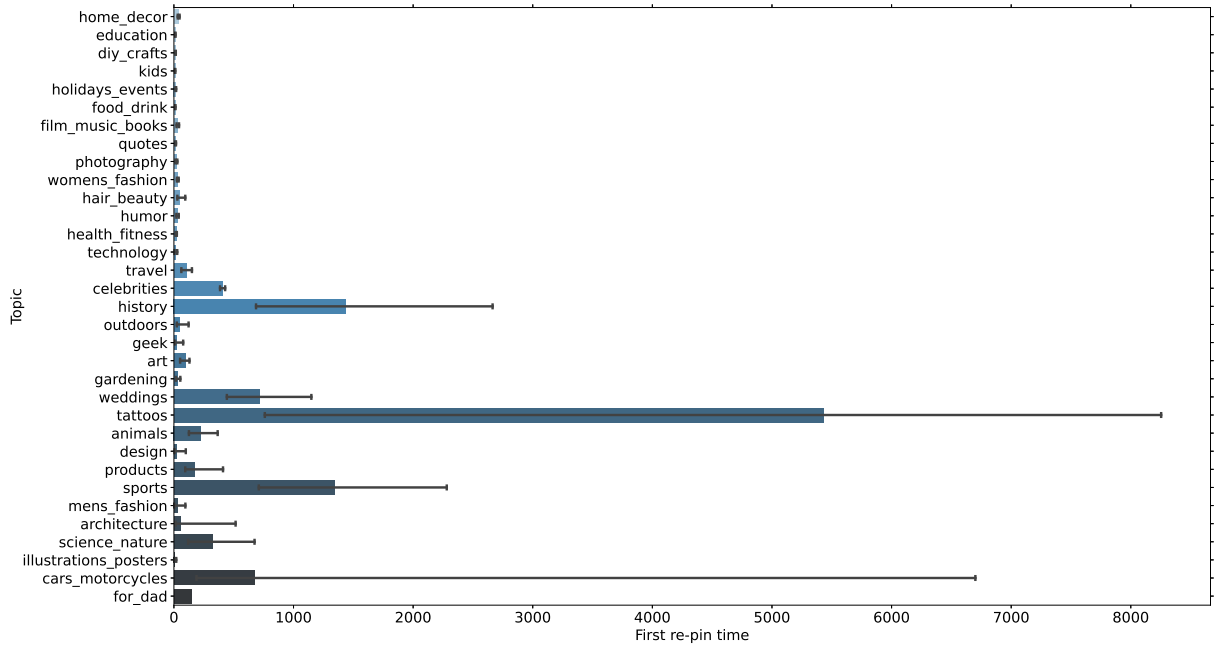
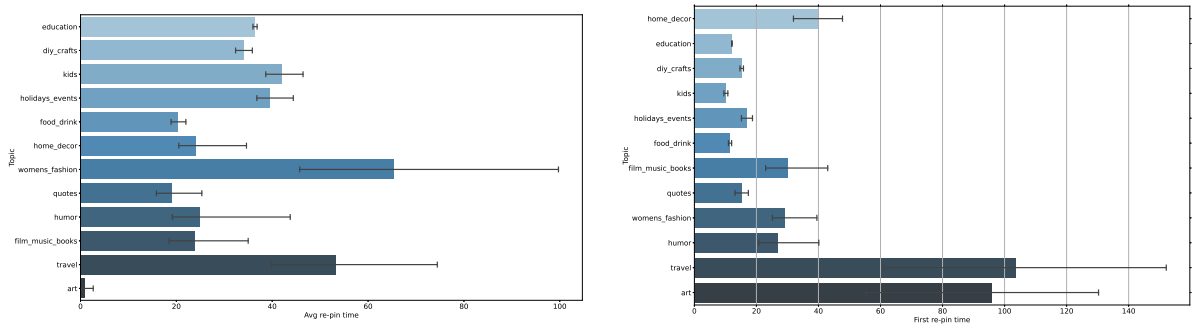


Figure 5.8: The median of the first re-pin time per topic.



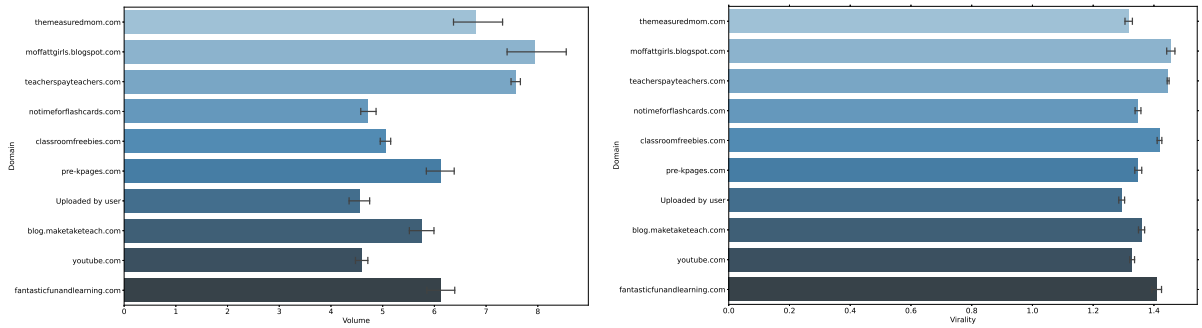
(a) Average re-pin time

(b) First re-pin time

Figure 5.9: The median of the velocity measures for the top topics.

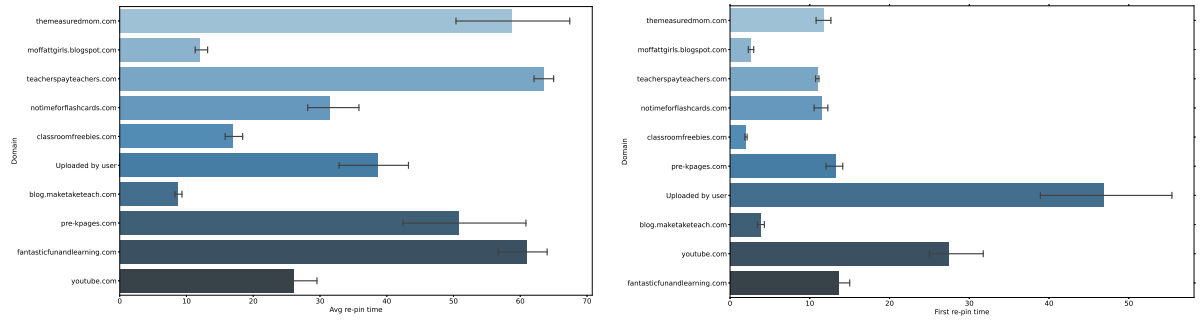
we can observe that pins from *moffattgirls.blogspot.com* have the highest volume. This blog is run by a former elementary school teacher who exclusively produces educational materials. Our further investigation reveals that this teacher is a prolific and influential user on *teacherspayteachers.com*.[3] Hence, it is no surprise that her educational materials are of interest to many others. Moreover, the pins from this domain have high virality and are diffused very fast. We have showcased an example pin from *moffattgirls.blogspot.com*, which has been re-pinned by 936 other users on Pinterest.

[3] `https://www.teacherspayteachers.com/Store/The-Moffatt-Girls`



(a) The average volume per domain (b) The average volume per domain

Figure 5.10: The average of the volume and virality for the top 10 domains of teacher-curated resources.



(a) Median of the average re-pin time per domain (b) The median of the first re-pin time per domain

Figure 5.11: The median of the velocity measures for the top 10 domains of teacher-curated resources.

These inspiring and active teachers signify the power of teachers in social media and how they can impact their fellow teachers in this digital age.

Materials from *teacherspayteachers.com* also have high volume and virality, which indicates the popularity of educational materials from this source. Interestingly, the velocity measures for pins from *teacherspayteachers.com* have a short first re-pin time while relatively longer average re-pin time. This is due to the artifact that pins from this popular source keep being diffused across Pinterest for a long time; thus, the average re-pin time becomes long.

