Modeling Online User Interactions and their Offline Effects on Socio-Technical Platforms

By

Hitkul

Under the supervision of

Dr. Rajiv Ratn Shah, IIIT Delhi

Prof. Ponnurangam Kumaraguru, IIIT Hyderabad

Indraprastha Institute of Information Technology Delhi

March, 2024

Modeling Online User Interactions and their Offline Effects on Socio-Technical Platforms

By

Hitkul

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

to
Indraprastha Institute of Information Technology Delhi

March, 2024

Certificate

This is to certify that the thesis titled "**Modeling Online User Interactions and their Offline Effects on Socio-Technical Platforms**" being submitted by **Hitkul** to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of Doctor of Philosophy, is an original research work carried out by her under our supervision. In our opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

Advisor
Dr. Rajiv Ratn Shah
Associate Professor
Indraprastha Institute of Information Technology, Delhi
New Delhi 110020

Advisor
Prof. Ponnurangam Kumaraguru
Professor
International Institute of Information Technology, Hyderabad
Hyderabad, Telangana 500032

Acknowledgements

First, I would like to thank my parents for their unconditional support. Their innumerable sacrifices have allowed me to pursue education and follow my dream. I would be unable to survive the hardships of the last six years without their love and support. Whatever I am or will become in my life is because of them. Thank you, Maa and Papa. I would also like to thank my sister Divya, brother-in-law Abhishek, and niece Abhidi. You guys were always a call away to help and guide me through any matter. Running around and playing with Abhidi during her yearly visits will always be some of the most cherished moments of my life.

I would never enroll in a PhD program if not for both of my advisors. Thank Rajiv, sir, for trusting me as one of the first students in his lab and motivating me to pursue a PhD. I can not acknowledge how many opportunities he has provided me, from internships to industry projects and the freedom to pursue my research interests. I am eternally grateful to him for the help and opportunity he extended, which helped me grow personally and professionally. When I felt lost in my journey, PK came as a rescue. He taught me to be an independent researcher, but more importantly, a kind person. Every conversation with him left me introspecting for hours and changed the direction of my work and life for good. Wishing not to disappoint him has been one of the biggest motivators that kept me through tough times. Sending him a cold email to take me in as his PhD. Student has been one of the best decisions of my life. I would also like to thank my committee members, Arun Balaji, and Raghava Mutharaju, for their comments and feedback. I would also like to congratulate TCS Research for supporting my PhD. I want to thank Ashwini Ma'am at TCS for her continued support and assistance through all types of funding queries.

I would also like to thank Hemank Lamba. I was lucky to be able to call him my mentor. Anything you read in this thesis that makes even an iota of sense is because of his guidance. He handholded and taught me the whole research process and has influenced how I work to date. I admire his sheer hard work and focus. If it were not for Hemank, I would never have finished my thesis.

During my PhD, I was fortunate to collaborate with a lot of amazing people - Kiran Garimella, Debanjan Mahata, Omkar Gurjar, Aanshul Sadaria, Kanay Gupta, Shashank Srikanth, Anjali Bhavan, Tanmay Bansal, Anmol, Aryamann Tomar, and Pranjal Kandhari. I have learned something valuable from each one, and I thank all of them.

One of the most exciting parts of my graduate journey was internships. I thank Prof. Shin'ichi Satoh for selecting me to attend the NII Tokyo, Japan internship program. The six months I spent in Tokyo were one of the best. I got to interact with many bright minds and experience the great country of Japan. I spent the summer of 2021 as a research intern at Goldman Sachs. I want to thank Raj Sharma for giving me the opportunity. It is rare to find people like Raj who contribute so much of their time and effort to help the student community. I would also like to thank my buddy and manager at Goldman Sachs, Sarang and Eliot Brenner. Along with technical, they taught me so much about working in the industry.

My late evening conversations with Sarang about research have been some of my most insightful discussions. I spent the second half of my 2023 at Flipkart. I got to work with great people like Abinaya and Soham. They taught me so much about getting work done while managing various stakeholders in the industry. I would also like to thank my managers and mentors at Flipkart, Satyajit Banerjee, and Muthusamy Chelliah, for their continuous motivation and support.

I would also like to thank the staff members at IIT Delhi for their continuous help in administrative and IT-related matters—especially Bhawani Shah, Ravi Bhasin, Rajeev sir, Jyoti, and Binu. I want to thank my amazing colleagues at MIDAS and Precog. Playful banter all day and the late-night hangouts at the Piano Man with Pakhi, Mohit, and Manraj made the lab feel like a second home to me. I admire and have learned so much from all of you guys. Chatting about work, life, and politics to Mann and Ritwik came as much-needed breaks from work. I would also like to thank Avinash from Precog; during the second wave of COVID-19, we were the only two on campus from the lab, and I am glad I got to know you so much better. I would also like to thank Prashant and Anmol from Precog; during my stay at IIIT Hyderabad, these guys were always around and never let me miss Delhi too much!

It would be highly unfair on my part not to thank my undergraduate mentor - Viomesh sir, for pushing me as a fresh CS grad and introducing me to Rajiv. Prof. Maheshwar Dwivedy, thank you for allowing me to pursue all the projects and for your unrestricted support. Finally, Prof. Sudip Sanyal introduced me to the world of research and was one of the first people who actively encouraged me to pursue graduate studies.

I can not emphasize the role friends play in my life, but I have met some fantastic people I am fortunate to call friends. First and foremost, Karmanya. You have been a friend, mentor, and also agony aunt. I can not believe it is been almost a decade since we met. I would not have survived this year without you. Love you bhai and thank you for everything. Tarun, I have never met, and I doubt I will ever meet someone as kind as you. Your passion and undivided attention toward your dream of Jujutsu inspires me daily to work hard towards my dreams. I am so fortunate to call you my brother and be a part of your life's inspiring story. I am also lucky to have met Anwesha through Tarun. You guys are the reason why I call Delhi my home. Hanging out with you two at Blue Tokai, Leo's, and Saket is what I looked forward to the most. Thank you for always welcoming me with open arms in your house; these last few years would have been much harder without you guys. I would also like to thank Anurag Joshi for being there for me since literally day one of my undergraduate. From our late-night walks around the campus to now, staying over at your house and chatting with you and Anusha. You keep me grounded and remind me what is essential in life.

I am grateful for my friend back home, Hitesh Sharma. Hitesh (Gatu mera bhai!) I do not even remember how long we have been friends. It is not possible to put into words what you mean to me. From dragging puncture cycles across the streets of Rewari to "cruising" in cars on NH8, you have been a part of all the ups and downs in my life. No matter what life brings,

life will be good if I hang out with you at Vaishali. Thank you for everything. Additionally, I would like to thank my school friend Rahul. Meeting you every time I went home took me back to the carefree school days and reminded me of how it all started. I would also like to thank Garima. When I shifted to Delhi in early 2018, I was going through a very rough time, and I would not have stayed back in the city if it were not for you. I never thought a random meeting with a classmate from secondary school would grow into such a great bond. I am forever grateful that I got to know you better and can call you a friend. I also thank Saurav Varma, Vaibhav Goyal, and Aayush Nagpal for the fabulous house parties. It is not an overstatement, but there was never a dull moment with you guys.

I attended my internship in Tokyo with Himanshu. Even though we were strangers, it felt like we had been friends for years since day one. From housemates in Tokyo to now neighbors in Bangalore, I cherish every moment I have spent with you and look forward to many more. I met Atish in the last year of my PhD, and soon, I spent more time in his lab than I did on my own. You are the best road trip partner one can ask for. I am looking forward to many more thousands of kilometers with you.

Last but not least, Shivangi. We both joined IIIT Delhi on the same day. From lab mates to friends, then girlfriend, and now my wife! This journey has been as much yours as it has been mine. I would never have completed this journey without your love and support, and I am grateful to have you with me for the rest of my life. Thank you for being in my life.

आप बस साथ बनाये रखियेगा अभी तो मै लम्बा चलूँगा।
जाकिर खान

Abstract

Do online interactions trigger reactions back in the offline world? How can these reactions be detected and quantified? Specifically, what insights can be extracted for users, platform owners, and policymakers to minimize the potential harm of such reactions?

Society functions based on the complex interactions between individuals, communities, and organizations. We communicate with each other to build family, friendship, and romantic relationships; to seek or provide advice and education; to execute trade and commerce. People unite to form organizations that drive economic activity, govern states, and provide social benefits. The advent of the Internet has enabled these interactions to move online. A website or an application that facilitates the digitization of social interactions is called a *socio-technical* platform. For instance, individuals converse with each other via direct messaging applications (e.g., WhatsApp, Telegram), share thoughts, and gather feedback from communities (e.g., Reddit, Twitter, Youtube). Trade of goods occurs via e-commerce (e.g., Flipkart, Amazon) and online marketplaces (e.g., Google Play store). At times interactions happening in the online world, trigger reactions in the offline world, which we call *overflow*. Such overflows can have either a positive or negative impact. Socio-technical platforms save every interaction and associated metadata, providing a unique opportunity to analyze rich data at scale. Discover interaction patterns, detect and quantify overflow of interactions, and extract insights for users and policymakers.

This thesis aims to study the interactions by keeping the individual as the focal point. We focus on three broad forms of interactions - i) the effect online community feedback can have on individual offline actions, ii) organizations leveraging individual customers' online presence to optimize business processes, and iii) how data from tracking platforms can be used to uncover the strategies behind successful users. In the first part, we work on three scenarios - (a) How does community feedback affect an individual future drug consumption frequency in a drug community forum?; (b) What changes does an individual undergo immediately after getting sudden popularity in Online social media? What actions help in maintaining popularity for longer?; (c) Dynamics of interactions in an online COVID-19 support group and what affects a user's longevity in the community. In the second part, we leverage online information about a user to improve the prediction of Return-to-Origin [1] orders in the e-commerce platform. Finally, in the third part, we leverage data from a habit-tracking platform to unveil what user actions lead to success in habit-building pursuits.

[1] https://easyinsights.ai/blog/return-to-origin-rto-why-is-it-a-crucial-metric-for-ecommerce-businesses/

Contents

List of Figures

List of Tables

Part I

Introduction and Background

Chapter 1

Introduction

Interactions are the bedrock of society. Life and communities structured around complex interactions. We communicate with each other to build family, friendship, and romantic relationships; to seek or provide advice and education; to execute trade and commerce. People unite to form organizations that drive economic activity, govern states, and provide social benefits. Writing allowed us to preserve ideas beyond the lifespan of an individual. It led to the inheritance of traditions, expansions of religion, and the evolution of ideologies. Writing combined with printing democratized knowledge and enabled the spread of ideas, leading to cultural exchange and scientific revolutions. Human interactions are multidimensional, and different methods are used to structure these interactions based on the study area. As part of our work, we categorize interactions based on the entities involved. Our three interest entities are 1) Individuals, 2) Communities, and 3) Organizations.

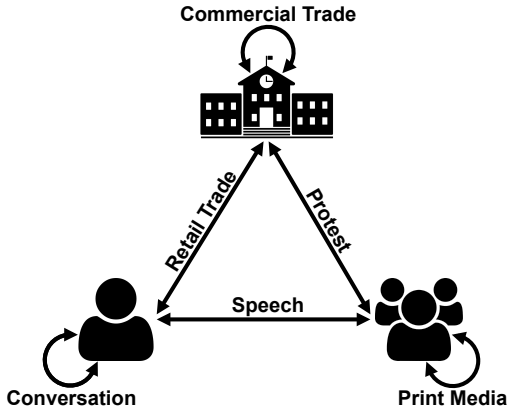


Figure 1.1: Different combinations of societal interactions. Each vertex represents an entity, individual, community, or organization. Edges depict a possible form of interaction between the entities.

We structure interactions in a trilateral arrangement as shown in Figure 1.1. Each vertex is an entity, and edges depict different possible interactions. Individuals indulge with other individuals in private conversations [Conversation edge in Fig. 1.1], trade with organizations [Retail Trade edge in Fig. 1.1], and impart their thoughts to a community via speeches/lectures [Speech edge in Fig. 1.1]. Communities come together to share ideas [Print Media edge in Fig. 1.1], conduct events, display their objection towards organizations via protests and help other communities as part of social service [Protest edge in Fig. 1.1]. Similarly, organizations (governments, companies, and NGOs) indulge in trade of goods and services, conduct social service, and govern communities [Commercial Trade edge in Fig. 1.1]. Researchers in the fields of Sociology, Psychology, Cognitive, and Economic sciences have long been interested in understanding the dynamics and characteristics of these interactions. The resulting literature has helped us design better communication structures, effective and efficient public policy, and derive economic growth.

Recently, an increase in accessibility to the Internet has led to the rise of platforms that move interactions online, e.g., WhatsApp, Twitter, Flipkart, and StickK. In our work, a website or an application that facilitates the digitization of social interactions is called *Socio-Technical* platform. The focus of the thesis is to study the impact and implications of interaction facilitated by socio-technical platforms.

1.1 Advent of Digital Interactions

Social networking websites such as Twitter, Reddit, and TikTok offer users the ability to share their thoughts and ideas through text, images, and videos. Users can engage with content by liking or commenting on it, as well as connect with other users through social connections like friends and followers or chatting. These platforms have had a significant impact on modern society, affecting areas such as politics, economics, and overall societal norms. Digital interactions are not limited to social networking systems. For example, Software developers use GitHub to collaborate with each other by posting their software code to repositories. People shop on e-commerce platforms like Flipkart and Amazon, consume content via streaming platforms (YouTube, Twitch), and use Fitbit and Apple Health to track and share exercise stats. Since, focus of the thesis is to study the interaction facilitated by socio-techincal platforms, we start by describing some categories and examples of the same.[1] Figure 1.2 shows offline interactions and their corresponding socio-technical platforms.

- **Individual centric**: Platforms centered around the interactions and conversations of two individuals, like *WhatsApp*, *Tinder*, and *Telegram*. There has also been a rise in tracking platforms like *step trackers* and *StickK* that can be put in this category. These platforms allow users to log data about their actions and share them with others.
- **Community centric**: Platforms that are used to build and facilitate communities like *Reddit*, *Change.org* and *Team-BHP*. Platforms like *Twitter* are also included in this category, allowing users to reach a broader audience and garner a following. Typically, these platforms would have a mechanism for an individual to present a thought and a feedback mechanism (like, share, and comments) for the community to react.
- **Organization centric**: Platforms that enable organizations to interact with their beneficiaries/customers/other organizations. E-commerce platforms like *Amazon* and *Flipkart*; B2B e-commerce like *BigCommerce* and *Moglix*; Online education platforms like *Byjus* and *Coursera*; and servicing platforms like *Uber* and *Swiggy*. State websites that allow users to interact with government departments such as *Parivahan Sewa* [2] also belong to this category.

[1] As platforms grow, it can be challenging to classify them into a single category. For Example, WhatsApp started as a direct messaging service, but now also supports groups, payments and shopping features. Categorization in this thesis is based on the primary function of the platform.

[2] Indian equivalent of DMV in the USA

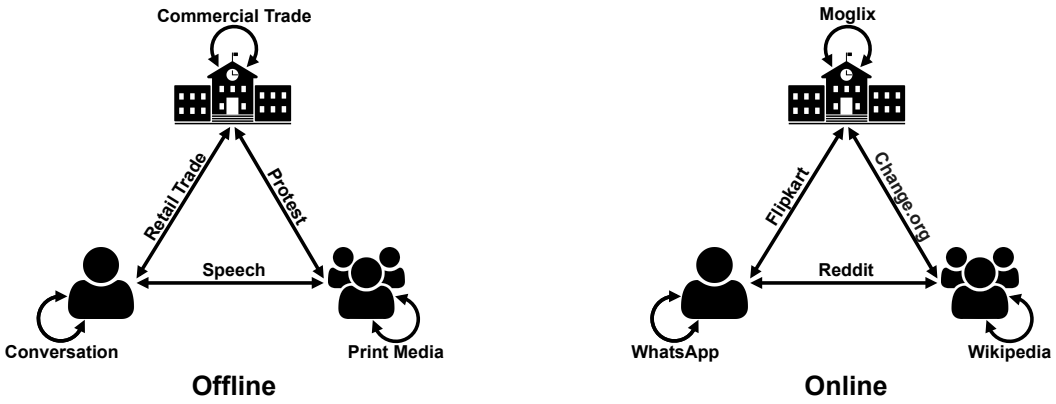


Figure 1.2: Offline interactions and corresponding online platforms.

Socio-technical platforms store each interaction occurring on their platform, including all types of user interactions with any other element (users, posts, items, comments, etc.) on the platform. Storing these interactions provides us with huge and rich data, which, before the existence of these platforms, was rarely available. The availability of such rich data at a large scale for platforms that govern key aspects of human life provides us with a unique opportunity to identify various previously unexplored and useful patterns. One key application of exploring interesting patterns in this data lies in the field of computational social science [103]. The existing sociological theories were developed over a small sample of humans and it is uncertain as to how they scale given such rich longitudinal data for different platforms. The availability of the interaction data allows us to argue about human behavior through the lens of these omnipresent social and technical systems. Mining useful and interesting patterns is also crucial to the platform owners. They aim to ensure that users remain engaged and use the system. Discovery based on how users are using the system can allow the platform owners to redesign and introduce features that can increase the longevity of users on the platform.

We discussed various types of offline interactions and their online counterparts. However, the two worlds are not mutually exclusive. Feedback and trends from online interactions regularly trigger reactions in the offline world, where individuals and organizations morph their actions to achieve a favorable online persona. In our work, we call the phenomenon of online reactions affecting the offline world as *overflow*. Sometimes, such overflows can lead to positive changes like alternative career options (e.g., content creator), monitory growth, and increased reach/awareness. On the other hand, overflows can lead to profound negative implications like self-harm (KiKi challenge, Blue Whale challenge), financial fraud, and social unrest. To maximize the positive overflow and minimize the negatives, it is important to study i) the loop between online interactions and offline actions, ii) devise algorithms to detect and quantify the overflow, and iii) suggest measures for involved entities, platforms, and policymakers.

In our work, we aim to study the interactions by keeping the individual as the focal point. The thesis, focuses on three broad forms of interactions - i) the effect online community feedback can have on individual offline actions, ii) organizations leveraging individual customers' online presence to optimize business processes, and iii) how data from tracking platforms can be used to uncover the strategies behind successful users.

1.2 Overview and Contributions

1.2.1 Individual-Community Interactions

Large online platforms like Reddit and X (previously Twitter) provide a unique opportunity to study user behavior at scale. The scale of these platforms not only enables us to study a wide variety of topics but also provides data for long temporal periods. As mentioned, such rich data can help us verify previously limitedly tested sociological theories and provide insights for platform/community owners. In this part of the thesis, we study the dynamics and behavior of a user within a community. We evaluate what affects the longevity of a user in a community, how community feedback affects users' future actions, and if a user receives elevated popularity in a community, what they can do to sustain it.

Virality is an interesting phenomenon on social media. *Being viral* is a colloquial term representing a user suddenly receiving disproportionately large numbers of impressions from the wider community. In Chapter 3, based on our paper [64], we explore how a user's behavior changes after receiving a popularity shock on a short video platform. After a popularity shock, users tend to increase their posting frequency and pivot their content to align more with the content piece that gained popularity. Further, the users who inculcate continuous engagement with their followers and maintain a narrow balance between unique content but in a similar scope as the viral content piece tend to maintain their popularity for the longest time. We also proposed a modified anomaly detection algorithm to detect popularity shocks and released the first-ever large-scale dataset of a short video platform containing the metadata related to thousands of users and millions of posts.

Another limitation of older sociology literature is the limited volume of studies around taboo topics like drugs, sex, racism, etc, because of the logistic, ethical, and legal challenges surrounding these topics. However, the pseudo-anonymous nature of social media allows users to converse about such topics, giving researchers an opportunity to study behavior around such topics. Chapter 4, based on our paper [78], we collect data from 10 drug-related subreddits spanning over 12 years. Grounded in theories like the Primacy effect, operant conditioning, and Edgework, we evaluate how feedback from the community affects users' future decisions to consume drugs. We also perform a parallel user study, helping us compare user opinions with statistical findings. We found out that about 80% of users participating in drug-related subreddits have posted content indicating drug consumption offline. Though in our user study, users say that positive feedback has little effect on their decision to consume drugs in the future, our causal experiments indicate that users who receive positive feedback

from the community on drug consumption activity tend to generate up to two times more drug consumption content in the future. We also released a manually annotated dataset of 4,000 samples and text classifier to detect user-generated content for drug consumption disclosures or not.

Online communities have also emerged as support groups. Traditional support groups are inaccessible to some people due to geographical or personal constraints. Online support communities can be immensely helpful for such people. Another time online communities came to the rescue was during the COVID-19 global pandemic. Since physical interactions were limited, online communities became a vital source of support for many people. For such communities to be effective and vibrant, they must have a base of knowledgeable, empathetic, and long-term members. In Chapter 5, based on our paper [77], we collect data from a COVID-19 support community *r/COVID19positive*. We classify data into different strands based on the type of support (emotional or informational), temporal phases (before, during, and after infections), and type of behavior (giving, seeking, and receiving support). We quantify different forms of interactions based on these strands and use survival analysis to uncover what leads to longevity in the community. We found out that users who give much support in the initial phases and seek help while suffering from COVID-19 tend to stay for longer. Contrary to common belief, our findings show that receiving emotional and informational support has little effect on users' tendency to stay in the community long term.

1.2.2 Individual-Organization Interactions

The most ubiquitous online interaction between organizations and individuals is online shopping, also known as e-commerce. The ability to shop online allows users to buy anything they like from the comfort of their homes, unlike offline shopping, where the items and deals in the surrounding geographical area limit users. This also allows sellers to reach a much wider audience, potentially higher revenues. Rich data available to e-commerce platforms allows for significant optimization in the entire supply chain. At a collective level, platforms can analyze the trends and inform sellers on what kind of items are more likely to sell, leading to a reduction in production waste. Complex routing and order clustering algorithms built on rich data can reduce the cost of shipping logistics. Finally, at an individual level, rich data enables the platform to provide the best product and deal that suits the user's requirements. However, the online nature of trade also exposes these platforms and sellers to potential abuse and fraud. E-commerce features like easy cancellations, returns, and refunds can be exploited by bad actors or uninformed customers, leading to revenue loss for the organization. Following this, it becomes essential for organizations to leverage the rich available data to build safety systems. In this part of the thesis, we study how e-commerce platform data can be used to build fraud detection models. Such models face a problem of cold start where the user is new and not enough data is available. We explore the potential use of cross-platform data and how mixing online public information about the user with internal data can improve model performance.

One such fraudulent problem e-commerce organizations face is Return to Origin (RTO), where the user cancels an order while it is in transit for delivery. In such a scenario, the platform faces logistics and opportunity costs. Traditionally, platforms analyze historical patterns at the user, seller, and product levels to predict the propensity of an order becoming RTO. However, such models have a high failure rate for new users. Further, social literature has shown clear links between socio-economic features and a user's potential to exploit a system, but often, such features are not available to the platforms. In Chapter 6, based on our paper [76], we use cross-platform data, where we mix the publicly available social data of the user with internal platform data. We propose a social re-identification method and verification scheme suitable for the unique case of e-commerce. We experiment with real-life data from the biggest e-commerce platform in India. Our system demonstrates a performance improvement in RTO detection of 3.1% and 19.9% on precision and recall, respectively. Our system directly impacts the bottom line revenue and shows the applicability of social re-identification in e-commerce.

1.2.3 Individual Centric Interactions

The most common types of individual-centric platforms are direct messaging and dating. Less common but quickly growing platforms in this category are tracking applications that allow users to log aspects of their lives, for example, Apple Health, Fitbit for tracking exercise, StickK to track habits, and Mint to track spending. These platforms are interesting because they record an aspect of user behavior that was previously primarily unrecorded. The concept of paper journals and budget books has existed, but logging data using manual methods was cumbersome and rarely used. Features like automatic and one-click logging have enabled much wider tracking application adoption. Though such data is often private to the users, some users decide to share this data publicly as part of a success celebration or public accountability. This allows computation social scientists to explore aspects of user behavior that were previously unexplored. In this part of the thesis, we demonstrate how non-conventional online data sources, like tracking applications, can be rich data sources, and help uncover actionable insights, and verify anecdotal evidence.

The pursuit of habit building is challenging, and most people struggle with it. Research on successful habit formation is mainly based on small human trials focusing on the same habit for all the participants, as conducting long-term heterogonous habit studies can be logistically expensive. In Chapter 7, based on our paper [79], we collect data from a popular habit-tracking platform, *StickK*, which allows users to track progress on habit-building attempts called commitments. Rooted in theories like the *Fresh Start Effect*, *Accountablity*, and *Loss Aversion*, we ask questions about how commitment properties like start date, external accountability, monetary stake, and pursuing multiple habits together affect the odds of success. We found that people tend to start habits on temporal landmarks, but that does not affect the probability of their success. Practices like accountability and stakes are not often used but are strong determents of success. Commitments of 6 to 8 weeks, weekly

reporting with an external referee, and a monetary amount at stake are most successful. Finally, around 40% of all commitments are attempted simultaneously with other goals. Simultaneous attempts to pursue commitments may fail early, but if pursued through the initial phase, they are statistically more successful than building one habit at a time.

1.3 Thesis Organization and Publications

The document is structured as follows. In Chapter 2, we provide the basic background information for commonly used ideas and techniques in the thesis. Part II includes Chapters 3,4 and 5, where we discuss the works related to the Interaction dynamics of a user in the community. Next is Part III, which focuses on the interaction of user and organization, discussing our work related to RTO detection. Following this, we present Part IV, looking at tracking platforms and their application in user behavioral understanding. Finally, Part V concludes the thesis by discussing the overall impact, limitations, and potential future extensions to our work.

1.3.1 Publications

List the publications that contribute to the thesis:

- Gurjar, O., Bansal, T., **Hitkul**, Lamba, H., and Kumaraguru, P. Effect of Popularity Shocks on User Behavior. *In Proceedings of the 16th AAAI International Conference on Web and Social Media (ICWSM' 22), June 6-9, 2022, Atlanta, Georgia, USA.*

- **Hitkul**, Shah, RR., and Kumaraguru, P. Effect of Feedback on Drug Consumption Disclosures on Social Media. *In Proceedings of the 17th AAAI International Conference on Web and Social Media (ICWSM' 23), June 5-8, 2023, Limassol, Cyprus.*

- **Hitkul**, Abinaya, Saha, S., Banerjee, S., Chelliah, M., and Kumaraguru, P. Social Re-Identification Assisted RTO Detection for E-Commerce. *In Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30-May 4, 2023, Austin, TX, USA*

- **Hitkul**, Pandey, T., Singhal, S., Kandhari, K., Tomar, K. and Kumaraguru, P. Together Apart: Decoding Support Dynamics in Online COVID-19 Communities. In Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2023 (ASONAM '23), November 6-9, 2023, Turkey

- **Hitkul**, Shah, RR., and Kumaraguru, P. Put Your Money Where Your Mouth Is: Dataset and Analysis of Real World Habit Building Attempts. *In Proceedings of the 18th AAAI International Conference on Web and Social Media (ICWSM' 24), June 3-6, 2024, Buffalo, New York, USA*

Additional publications while at IIIT-Delhi:

Peer-Reviewed publications

- Agarwal, A., Priyadarshi, PP., Gupta, S., **Hitkul**, Sinha, S., Kumaraguru, P., and Garimella, K. Television Discourse Decoded: Comprehensive Multimodal Analytics at Scale. *Under Review at a Top Tier Conference.*
- Goyal, S., Bhagat, S., Uppal, S., **Hitkul**, Yu, Y., Yin, Y., and Shah, R. R. Emotionally enhanced talking face generation. *In Proceedings of the 1st International Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice (pp. 81-90).*
- **Hitkul**, Gurjar, O., Sadaria, A., Gupta, K., Srikanth, S., Shah, R. R., and Kumaraguru, P. Are Bots Humans? Analysis of Bot Accounts in 2019 Indian Lok Sabha Elections. *In Proceedings of the IEEE Sixth International Conference on Multimedia Big Data 2020 (BigMM'20).*
- **Hitkul**, Shah, R. R., Kumaraguru, P., and Satoh, S. I. Maybe Look Closer? Detecting Trolling Prone Images on Instagram. *In Proceedings of the IEEE Fifth International Conference on Multimedia Big Data 2019 (BigMM'19), September 11-13, 2019, Singapore.*
- Bhavan, A., Chauhan, P., **Hitkul**, and Shah, R. R. Bagged support vector machines for emotion recognition from speech. *Knowledge-Based Systems 184 (2019)*
- **Hitkul**, Singhal, S., Shah, R. R., and Zimmermann, R. Aspect-based financial sentiment analysis using deep learning. *In Companion Proceedings of the ACM Web Conference 2018 (WWW '18 Companion), April 23-27, 2018, Lyon, France*

Technical Reports and Book Chapters

- **Hitkul**, Prabhu, A., Guhathakurta, D., Subramanian, M., Reddy, M., Sehgal, S., Karandikar, T., ... and Kumaraguru, P. Capitol (Pat) riots: A comparative study of Twitter and Parler. *arXiv preprint arXiv:2101.06914.*
- **Hitkul**, Aggarwal, K., Bamdev, P., Mahata, D., Shah, R. R., and Kumaraguru, P. Trawling for trolling: A dataset. *arXiv preprint arXiv:2008.00525.*
- **Hitkul**, Shahid, S., Singhal, S., Mahata, D., Kumaraguru, P., and Shah, R. R. Aspect-based sentiment analysis of financial headlines and microblogs. *In Deep Learning-Based Approaches for Sentiment Analysis (2020): 111-137.*
- Bhavan, A., Sharma, M., Piplani, M., Chauhan, P., **Hitkul**, and Shah, R. R. Deep learning approaches for speech emotion recognition. *In Deep Learning-Based Approaches for Sentiment Analysis (2020): 259-289.*

- Rajput, K., Kapoor, R., Mathur, P., **Hitkul**, Kumaraguru, P., and Shah, R. R. Transfer Learning for Detecting Hateful Sentiments in Code Switched Language. *In Deep Learning-Based Approaches for Sentiment Analysis (2020): 159-192.*
- Mahata, D., Friedrichs, J., Shah, R. R, and **Hitkul**. # phramacovigilance-Exploring Deep Learning Techniques for Identifying Mentions of Medication Intake from Twitter. *arXiv preprint arXiv:1805.06375.*

Chapter 2

Preliminaries and Background

The main aim of the thesis is to study various online user interactions and their effect on users' online/offline behavior. Research has shown that large volumes of social media data are adequate for studying user behaviors [164]. However, to ensure the validity and impact of such works, it is necessary to account for the past social science literature, ensuring online data provides a correct proxy to the phenomenon being studied and not confusing meer correlations as causations [103, 194]. We begin with an overview of the platforms, previous sociology literature, and computational techniques used throughout the thesis. We will provide further necessary details, if needed, in each chapter.

2.1 Online Social-Technical Platforms

User activities acquired from online social-technical platforms are the primary data used in the thesis. Hence, we start by defining our various platforms throughout the thesis.

TikTok[1]**:** TikTok is a short-form video hosting platform. Users can upload vertical videos, ranging from 3 seconds to 10 minutes.[2] Other users can Like, Comment, or Share these videos. TikTok also allows for directed follow/follower relationships.

Reddit[3]**:** Reddit is a vast online community-driven platform that functions as a social news aggregator, discussion forum, and content-sharing website. Reddit's core structure revolves around "subreddits," which are individual communities focused on specific topics. Users can subscribe to these communities to view, share, and discuss their interests. The up-voting/downvoting system influences post visibility. Each post on Reddit allows for threaded discussions via comments. Reddit features a karma system reflecting user contributions and awards for exceptional content. Reddit also allows each subreddit to have specific rules, and the community appoints moderators to ensure adherence.

Flipkart[4]**:** Flipkart is an Indian e-commerce company that allows users to sell or buy products online. It also provides features like cash on delivery, easy returns, and open-box deliveries to facilitate users' shopping experience. Users can also rate or review purchased products through star ratings, text, images, and videos.

[1] https://www.tiktok.com/en/

[2] Duration was initially limited to 60 seconds, hence the name short-form.

[3] https://www.reddit.com/

[4] https://www.flipkart.com/

LinkedIn[5]**:** LinkedIn is a professional networking platform for career development, business connections, and industry insights. Users create profiles showcasing their professional experience, skills, and education. The platform facilitates networking through connections, allowing users to expand their professional circles. LinkedIn provides job listings, company profiles, and industry-specific content. Users can share articles and updates and discuss through posts and comments.

StickK[6]**:** StickK is a goal-setting platform designed to help individuals achieve personal objectives by creating commitment contracts. Users set specific goals, from fitness targets to quitting habits, and stake a financial amount to stay committed. The platform uses behavioral economics principles, allowing users to appoint a referee or choose a charity recipient to keep them accountable. If goals are met, the user can retain the pledged money. StickK provides support through progress tracking and reminders, fostering motivation and accountability.

2.2 Sociology Theories

We ground our research questions in the sociology theories proposed and studied in the past and build upon them. This section provides a primer on various theories utilized in this thesis.

2.2.1 Behavioral Theories

This section elaborates on theories describing the reasoning, rationale, and incentives driving users' behaviors and actions.

Economies of Online Cooperation: The *Economies of Online Cooperation* [93] delves into the "gift economy" concept prevalent in many online environments, where individuals contribute without direct/immediate monetary compensation, fostering collaboration. Authors argue that such gift economies create public goods and foster a sense of community, reputation, and reciprocal social relationships among contributors. Furthermore, the theory discusses the role of social norms, peer recognition, and reputation systems in sustaining these gift economies. It explores how the community's desire for social recognition and status motivates individuals to contribute, resulting in the collective development of valuable resources that benefit a wider audience. Theory hypothesized that there are three significant reasons for users to keep on contributing to the social community - 1) *Anticipated Reciprocity*: The user is generally motivated to contribute or stay as an active participant in online communities in the expectation that the user will receive helpful information when they are in need. 2) *Sense of Efficacy*: The users might contribute information because they are rewarded with the sense that they contributed something to the community [14]. The efficacy can also result in the self-belief that they have a high impact on the community, hence validating their self-image as an efficacious person. 3) *Reputation*: Most users want recognition for their contributions or efforts. As quantified by the number of unique impressions of their content,

[5]https://www.linkedin.com/

[6]https://www.stickk.com/

popularity validates their content. This can be seen as an increase in reputation for the user based on the high number of people that follow or subscribe to them.

Impression Management: According to Goffman's *Impression Management* theory, individuals tend to present an idealized version of themselves rather than an authentic one to shape how others perceive them [59]. This theory suggests that people use various techniques to create a favorable image or achieve specific social goals. Goffman uses the analogy of a stage to represent social interactions, with the idea of "front stage" and "backstage" behaviors. The front stage refers to the public self that individuals present to others, carefully managing their behavior, appearance, and speech to convey a desired image. On the other hand, the backstage represents private settings where individuals can relax and drop their public facades. The Impression Management theory highlights several techniques used to control the impressions that people create, such as impression formation (creating initial perceptions), maintenance (sustaining the desired image), and repair (addressing inconsistencies or disruptions in the established impression). These techniques are applicable in various social contexts, including business, social media, and everyday interactions. Goffman's work provides insight into how individuals manipulate impressions to influence social perceptions and interactions. Hogan [70] extended Goffman's concept of impression direction to online social media websites and considered online social media platforms as a stage that allows users to control their impressions via status messages, pictures posted, and social media profiles.

