Privacy and security implications of active participation in online social networks: An information diffusion based approach to modeling user behavioral patterns

by

Abiola Afolake Osho

B.S., Olabisi Onabanjo University, Nigeria, 2010

M.S., University of Ibadan, Nigeria, 2016

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Computer Science
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2022

Abstract

Due to their wide reach, oversimplified conversations, and ability to provide quick blasts of information, online social networks (OSNs) have become an avenue where users connect to share information, news, and events around the world. Third-party recommender systems, spammers, and manipulators can learn a user's online behavior and interests through their connections and interactions because users would often leave breadcrumbs on their interests and personality through activities - like, comments, share, repost, etc. With access to user data and sophisticated learning models, manipulation through inferential attack are now easier to achieve, causing users to struggle with privacy loss as a consequence of their participation in OSNs. Given that some users have a higher propensity for disclosure than others, a one-size-fits-all technique for limiting manipulation and privacy loss proves insufficient. In the search to find a balance between privacy preservation and social influence, we propose a solution that uses the information spread behavior as a basis for estimating the possible exposure of users to abuse and misinformation in the network. We focus on the information spread behavior and explore how it can be used for manipulation purposes. We explore a microscopic follower-followee relationship to show how direct interactions can contribute to targeted manipulation based on inferential attack. The proposed model utilizes the user's probability of engaging with a post as a way to measure their sensitivity to privacy loss. With this knowledge, the user can then implement a privacy preservation mechanism to minimize their privacy loss by adding noise to their profile to muddle up an attacker's opinion of them. The result from experiments on real-world Twitter data showed that even though there will be costs to participating in OSNs, these costs can be minimized relative to the disclosure threshold set by the user as their maximum privacy loss. Additionally, we report attributes that can be tweaked to minimize the user's exposure.

Privacy and security implications of active participation in online social networks: An information diffusion based approach to modeling user behavioral patterns

by

Abiola Afolake Osho

B.S., Olabisi Onabanjo University, Nigeria, 2010

M.S., University of Ibadan, Nigeria, 2016

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Computer Science
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2022

Approved by:

Major Professor
George Amariucai, PhD

Copyright

Abstract

Due to their wide reach, oversimplified conversations, and ability to provide quick blasts of information, online social networks (OSNs) have become an avenue where users connect to share information, news, and events around the world. Third-party recommender systems, spammers, and manipulators can learn a user's online behavior and interests through their connections and interactions because users would often leave breadcrumbs on their interests and personality through activities - like, comments, share, repost, etc. With access to user data and sophisticated learning models, manipulation through inferential attack are now easier to achieve, causing users to struggle with privacy loss as a consequence of their participation in OSNs. Given that some users have a higher propensity for disclosure than others, a one-size-fits-all technique for limiting manipulation and privacy loss proves insufficient. In the search to find a balance between privacy preservation and social influence, we propose a solution that uses the information spread behavior as a basis for estimating the possible exposure of users to abuse and misinformation in the network. We focus on the information spread behavior and explore how it can be used for manipulation purposes. We explore a microscopic follower-followee relationship to show how direct interactions can contribute to targeted manipulation based on inferential attack. The proposed model utilizes the user's probability of engaging with a post as a way to measure their sensitivity to privacy loss. With this knowledge, the user can then implement a privacy preservation mechanism to minimize their privacy loss by adding noise to their profile to muddle up an attacker's opinion of them. The result from experiments on real-world Twitter data showed that even though there will be costs to participating in OSNs, these costs can be minimized relative to the disclosure threshold set by the user as their maximum privacy loss. Additionally, we report attributes that can be tweaked to minimize the user's exposure.