Another observation is the long first re-pin time for pins whose domain is *Uploaded by User*. We believe this is because of the following reason. Since these pins have essentially no domain (i.e., they do not come from a website on the Internet), other users might hesitate to adopt them quickly, e.g., due to the lack of trust. Nevertheless, once they accumulate some initial popularity,

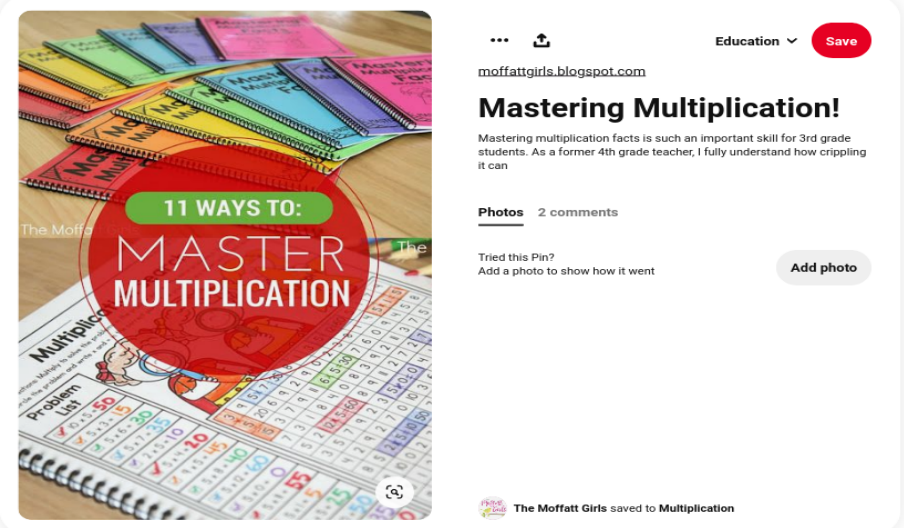


Figure 5.12: A showcase of a popular pin from *moffattgirls.blogspot.com* adopted by 936 other users.

they become more popular and diffuse across the network.

5.3.3 Teacher Attributes and Diffusion

In Section 5.3.2, we demonstrated that attributes of a teacher-curated resource, namely its topic and domain, are related to its diffusion. In addition to the resource itself, a resource producer (here a teacher) can also affect the diffusion process [169]. For instance, there is a rich literature on identifying influential spreaders in social media based on the spreader's attributes [170, 171, 172]. Hence, in this part, we investigate whether teacher attributes are related to the diffusion of the resources they curate. To this end, we consider ten teacher-related attributes and inspect their relationship to diffusion measures as explained in the following.

We consider the number of pins and the number of boards for the first two attributes, respectively. The reason to include these two attributes is to assess whether a teacher's activity level leads to the widespread and fast diffusion of their materials. For the next attribute, we consider the number of followers. The reason to include this attribute is that the resources of a teacher who has more followers might enjoy a higher chance of being disseminated in the network. Following this argument, we also include the number of followees and the total number of friends (i.e., followers

and followees combined) to discover how these two attributes affect the diffusion measures. In Section 4.4.1, we defined reciprocity, which captures the proportion of bidirectional connections of a user. Here, we include reciprocity as a teacher's attribute to investigate its relationship to diffusion. The idea is to investigate whether having a stronger connection between a teacher and their online friends, measured via reciprocity, affects the diffusion. Moreover, we include three crucial centrality attributes, namely eigenvector, betweenness, and closeness centrality. As mentioned before, centrality attributes calculate the structural importance (or influence) of a node in a network. The question is, do resources of the more central teachers have a higher chance to be adopted by other users and perhaps at a faster rate? The final attribute includes the local clustering coefficient, which quantifies how close a user's neighbors are to a complete graph (a clique). Including the local clustering coefficient is particularly important since previous studies [25, 173] have shown that cliques in school-level teacher networks can lead to a better diffusion of information.

To investigate the relationship between teacher attributes and diffusion measures, we perform four regression analyses. In each analysis, teacher attributes are the independent variables used to predict the corresponding diffusion measure, i.e., the dependent variable. Thus, we attempt to determine how much information in each teacher's attribute can explain a diffusion measure. Note that we only investigate teachers who are the roots of the diffusion trees since we aim to determine the attributes of the producers of pins, not others who further do the re-pinning. Given this, a teacher can be associated with more than one diffusion tree as the root. Hence, to perform a teacher-level analysis, we aggregated values of each diffusion measure for all diffusion trees associate with a teacher. For the volume and virality, we computed the mean values. For the velocity measures, we computed the median since it offers a better estimation than the mean, as discussed in Section 5.3.2. Finally, we used the statsmodels package [174] in Python and fit ordinary least squares (OLS) for each regression analysis.

Tables 5.2, 5.3, 5.4, and 5.5 show the results of the regression analysis for the volume, the virality, the average re-pin time, and the first re-pin time, respectively. Each table has four columns. The first column is the coefficient of each teacher attribute in the regression analysis. The larger

Table 5.2: Regression analysis results of predicting volume using teacher attributes.

Attribute	Coefficient	Std error	t	P > |t|
#Pins	4.449e-07	2.44e-06	0.182	0.855
#Boards	-0.0011	0.000	-3.049	0.002
#Followers	-0.0005	0.000	-3.590	0.000
#Followees	0.0003	0.000	2.539	0.011
#Friends	-0.0003	9.35e-05	-2.976	0.003
Reciprocity	0.2359	0.084	2.816	0.005
Eigenvector Centrality	54.8268	12.913	4.246	0.000
Betweenness Centrality	28.2263	75.308	0.375	0.708
Closeness Centrality	7.4408	0.190	39.110	0.000
Local Clustering Coefficient	1.7113	0.283	6.039	0.000
	Mean squared error: 2.19 Adj. R-squared: 0.539			

Table 5.3: Regression analysis results of predicting the virality using teacher attributes.

Attribute	Coefficient	Std error	t	P > |t|
#Pins	-2.419e-06	2.29e-07	-10.574	0.000
#Boards	-0.0001	3.31e-05	-3.861	0.000
#Followers	-1.318e-05	1.41e-05	-0.936	0.349
#Followees	-3.879e-05	9.63e-06	-4.029	0.000
#Friends	-5.197e-05	8.77e-06	-5.923	0.000
Reciprocity	0.0894	0.008	11.379	0.000
Eigenvector Centrality	2.5898	1.211	2.138	0.033
Betweenness Centrality	24.0530	7.065	3.405	0.001
Closeness Centrality	3.5286	0.018	197.698	0.000
Local Clustering Coefficient	0.6131	0.027	23.064	0.000
	Mean squared error: 0.04 Adj. R-squared: 0.965			

the magnitude of the coefficient, the more impact the attribute has on the corresponding predicted diffusion measure. The second column shows the standard error of the coefficient. The third column is the t-value retrieved from a t-test. It essentially calculates the difference between the predicted value of a diffusion measure using an independent variable and the actual value of the measure. The last column is the p-value of the t-test, which is a measurement of how likely a coefficient measured through a regression model is by chance. In the last row of each table, we have included two major pieces of information. The first one is the mean squared error, which measures the difference between the actual values of a diffusion measure and the predicted values. The second one is the adjusted R-squared, which measures how much of the independent variables

Table 5.4: Regression analysis results of predicting the average re-pin time using teacher attributes.

Attribute	Coefficient	Std error	t	P > |t|
#Pins	-0.0015	0.000	-4.077	0.000
#Boards	0.1079	0.061	1.781	0.075
#Followers	0.0096	0.022	0.440	0.660
#Followees	0.0020	0.016	0.127	0.899
#Friends	0.0116	0.014	0.822	0.411
Reciprocity	-143.8068	17.347	-8.290	0.000
Eigenvector Centrality	-8090.1645	1884.586	-4.293	0.000
Betweenness Centrality	8936.4525	1.1e+04	0.811	0.417
Closeness Centrality	607.2976	36.660	16.566	0.000
Local Clustering Coefficient	410.1702	62.894	6.522	0.000
	Mean squared error: 85667.98 Adj. R-squared: 0.263			

Table 5.5: Regression analysis results of predicting the first re-pin time using teacher attributes.

Attribute	Coefficient	Std error	t	P > |t|
#Pins	0.0308	0.005	-6.450	0.000
#Boards	1.3398	0.691	1.940	0.052
#Followers	0.3589	0.294	1.223	0.222
#Followees	-0.3620	0.201	-1.804	0.071
#Friends	-0.0031	0.183	-0.017	0.986
Reciprocity	-1005.0319	163.868	-6.133	0.000
Eigenvector Centrality	-8.951e+04	2.53e+04	-3.544	0.000
Betweenness Centrality	1.404e+05	1.47e+05	0.953	0.340
Closeness Centrality	85375.8527	372.131	14.446	0.000
Local Clustering Coefficient	4871.5493	554.231	8.790	0.000
	Mean squared error: 18323220.30 Adj. R-squared: 0.143			

(i.e., teacher attributes) is explained by changes in a dependent variable (i.e., a diffusion measure). We make the following observations based on the results presented in these four tables.