Edgework: Lyng's *Edgework* theory explores the motivations and experiences of individuals actively seeking risky or adrenaline-inducing activities [118]. This theory delves into the psychological and sociological aspects of risk-taking behaviors, emphasizing the thrill and excitement individuals derive from engaging in potentially dangerous or unconventional activities. The framework defines Edgework activities as those with a "clearly observable threat to one's physical or mental well-being", such as rock-climbing, auto-racing, criminal behavior, etc. At the core of Edgework is the notion that individuals engage in such risky behaviors not solely for the risks themselves but for the unique experiences and sensations they offer. These experiences often occur in spaces or situations with a fine line between safety and danger, providing an adrenaline rush and a sense of "illusion of control". While initially developed to explain thrill-seeking activities, Edgework theory's concepts can be applied to understand the motivations and experiences underlying self-harm behaviors that involve risk-taking, e.g., drug consumption, engaging in unprotected sex, etc. Treating an illusory sense of control as a factor, Lyng observed that Edgework is more common among young people than older people and males than females. Other studies have found similar evidence related to the gender and age of the risk-takers [47, 105].

Behavioral Momentum: Focuses on understanding and predicting the persistence and resistance of behavior when faced with disruption or change [139]. This theory draws upon principles from behavioral psychology and describes behavior as having varying degrees of momentum analogous to physical momentum. The core concept involves differentiating

between high and low momentum behaviors. “High momentum” behaviors are frequently reinforced and less susceptible to change, e.g., smoking addiction and regular exercise routine. In contrast, “low momentum” behaviors, like newly adopted dietary changes and learning a new skill, have fewer reinforcements and are more easily disrupted or changed. The theory emphasizes that behaviors with a higher rate of reinforcement, even in the face of adverse conditions or changes in the environment, tend to persist and are more resistant to extinction. In contrast, behaviors with a lower rate of reinforcement are more prone to disruption or cessation. The theory has practical applications in various fields, particularly in understanding the persistence of behaviors in applied settings such as behavioral interventions, addiction treatment, and education. For instance, in behavioral therapy, interventions that aim to reduce problematic behaviors might benefit from understanding and manipulating the momentum of behaviors by altering reinforcement schedules or introducing alternative reinforcement strategies.

2.2.2 Feedback Theories

This section elaborates on theories that describe different kinds of feedback and users' responses to them.

Law of Effect: The *Law of Effect* [185] is a fundamental principle in behavioral psychology that suggests the consequences of an action determine the likelihood of that action being repeated in the future. This law asserts that behaviors followed by favorable or satisfying consequences are more likely to be repeated, while behaviors followed by unfavorable or unsatisfying consequences are less likely to recur. The author's experiments with animals, particularly with puzzle boxes and cats, formed the basis for the Law of Effect. They observed that animals learned to perform actions that led to desirable outcomes, such as escaping the box and obtaining food, while avoiding actions that resulted in unpleasant consequences, such as confinement without reward. The Law of Effect underlines the idea that positive (rewarding) or damaging (removing an undesirable stimulus) reinforcement strengthens the association between a behavior and its consequences. This association influences the likelihood of the behavior occurring again in similar circumstances. This theory has been influential in the development of operant conditioning principles.

Operant Conditioning: Based on the principles of the Law of Effect, B.F. Skinner proposed *Operant Conditioning*, a comprehensive framework elucidating how behaviors are acquired, shaped, and modified through consequences. Unlike classical conditioning, which focuses on involuntary responses to stimuli, operant conditioning centers on voluntary behaviors influenced by their outcomes. The theory revolves around four primary components: “reinforcement” (positive and negative) and “punishment” (positive and negative). Positive reinforcement involves adding a desirable stimulus after a behavior, strengthening the likelihood of its repetition. Negative reinforcement entails removing an aversive stimulus, increasing the probability of a behavior's recurrence. Conversely, positive punishment introduces an unfavorable consequence after a behavior, aiming to decrease frequency. In

contrast, negative punishment involves the removal of a desirable stimulus, also intended to decrease the likelihood of a behavior occurring again. Skinner identified diverse schedules of reinforcement, fixed or variable, ratio or interval, that impact how often and predictably reinforcement is provided. Moreover, shaping involves reinforcing successive approximations of a desired behavior to guide subjects toward the target behavior gradually. Operant conditioning underscores the significance of trial-and-error learning, where organisms learn to associate their actions with specific consequences. These principles find broad applications across diverse domains, such as education, therapy, and organizational management. By manipulating reinforcement and punishment contingencies, individuals and institutions can effectively modify behaviors, fostering desired actions while discouraging unwanted ones.

Loss Aversion: *Loss aversion* is a fundamental aspect of behavioral economics [83]. It refers to the psychological tendency where individuals prefer to avoid losses more than the inclination to acquire equivalent gains. Authors found that the pain of losing is psychologically twice as powerful as the pleasure of gaining. The theory revealed that losses loom more prominent in people's minds than potential gains of equal value, and this asymmetry significantly influences decision-making and risk assessment. This bias is embedded in *Prospect Theory* [83], which proposes that people tend to weigh potential losses more heavily than gains when evaluating options. This tendency results in risk-averse behavior, where individuals might opt for safer choices to avoid potential losses, even if the potential gains are equal or more significant. Loss aversion impacts various aspects of decision-making, including financial choices, investment strategies, and consumer behavior. It explains why individuals might hold onto depreciating investments to avoid realizing losses or might be more willing to take risks to avoid potential losses than to pursue equivalent gains.

2.2.3 Bias Theories

This section elaborates on theories describing different cognitive biases users can have.

Primacy Effect: The *Primacy Effect* is a phenomenon that affects how people remember and prioritize information presented to them [7]. According to this cognitive bias, individuals are more likely to recall and give more weight to the initial items or experiences they encounter than those presented later. Researchers have found that the first information presented often shapes our impressions, attitudes, and memory retention more than the subsequent ones. The reason behind this effect is how the human brain processes and stores information where the initial items are more deeply encoded and have a more substantial impact on memory due to their position at the beginning of a sequence. Given its implications across various areas, such as advertising, persuasion, and decision-making, it is crucial to understand the primacy effect. For example, advertisers often place their key messages or brand information at the beginning of advertisements to take advantage of this cognitive bias and ensure that important details are more likely to be remembered by consumers. Understanding the primacy effect is crucial since it affects how individuals perceive, retain, and utilize information.

Anchoring Bias: *Anchoring bias* is a cognitive bias that describes the tendency for individuals to rely too heavily on the initial piece of information (the "anchor") when making subsequent judgments or decisions, even if the anchor is irrelevant or arbitrary [187]. This bias occurs because the initial information sets a reference point that influences subsequent evaluations, often leading individuals to adjust insufficiently from that anchor when making estimations or assessments. Research shows that once an anchor is established, individuals tend to assimilate their judgments or decisions around that reference point, even if it is entirely unrelated to the context. Anchoring bias often works in conjunction with the primacy effect. The relationship emerges in how both biases highlight the significance of the initial information presented. Anchoring bias focuses on how the initial anchor affects subsequent judgments, guiding individuals' estimations or decisions, while the primacy effect emphasizes the memory advantage and increased importance given to information encountered first. Both biases underscore the human tendency to be disproportionately influenced by initial information, whether in decision-making or memory retention. They showcase how the sequence or order in which information is presented can significantly impact perceptions, judgments, and subsequent decision-making processes.

Fresh Start Effect: The *Fresh Start Effect* is a psychological phenomenon that describes the tendency for individuals to be more motivated to pursue new goals, make changes, or initiate positive behaviors following temporal landmarks or significant dates, such as the start of a new year, a birthday, the beginning of a week, or even a special anniversary [35]. These moments act as "temporal landmarks" that create a sense of new beginnings or fresh starts, prompting individuals to perceive these times as opportunities to leave behind past shortcomings and embrace renewed commitment to their goals. Research suggests that temporal landmarks serve as mental breakpoints, separating past failures or shortcomings from future endeavors. The psychological separation created by these markers encourages individuals to distance themselves from past setbacks, fostering motivation, optimism, and commitment toward pursuing new behaviors or goals. The Fresh Start Effect is significant because it highlights the psychological impact of temporal landmarks in promoting behavior change and goal pursuit.

2.3 Computational Methods

We leverage various computation methods in our research, and specifics of experiments are provided in respective chapters. This section discusses the general overview of some techniques used multiple times in our work.

2.3.1 Regression Discontinuity Design

Regression Discontinuity Design (RDD) [184] is a quasi-experimental research design used to estimate causal effects by taking advantage of a naturally occurring cutoff point or threshold. RDD is employed when individuals or units on one side of a cutoff point

experience a treatment or intervention. At the same time, those do not, solely because they fall above or below a specific threshold. The critical principle of RDD is that individuals or units close to the cutoff point are very similar in characteristics, except for their proximity to the threshold. The treatment assignment is random at the threshold, allowing researchers to attribute any differences in outcomes between the two groups to the treatment. The design involves comparing outcomes of units just below and above the threshold. Suppose there is a significant difference in outcomes between these two groups. In that case, it can be attributed to the treatment, assuming that other confounding factors are well-controlled or balanced on both sides of the cutoff.

The mathematical formulation of RDD involves estimating a regression model that captures the relationship between the outcome variable (Y) and the assignment variable (X), while accounting for the discontinuity at the cutoff point. The basic model for RDD can be represented as:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \varepsilon_i \tag{2.1}$$

Where:

- Y_i represents the outcome variable for individual or unit.
- X_i is the assignment variable that determines treatment eligibility, usually continuous.
- D_i is a binary variable indicating whether an individual is just above or below the cutoff point (0 for below the cutoff, 1 for above the cutoff).
- β_0 is the intercept term.
- β_1 is the coefficient representing the slope of the relationship between the outcome and the assignment variable.
- β_2 represents the discontinuity effect or treatment effect at the cutoff.
- ε_i is the error term.

The coefficient β_2 captures the local treatment effect or discontinuity at the cutoff point. If there is a statistically significant difference in outcomes between individuals just below and above the cutoff after controlling for the assignment variable (X), it suggests that the treatment has a causal effect on the outcome. To estimate the effect more precisely, researchers often use local polynomial regression techniques, such as local linear regression or local quadratic regression, to model the relationship around the cutoff. These techniques involve fitting separate regression lines or curves on both sides of the threshold to better capture the discontinuity. RDD is commonly applied in various fields, including economics, education, public policy, and health sciences, to evaluate the impact of interventions or policies where eligibility depends on a cutoff score, age limit, income level, or other criteria.

2.3.2 Propensity Score Matching

Propensity score matching [158] is a statistical technique used in observational studies to reduce selection bias and mimic some aspects of randomization by creating comparable groups based on their likelihood (propensity) to receive a treatment or intervention. Propensity score matching usually involves four steps:

1. **Propensity Score Estimation:** A logistic regression or other modeling technique is initially used to estimate the propensity scores. The propensity score represents the likelihood or probability that an individual or unit receives the treatment based on observed covariates (variables that influence treatment assignment but do not cause the outcome directly). The model predicts the probability of receiving the treatment based on these covariates.

2. **Matching:** After obtaining the propensity scores, individuals or units are matched between the treatment (those who received the intervention) and control (those who did not) groups based on their propensity scores. Matching methods could include one-to-one matching, nearest-neighbor matching, or other algorithms to pair treated and untreated units with similar propensity scores.

3. **Balance Assessment:** The effectiveness of the matching process is assessed by checking for covariate balance between the treatment and control groups after matching. Covariates should be well-balanced between the groups, indicating that individuals with similar characteristics are now present in both groups.

4. **Outcome Analysis:** Once the matched groups are formed and balanced, the outcome of interest is analyzed and compared between the treatment and control groups. The comparison allows to estimate the treatment effect in a manner that reduces the influence of observed confounders, making it more similar to a randomized controlled trial.

Propensity score matching is valuable in observational studies where randomization (as in randomized controlled trials) is not feasible or ethical. By creating matched groups based on their likelihood to receive treatment, researchers aim to reduce the impact of selection bias and confounding variables, thereby providing more accurate estimates of the treatment effect. However, it is crucial to use appropriate statistical methods and carefully interpret the results to ensure the validity and reliability of the findings.

2.3.3 Survival Analysis

Survival analysis [131] is a statistical method used to analyze the time until the occurrence of an event, such as death, failure, recovery, or any other endpoint of interest. It is widely used in medical research, social sciences, engineering, and other fields to study duration or times-to-event. Critical components of survival analysis:

Survival Function: The survival function, denoted as $S(t)$, represents the probability that an event has not occurred by time t. It indicates the proportion of individuals or units that survive beyond a specified time. Mathematically, the survival function is expressed as:

$$S(t) = P(T > t) \tag{2.2}$$

Where, T represents the time-to-event variable.

Hazard Function: The hazard function, denoted as $\lambda(t)$, describes the instantaneous rate at which an event occurs at a specific time t, given that the individual or subject has survived up to that time. Mathematically, the hazard function is expressed as the ratio of the probability of experiencing the event in an infinitesimally small time interval Δt to the survival probability at time t:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{2.3}$$

Kaplan-Meier (KM) Curves: The Kaplan-Meier (KM) [85] estimator is a non-parametric method used to estimate the survival function for censored data. Censoring occurs when the event of interest has not occurred for some individuals by the end of the study or when they are lost to follow-up. KM curves display the estimated survival probability over time and are commonly used to compare survival between different groups or treatments. KM curves are step functions that estimate survival probabilities at specific time points based on observed event times and censored data. At each event time, the curve decreases based on the occurrence of an event or remains constant if censored. The mathematical formulation of the Kaplan-Meier estimator involves calculating stepwise survival probabilities at distinct event times. Algorithm 1 provides a step by step functioning of KM estimator.

Algorithm 1 Kaplan-Meier Estimator

Initialize variables:
    $t_0 \leftarrow 0$ ▷ Initial time
    $S(0) \leftarrow 1$ ▷ Survival probability at time t_0
Sort event times $t_1, t_2, \ldots, t_k$
for $i \leftarrow 1$ to k **do** ▷ Loop through event times
    Obtain d_i (number of events) and n_i (number at risk)
    Calculate survival probability:
        $S(t_i) \leftarrow S(t_{i-1}) \times (1 - \frac{d_i}{n_i})$
end for
Output: Kaplan-Meier survival probabilities $S(t_1), S(t_2), \ldots, S(t_k)$

Cox Proportional Hazards Regression: Cox regression [30] is a widely used method in survival analysis to assess the association between covariates (independent variables) and the hazard rate (risk of an event) while adjusting for other variables. The Cox proportional hazards model assumes that a particular variable's hazard ratio (HR) is constant over time.

The model estimates the hazard ratio, indicating how much a change in a predictor variable affects the hazard or risk of experiencing the event. A hazard ratio greater than 1 signifies an increased risk of the event, while a ratio less than 1 indicates a decreased risk, holding other variables constant. The hazard function for an individual i at time t in Cox regression is given by:

$$\lambda_i(t) = \lambda_0(t) \times e^{(\beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip})} \tag{2.4}$$

Where:

- $\lambda_i(t)$ represents the hazard for individual i at time t.
- $\lambda_0(t)$ is the baseline hazard function, representing the hazard when all covariates $(X_{i1}, X_{i2}, ..., X_{ip})$ are zero.
- $e^{(\beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip})}$ denotes the hazard ratio associated with the covariates.
- $\beta_1, \beta_2, ..., \beta_p$ are the coefficients representing the effects of the covariates on the hazard rate.
- $X_{i1}, X_{i2}, ..., X_{ip}$ are the values of covariates for individual i.

In Cox regression, the model does not make specific assumptions about the shape of the baseline hazard function $\lambda_0(t)$; instead, it assumes that the hazard ratios for the covariates are constant over time, implying that the proportional hazards assumption holds. The coefficients $\beta_1, \beta_2, ..., \beta_p$ are estimated using partial likelihood estimation methods, aiming to maximize the likelihood of observing events conditional on the observed data.

Survival analysis allows researchers to examine the time-to-event data, handle censored observations, and understand factors influencing the timing of events. Kaplan-Meier curves provide visual representations of survival probabilities, while Cox regression helps identify the relationship between covariates and survival outcomes while accounting for potential confounders.

2.3.4 Dense Representations for Text

The predominant modality across the majority of our datasets is text, presenting one of the foremost challenges encountered in effectively classifying and clustering textual data. While classical Natural Language Processing (NLP) techniques rely on bag-of-words representations and rule-based algorithms, recently, a popular methodology in NLP is learning dense text representations. Mikolov et al. [130] proposed a neural algorithm to learn text representations based on word co-occurrence, outperforming classical token-based representation in various tasks. Vaswani et al. [190] proposed an improved model architecture called Transformers based on self-attention [169] to learn contextually aware dense text representations. Transformer-based large pre-trained models [44, 113] have provided an

efficient base to perform multiple tasks like classification, topic modeling, retrieval, and clustering on various data sources.

Part II

Individual-Community Interactions

Chapter 3

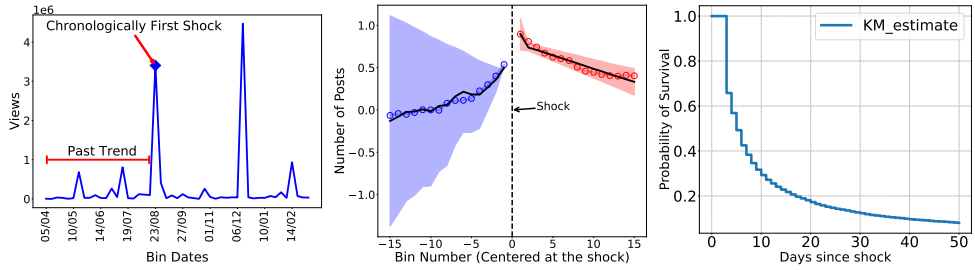
Effect of Popularity Shocks on User Behavior

Users often post on content-sharing platforms in the hope of attracting high engagement from viewers. Some posts receive unusual attention and go “viral", eliciting a significant response (likes, views, shares) to the creator in the form of *popularity shocks*. Past theories have suggested a sense of reputation as one of the key drivers of online activity and the tendency of users to repeat fruitful behaviors. Based on these, we theorize popularity shocks to be linked with changes in the behavior of users. In this chapter, we propose a framework to study the changes in user activity in terms of frequency of posting and content posted around popularity shocks. Further, given the sudden nature of their occurrence, we look into the survival durations of effects associated with these shocks. We observe that popularity shocks lead to an increase in the posting frequency of users, and users alter their content to match with the one which resulted in the shock. Also, it is found that shocks are tough to maintain, with effects fading within a few days for most users. High response from viewers and diversification of content posted is found to be linked with longer survival durations of the shock effects. We believe our work fills the gap related to observing users' online behavior exposed to sudden popularity and has widespread implications for platforms, users, and brands involved in marketing on such platforms.

This chapter is partly a reproduction of paper published at the AAAI International Conference on Web and Social Media (ICWSM) 2022 [64].

3.1 Introduction

Social media platforms have emerged or transformed themselves to focus more on content creation and sharing, e.g., TikTok, Instagram, Twitch, YouTube, etc. These social media platforms, focusing on content/multimedia sharing, have enabled users to express themselves in unique ways (text, photos, videos, etc.) to their followers (subscribers). To continue content creation and also engagement, most of the platforms have also launched creators' funds and also allow content creators (users) to get incentives/money to create such content [94, 191]. With social media content creation becoming an alternate source for revenue generation, users



(a) Example of Shock (b) RDiT (c) Survival Analysis

Figure 3.1: Our work discusses (a) detecting points of sudden increase in response known as *popularity shocks* on users' timelines; (b) Quantifying behaviour change due to popularity shock in terms of change in posting frequency using RDiT(Regression Discontinuity in Time); (c) Short-lived survival duration of effect of shocks and factors affecting it.

are also focusing on creating exciting content and eliciting attention and thus engagement from other users.

Users who become popular on these social media platforms are often termed as "influencers" or "micro-celebrities" [191, 54, 80]. Influencers, due to their popularity, have a broad reach and have been studied in the past on swaying/forming attitudes about consumer purchase intention [112], brand's image [68] and perceived uniqueness [42], along with even dietary behavior in children [177]. Influencers are also often contacted by different brands for endorsing their products [191, 188]. Many studies have been done in understanding why certain content and users who post them become popular [50, 51, 56, 123]. However, there has been little or no study on addressing how viral users respond to their newly achieved popularity.

On content sharing platforms, receiving sudden popularity due to specific content (or a series of content getting viral) can be termed as *popularity shocks*. Popularity shocks can be characterized as a sudden increase in feedback (i.e., views, likes, etc.). Previously, popularity shocks have been studied towards Wikipedia pages because of an associated event [204, 88], and Github repositories due to being highlighted by the platform [119]. However, the effect of popularity shocks on users' content creating behavior has not been studied in detail. Similarly, much work has been done in predicting posts that will go viral or will become popular using initial dynamics [49, 207, 199]. However, little work has been done in analyzing the after-effects of a post becoming viral or a user becoming popular. *Do users become more active on the platform after getting popular? Do users alter their content or stick to the content that made them popular? How long does the popularity shock last? Is popularity short-lived, or can it be long-term based on how user conducts themselves?* Answering these questions could have wide-reaching implications for all three - the users, potential brands seeking influencers to partner with, and also the social media platform itself. Studying users' response to popularity shock can be insightful for (a) users, who want to continue engagement, (b) brands, for identifying new influencers which align

with their values, and (c) social media platforms, for guiding new popular users on specific interventions that can be related to education, design changes or guidelines.

We ground our work in sociological theories related to social reinforcement and a sense of reputation. A reputable theory in the field of behavioral psychology has been *Operant Conditioning* [175]. Under this theory, an activity that earns rewards prompts an individual to repeat that activity, and similarly, an activity that earns punishment makes the individual more inclined to repeat that activity. In our context, if we treat receiving popularity, which is quantified with high engagement from the community on users' content, as positive feedback (or reward), the user ideally will keep repeating the same behavior. Alternatively, if the user received a popularity shock in a negative context, i.e., they were a recipient of a firestorm [99], they might stop posting similar content. We also draw on the theoretical work carried out in a more specific context of online communities [93, 155]. In one of the earliest analyses of an online community, Rheingold hypothesized that desire for prestige is one of the key motivators for individuals' contribution to the community. Kollack re-emphasized this [93], highlighting that increased reputation is one of the three reasons for individuals to contribute content on online platforms. Contextualizing this in our work, popularity shock can be viewed as a signal of increasing reputation and might prompt users to continue contributing to the platform. Though these theories were proposed some time ago, rigorous empirical evaluation/validation of these theories in the context of popularity on online social media platforms have not yet been conducted.

In this chapter, we study how do users' behavior changes after a popularity shock in terms of (a) frequency of posting, (b) the content, itself and (c) how long do they continue with their altered behavior. We first characterize what should be considered as a popularity shock and develop a method to identify popularity shocks from a user timeline. Using popularity shock as an intervention, we use causal inference techniques to examine the change in behavior from pre-and-post popularity shock. Next, we study the change in the content posted by users under the effect of popularity shock. We leverage document embeddings [104] to model the posted content mathematically. Finally, we investigate the expected duration for shock's effect and its dependence on other factors using survival analysis techniques.

Data and Code: We released the anonymized version of our data available at:
`https://precog.iiit.ac.in/research/effect-popularity/`

3.2 Related Work

Since our work is related to users' response towards increased attention, our related work flows from three main directions - (a) Effect of social feedback, (b) Attention Shocks and (c) Popularity/ Virality Prediction.

Effect of social feedback: Positive reinforcement or feedback has been a popular area of study among social scientists [10, 127, 148, 163, 165]. Rushton et al. [163] demonstrated through experiments on around 60 children through a bowling game that positive reinforce-

ment led to improvement in altruistic behavior in children, while punishment led to the opposite. This framework has been studied extensively in various settings, such as effect of positive feedback on promoting safe behaviours in housekeeping [165] and effect on compliance following transgression [127] as well as simulating motivations and future play of a brain training game [20]. In the domain of online world, however opposite effect has been observed in the case of low quality comments [25], where it was observed that negative feedback prompted users to continue with writing low quality comments on news articles. Further, [24] how the community perception of helpfulness of online reviews, influences consumer purchase decisions, and how this helpfulness vote is itself determined by evaluations of the same product by the community [39, 173]. Similar study has also been conducted for the effect of social feedback on weight loss community [33].

Though there has been a lot of studies discussing social feedback, however very few have tried to characterize how do users or actors in turn respond to extremely high and sudden feedback in data-oriented fashion on a large-scale data.

Attention Shocks: Attention shocks are characterized as sudden attention being drawn towards a specific entity (any author/artefact on social media platform). Examples include, death of a celebrity leading to increased attention towards the celebrity's wikipedia page [204]. Danaja et al. [119] use the lens of organisation change to study the dynamics of change in behaviour of contributors of a GitHub repository experiencing increased attention as a result of being listed on the trending page. On Wikipedia, [204] observe increased participation of new comers and study collaborator dynamics on pages in times of shock detected through Google Trends, while [203] look into the changes in collaborative behaviour of editors due to shock resulting from imposition of censorship in mainland China. Other works like [87] study similar changes in case of breaking news articles on Wikipedia. Lamba et al. [99] analyse shocks in form of sudden bursts of negative attention towards controversial events called 'firestorms', and use Twitter data to characterize the size and longevity of these firestorms.

Other works study the effect on online network structures under shocks. Keegan et al. [88] suggest the formation of complex but temporary collaboration networks of users during increased editing activity on Wikipedia page of a diseased person and study their dynamics. Further, [86] introduce a method of capturing collaboration structure of co-authors of a Wikipedia articles and highlight the difference between such networks for breaking news articles, as compared to traditional ones based on pre-existing knowledge.

Though attention shocks have been studied on online social media platform, to the best of our knowledge, our work is the first attempt to study the behaviour of users whose posts goes viral (i.e. the user who gets the shock). A minor characteristic that differs us from other studies is that we are looking at shock as a sudden virality of the post, and the virality of the post is mostly algorithm-driven (i.e. probably a mixture of recommendation algorithm and "rich-gets-richer" theory). In comparison, other studies looked at shock which was more exogenous i.e. appearance on GitHub trending page or death of a celebrity. Lastly, there are inherent differences in nature of platforms being studied. While Github and Wikipedia are

collaborative platforms where users are often driven by non-monetary motivations such as reputation and collective identity Danaja et al. [119], users on such content sharing platforms are driven by monetary causes and for self-satisfaction. Thus there is clear distinction in intent of use, due to which we can expect difference in user behaviour as well.

Though it is not highly aligned with our work, however there has been significant amount of work done for predicting if a post is going to get popular or not, and hence we mention about some of the efforts done to solve that problem.

Content Virality Most work in this domain is focused on predicting and characterising virality of online content. [49] understands popularity trends for online user generated content (UGC) in the form of online videos, and proposes a prediction model based on extremely random ensemble tree to predict the popularity trends for Youtube videos. The Seismic model proposed by [207] predicts the final number of reshares a post will recieve based on the past history. The problem is modelled as predicting the final size of an information cascade and performance is validated on a month of Twitter data. Other models like [196], [128] have tackled the problem of virality prediction on Twitter and Flickr respectively.

Other works are inclined more towards characterizing virality and viral content. Lilian et al. [199] studies the virality and diffusion of memes on online networks. Masoud et al. [123] seeks to identify features in posts which are related to its popularity using a multi-modal approach. Flavio et al. [51] aims to characterize and understand popularity growth of videos, and what kinds of mechanisms contribute towards popularity. The work also mentions presence of sudden bursts of popularity on top listed videos.

Table 3.1: Number of unique users for each category (arranged in alphabetical order of Category).

Category	**Hashtags**	**Unique Users**
Animals	cats, dogs, pets	1666
Beauty/Makeup	beauty, makeup, naturalbeauty, skincare	3052
Craft/DIY	5_min_craft, craftchallenge, diycraft, easycraft	1128
Dance	dance, dancechallenge, dancekpop	2144
Education	careergoals, education, learning, mindpower	2429
Entertainment	entertainment	449
Fitness	fitness, fitnessgoals, gym, weightloss, workout	3911
Food	food, foodislove, foodrecipe, healthyfood, myrecipe	2815
Funny	comedy, funny, meme	2632
Health	wellness	558
Motivational	advice, inspirational, lifehacks	2146
Music	hiphop, music	1323
Pranks	prank	667
Sports	cricket, football, sports, tennis	2341

3.3 Theory and Research Questions

Kollock [93] hypothesized that there are three significant reasons for users to keep on contributing to the social community - (a) anticipated reciprocity; user is generally motivated to contribute or stay as an active participant in online communities in the expectation that the user will receive helpful information when they are in need, (b) sense of efficacy; the users might contribute information because they are rewarded with the sense that they contributed something to the community [14]. The efficacy can also result in the self-belief that they have a high impact on the community, hence providing the validation of their self-image as an efficacious person, and (c) Reputation; most users want recognition for their contributions or their efforts. As quantified by the number of unique impressions of their content, popularity validates their content. This can be seen as an increase in reputation for the user based on the high number of people that follow or subscribe to them. On the lines of Kollock, we hypothesize that receiving a popularity shock (i.e., increase in reputation) will prompt users to increase their activity on online social media platforms. [1] Therefore, we ask the following question:

RQ1. [Engagement Response to Popularity] *Do users increase their posting behavior after receiving popularity shock?*

Another social theory framework that fits very well with our setting is that of *operant conditioning* [175]. Skinner theorized that the reward for action leads the agent to keep on performing the same action in anticipation of reward, and a punishment hinders the user's propensity to take that action. Again, operationalizing reward as the popularity shock, we can hypothesize that users who received popularity shock will continue with the same behavior that earned them the reward even in our setting. This brings us to the following research question:

RQ2. [Content Response to Popularity] *Do users alter their content post receiving popularity shock?*

In network science, the transition of network states and dynamics due to an external event has been a topic of interest [204, 119, 88]. Momin at al. [120] argue that some of the network transitions, and along with it changes in user behavior in these networks, are more permanent. Moreover, some studies argue that networks bounce back after the event, and normal communication ensues [99]. In our setting, we were interested in understanding how long the popularity shock lasts.

RQ3. [Longevity of Effect] *How long do the effects of popularity shock last?*

For users who receive the popularity shock, it is imperative to understand what users can do or how they should maintain their activity that can prolong the shock's effect. Therefore we ask the following question:

[1] In this work, we discovered that the popularity shocks were positive, analysis can be done if this popularity instead was negative too.

RQ4. [Sustained Shock Effect] *What type of activity characterizes long-term sustainability of effects of popularity shock?*

3.4 Data Collection

Background: We collect data from popular multimedia sharing social media platform.[2] On the platform, users can post multimedia content (images/videos) along with an associated caption. Depending upon the privacy setting of the post and the user's profile, other users can view their content and engage with the content using platform-provided mechanisms such as liking the content, commenting on the post, or resharing the post. By liking, a user can express their positive response or acknowledgment, sharing works to amplify the reach of content, and viewers can also express their opinions in the form of comments. Like all other social networking platforms, the social platform understudy also provides functionality that allows users to 'follow' other users on the platform. Besides this, the platform can also grant a special 'verified' status to specific users based on their strong influence on the platform or in the real world. Though we study a specific platform, we believe that a similar methodology can be applied to any social media platform with similar mechanisms in place. A cross-platform study on measuring this behavior and ensuring generalizability is one of the promising future directions of this work.

Data Collection: We identify 14 generic categories related to commonly posted content on the platform. From the list of these 14 categories, we curated a list of 43 popular hashtags. The hashtag selection was made keeping the goal of generalization in mind, and hence no hashtags related to specific entities (e.g., #ronaldocr7) were considered. The selected categories and hashtags are described in Table 3.1. Approximately $4,000$ posts per hashtag were collected, coming from $21,224$ unique users. Next, we collect posts liked by these users and add the authors of the posts to our dataset to minimize any sampling bias due to the collection strategy(which might be due to bias in the platform's search functionality) [3]. Finally, we had a total of $33,490$ users. We collected the entire timeline of these users and filtered out users who had less than 200 posts in their entire lifetime to ensure we had substantial data for our analysis.

Following the filtering, our final dataset contains a total of $30,969$ users. We describe the data statistics in Table 3.2 along with distribution of number of posts across users in Figure 3.2.

For each post, we collected the following details of the post - (a) post id, unique identifier for the post, (b) timestamp of when the post was published, (c) caption of the post, (d) number of views the post received, (e) number of likes the post received, (f) number of times the post was reshared, (g) number of comments the post received and (h) user information - all key statistics such as name, bio, etc. of the user who created the post.

[2] name of the platform suppressed to retain anonymity and non-public API access.

[3] Data collection was done when the first and second authors were students at their respective institutes

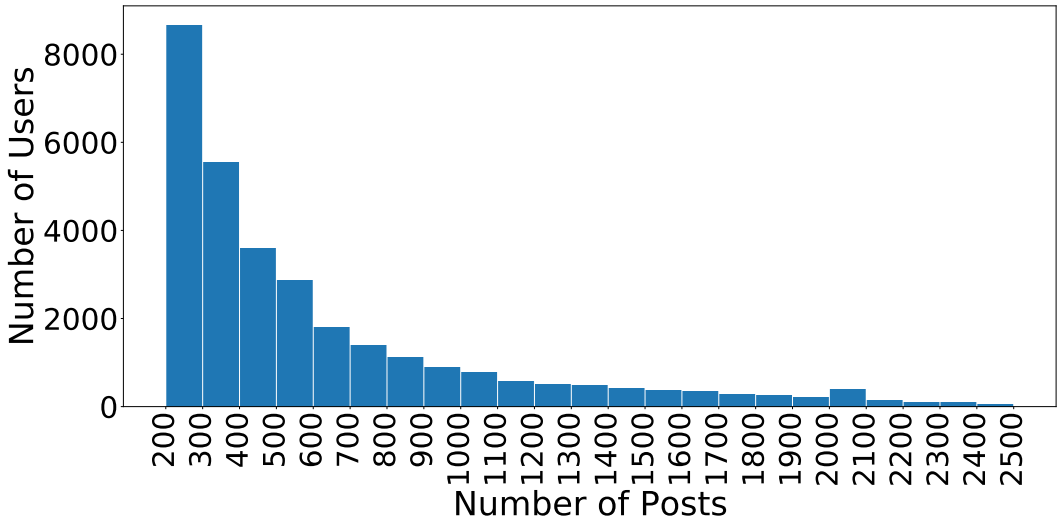


Figure 3.2: Distribution of Users' Total Posts (Follows Power Law).

Table 3.2: Dataset Details.

Number of Users	30,969
Number of Posts	18,911,417
Timestamp of First Post	7^{th} Jan 2015
Timestamp of Last Post	31^{st} Dec 2021

3.5 Detecting Popularity Shocks

To answer any of the research questions mentioned above, we first need an algorithm that can identify popularity shocks from a user's timeline. Before going into the details of the algorithm, we describe the assumptions we made to define popularity shock.

- We use the number of views as a proxy for popularity. Views give a more objective metric of the reach or engagement as it is implicit, unlike other metrics such as the number of likes, shares, or comments which require explicit action from the audience.
- A user might receive multiple popularity shocks throughout their career. However, we only study effects due to the chronologically first shock the users receive. We do not consider later shocks as the user would have already experienced some popularity until that point. In this paper, we want to characterize the effect of the first popularity shock when the sudden growth in popularity is unexpected for the user.

A desired shock detection algorithm should detect a sudden percentage increase in views of the user, we should also account for absolute thresholds to avoid false positives caused by the base effect. The first natural candidates for the task are time-series anomaly detection algorithms like Z-score [23] or Facebook's Prophet [181]. However, these algorithms consider

Table 3.3: Shock detection accuracy against the manually annotated ground truth. Proposed algorithm outperforms other baselines.

Algorithm	Accuracy
Z-Score	23.6%
Prophet	42.5%
Proposed	66.6%

time-series signals in isolation and do not account for global thresholds. To curb this, we also experiment with a custom algorithm as presented in Algorithm 2.