Table of Contents

List of Figures x

List of Tables xi

Acknowledgements xiii

Dedication xiv

Preface xv

1 Introduction 1

1.1 Online Social Networks 1

1.2 Security Challenges Associated with OSNs 2

1.3 Purpose of the Research 3

1.4 Concepts and Terms 5

1.5 Data Collection and Ethics 6

1.6 Contributions 7

1.7 Overview 9

2 Literature Study 10

2.1 Information Diffusion and Prediction Models 10

2.1.1 Feature Selection for Information Propagation 11

2.2 Rumor Propagation 12

2.2.1 Feature-Based Rumor Detection 13

2.2.2 Crowdsourcing Techniques for Misinformation 14

2.3 Models for Abusive Behavior Detection . . . 15
2.3.1 Crowdsourcing Techniques for Offensive Behavior Identification . . . 16
2.4 Privacy and Information Leakage . . . 17

3 MIDMod-OSN: A Microscopic-level Information Diffusion Model for Online Social Networks[1] . . . 20
3.1 Overview . . . 21
3.2 Dataset Description . . . 22
3.3 Learning and Feature Estimation Models . . . 24
3.4 Model Evaluation . . . 25
3.5 The Diffusion Prediction Experiment . . . 25
3.5.1 Experimental setup . . . 27
3.5.2 Diffusion prediction model . . . 28
3.5.3 Cross testing between models . . . 30
3.6 Feature Selection Framework . . . 31
3.7 Time to Tweet . . . 33
3.8 Crowdsourcing for Early Trending Topic Detection . . . 33
3.8.1 The early detection of trending topics . . . 34
3.9 Using MIDMod-OSN for Crowdsourcing . . . 34
3.9.1 Experiment design and results . . . 35

4 An Implicit Crowdsourcing Approach to Rumor Identification in Online Social Networks[2] . . . 37
4.1 Overview . . . 38
4.2 Features for Rumor Propagation and Identification . . . 39
4.2.1 Network-based features . . . 40
4.2.2 Interaction-based Features . . . 40
4.2.3 Message-based Features . . . 41

4.3 Experiment Setup . . . 42
4.4 Data Collection . . . 43
4.5 Prediction Models . . . 45
4.5.1 Predicting credibility of posts . . . 45
4.5.2 Predicting rumor propagation . . . 46
4.6 Baseline . . . 46
4.6.1 Emotion-based . . . 46
4.6.2 Implicit-link . . . 47
4.6.3 User-behavior . . . 47
4.7 Experimental Results . . . 47
4.7.1 Predicting credibility by implicit crowdsourcing . . . 47
4.7.2 Features analysis for rumor propagation . . . 48
4.7.3 Predicting Rumor Propagation . . . 50

5 Implicit Crowdsourcing for Identifying Abusive Behavior in Online Social Networks[3] 51
5.1 Overview . . . 52
5.2 Model and Method . . . 53
5.3 Data Description . . . 53
5.4 Task 1: Implicit Crowdsourcing for Predicting the Abuse Level of a Tweet . 54
5.5 Task 2: Estimating Features Contributing to Abusive Tweet Propagation . . 55
5.5.1 Task 2.1: Evaluating a user's tendency to post or reactive to abusive posts . . . 56
5.5.2 Task 2.2: Identifying features for abusive post propagation . . . 57
5.6 Results . . . 57
5.6.1 Predicting Abuse Level using Implicit Crowdsourcing . . . 57
5.6.2 Features for Abusive Behavior Propagation . . . 58

6 Heuristic Gradient Optimization Approach to Controlling Susceptibility to Manipulation in Online Social Networks . . . 61
6.1 Model . . . 63
6.1.1 Dataset . . . 63
6.1.2 Attributes . . . 63
6.1.3 Learning Model . . . 64
6.2 Experiment . . . 67
6.2.1 Gaussian Process Classification Model . . . 68
6.2.2 Gradient Optimization . . . 69
6.2.3 Optimization over Multiple Connections . . . 70
6.2.4 Constraints over Social Influence . . . 72
6.3 Results . . . 73
6.3.1 GPC Model & Calibration . . . 73
6.3.2 Gradient Optimization and Manipulation . . . 73
6.3.3 Attribute Disclosure . . . 75
6.3.4 Optimization over Multiple Connections . . . 77
6.3.5 Constraints over Social Influence . . . 80