- The adjusted R-squared is high for the volume and virality. Nevertheless, its value is very low for the velocity measures. This means teacher attributes can explain the volume and virality while failing to do so for the velocity measures. This can also be inferred by looking at the mean squared errors whose values are low for the volume and virality while they are high for the velocity measures.

- As far as the number of pins is concerned, their coefficients are generally low for all measures.

This indicates that the diffusion of pins is not related to the high activity rate of their producer. This seems logical since merely saving more pins and creating more boards on Pinterest does not guarantee that these resources will be diffused widely. The only exception is the number of boards for the first re-pin time, for which the coefficient is positive and relatively large. We think the reason is that Pinterest users can independently follow a board without even following its curator. Hence, the more boards a user has, the higher chance someone can quickly re-pin from any of these board. However, based on our results, such rapid adoption has not necessarily led to the pin's popularity (high volume) and virality.

- The coefficients of the number of connections (i.e., number of followers, followees, and friends) are very low. While the coefficient of the number of followers is relatively high for the first re-pin time, it is not statistically significant– See the P>|t| column. We think the low value of coefficients of the number of connections can be explained by the fact that Pinterest is essentially a social curation website where users can re-pin resources from others without following them.

- Reciprocity has a low coefficient for the virality while it is relatively high for the volume. A teacher with high reciprocity has a strong relationship with their online friends. Consequently, their friends trustfully re-pin their resources. Nevertheless, virality is a complicated measure that cannot be explained adequately by reciprocity. Moreover, as far as the velocity measures are concerned, reciprocity coefficients have a large magnitude and negative sign. The explanation behind this requires further investigation.

- The most critical finding of this part is the relationship between the centrality metrics and the volume and virality. Except for betweenness centrality for the volume, the centrality metrics can perfectly explain the virality and volume. The reason is that centrality metrics take into account the network's structure, which plays an essential role in how information is diffused. For instance, a high eigenvector centrality means a teacher is connected to other high central users. Therefore, after these central neighbors re-pin a resource, there is a bigger chance

the resource will be widely diffused because of their own high structural influence. Also, closeness and betweenness centrality have to do with the shortest paths in the graph, which play an essential role in the diffusion of information [175].

- As mentioned before, the local clustering coefficient is an important factor in the diffusion of information in school-level teacher networks [25, 173]. Moreover, our results in the part of the dissertation indicate that this attribute is also crucial in the diffusion of information across the network of teachers on Pinterest.

Based on the above observations, we can conclude that teacher attributes significantly affect the volume and virality of their curated resources. In particular, the network-level structural characteristics of a teacher play an essential role in determining the volume and virality of their resources. In contrast, how fast these resources are diffused cannot be adequately determined by teacher attributes.

CHAPTER 6

CONCLUSION AND FUTURE DIRECTIONS

In this chapter, we provide a summary of our research results and further present promising future research directions.

6.1 Summary

In this dissertation, we proposed novel research in the three primary directions of teachers in social media: automatic teacher identification, teacher gender analysis, and diffusion of teacher-curated resources. Next, we summarize our contributions in each direction one by one.

To supplement their students' educational needs and improve their teaching quality, many teachers turn to online social media platforms where an enormous number of educational resources have been curated. Such resources are precious materials for teachers and students, especially with the COVID-19 pandemic affecting traditional education. Hence, for the past few years, teachers in social media have been the subject of many educational studies. Despite the progress in this line of research, one of the major obstacles is the limited number of teachers being investigated. More specifically, the current studies usually suffice to at most a few hundred surveyed/interviewed teachers. However, to offer a better picture of teachers in online social media and enable modern data science approaches to find meaningful patterns in the teacher-related data, we need to identify more teachers. Thus, in the third chapter of this dissertation, we proposed a framework to automatically identify teachers on Pinterest-an image-based social media popular among teachers. For the first time, our framework formulated the teacher identification as a positive unlabelled learning task where positive samples are a small set of surveyed teachers, and unlabelled samples are their friends on Pinterest. We performed extensive experiments on a real dataset of teachers on Pinterest and showed that our framework outperforms strong baselines. Moreover, using our framework, we reliably identified thousands of other teachers on Pinterest. Finally, we believe our proposed framework can improve the quality of many research endeavors concerned with studying teachers

in social media.

Some studies in the education literature have shown that the gender of teachers affects their behavior [43, 176, 177, 178]. For instance, male and female teachers may differ in the way they structure their classroom, selecting topics and examples in their pedagogical practices [43]. However, while such behavioral differences between males and females have been investigated before, there is a lack of study on how male and female teachers behave on social media. Such investigation is crucial since online social media is now an integral part of the teacher's professional career development. Perhaps, one reason for the lack of such study had been the unavailability of a rich and large-scale dataset of teachers on social media. Nevertheless, we addressed this issue in the third chapter and identified a large dataset of teachers on Pinterest. Hence, we used this dataset in the fourth chapter and performed a thorough exploratory analysis of male and female teachers on Pinterest. We performed our study in two main parts: online activity analysis of male and female teachers and their social network analysis. In the first part, we discovered that male and female teachers curate similar types of resources and mainly utilize Pinterest for educational purposes. Moreover, we performed a thorough investigation on the topic and domain of the resources curated by teachers using several novel measures: the topic/domain entropy and the topic/domain oscillation. As a result, we found out that male teachers are more focused on their resource curation process while females are more receptive to exploring non-educational content. We also identified the unique patterns that male and female teachers use social media in terms of the time of their access. In the social network analysis part, we found out that male teachers tend to have more connections and more actively follow other users on social media. Moreover, male and female teachers showed having very similar structural centrality scores. Eventually, we investigated gender homophily and identified some homophilic behavior, primarily by male teachers. In conclusion, as far as teachers' professional activities are concerned, based on our findings, males and females behave very similarly and see social media as a means to advance their careers.

Previous studies have referred to the widespread and fast diffusion of online educational resources on social media [4, 133, 45, 78]. Nevertheless, they have shown this using qualitative

analysis (interviews/surveys with teachers) or anecdotal reports. Therefore, we recognized a need to investigate the diffusion of resources on social media through a data-driven investigation. Thanks to our proposed method in Chapter 3, we had access to a large set of teachers and their diffused resources on Pinterest. Given this, in the fifth chapter, we performed an analysis of the diffusion of teacher-curated resources on Pinterest. First, we built a large set of diffusion trees of these resources on Pinterest. Then, we defined three crucial measures to characterize the diffusion process, namely, i.e., volume, virality, and velocity. Our analysis of the diffusion of teacher-curated resources showed that educational materials are disseminated across the network widely and in a rapid manner. Eventually, we performed several regression analyses to determine what teacher attributes affect the diffusion process. Our study showed that the structural attributes significantly impact the diffusion of teacher-curated resources on Pinterest. We believe our large-scale data-driven study in this chapter of the dissertation has deepened our understanding of the diffusion of teacher-curated resources materials on Pinterest and can foster further research.

6.2 Future Directions

In this section, we present several possible future directions across the major areas of teachers in social media and how data science can help these critical directions.

- **Closing the loop in the automatic teacher identification.** In the first chapter, we proposed a method to identify teachers on Pinterest automatically. We showed that by using this method, we could answer interesting research questions. However, it would be valuable to conduct surveys/interviews with these newly identified teachers. The main reason to do this is that we can benefit from both worlds: direct data of teachers in online social media as well as obtaining detailed and controlled information via surveys/interviews. In fact, educational researchers have recommended using both types of data to reach better conclusions [34]. It is worth mentioning that conducting surveys/interviews with automatically identified teachers is related to snowball sampling, where researchers recruit their samples (teachers) via broadcasting an invitation on social media [72]. Snowball sampling, however, has a low

response rate; while using our method, we can precisely control whom we should contact for the survey/interview.

- **Teachers in multiple social media platforms.** Most current social media studies, including this dissertation, have investigated teachers in a single platform. However, teachers, like other people, use different platforms and probably for different purposes. Therefore, it would be interesting to examine teachers on multiple social media platforms. For instance, it would be interesting how information diffuses in different social media platforms, e.g., Pinterest vs. Twitter. One of the challenges for this direction is identifying the same teachers on multiple social media platforms. A preliminary solution to this challenge is to use network alignment [138], where the same nodes across the two networks are mapped together.