We preprocess the timeline by binning the posts, where each bin is a period of consecutive D (bin size) days. For each user, we iterate over the bins in chronological order (Line 6). We maintain a running average of views of all the bins encountered so far (Line 13). Once we have processed the bin (i.e., no more posts need to be counted for that bin), we compute the ratio of views of the bin to the running average of bins before it. Note that we ignore bins with no posts while computing the running average. This ratio needs to be higher than a ratio threshold θ for it to be considered a shock candidate. To account for the cases where the running average is very low, we also consider the difference between current views and the running average, which needs to be greater than the base threshold η. Therefore, the first bin satisfying these two conditions is classified as the popularity shock for the user. If no point satisfies these conditions, we consider the user is without a popularity shock. We show results across a variety of θ and η values.

Ideally, keeping consistent with our shock assumptions, we want to capture the first post at which user perceives they might have gotten popular. To evaluate our detection algorithm, we conduct a verification experiment. We solicited annotations from long-term social media users, who were asked to independently look at the view timeline of 100 users and mark what they deem as the first instance a user would have felt popularity shock. The annotators had a Fleiss' Kappa score [52] of 0.60, which indicates moderate agreement [102]. Each sample was annotated by 3 annotators, and a clear majority was received in 93 instances out of 100. We compared the efficacy of our proposed approach with baselines of z-score and Prophet algorithm using the ground truth set. Predictions were obtained across a range of hyper-parameters for all algorithms, best achieved results are shown in Table 3.3. Our proposed shock detection algorithm performs the best and is used to detect popularity shocks in our further experiments.

The percentage of users we can discover having a popularity shock with different values of θ and η is presented in Figure 3.3. Note that, we report results with hyper-parameters $D = 1$, $\theta = 50$, $\eta = 1.5$M, unless specified. However, we experimented with multiple values of θ and η, and results stay consistent over reasonable values of these thresholds.

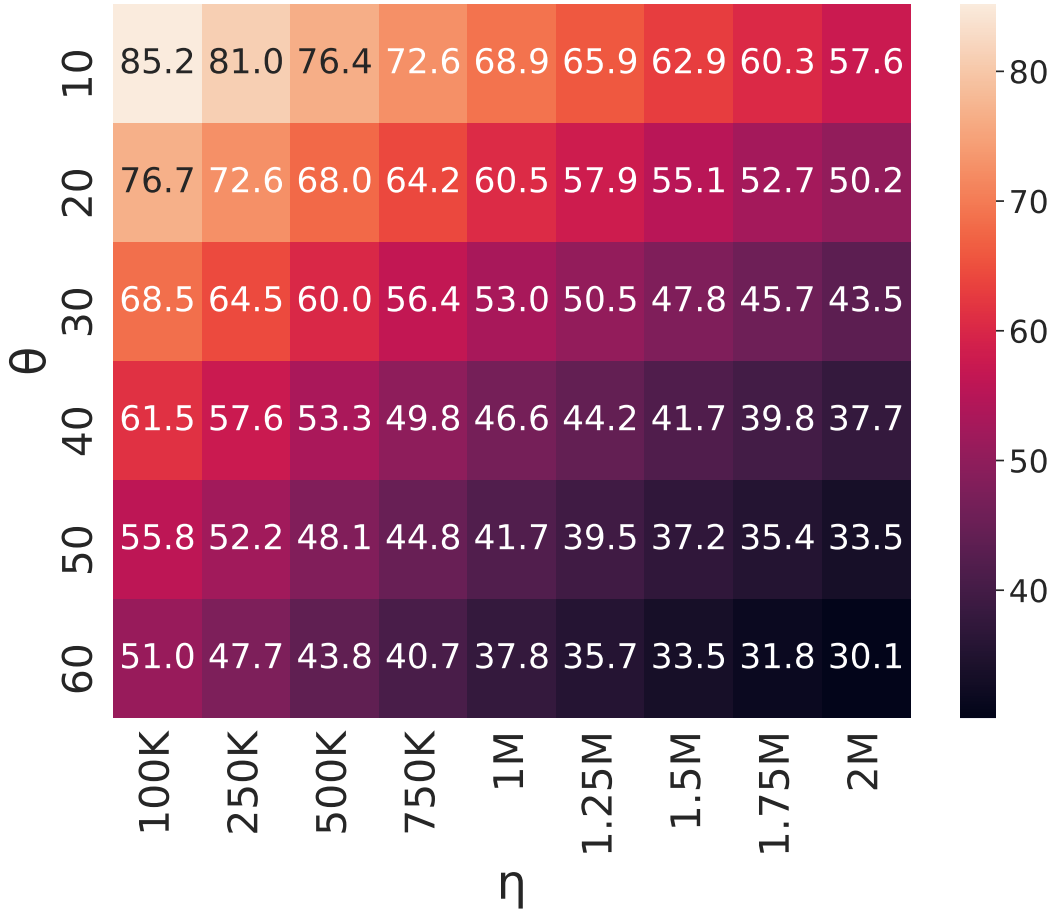


Figure 3.3: Heatmap representing percentage of users detected with a shock for different values of θ and η for $D = 1$. θ is the minimum ratio of views in the bin to the running average, while η is the minimum difference between the two, for detecting shocks.

3.6 Effect of Popularity

RQ1 seeks to quantify the change in posting frequency of a user due to the shock received. We do this using a causal inference technique called Regression Discontinuity in Time (RDiT) [67].

Regression Discontinuity Design (RDD): Introduced by [184], RDD is a quasi-experimental technique to measure the effects of a treatment or intervention. The population receives the treatment having the value of running variable X above a certain threshold known as the 'cut-off' point, and data is checked for any jumps or discontinuities in the outcome variable Y around the cut-off. Previously, RDD has been widely used in fields such as Economics [106] and Psychology [29]. Specifically, on social media studies, RDD has been analyzed previously to quantify the effect of obtaining a GitHub badge on users' posting frequency [145], on the effect of the introduction of Facebook "People you may know" feature [120], and also on the effect of averaging rounding stars on Yelp [110].

Algorithm 2 Shock Detection Algorithm

```
function DETECTSHOCK(posts, D, θ, η)
    bins ← bin_data(posts, bin_size=D)
    shock ← −1
    n ← length(bins)                                  ▷ Total number of bins
    run_avg = views(bins[1])
    for i in 2 to n do
        ratio ← views(bins[i])/run_avg
        diff ← views(bins[i]) − run_avg
        if ratio ≥ θ and diff ≥ η then
            shock ← i
            break                                     ▷ break at the earliest shock
        end if
        run_avg ← mean(views(bins[1:i]))
    end for
    return shock                                      ▷ if shock is -1, no shock found
end function
```

Acknowledging that time being the running variable might cause some of the assumptions of traditional RDD not to hold, we use a variation of the RDD framework called **Regression Discontinuity in Time (RDiT)** proposed in [67], in which time is the running variable and a fixed point in time is taken as the threshold. RDiT conceptually differs from the regular RDD on the following fronts:

- While RDiT aligns with the 'discontinuity at cut-off' interpretation of RDD, the 'local randomization' interpretation may not hold as the time assignment can not be taken as entirely random around the cut-off.
- Unlike RDD, sample size can not be grown arbitrarily with smaller bandwidths. Due to this, data points far from the cut-off need to be included, which can introduce biases due to changes in unobserved confounders over time.
- Including covariates becomes far more critical to control biases since the assignment of treatment and control groups is not entirely random around the cut-off.

Our methodology: To model our problem using RDiT, we define our running or forcing variable X as the bin index (signifying time) and outcome variable Y as the number of posts done by the user in the bin X. The shock bin is assigned index 0; subsequently, index $+i$ denotes the i^{th} bin after the shock, while the index $-i$ denotes i^{th} bin preceding the shock. The cut-off point c is $X = 0$, where the shock occurs. Then, treatment group is defined as $\{(X_i, Y_i)\ s.t.\ X_i > 0\}$ and control group as $\{(X_i, Y_i)\ s.t.\ X_i < 0\}$. We also control for the following covariates in our regression design:

- **Intensity of shock:** To account for variation in treatment, we control for the intensity of shock obtained in the preceding bin. The intensity is the value of the *ratio* variable for the bin as in Algorithm 2. We take the logarithm of this variable.
- **Age of User:** As receiving a popularity shock at different stages of users' online life might have different effects. We control for the number of days since the user's first post.

We then fit models separately on the two groups using regression. We only use W bins before and after the shock bin to fit the lines to avoid any effects of future shocks. On obtaining the equations of the two lines, their values at the cut-off point are predicted, which are used to calculate the discontinuity at the cut-off. Formally, let $Y_{t,0}$ and $Y_{c,0}$ be the values at the cut-off for the treatment and control lines respectively, then discontinuity at the shock d is given by $d = Y_{t,0} - Y_{c,0}$. From the equation, it can be seen that a positive d corresponds to an increase in the frequency of posting after the shock as compared to before and vice versa.

3.6.1 Effect on Posting Frequency

We tried to estimate the effect of popularity shock on the posting frequency of user post-shock using RDiT. We quantified the intervention to occur at the time-point where we detected the popularity shock. Further, we count the total number of views that the user received each day before and after the shock. Note that this corresponds to setting $D = 1$ in Algorithm 2.[4] In Figure 3.1b, we visualize the effect on posting frequency. The x-axis clearly shows the time before and after the shock. To aggregate the effect across all users, we compute the number of posts done by the user each day subtracted by the average number of posts done by the user in the past 15 days (this is done to maintain a consistent scale across users). Then, the average is taken across all users (including covariates) and curves as fit. The vertical dashed line shows the day on which popularity shock was observed. As mentioned above, we fit two linear regression models.[5] The first model is for the average number of posts done before the popularity shock, and the second one is for the average number of posts after the popularity shock. We see a significant difference between the intercept and the slope for both the regression models. The discontinuity at shock (d) estimates how users are changing their posting behavior pre- and post-shock. This is measured as the difference of the predicted number of posts done at the shock by the two regression models (intercept of the second model - intercept of the first model). We note that for all values of W, we observe positive discontinuity, implying a positive effect on the number of posts made by the user after receiving popularity shock. Both of these slopes are significantly different and hint

[4]Important to note that here, we also experimented with various values of D, θ and η and achieve similar results.

[5]We also experimented with higher order polynomial regression models, and results were consistent. Although we do observe overfitting in some cases.

towards a significant effect due to shock. Looking at the regression fits and the magnitude of discontinuity, we make the following observation:

Observation 3.1 (Increased Posting) *Users increase their posting behavior post shock.*

Observation 3.2 (Short-Term Gains) *Though users increase their posting behavior post shock, it also quickly decays off, as time progresses.*

Note that while the trend of the fit of the model pre-shock is positive and post-shock is negative - this could be due to the sensitivity towards our shock detection algorithm. Our shock detection algorithm works by binning the posts and classifying if a particular bin is a shock bin or not, and also, the algorithm takes into account total views rather than the average number of views. Therefore, users might be posting a high number of posts that were getting a sizable number of views (lesser than our threshold) until eventually tipping on the next bin and satisfying our threshold.

3.6.2 Significance of Result

We perform following checks as mentioned by [67]. 1) We control for observable confounders to remove biases and account for variation in treatment. 2) We perform a Placebo Test to ensure no discontinuity at points where there should not be any. Guido at al. [72] suggests checking for any discontinuities at the median values of the running variable for the sub-samples corresponding to either side of the cut-off and using standard errors to test for no discontinuity. We do this test only for the sub-sample below the cut-off, as the points above our cut-off may have discontinuities due to potential future shocks. Say the shock occurs at the s^{th} bin from the start, then we check for any discontinuity at $\frac{s}{2}^{th}$ bin. We observe significantly less discontinuity and overlap between 99% CI intervals, implying no observable discontinuity. 3) We check for robustness of our results towards window size and polynomial order. 4) We fit regression lines without controlling for covariates and observe similar results, indicating no time-varying treatment effects.
Note that, as suggested in [67], the McCrary density test [124] is not valid when time is the forcing variable. However, we argue that there is no manipulation in our case as users' can not preempt an imminent shock due to lack of knowledge of platform recommendation algorithm and the large magnitude of our shocks (50x more views with 1.5M difference).

3.6.3 Effect on Posted Content

In **RQ2**, we aim to determine if users alter the content they post after receiving a popularity shock. We characterize the content by using the posts' captions. The posts' captions can be noisy, so we take appropriate steps to develop a consistent representation from the captions. First, we preprocess the hashtags present in the caption by removing the '#' symbol from every hashtag and then use wordsegment[6] library to segment these hashtags into separate

[6]http://www.grantjenks.com/docs/wordsegment/

Table 3.4: Results showing similarity of content for before and after the shock to the shock($^{***}p < 0.001$).

Time Period	All Users	
	Sim(Pre, Shock)	Sim(Post, Shock)
7	0.625 ± 0.22	$0.714 \pm 0.17^{***}$
30	0.656 ± 0.21	$0.699 \pm 0.20^{***}$
	High Discontinuity Users	
7	0.645 ± 0.24	$0.730 \pm 0.20^{***}$
30	0.670 ± 0.21	$0.732 \pm 0.19^{***}$

words in order to extract their semantic meaning. Following this, we compute the similarity between the content posted in two time periods (set of bins). We represent the captions of all the posts done in that bin duration using a single feature vector and then measure their similarity. We use document embeddings to come up with the representation. We convert every post into a single vector using the document embedding of its caption. We leverage doc2vec [104] to generate embeddings.

Subsequently, we obtain a single vector representation for a time period by averaging the document embedding vectors corresponding to a set of posts from that temporal bin. We use cosine similarity to compare vectors formed using document embeddings. Cosine Similarity yields a score between 0 and 1, with 1 representing the same vectors. With the above experimental framework, we compare content posted in the shock bin with that of W bins just before and after the shock to capture the change around the shock. We also perform the analysis for users whose discontinuity in posting frequency lied in Top 25 percent. Based on the results in Table 3.4, we make the following observations:

Observation 3.3 (Post Shock Similarity) *Users, post-shock generate more similar content to the shock inducing posts.*

Observation 3.4 (High Discontinuity Similarity) *Users who increase their positing frequency more, also tend to stay more closer content-wise to the shock related posts*

We can observe from Table 3.4 that similarity of *doc2vec* embeddings between post-shock and shock is significantly higher than similarity between pre-shock and shock. We use significance test and obtain $p < 0.05$ to show that these two values over all users is significantly different.

3.7 Sustainability of Popularity

In both **RQ 3** and **RQ 4**, we try to answer the questions related to the sustainability of the popularity shock. For both of the questions, we leverage *survival analysis* [131]. Survival

Analysis is a popular multivariate event history modeling technique that focuses on estimating the average hazard rate of an event under consideration at a given time and also corresponding relative strength of the effect of different factors on this hazard rate, where hazard rate can be defined by $h(t) = \frac{P(T<t+\delta|T\geq t}{\delta t}$. Cox proportional hazard model [30] can be used to estimate this probability and the coefficients of the regression $h(t,X) = \theta(t)\exp(\beta^T X)$ using partial likelihood, without making any assumptions about the baseline hazard rate.

Our observation period for a particular user starts from the bin where the shock occurs. We define our event of interest as the point in time post the shock where there is no difference in activity level compared to pre-shock level. Specifically, we rely on the number of views to compare post-shock and pre-shock levels. We say that the increased response due to shock has faded away if we discover B consecutive bins with the number of views less than K. We set K as the 10% of the views obtained in the shock bin. We set the value of B as 3.

3.7.1 RQ3: Longevity of Shock Effect

In RQ3, we study how long the effect of a popularity shock lasts. We plot in Figure 3.1c, the survival curve for users to demonstrate the longevity of effect on shocks. From the curve, we observe that the effect of shock dies down rather quickly for most users. For 50% of the users the effect fades away in the first 5 days itself, while it ends for 90% of the users within 39 days of the shock. This implies that it is extremely difficult to maintain response levels observed during the shock for an extended period.

Observation 3.5 (Shock Longevity) *Popularity shocks are short-lived. The increased response received by users goes down to pre-shock level very quickly after the shock.*

3.7.2 RQ4: Sustaining Shock Effect

In RQ 4, we model the factors on which the longevity of shock effect depends as well as the effect and extent of the dependence. To do this, we build on existing survival model, and use Cox Proportional Hazards regression model [30] to quantify the effect of different factors on survival.

Factors affecting survival: We are specifically interested in understanding what a user can do to prolong the effect of popularity shock. We hypothesize the following factors:

1. *Posting frequency*: The frequency of posting represents how eager a user is to create and post more content after the popularity shock. It can be hypothesized that high posting frequency could indicate users trying to be more active on the platform and trying to engage highly with the new audience that the user has got access to. We operationalize this by the total number of posts a user does in a bin.

2. *Similarity in Consecutive Posts*: The change or variation in the content that users post could be indicative of how versatile the user is in adapting their content to the needs of their audience. A user might have got popular due to a specific type of content and keep

Table 3.5: Dependence of Shock Effect survival on other variables using Cox Regression ($^{***}p < 0.001$).

Covariate	**HR (St Err)**	**LR Chisq**
Avg. Likes	0.90 (0.01)***	292.43***
Shock Intensity	1.13 (0.03)***	80.59***
Posting Frequency	0.86 (0.01)***	1047.9***
Similarity between consecutive posts	6.54 (0.03)***	2734.31 ***
Similarity of posts with shock post	0.38 (0.04)***	37.57 ***

posting it in the hope of a similar response. However, this may lead to repetitiveness in content, and the audience might lose interest. Our analysis operationalizes this by the average cosine similarity between all posts in consecutive bins.

3. *Similarity with the shock content*: The similarity between the shock-related content and the current content is an indicator of how much the user has digressed from the content, which leads to their popularity. Viewers often start associating users with a specific type of content, and thus deviating too much from that may cause disengagement from their audience. We model this as the average cosine similarity of content posted in a bin with the shock content.

Though these are the factors that we are interested in, we also control for the following variables, which could affect the longevity of the effect.

- *Effect of feedback*: The amount of feedback received by a user on the posts user created after popularity shock is indicative of the engagement levels of the user's audience. We measure this by introducing three variables - (a) Number of likes, (b) Number of shares, and (c) Number of comments. Since these variables are highly correlated, we only use the average number of likes in the regression model.

- *Intensity of shock*: Another factor that needs to be controlled as to what was the magnitude of the shock. Higher the intensity of the shock, higher will be the survival chance for it.

We report the results of Cox proportional hazard regression model in Table 3.5.

Observation 3.6 (Constant Posting) *Maintaining high posting frequency helps keep retaining the long-term effect.*

Observation 3.7 (Similarity in Content) *Users deviating away from the content which got them to the shock have shorter survival times of shock effect, at the same time having high*

similarity in consecutive posts can lead to repetitiveness which again causes the survival to go down.

Observation 3.8 (Engagement) *On audience side, high engagement from audience helps maintain the effect of popularity shocks.*

3.8 Discussion and Implications

3.8.1 Research Questions

In this chapter, we focused our analysis on popularity shocks. We started with four research questions related to the effects of popularity shocks, longevity, and sustainability of the shock. Specifically, **RQ1** tries to study the effect of popularity shock on users posting frequency. From the RDiT results, we discover that users increase their posting frequency after the shock compared to before. However, as time passes, the posting frequency starts to decrease. **RQ2** is aimed at analyzing how does a user changes the content that they post after popularity shock. We find that not only do users alter their content after the shock, the post-shock content is also more similar to the content which leads to the shock, as compared to before. Thus, we conclude that popularity shocks indeed induce a behavior change in users who experience them. We are interested in understanding the longevity of the popularity shock, and hence we ask the **RQ3**. We used survival analysis to answer this question. We observe that most shocks are short-lived, i.e., the shocks reduce to 10% of their shock intensity within 5 days for 50% users. For **RQ4**, we were interested in knowing the factors that enhance the sustainability of popularity shock effects. We discover that repeatedly posting the same content as well as deviating away from the shock content cause low shock survival. Finally, high posting frequency and high response received from the users lead to more prolonged shock effect survival.

It is also worth discussing that a popularity shock or virality may not always occur in a positive connotation. Such shock can also indicate hate or networked harassment (i.e. negative attention) towards the creator [108]. Similarly, increased content posting frequency can be attributed to the author apologizing, explanation, or clarifications. Such hateful phenomenons can adversely affect the mental health of the creator [147] and cause instability in the community [17]. Though our work is centered only on positive popularity shocks, a potential extension to our work can be to categorize shocks into positive or negative and analyze their effect on the creator's behavior.

3.8.2 Implications

Our work provides numerous insights and observations into phenomena of popularity shocks. These insights form the basis for several implications for all three - (a) advertisers, (b) platform designers, and the (c) users.

Advertisers, or brands can adjust their marketing campaigns by understanding which users are behaving in a particular fashion that will lead to lasting popularity levels. They can also use topical information to identify if popular users identify more with their brand's content or not.

Platforms can utilize the insights from the study to devise algorithms for their trending pages. As popularity shock is found to increase users' engagement with the platform, enhancing attention towards dormant users can cause them to resume to increase their activity. Our content similarity results also show that such shocks can cause homogenization of content on the platform.

Users can learn the behaviors which lead to sustaining the effect of popularity shock. This can help them keep their increased engagement and benefit from the shock for a longer duration.

3.8.3 Threats to Validity

Like any quantitative study, our work is subject to multiple threats to validity. In this section, we attempt to list biases, data issues, and threats to the validity of our study by following the framework proposed by [146]. First, our work is based on a single social platform, and though it works and leverages features available on multiple social platforms, similar results do not have to hold. One possible point of differentiation would be that each platform has a different recommendation algorithm for recommending content to its users. However, the effect of recommendation algorithms on our results should be minimal since we study the effect of receiving a popularity shock by the user whereas, the recommendation algorithms primarily determines who and how big of a shock user will get. Our data can also suffer from representativeness - we use just a limited set of users who posted using a limited set of hashtags. This data representation could be significantly different from the general population on the platform. Another data issue that theoretically casts clouds on the analysis is that the number of views, likes, and comments are retrospective, i.e., they are not computed in real-time while they are the numbers on the platform at the time of data collection. Though we believe the practical effect on our results is limited since the majority of impressions on social media posts are received soon after posting [207]. For further validation, we tracked daily view counts of $1,374$ randomly sampled posts for the first 10 days after posting and found that 70% of total views were received in the first 2 days. Additionally, we did perform two analyses - regression discontinuity and survival analysis. We ensured that our data and modeling choices hold the assumptions, but there might be some unobserved confounders that we might not have considered. Finally, our statistical modeling required multiple parameters related to the operationalization of theories in sociology literature. Some of these parameters might not be capturing the factors that we intended to capture or that the theories proposed.

This work forms the basis for various future works related to popularity shocks. First of all, the work can be extended to a more generalized population and more social media platforms. Similarly, extending to different users could also open the potential to study the

effect of user personality or user type on how they respond to popularity shocks. Another significant improvement in this work could be by leveraging matching techniques to match users who got popular with similar content with users who did not get popular and then record average responses. This was not possible in our current work due to multiple reasons - (a) limited data and (b) the presence of too many confounders to create a propensity model for popularity prediction.

3.9 Conclusion

We performed a large-scale analysis of the effect of popularity shocks on users. Grounded in operant conditioning and increased sense of reputation, our results confirm the extent to which popularity shock leads users to post more and modify their future content to be more similar to the content that made them famous. Similarly, on analyzing the longevity of this shock, we discovered the short-lived nature of the shocks and the effects of various posting behaviors on shock longevity. We also provide factors that users could leverage for sustaining increased engagement post-popularity shock.

Chapter 4

Effect of Feedback on Drug Consumption Disclosures

Deaths due to drug overdose in the US have doubled in the last decade. Drug-related content on social media has also exploded in the same time frame. The pseudo-anonymous nature of social media platforms enables users to discourse about taboo and sometimes illegal topics like drug consumption. User-generated content (UGC) about drugs on social media can be used as an online proxy to detect offline drug consumption. UGC also gets exposed to the praise and criticism of the community. *Law of effect* proposes that positive reinforcement on an experience can incentivize the users to engage in the experience repeatedly. Therefore, we hypothesize that positive community feedback on a user's online drug consumption disclosure will increase the probability of the user doing an online drug consumption disclosure post again. To this end, we collect data from 10 drug-related subreddits. First, we build a deep learning model to classify UGC as indicative of drug consumption offline or not, and analyze the extent of such activities. Further, we use matching-based causal inference techniques to unravel community feedback's effect on users' future drug consumption behavior. We discover that 84% of posts and 55% comments on drug-related subreddits indicate real-life drug consumption. Users who get positive feedback generate up to two times more drugs consumption content in the future. Finally, we conducted an anonymous user study on drug-related subreddits to compare members' opinions with our experimental findings and show that user tends to underestimate the effect community peers can have on their decision to interact with drugs.

This chapter is partly a reproduction of paper published at the AAAI International Conference on Web and Social Media (ICWSM) 2023 [78].

4.1 Introduction

In 2019, 70,630 people died due to drug [1] overdose in the US alone; this number has almost doubled from 38,329 in 2010 [141]. The US president declared the drug crisis as a national public health emergency in 2017.[2]

A similar increase has also been observed in drug-related user-generated content on social media. The number of unique users in *r/Drugs* has gone up by 324% between 2012 and 2017 [116]. Anonymity and limited content moderation make Reddit[3] an appealing platform for participating in unfiltered conversations on shared interests.

Though drug-related conversations on Reddit vary widely in their purpose, we are particularly interested in content that indicates offline drug consumption by a user.[4] These can be content where a user directly talks about their experience with consuming drugs, e.g., *Just downed this bad boy! 473mg tonight, wish me luck boys!* Sometimes content may not talk about a drug experience directly but indicate the intent of drug consumption, e.g., *I recently got two orange pyramid geltabs and was wondering if I should never handle them like tabs or if they are ok to touch a little bit.* These content pieces are interesting because they are online proxies for authors consuming drugs offline. Hereafter, we call user-generated content (post or comments) like these *drug consumption activity*.

An increasing amount of research has used Reddit to study various drug-related problems like drug abuse [71], forecasting drug overdose [129], transition into drug addiction [116], patterns of drug use and consumption methods [12], and geospatial patterns in drug use [11]. Though all these studies shine a light on the various patterns of drug consumption using digital data, none of them quantify the effect of the platform and community itself on drug consumption behavior. Research has shown online community feedback has an effect on multiple facets of users offline behavior like weight loss [32], physical activity [3], smoking and drinking relapses [180], quality of user-generated content [25] and involvement in open-source projects [189].

To fill this gap, we seek to quantify the effect of the platform and community on drug consumption behaviour. We collect data from 10 drug-related subreddit; develop a deep learning classifier to label activity as indicative of drug consumption or not, to quantify the extent of drug consumption activities. Further, grounded in *Primacy Effect* [7] and *Operant Conditioning Theory* [176], we use propensity score matching [179] to quantify the impact of community feedback on the magnitude of future drug consumption activity posted by a user. Finally, we conducted an anonymous user study on our subreddits of interest to collect

[1] In this chapter, the term "drug" represents illicit substances and not generic medical drugs.

[2] `https://www.cms.gov/About-CMS/Agency-Information/Emergency/EPRO/Current-Emergencies/Ongoing-emergencies`

[3] `https://www.reddit.com`

[4] Disclaimer: We do not oppose the existence or the way these subreddits function - as they can be helpful for support and harm reduction. Similarly, we do not view drug consumption negatively or condone it, as a sizable population might be indulging in it due to therapeutic or other social factors.

members' acknowledgment of drug consumption and opinion on the effect of community feedback on their subreddit and drug consumption behavior.

We discover that (1) deep learning classifiers can identify Reddit content indicative of drug consumption (macro F1 79.54), (2) 80.29% of users in drug-related subreddits have online activity indicating drug-consumption offline, which is in line with the response received in our user study, (3) 84.2% and 54.4% of all posts and comments posted on drug-related subreddits are indicative of drug consumption; (4) users' who receive positive feedback (comments or score) from the community on drug consumption activity tend to generate up to two times more drug consumption content in future, and finally (5) user's under-estimate the effect of community feedback can have on their decision to interact in drugs.
In summary, our main contributions are:

1. To reveal (using 10 subreddits) the causal effect online community feedback has on users' offline drug behavior.

2. A manually annotated dataset (4,000 samples) and deep learning classifier to detect UGC indicative of offline drug consumption.

3. An anonymous user study of drug-related subreddits members to compare community opinion with our statistical findings.

Our work impacts researchers, platform owners, and community moderators, providing a fertile base for developing harm-reduction research and tools. Our classifiers can be used to detect social media content indicative of drug consumption, providing opportunities for demographic-specific censoring or intervention. Our causal inference results and experiment setup can help platforms/communities design different methods of showing and providing feedback that can assist in harm reduction.

Data and Code: Reddit data is available via Pushshift API. [5] Our annotated dataset, user study responses, and modeling code is available at *https://precog.iiit.ac.in/research/drug-feedback/*.

4.2 Theories and Research Questions

Individuals prefer to present an idealized version of themselves; this phenomenon is known as *Impression Management* and is used to improve social standing among peers [59]. Leary et al. [105] showed that individuals indulge in voluntary risk-taking activities like consumption of drugs, distracted driving, unprotected sex to improve impression among peers. Hogan [70] extends the concept of impression management to social media. He states that social media users can use status messages and media posted by them as a tool for impression management.

[5] `https://github.com/pushshift/api`

Subreddits are communities where having a positive impression/reputation can lead to various tangible and non tangible benefits like status, moderator privileges, Karma[6] and trophies. Thus we expect users could post drug consumption content to improve their impressions. Hence, we ask our first question:

RQ1. [Extent] *What is the extent (i.e. percentage of content, and users) of content indicating offline drug consumption in drugs-related subreddits?*

Our second research question is grounded in the *Primacy effect*, the tendency to remember the first piece of information [7]. For e.g., people's impression of an individual is dependent on the first traits they encounter [7]; probability of recalling initial items in a list is higher [135]; people have a more vivid memory of their first romantic encounter, achievements, and even losses [46]. The primacy effect can cause *anchoring bias*, leading to skewed decisions relaying heavily on the initial information [187]. Building on these theories, [170] proposed *outcome primacy*, proving long-lasting effects of the first experience. We hypothesize that the community feedback on the first drug consumption post can affect the user's future drug consumption and posting behavior.

RQ2. [First Experience] *How does the community feedback on first drug consumption post affect users' future drug consumption?*

Besides feedback on the first experience, user experience can also be dependent on *law of effect*, actions that are closely followed by satisfaction are more likely to reoccur [185]. Based on this principle, Skinner et al. proposed *Operant Conditioning* [176]. It states the probability of acting in the future is a function of the outcomes received in the past. Positive reinforcement will incentivize the user to repeat an action in the future. Similar behavior is observed in the context of social media, e.g., more number of comments on post leads to higher weight loss [32], increased social media interactions lead to higher steps in activity tracking apps [3] and community feedback affects the quality of future posts [25]. Grounded in these theories, we expect continued positive feedback can affect a user's future drug consumption activity. We therefore ask:

RQ3. [Feedback] *How does continuous positive community feedback affect users' future drug consumption?*

Before studying the causal effect of feedback, we need to be able to detect drug consumption activity. An essential prerequisite to our work is building a classifier that can predict users' drug consumption in the offline world via user-generated textual content. A popular methodology in Natural Language Processing (NLP) is to learn dense representations of text. Mikolov et al. [130] proposed a neural algorithm to learn text representations based on word co-occurrence, which outperformed classical token-based representation in a variety of classification tasks. Vaswani et al. [190] proposed an improved model architecture called Transformers based on self-attention [169] to learn contextually aware dense text representations. Transformer-based large pre-trained models [44, 113] have provided an efficient base

[6] `https://reddit.zendesk.com/hc/en-us/articles/204511829-What-is-karma-`

to perform classification on a variety of tasks and data sources. We build a deep learning classifier based on these architectures, asking:

RQ4. [Detection] *Can we use Reddit textual data to classify between drug consumption and non-drug consumption content? How accurate is such a classifier?*

4.3 Related Work

Our work is about the effect of social media community feedback on users' drug consumption behavior. Our related work flows from three directions - (1) Drug studies leveraging social media, (2) Causal inference using online data, and (3) Self-harm behavior on social media.

Drug Studies on Social Media: Ease of data availability and many active communities around drugs have enabled a variety of related research. John et al. [116] built a machine learning classifier trained on textual features to identify users at risk of addiction and transition into drug recovery. They further use survival analysis to identify how much time it will take to undergo the transition. Duilio at al. [12] used Word2Vec [130] similarity to curate a list of words used by Reddit users for different drugs, Routes of Administrations (ROA), and drug tampering techniques. Using the list, they rank the popularity of various drugs and ROAs. They report that between 2014 to 2018, the popularity of synthetic drugs like Fentanyl and unconventional ROAs like rectal administration of drugs has increased, whereas a decline has been observed in conventional ROAs like inhaling and injecting. Duilio at al. [11] filtered all activities of users on drug subreddits to extract location information and study the geospatial patterns of drug consumption in the US.

Besides Reddit, [71] used deep learning ensemble models to detect drug abuse in tweets and [129] used community attentive neural networks to forecast drug overdoes using information about crime dynamics.

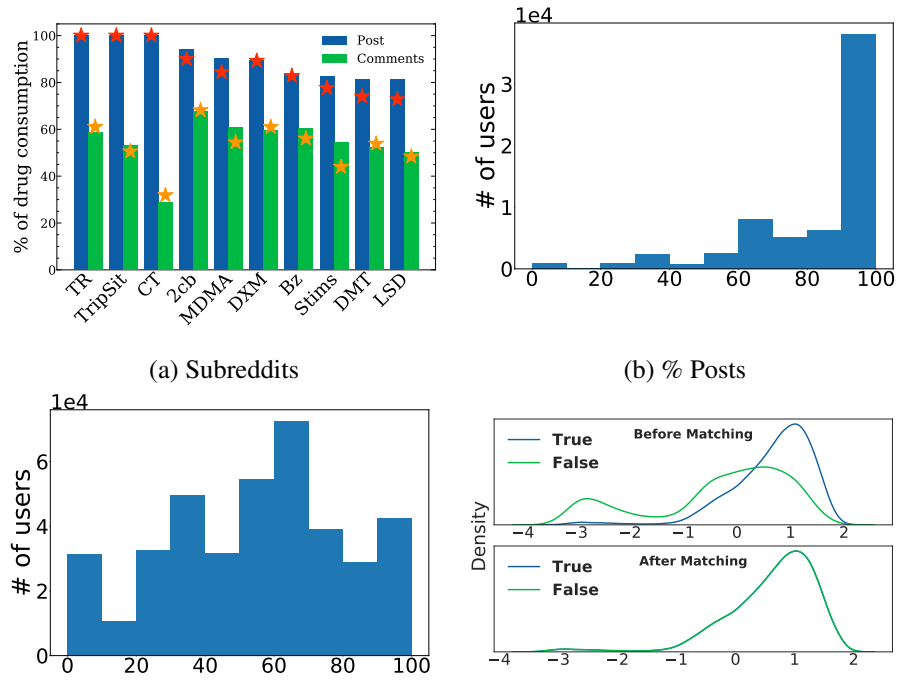
Causal inference using online data: Traditionally, researchers have established randomized controlled trials to establish causations. However, due to logistically and ethical concerns, such trials are not always feasible; e.g., it is not ethical and legal to make subjects consume illicit substances to study feedback's effect. For such studies, research has utilized publicly available online data. Additionally, the Internet provides a large volume of data, which is logistically impossible to obtain from controlled physical experiments. Careful filtration and analysis of large online data can help us simulate a randomized control trial [158].

[32] showed that positive feedback from the online community could help users lose more weight. Althoff at al. [3] studied data from an exercise logging application and found increased social connections on the platform caused higher physical activity in the offline world. Tamersoy at al. [180] used survival regression to establish a causal relation between linguistic cues from user-generated content and smoking or drinking relapse. Kiciman at al. [91] used social media posting behavior to identify alcohol consumption and academic success of college students. Their analysis proved a causal relationship between high alcohol

consumption and poor academic performance. Choudhury at al. [41] unveiled the causal relation between user's vocabulary and suicidal tendency.