- **Quality assessment of online educational materials.** As shown in this dissertation, online educational resources are widely diffused across social media and are adopted by many teachers and used in classroom activities. Given this, an important research direction is to characterize the content of these educational resources and assess their quality. Moreover, with the proliferation of online misinformation [179, 180], assessing the quality of online educational material is particularly important. Moreover, as we press onward to utilize machine learning models for developing practical educational applications, e.g., recommending educational materials to teachers and students, quality assurance of the disseminated online materials is critical. In particular, machine learning algorithms have been shown to be vulnerable to biased and low-quality content [181, 182].

- **Unifying theoretical frameworks.** Teachers in social media have been studied from the perspective of several theories, e.g., the affinity space, the community of practice, the professional learning network [46]. Despite this, there is no conclusive framework that could delineate *what social media is to teachers?* We believe empirical evaluation using a large-scale data-driven analysis, similar to this dissertation, can be a great help to demystify the core notion of social media to teachers.

APPENDIX

THE ANNOTATION PROCEDURE

As mentioned in Chapter 3, we annotated some users to teachers and non-teachers. In this appendix, we describe the annotation procedure. Figure .1 illustrates the flowchart of this procedure. In the following, we describe each component of the procedure.

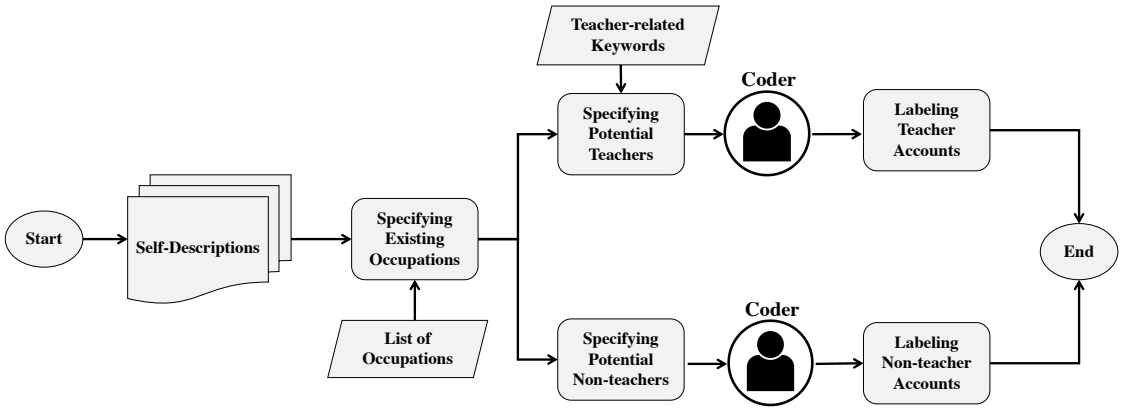


Figure .1: The flowchart of the annotation procedure.

A.1 Specifying Existing Occupations

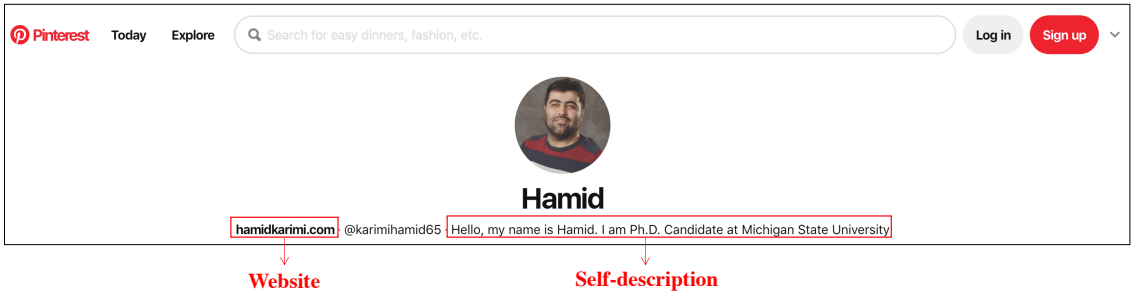


Figure .2: An example of self-description and website URL in a Pinterest's account.

On Pinterest, users can include a maximum of 160 characters about themselves. This textual information, which we call self-description, is the basis of our annotation procedure. In addition, users are allowed to include an URL to their website, which is visible in their profile. Figure .2 illustrates an example of self-description and website address in an account. We acquired these two pieces of information through the provided Pinterest API (refer to Section 2.2). In our dataset,

19,588 users had included the self-description (around 21% of users). In the following, we explain how we used self-descriptions in our annotation procedure.

Sometimes users mention their occupation in their self-description, e.g., "I am an accountant living in ...". Self-declared occupations are trustworthy information that can help us determine the person's occupation effectively. To this end, first, we acquired a list of 965 common human occupations from an online repository.[1] This popular repository maintains lists of words related to different entities, e.g., travel, sport, religion. Our investigation discovered that their maintained list of occupations contains typical occupations in our modern society. Second, we performed keyword matching and determined those occupations that appeared in the self-descriptions of users. We found the match for 102 occupations, e.g., *designer*, *singer*, *therapist*.

A.2 Specifying Potential Teachers

After determining the existing occupations in self-descriptions, we specified *potential teachers*. The process is as follows. We marked a user as a potential teacher if two conditions were met: 1) the self-description contained several teacher-related keywords, and 2) the self-description did not contain any of the existing occupations except 'teacher' and 'instructor'. The reason for enforcing the first condition was to comprehensively consider those accounts that potentially belong to teachers. Moreover, the purpose of the second condition was to exclude those accounts that mentioned other occupations, e.g., *designer*. For the first condition, we used the following keywords: 'teacher', 'teachers', 'teach', 'teaching', 'teaches', 'math', 'educator', 'instructor', 'kindergarten', 'grade', 'school', 'classroom', and 'teacherspayteachers' and 'tpt'. The selection of the keywords is based on the common words that appeared in the self-descriptions of the surveyed teachers. Notably, we included 'teacherspayteachers', and 'tpt' since active (American) teachers usually mention their *teacherspayteachers.com* account on Pinterest, e.g., "visit my store at TeachersPayTeachers http://www.teacherspayteachers.com/Store/X" or "find me on TPT: https://www.teacherspayteachers.com/Store/X". Two occupations were exempt from the second

[1] `https://github.com/dariusk/corpora/blob/master/data/humans/occupations.json`

condition, namely 'teacher' and 'instructor', since they are obviously related to teachers. After applying these two conditions, we ended up with a list of 3,624 potential teachers. Two human coders manually processed this list and determined the final labels, as will be explained later.

A.3 Specifying Potential Non-teachers

To specify *potential non-teacher* accounts, we did the opposite of what we performed for potential teachers. More specifically, two conditions had to be met to mark an account as a potential non-teacher: 1) if one of the existing occupations appeared in the self-description except for 'teacher' and 'instructor', and 2) none of the previously mentioned teacher-related keywords (i.e., 'teacher', 'teachers', 'teach', etc.) were in the self-description. The primary purpose of these two conditions was to enhance the confidence in marking accounts as potential non-teachers. As a result, we marked a list of 2,503 users as potential non-teachers. Similar to potential teachers, two human coders manually processed this list and determined the final labels, as will be explained next.

A.4 Labeling Potential Teachers and Non-Teachers

After preparing the lists of potential teachers and non-teachers, two coders manually labeled the users. The first coder is the author of this dissertation and the second coder is a senior computer science undergraduate student. The categories for labeling were *teacher*, *non-teacher*, and *uncertain*. The first source of information for labeling was reading the self-description and looking for definite cues about the person's occupation. For instance, in self-descriptions, "*I am a middle school teacher, I welcome creative ideas and pins!*" and "*Wedding | Portrait | Real Estate photographer in Lansing and serving all surrounding areas*", the occupations are teacher and photographer, respectively. The second source of information was the included website of a user if any, e.g., *hamidkarimi.com* in Figure .2. Therefore, the coders were allowed to refer to the user's website to infer their occupation, e.g., reading the 'About' page on the website. This was particularly useful for potential teachers since teachers often mention their websites or blogs. The third source

of information was other social media accounts mentioned in the self-descriptions, if any, e.g., Instagram or Facebook accounts. Therefore, the coders were allowed to visit a user's other social media accounts to determine the label. The fourth source of information was resources curated by a user (i.e., pins and boards). For instance, boards such as 'back to school', 'for the classroom', 'second-grade math' in a user's account indicate that the person's occupation is teacher/educator. Eventually, the coders were allowed to use other external sources such as Google search to specify the label of an account.

The coders performed their labeling independently. Regarding labeling the list of potential teachers, both coders agreed on 3,508 users to be teachers. For the list of potential non-teachers, the agreement on the non-teachers was 2,079. Those users for which the coders' labels did not agree were added back to the unlabeled users.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Statista. Number of social network users worldwide from 2017 to 2025 (in billions). https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/, 2020.

[2] Barry Wellman. Computer networks as social networks. *Science*, 293(5537):2031–2034, 2001.

[3] Tiffany A Pempek, Yevdokiya A Yermolayeva, and Sandra L Calvert. College students' social networking experiences on facebook. *Journal of applied developmental psychology*, 30(3):227–238, 2009.

[4] Kaitlin Torphy and Corey Drake. Educators meet the fifth estate: The role of social media in teacher training. *Teachers College Record*, 121(14):1–26, 2019.

[5] Kaitlin Torphy, Yuqing Liu, Sihua Hu, and Zixi Chen. Sources of professional support: Patterns of teachers' curation of instructional resources in social media. *American Journal of Education*, 127(1):13–47, 2020.

[6] Martin Rehm and Ad Notten. Twitter as an informal learning space for teachers!? the role of social capital in twitter conversations among teachers. *Teaching and Teacher Education*, 60:215–223, 2016.

[7] Fernando Rosell-Aguilar. Twitter: A professional development and community of practice tool for teachers. *Journal of Interactive Media in Education*, 1, 2018.

[8] Hamid Karimi, Kaitlin T Torphy, Tyler Derr, Kenneth A Frank, and Jiliang Tang. Understanding and promoting teacher connections in online social media: A case study on pinterest. In *2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, pages 536–541. IEEE, 2020.

[9] Hamid Karimi, Kaitlin T Torphy, Tyler Derr, Kenneth A Frank, and Jiliang Tang. Characterizing teacher connections in online social media: A case study on pinterest. In *Proceedings of the Seventh ACM Conference on Learning@ Scale*, pages 249–252, 2020.

[10] Hamid Karimi, Tyler Derr, Kaitlin Torphy, Kenneth Frank, and Jiliang Tang. A roadmap for incorporating online social media in educational research. *Teachers College Record Year Book*, (14), 2019.

[11] Kaitlin Torphy, Hamid Karimi, Sihua Hu, Frank Kenneth, and Jiliang Tang. Educational research in the 21st century: Leveraging big data to explore teachers' professional behavior and educational resources accessed within pinterest. *The Elementary School Journal*, 2021.

[12] Lauren M Bagdy, Vanessa P Dennen, Stacey A Rutledge, Jerrica T Rowlett, and Shannon Burnick. Teens and social media: A case study of high school students' informal learning

practices and trajectories. In *Proceedings of the 9th International Conference on Social Media and Society*, pages 241–245, 2018.

[13] Christine Greenhow, Beth Robelia, and Joan E Hughes. Learning, teaching, and scholarship in a digital age: Web 2.0 and classroom research: What path should we take now? *Educational researcher*, 38(4):246–259, 2009.

[14] Stacey Rutledge, Vanessa Dennen, and Lauren Bagdy. Exploring adolescent social media use in a high school: Tweeting teens in a bell schedule world. *Teachers College Record*, 121 (14):1–30, 2019.

[15] Vimala Balakrishnan and Chin Lay Gan. Students' learning styles and their effects on the use of social media technology for learning. *Telematics and Informatics*, 33(3):808–821, 2016.

[16] R Arteaga Sánchez, Virginia Cortijo, and Uzma Javed. Students' perceptions of facebook for academic purposes. *Computers & Education*, 70:138–149, 2014.

[17] Hamid Karimi, Tyler Derr, Jiangtao Huang, and Jiliang Tang. Online academic course performance prediction using relational graph convolutional neural network. *International Educational Data Mining Society*, 2020.

[18] Alan Daly, Yi-Hwa Liou, Miguel Del Fresno, Martin Rehm, and Peter Bjorklund Jr. Educational leadership in the twitterverse: Social media, social networks, and the new social continuum. *Teachers College Record*, 121(14):1–20, 2019.

[19] Maeve Duggan, Amanda Lenhart, Cliff Lampe, and Nicole B Ellison. Parents and social media. *Pew Research Center*, 16, 2015.

[20] Uğur Gündüz. The effect of social media on identity construction. *Mediterranean Journal of Social Sciences*, 8(5):85–85, 2017.

[21] Cam Escoffery, Melissa Kenzig, Christel Hyden, and Kristen Hernandez. Capitalizing on social media for career development. *Health promotion practice*, 19(1):11–15, 2018.

[22] Jeffrey Carpenter. Preservice teachers' microblogging: Professional development via twitter. *Contemporary Issues in Technology and Teacher Education*, 15(2):209–234, 2015.

[23] Jeffrey P Carpenter and Daniel G Krutka. Engagement through microblogging: Educator professional development via twitter. *Professional development in education*, 41(4):707–728, 2015.

[24] Catharyn C Shelton and Leanna M Archambault. Who are online teacherpreneurs and what do they do? a survey of content creators on teacherspayteachers. com. *Journal of Research on Technology in Education*, 51(4):398–414, 2019.

[25] Kenneth Frank, Yun-jia Lo, Kaitlin Torphy, and Jihyun Kim. Social networks and educational opportunity. In *Handbook of the Sociology of Education in the 21st Century*, pages 297–316. Springer, 2018.

[26] Sihua Hu, Kaitlin T Torphy, Kim Evert, and John L Lane. From cloud to classroom: Mathematics teachers' planning and enactment of resources accessed within virtual spaces. *Teachers College Record*, 122(6):n6, 2020.

[27] John Lane, Brian Boggs, Zixi Chen, and Kaitlin Torphy. Conceptualizing virtual instructional resource enactment in an era of greater centralization, specification of quality instructional practices, and proliferation of instructional resources. *Teachers College Record*, 121(14): 1–36, 2019.

[28] Kenneth Frank, Diana Brandon, Alan Daly, Christine Greenhow, Sihua Hu, Martin Rehm, and Kaitlin Torphy. Welcome to cloud2class: social media in education. *Teachers College Record*, 121(14):1–12, 2019.

[29] James Robson. Performance, structure and ideal identity: Reconceptualising teachers' engagement in online social spaces. *British Journal of Educational Technology*, 49(3): 439–450, 2018. doi: https://doi.org/10.1111/bjet.12551.

[30] Madeline Will. Looking for more support, new teachers turn to online communities. *Education Week*, 2016.

[31] Ying Zhao, Yong Guo, Yu Xiao, Ranke Zhu, Wei Sun, Weiyong Huang, Deyi Liang, Liuying Tang, Fan Zhang, Dongsheng Zhu, et al. The effects of online homeschooling on children, parents, and teachers of grades 1–9 during the covid-19 pandemic. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, 26:e925591–1, 2020.

[32] Delroy L Paulhus. Measurement and control of response bias. 1991.

[33] Kamal Mahtani, Elizabeth A Spencer, Jon Brassey, and Carl Heneghan. Catalogue of bias: observer bias. *BMJ evidence-based medicine*, 23(1):23, 2018.

[34] Kenneth Frank and Kaitlin Torphy. Social media, who cares? a dialogue between a millennial and a curmudgeon. *Teachers College Record*, 121(14):1–24, 2019.

[35] Mark LaVenia. The state of the instructional materials market: 2019 report. www.edreports.org/resources/article/2019-state-of-the-market-report, 2020.

[36] V. Darleen Opfer, Julia H. Kaufman, and Lindsey E. Thompson. *Implementation of K-12 state standards for mathematics and English Language Arts and literacy: Findings from the American Teacher Panel*. RAND Corporation, Santa Monica, CA, 2016. doi: 10.7249/RR1529-1.

[37] Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760, 2020.

[38] Daniel Voyer and Susan D Voyer. Gender differences in scholastic achievement: a meta-analysis. *Psychological bulletin*, 140(4):1174, 2014.

[39] Marcus A. Winters, Robert C. Haight, Thomas T. Swaim, and Katarzyna A. Pickering. The effect of same-gender teacher assignment on student achievement in the elementary and secondary grades: Evidence from panel data. *Economics of Education Review*, 34:69–75, 2013. ISSN 0272-7757. doi: https://doi.org/10.1016/j.econedurev.2013.01.007.

[40] Alison Kelly. Gender differences in teacher–pupil interactions: a meta-analytic review. *Research in education*, 39(1):1–23, 1988.

[41] Jere Brophy. Interactions of male and female students with male and female teachers. *Gender influences in classroom interaction*, pages 115–142, 1985.

[42] Thomas S Dee. Teachers and the gender gaps in student achievement. *Journal of Human resources*, 42(3):528–554, 2007.