Online data have also shown effects in opposition to expectations, e.g., [25] showed that negative community feedback leads users to create even worst quality posts in the future rather than improving. Repercussions of feedback are topic and community dependent. Lack of literature analyzing feedback in drug and self-harm communities makes it an important area to study.

Self-harm behavior on social media: In the context of impression management, it has been shown that users tend to take part in self-harm activities like drug consumption and unsafe sex to improve social standings [105]. An increasing number of users are getting involved in dangerous social media challenges like the KiKi Challenge [9], the Salt and Ice Challenge [160], the Cinnamon Challenge [62], Tide Pod Challenge [136], and the Fire Challenge [1]. Lamba at al. [100] analyzed public Snapchat data from 173 cities around the world, revealing 23.5% of total 6.4 Million samples were examples of distracted driving. They performed demographic analysis to reveal that young males from the Middle Eastern and Indian subcontinent are more likely to produce distracted driving content. Similarly, [138, 96] analyzed deaths caused by taking selfies in dangerous situations like elevation, near water-bodies, or with firearms.



(a) Subreddits

(b) % Posts

(c) % Comments

(d) Propensity Logits

Figure 4.1: (a) Perecentage of drug consumption content across subreddits. Values derived from proposed model are indicated by bars, and ⋆ shows values from manual annotation. (b)&(c) are distribution of % posts and % comments indicating real world drug consumption per user. (d) Distribution of propensity logits before (top) and after (bottom) matching.

Table 4.1: Statistics about the data collected.

Subreddit	# of Post	# of Comments	# of Users	# of Users with Post
LSD	343,346	2,658,323	266,185	138,073
MDMA	113,030	1,022,810	103,900	55,149
Benzodiazepines	107,264	794,141	55,823	32,887
Stims	84,049	710,692	51,848	24,440
DMT	81,860	753,570	79,215	35,005
DXM	60,555	486,052	30,989	18,795
Currentlytripping	17,388	50,757	19,540	6,650
2cb	9,258	83,642	11,348	5,317
TripSit	7,780	76,329	16,267	5,609
TripReports	2,148	10,991	3,791	1,659

4.4 Data Collection and Dataset

We use Reddit, a widely used social media platform. Reddit is formed by a collection of communities called *subreddits*. As of October, 2021, Reddit has 52 Million daily active users and 3 Million subreddits.[7] Subreddits are largely allowed to moderate their own community posts and the anonymity allowed, makes it a suitable platform for relatively unfiltered discourse compared to other social media platforms. Each subreddit is built around a specific topic. Users post content related to their interest and fellow users can *comment* on these posts, which creates a *thread*. Users can also *upvote* and *downvote* a post or comment, though only the total aggregate of votes is visible called *score*.

Reddit has several subreddits built around the topic of drugs. Wiki page of *r/Drugs* maintains a list of popular drug-related subreddits.[8] These subreddits contain different facades of drugs like addiction, recovery, cultivation, and experience. Some are drug agnostic like *r/tripreports* whereas others are drug specific like *r/MDMA* or *r/LSD*. We manually audited all the subreddits in the list and filtered 10 subreddits (see Table 4.1 in appendix), which is either (1) based around users sharing personal drug consumption experiences or (2) has a popular *flair*[9] indicating offline drug consumption.

To obtain the data from Reddit, we use the Pushshift API. For each subreddit, we collected all the threads made from the inception of the subreddit. Each thread contains the original post, the comments made, and scores for all activity in the thread. In total, we collected 826,905 posts and 6.6 Million comments made by 493,906 unique users. Only 269,059 unique users at least have one post. Table 4.1 provides a summary of statistics for each subreddit.

[7] `https://backlinko.com/reddit-users`

[8] `https://www.reddit.com/r/Drugs/wiki/subreddits`

[9] `https://www.reddit.com/r/help/comments/3tbuml/whats\_a\_flair/`

The impact of our research can be dependent on two factors: 1) Do the users actually consume drugs in real life, and 2) Analyzing causal inference results in light of members' perception since it can dictate the design of the effective intervention and education strategy.

To this end, we conduct a voluntary anonymous user study with members of 10 subreddits we are studying. Necessary permissions from the Institute's Review Board and moderators of subreddits were obtained before conducting the user study. Firstly, participants were asked to acknowledge (Yes or No) if they consumed drugs during their active period on the subreddit. Later, they were asked a series of questions about how much impact community feedback, number of comments, and score have on their future participation in subreddit and drug consumption. We wanted a quantitative understanding of users' perceptions rather than a simple yes/no answer while keeping the study's cognitive load low. Hence, we opted for the 5-point Likert scale [111], 1 being *No Impact* and 5 being *Essential*. User study questioner can be accessed at *https://forms.gle/yRqRriSPbgG9p2gN8*. Total 45 users participated in our study. Results of each component are presented with the corresponding computational results.

4.5 Detecting Drug Consumption Content

To understand the extent of drug consumption behavior (**RQ1**), we first need to identify which user-generated content indicates drug consumption in real life (**RQ4**). Past research has assumed being active on drug-related subreddit as a proxy of drug consumption [116, 12, 11]. Though this may be true in most cases, users can also join the community as bystanders, for research purposes, or to help others. Hence, considering mere participation as a proxy of drug consumption is a weak assumption. Some subreddits have flairs that indicate drug consumption, but adding flairs to post is voluntary, and users may choose not to do so. Moreover, comments do not have a flair but still can indicate drug consumption. Towards solving this, we build a classifier that can mark posts and comments as indicative of drug consumption or not. Henceforth, we will use the term *activity* to represent user-generated posts or comments.

4.5.1 Ground Truth Annotation

To build a classification model, we need to have a ground truth dataset of activities labeled as drug consumption or non-drug consumption. The goal is to mark a sample as positive if it indicates the author consuming drug offline. We sample 4,000 user activities for annotation. To ensure a well-distributed ground truth, half of the samples were posts, and half were taken from comments. Further, a uniform split is maintained across all 10 subreddits.

Annotation Guidelines: Annotators were provided with the text content and title (in case of posts) of an activity. An activity should be annotated as drug consumption in case of self-disclosure by the author, or if clear indication of author's possession/intent to consume drugs is present. For example:

- Self-disclosure: *Haha I had a bad trip off 30mg and weed first time but can't wait to try smaller doses.*
- Intent to consumption: *I'd be up for a distanced experience with a stranger (s) Just itching to get out of this awkward routine....*
- Drug possession: *I'm thinking about dissolving it in some alcohol and putting it in empty caps, not sure it will be better...*

It is important to note that just the presence of drug-related words does not imply drug consumption. The following examples contain drug-related words but do not indicate drug consumption:

- *My fourteen year old niece is smoking pot. What would /r/trees tell a fourteen year old about the effects of Marijuana? She might believe YOU and sources you cite.*
- *Mdma tolerance information. Does anyone have any information on immediate mdma tolerance or articles about the subject*

Annotators were also given a list of drugs *street* names and slangs used in drug-related subreddits to assist the annotation process [12]. Each sample was annotated by 3 annotators independently. We obtain a Fleiss-kappa [52] agreement rate of 0.69, which signifies substantial agreement [102]. An activity was marked as drug consumption if 2 or more annotators agreed.

Dataset: 2,614 (65.32%) of 4,000 samples were marked as drug consumption, 79.35% of posts and 51.30% comments were marked as positive, respectively. Since comments are made in response to posts providing specific information, feedback, or expressing gratitude, a lesser positivity rate of drug indication than posts is expected. We make our annotated data public for future use. [10]

4.5.2 Deep Learning Classifier

We randomly split the manually labeled dataset into a train and test set of 3,200 (2,091 drug consumption, 1,109 non-drug consumption) and 800 (523 drug consumption, 277 non-drug consumption). Five-fold cross-validation is performed on train set to tune models, and final models are evaluated on the test set.

Model: Performing text classification with a combination of neural models and dense text representations has become a norm in NLP. Following the same, we experiment with different types of neural network models combined with contextual and non-contextual text embeddings. Our first model is a single channel one-dimensional convolutional neural network (Text-CNN) [92]. Input text for the Text-CNN model is vectorized using pre-trained Google News corpus Word2Vec embedding [130].

[10] `https://precog.iiit.ac.in/research/drug-feedback/`

Transformer-based large pre-trained models with their ability to capture sentence context have achieved state-of-the-art performance on a variety of NLP tasks [190]. Leveraging that, we experimented with BERT, a model built using bidirectional transformers and pre-trained on masked language model, and next sentence predictions tasks [44]. We also built a classifier based on RoBERTa [113], an optimized version of BERT. Table 4.2 reports 5-fold cross-validation performance of all the models.

Training Details: Our model is trained using Adam optimizer with the learning rate of $3X10^{-4}$, batch size 64, and utilized dropouts for regularization. We train models for 100 epochs with early stopping and checkpointing the best-performing model on the validation set. The training was performed on an Nvidia RTX 3090 GPU. Our code is available publicly for reproducibility and future use purposes.[10]

Validation and Robustness of Classifier: To further validate the generalizability of our models, we validate its performance on the test set (not used in the training step). Table 4.2 provides performance of all models on test set. Our best model achieve a macro F1 score of **79.54**. Table 4.3 in appendix provides performance numbers of our best model across subreddits.

4.6 Extent of Drug Consumption

We want to discover the extent of content on drug-related subreddits that indicates offline drug consumption by the user (**RQ1**). We use the proposed classifier to generate predictions for all the activities (posts or comments) that are not already marked as drug consumption by a flair or subreddit. We found that 84.2% of all posts and 54.4% of all comments indicate drug consumption by the user. Figure 4.1(a) shows the percentage of drug consumption posts and comments for each subreddit individually. $\star$ in the Figure indicates the drug

Table 4.2: Drug consumption classification performance.

Model	**Accuracy**	**Macro Precision**	**Macro Recall**	**Macro F1**
5-fold cross validation				
Text CNN	73.51 ± 3.10	78.79 ± 1.82	63.28 ± 5.78	62.34 ± 7.28
BERT	83.79 ± 1.03	82.13 ± 1.15	82.22 ± 1.29	82.14 ± 1.16
RoBERTa	83.16 ± 0.95	81.72 ± 1.32	80.96 ± 1.15	81.22 ± 0.98
Test set				
Text CNN	78.65	77.52	73.71	74.89
BERT	81.27	79.51	78.67	79.05
RoBERTa	**81.90**	**80.43**	**78.89**	**79.54**

Table 4.3: Performance of proposed model across subreddits on test set. Sorted by Macro F1 score.

Subreddit	Accuracy	Macro Precision	Macro Recall	Macro F1
TripSit	88.24	88.77	88.24	88.19
DMT	89.53	88.14	87.35	87.73
Stims	84.72	84.70	83.28	83.82
Currentlytripping	82.35	81.60	84.47	81.79
LSD	82.93	82.11	81.21	81.60
2cb	84.71	78.58	81.35	79.77
DXM	85.86	77.92	81.80	79.55
TripReports	78.57	78.15	78.47	78.26
Benzodiazepines	82.35	86.24	71.17	74.09
MDMA	76.92	74.44	69.30	70.68

consumption percentage observed in our manual annotation. A consistent slight difference between predicted and annotated drug consumption percentages shows the proposed model's robustness across subreddit and content types.

Once we have drug consumption prediction for all the activities, we aggregate them based on user ids and observe what percentage of users have activities for whom we have a positive prediction. We found that across 10 subreddits, 80.29% of all users in our dataset have consumed drugs. This is echoed in our user study findings too, where 84.4% participants (38 out of 45) acknowledged consumption of drug. As shown in Figure 4.1(b) $90\% - 100\%$ of posts for most users are indicative of drug consumption offline. The distribution of user's comments is less skewed, centered around the 60%-70% (Figure 4.1(c)). This proves a strong proxy between user activity on drug-related subreddit and drug consumption in the offline world and signifying the importance of studying the platform's impact on users' future online and offline activity.

Observation 4.1 (Extent) *User-generated content of about 80% of total users in drug-related subreddits indicates drug consumption in real life. This is inline with the data received via our user study.*

Observation 4.2 (Extent) *84% user-generated posts and 54% comments indicate drug consumption. For majority user 90%-100% of their posts and 60%-70% of comments indicate offline drug consumption.*

4.7 Causal Analysis

In **RQ2** and **RQ3**, we aim to understand the causal effect that receiving positive feedback on drug consumption posts has on the users' future drug consumption activity. To this end, we use Propensity Score matching, a causal inference model shown to reduce bias compared to the naive correlation analysis [73].

In the potential outcome framework [140], the "effect" of an experience on the outcome is formalized as an outcome $Y_i(T=1)$ after a person i had the target experience T, i.e., treated,[11] and outcome $Y_i(T=0)$ when the same person in the same circumstances has not received the treatment. The causal effect of the experience T is estimated as $Y_i(T=1) - Y_i(T=0)$. However, it is impossible to have the same individual receive and not receive treatment simultaneously. Propensity score matching attempts to overcome this challenge by observing the outcome on two different individuals, one treated and the other control but having similar treatment probability and confounders.

Feedback Threshold: In our case, treatment is the feedback received on a drug consumption post which is measured by the number of comments and scores [12] received. Conventionally, treatment is a binary variable (e.g., vaccine administered or not), and hence the assignment of treatment is trivial, but in our case, treatment is a continuous variable. We use hard thresholds (θ) to divide feedback into positive or negative and present results across various values of θ. It means, a post is considered to have positive feedback if it receives greater than θ number of comments or score. Averaged across our 10 subreddits, 80% of drug consumption activities receive less than 1.1 ± 0.3 comments and 2.9 ± 1.13 scores. To ensure robustness and generalizability in results, we experiment by varying our θ from 2 to 6, both inclusive.

Group Assignment: In **RQ2**, we analyze the treatment outcome on a user's first-ever drug consumption post. User is assigned to the treatment group if their first drug consumption post receives positive feedback. Additionally, in **RQ3**, we aim to study the effects of continuous feedback. A user at their n^{th} drug consumption post is assigned to the treatment group if all their past drug consumption posts, including n^{th}, have individually received positive feedback. We experiment with values of n between 1 to 6, both inclusive.

Propensity Model and Matching: After group assignment, we need to find pairs of users who have a similar likelihood of receiving treatment, but one is treated, and the other is not. In our case, given drug consumption post n and feedback threshold θ, propensity model estimates $P(n_{feedback} \geq \theta)$. Latent confounders encoded in linguistic and content characteristics, past feedback, and volumes can affect a post's feedback. In our experiment, we account for all these confounders while matching to create balanced treatment and control

[11] In causal analysis literature, the subject who received the target experience is called treated and becomes part of the Treatment group. Whereas users who do not receive the target experience are referred as Control group.

[12] Score is an aggregate of number of up votes and down votes received. Only the aggregate is reported by Reddit not the individual values.

groups. Choices of our confounders are inspired by previous work using causal inference on social posts like [25, 167, 180], and can be divided into 3 broad categories:

- *Content text*: User-generated textual content can give a measure of multiple confounders. In our case also, text of the drug consumption post is the main confounder and is being used for propensity score prediction.
- *Past activity*: Apart from text, we also use users' frequency of past activity as a confounder. While performing matching, only users with a similar number of posts and comments done in the past are paired together.
- *Past feedback*: Another important confounder regularly used in literature is the feedback (scores and comments in our case) received by a user in the past.

Multiple recent social media causal inference studies have used text-based models for propensity estimation [25, 178, 41, 91, 167]. Most of these studies use a combination of n-gram features and Logistic Regression to train the propensity model [89]. However, recently [198] showed that choice of model architecture for text propensity model could induce bias in causal inference results. They experimented with a wide range of text representations (n-grams, LDA, contextual embeddings) and architectures (Logistic Regression, Simple NN, and BERT-derivatives) and found that BERT-based models were least prone to induce bias.

Considering [198] findings, we use pre-trained RoBERTa [113] model for propensity estimations. Since subreddits may have different community dynamics and rules, separate models are trained for each subreddit across the range of feedback thresholds.[13] Size of training data was capped at 10,000 samples. An $80:20$ train test split was used for evaluation. Accuracy and macro F1 of our propensity models varied for subreddits between 57.8% to 89.2% and 43.3% to 70.1% respectively. It is important to note that a propensity model aims to build a descriptive selection model and not a predictive model [158], and hence, the importance of classification performance is secondary [91]. Further, [198] demonstrated that a highly accurate propensity model could induce bias in the estimation of the causal effect. Therefore, we move forward with propensity models having moderate performance.

Matching: A user in treatment group is matched with one user from control group when posts made by both have a similar propensity score. Generally, given a propensity score p, matching is done on $logit(p)$ (Equation 4.1). A pair is considered as a good match, if difference of $logit(p)$ is less than a *caliper* value as defined in Equation 4.2 [66].

$$logit(p) = ln\left(\frac{p}{1-p}\right) \tag{4.1}$$

$$caliper = 0.25 \times \sigma(logit(p)) \tag{4.2}$$

[13]Training parameters were similar to those presented in Section Detecting Drug Consumption Content. Training code is present in our code repository `https://precog.iiit.ac.in/research/drug-feedback/`

For a given treatment user, we filter all control users with $logit(p)$ difference less than $caliper$ value and then conduct a greedy search to find the nearest value. Matching is done in a one-to-many fashion.

Apart from propensity score, number of past activities and feedback received in past should also be balanced as confounders [25]. We ensure balance by matching the n^{th} drug consumption post made by both users, and treatment is only assigned if all the drug consumption posts from 1 to n individually receive positive feedback.

Quality of matching: Finally, to ensure the treatment and control group after matching are statistically similar, we use standardized mean difference (SMD) also known as Cohen's D [179] defined as: -

$$SMD = \frac{\bar{x}_{treatment} - \bar{x}_{control}}{\sqrt{\frac{\sigma^2_{treatment} + \sigma^2_{control}}{2}}} \tag{4.3}$$

Here, $\bar{x}$ and σ represent mean and standard deviation, respectively. To ensure matching quality SMD is preferred over p-value hypothesize testing since it conflates changes in balance with changes in statistical power [179].

In literature, where text propensity models are built on n-gram features, SMD balance check is conducted on n-gram vectors [32, 91, 167]. Since our propensity model is deep learning-based, we use feature vectors extracted from the last hidden layer of our model along with the frequency of user past activity to conduct a balance check [84, 115, 81]. We evaluate the SMD distribution of feature vectors before and after matching for treatment and control users. A confounder is considered to be balanced if SMD is less than 0.25 [179].

Effect Size: Once we have our treatment and control groups statistically balanced upon confounders, effect of treatment can be calculated on the matched pairs. Estimated average treatment effect ($EATE$) is calculated as:-

$$EATE = \frac{\sum_{i=1,j=1}^{N} \frac{(Y_i(T=1) - Y_j(T=0)) * 100}{Y_j(T=0)}}{N} \tag{4.4}$$

$EATE$ gives an average percentage increase in the treatment group's outcome compared to the control group's outcome. Since the distribution of the treatment effect can be skewed, we report median values instead of mean.

4.7.1 Feedback on First Drug Consumption Post

We study the effect number of comments received by the first drug consumption post has on future drug consumption activity volume (**RQ2**). A user is assigned to a treatment or control group based on the number of comments received on their first drug consumption post. [14] We experiment with comment thresholds (θ) between 2 to 6.

[14] Note that the first drug consumption post here represents the first post of the user which indicative of offline drug consumption in the subreddit. We do not claim this to be the user's first encounter with drugs in life.

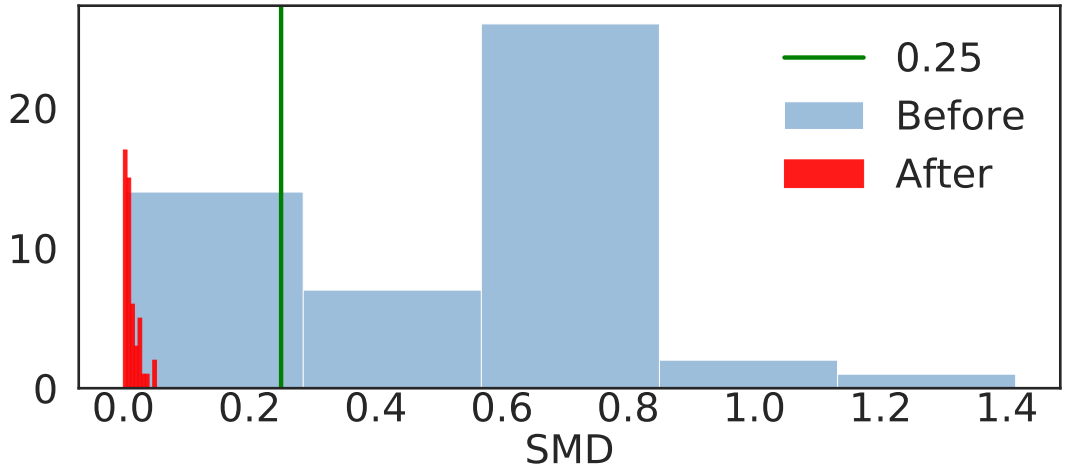


Figure 4.2: Matching quality for r/LSD, n_1, $\theta = 4$. Distribution of confounders' SMD before and after matching. After matching *SMD* for all confounders in ≤ 0.25 indicating good quality matching.

We discover users who received positive feedback on first drug consumption post, generated upto 100% more drug consumption content in the future compared to the users in the control group. These results are statistically significant (p<.001), evaluated using Kolmogorov-Smirnov test [122] and consistent across different treatment thresholds and subreddits. Table 4.4 shows $EATE$ of n_1 for all the subreddits calculated on $\theta = 4$. Figure 4.2 shows change in confounders *SMD* and Figure 4.1(d) changes in $logit(p)$ distributions before and after matching for *r/LSD* n_1, $\theta = 4$.

4.7.2 Continuous Feedback on Drug Consumption Posts

Additionally, we check the causal effect when a user continuously receives positive feedback on drug consumption posts (**RQ3**). We repeat the matching experiments to evaluate the $EATE$ of the same outcome when the user receives consecutive positive feedback on their first n drug consumption posts i.e. all 1 to n drug consumption posts got positive feedback individually. Averaged across our 10 subreddits, we observe 80% of the users posts less than 6.9 ± 2.5 drug consumption activities in our time of observation. We experiment with values of n between 2 to 6. Table 4.4 shows the results for $\theta = 4$. We observe that treated users performed a higher number of drug consumption activities in the future. Our results are statistically significant. However, we do get insignificant results for experiment configurations with high values of θ and n due to the lack of enough matching pairs. This is more pronounced in smaller subreddits. However, we never receive a statistically significant result that conflicts with our hypothesis.

4.7.3 Score as Feedback

We also conduct all configurations of our experiments with the score as the treatment variable. Just as with comments, we receive consistent and statistically significant results; an increase in future drug consumption activity for treated users. Table 4.4 show results for $\theta = 4$. To ensure robustness we experiment across a wide range of parameters ($\theta = [2,6], n = [1,6]$, comments and score as feedback) for each subreddit, leading to ≈ 600 experiment configurations. Results across the various experiment configuration are inline with our hypothesis and statistically significant. Complete results and statics of matching quality (before and after confounder *SMD* distributions), $EATE$, number of treatment control pairs, and statistical significance across all configurations are available at *https://precog.iiit.ac.in/research/drug-feedback/*.

Observation 4.3 (Increased Volume) *Positive community feedback on drug consumption posts (first and continuous) causes an increase in future drug consumption activity.*

Though causal inference shows a significant impact of community feedback on users' future participation and drug consumption, the impression of community members in our user study differs. Participants, on average, reported a *little* to *moderate* impact of community feedback on their behavior. On a 5 point Likert scale (1=*No Impact*, 5=*Essential* the average response was $2.28/5$ for scores and $2.53/5$ for comments. Such phenomenon of users under-estimating the effect of external factors on their participation in self-harm activity to maintain an "illusion of control" is well studied in the social psychological theory Layng's edgework [118]. Understanding the contrast between user opinion and statistical findings is vital to designing effective intervention and harm-reduction strategies.

Observation 4.4 (Effect Underestimation) *Users on drug-related subreddits tend to underestimate the effect community feedback has on their future engagement and drug consumption.*

4.8 Discussion

4.8.1 Research Questions

We begin our analysis with **RQ1** which aims to understand the extent of content in drug-related subreddits indicating drug consumption by a user in the offline world. Such content pieces provide a strong proxy for online-offline interaction of drug consumption and help quantify the prevalence of such self-harm behavior on social media. We discover that 84.2% of all posts and 54.4% of all comments posted on our observed subreddits indicate offline drug consumption. According to our model predictions, 80% of users have indulged in drug consumption, which is in line with the user acknowledgment we obtained from our user

Table 4.4: $EATE$ of feedback threshold (θ) 4 on the number of future drug consumption activities. n_i represents the i^{th} drug consumption activity done by an user. Positive feedback consistently leads to a higher volume of future drug consumption activity. Lack of enough treatment users lead to statically insignificant results in some configurations.

	Comment ≥ 4						**Score** ≥ 4					
Subreddit	n_1	n_2	n_3	n_4	n_5	n_6	n_1	n_2	n_3	n_4	n_5	n_6
LSD	50.0***	44.4***	35.6***	25.0***	35.0*	37.0**	50.0***	33.3***	37.5***	33.3*	53.9*	0.0
MDMA	75.0***	52.9***	50.0***	27.6*	50.0***	71.4***	41.4***	53.8***	50.0*	41.1	64.0***	129.9
Benzodi-azepines	75.0***	50.0***	35.0***	52.7***	33.3*	30.0*	50.0***	75.0***	66.7**	133.3**	183.3*	266.6*
Stims	82.6***	63.6***	42.8***	40.0***	37.5***	68.4***	38.4***	30.0*	51.9**	12.3	-13.3	158.7**
DMT	66.7***	40.0***	30.0***	20.5*	25.0***	26.7*	43.6***	32.2*	33.3*	52.3	28.2	47.0
DXM	66.6***	33.3***	45.5***	33.3***	20.0	47.2	33.3***	27.3	41.7	58.9	109.4*	85.7
Currently tripping	60.0***	100.0***	255.0**	31.6	465.3	1033.3	50.0***	60.0***	50.0***	100.0***	37.0	142.9*
2cb	100.0***	50.0***	50.0*	21.5	0.16	29.9	100.0***	83.3***	266.7	167.8	281.0	206.4
TripSit	80.0***	50.0**	33.5	14.3	25.0	17.5	33.3*	50.0	-9.4	276.3	20.0	N/A
TripReports	100.0***	44.4	41.7	21.4	-62.5	N/A	14.3	150.0	350.0	233.3	N/A	N/A

Note:$^{***}p <= .001$,$^{**}p <= .01$,$^{*}p <= .05$. N/A means no matching pairs for the configuration.

study. This distribution is consistent across subreddits irrespective of the subreddits theme (drug experience or not) or drug type. In fact, for most users, between 80% to 100% of their posts indicate drug consumption.

Primacy effect is a cognitive bias that explains people's tendency to depend on first experiences and impressions while making decisions. We validate does primacy effect holds for the users of drug-related subreddits. For social media users, feedback from the community can provide tangible and intangible benefits like gratification, a sense of belonging, special moderator status in the community. Thus in **RQ2**, we use propensity score matching to infer the causal effect positive feedback on first drug consumption post has on future drug consumption. Validated across different thresholds, we found that users who receive a high number of comments on first drug consumption post showed up to 100% increase in drug consumption indicative activity in the future.

Operant conditioning framework further expands the effect of feedback stating positive reinforcements can lead to repeated actions and habit building. In **RQ3**, we validate this by expanding our causal inference experiments to include continuous feedback received on drug consumption posts generated later in the timeline. We observe, similar to the first experience, receiving a continuous positive community feedback on drug consumption posts leads to an increase in magnitude of drug consumption activity. Observing **RQ2** and **RQ3** in tandem, we hypothesize that the feedback on the first drug consumption post can act as a "gateway" for the user; continuous feedback on later instances "reinforces" the habit. Together, positive feedback incentivizes a user to produce higher volumes of drug consumption content and, as a proxy, increased self-harm in the offline world.

Our user study unveiled, users perception of community feedback's impact on their behavior is less than what is observed statistically. Discrepancies like this have been studied in psychology literature [118] and can pose a danger to users well-being.

Finally, to answer our research questions, we need to classify subreddit activities (posts or comments) into indicative of drug consumption or not. Leveraging the large scale data available, to answer **RQ4** we train a deep learning classifier capable of classifying activities into offline drug consumption or not with high precision and recall. We further validate the robustness of the proposed model by evaluating performance on the test set spread across subreddits.

4.8.2 Implications and Ethical Considerations

All subreddits involved in our work list harm reduction as one of the community's primary goals. We believe our models and findings have direct implications for community moderators and platform designers involved in harm reduction interventions.

Feedback based: One of our key insights is increased drug consumption activity by users who received positive community feedback. Thus communities can experiment with different strategies of showing feedback, like only showing counts, partial, or rate limited feedback and quantify the reduction in said effect. Our insight and models can also help design community feedback guidelines regarding limiting community interactions on specific activities.

Intervention based: User's feedback history combined with our proposed deep learning classifier can help in monitoring drug consumption activity at an user or cohort level. High-risk individual(s) can be detected, and timely interventions like notifying, community reach outs, or restricted activity can help in reducing overall self-harm.

Such interventions may also have adverse effects; hence, more experimentation is required before moving forward. We acknowledge that tracking user data and restricting platform usage patterns can violate privacy and freedom of expression. However, our work does not aim at providing specific intervention methods. Instead, we provide necessary insights, data, and models that researchers and community moderators can use for further work based on every community's rules and ethics.

Resource based: A variety of research can be conducted on these platforms to understand and prevent the harms caused by drug consumption. However, the validity of any such work is dependent on ensuring that the online content provides a strong proxy for offline drug consumption. We open-source a manually annotated dataset and our pre-trained models from drug consumption classification to enable further research.

4.8.3 Threats to Validity

It is always challenging to ensure generalizability while analyzing pseudo anonymous online data. Our analysis is also susceptible to these challenges. Firstly, our data is collected through Reddit, which can have biased representations in terms of geography, gender, and age. Further,

though we experiment with 10 different drug-related subreddits varying across size, time, drug, and community objective, some other subreddits or social media platforms may not follow our insights. Finally, the users posting about drug consumption online may themselves not be a fair representation of the population engaging in drug consumption. However, since these people are consuming drugs and publicly generating content about it, we believe it is an important demographic to study if we aim to understand the online-offline connection of drug consumption behavior.

In our analysis, user-generated drug consumption content is used as a proxy for offline drug consumption by the user. Since our data source is online, we do not have any way to ensure that the user did consume the drugs. We use data spread across various communities and long timelines adding up to millions of activities reducing the possibility of large scale tampered data. Further we perform a voluntary and anonymous user study in same communities to get acknowledgment of drug consumption. Our analysis and user study responses are based on the belief that users are not putting out false experiences. Additionally, it is necessary to note that the absence of online drug consumption content is insufficient for proving users not congesting drugs offline. Our study does not aim to make conclusions about drug consumers who do not actively interact with the platform.

Our experiments do not account for the sentiment of comments received to prevent errors in sentiment identification propagating to causal inference results. Due to drug/self-harm content dynamics, off-the-shelves sentiment models can cause unforeseen biases. A potential future work can be to train topic-specific sentiment models and observe their effect on the outcome.

Additionally, we control multiple contents, user, and community confounders while setting up our causal inference pipeline. However, there is always a possibility of unaccounted variables leaking into the causal inference outcomes. Finally, the sample size of our user study is small. Though this does not affect the primary statistical findings of our work, a more extensive and exhaustive study is desirable.

4.9 Conclusion

Our study investigates user-generated content indicative of drug consumption in the offline world. Specifically, we collect publicly available data from 10 drug-related subreddits and analyze the extent of drug consumption activity in these communities. First, we build a text-based deep learning model to classify user activities into drug consumption or not. Adapting from the sociology literature of feedback, we aim to test if the theories proposed for the offline world are also applicable to the behavior of posting drug consumption content on a social media platform. We put forth multiple RQs related to feedback's extent and causal effect on such behavior.

In summary, we observe that the majority of content posted on drug-related subreddits indicates drug consumption in the offline world. Further, we discover that users who receive

positive community feedback on drug consumption content tend to generate higher volumes of similar content in the future, though users seem to underestimate this effect as shown by our user study. We believe that the observation made in our work can help to design online feedback mechanisms and interventions to reduce self-harm.

Chapter 5

Together Apart: Decoding Support Dynamics in Online COVID-19 Communities

The COVID-19 pandemic that broke out globally in December 2019 put us all in an unprecedented situation. Social media became a vital source of support and information during the pandemic, as physical interactions were limited by people staying at home. This chapter investigates support dynamics and user commitment in an online COVID-19 community of Reddit. We define various support classes and observe them along with user behavior and temporal phases for a coherent in the community. We perform survival analysis using Cox Regression to identify factors influencing a user's commitment to the community. People seeking more emotional and informational support while they are COVID-positive stay longer in the community. Surprisingly, people who give more support in their early phases are less likely to stay. Additionally, contrary to common belief, our findings show that receiving emotional and informational support has little effect on users' longevity in the community. Our results lead to a better understanding of user dynamics related to community support and can directly impact moderators and platform owners in designing community guidelines and incentive structures.

This chapter is partly a reproduction of paper published at the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) 2023 [77].

5.1 Introduction

As of July 2023, the number of people affected by COVID-19 worldwide stands at 690 million, with the death toll reaching 6.9 million. To tackle the pandemic, authorities imposed travel bans, movement restrictions, and closed public places to reduce the spread. Such drastic changes in the daily routine and uncertainties of the pandemic took a toll on people's mental health. Studies show that COVID-19 has a consistent negative impact on mental health, which has led to an increase in anxiety, depression, and Post-traumatic stress syndrome (PTSS) [144, 193, 114]. It has also increased the number of people looking for support

and mental healthcare [161]. Lack of physical interaction and increased distress resulted in people sitting at home, spending more time on social media to maintain relationships, get information/support during the lockdown [182, 90]. People use social media to share their personal stories [137], look for information on COVID-19 [28], and seek support from others during this challenging time in their lives [21].

Many people turn to online communities to seek social support [27]. Previous studies have shown that social support in online communities can help people feel better [48, 15, 192, 152] and positively affect a user's mental health [166]. It has helped users in battling drug addiction [117], dealing with cancer [197], losing weight [32], and curbing depression [40]. Analyzing the kind of support people seek, kind of support they receive, and how it affects a user's behavior can be instrumental for community moderators and platform designers.

Benefits provided by an online community are likely to be more accessible to people who stay longer [197]. To study the extent of online communities' role in providing support, we need to understand what influences a user's decision to participate longer in the community. Analysis of user longevity can give us valuable insights into the dynamics of online social support.

Previous works have analyzed people's sentiments during COVID-19 [201]. Han et al. [21], studied public opinion, while another study [201] used topic modeling techniques to identify discussion topics and analyze emotions. However, more work needs to be done in analyzing dynamics of community support on social media during COVID-19. Moreover, not much work has tried to study user commitment and its effects in a COVID-19 based online community.

We address this gap in our this chapter by doing a coherent study of the social support community subreddit named *COVID-19positive* on Reddit. We study two popular support categories - emotional and informational support. We examine these support classes in two dimensions - user behavior and temporal phases. Using survival analysis, we then study the support factors that influence a user's longevity in these communities, precisely what compels a user to stay in the community even after recovering.