[43] Dario Sansone. Why does teacher gender matter? *Economics of Education Review*, 61: 9–18, 2017.

[44] Jim Duffy, Kelly Warren, and Margaret Walsh. Classroom interactions: Gender of teacher, gender of student, and classroom subject. *Sex roles*, 45(9):579–593, 2001.

[45] Yuqing Liu, Kaitlin T Torphy, Sihua Hu, Jiliang Tang, and Zixi Chen. Examining the virtual diffusion of educational resources across teachers' social networks over time. *Teachers College Record*, 122(6):n6, 2020.

[46] Christine Greenhow, Sarah M Galvin, Diana L Brandon, and Emilia Askari. A decade of research on k-12 teaching and teacher learning with social media: Insights on the state of the field. *Teachers College Record*, 122(6):n6, 2020.

[47] Camille Rutherford. Facebook as a source of informal teacher professional development. 2010.

[48] Emrah Cinkara and Fadime Yalçin Arslan. Content analysis of a facebook group as a form of mentoring for efl teachers. *English Language Teaching*, 10(3):40–53, 2017.

[49] Maria Ranieri, Stefania Manca, and Antonio Fini. Why (and how) do teachers engage in social networks? an exploratory study of professional use of f acebook and its implications for lifelong learning. *British journal of educational technology*, 43(5):754–769, 2012.

[50] Rene Dubos. *Social capital: Theory and research*. Routledge, 2017.

[51] Hüseyin Bicen and Hüseyin Uzunboylu. The use of social networking sites in education: A case study of facebook. *J. UCS*, 19(5):658–671, 2013.

[52] Andrea K Veira, Coreen J Leacock, and S Joel Warrican. Learning outside the walls of the classroom: Engaging the digital natives. *Australasian Journal of Educational Technology*, 30(2), 2014.

[53] Evren Sumuer, Sezin Esfer, and Soner Yildirim. Teachers' facebook use: their use habits, intensity, self-disclosure, privacy settings, and activities on facebook. *Educational Studies*, 40(5):537–553, 2014.

[54] Fu Wen Kuo, Wen Cheng, and Shu Ching Yang. A study of friending willingness on snss: Secondary school teachers' perspectives. *Computers & Education*, 108:30–42, 2017.

[55] Alona Forkosh-Baruch, Arnon Hershkovitz, and Rebecca P Ang. Teacher-student relationship and sns-mediated communication: Perceptions of both role-players. *Interdisciplinary Journal of e-Skills and Lifelong Learning*, 11:273–289, 2015.

[56] Christa S.C. Asterhan and Hananel Rosenberg. The promise, reality and dilemmas of secondary school teacher–student interactions in facebook: The teacher perspective. *Computers & Education*, 85:134–148, 2015. ISSN 0360-1315. doi: https://doi.org/10.1016/j.compedu.2015.02.003.

[57] Baruch Schwarz and Galit Caduri. Novelties in the use of social networks by leading teachers in their classes. *Computers & Education*, 102:35–51, 2016. ISSN 0360-1315. doi: https://doi.org/10.1016/j.compedu.2016.07.002.

[58] Shari Tishman, Eileen Jay, and David N Perkins. Teaching thinking dispositions: From transmission to enculturation. *Theory into practice*, 32(3):147–153, 1993.

[59] C Matt Seimears, Emily Graves, M Gail Schroyer, and John Staver. How constructivist-based teaching influences students learning science. In *The Educational Forum*, volume 76, pages 265–271. Taylor & Francis, 2012.

[60] Ron Blonder and Shelley Rap. I like facebook: Exploring israeli high school chemistry teachers' tpack and self-efficacy beliefs. *Education and Information Technologies*, 22(2): 697–724, 2017.

[61] Radzuwan Ab Rashid. Dialogic reflection for professional development through conversations on a social networking site. *Reflective Practice*, 19(1):105–117, 2018.

[62] Esteban Vázquez Cano. Mobile learning with twitter to improve linguistic competence at secondary schools. *New Educational Review*, 29(3):134–147, 2012.

[63] Carol Van Vooren and Corey Bess. Teacher tweets improve achievement for eighth grade science students. *Journal of Education, Informatics & Cybernetics*, 11(1), 2013.

[64] Anna Noble, Patrick McQuillan, and Josh Littenberg-Tobias. "a lifelong classroom": Social studies educators' engagement with professional learning networks on twitter. *Journal of Technology and Teacher Education*, 24(2):187–213, 2016.

[65] Pamela M Wesely. Investigating the community of practice of world language educators on twitter. *Journal of teacher education*, 64(4):305–318, 2013.

[66] Etienne Wenger. *Communities of practice: Learning, meaning, and identity*. Cambridge university press, 1999.

[67] Virginia G Britt and Trena Paulus. "beyond the four walls of my building": A case study of# edchat as a community of practice. *American Journal of Distance Education*, 30(1):48–59, 2016.

[68] Kathryn Holmes, Greg Preston, Kylie Shaw, and Rachel Buchanan. " follow" me: Networked professional learning for teachers. *Australian Journal of Teacher Education*, 38(12):n12, 2013.

[69] Jeffrey P Carpenter and Daniel G Krutka. How and why educators use twitter: A survey of the field. *Journal of research on technology in education*, 46(4):414–434, 2014.

[70] Jeffrey P Carpenter and Daniel G Krutka. Engagement through microblogging: Educator professional development via twitter. *Professional development in education*, 41(4):707–728, 2015.

[71] Ryan D Visser, Lea Calvert Evering, and David E Barrett. # twitterforteachers: The implications of twitter as a self-directed professional development tool for k–12 teachers. *Journal of Research on Technology in Education*, 46(4):396–413, 2014.

[72] Torrey Trust, Daniel G Krutka, and Jeffrey Paul Carpenter. "together we are better": Professional learning networks for teachers. *Computers & education*, 102:15–34, 2016.

[73] Kerry Davis. Teachers' perceptions of twitter for professional development. *Disability and rehabilitation*, 37(17):1551–1558, 2015.

[74] Carrie R Ross, Robert M Maninger, Kimberly N LaPrairie, and Sam Sullivan. The use of twitter in the creation of educational professional learning opportunities. *Administrative Issues Journal*, 5(1):6, 2015.

[75] Joshua M Rosenberg, Spencer P Greenhalgh, Matthew J Koehler, Erica R Hamilton, and Mete Akcaoglu. An investigation of state educational twitter hashtags (seths) as affinity spaces. *E-Learning and Digital Media*, 13(1-2):24–44, 2016. doi: 10.1177/2042753016672351.

[76] James Paul Gee. *Situated language and learning: A critique of traditional schooling*. routledge, 2012.

[77] Jeffrey P Carpenter and Daniel G Krutka. How and why educators use twitter: A survey of the field. *Journal of research on technology in education*, 46(4):414–434, 2014.

[78] Sihua Hu, Kaitlin T Torphy, Amanda Opperman, Kimberly Jansen, and Yun-Jia Lo. What do teachers share within socialized knowledge communities: A case of pinterest. *Journal of Professional Capital and Community*, 2018.

[79] Kaitlin Torphy, Sihua Hu, Yuqing Liu, and Zixi Chen. Teachers turning to teachers: teacherpreneurial behaviors in social media. *American Journal of Education*, 127(1):49–76, 2020.

[80] Jeffrey Carpenter, Amanda Cassaday, and Stefania Monti. Exploring how and why educators use pinterest. In *Society for Information Technology & Teacher Education International Conference*, pages 2222–2229. Association for the Advancement of Computing in Education (AACE), 2018.

[81] John S Kendall. *Understanding common core state standards*. ASCD, 2011.

[82] V. Darleen Opfer, Julia H. Kaufman, and Lindsey E. Thompson. *Implementation of K-12 State Standards for Mathematics and English Language Arts and Literacy: Findings from the American Teacher Panel*. RAND Corporation, Santa Monica, CA, 2016. doi: 10.7249/RR1529-1.

[83] Jinyoung Han, Daejin Choi, A-Young Choi, Jiwon Choi, Taejoong Chung, Ted Taekyoung Kwon, Jong-Youn Rha, and Chen-Nee Chuah. Sharing topics in pinterest: understanding content creation and diffusion behaviors. In *Proceedings of the 2015 ACM on Conference on Online Social Networks*, pages 245–255, 2015.

[84] Eric Gilbert, Saeideh Bakhshi, Shuo Chang, and Loren Terveen. " i need to try this"? a statistical overview of pinterest. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2427–2436, 2013.