We discover that (1) In a COVID-19 community, emotional support involves discussing recovery, the status of family and loved ones. Emotions such as gratitude, prayer, and hope are expressed. Informational support involves discussion around research, infections, finance, and tests. (2) People who stay longer seek more informational and emotional support from the community. They also (3) give more support. Surprisingly, (4) the amount of support a user receives from the community is independent of the user's decision to stay. Furthermore, factors like talking about symptoms and recovery and interacting with more users in the community promote a longer stay. Through our work, we make the following contributions:

- Investigate support dynamics in a COVID-19-based online community.
- Characterize the factors influencing a user's longevity in the community.

5.2 Related Work

Social isolation is associated with increased morbidity and mortality in a host of medical illnesses [16]. Online healthcare communities have extensively studied support for diseases such as cancer and depression. Emotional support in a community can help build relationships, improving their commitment to the group. Wang, Kraut, and Levine [197] conducted a survival analysis to predict how emotional or informational support exposure affects the length of subsequent participation and user commitment. Emotional support was positively associated with how long members remained in the group.

Numerous studies have explored various aspects of social media and COVID-19. For instance, [149] conducted a sentiment analysis of Twitter data to examine changes in public sentiment overtime during the COVID-19 outbreak. The study found a significant increase in negative sentiment during the initial outbreak of COVID-19. Similarly, a study by [206] examined public opinion on COVID-19 in China and found that various factors, including government policies, media coverage, and social media, influenced public opinion. Additionally, studies have shown that social media can be a platform for spreading misinformation about COVID-19 [161].

Despite the importance of social support in health-related communities, the dynamics of support have yet to be explored in detail for COVID-19 communities. Li et al. [109] examined the association between social support and mental health in COVID-19 patients but did not specifically focus on support dynamics in online communities. In contrast, our study aims to fill this gap by examining the support dynamics of users in COVID-19 communities, focusing on emotional and informational support before getting tested positive, during the quarantine, and after recovering.

Existing literature has highlighted the critical role of social support in online health communities. For instance, [208] conducted a study on loneliness in online health communities and found it a prevalent problem. However, social support can help mitigate the adverse effects of loneliness. Similarly, [200] found that social support can buffer the negative impact of COVID-19 related stressors on mental health outcomes. In addition, previous research has shown that understanding users' trajectories in online communities can provide valuable insights into how individuals engage with and benefit from online communities. In their recent research, [205] examined user trajectories in online health communities and found that participation levels can vary significantly. This finding highlights the importance of understanding factors contributing to diverging user trajectories.

In summary, our study aims to contribute to the growing body of literature on social media and COVID-19 by examining the support dynamics of user communities. We build upon existing research on social support in online health communities and aim to provide insights into how online communities can be leveraged to improve mental health outcomes during the pandemic. Additionally, by exploring the trajectories of users in these communities, we

hope to understand how seeking/giving/receiving support can affect users longevity in the communities.

5.3 Dataset

We focus on Reddit as our social media platform. The content on Reddit is organized in communities by topics of interest called subreddits. For our study, we look at the */r/COVID19positive* subreddit[1] where people ask questions and share their stories and experiences around the COVID-19 pandemic. The data used in our analysis were collected using the Pushshift API[2] and PRAW (Python Reddit API Wrapper).[3] Choosing */r/COVID19positive* as the subreddit to study has the following advantages:

- People ask questions, share experiences, and gain information from others on how to cope with the disease, making it a rich source of data for studying support.
- The data is classified using flairs. These flairs allow us to study the data in a structured manner. Each submission can be assigned to a predefined category that the admin of the subreddit has defined. Some popular flairs in */r/COVID19positive* subreddit are Tested Positive - θ where θ can be Me, Family, Friends, etc., Question-to those who tested positive, and Question for medical research.
- The subreddit follows strict guidelines, and the community is well-moderated. Hence there is less possibility of falsely labelled data.

We collected posts, comments, and metadata like usernames, timestamps, and scores. We obtained a total of 93,576 posts and 9,93,030 comments. This data was generated by 104,818 unique active users, of which 37,762 (36.03 %) wrote at least one post and 94,469 (90.13 %) wrote at least one comment. Table 5.1 shows basic statistics of */r/COVID19positive* dataset.

5.4 Data Classification

Researchers have investigated online social support in a variety of ways. Social support has been conceptualized earlier either in terms of its functional content (the division of support into different categories like emotional and informational), being active (giving), or being passive (receiving) [5]. Some studies also analyze a third behavioral category which is seeking support [69, 55]. Based on these, we study support along the two dimensions, functional content, behavioral aspect and add another dimension, i.e., temporal classification. In the context of COVID-19, a user's timeline can be divided into three phases: before the user tests positive, during the 15 days of quarantine in which a user is positive, and after the

[1] https://www.reddit.com/r/COVID19positive/
[2] https://github.com/pushshift/api
[3] https://pypi.org/project/praw/

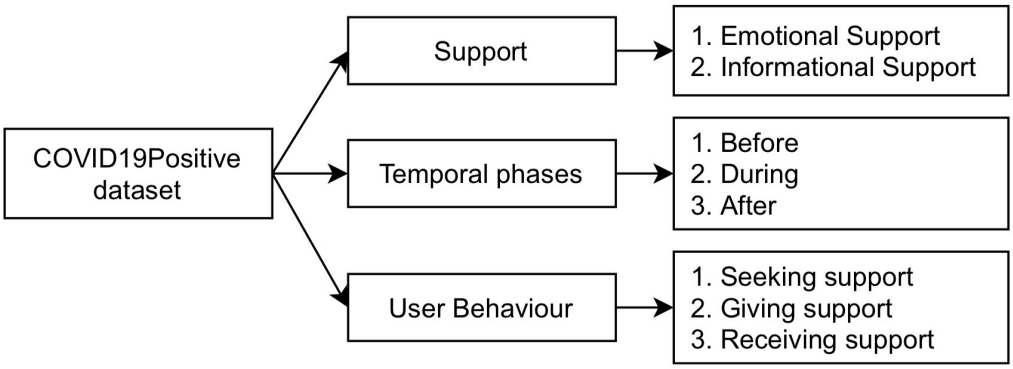


Figure 5.1: Data classification of */r/COVID19Positive* dataset. We divide data across three dimensions: Types of support, temporal phases in the context of COVID-19, and user behavior.

quarantine is over. Analyzing support in these three phases can give us more insights into the different support dynamics in each phase. Considering the above points, in the following sections, we classify data on three dimensions as seen in Figure 5.1 - (1) different categories of support - *emotional and informational support*; (2) the three user behaviors - *seeking, giving, and receiving support*; (3) the three temporal phases in the context of COVID-19 *before, during, and after phase*.

5.4.1 Social Support Categories

Many support categories have been identified [95, 159, 26], but two have been most talked about in online communities:

Emotional Support: Defined as having others sharing care, concern, sympathy, empathy, encouragement, and validation [13]. It can be crucial in the scenario of COVID-19. We define emotional support as posts with flair - Tested positive + θ and the comments received on such posts, where θ can be *Me, Family, Friends*. We define this group as emotional flair set in Table 5.1 because posts with these flairs often include people sharing their experiences with COVID-19 drawing other people's interest. People commenting on such posts also show concern, sympathy and offer condolences to those affected.

Informational Support: This is defined as sharing suggestions and information [13], which for COVID-19 includes information about symptoms, treatment, side effects, disease development, preventive measures, etc. This support is vital during COVID-19 because there was a lot of misleading information in the beginning of the pandemic. We define informational support to include posts with flairs - "Verified Research", "Question-to those who tested positive", "Question-for medical research".

	Flair	Posts	Comments
Emotional flair	TP	4,219 (4.51%)	46,302 (4.95%)
	TP - Me	26,975 (28.82%)	266,806 (28.53%)
	TP - Family	7,452 (7.96%)	85,243 (9.12%)
	TP - Friends	1,575 (1.68%)	15,588 (1.67%)
	TP - LongHauler	785 (0.84%)	6,860 (0.73%)
	TP - Unvaccinated	269 (0.29%)	3,237 (0.35%)
	TP - Breakthrough	1,034 (1.1%)	10,431 (1.11%)
Informational flair	Verified Research	183 (0.19%)	723 (0.07%)
	Question - to those who tested positive	25,439 (27.18%)	212,748 (22.75%)
	Question - for medical research	4,311 (4.61%)	38,055 (4.07%)

Table 5.1: Total posts and comments for different flairs on the r/COVID19positive dataset. TP = Tested Positive.

5.4.2 Behaviour

LaCoursiere [97] presented a holistic theory for online social support in health communities. She defined three main channels by way of which online social support can occur:

- Perceptual: When an individual's need for social support arises due to emotional factors such as stress.
- Cognitive: When an individual seeks information about particular medical entities like medication, symptoms, procedures, etc.
- Transactional: This is when an individual evaluates the social support they receive from the community.

This theory can be helpful in our social support analysis by defining users' various behaviors in the context of support. We define three kinds of user behavior on the subreddit:
Seeking Behaviour: This behavior can be described as someone asking for support on the subreddit. A user can seek two kinds of support - Emotional support and Informational support, as defined in the previous section. Emotional support seeking is defined as a user uploading a post with a flair from emotional flair set. Informational support seeking is defined as a user uploading a post with a flair from informational flair set. Emotional support seeking is a example of the perceptual channel of online social support, whereas, informational support seeking is an example of the cognitive channel of online social support.
Receiving Behaviour: This behavior can be described as the community's response to someone seeking support. We consider received support to be the comments on the posts

seeking that particular kind of support. Therefore, emotional support receiving are the comments on the posts seeking emotional support. Informational support receiving are the comments on the posts seeking informational support. Since we are considering the community's response here, we remove the comments made by a user on their own post. Receiving behavior is a direct consequence of the transactional channel of online social support. In their 2006 study, Moreland and Levine [133] analyzed the antecedents and consequences of individuals' group involvement. The authors put forth a group socialization model, according to which members assess the group's ability to fulfill their needs. They consider the group's past and potential future benefits during this evaluation. Members' level of commitment to the group is determined by the outcome of this assessment, which in turn influences their inclination to remain in the group and actively work towards collective goals. Primary determinant of whether the user's needs are being met is the support received from the community, which consequently influences the user's persistence in the group. This information can help understand the dynamics of group engagement and its impact on individual behavior.

Giving Behaviour: This behavior can be described as a user supporting others in the community by giving back. Preece and Shneiderman [186] found that people who receive support, start to reciprocate it back to other community members. We consider giving support to be comments made on other people's posts seeking that kind of support. Therefore, emotional support giving is the comments made on the posts seeking emotional support. Informational support giving is the comments made on the posts seeking informational support.

5.4.3 Phases

Let the time at which the user tested positive be t. Based on when a user tested positive, we can divide our data into the following three phases:

Before Phase: Any time spent by a user in this community before time t is before phase of the user.

During Phase: Time between t and $t+15$ days is the during phase. This is the period when a user is COVID-19 positive. This period coincides with the quarantine period in most countries before a user gets tested again.

After Phase: This is the time spent by a user in the community after $-t+15$ days.

How to decide when a user tests positive? We assume that a user tested positive the day they uploaded the first post using the indicative flair. We do this because there is no definite indicator in an online community to decide when a user tests positive.

5.5 Support Analysis

We analyze the differences between defined support classes in two steps. Firstly, we analyze the linguistic differences between support classes across phases using topic modeling and

odds of topics. This helps us understand the functional differences between support and behavioral classes. Then, we analyze support in phases this helps us understand how many users move in and out of the community. Next, we analyze which support classes influence a user's decision to stay in the community for longer.

5.5.1 Differentiating support classes

Topic Modelling: We use topic modeling to get a high-level overview of the content in support classes and observe the differences between them. This also acts as a validation for flair based classification. Using topic modeling, we represent a document as a collection of topics, giving an idea of what the document is about. We use a dense representation topic modeling framework [4] and a class-based TF-IDF to create dense clusters allowing for easily interpretable topics while keeping essential words in the topic descriptions. This algorithm starts with creating document embeddings from a set of documents useing Doc2Vec [104]. We cluster these embeddings together using HDBSCAN [125]. Since HDBSCAN is prone to the curse of dimensionality, we first reduce the dimensions using UMAP [125], which preserves local structure well, after which we can use HDBSCAN to cluster similar documents. We find cluster descriptors, i.e., words that describe each cluster the best using TF-IDF. We use class-based TF-IDF, which considers each cluster a single document and then applies the standard TF-IDF algorithm.

Odds of Topics: For the support classes, we found most frequent topics to be quite similar, which is understandable since all the classes are talking about COVID-19. To find discriminative topics for each class, we use the odds ratio, a statistical metric to measure the association between the presence of one property with another. We define Odds of Topics, where we find a topic for a class that has the least overlap with the topics of the other class. We consider only those topics with a frequency of occurrence greater than 2 to avoid scarce topics. The discriminative topics for each support class are given in Table 5.2. We see a clear distinction between the two support classes in seeking and giving behaviors.

	Support Seeking			
	Topic	**Words**	**Topic**	**Words**
Informational	Symptoms	smell, taste, body, temp, headache, breathe, shortness, asthma, insomnia	Nutritional and Lifestyle advice	appetite, eat, weight, exercise, take, taking, ivermectin, zinc, quercetin, paxlovid, vitamin, supplement
	Test	test, tested, positive, pulse, levels, oxygen, oximeter, day, vaccinated, antibody, blood	Other related topics	menstrual cycle, periods, urination, bladder, alcohol, smoke, smoking, dog, cats, pets

[4] BERTopic: Neural topic modeling with a class-based TF-IDF procedure - Maarten Grootendorst

Emotional	Sickness and Family	dad, mom, father, hospital, baby, son, kid, cough, daughter, husband, fever, toddler, family	Recovery	contagious, still, positive, quarantine, resting, test, tested, exercise, run, workout, walk, recovered, day longer, isolate
	Mental health and Anxiety	anxiety, feel, pain, fear, panic, scared, nausea, anyone, brain, fog, memory, focus	Personal and Health Concerns	fever, cough, taste, fatigue, breath, job, loss, pay, employer, manager, living
Support Giving				
Informational	Infection	swab, allergy, response, nose, itchy, immune, system, fever, breath, fatigue, shortness, headache	Wellness and Rest	chicken, soup, fruit, immunity, taking, b12, daily, salt, appetite, gargle, honey, workout electrolyte, hydrated, exercise
	Research and Facts	covid19, science, data, study, studies, evidence, research, scientific, scientist, theory	COVID-19 related topics	pcr, antigen, mask, air, n95, quarantine, omicron, delta, hair, variant, body, response
Emotional	Love and Care	love, hug, hugs, sending, virtual, glad, feel, sorry, loss, grief, supportive, healthcare	Coping Strategies	pfizer, moderna, shot, steroid, antibiotic, netflix, watch, binge, watching, game, podcasts, book, tv, show, green, tea, honey, ginger, lemon, water, cayenne, manuka
	Gratitude	please, thank, thanks you, so, much, contribution	Pray and Hope	wish, speedy, recovery, better, pray, sending, you, strong, hope, hopeful, crossed, fingers, miracles

Table 5.2: Topic Modelling results. Informational support seeking has topics consistent with asking for information- curiosity, help, details about infection and the testing process. In contrast, Emotional Support Seeking has content describing mild symptoms, recovery, sickness in the family and anxiety about health of family members. Information Support Giving provides information related to finance, infection, research and severe symptoms. On the other hand, Emotional Support Giving includes showing gratitude, love and hope, along with recommending rest. We see a clear distinction between the two support classes, in both the seeking and giving behaviours

5.5.2 Support in Phases

We study the number of users in each phase and intersection of users between phases. Out of all the users who tested positive, 49% of positive users became active on the subreddit in the *during phase*, which means the very first post they did on the subreddit was presumably about them testing positive. About 41% of people did some activity before they tested positive, and 30% continued to be active in *after phase*, with 21% of users being in all three phases. The Venn diagram in Figure 5.2 shows the exact distribution of 24,644 users who tested positive in each phase. We observe two subsets of users, one in *before and during phase* but never entered *after phase*. They never returned to do any activity (post or comment) in their *after phase*. Other subset is the users who were present in all three phases. This compels us to ask the question *What makes a user stay in the community?*

We analyzed the support types provided by each group and found that those who stayed in the community were likelier to seek and provide information and emotional support than those who did not. In the during phase, users who stayed received significantly more emotional support than those who did not.

Figure 5.3 illustrates the values for each support class for both subsets. Our analysis revealed that the average number of users who sought information in before phase was 0.18 for the "*Before During After*" group, while it was only 0.012 for the "*Before During NOT After*" group. Additionally, the average number of users who gave emotional and informational support in before phase and during phase were 2.28, 1.43, 3.11, and 1.72, respectively, for the "*Before During After*" group. The corresponding values for the "*Before During NOT After*" group were 0.08, 0.04, 0.1, and 0.05, respectively.

Moreover, the average number of users who received emotional support in the during phase was much higher for the "*Before During After*" group (20.27) compared to the "*Before During NOT After*" group (1.94). Also, emotional support seeking and receiving was not defined for the before phase. In summary, our findings suggest that users who stay in the community are more engaged and active in seeking and providing support to others. We test the validity of this hypotheses by performing causal inference using survival analysis.

5.6 Survival analysis: Relationship between support and longevity

Survival analysis [131] is a statistical approach that assesses the likelihood and timing of an event's occurrence, and how various factors influence it. We use survival analysis to find whether there is a causation between how long a tested positive user participates in the group and the amount of support they seek, give, or receive. In our case, the event of interest is whether a user is active in the after phase. Our goal is to understand whether the amount of emotional or informational support a user gives, seeks, or receives has any role in increasing the length of participation. We also try to discover what other factors may be responsible for their behavior to remain active in later phase.

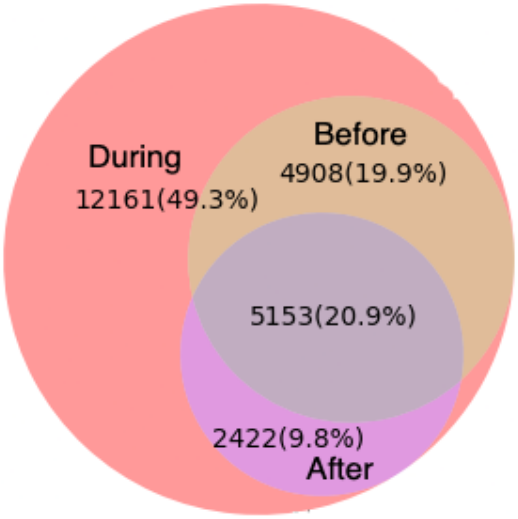


Figure 5.2: Distribution of users doing some activity (posting or commenting) in each of the phases. Note that we do not have people just in the before phase or the after phase because we define the phases with respect to the first post of a user made by a user using the flair in emotional flair set which is in the during phase. Any activity before this post lies in the before phase, any activity 15 days after this post is in the after phase

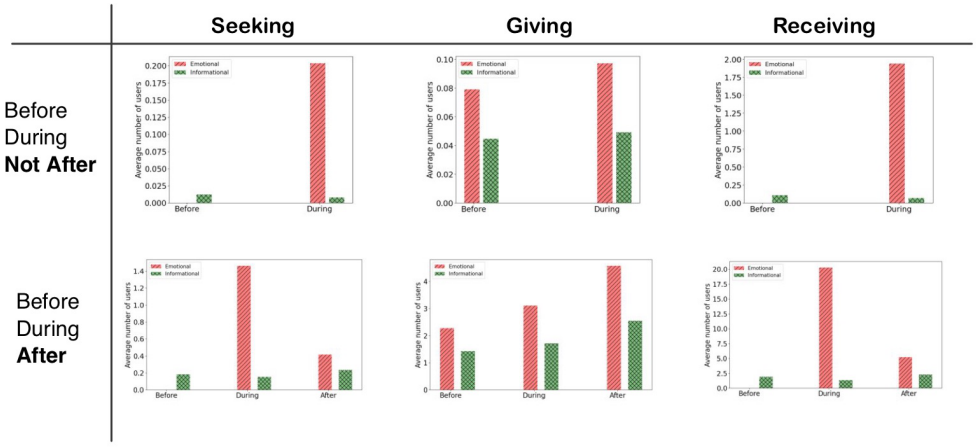


Figure 5.3: Support in phases. Users who stay tend to seek double the information support in before and during phases than those who don't. They give three to four times more support, both emotional and informational, in before and during phases. They also receive 1.6 times more emotional support in the during phase.

5.6.1 Data and Methods

To conduct the analysis, we included only users who contributed more than one post. Failure event is defined as the user's last login date. If a user does not return in the after phase, failure event is true.

Survival time of a user is defined in days as the time between their first activity on the subreddit and their last activity of interest. If a user returns in after phase (failure event is false), survival time is the number of days between the first post in after phase and the first post ever created. If a user does not return in after phase (failure event is true), survival time is the number of days between the last post in during phase and the first post created.

Covariate	**coef**	**HR**	**se(coef)**
Before Info Seeking***	0.08	**1.09**	0.02
During Emo Seeking***	-0.07	**0.93**	0.02
During Info Seeking***	-0.07	**0.94**	0.04
Before Emo Giving	0.12	1.13	0.01
Before Info Giving***	0.01	**1.01**	0.01
During Emo Giving	0.14	1.15	0.01
During Info Giving	-0.02	0.98	0.01
Before Info Receiving***	0.0	1.0	0.0
During Info Receiving	0.0	1.0	0.0
During Emo Receiving***	0.0	1.0	0.0
Num Self Comments***	-0.02	**0.98**	0.0
Outdegree	-0.14	0.87	0.01
Degree***	-0.02	**0.98**	0.0
Avg Post Length***	0.0	1.0	0.0
Avg Comment Length***	0.0	1.0	0.0
Avg Post Time Diff***	0.0	1.0	0.0
Topic Post Family***	-0.13	**0.88**	0.02
Topic Post Symptom***	-0.24	**0.79**	0.01
Topic Comment Gratitude***	0.01	**1.01**	0.01
Topic Comment Recovery***	0.0	1.0	0.01

Table 5.3: Results from survival analysis. The covariates marked $***$ have a significant positive effect ($p < 0.005$) on survival.

Because people who logged in close to the end of data collection might still be participating, we considered those who last logged in within 15 days of data collection as right censored.

5.6.2 Cox Regression

Cox proportional hazards regression [30] is a method for investigating the effect of covariates on time a specified event takes to happen. In such cases, the conditional survival function is calculated

$$S(t|x) = P(T > t|x) \tag{5.1}$$

Where x denotes the covariates and t denotes the time till the event of interest occurs. Cox proportional hazard is represented as:

$$h(t) = h_0(t)^{b_1x_1+....+b_kx_k} \tag{5.2}$$

where,
$h(t)$: hazard at time t
$h_0(t)$: hazard for a person with value of 0 for all independent variables
b: regression coefficient for independent variable x
x_i: independent variables
For our model the survival event is whether a user is active on the community in their after phase. We use Cox Regression to see what are the factors that affect this survival.

5.6.3 Covariates

Support seeking/giving/receiving is divided into before and during phases of the user. There is no emotion seeking in the before phase due to our definition: The first post with "Tested positive" (emotional support) flair starts the during phase.

- `Emotional support seeked`: Total emotional seeking posts made in the corresponding phase.
- `Informational support seeked`: Total informational seeking posts made in the corresponding phase.
- `Emotional support given`: Total emotional giving comments made in the corresponding phase.
- `Informational support given`: Total informational giving comments made in the corresponding phase.
- `Emotional support received`: Total emotional comments received in the corresponding phase.
- `Informational support received`: Total informational comments received in the corresponding phase.

Total support received is the summation of all the comments received by that user. We also used covariates found from the topic modeling like number of posts with topic `family`, `symptoms`, `gratitude`, and `recovery`. Other covariates considered were `Num Self comments`, number of unique people a user interacts with (`Degree`), `Avg Post Length`, `Avg Comment Length`, and average time interval between consecutive posts.

5.6.4 Analysis

Table 5.3 shows the results obtained from the Cox Proportional Hazard model. The *coef* column represents the covariate's effect on a user's survival. If the coef is negative, hazard is less; therefore, the particular covariate will positively affect survival. This means that if we increase the value of this parameter, survival will increase and vice-versa.

The Hazard Ratio (HR) is the effect of an explanatory variable on risk or probability of participants' leaving the group. The HR value for Before Info Seeking is 1.09, indicating that users' survival in the group is 9% (100-(100*1.09)) lesser for those seeking information support than those not. Similarly, people seeking other forms of support during emotional seeking (7% more survival) and informational seeking (6% more survival) are more likely to remain in the group. Hence more support-seeking behavior in *during phase* is associated with more probability of staying, and surprisingly, more support-seeking behavior in *before phase* is associated with a lesser probability of staying in the community.

For support-giving behavior, before informational-giving (1% less survival), people are less likely to stay. Hence, people giving more informational support are less likely to stay. Also, before emotional giving, during emotional giving, and informational giving, behavior is not significantly associated with the user's probability of staying in the community. However, contrary to our assumption, support-receiving behavior did not significantly affect the user's stay in the community.

Other factors that promote users staying are if a user posts about their symptoms (21% more survival) or posts about their family (12% more survival). Also, users with more self-comments (2% more survival) and users who interact more with other users (2% more comments) are more likely to stay in the community.

5.7 Discussion

Our work aims to analyze online COVID-19 communities for support dynamics and factors affecting the longevity of the users. We collect data from */r/COVID19positive* subreddit. First, we classify all the activities into different types and phases of support like Information or Emotional giving/seeking/receiving. We also divide a user timeline into before being contacted with COVID-19, during, and after the recovery.

We see that higher the support seeking, higher the probability of survival. Contrary to common belief, support receiving volumes did not significantly affect a user's stay in the community. Our results leads to a better understanding of user dynamics related to

community support and can directly impact moderators and platform owners in designing community guidelines and incentive structures.

5.8 Limitations

Our work provides direct insights into the support dynamics of COVID-19 communities, which can assist moderators and platform designers in setting guidelines and incentive structures. However, our analysis also has some cavities which should be accounted for while building upon our work. Firstly, our definition of during phase is based on when the users post about it. We can't be sure if the user tested positive the same day or earlier. Further, our analysis does not consider that a user might get tested positive multiple times during their stay on the subreddit since data available for such cases was limited.

Part III

Individual-Organization Interactions

Chapter 6

Social Re-Identification Assisted RTO Detection for E-Commerce

E-commerce features like easy cancellations, returns, and refunds can be exploited by bad actors or uninformed customers, leading to revenue loss for organization. One such problem faced by e-commerce platforms is Return To Origin (RTO), where the user cancels an order while it is in transit for delivery. In such a scenario platform faces logistics and opportunity costs. Traditionally, models trained on historical trends are used to predict the propensity of an order becoming RTO. Sociology literature has highlighted clear correlations between socio-economic indicators and users' tendency to exploit systems to gain financial advantage. Social media profiles have information about location, education, and profession which have been shown to be an estimator of socio-economic condition. We believe combining social media data with e-commerce information can lead to improvements in a variety of tasks like RTO, recommendation, fraud detection, and credit modeling. In our proposed system, we find the public social profile of an e-commerce user and extract socio-economic features. Internal data fused with extracted social features are used to train a RTO order detection model. Our system demonstrates a performance improvement in RTO detection of 3.1% and 19.9% on precision and recall, respectively. Our system directly impacts the bottom line revenue and shows the applicability of social re-identification in e-commerce.

This chapter is partly a reproduction of paper published at the ACM Web Conference (WWW) 2023 [76].

6.1 Introduction

Over the last decade, e-commerce adoption has proliferated rapidly [143]. Such growth is fueled by convenience that e-commerce can provide over brick and mortar, e.g., large product selection, lower prices, same-day shipping, and hassle-free returns and cancellations. Though convenience features attract customers, they can sometimes cause significant business challenges; one such case is Return-to-Origin (RTO). RTO as depicted in Figure 6.1 is a scenario when a customer orders a product and then cancels while it is en route. RTO leads to two kinds of losses in a system:-

- **Logistical cost**: This is the cost of shipping the product till the point of cancellation in the supply chain and then returning it to the warehouse safely and restocking it.
- **Opportunity cost**: In the time while the product was ordered and canceled, this product unit became unavailable to order by another customer who would accept the delivery.

Though business accounts for potential revenue loss while offering functionality like RTO, an increased rate of RTO by uninformed customers or bad actors can cause unanticipated revenue losses totaling double-digit million dollars annually. Hence it becomes necessary to develop a real-time system that can predict the likelihood of the order being subjected to RTO at the time of checkout. Prediction of the model combined with other attributes like customer history, and available stock of the product can be used to initiate precautionary measures that can mitigate RTO risk. Naturally, the data used to build such a system would be, the historical pattern of RTOs at a user and product level. However, a system built on these features is limited in its capability, especially for new users and product categories.

Literature has shown that socio-economic attributes of customer can be an indicator to identify the likelihood of a person being involved in activities like electricity theft [150], false insurance claims [183], or mortgage fraud [22]. Public social media profiles can be used to estimate socio-economic features [101]. Adding features from social media profiles has shown improved results in a variety of tasks, e.g., identifying transaction fraud [75], the credibility of online information [63], hate speech [34], and propensity to participate in risk-taking activities [100, 105]. Grounded in the aforementioned literature, we hypothesize that enriching historical data with publicly available social data of a consumer will lead to a performance improvement in RTO prediction.

The first step for our experiments is to re-identify the social profiles of a given user. The problem of social re-identification is studied widely [74, 202, 171, 132, 134]. Though most literature relates to retrieving matches between two social media platforms with a notable

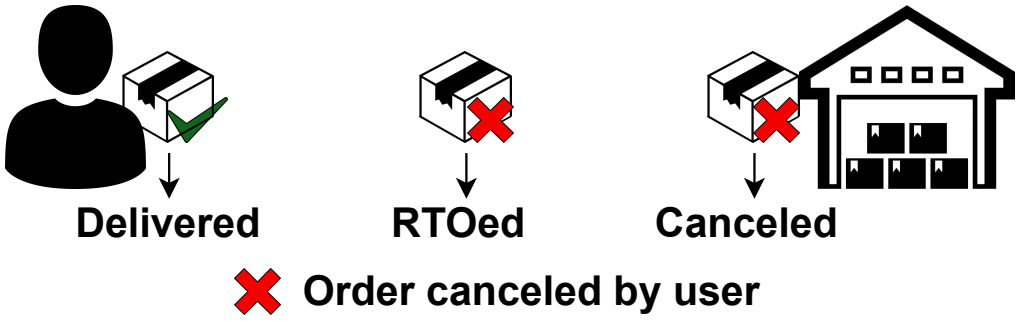


Figure 6.1: An order becomes Return to Origin (RTO) when the user cancels an order after it has been shipped from the source location.

exception of [60], our task is slightly different, where we need to match profiles between a social and an e-commerce platform.

In this chapter, given an e-commerce user, we find the relevant public social profile and show that the fusion of social information with historical trend data improves the performance of RTO prediction by 3.1% increase in precision and 19.9% increase in recall. Our work has direct implications for e-commerce platforms where a system like this can prevent loss of revenue. Additionally, our study demonstrates that combining social information with internal platform data can be a valuable tool for improving downstream tasks like RTO.

6.2 Data and Social Re-Identification

In this section, we first provide the details of our ground truth RTO dataset, followed by social re-identification candidate extraction (§ 6.2.2) and validation (§ 6.2.3) steps.

6.2.1 Ground Truth Data

We can extract ground truth from all past orders and their subsequent outcomes of the e-commerce platform. Orders are subjected to multiple internal models during checkout, which can induce unintended biases in the data. To prevent this, 5% of all orders are randomly set aside, as the *control set*, where no intervention is applied. Further, we extracted the cash on delivery orders from the control set, because we observed that orders with cash on delivery are more prone to RTO. All our experiments and benchmarking are performed against this set. Our experiments are performed on 6 months (November 2021 - April 2022) of data. First 5 months of data is used for training, and the following 1 month is used as a test set.

We ensure that our study design does not breach the privacy terms and conditions of the platform, or of the social media platforms used. As an extra layer of prevention, experiments shown in this work are performed only on users who explicitly decided to make their name and city locations[1] public on the platform. After all filtration, our final dataset includes 6,881 orders placed by 2,121 unique users. Out of all, 2,201 (32%) orders were RTOed.

6.2.2 Potential Candidate Extraction

The initial step of user re-identification is to reduce the infinite search space of social profiles to a few candidate profiles for a given user. Querying social media platform's search engine using the *name* and *location* of a user has been shown to narrow the candidate pool effectively [74, 60]. For every unique user in our dataset, we create a search query of format <*user name*> <*city name*> and retrieve results from the social platform's search engines and a leading web search engine. Top 10 results of the query are used as candidate profiles.

We use a popular professional networking social media platform as a source of our social data; since, along with general information, such platforms have specific information that

[1] Used for social re-identification, see § 6.2.2

can reflect socio-economic indicators. Only data explicitly made public on the platform by the user is collected and used. Out of total 2,121 unique users, we found potential candidate profiles for 1,091 users.

6.2.3 Social Re-Identification

Literature shows that different social profile attributes like name, location, network, and language features can be used to find a match from candidate profiles [171]. Considering the asymmetry between e-commerce and social media platforms (like unavailability of a connection among user profiles), all these attributes are not available on both the platforms. However, we are in a unique position to access various locations a user has ordered from in the past. [60, 156, 195, 57] showed that matching various location information in a user's profiles with candidate profiles can find correct matches with a high probability.

We perform candidate filtration using two attributes viz. names and locations. Firstly, any candidate profiles whose names do not match the source user are rejected. In the second step, given a source user u, we extract from the orders history a set L_u, defined as $\{l_u^1, l_u^2, ..., l_u^n\}$ where l_u^i is the i^{th} city u placed a order at. For each potential candidate profile of u, a similar location set L_u^c is defined as $\{l_c^1, l_c^2, ..., l_c^m\}$ where c denotes a candidate profile and l_c^i is a city location mentioned in c's social profile.

The Match score of candidate profile c with u (α_u^c) is defined as the ratio of location in social profiles also present in the source user location set. While calculating the intersection between the set of city names fuzzy matching was used to account for slight variation in spellings and syntax of city names. E.g., Delhi vs. New Delhi, or Bangalore vs. Bengaluru.

$$\alpha_u^c = \frac{|L_u^c \cap L_u|}{|L_u^c|} \tag{6.1}$$

A candidate profile is considered a match if α_u^c is above a predefined threshold θ. A user can be classified into three categories based on the number of matches received. 'No match' for users where no candidate profiles had a score above θ, an 'Exact match' where exactly 1 candidate profile had the matching score above θ, and 'Multiple matches' in which case we found more than one candidate profile who had match score above the threshold. Table 6.1 shows the percentage of users in each of three categories for different values of θ. Users in the 'No match' category were removed from the modeling step. In case of 'Multiple match', final feature value is obtained by averaging over all the matches. Results shown in this chapter are calculated using $\theta = 0.6$, results for varying values of θ were consistent and are omitted due to lack of space.

6.3 RTO Model

We discuss the features used by our proposed model, the types of modeling techniques we experimented with, and the evaluation metrics used.

6.3.1 Features

We broadly divide the features used into three categories; 1) past trends, 2) social profile quantitative, and 3) social profile abstractive. The first category is derived from internal data, and the other two are extracted from social profiles.

Past trends: These features are derived from historical data. Each sample includes the ratio of RTO vs. total orders over the last 3 months and 1 year for the user, products in order, seller, and product category. Apart from this, location is also a robust socio-economic indicator; therefore, we extract the same trends for pin code, street, and city mentioned in the delivery address. Additionally, we noticed a correlation between the RTO rate and the order time, specifically the hour and weekday. Therefore, the hour of the day, weekday, and respective past trends are incorporated into the feature list.

Social profile quantitative: As we identify social profiles for a user, we extract if the user is a student, number of jobs, number of educational degrees, and number of friends and followers. The count of jobs/degrees may not always be a good indicator of someone's professional position since some people spend a long time in the same jobs, whereas others often switch jobs. Pertaining to that, we add two features counting the total years a user has spent working and in education.

Social profile abstractive: We have extracted social features related to the quantity of experience and education of users. Research has shown that institutions of education and programs studied can significantly impact career success [157]. Similarly, two people with the same years of job experience can have widely different buying propensities based on what roles they are pursuing at which organizations. We hypothesize features capturing user's education institutes and job roles can assist in RTO prediction. Recently, contextual language models pretrained on large volumes of data, have captured and exploited complex relations

Table 6.1: Results of social re-identification for varying values of matching threshold θ.

Match Threshold θ	Exact Match	Multiple Match	No Match
0.1	81.49	18.51	0.00
0.2	81.31	18.51	0.18
0.3	79.58	18.51	1.91
0.4	76.21	18.41	5.38
0.5	71.01	18.41	10.57
0.6	68.92	17.68	13.40
0.7	65.91	17.50	16.59
0.8	64.63	17.41	17.96
0.9	64.36	17.41	18.23
1.0	42.57	17.41	40.02

well for downstream tasks [45, 19]. Following this, we extract the latest education institute, and the course pursued by a given user and pass this textual information via a pretrained Sentence-BERT [153] model to generate 387 dimension vectors. A similar vector is also created for the Job organization and designation the user had while placing the order.

6.3.2 ML Modeling

Most of our data is tabular making tree-based ensemble methods like Random forest and XGBoost the default choice. Recently, attention-based architecture like Tabnet [6] has been proposed claiming to outperform traditional tree-based models. We present results on both types of models.

Figure 6.2 shows our training setup. In the tree-based models, 387 dimension vector obtained for job and education are decomposed to lower dimensions using UMAP [126] to prevent overfitting. The final dimension after decomposition is treated as a hyperparameter. Finally, decomposed vectors are added to the table of quantitative features as columns and fed into the model. When experimenting with deep learning-based models, tabular features are passed through Tabnet to generate a feature embedding. Generated embedding is concatenated with sentence-BERT embeddings (see § 6.3.1) and passed into a series of fully connected layers. All models are hyperparameter tuned using random search over; 4-fold cross-validation over the training data is used for parameter selection.

6.3.3 Evaluation

We use precision and recall to evaluate the performance of our models, but at a large scale, even very small improvements in model performance can lead to measurable revenue benefits. Additionally, traditional metrics may not always fit well in business discourse. Highlighting this, we define a metric named *Goodness* on which our models are evaluated.

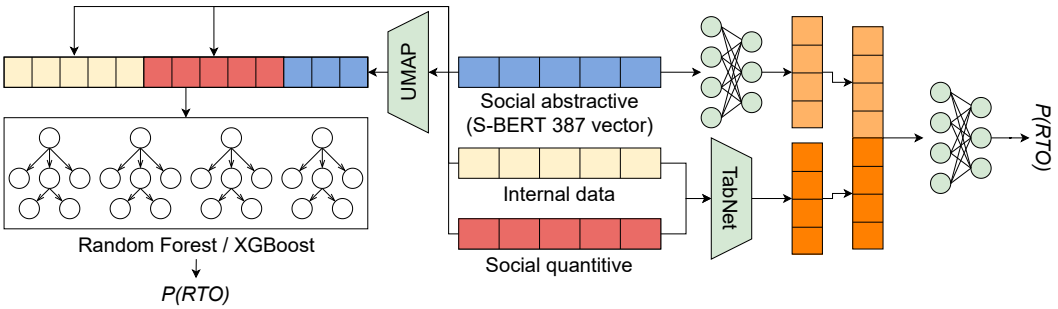


Figure 6.2: Our training architecture. In the case of tree-based models (on the left), all three feature sets are concatenated to form the input. While training deep learning models (on the right), tabular features are encoded via Tabnet and concatenated with S-BERT embeddings before being passed into a feed-forward neural network.

Goodness : It reflects the improvement in recall performance. Defined in Equation 6.2, it calculates the reduction in the ratio of RTO orders after being evaluated by the model. Multiplication with 10^4 is performed to convert value into Basis Points (bps), this improves readability even while observing quantitatively small improvements. A higher value is better.

$$Goodness = \left(\frac{|P|}{|P|+|N|} - \frac{|P| - |P_{Pred\ and\ True}|}{|P|+|N|-|P_{Pred}|} \right) \times 10^4 \tag{6.2}$$

$$FPR = \frac{|P_{Pred}| - |P_{Pred\ and\ True}|}{|P_{Pred}|} \tag{6.3}$$

Here, P is set of RTO orders, and N is set of Delivered orders. P_{Pred} is set of orders predicted as RTO by a model, and $P_{Pred\ and\ True}$ is set of true positive RTO predictions.

Our aim is to choose a classification threshold that maximizes *Goodness* while maintaining the false positive rate (FPR) below a fixed value. [2] A high FPR means increased false interventions, reducing customer experience. Just like precision and recall, *Goodness* and FPR are a trade-off balance. High *Goodness* comes with an increase in FPR.

6.4 Results

Table 6.2 shows performance of various RTO models on our test set. The random forest provides the overall best performance. As hypothesized, adding social features with past trends improves goodness by 300 bps, and adding contextual embeddings representing education and professional information improves the goodness further by 328 bps. This model has direct implications for improving the bottom-line revenue performance of an e-commerce organization.

Contrary to intuition, deep learning based models performed the worse. Comparative studies has shown that this behaviour is common in case of tabular data [61, 82, 172]. Studies compared the performance of Tabnet, and its contemporaries on a large variety of tabular data tasks, and concluded that these neural architectures do not perform consistently and are very sensitive to parameter tuning.

6.5 Conclusion and Future Work

Our study aims to improve the performance of a critical e-commerce problem RTO, where a user places an order and then cancels while the product is in transit, leading to logistics and opportunity cost. We hypothesize that fusing a users' social data with past RTO trend data can lead to improvements in performance. Towards this, we build a system to extract social profiles from popular professional networking social media platforms for a given user. Location-based matching is used to filter from the candidate matches. Finally, we extract quantitative and contextual features of matched profiles and demonstrate improvements of

[2] FPR threshold is decided based on product requirement.

Table 6.2: RTO detection performance on the test set. Random forest performs the best. The addition of social features with past trend data increases goodness by 628 bps.

Model	Features	Precision (%)	Recall (%)	Goodness (bps)
Random Forest	Past Trends	85.7	40.3	1,005.7
	Past Trends + Social quantitative	85.7	50.4	1,305.6
	Past Trends + Social quantitative + Social abstractive	**88.8**	**60.2**	**1,633.7**
XGBoost	Past Trends	80.0	33.6	809.3
	Past Trends + Social quantitative	82.2	39.7	994.1
	Past Trends + Social quantitative + Social abstractive	86.8	44.5	1,129.4
TabNet	Past Trends	82.4	39.4	977.0
	Past Trends + Social quantitative	78.2	30.2	716.2
	Past Trends + Social quantitative + Social abstractive	64.2	15.1	320.0

3.1%, and 19.9% precision and recall, respectively, in the RTO detection task. Our work has direct implications for improving the bottom-line revenue of an e-commerce organization. Potential future directions of our work can be to experiment with transfer learning or multitask setup to see if social re-identification can help in other facets of e-commerce experience like review credibility or credit modeling. We would also like to extend our experiments to include data from a broader type of social media platforms.

Part IV

Individual centric Interactions

Chapter 7

Put Your Money Where Your Mouth Is: Dataset and Analysis of Real World Habit Building Attempts

The pursuit of habit building is challenging, and most people struggle with it. Research on successful habit formation is mainly based on small human trials focusing on the same habit for all the participants as conducting long-term heterogonous habit studies can be logistically expensive. With the advent of self-help, there has been an increase in online communities and applications that are centered around habit building and logging. Habit building applications can provide large-scale data on real-world habit building attempts and unveil the commonalities among successful ones. We collect public data on *stickk.com*,[a] which allows users to track progress on habit building attempts called commitments. A commitment can have an external referee, regular check-ins about the progress, and a monetary stake in case of failure. Our data consists of 742,923 users and 397,456 commitments. Rooted in theories like *Fresh Start Effect*, *Accountablity*, and *Loss Aversion*, we ask questions about how commitment properties like start date, external accountability, monitory stake, and pursuing multiple habits together affects the odds of success. We found that people tend to start habits on temporal landmarks, but that does not affect the probability of their success. Practices like accountability and stakes are not often used but are strong determents of success. Commitments of 6 to 8 weeks in length, weekly reporting with an external referee, and a monetary amount at stake tend to be most successful. Finally, around 40% of all commitments are attempted simultaneously with other goals. Simultaneous attempts of pursuing commitments may fail early, but if pursued through the initial phase, they are statistically more successful than building one habit at a time.

[a] https://www.stickk.com/

This chapter is partly a reproduction of paper published at the AAAI International Conference on Web and Social Media (ICWSM) 2024 [79].

7.1 Introduction

In their endeavor to quit smoking, 92.5% of individuals experience failure [31]. Only 20% are able to lose weight [4], and 9% can stick to their new year's resolutions [142]. Building new habits is hard; most people struggle with it.

Recently we have also witnessed a rise in communities and applications centered on helping people in pursuit of habit formations. Reddit communities like *r/loseit*[1] and *r/stopsmoking*.[2] Applications like *Streaks,*[3] *Strides,*[4] *and Way of Life*[5] help users create goals and track progress. Some advanced applications like *StickK, Habitica,*[6] *Habitshare*[7] leverage previous research to build features about accountability, gamification, and incentives which can improve users' propensity to succeed in their goals.

Past research related to habit-forming can be widely divided into two parts: i) Sociology theories relating to behavior change like Operant conditioning [176], loss aversion [83], and fresh start effect [35]; ii) application/effect of these theories in a specific scenario. Skarupski at al. [174] used peer accountability to improve writing habits. Gine et al. [58] showed the effectiveness of loss aversion in quitting smoking, and [35] demonstrated that the commitment towards goals, increases when they are started on a new week or month. Habit formation has also been of interest in the computational social science community, with research evaluating the effect of online communities on various habits like weight loss [32], physical activity [3], smoking and drinking relapses [180], drug consumption [78], quality of user-generated content [25] and involvement in open-source projects [189].

Though current literature explores various habits, none evaluates large-scale heterogeneous attempts of habit building, unveiling the prevalence and effectiveness of guidelines suggested by social science literature. Historically, studying habits via human trials had a high logistic and monetary cost. However, the advent of habit-tracking applications provides us with data on real-world habit building attempts.

In this study, we collect publicly available data about users' past attempts at habit-forming on *stickk.com*. Grounded in theories of loss aversion [83], fresh start effect [35], and accountability, we perform characterization on temporal patterns (when commitments starts), incentive structures (type and amount of stake), and reporting habits (frequency and length of reporting). Further, we use an unsupervised retrieval method based on Word2Vec [130] and relevance feedback [168] to classify commitments into different categories. Finally, we use survival regression to analyze the effect of these confounders, like start date, stake, length of commitment, and category, on the commitment's success rate.

[1] https://www.reddit.com/r/loseit/
[2] https://www.reddit.com/r/stopsmoking/
[3] https://streaksapp.com/
[4] https://www.stridesapp.com/
[5] https://wayoflifeapp.com/
[6] https://habitica.com/static/home
[7] https://habitshareapp.com/

We discover 1) the fresh start effect is very prevalent. Users are 40% more likely to start a commitment on 1^{st} of a month, or Monday, and compared to an average day in the year, four times more commitments are started on New Year's. Though commitments started on New Year's are more likely to fail, commitments started on other temporal landmarks do not affect the likelihood of success; 2) Only 29% and 19% of total commitments have monitor stakes and external accountability attached respectively, but doing so significantly increases the success rate; 3) Success rate increases if a stake is given to an anti-charity instead of a charity on failure; 4) A critical factor in increasing success rate is to keep short-term commitments (7 weeks is the base hazard) and do frequent check-ins on the application; 5) Users pursuing multiple commitments simultaneously are more likely to succeed than users with singular goals.
In summary, our main contributions are:

1. To quantify the prevalence and patterns of parameters like start date, accountability, monetary stakes in habit building attempts.
2. To quantify the effect parameters mentioned above have on users' success rate.
3. A large-scale heterogeneous dataset of habit building attempts with detailed information about associated parameters and success rates.

Our work impacts researchers, users, and platform owners by providing a fertile base for developing future research or tools. Our dataset can be used to evaluate the effect of more complex factors like the types of habit overlaps or streaks on commitment success. Our results provide users with actionable insights such as not waiting for key dates, using external referees, and monitory stakes for their future habit-forming attempts, setting themselves up for a higher success rate. Platform owners can use our results to create new features or timed interventions that prevent users from derailing their goals.

Data and Code: Dataset, and code is available at *https://precog.iiit.ac.in/research/put-your-money/.*

7.2 Theories and Research Questions

The *Fresh Start Effect* [35] is defined as the human tendency to take action towards your goals starting a specific key date. Dates that stand out as more meaningful, such as the new week or month, birthday, or a holiday, signal the start of a distinct period. These "temporal landmarks" make it easier for people to mentally separate from their past imperfections and failures. Dai at al. [36] showed that the searches related to dieting, visits to the gym, and self-reported motivation towards the goals increases after temporal landmarks. Another study by [2] showed that the rate of exercise, extramarital affairs, and suicide increased when adults approach a new decade in their age, i.e., ages 29, 39, 49. Authors claim that certain numerical ages show a greater sense of self-reflection than others. Thus we expect users would be more likely to start new commitments on key dates and ask our first question:

RQ1. [Key Dates] *What is the prevalence of the fresh start effect in commitment start dates? Do commitments started on these dates have a higher rate of success?*

Best-selling books Tiny Habits [53] and Better than Before [162] claim accountability to be a key component in successful habit building. Past research has also shown a consensus with these claims. A 15-week study of 704 participants showed increased weight loss when paired with a support buddy [37]. Similarly, people working in an accountability group showed improvement in writing habits [174]. Renfree et al. [154] showed that check-in reminders in habit formation applications improve adherence, though they also create a dependency. Accountability can be of many kinds, like self (check-in to an application), external (buddy to validate your progress), and social (support community or social media announcements). Grounded in these, we ask:

RQ2. [Accountability] *What is the extent of different accountability methods (external and social) and their effect on commitment success?*

People are motivated or deterred from doing things based on the power of incentive or fear of loss, colloquially known as the method of "Carrots and Sticks" [8]. This effect is rooted in *Operant Conditioning*, stating the probability of acting in the future is a function of the stimuli received in the past [176]. Stimuli can be *appetitive* or *aversive*. Appetitive stimuli are those one voluntarily approaches (e.g. food treat), while aversive stimuli are those one try to avoid or escape (e.g. electric shock). Analyzing aversive stimuli, [83] found that the pain of losing is psychologically twice as powerful as the pleasure of gaining. This was termed as *Loss aversion*, and have shown wide applications ranging from designing insurance products [151], to help people abstain from smoking [58]. Methods of reward and punishment have been at the center of habit building recommendations. In the context of habit building, the most common method of incentive/loss is putting monetary amounts at stake to an entity in case of failure. Considering this, we ask:

RQ3. [Stake] *Does monetary stake affect the probability of success in a commitment? Does the nature of the entity with which money is staked affect the success rate?*

Guidelines on simultaneous habit forming are conflicting. On the one hand, it is suggested that an individual should focus on one goal at a time [38] and strive to reach the state of "automaticity" [98]. This means, initially, a new habit needs conscious effort, but after a while, it becomes an automatic routine, after which the person can move on to other commitments. On the other hand, concepts of habit stacking/anchoring [53] talk about linking a series of habits to one after the other. For example, working out makes you more likely to have a healthy meal. This is rooted in Behavioral Momentum Theory [139]. Once you are in a flow of doing things, momentum will carry you through the series of habits. To evaluate the effect simultaneity has on success, we ask:

RQ4. [Simultaneity] *What is the extent of the user trying to pursue simultaneous goals? How does it affect the success rate of a user?*

7.3 Related Work

Literature related to habit building is vast and has been a topic of interest in areas like sociology and behavioral economics. In §Theories and Research Questions, we discussed multiple habit building principles and associated literature. In this section we focus on studies in the field of computational social science that tried to measure the human behavior related to habit building.

A common theme has been the effect being part of a habit specific community can have. Positive community feedback has been shown to help with weight loss [32], and project contribution [25]. Althoff et al. [3] studied data from an exercise logging application and found that an increase in social connections on the platform led to higher physical activity. Being part of a community can also induce adverse effects like increased drug consumption [78] or deterioration in the quality of writing [25].

Commitment contracts are exercises where users put money toward a goal, which is returned only after successfully completing it. Such contracts have been shown to help people improve health [65], quit smoking [58], and help users reduce their carbon footprint [121]. Lee et al. utilized StickK data to analyze commitment properties and differences in commitments which are meant to start something new vs. to stop something the user is already doing [107]. Their annotation concluded that 82.3% of commitments are aimed towards starting something new. However, their analysis was limited to only 1,000 commitments.

7.4 Data Source and Description

We use data from the habit building platform *stickk.com*. We chose StickK because of the availability of large-scale public data, heterogeneous types of habits, and a diversity of features like allowing users to have referee and put money on stake. A user can have multiple habits called commitments on the platform. Commitments can be active, i.e., being pursued right now or completed. We only use the completed commitments in this study. Optionally, users can put money on stake for each commitment, which needs to be paid in case of failure. We use the entire historic dump of the platform to create a final dataset of 742,923 users who created 397,456 commitments with a collective $35.5 Million on stake. Table 7.1 shows our dataset statistics.

Each user profile has a unique numeric id, username, and joining date. Optionally, users can also add display pictures, location, interests, and a bio message. Users can decide to keep their profiles public or private. Out of the total platform users, 6.4% have designated their profiles as private. The unique numeric id allows us to identify the chronological order of

Total # of users	742,923
# of public users	655,750
# of private users	47,716
# of deleted users	39,457
# of total commitments	397,456
# of users with commitments	244,313
Total $ at stake	$35,598,253.74
Minimum $ at stake	$0.5
Maximum $ at stake	$20,000
Date range	19 Oct 2007 - 17 Aug 2023

Table 7.1: Dataset details. 32.8% of total users have created commitments, with total $35 Million at stake.

private or deleted profiles but not any information associated with them. For such users, we approximate their joining date using a public profile before or after them in the sequence. [8]

Figure 7.1 shows the cumulative distribution function (CDF) of the joining dates for users. We observe a linear growth in the number of users over the year, with a slight bump in the rate of private account creation between 2011 to 2013. We also observed an abnormal number of account deletions in 2011, potentially caused by a platform-level glitch. Since our experiments are only performed over public profiles, this does not affect our findings.

Our dataset has 397,456 commitments created by 244,313 unique users (32.8% of total users). Among users with commitments, 187,333 (76.6% of users with commitment) have only one completed commitment; 95% of users have less than four completed commitments in the past. Each commitment has a title, optional description, length, and reporting interval, which defines how often users would report their status. At every reporting period, the user declares whether or not they succeeded in achieving the goal. The user decides the total length and reporting interval.

Users can optionally assign referees and supporters to a commitment. A referee's job is to audit the user's performance and mark the status for each reporting period. [9] Whereas supporters can provide encouragement and social accountability to the user. Unlike referee, supporters can not audit a users performance. Each commitment can have only one referee

[8] For most cases joining dates of the public profile before and after is the same, ensuring the deleted/private profile was also created on the same date. In cases deviating from this pattern, we assign the joining date of the previous public profile.

[9] The platform does not provide a method for referees to audit. It is managed offline between the user and the referee. The latter's report is considered in case of a conflict.

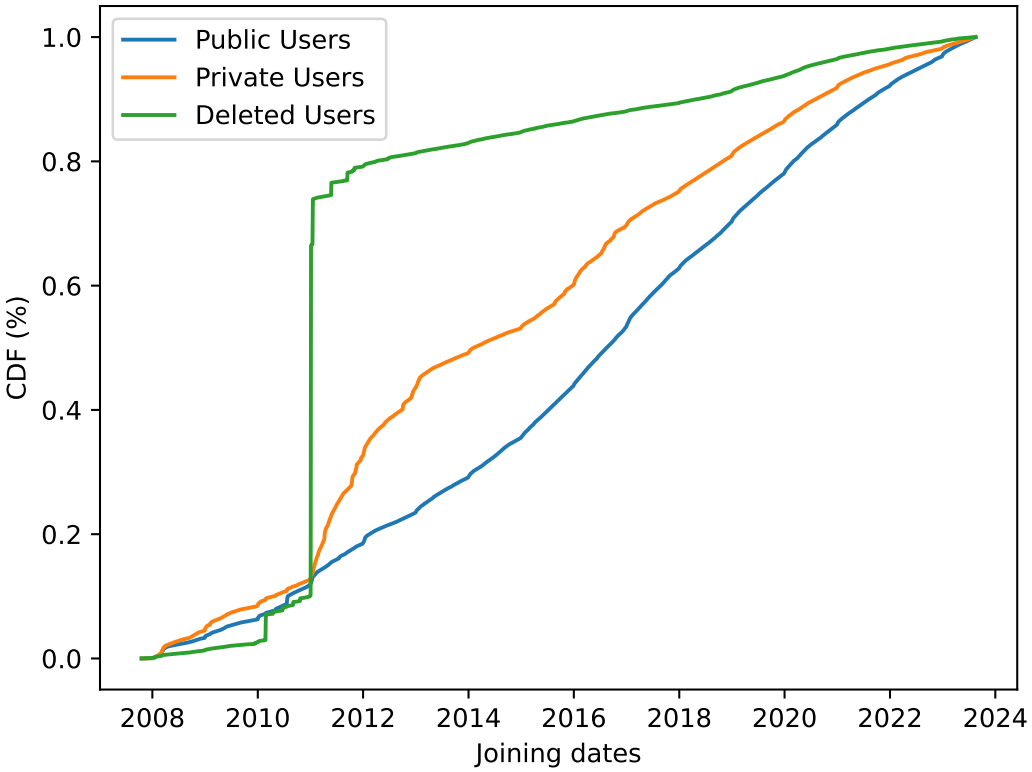


Figure 7.1: CDF of joining date's to the platform. We see a linear increase in number of users over the year. 2011 shows a spike in the number of deleted users potentially caused by platform-level data loss/malfunction.

but any number of supporters. Only 19% and 8.6% of total commitments had referee and supporters assigned, respectively.

Observation 7.1 (RQ2: Accountability Extent) *Most users do not utilize accountability methods in their pursuit of building habits. Only 19% of total commitments had external accountability (referee), and 8.6% had social accountability (supporters) attached to them.*

Finally, the user can also attach a monetary stake to the commitment. The user chooses a stake amount per reporting period, leading total money at stake to be *stake per period* $\times$ *# of reporting periods*. When a user fails to achieve the goal during a reporting period, the stake for that period is awarded to a selected entity (charity, anti-charity, friend, or StickK). It is worth noting that the complete freedom in allowing users to set parameters of the commitments does lead to outlier cases in the dataset, e.g., commitments with unusually long lengths or very high stakes. All analysis in this chapter is performed after removing outliers from the dataset using the Interquartile range (IQR) method [43], as most attributes in our data follow a skewed distribution.

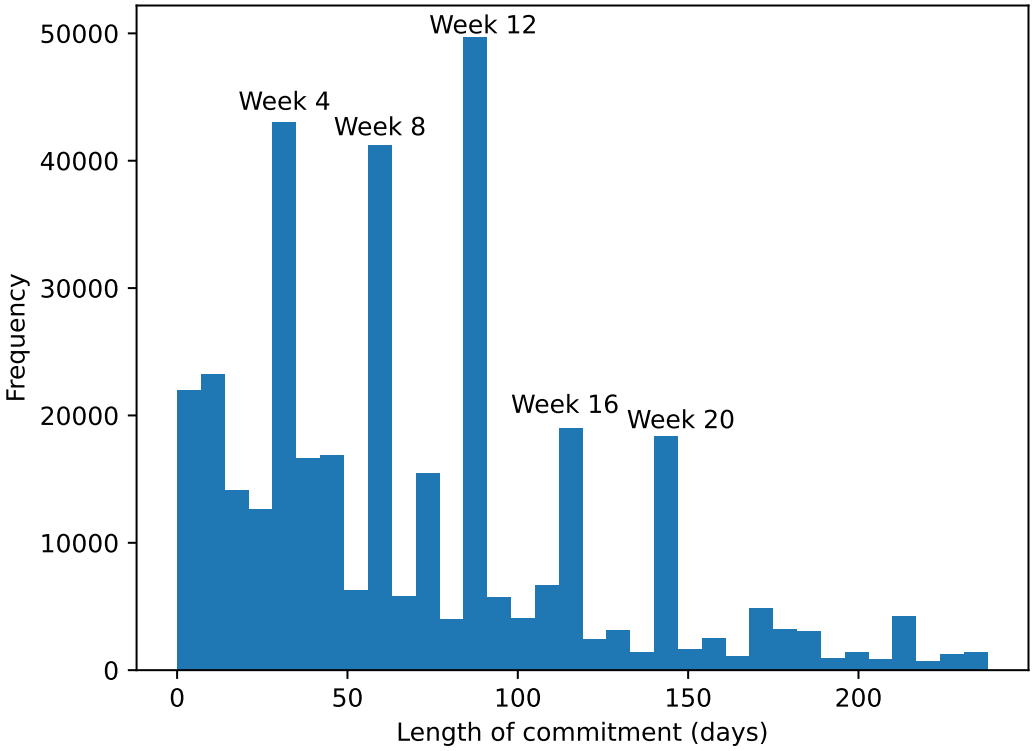


Figure 7.2: Distribution of length of commitments. Lengths in quantum of months like week 4, 8, 12 are most frequent.

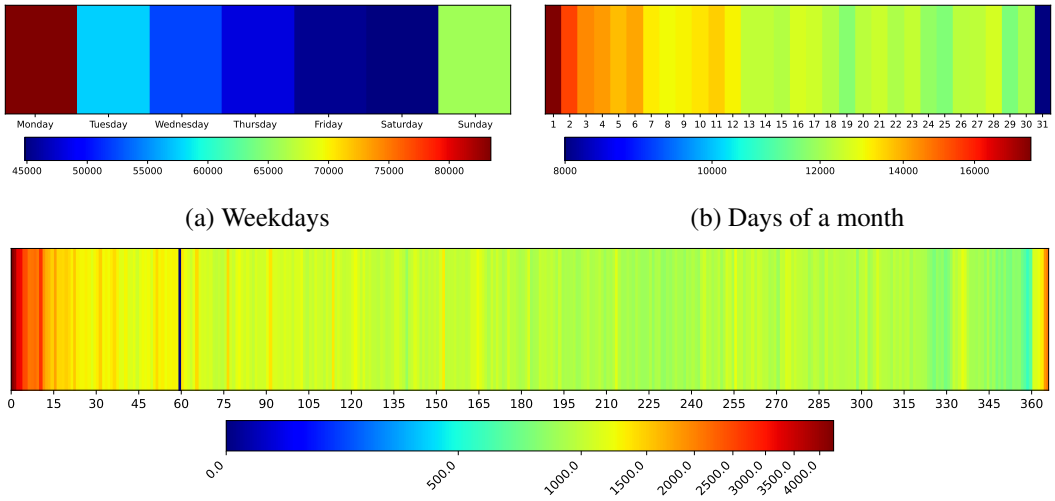
7.5 Temporal Analysis

Temporal patterns have proven to help analyze trends. This section first discusses the temporal patterns observed in the commitments' total length and reporting period. Then in the context of **RQ1**, we look at the relations between the commitment start date and temporal landmarks.

7.5.1 Commitment Length and Reporting Interval

Any habit building exercise aims to reach the state of automaticity [98]. Initially, the user must invest effort towards the commitment until it becomes routine. This makes the initial length of commitment an essential parameter for habit building. Figure 7.2 shows a histogram of lengths of commitments in our dataset. We observe prominent peaks at weeks 4, 8, and 12 and relatively high frequencies at weeks 16 and 20. This shows that users think about habit building planning more often in a quantum of months, with three months being the most popular choice.

Further, we look at the length of reporting intervals chosen by users. Out of the total commitments, 72% (286,206) and 10.6% (42,263) are set up to report progress weekly and daily, respectively. We compare the median commitment length for commitments with daily and weekly reporting intervals. On average, commitments with daily reporting had a length



(a) Weekdays

(b) Days of a month

(c) Days of a year

Figure 7.3: Distribution of commitment start dates. Users are four times more likely to start a commitment on New Year (c) and 40% more likely to start on Monday (a) or 1^{st} of the month (b).

of 35 days, compared to 84 days (12 weeks) for the ones with weekly reporting, showing users shift to more coarser reporting intervals as the total commitment length increases.

Observation 7.2 (Commitment Length) *Users plan habit building in the quantum of months, i.e., 4, 8, and 12 weeks, with the daily or weekly reporting intervals being the most popular.*

7.5.2 Commitment Start Date

The Fresh Start Effect [35] is a cognitive bias which is the user's tendency to start a new goal on specific dates known as temporal landmarks. These landmarks can be general, like New Year, new week/month, or specific, like birthdays, start of a new job/semester. Considering the data available, we keep our analysis limited to the effects of general temporal landmarks, specifically, the start of a new week i.e. Monday, 1^{st} of a new month, and New Year.

Figure 7.3 shows the distribution of commitment start dates across (a) weekdays, (b) days of the months, and (c) days of the year. A skew towards Monday, 1^{st} date, and New Year is apparent. Activities stay relatively high for the first 15 days of the year, with the highest on 1^{st} January. Users are four times more likely to start a commitment on New Year than on an average day of the year. Similar observations are made in the patterns of new weeks and months too. Compared to an average day, users are 40% and 38% are more likely to start a commitment on Monday or 1^{st} of a month, respectively.

Observation 7.3 (RQ1: Key Dates Extent) *The fresh start effect is prevalent among users starting new habits. Users are 40% more likely to start on Monday or 1^{st} day of the month. Four times more commitments are started on New Year's than on an average day.*

Type of stake	Frequency	Percentage
No stake	282,366	71.0%
Anti-charity	60,368	15.2%
Charity	27,053	6.8%
Friend	21,711	5.5%
StickK	5,958	1.5%

Table 7.2: Distribution of different types of stakes. 71% of total commitments do not have any stakes attached to them.

7.6 Stake Analysis

This section discusses the extent part of **RQ3**. The theory of loss aversion tells us that the physiological pain of losing is very powerful and can induce behavioral change [83]. Stickk allows its users to leverage this in their habit building journey by allowing users to attach a monetary stake to the commitments if they wish to do so. In case of failure to achieve the goal for a specific reporting interval, the stake is transferred to an entity chosen prior by the user. StickK allows users to choose from 4 different kinds of entities:

- **StickK**: Stake is passed on to the platform itself.
- A **friend** chosen by the user.
- **Charity** of user's choice.
- **Anti-charity**: An Anti-charity[10] is an organization whose views user strongly oppose. The assumption being a user would want to avoid extending monetary value to such an organization, increasing the motivation to succeed in the commitment.

Table 7.2 shows a distribution of different types of stakes in our dataset. For 71% of total commitments, there are no stakes attached. Anti-charity is the most popular for the commitments with stakes, followed by charity, friend, and StickK. Further, Figure 7.4 depicts the CDF of stake amount per period for different stake categories. $5 is the most common stake amount, followed by $10, $50, and $20, irrespective of the stake type. The slope for the CDF curves of charity and StickK is steep for the smaller amounts, indicating users tend to put less money towards those types.

Observation 7.4 (RQ3: Stake Extent) *29% of the total commitments have a monetary stake attached to them. Anti-charity is the most common stake type and $5, $10 the most common stake amounts per period.*

[10]https://stickk.zendesk.com/hc/en-us/articles/206833337

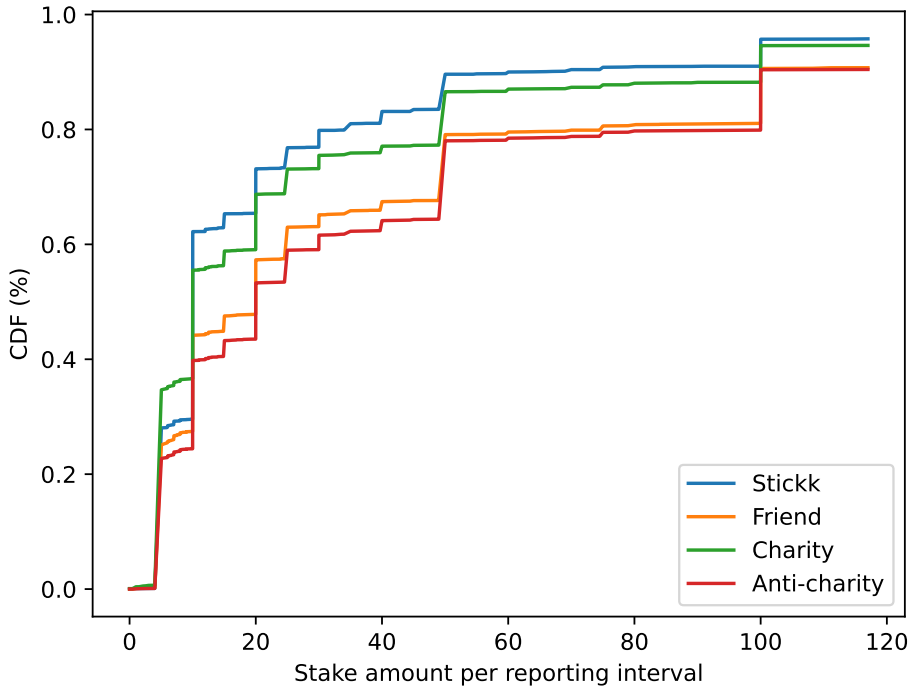


Figure 7.4: CDF of stake per period for different types of stakes. Users tend to put less amount of stake towards charity (green) than anti-charity (red). $5 is the most common stake amount.

7.7 Commitment Classification and Simultaneity

In this section, we start with classifying commitments into different classes of habits. This helps us understand what habits and goals are prevalent among users and how they have changed over time. Further, as part of **RQ4**, how common it is for users to pursue multiple commitments simultaneously.

Title of a commitment talks about what habit users are trying to build. These are generally short combinations of tokens like *"Lose weight", "Study two hours"*, or *"exercise 3x a week"* instead of complete sentences. Considering this semantic property of titles, we used a Word2Vec-based relevance feedback [12] method for retrieval/classification rather than a supervised neural classifier. [11] We start with a set of classes and respective query terms based on common occurrences observed in the manual inspection of the dataset. While retrieval, we match a title to the class if it has query terms or terms semantically similar to query terms. Pretrained Word2Vec [130] embeddings were used for semantic matching. Similar terms were added to the query of the respective class as part of the feedback for the subsequent retrieval cycle. These steps are repeated till convergence. Table 7.3 contains the initial list of classes and related query terms.

[11] A neural classifier would probably work well for this task. However, relevance feedback works well enough for such a simple task without requiring large annotations and computational power.

Class	Query
Weight	weight, diet, fat, pound, kg, calories, kilos, pounds, kcal
Exercise	exercise, run, walk, race, cycling, work-out, workout, bicyclng, gym, km, steps, miles, fitness, yoga, cardio, squats, deadlift, climbing, hike, pushup, pullup, healthy
Study	study, exam, diploma, phd, assignment, math, gmat, homework, gre, sat, school, learn, thesis, degree, certification, preparation, dissertation, class, course, english, french, spanish, java, experiments
Food	eat, chocolate, water, food, sugar, softdrinks, candy, desserts, veggies, gluten, lactose, snacking, coffee, beverage, shakes, caffeine
Smoke	smoking
Sleep	sleep, bed, wake, asleep, nap
Read	read, book
Meditate	meditate, journal
Money	money, finance, saving, expense, spending, earn, save, budget, buy, invest, cash, debt
Write	write, draft, screenplay, scripts, copywriting
Business	client, job, business, network, inbox, emails, career
Alcohol	alcohol, drink, beer, wine, booze
Digital	internet, electronics, tv, phone, mobile, games
Porngraphy	mastrubate, mastrubation, porn, masturbation, nofap, fap, porngraphy
Self-care	nail, hair, brush, floss, shower

Table 7.3: List of classes and related query terms.