[85] Jinyoung Han, Daejin Choi, Byung-Gon Chun, Ted Kwon, Hyun-chul Kim, and Yanghee Choi. Collecting, organizing, and sharing pins in pinterest: interest-driven or social-driven? *ACM SIGMETRICS Performance Evaluation Review*, 42(1):15–27, 2014.

[86] Shuo Chang, Vikas Kumar, Eric Gilbert, and Loren G Terveen. Specialization, homophily, and gender in a social curation site: Findings from pinterest. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 674–686, 2014.

[87] Daehoon Kim, Jae-Gil Lee, and Byung Suk Lee. Topical influence modeling via topic-level interests and interactions on social curation services. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 13–24. IEEE, 2016.

[88] Regina Kasakowskij, Thomas Kasakowskij, and Kaja Fietkiewicz. Pinterest: A unicorn among social media? an investigation of the platform's quality and specifications. In *ECSM 2020 8th European Conference on Social Media*, page 399. Academic Conferences and publishing limited, 2020.

[89] Stephanie Schroeder, Rachelle Curcio, and Lisa Lundgren. Expanding the learning network: How teachers use pinterest. *Journal of Research on Technology in Education*, 51(2):166–186, 2019. doi: 10.1080/15391523.2019.1573354.

[90] Amanda Sawyer, Lara Dick, Emily Shapiro, and Tabitha Wismer. The top 500 mathematics pins: An analysis of elementary mathematics activities on pinterest. *Journal of Technology and Teacher Education*, 27(2):235–263, 2019.

[91] Kaitlin T Torphy, Diana L Brandon, Alan J Daly, Kenneth A Frank, Christine Greenhow, S Hua, and Martin Rehm. Social media, education, and digital democratization. *Teachers College Record*, 122(6):1–7, 2020.

[92] Christine Greenhow, Sarah Galvin, Emilia Askari, and Diana Brandon. # cloud2class: The disruption and reorganization of educational resources with social media. *American Journal of Education*, 127(1):1–11, 2020.

[93] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

[94] Lev Muchnik, Sen Pei, Lucas C Parra, Saulo DS Reis, José S Andrade Jr, Shlomo Havlin, and Hernán A Makse. Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Scientific reports*, 3(1):1–8, 2013.

[95] Fantine Mordelet and J-P Vert. A bagging svm to learn from positive and unlabeled examples. *Pattern Recognition Letters*, 37:201–209, 2014.

[96] Jinfeng Yi, Cho-Jui Hsieh, Kush Varshney, Lijun Zhang, and Yao Li. Scalable demand-aware recommendation. *arXiv preprint arXiv:1702.06347*, 2017.

[97] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–220, 2008.

[98] Fengxiang He, Tongliang Liu, Geoffrey I Webb, and Dacheng Tao. Instance-dependent pu learning by bayesian optimal relabeling. *arXiv preprint arXiv:1808.02180*, 2018.

[99] Hwanjo Yu, Jiawei Han, and Kevin Chen-Chuan Chang. Pebl: positive example based learning for web page classification using svm. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 239–248, 2002.

[100] Dino Ienco and Ruggero G Pensa. Positive and unlabeled learning in categorical data. *Neurocomputing*, 196:113–124, 2016.

[101] Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069*, 2018.

[102] Courtland VanDam, Pang-Ning Tan, Jiliang Tang, and Hamid Karimi. Cadet: A multi-view learning framework for compromised account detection on twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 471–478. IEEE, 2018.

[103] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[104] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[105] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[106] Marthinus Du Plessis, Gang Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In *International conference on machine learning*, pages 1386–1394. PMLR, 2015.

[107] Ryuichi Kiryo, Gang Niu, Marthinus C du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. *arXiv preprint arXiv:1703.00593*, 2017.

[108] Liwei Jiang, Dan Li, Qisheng Wang, Shuai Wang, and Songtao Wang. Improving positive unlabeled learning: Practical aul estimation and new training method for extremely imbalanced data sets. *arXiv preprint arXiv:2004.09820*, 2020.

[109] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

[110] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65, 2018.

[111] Hamid Karimi, Tyler Derr, and Jiliang Tang. Characterizing the decision boundary of deep neural networks. *arXiv preprint arXiv:1912.11460*, 2019.

[112] Kathleen Elizabeth Kaump Truitt. The relationship between elementary school administrators' and teachers' perceptions of the influence of male teachers and schools' male student achievement growth in english language arts. 2019.

[113] Yena Kim and Allyson J Weseley. The effect of teacher gender and gendered traits on perceptions of elementary school teachers. *Journal of Research in Education*, 27(1):114–133, 2017.

[114] Margaret H Cooney and Mark T Bittner. Men in early childhood education: Their emergent issues. *Early Childhood Education Journal*, 29(2):77–82, 2001.

[115] Shaaista Moosa and Deevia Bhana. Men teaching young children:"you can never be too sure what their intentions might be". *Oxford Review of Education*, 46(2):169–184, 2020.

[116] Bryan G Nelson. The importance of men teachers: And reasons why there are so few. a survey of members of naeyc. 2002.

[117] National Center for Education Statistics. Teacher characteristics and trends. https://nces.ed.gov/fastfacts/display.asp?id=28, 2020. Accessed: 2021-03-20.

[118] Organisation for Economic Co-operation and Development. Distribution of teachers by age and gender. https://stats.oecd.org/Index.aspx?DataSetCode=EAG_PERS_SHARE_AGE, 2018. Accessed: 2021-02-20.

[119] Simon Brownhill. 'build me a male role model!'a critical exploration of the perceived qualities/characteristics of men in the early years (0–8) in england. *Gender and Education*, 26(3):246–261, 2014.

[120] Kevin McGrath and Mark Sinclair. More male primary-school teachers? social benefits for boys and girls. *Gender and Education*, 25(5):531–547, 2013.

[121] Hootsuite DataReportal and We Are Social. Distribution of pinterest users worldwide as of january 2021, by gender. 2021. Accessed: April 08, 2021.

[122] Emily S Johnson. *Feminism, Self-presentation, and Pinterest: The Labor of Wedding Planning*. Lexington Books, 2020.

[123] Elana Levine. *Cupcakes, pinterest, and ladyporn: Feminized popular culture in the early twenty-first century*. University of Illinois Press, 2015.

[124] Amanda Friz and Robert W Gehl. Pinning the feminine user: gender scripts in pinterest's sign-up interface. *Media, Culture & Society*, 38(5):686–703, 2016.

[125] Raphael Ottoni, Joao Paulo Pesce, Diego Las Casas, Geraldo Franciscani Jr, Wagner Meira Jr, Ponnurangam Kumaraguru, and Virgilio Almeida. Ladies first: Analyzing gender roles and behaviors in pinterest. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, 2013.

[126] Spencer P Greenhalgh and Matthew J Koehler. 28 days later: Twitter hashtags as "just in time" teacher professional development. *TechTrends*, 61(3):273–281, 2017.

[127] Paul T Costa Jr, Antonio Terracciano, and Robert R McCrae. Gender differences in personality traits across cultures: robust and surprising findings. *Journal of personality and social psychology*, 81(2):322, 2001.

[128] Madhura Ingalhalikar, Alex Smith, Drew Parker, Theodore D Satterthwaite, Mark A Elliott, Kosha Ruparel, Hakon Hakonarson, Raquel E Gur, Ruben C Gur, and Ragini Verma. Sex differences in the structural connectome of the human brain. *Proceedings of the National Academy of Sciences*, 111(2):823–828, 2014.

[129] Sudip Mittal, Neha Gupta, Prateek Dewan, and Ponnurangam Kumaraguru. The pin-bang theory: Discovering the pinterest world. *arXiv preprint arXiv:1307.4952*, 2013.

[130] William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38, 2010.

[131] J Nguyen. A teacher's new best friend: Amazon inspire. *Edudemic: Connecting*, 2016.

[132] Hailley Griffis. Marketing: Common words, popular times, plus 4 experiments to try. 2021. Accessed: May 12, 2021.

[133] K Torphy and S Hu. Social media in education: Curation within and outside the schoolhouse. *Handbook on Social Media Analytics: Advances and Applications*.

[134] William R Penuel, Margaret Riel, Ann Krause, and Kenneth A Frank. Analyzing teachers' professional interactions in a school as social capital: A social network approach. *Teachers college record*, 111(1):124–163, 2009.

[135] Hamid Karimi, Courtland VanDam, Liyang Ye, and Jiliang Tang. End-to-end compromised account detection. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 314–321. IEEE, 2018.