Habit class	Frequency	Percentage
Weight	102,584	25.81%
Exercise	88,659	22.31%
Study	22,040	5.55%
Food	14,202	3.57%
Smoke	10,251	2.57%
Sleep	8,956	2.25%
Read	8,735	2.19%
Meditate	7,886	1.98%
Money	7,639	1.92%
Write	5,981	1.50%
Business	4,535	1.14%
Alcohol	4,194	1.05%
Digital	2,826	0.71%
Pornography	2,486	0.62%
Self-care/ Personal hygiene	1,761	0.44%

Table 7.4: Fifteen most common habit classes in our dataset. Habits related to health (Weight, Exercise, Food, Sleep) make 53.94% of total commitments.

Table 7.4 shows the frequency of the top 15 classes identified, and Figure 7.5 shows the proportion of these classes over the year. In total, our Word2Vec-enabled relevance feedback algorithm was able to classify 73% of total commitments successfully. Habits related to health make 53.94% of total commitments. Weight-related commitments were most common at 25.81%, followed by exercise (22.31%) and study (5.54%).

From Figure 7.5, we can observe that though the most famous, proportion of commitments related to weight has gone down over the years, an increase in commitments related to food has gone up. This probably indicates a shift in user mindset about how they perceive food and weight management. We also observe a rise in the proportion of commitments related to reading, sleep, meditation, and digital (limiting phone and internet usage), indicating a shift in emphasis towards mental wellness and self-development.

Observation 7.5 (Habit Classification) *Habits related to health comprise 53.94% of total commitments, with weight being the most common. Over the years, the proportion of habits related to self-development and mental well-being, like sleep, meditation, reading, and digital, has gone up.*

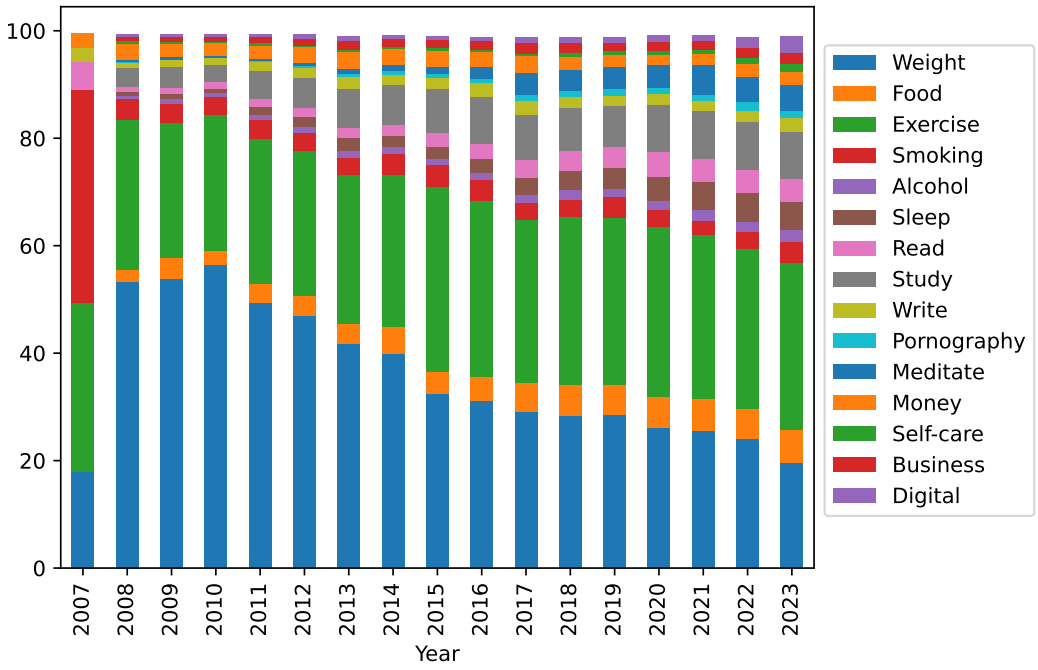


Figure 7.5: The proportion of types of habit over the years. We observe a decline in weight-related habits and an increase in habits related to sleep (brown), meditation (blue), reading (pink), and digital technology (purple).

7.7.1 Simultaneity

In our effort to answer **RQ4**, we want to see how common it is for users to pursue multiple habits in parallel. During our manual inspection, we observed two ways users were structuring multiple habits on the platform. 1) Multiple commitments running during the same time period, 2) User's listed multiple goals in a singular commitment, e.g., *"Lose weight and exercise"* or *"quit smoking and study for finals"*. Our analysis considers both types of structures since our relevance feedback method can perform multi-label classification. Despite being advised against [38], 42% of total commitments (167,511) are pursued with other commitments, with similar habits like weight and exercises often paired together. §Survival Analysis discusses the effect of the simultaneity on commitment success in detail.

Observation 7.6 (RQ4: Simultaneity Extent) *Users tend to pursue multiple habits together, with 41% habit building attempts paired with other goals.*

7.8 Survival Analysis

A key component common across all our research questions is measuring the effect of commitment properties (e.g., start date and length) on users' success rate in the habit building pursuit. We perform survival analysis [131] on our data to answer the "effect" part in **RQ1-4**.

Specifically, we used the Kaplan–Meier estimator [85] for measuring the effect of categorical variables and Cox regression [30] for continuous variables. In this section, we first define how we measure the success rate of a commitment, followed by the details about our survival analysis experiment and the variables used.

7.8.1 Commitment Success Rate

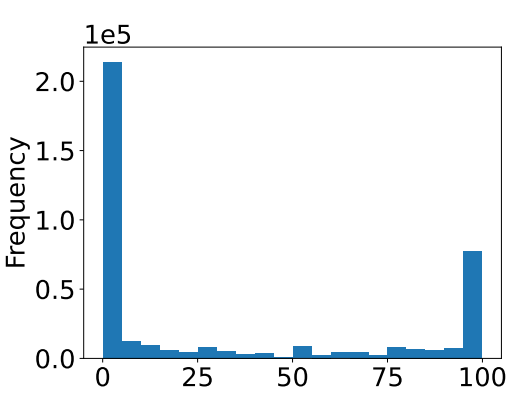
An advantage of using data from StickK is that success for a component is not binary. During a commitment, users check-in at pre-chosen intervals (weekly and daily, most common) to update if they could stick with the habit. A time-stamped record of this checks-in is maintained and is available in our data. A reporting interval can have three statuses, *successful*, *not successful*, or *not reported*. Figure 7.6 shows the proportion distribution of all three statutes across commitments. We observe that the peak frequencies for successful status are in $< 5\%$ bucket or $> 95\%$ bucket, showing that users tend to fail early or do well. Interestingly, the frequency of high rates of not successful is low, but not reported is high, indicating the users who had trouble pursuing the commitments tend to discard the pursuit. Historically, of the total \$35.5 Million on stake, \$4.2 Million and \$1.5 Million have been lost due to intervals being marked as *not reported* and *not successful*, respectively.

Observation 7.7 (Success Rate) *Users either fail early or stick through the entire commitment, indicating that the initial phase of habit building is the most critical. Users with low success levels tend to refrain from returning to the platform for reporting.*

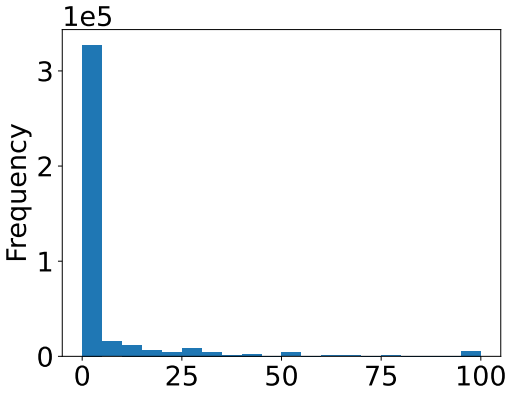
7.8.2 Experiment Details

Typically data for survival analysis experiments are set up in terms of the time it took for an event of interest to occur, e.g., in a study of patients with critical cancer, how many days did a patient survive after the initial diagnosis? In our case, the lengths of commitments are widely different. Hence, to standardize the comparison, instead of measuring success in an absolute number of days, we measure it in terms of the proportion of reports marked as successful. In the terminology of survival analysis, the timeline of our experiment becomes 0 to 100, and the event of interest for a commitment occurs at a timestamp represented by the proportion of reports marked as a success by the user. Commitments with a 100% success rate are marked as right censored entries. In order to prevent our experiments from being biased by commitments that were made frivolously, we excluded all commitments which did not have even one report marked as successful.

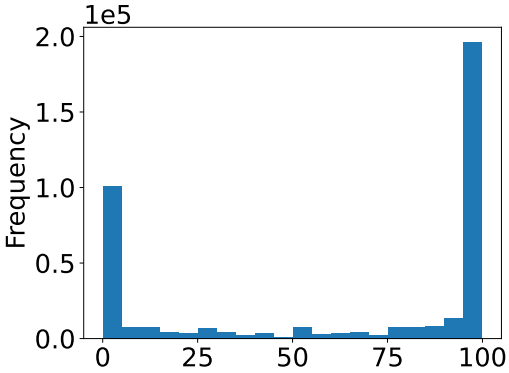
We used Kaplan–Meier estimator [85] to study the effect on survival in cases where the treatment variable is categorical. It is used to estimate the survival function, which measures the fraction of users surviving for a particular time after treatment. To measure the effect of a categorical variable, a comparison is made across the Kaplan–Meier estimates for shards of data divided based on the values the variable can take. Logrank test [18] is used to validate



(a) Proportions of status marked *successful*.



(b) Proportions of status marked *not successful*.



(c) Proportions of status marked *not reported*.

Figure 7.6: Distribution of reporting interval statuses. Users tend to either fail early or do really well (a). The frequency of high rates of not successful is low (b), but not reported is high (c), indicating user's abandonment of the commitment.

if estimates generated from different variable values are statistically significant. In cases where the treatment variable is continuous, we use the Cox proportional hazard model [30], which fits a regression model to evaluate how changes in a continuous variable's value affect a user's survival.

In the context of **RQ1**, we conduct three experiments where the start date of the commitment is the treatment variable. Specifically, we compare the success rate of commitments that started on New Year's, 1^{st} of any month, and Monday with their respective counterparts. Extending on the theme of temporal properties, we also evaluate the effect commitment length and reporting interval have on success rate. Though reporting interval is a continuous variable, 82.6% of all the commitments are set up to report weekly or daily. Hence, we treat reporting interval as a categorical variable with possible values of weekly, daily, and others. For **RQ2**, we compare the success rates of commitments with an external referee vs. self-referring and the effect of the number of supporters (social support) on commitment success. To answer **RQ3**, we measure the effects changes in monetary value of stake per reporting period and who it is bet against (Charity, Anti-charity, Friend, or StickK) has

on the commitment success rate. Finally, for **RQ4**, we compare the survival functions of commitments pursued in simultaneity with others vs. individually.

7.8.3 Results

Figure 7.7 and Table 7.5 show all the results for our survival analysis. As seen in Figure 7.7(b), 7.7(c), commitments started on temporal landmark days like 1^{st} of a month or Monday did not perform any better compared to the ones started on any other day. Conversely, commitments that are started on New Year's have a much worse survival rate than those started later in the year (Figure 7.7(a)).

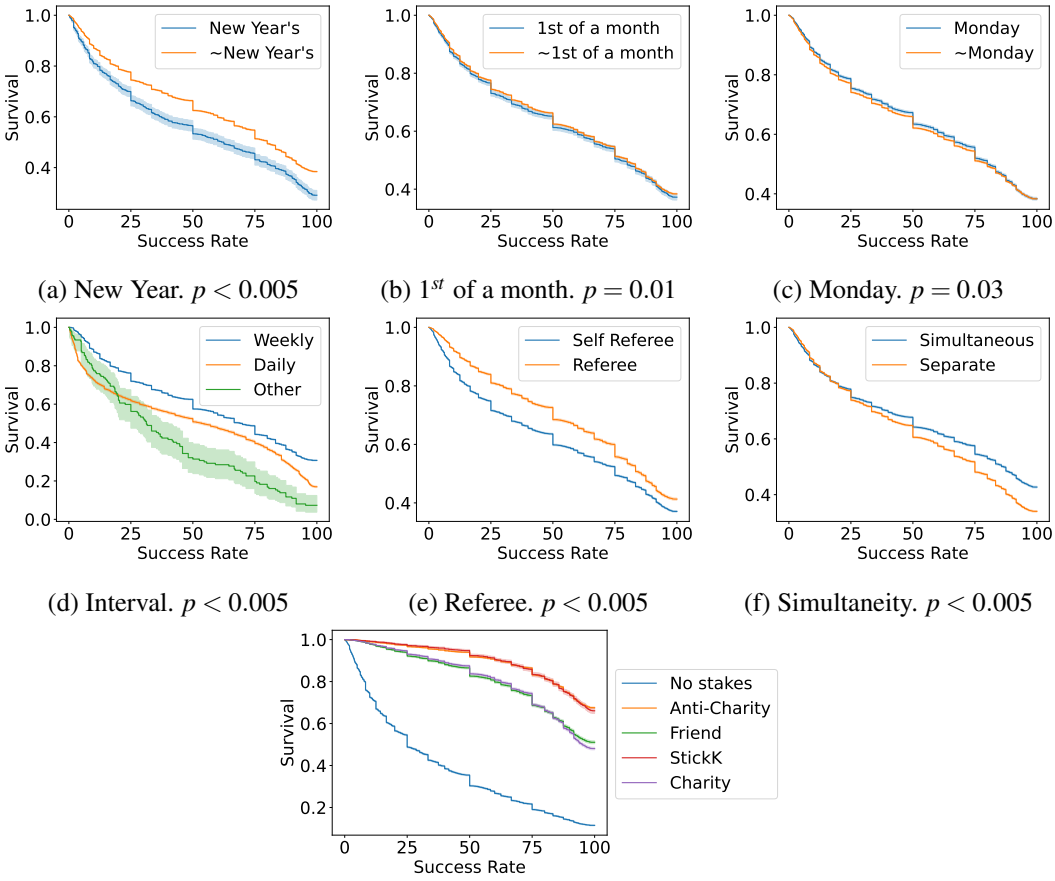
Increased commitment length and number of reporting intervals lead to a decrease in success rate (Table 7.5). The baseline hazard for commitment length was found to be at seven weeks. Regarding reporting intervals, commitments with weekly check-ins have the highest success rate, followed by daily and then others. We observe a sharp decline in the survival function for commitments with daily reporting, indicating that many commitments of this type fail with less than a 20% success rate. However, commitments that pass this threshold tend to achieve greater success rates, as depicted by a reduction in the slope of the survival curve later on (Figure 7.7(d)). Both external and social accountability have a significant positive impact on success rates. Commitments with external referees achieve higher success rates compared to one self-referred (Figure 7.7(e)). Similarly, increasing the number of supporters (social accountability) on the commitment reduces the hazard.

Observation 7.8 (RQ1: Key Dates Effect) *Starting a commitment on a temporal landmark day like Monday or* 1^{st} *of a month does not affect the success rate of habit building.*

Observation 7.9 (RQ2: Accountability Effect) *Having external and social accountability attached to a habit building pursuit in terms of referee and supporters significantly increase the odds of success.*

Commitments with no monetary stakes perform much worse than those with money on the line. An increase in the stake amount per period positively affects the success rate. Interestingly, along with the amount, the entity which the stake is a bet against, also strongly influences users' success rate. Commitments where lost money went to anti-charity or the platform (StickK) had better survival than those with money going to charity or a friend (Figure 7.7(g)). Finally, Figure 7.7(f) shows the effect of pursuing simultaneous commitments on success. Though initially, users pursuing individual goals have a mildly better survival function, but pass a success threshold (approximately 20% success rate), users pursuing multiple commitments tend to have statistically better survival than those pursuing an individual goal.

Observation 7.10 (RQ3: Stake Effect) *Amount of monetary stake and who it is bet against strongly affect the success of habit building pursuit. Adherence increases with the amount*



(a) New Year. $p < 0.005$

(b) 1^{st} of a month. $p = 0.01$

(c) Monday. $p = 0.03$

(d) Interval. $p < 0.005$

(e) Referee. $p < 0.005$

(f) Simultaneity. $p < 0.005$

(g) Type of stake. $p < 0.005$

Figure 7.7: Survival analysis results (Kaplan–Meier curves). $\sim$ represents negation. (a) Commitments started on New Year's have a lower survival probability. Starting on 1^{st} of a month (b) or Monday (c) does not affect the survival function. (d) Reporting every week increases survival. Commitments with an external referee (e) and monitory stake have better survival (f). (g) Finally, Pursuing multiple goals may fail early, but if pursued, it is better than pursing one commitment at a time.

at stake and when it is a bet against an entity that may induce a greater sense of loss, like anti-charities.

Observation 7.11 (RQ4: Simultaneity Effect) *Though sustaining multiple habits initially may be challenging, passing a threshold, users pursuing multiple habits together tend to perform better than those building one habit at a time.*

Covariate	HR	95% CI
# of reports	1.0447***	[1.0437, 1.0457]
Length of commitment	1.0045***	[1.0043, 1.0046]
# of supporters	0.9797***	[0.9725, 0.9870]
$ on stake per period	0.9474***	[0.9467, 0.9482]

$^{***} p <= 0.005$

Table 7.5: Cox Regression results. An increase in monetary stake and the number of supporters increase the success rate. In contrast, increasing the length and number of reports is more hazardous for the user.

7.9 Discussion and Conclusion

7.9.1 Research Questions

In **RQ1**, we analyze the *Fresh Start Effect* [35], a cognitive bias that defines the human tendency to start taking action towards a goal on specific dates, also known as temporal landmarks. These dates can be general, e.g., New Year's, the start of a new month/week, or specific to the user, like birthday or the start of a new job/semester. In this chapter, we study the effects of only the general landmarks. We compare the likelihood of a commitment starting on temporary landmarks and how the success rate of such commitments is different. We found that the fresh start effect is highly prevalent, with 40% more commitments starting on 1^{st} date or Mondays compared to the average day of the month or week. Our analysis does not show any benefits of taking action on these landmark days, with the success rate of commitments started on these dates statistically the same as others. However, such behavior does add an opportunity cost for the user by introducing a delay between the decision to pursue a habit and taking action toward it. These delays can lead to overindulgence (justified as "one last time"), distractions, or loss of motivation. In the absence of any statistical edge, users should take immediate action and not wait for specific days. Further, users are four times more likely to start a commitment on New Year's than on an average day, but commitments started on New Year's are much more likely to fail.

Accountability is an essential factor for successful habit building. In **RQ2**, we study the two accountability options in our data. Referee adds external accountability to a commitment by verifying the user's progress, and supporters can provide social accountability to the user. We found that though users leveraging accountability is rare, with only 19% commitments having a referee and 8.6% having supporters, it is a strong determiner of success. The presence of an external referee and an increased number of supporters lead to a higher success rate. Often, habit building is perceived as an individual pursuit. Adherence increases when others

are involved, maybe because users attach success to social standing, and individuals try to present an idealized version of themselves [59].

Theory of *Loss Aversion* tells us that the psychological pain of losing is twice as powerful as the pleasure of gain. In **RQ3**, we explored how the deterrence of loss can help users in habit building. On StickK, this is manifested by allowing users to assign an optional monetary stake to the commitment. In case of failure to achieve the goal during a reporting period, the amount on stake is passed on to a pre-chosen entity, which can be a friend, the platform itself, or a choice of charity/anti-charity. Like accountability, users' leveraging stakes is uncommon, with only 29% of commitments having it, but it is a strong determinant of success rate. An increase in stake amount causes an increase in success rates. Commitments with no stakes have a steep decaying survival function. Effects of loss aversion are also observed in the types of stakes, with entities that induce a higher sense of loss like anti-charity or platform, leading to a statistically prolonged survival (higher success rate) function for the commitment.

We used a Word2Vec-enabled relevance feedback system to classify commitment into various classes. Fifty four percent of all habits are related to the health of the user. Over the years, we have observed a reduction in the proportion of commitments related to weight and an uptick in the habits related to food, sleep, meditation, and reading. This shows a shift in users' perspective towards a more holistic approach to health.

Finally, in answering **RQ4**, we found that users lean towards developing multiple habits at a time, with 41% of total habit building attempts being made in simultaneity with others. The effect of simultaneous habit building has been unclear. On the one hand, some research suggests users should focus on one thing at a time [38], but theories like habit stacking/anchoring [53] suggest using a combination of related habits at a time. Survival analysis of our data showed that habits practiced in simulating lead to a higer success rate. Through simulations, habits initially have a mildly lower success rate; if a user survives this phase, later on, the success rate is statistically higher, showing the benefit of behavioral momentum [139] in habit building. Most people want to create multiple habit changes, and our analysis shows that this is not only possible, but it can be a better approach. The worst survival of simulation commitments in regions with low success rates indicates that starting with multiple habits together can be challenging and requires proper planning. Further analysis is required to answer planning-related questions such as what kinds of habits go together well or the optimal number of habits to build simultaneously.

7.9.2 Implications

Any habit-tracking tool aims to enable the users in achieving their goals. We believe the findings in this chapter provide direct, actionable insights for both users and platform owners.
Intervention based: Our large-scale analysis reveals practices that increased users' success rate in their habit building pursuits. Users can use these insights to plan their goals, giving them a higher probability of success. Platforms can also use these insights in designing features and interventions to steer the users towards better choices.

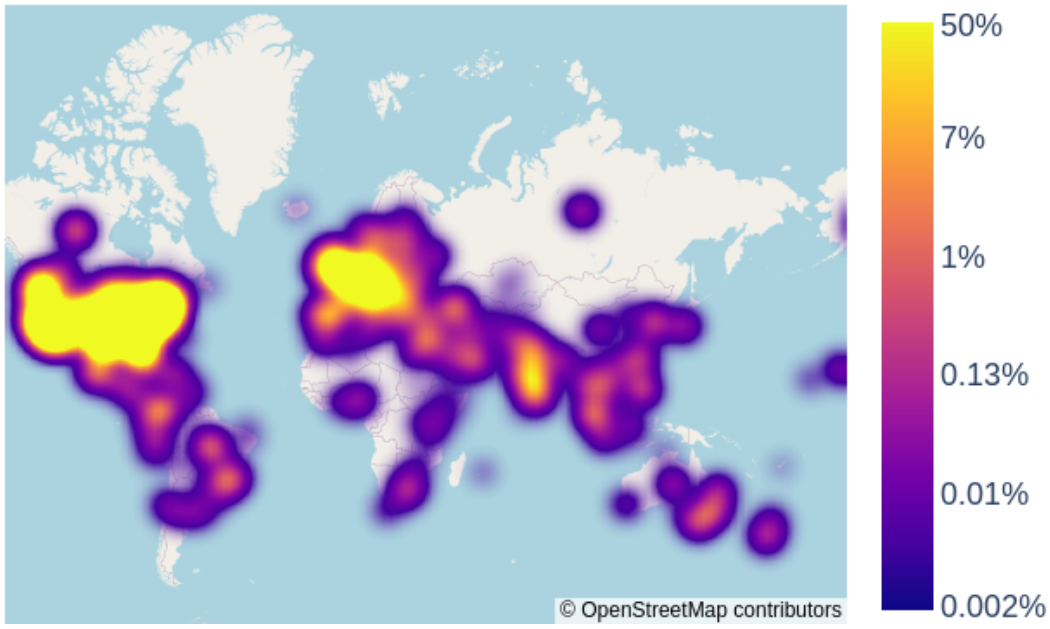


Figure 7.8: Geographical heatmap of user locations. About 50% of all users are located in the USA.

Resource based: Research related to habit building exercises can be logistically challenging. Most of the past literature is based on small-scale single-habit trials. Our data will enable researchers to observe patterns and validate theories on a large-scale dataset of real-life heterogeneous habit building attempts. §Future Work talks about potential use cases of this data in detail.

Niche platform: Though not the first one to do it, our work demonstrates the use of smaller and niche platforms to study characteristics of human behavior. This becomes even more important with popular social media platforms like Twitter and Reddit imposing data collection limitations.

7.9.3 Threats to Validity

Ensuring generalizability and data accuracy is always challenging while using online data to analyze offline human behavior. Our analysis is also susceptible to these challenges. Firstly, we define success rate as the proportion of reporting periods marked as successful by the user (by the referee in case one is present). We do not have a way to validate the authenticity of these report, ensuring that the user was able to achieve the goal. The converse is also true; lack of a report does not necessarily mean a relapse in habit but can be a function of other factors such as not liking the platform. We use large-scale longitudinal data consisting of hundreds of thousands of commitments, reducing the possibility of large-scale tampered data. Success at habit building is also affected by the user's environmental factors, like social/family support and motivation behind the pursuit, which are not captured in our data.

Our data is collected online, not leading to a representative sample. A total of 38,828 users, which accounts for 5.22% (as of August 2023) of all users, have opted to make their location public on the platform. Figure 7.8 shows the geographical heatmap of user locations. Most of our users are in the USA. Followed by Europe and Southeast Asia. Regions like Russia, Africa, China, Mongolia, and South America have limited presence in our data. This is probably caused by various cultural, political, or economic reasons. Further work is required to ensure our findings are applicable globally. Lastly, our unsupervised relevance feedback-based classifier could not assign labels to 27% of the data. In this study, we curated classes and associated keywords (Table 7.3) based on frequent occurrences during manual inspection. A structured recursive annotation process will be required to ensure better coverage and validity of labels.

7.10 Future work

One significant contribution of our work is to release a large-scale dataset of heterogeneous habit building attempts. In this chapter, we measure the extent and effects of various commitment properties on success rate. However, this data allows us to study multiple aspects of habit building not touched on in our analysis. Firstly, we define success rate as the proportion of the reporting period marked as successful. We only account for the proportion of successful reports, not the chronology of the reports. It would be interesting to explore the relationship between user success rate and their commitment stage. Further, how do streaks relate to the overall success rate of the attempt? Second, since we find simultaneous habit building attempts are very prevalent, their scope of exploring the specifics of such attempts. What commitments are most often linked together? Are certain combinations more favorable than others? Finally, since weight management is one of the most common commitments, the platform allows users to add specific information to such commitments like start weight, end weight, rate of change in weight, etc. Though not utilized in our analysis, this information is in our data. It can enable researchers to explore weight management specific questions such as what is the optimal rate of weight loss from the perspective of adherence?

Part V

Conclusion and Future Work

Chapter 8

Conclusion

In this thesis, we worked on the problem of characterizing and quantifying online-offline interactions on social-technical platforms and the impact such interactions have on user behavior. While keeping the individual user as the focal point, we leverage large-scale interaction data from social-technical platforms. The works presented in this thesis can be broadly categorized into three forms of interactions - i) the effect online community feedback can have on individual offline actions, ii) organizations leveraging individual customers' online presence to optimize business processes, and iii) how data from tracking platforms can be used to uncover the strategies behind successful users. The first part quantifies how community reactions towards a user affect their future behavior on the platform, motivation to indulge in various activities in the offline world, and longevity in the community itself. The second part explores the possibility of utilizing cross-platform data in cases of absence or limited availability of required interaction data on the host platform. Finally, in part three, we demonstrate how non-conventional online data sources, like tracking applications, can be rich data sources, help uncover actionable insights, and verify anecdotal evidence.

8.1 Contributions

8.1.1 Individual-Community Interactions

- **Effect of popularity shocks:** In Chapter 3, we explore the changes that occur in user's actions on the platform after receiving sudden and unexpected engagement/attention from the community. We collect user timelines of 30,969 users from a popular short video platform. We propose an algorithm to identify the first popularity shock from a user's timeline. Using popularity shock as an intervention, we use causal inference techniques to examine the change in behavior from pre-and-post popularity shock. We observe that popularity shocks lead to an increase in the posting frequency of users, and users alter their content to match the one that resulted in the shock. Also, shocks are found to be challenging to maintain, with effects fading within a few days for most users. High response from viewers and diversification of content posted are found to be linked with longer survival durations of the shock effects.

- **Feedback's effect on drug consumption:** In Chapter 4, we quantify how positive feedback from the community can lead to an increase in user activity indicative of offline drug consumption. The work bases the hypothesis on the famous sociology

theory of *primacy effect*, *operant conditioning*, and *edgework*. We collect historical data from 10 drug-related subreddits totaling 826,905 posts and 6.6 Million comments made by 493,906 unique users. First, we built a deep-learning model to classify user-generated content as indicative of offline drug consumption and achieved a macro F1 score of 79.54. We discover that 84% of posts and 55% comments on drug-related subreddits indicate real-life drug consumption. Further, we use matching-based causal inference techniques to unravel community feedback's effect on users' future drug consumption behavior. Users who receive positive feedback from the community on drug consumption activity tend to generate up to two times more drug consumption content in the future. Finally, we conducted an anonymous user study on drug-related subreddits to compare members' opinions with our experimental findings and show that user tends to underestimate the effect community peers can have on their decision to interact with drugs.

- **Support dynamics in communities:** In Chapter 5, we study how support from the community affects a user's longevity in the community. To this end, we collect all the activities in the popular online support group *COVID-19positive* since its inception. We define various support classes and observe them along with user behavior and temporal phases for a coherent community. We perform survival analysis using Cox Regression to identify factors influencing a user's commitment to the community. In a COVID-19 community, emotional support involves discussing recovery and the status of family and loved ones. Emotions such as gratitude, prayer, and hope are expressed. Informational support involves discussion around research, infections, finance, and tests. People who stay longer seek more informational and emotional support from the community. They also give more support. Surprisingly, the amount of support a user receives from the community is independent of the user's decision to stay. Furthermore, talking about symptoms and recovery and interacting with more users in the community promote a more extended stay.

8.1.2 Individual-Organization Interactions

- **Re-Identification Assisted RTO Detection:** In Chapter 6, we explore the problem of Return To Origin (RTO) faced by E-commerce platforms, where the user cancels an order while it is in transit for delivery. In such a scenario, the platform faces logistics and opportunity costs. Sociology literature has highlighted clear correlations between socioeconomic indicators and users' tendency to exploit systems to gain financial advantage. We leverage public information available in social media profiles like location, education, and profession as an estimator of socioeconomic condition. We propose using location information available in e-commerce platforms to perform and validate social re-identification of user profiles. Internal data fused with extracted social features are used to train an RTO order detection model. Our system, when

trained and evaluated on real-life data from one of the largest e-commerce platforms of India, demonstrates a performance improvement in RTO detection of 3.1% and 19.9% on precision and recall, respectively.

8.1.3 Individual centric Interactions

- **Characterization of Habit Building Attempts:** In Chapter 7, we analyze past attempts at habit building by people to uncover what chratersticks lead to success. We base our work on theories like the Fresh Start Effect, Loss aversion, and Behavioral Momentum. We collect data of 397,456 habit-building attempts made by 244,313 users on stickk.com. Habits related to health (weight, Exercise, Food, Sleep) comprise 53.94% of total commitments, with weight being the most common. Over the years, the proportion of habits related to self-development and mental well-being, like sleep, meditation, reading, and digital, has gone up. We ask questions about how commitment properties like start date, external accountability, monetary stake, and pursuing multiple habits together affect the odds of success. Users are 40% more likely to start a commitment on 1^{st} of a month, or Monday, and compared to an average day in the year, four times more commitments are started on New Year's. Though commitments started on New Year's are more likely to fail, commitments started on other temporal landmarks do not affect the likelihood of success. Only 29% and 19% of total commitments have monitor stakes and external accountability attached, respectively, but doing so significantly increases the success rate. The success rate increases if a stake is given to an anti-charity instead of a charity on failure. Users who keep short-term commitments (7 weeks is the base hazard) do frequent check-ins on the application, and leverage behavioral momentum for pursuing multiple commitments simultaneously are more likely to succeed.

8.2 Limitation

This section highlights three overarching limitations encountered across the studies: Data Source and Representation, Hidden Confounders, and Weak Proxies.

- **Data Source and Representation:** The reliance on specific data sources and limited representation poses a significant challenge. Data sourced from singular platforms or restricted user subsets may not fully capture the diversity of behaviors and demographics present in the broader population. This limitation restricts the generalizability of findings beyond the confines of the studied platforms or user samples.

- **Hidden Confounders:** The presence of unobserved variables, or hidden confounders, introduces potential biases into the analyses. Despite efforts to control for various factors, the influence of unaccounted variables on the observed outcomes remains

a concern. These hidden confounders could confound the interpretation of results, leading to inaccurate or incomplete conclusions.

- **Weak Proxies:** Another limitation arises from the use of proxies that inadequately represent the constructs of interest. Whether utilizing user-generated content as a proxy for offline behavior or employing certain metrics as indicators of underlying phenomena, the validity and reliability of these proxies may be compromised. Weak proxies hinder the accuracy and robustness of the analyses, potentially skewing the interpretation of findings.

Chapter 9

Future Work

The work presented in this thesis is a preliminary step in utilizing large-scale online data to understand user interactions and their effects on users' actions online and offline. Building on the work presented in the thesis, there are four different streams of future work that could be seen as the next logical step for this dissertation. The first stream focuses on the stream of computational social science (§ 9.1). It would involve exploring causal inference methods and mixed-method studies to establish more robust proxies between online interactions and offline actions. The next opportunity is the development of accessible tools that can help community moderators and platform designers analyze interactions and their effects (§ 9.2). The third stream focuses on sociology and leveraging large-scale online data to develop/improve theories of human behavior. Finally, we aim to extend previously developed methods and produce new methods for analyzing large-scale archival data in other exciting domains (§ 9.3).

9.1 Computational Social Science

A promising direction is to continue leveraging the large-scale data the social-technical platforms provide to mine patterns in human behavior. Specifically, we find the following directions interesting:

9.1.1 Causal Inference

Most social media-based studies are only focused on correlation. However, given the vast implications of these studies, it is necessary to argue whether the factors discovered to predict the behavior are causal. However, many previously developed causal methods are not apt for making causal inferences from social media data, which is highly unstructured and multimodal. One of the promising directions is to develop/adapt causal inference methods that can handle multimodal covariates like text, images, and graphs.

9.1.2 Mixed Method Studies

Most of the work presented in this thesis focuses on the quantitative aspects of user analysis. However, qualitative analysis is an equally important aspect of developing impactful insights. An essential aspect of future work in this direction is to design mixed-method studies where qualitative data can be collected through user surveys and fed into the quantitative

analysis. A common problem with building causal inference models on online data is hidden confounders; user survey data can be instrumental in filling such gaps in the observed data and producing more robust results.

9.1.3 Establishing Stronger Online Offline Proxy

All the work presented in this thesis uses some online action as a proxy for offline behavior. Most of the time, these proxies are weak. Such proxies are hard to validate, e.g., when a user marks a commitment as successful or claims about consuming drugs, we do not have a way of validating that. Secondly, such proxies are incomplete; e.g., if a user stops posting their habit progress, it does not necessarily mean they stopped pursuing the goal. In order to increase the varsity of the findings made by such studies, it is necessary to develop methods to establish stronger proxies. Interesting directions to pursue in this stream can be to consolidate multiple weak proxies or use mixed method studies to measure the discrepancy ratio between reality and online data, which can then be used to adjust the proxy estimate.

9.2 Tools

Platform developers perform extensive A/B testing to understand the effects of features on business metrics. However, community moderators rarely have access to perform such experiments, which can help understand and improve community behavior. In such cases, the next best option is to mine insights from past data, which requires extensive technical and scientific know-how. An exciting future direction can be to develop accessible tools that can help community moderators and researchers perform such analysis with ease.

Some common steps in most of this analysis are data ingestion, pre-processing, treatment selection, effect, covariates, choosing appropriate algorithms, and evaluation. The future aim would be to develop a tool that can help perform all of these steps while abstracting the technicalities of the end user. Some of the challenges that should be focused on while developing such a tool are:

- **Automatic data injection:** Online data comes in various data types and formats. These structures can vary from platform to platform. For a tool to be widely useful, it should have adapters to collect data from various popular sources like Twitter, Reddit, Wikipedia, and GitHub and store it in a consistent format.
- **Appropriate abstraction:** A significant challenge while creating tools like these is the level of technical abstraction. Performing classification, clustering, and causal inference on large-scale unstructured data is complex, and standardized solutions do not exist. Build a very abstract tool, and it will perform poorly in complex scenarios. On the other hand, if the tool requires too much custom formulation, it would alienate many users. Finding the right balance between solid abstraction for regular users while allowing flexible tweaking mechanisms to power users is necessary.

9.3 Sociology

Work done in this thesis highlights different scenarios and discrepancies and similarities in user behavior based on those scenarios. On one hand, community feedback has a strong effect on future drug consumption; on the other side, community support has little effect on user's longevity in Covid-19 support groups. Such discrepancies have also been observed in the literature. Accountability from offline peer groups improves writing quality [174], whereas feedback from the online community can have an adverse effect on the writing quality [25]. The monetary stake has a more substantial effect on success when attributed to anti-charity vs. charity. User behavior has a complex bearing on multiple variables like type, strength, schedule of treatment, and who it is coming from. Large-scale data from online social and technical platforms can provide a fertile ground to evaluate, propose, and develop sociology theories that can explain different scenarios and their effect on user behavior.

Bibliography

[1] Nancy R Ahern, Penny Sauer, and Paige Thacker. Risky behaviors and social networking sites: how is youtube influencing our youth? *Journal of psychosocial nursing and mental health services*, 53(10):25–29, 2015.

[2] Adam L Alter and Hal E Hershfield. People search for meaning when they approach a new decade in chronological age. *Proceedings of the National Academy of Sciences*, 111(48):17066–17070, 2014.

[3] Tim Althoff, Pranav Jindal, and Jure Leskovec. Online actions with offline impact: How online social networks influence online and offline user behavior. In *WSDM*, 2017.

[4] James W Anderson, Elizabeth C Konz, Robert C Frederich, and Constance L Wood. Long-term weight-loss maintenance: a meta-analysis of us studies. *The American journal of clinical nutrition*, 74(5):579–584, 2001.

[5] Carol A Anson, Douglas J Stanwyck, and James S Krause. Social support and health status in spinal cord injury. *Spinal Cord*, 31(10):632–638, 1993.

[6] Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6679–6687, 2021.

[7] S. E. Asch. Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, 41(3):258–290, 1946.

[8] Ian Ayres. *Carrots and sticks: Unlock the power of incentives to get things done*. Bantam, 2010.

[9] Nupur Baghel, Yaman Kumar, Paavini Nanda, Rajiv Ratn Shah, Debanjan Mahata, and Roger Zimmermann. Kiki kills: Identifying dangerous challenge videos from social media. *arXiv preprint arXiv:1812.00399*, 2018.

[10] Fabricio Balcazar, Bill L Hopkins, and Yolanda Suarez. A critical, objective review of performance feedback. *Journal of Organizational Behavior Management*, 7(3-4):65–89, 1985.

[11] Duilio Balsamo, Paolo Bajardi, and Andr?? Panisson. Firsthand opiates abuse on social media: monitoring geospatial patterns of interest through a digital cohort. In *The World Wide Web Conference*, 2019.

[12] Duilio Balsamo, Paolo Bajardi, Alberto Salomone, and Rossano Schifanella. Patterns of routes of administration and drug tampering for nonmedical opioid consumption: Data mining and content analysis of reddit discussions. *Journal of Medical Internet Research*, 2021.

[13] Antonina Bambina. *Online social support: the interplay of social networks and computer-mediated communication*. Cambria press, 2007.

[14] Albert Bandura. Self-efficacy: toward a unifying theory of behavioral change. *Psychological review*, 84(2):191, 1977.

[15] Christopher E Beaudoin and Chen-Chao Tao. Benefiting from social capital in online support groups: An empirical study of cancer patients. *CyberPsychology & Behavior*, 10(4):587–590, 2007.

[16] Lisa F Berkman. The role of social relations in health promotion. *Psychosomatic medicine*, 57(3):245–254, 1995.

[17] Michał Bilewicz and Wiktor Soral. Hate speech epidemic. the dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, 41:3–33, 2020.

[18] J Martin Bland and Douglas G Altman. The logrank test. *Bmj*, 328(7447):1073, 2004.

[19] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[20] Christian Burgers, Allison Eden, Mélisande D van Engelenburg, and Sander Buningh. How feedback boosts motivation and play in a brain-training game. *Computers in Human Behavior*, 2015.

[21] Wenjun Cao, Ziwei Fang, Guoqiang Hou, Mei Han, Xinrong Xu, Jiaxin Dong, and Jianzhong Zheng. The psychological impact of the covid-19 epidemic on college students in china. *Psychiatry research*, 287:112934, 2020.

[22] Andrew T Carswell and Douglas C Bachtel. Mortgage fraud: A risk factor analysis of affected communities. *Crime, law and social change*, 52(4):347–364, 2009.

[23] Mehmet Cem Catalbas, Tomaz Cegovnik, Jaka Sodnik, and Arif Gulten. Driver fatigue detection based on saccadic eye movements. In *2017 10th International Conference on Electrical and Electronics Engineering (ELECO)*, pages 913–917. IEEE, 2017.

[24] Pei-Yu Chen, Samita Dhanasobhon, and Michael D Smith. All reviews are not created equal: The disaggregate impact of reviews and reviewers at amazon. com. *Com (May 2008)*, 2008.

[25] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. How community feedback shapes user behavior. In *ICWSM*, 2014.

[26] Jinmyoung Cho, Marcia G Ory, and Alan B Stevens. Socioecological factors and positive aspects of caregiving: Findings from the reach ii intervention. *Aging & mental health*, 20(11):1190–1201, 2016.

[27] Wei-Po Chou, Peng-Wei Wang, Shiou-Lan Chen, Yu-Ping Chang, Chia-Fen Wu, Wei-Hsin Lu, and Cheng-Fang Yen. Risk perception, protective behaviors, and general anxiety during the coronavirus disease 2019 pandemic among affiliated health care professionals in taiwan: Comparisons with frontline health care professionals and the general public. *International journal of environmental research and public health*, 17(24):9329, 2020.

[28] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The covid-19 social media infodemic. *Scientific reports*, 10(1):1–10, 2020.

[29] Thomas D Cook. "waiting for life to arrive": a history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, pages 636–654, 2008.

[30] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1972.

[31] MeLisa R Creamer, Teresa W Wang, Stephen Babb, Karen A Cullen, Hannah Day, Gordon Willis, Ahmed Jamal, and Linda Neff. Tobacco product use and cessation indicators among adults—united states, 2018. *Morbidity and mortality weekly report*, 68(45):1013, 2019.

[32] Tiago Cunha, Ingmar Weber, and Gisele Pappa. A warm welcome matters! the link between social feedback and weight loss in/r/loseit. In *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017.

[33] Tiago Oliveira Cunha, Ingmar Weber, Hamed Haddadi, and Gisele L Pappa. The effect of social feedback in a reddit weight loss community. In *Proceedings of the 6th International Conference on Digital Health Conference*, pages 99–103, 2016.

[34] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, pages 693–696. Springer, 2013.

[35] Hengchen Dai, Katherine L Milkman, and Jason Riis. The fresh start effect: Temporal landmarks motivate aspirational behavior. *Management Science*, 60(10):2563–2582, 2014.

[36] Hengchen Dai, Katherine L Milkman, and Jason Riis. Put your imperfections behind you: Temporal landmarks spur goal initiation when they signal new beginnings. *Psychological science*, 26(12):1927–1936, 2015.

[37] Rene Dailey, Lynsey Romo, Sarah Myer, Cathy Thomas, Surabhi Aggarwal, Kelly Nordby, Madison Johnson, and Carolyn Dunn. The buddy benefit: Increasing the effectiveness of an employee-targeted weight-loss program. *Journal of health communication*, 23(3):272–280, 2018.

[38] Amy N Dalton and Stephen A Spiller. Too much of a good thing: The benefits of implementation intentions depend on the number of goals. *Journal of Consumer Research*, 39(3):600–614, 2012.

[39] Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian Lee. How opinions are received by online communities: a case study on amazon. com helpfulness votes. In *WWW*, 2009.

[40] Munmun De Choudhury and Sushovan De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 71–80, 2014.

[41] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the CHI 2016*, pages 2098–2110, 2016.

[42] Marijke De Veirman, Veroline Cauberghe, and Liselot Hudders. Marketing through instagram influencers: the impact of number of followers and product divergence on brand attitude. *International journal of advertising*, 36(5):798–828, 2017.

[43] Frederik Michel Dekking, Cornelis Kraaikamp, Hendrik Paul Lopuhaä, and Ludolf Erwin Meester. *A Modern Introduction to Probability and Statistics: Understanding why and how*, volume 488. Springer, 2005.

[44] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[45] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[46] Jay Dixit. Heartbreak and home runs: The power of first experiences, Jan 2010.

[47] James A Doyle. *The male experience*. Brown & Benchmark, 1995.

[48] Christine Dunkel-Schetter. Social support and cancer: Findings based on patient interviews and their implications. *Journal of Social issues*, 40(4):77–98, 1984.

[49] Flavio Figueiredo. On the prediction of popularity of trends and hits for user generated videos. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 741–746, 2013.

[50] Flavio Figueiredo, Jussara M Almeida, Fabrício Benevenuto, and Krishna P Gummadi. Does content determine information popularity in social media? a case study of youtube videos' content and their popularity. In *CHI*, 2014.

[51] Flavio Figueiredo, Fabrício Benevenuto, and Jussara M Almeida. The tube over time: characterizing popularity growth of youtube videos. In *WSDM*, 2011.

[52] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[53] Brian J Fogg. *Tiny habits: The small changes that change everything*. Eamon Dolan Books, 2019.

[54] Karen Freberg, Kristin Graham, Karen McGaughey, and Laura A Freberg. Who are the social media influencers? a study of public perceptions of personality. *Public relations review*, 2011.

[55] Eline Frison and Steven Eggermont. The impact of daily stress on adolescents' depressed mood: The role of social support seeking through facebook. *Computers in Human Behavior*, 44:315–325, 2015.

[56] Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. Social clicks: What and who gets read on twitter? In *ACM SIGMETRICS ICMMCS*, 2016.

[57] Xing Gao, Wenli Ji, Yongjun Li, Yao Deng, and Wei Dong. User identification with spatio-temporal awareness across social networks. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1831–1834, 2018.

[58] Xavier Giné, Dean Karlan, and Jonathan Zinman. Put your money where your butt is: a commitment contract for smoking cessation. *American Economic Journal: Applied Economics*, 2(4):213–235, 2010.

[59] E Goffman. The presentation of self in. butler, bodies that matter. *The Presentation of Self in. Butler, Bodies that Matter*, 1959.

[60] Oana Goga, Howard Lei, Sree Hari Krishnan Parthasarathi, Gerald Friedland, Robin Sommer, and Renata Teixeira. Exploiting innocuous activity for correlating users across sites. In *Proceedings of the 22nd international conference on World Wide Web*, pages 447–458, 2013.

[61] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34:18932–18943, 2021.

[62] Amelia Grant-Alfieri, Judy Schaechter, and Steven E Lipshultz. Ingesting and aspirating dry cinnamon by children and adolescents: the "cinnamon challenge". *Pediatrics*, 131(5):833–835, 2013.

[63] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. Tweetcred: Real-time credibility assessment of content on twitter. In *International conference on social informatics*, pages 228–243. Springer, 2014.

[64] Omkar Gurjar, Tanmay Bansal, Hitkul Jangra, Hemank Lamba, and Ponnurangam Kumaraguru. Effect of popularity shocks on user behaviour. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):253–263, May 2022.

[65] Scott D Halpern, David A Asch, and Kevin G Volpp. Commitment contracts as a way to health. *Bmj*, 344, 2012.

[66] Heather Harris and S Jeanne Horst. A brief guide to decisions at each step of the propensity score matching process. *Practical Assessment, Research, and Evaluation*, 21(1):4, 2016.

[67] Catherine Hausman and David S Rapson. Regression discontinuity in time: Considerations for empirical applications. *Annual Review of Resource Economics*, 10:533–552, 2018.

[68] Atika Hermanda, Ujang Sumarwan, and Netti Tinaprillia. The effect of social media influencer on brand image, self-concept, and purchase intention. *Journal of Consumer Sciences*, 4(2):76–89, 2019.

[69] Audrey Hickert, Hanneke Palmen, Anja Dirkzwager, and Paul Nieuwbeerta. Receiving social support after short-term confinement: How support pre-and during-confinement contribute. *Journal of Research in Crime and Delinquency*, 56(4):563–604, 2019.

[70] Bernie Hogan. The presentation of self in the age of social media: Distinguishing performances and exhibitions online. *Bulletin of Science, Technology & Society*, 2010.

[71] Han Hu, NhatHai Phan, James Geller, Stephen Iezzi, Huy T Vo, Dejing Dou, and Soon Ae Chun. An ensemble deep learning model for drug abuse detection in sparse twitter-sphere. In *MedInfo*, 2019.

[72] Guido W Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. *Journal of econometrics*, 142(2):615–635, 2008.

[73] Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

[74] Paridhi Jain, Ponnurangam Kumaraguru, and Anupam Joshi. @ i seek'fb. me' identifying users across multiple online social networks. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1259–1268, 2013.

[75] Soheil Jamshidi and Mahmoud Reza Hashemi. An efficient data enrichment scheme for fraud detection using social network analysis. In *6th International Symposium on Telecommunications (IST)*, pages 1082–1087. IEEE, 2012.

[76] Hitkul Jangra, Abinaya K, Soham Saha, Satyajit Banerjee, Muthusamy Chelliah, and Ponnurangam Kumaraguru. Social re-identification assisted rto detection for e-commerce. In *Companion Proceedings of the ACM Web Conference 2023*, pages 854–858, 2023.

[77] Hitkul Jangra, Tanisha Pandey, Sonali Singhal, Pranjal Kandhari, Aryamann Tomar, and Ponnurangam Kumaraguru. Together apart: Decoding support dynamics in online covid-19 communities. In *2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2023.

[78] Hitkul Jangra, Rajiv Shah, and Ponnurangam Kumaraguru. Effect of feedback on drug consumption disclosures on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 435–446, 2023.

[79] Hitkul Jangra, Rajiv Shah, and Ponnurangam Kumaraguru. Put your money where your mouth is: Dataset and analysis of real world habit building attempts. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2024.

[80] S Venus Jin, Aziz Muqaddam, and Ehri Ryu. Instafamous and social media influencer marketing. *Marketing Intelligence & Planning*, 2019.

[81] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029. PMLR, 2016.

[82] Arlind Kadra, Marius Lindauer, Frank Hutter, and Josif Grabocka. Regularization is all you need: Simple neural nets can excel on tabular data. *arXiv preprint arXiv:2106.11189*, 2021.

[83] Daniel Kahneman. Prospect theory: An analysis of decision under risk. *Journal of Political Economy*, 85:97–122, 1977.

[84] Nathan Kallus. Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. In *International Conference on Machine Learning*, pages 5067–5077. PMLR, 2020.

[85] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[86] Brian Keegan, Darren Gergle, and Noshir Contractor. Staying in the loop: Structure and dynamics of wikipedia's breaking news collaborations. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, pages 1–10, 2012.

[87] Brian Keegan, Darren Gergle, and Noshir Contractor. Hot off the wiki: Structures and dynamics of wikipedia's coverage of breaking news events. *American behavioral scientist*, 2013.

[88] Brian C Keegan and Jed R Brubaker. 'is' to'was' coordination and commemoration in posthumous activity on wikipedia biographies. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 533–546, 2015.

[89] Katherine Keith, David Jensen, and Brendan O'Connor. Text and causal inference: A review of using text to remove confounding from causal estimates. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5332–5344, Online, July 2020. Association for Computational Linguistics.

[90] Steven A Kemp, Dami A Collier, Rawlings P Datir, Isabella ATM Ferreira, Salma Gayed, Aminu Jahun, Myra Hosmillo, Chloe Rees-Spear, Petra Mlcochova, Ines Ushiro Lumb, et al. Sars-cov-2 evolution during treatment of chronic infection. *Nature*, 592(7853):277–282, 2021.

[91] Emre Kiciman, Scott Counts, and Melissa Gasser. Using longitudinal social media analysis to understand the effects of early college alcohol use. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[92] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 EMNLP*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics.

[93] Peter Kollock et al. The economies of online cooperation: Gifts and public goods in cyberspace. *Communities in cyberspace*, 239, 1999.

[94] Susanne Kopf. "rewarding good creators": Corporate social media discourse on monetization schemes for content creators. *Social Media+ Society*, 2020.

[95] Neal Krause and Elaine Borawski-Clark. Social class differences in social support among older adults. *The Gerontologist*, 35(4):498–508, 1995.

[96] Yusuf Kurniawan, Sri Kusumo Habsari, and Ismi Dwi Astuti Nurhaeni. Selfie culture: Investigating the patterns and various expressions of dangerous selfies and the possibility of government's intervention. *The 2nd Journal of Government and Politics*, 324, 2013.

[97] Sheryl Perreault LaCoursiere. A theory of online social support. *Advances in Nursing Science*, 24(1):60–77, 2001.

[98] Phillippa Lally, Cornelia HM Van Jaarsveld, Henry WW Potts, and Jane Wardle. How are habits formed: Modelling habit formation in the real world. *European journal of social psychology*, 40(6):998–1009, 2010.

[99] Hemank Lamba, Momin M Malik, and Jurgen Pfeffer. A tempest in a teacup? analyzing firestorms on twitter. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2015.

[100] Hemank Lamba, Shashank Srikanth, Dheeraj Reddy Pailla, Shwetanshu Singh, Karandeep Singh Juneja, and Ponnurangam Kumaraguru. Driving the last mile: Characterizing and understanding distracted driving posts on social networks. In *Proceedings of the ICWSM*, volume 14, pages 393–404, 2020.

[101] Vasileios Lampos, Nikolaos Aletras, Jens K Geyti, Bin Zou, and Ingemar J Cox. Inferring the socioeconomic status of social media users based on behaviour and language. In *European conference on information retrieval*, pages 689–695. Springer, 2016.

[102] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[103] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Computational social science. *Science*, 323(5915):721–723, 2009.

[104] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, 2014.

[105] Mark R Leary, Lydia R Tchividijian, and Brook E Kraxberger. Self-presentation can be hazardous to your health: impression management and health risk. *Health Psychology*, 13(6):461, 1994.

[106] David S Lee and Thomas Lemieux. Regression discontinuity designs in economics. *Journal of economic literature*, 2010.

[107] Hyunsoo Lee, Uichin Lee, and Hwajung Hong. Commitment devices in online behavior change support systems. In *Proceedings of Asian CHI Symposium 2019: Emerging HCI Research Collection*, pages 105–113, 2019.

[108] Rebecca Lewis, Alice E Marwick, and William Clyde Partin. "we dissect stupidity and respond to it": Response videos and networked harassment on youtube. *American Behavioral Scientist*, 65(5):735–756, 2021.

[109] Fugui Li, Sihui Luo, Weiqi Mu, Yanmei Li, Liyuan Ye, Xueying Zheng, Bing Xu, Yu Ding, Ping Ling, Mingjie Zhou, et al. Effects of sources of social support and resilience on the mental health of different age groups during the covid-19 pandemic. *BMC psychiatry*, 21:1–14, 2021.

[110] Xitong Li. How does online reputation affect social media endorsements and product sales? evidence from regression discontinuity design. In *WISE*, 2013.

[111] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[112] Xin Jean Lim, AM Radzol, J Cheah, and Mun W Wong. The impact of social media influencers on purchase intention and the mediation effect of customer attitude. *Asian Journal of Business Research*, 7(2):19–36, 2017.

[113] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[114] Yuan Liu, Zhi Ning, Yu Chen, Ming Guo, Yingle Liu, Nirmal Kumar Gali, Li Sun, Yusen Duan, Jing Cai, Dane Westerdahl, et al. Aerodynamic analysis of sars-cov-2 in two wuhan hospitals. *Nature*, 582(7813):557–560, 2020.

[115] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

[116] John Lu, Sumati Sridhar, Ritika Pandey, Mohammad Al Hasan, and Georege Mohler. Investigate transitions into drug addiction through text mining of reddit data. In *KDD*, 2019.

[117] Wen Lu, Hang Wang, Yuxing Lin, and Li Li. Psychological status of medical workforce during the covid-19 pandemic: A cross-sectional study. *Psychiatry research*, 288:112936, 2020.

[118] Stephen Lyng. Edgework: A social psychological analysis of voluntary risk taking. *American Journal of Sociology*, 95(4):851–886, 1990.

[119] Danaja Maldeniya, Ceren Budak, Lionel P Robert Jr, and Daniel M Romero. Herding a deluge of good samaritans: How github projects respond to increased attention. In *Web Conference*, 2020.

[120] Momin M Malik and Jürgen Pfeffer. Identifying platform effects in social media data. In *Tenth International AAAI Conference on Web and Social Media*, 2016.

[121] Richard W Malott. I'll save the world from global warming—tomorrow: Using procrastination management to combat global warming. *The Behavior Analyst*, 33(2):179, 2010.

[122] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[123] Masoud Mazloom, Robert Rietveld, Stevan Rudinac, Marcel Worring, and Willemijn Van Dolen. Multimodal popularity prediction of brand-related social media posts. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 197–201, 2016.

[124] Justin McCrary. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics*, 142(2):698–714, 2008.

[125] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017.

[126] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[127] David L McMillen and James B Austin. Effect of positive feedback on compliance following transgression. *Psychonomic Science*, 1971.

[128] Philip J McParlane, Yashar Moshfeghi, and Joemon M Jose. " nobody comes here anymore, it's too crowded"; predicting image popularity on flickr. In *Proceedings of international conference on multimedia retrieval*, pages 385–391, 2014.

[129] Ali Mert Ertugrul, Yu-Ru Lin, and Tugba Taskaya-Temizel. Castnet: Community-attentive spatio-temporal networks for opioid overdose forecasting. *arXiv e-prints*, 2019.

[130] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[131] R.G. Miller. *Survival Analysis*. Wiley Classics Library. Wiley, 2011.

[132] Ravita Mishra. Entity resolution in online multiple social networks (@ facebook and linkedin). In *Emerging Technologies in Data Mining and Information Security*, pages 221–237. Springer, 2019.

[133] Richard L Moreland and John M Levine. Socialization in organizations and work groups. In *Groups at work*, pages 69–112. Psychology Press, 2014.

[134] Ahmet Anıl Müngen, Esra Gündoğan, and Mehmet Kaya. Identifying multiple social network accounts belonging to the same users. *Social Network Analysis and Mining*, 11(1):1–19, 2021.

[135] Bennet B Murdock Jr. The serial position effect of free recall. *Journal of experimental psychology*, 64(5):482, 1962.

[136] Ryan H Murphy. The rationality of literal tide pod consumption. *Journal of Bioeconomics*, 21(2):111–122, 2019.

[137] Aimee K Murray. The novel coronavirus covid-19 outbreak: global implications for antimicrobial resistance. *Frontiers in microbiology*, 11:1020, 2020.

[138] Vedant Nanda, Hemank Lamba, Divyansh Agarwal, Megha Arora, Niharika Sachdeva, and Ponnurangam Kumaraguru. Stop the killfies! using deep learning models to identify dangerous selfies. In *Companion Proceedings of the The Web Conference 2018*, pages 1341–1345, 2018.

[139] John A Nevin and Timothy A Shahan. Behavioral momentum theory: Equations and applications. *Journal of Applied Behavior Analysis*, 44(4):877–895, 2011.

[140] Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

[141] NIH. Overdose death rates. *National Institutes of Health*, 2021.

[142] John C Norcross and Dominic J Vangarelli. The resolution solution: Longitudinal examination of new year's change attempts. *Journal of substance abuse*, 1(2):127–134, 1988.

[143] Shrey Nougarahiya, Gaurav Shetty, and Dheeraj Mandloi. A review of e–commerce in india: The past, present, and the future. *Research Review International Journal of Multidisciplinary*, 6(03):12–22, 2021.

[144] Hans Oh, Caitlin Marinovich, Ravi Rajkumar, Megan Besecker, Sasha Zhou, Louis Jacob, Ai Koyanagi, and Lee Smith. Covid-19 dimensions are related to depression and anxiety among us college students: Findings from the healthy minds survey 2020. *Journal of affective disorders*, 292:270–275, 2021.

[145] Hüseyin Oktay, Brian J Taylor, and David D Jensen. Causal discovery in social media using quasi-experimental designs. In *Proceedings of the First Workshop on Social Media Analytics*, pages 1–9, 2010.

[146] Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. What to expect when the unexpected happens: Social media communications across crises. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 994–1009, 2015.

[147] Justin W Patchin and Sameer Hinduja. *Cyberbullying prevention and response: Expert perspectives*. Routledge, 2012.

[148] Cynthia M Pavett. Evaluation of the impact of feedback on performance and motivation. *Human Relations*, 36(7):641–654, 1983.

[149] Ana Luisa Pedrosa, Letícia Bitencourt, Ana Cláudia Fontoura Fróes, Maria Luíza Barreto Cazumbá, Ramon Gustavo Bernardino Campos, Stephanie Bruna Camilo Soares de Brito, and Ana Cristina Simões e Silva. Emotional, behavioral, and psychological impact of the covid-19 pandemic. *Frontiers in psychology*, 11:566212, 2020.

[150] Jonatas Pulz, Renan B Muller, Fabio Romero, André Meffe, Álvaro F Garcez Neto, and Aldo S Jesus. Fraud detection in low-voltage electricity consumers using socio-economic indicators and billing profile in smart grids. *CIRED-Open Access Proceedings Journal*, 2017(1):2300–2303, 2017.

[151] Daniel S Putler. Incorporating reference price effects into a theory of consumer choice. *Marketing science*, 11(3):287–309, 1992.

[152] Jianyin Qiu, Bin Shen, Min Zhao, Zhen Wang, Bin Xie, and Yifeng Xu. A nationwide survey of psychological distress among chinese people in the covid-19 epidemic: implications and policy recommendations. *General psychiatry*, 33(2), 2020.

[153] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[154] Ian Renfree, Daniel Harrison, Paul Marshall, Katarzyna Stawarz, and Anna Cox. Don't kick the habit: The role of dependency in habit formation apps. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, page 2932–2939, New York, NY, USA, 2016. Association for Computing Machinery.

[155] Howard Rheingold. *The virtual community: Finding commection in a computerized world*. Addison-Wesley Longman Publishing Co., Inc., 1993.

[156] Christopher Riederer, Yunsung Kim, Augustin Chaintreau, Nitish Korula, and Silvio Lattanzi. Linking users across domains with location data: Theory and validation. In *Proceedings of the 25th international conference on world wide web*, pages 707–719, 2016.

[157] Georgeanna FWB Robinson, Lisa S Schwartz, Linda A DiMeglio, Jasjit S Ahluwalia, and Janice L Gabrilove. Understanding career success and its contributing factors for clinical and translational investigators. *Academic medicine: journal of the Association of American Medical Colleges*, 91(4):570, 2016.

[158] Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 04 1983.

[159] Wolff-Michael Roth. Emotion at work: A contribution to third-generation cultural-historical activity theory. *Mind, culture, and activity*, 14(1-2):40–63, 2007.

[160] Lauren O Roussel and Derek E Bell. Tweens feel the burn:"salt and ice challenge" burns. *International journal of adolescent medicine and health*, 28(2):217–219, 2016.

[161] Deblina Roy, Sarvodaya Tripathy, Sujita Kumar Kar, Nivedita Sharma, Sudhir Kumar Verma, and Vikas Kaushal. Study of knowledge, attitude, anxiety & perceived mental healthcare need in indian population during covid-19 pandemic. *Asian journal of psychiatry*, 51:102083, 2020.

[162] Gretchen Rubin. *Better than before: Mastering the habits of our everyday lives*. Hachette UK, 2015.

[163] J Philippe Rushton and Goody Teachman. The effects of positive reinforcement, attributions, and punishment on model induced altruism in children. *Personality and Social Psychology Bulletin*, 1978.

[164] Derek Ruths and Jürgen Pfeffer. Social media for large studies of behavior. *Science*, 346(6213):1063–1064, 2014.

[165] Jorma Saari and Merja Näsänen. The effect of positive feedback on industrial housekeeping and accidents; a long-term study at a shipyard. *International Journal of Industrial Ergonomics*, 4(3):201–211, 1989.

[166] Koustuv Saha and Amit Sharma. Causal factors of effective psychosocial outcomes in online mental health communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 590–601, 2020.

[167] Koustuv Saha, Benjamin Sugar, John Torous, Bruno Abrahao, Emre Kıcıman, and Munmun De Choudhury. A social media study on the effects of psychiatric medication use. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 440–451, 2019.

[168] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *Journal of the American society for information science*, 41(4):288–297, 1990.

[169] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

[170] Hanan Shteingart, Tal Neiman, and Yonatan Loewenstein. The role of first impression in operant learning. *Journal of Experimental Psychology: General*, 142(2):476, 2013.

[171] Kai Shu, Suhang Wang, Jiliang Tang, Reza Zafarani, and Huan Liu. User identity linkage across online social networks: A review. *Acm Sigkdd Explorations Newsletter*, 18(2):5–17, 2017.

[172] Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

[173] Ruben Sipos, Arpita Ghosh, and Thorsten Joachims. Was this review helpful to you? it depends! context and voting patterns in online content. In *WWW*, 2014.

[174] Kimberly A Skarupski and Kharma C Foucher. Writing accountability groups (wags): A tool to help junior faculty members build sustainable writing habits. *The Journal of Faculty Development*, 32(3):47–54, 2018.

[175] BF Skinner. The behavior of animals: An experimental analysis. *New York: Appleton-Century Crofts*, 1938.

[176] Burrhus Frederic Skinner. The behavior of organisms: an experimental analysis. In *Appleton-Century*, 1938.

[177] Crystal R Smit, Laura Buijs, Thabo J van Woudenberg, Kirsten E Bevelander, and Moniek Buijzen. The impact of social media influencers on children's dietary behaviors. *Frontiers in psychology*, 10:2975, 2020.

[178] Dhanya Sridhar and Lise Getoor. Estimating causal effects of tone in online debates. *arXiv preprint arXiv:1906.04177*, 2019.

[179] Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

[180] Acar Tamersoy, Duen Horng Chau, and Munmun De Choudhury. Analysis of smoking and drinking relapse in an online community. In *Proceedings of the 2017 international conference on digital health*, pages 33–42, 2017.

[181] Sean J Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.

[182] Steven Taylor, Caeleigh A Landry, Michelle M Paluszek, Thomas A Fergus, Dean McKay, and Gordon JG Asmundson. Covid stress syndrome: Concept, structure, and correlates. *Depression and anxiety*, 37(8):706–714, 2020.

[183] Sharon Tennyson. Economic institutions and individual ethics: A study of consumer attitudes toward insurance fraud. *Journal of Economic Behavior & Organization*, 32(2):247–265, 1997.

[184] Donald L Thistlethwaite and Donald T Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, 1960.

[185] Edward L Thorndike. Animal intelligence. *Nature*, 58(1504):390–390, 1898.

[186] Hsien-Tung Tsai and Peiyu Pai. Explaining members' proactive participation in virtual communities. *International Journal of Human-Computer Studies*, 71(4):475–491, 2013.

[187] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131, 1974.

[188] Ebru Uzunoğlu and Sema Misci Kip. Brand communication through digital influencers: Leveraging blogger engagement. *International journal of information management*, 2014.

[189] Marat Valiev, Bogdan Vasilescu, and James Herbsleb. Ecosystem-level determinants of sustained activity in open-source projects: A case study of the pypi ecosystem. In *ESEC/FSE*, 2018.

[190] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[191] I Veissi. *Influencer Marketing on Instagram. Yayınlanmış Lisans Tezi, Haaga-Helia University Of Applied Sciences*. PhD thesis, Bachelor's Thesis Degree Programme in International business, 2017.

[192] Ruvanee P Vilhauer. Perceived benefits of online support groups for women with metastatic breast cancer. *Women & health*, 49(5):381–404, 2009.

[193] Nina Vindegaard and Michael Eriksen Benros. Covid-19 pandemic and mental health consequences: Systematic review of the current evidence. *Brain, behavior, and immunity*, 89:531–542, 2020.

[194] Hanna Wallach. Computational social science != computer science + social data. *Commun. ACM*, 61(3):42–44, feb 2018.

[195] Huandong Wang, Yong Li, Gang Wang, and Depeng Jin. Linking multiple user identities of multiple services from massive mobility traces. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(4):1–28, 2021.

[196] Ke Wang, Mohit Bansal, and Jan-Michael Frahm. Retweet wars: Tweet popularity prediction via dynamic multimodal regression. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1842–1851. IEEE, 2018.

[197] Yi-Chia Wang, Robert Kraut, and John M Levine. To stay or leave? the relationship of emotional and informational support to commitment in online health support groups. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 833–842, 2012.

[198] Galen Weld, Peter West, Maria Glenski, David Arbour, Ryan Rossi, and Tim Althoff. Adjusting for confounders with text: Challenges and an empirical evaluation framework for causal inference, 2021.

[199] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. Virality prediction and community structure in social networks. *Scientific reports*, 2013.

[200] Yang Wu, Laura E Schulz, Michael C Frank, and Hyowon Gweon. Emotion as information in early social learning. *Current Directions in Psychological Science*, 30(6):468–475, 2021.

[201] Chang Yan, Zhuo Luo, Wen Li, Xue Li, Robert Dallmann, Hiroshi Kurihara, Yi-Fang Li, and Rong-Rong He. Disturbed yin–yang balance: Stress increases the susceptibility to primary and recurrent infections of herpes simplex virus type 1. *Acta Pharmaceutica Sinica B*, 10(3):383–398, 2020.

[202] Reza Zafarani and Huan Liu. Connecting users across social media sites: a behavioral-modeling approach. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 41–49, 2013.

[203] Ark Fangzhou Zhang, Danielle Livneh, Ceren Budak, Lionel Robert, and Daniel Romero. Shocking the crowd: The effect of censorship shocks on chinese wikipedia. In *ICWSM*, 2017.

[204] Ark Fangzhou Zhang, Ruihan Wang, Eric Blohm, Ceren Budak, Lionel P Robert Jr, and Daniel M Romero. Participation of new editors after times of shock on wikipedia. In *ICWSM*, 2019.

[205] Jason Shuo Zhang, Brian Keegan, Qin Lv, and Chenhao Tan. Understanding the diverging user trajectories in highly-related online communities during the covid-19 pandemic. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 888–899, 2021.

[206] Yingfei Zhang and Zheng Feei Ma. Impact of the covid-19 pandemic on mental health and quality of life among local residents in liaoning province, china: A cross-sectional study. *International journal of environmental research and public health*, 17(7):2381, 2020.

[207] Qingyuan Zhao, Murat A Erdogdu, Hera Y He, Anand Rajaraman, and Jure Leskovec. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1513–1522, 2015.

[208] Shuang-Jiang Zhou, Li-Gang Zhang, Lei-Lei Wang, Zhao-Chang Guo, Jing-Qi Wang, Jin-Cheng Chen, Mei Liu, Xi Chen, and Jing-Xu Chen. Prevalence and socio-demographic correlates of psychological health problems in chinese adolescents during the outbreak of covid-19. *European child & adolescent psychiatry*, 29:749–758, 2020.