[136] Aaron Brookhouse, Tyler Derr, Hamid Karimi, H Russell Bernard, and Jiliang Tang. Road to the white house: Analyzing the relations between mainstream and social media during the us presidential primaries. *arXiv preprint arXiv:2009.09307*, 2020.

[137] Hamid Karimi, Tyler Derr, Aaron Brookhouse, and Jiliang Tang. Multi-factor congressional vote prediction. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 266–273, 2019.

[138] Tyler Derr, Hamid Karimi, Xiaorui Liu, Jiejun Xu, and Jiliang Tang. Deep adversarial network alignment. *arXiv preprint arXiv:1902.10307*, 2019.

[139] Elizabeth Homan. The shifting spaces of teacher relationships: Complementary methods in examinations of teachers' digital practices. *Journal of Technology and Teacher Education*, 22(3):311–331, 2014.

[140] Fei Gao and Lan Li. Examining a one-hour synchronous chat in a microblogging-based professional development community. *British Journal of Educational Technology*, 48(2): 332–347, 2017.

[141] Stephen J Aguilar, Joshua Rosenberg, Spencer Greenhalgh, Tim Fütterer, Alex Lishinski, and Christian Fischer. A different experience in a different moment? teachers' social media use before and during the covid-19 pandemic. 2021.

[142] Hamid Karimi, Tyler Derr, Kaitlin T Torphy, Kenneth A Frank, and Jiliang Tang. Towards improving sample representativeness of teachers on online social media: A case study on pinterest. In *International Conference on Artificial Intelligence in Education*, pages 130–134. Springer, 2020.

[143] Martijn P van den Heuvel and Olaf Sporns. Network hubs in the human brain. *Trends in cognitive sciences*, 17(12):683–696, 2013.

[144] Tracy C Russo and Joy Koesten. Prestige, centrality, and learning: A social network analysis of an online class. *Communication Education*, 54(3):254–261, 2005.

[145] Stephen P Borgatti. Centrality and network flow. *Social networks*, 27(1):55–71, 2005.

[146] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[147] Paul F Lazarsfeld, Robert K Merton, et al. Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society*, 18(1):18–66, 1954.

[148] Luca Maria Aiello, Alain Barrat, Rossano Schifanella, Ciro Cattuto, Benjamin Markines, and Filippo Menczer. Friendship prediction and homophily in social media. *ACM Transactions on the Web (TWEB)*, 6(2):1–33, 2012.

[149] Halil Bisgin, Nitin Agarwal, and Xiaowei Xu. A study of homophily on social media. *World Wide Web*, 15(2):213–232, 2012.

[150] Luis Abreu and Doh-Shin Jeon. Homophily in social media and news polarization. 2020.

[151] Gabriele Chierchia and Giorgio Coricelli. The impact of perceived similarity on tacit coordination: propensity for matching and aversion to decoupling choices. *Frontiers in behavioral neuroscience*, 9:202, 2015.

[152] Noah P Mark. Culture and competition: Homophily and distancing explanations for cultural niches. *American sociological review*, pages 319–345, 2003.

[153] Damon Centola, Robb Willer, and Michael Macy. The emperor's dilemma: A computational model of self-enforcing norms. *American Journal of Sociology*, 110(4):1009–1040, 2005.

[154] Munmun De Choudhury, Hari Sundaram, Ajita John, Doree Duncan Seligmann, and Aisling Kelliher. " birds of a feather": Does user homophily impact information diffusion in social media? *arXiv preprint arXiv:1006.1702*, 2010.

[155] Mustafa Yavaş and Gönenç Yücel. Impact of homophily on diffusion dynamics over social networks. *Social Science Computer Review*, 32(3):354–372, 2014.

[156] Fariba Karimi, Mathieu Génois, Claudia Wagner, Philipp Singer, and Markus Strohmaier. Homophily influences ranking of minorities in social networks. *Scientific reports*, 8(1):1–12, 2018.

[157] Damon Centola, Juan Carlos Gonzalez-Avella, Victor M Eguiluz, and Maxi San Miguel. Homophily, cultural drift, and the co-evolution of cultural groups. *Journal of Conflict Resolution*, 51(6):905–929, 2007.

[158] David Laniado, Yana Volkovich, Karolin Kappler, and Andreas Kaltenbrunner. Gender homophily in online dyadic and triadic relationships. *EPJ Data Science*, 5:1–23, 2016.

[159] Thomas Bedorf. *Dimensionen des Dritten: sozialphilosophische Modelle zwischen Ethischem und Politischem*. Wilhelm Fink, 2003.

[160] George Herbert Mead. *Mind, self and society*, volume 111. Chicago University of Chicago Press., 1934.

[161] Jinyoung Han, Daejin Choi, Jungseock Joo, and Chen-Nee Chuah. Predicting popular and viral image cascades in pinterest. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 2017.

[162] Thi Bich Ngoc Hoang and Josiane Mothe. Predicting information diffusion on twitter–analysis of predictive features. *Journal of computational science*, 28:257–264, 2018.

[163] Bo Wu and Haiying Shen. Analyzing and predicting news popularity on twitter. *International Journal of Information Management*, 35(6):702–711, 2015.

[164] Andrey A Dobrynin, Roger Entringer, and Ivan Gutman. Wiener index of trees: theory and applications. *Acta Applicandae Mathematica*, 66(3):211–249, 2001.

[165] Harry Wiener. Structural determination of paraffin boiling points. *Journal of the American chemical society*, 69(1):17–20, 1947.

[166] Christine Greenhow, Sarah M Galvin, and K Bret Staudt Willet. What should be the role of social media in education? *Policy Insights from the Behavioral and Brain Sciences*, 6(2): 178–185, 2019.

[167] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. Virality prediction and community structure in social networks. *Scientific reports*, 3(1):1–6, 2013.

[168] Piia Varis and Jan Blommaert. Conviviality and collectives on social media: Virality, memes, and new social structures. *Multilingual Margins: A journal of multilingualism from the periphery*, 2(1):31–31, 2015.

[169] Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A Zighed. Information diffusion in online social networks: A survey. *ACM Sigmod Record*, 42(2):17–28, 2013.

[170] Qian Li, Tao Zhou, Linyuan Lü, and Duanbing Chen. Identifying influential spreaders by weighted leaderrank. *Physica A: Statistical Mechanics and its Applications*, 404:47–55, 2014.

[171] Ling-ling Ma, Chuang Ma, Hai-Feng Zhang, and Bing-Hong Wang. Identifying influential spreaders in complex networks based on gravity formula. *Physica A: Statistical Mechanics and its Applications*, 451:205–212, 2016.

[172] Frank Bauer and Joseph T Lizier. Identifying influential spreaders and efficiently estimating infection numbers in epidemic models: A walk counting approach. *EPL (Europhysics Letters)*, 99(6):68007, 2012.

[173] William Penuel, Kenneth Frank, Min Sun, Chong Kim, and Corinne Singleton. The organization as a filter of institutional diffusion. *Teachers college record*, 115(1):1–33, 2013.

[174] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[175] Anastasia Mochalova and Alexandros Nanopoulos. On the role of centrality in information diffusion in social networks. 2013.

[176] Mary E Yepez. An observation of gender-specific teacher behavior in the esl classroom. *Sex Roles*, 30(1):121–133, 1994.

[177] Kelly Jones, Cay Evans, Ronald Byrd, and Kathleen Campbell. Gender equity training and teacher behavior. *Journal of Instructional Psychology*, 27(3):173–173, 2000.

[178] Heather Antecol, Ozkan Eren, and Serkan Ozbeklik. The effect of teacher gender on student achievement in primary school: Evidence from a randomized experiment. 2012.

[179] Hamid Karimi and Jiliang Tang. Learning hierarchical discourse-level structure for fake news detection. *arXiv preprint arXiv:1903.07389*, 2019.

[180] Hamid Karimi, Proteek Roy, Sari Saba-Sadiya, and Jiliang Tang. Multi-source multi-class fake news detection. In *Proceedings of the 27th international conference on computational linguistics*, pages 1546–1557, 2018.

[181] Haochen Liu, Wei Jin, Hamid Karimi, Zitao Liu, and Jiliang Tang. The authors matter: Understanding and mitigating implicit bias in deep text classification. *arXiv preprint arXiv:2105.02778*, 2021.

[182] Haochen Liu, Yiqi Wang, Wenqi Fan, Xiaorui Liu, Yaxin Li, Shaili Jain, Anil K Jain, and Jiliang Tang. Trustworthy ai: A computational perspective. *arXiv preprint arXiv:2107.06641*, 2021.