7 Conclusion and Recommendations . . . 82
7.1 Summary and Conclusions . . . 82
7.2 Limitations and Future Work . . . 84

Bibliography . . . 86

List of Figures

1.1 Implication of active participation in online social networks. 3

1.2 Figure describing the blueprint of the models used throughout this work. . . 5

4.1 An illustration of experiment setup for the credibility prediction task. 43

6.1 Probability of response when receiver responds to random post and when sender and message attributes are optimally chosen to manipulate receiver. . 74

6.2 Probability of response when receiver responds to random post and when receiver optimize their attributes while being targeted. 75

6.3 Estimated manipulation gain when receiver does nothing but sender and message attributes are optimized compared with manipulation gain when receiver optimizes their attributes while being targeted. 76

6.4 Probability of response when sender optimize attributes over multiple followers compared with when receiver responds to random post and when sender and message attributes are optimally chosen to manipulate receiver. 78

6.5 Probability of response when receiver optimize attributes by considering multiple connections compared with when receiver responds to random post and when receiver optimize their attributes while being targeted. 79

6.6 Estimated manipulation gain when optimization is done over multiple connections. 79

List of Tables

3.1 Data distribution. 23

3.2 Network features extracted for each user to serve as input variables to the learning model. 25

3.3 Interaction features extracted for each user to serve as input variables to the learning model. 26

3.4 Semantic features extracted for each user to serve as input variables to the learning model. 26

3.5 Temporal features extracted for each user to serve as input variables to the learning model. 27

3.6 Performance evaluation of MIDMod-OSN in predicting diffusion of posts from different event types. 28

3.7 Prediction accuracy using proposed model with different number of features and state-of-the-art. 29

3.8 Ranking of the top 15 optimal features that should be maximized for maximum diffusion or minimized for containment. 30

3.9 Performance evaluation of MIDMod-OSN in predicting the trending status of a message without counting reactions. 36

4.1 Network-based features. 40

4.2 Interaction-based features. 41

4.3 Message-based Features. 42

4.4 Topics identified from Snopes, along with the associating keywords used in querying the Twitter seach API. 44

4.5 Model performance for predicting credibility of a tweet using crowdsourcing techniques. 48
4.6 Top 20 features for efficient diffusion prediction of *True* and *False* posts selected using Random Forest classifiers. 49
4.7 Model performance for predicting diffusion of *True* and *False* posts of a post. 50
4.8 Model performance for inter-topic, inter-credibility diffusion prediction. . . . 50

5.1 Model performance in abuse level of a post. 58
5.2 Top 10 features for predicting propagation of *abusive* and *normal* posts selected using Random Forest classifiers. 60

6.1 Data Distribution. 63
6.2 Summary of Notation. 64
6.3 Attribute Description. 65
6.4 Expected Calibration Error (ECE) reported for GPC model using various calibration techniques. 73
6.5 Observed changes in disclosed attributes post-optimization. 77
6.6 Observed changes in disclosed attributes with multi-connection targeting post-optimization. 80

Acknowledgments

I express my sincere appreciation to all those who have helped me throughout the program.

I owe a tremendous debt of gratitude to my advisor, Dr George Amariucai, whose talents extend well beyond his field of expertise into the realm of guidance, motivation, mentorship and leadership. Your outstanding impressions will last a lifetime.

I would also like to thank the rest of my committee, Drs Doina Caragea, Eugene Vasserman, and Raluca Cozma, who have all given me sound advice and help at one time or another during this process, and who did a wonderful job providing me feedback on this document.

I also want to thank Professor Shuangqing Wei of Louisiana State University School of Electrical Engineering and Computer Science, for spending time helping through a critical part of this work.

Special thanks to Sheryl Cornell, Theresa Hogenkamp and Wynne Reichart for always answering my questions and keeping me straight with respect to administrative issues.

Most importantly, I thank my husband - Kehinde Osho, and children - Oluwasemilore and Oluwapelumi, who were very supporting and understanding throughout the process.

Finally, I'd like to thank my siblings - Omowunmi, Olajide, Omorinola, Damilola; Adedolapo Okanlawon; and everyone at RCCG DGC for being the *village* I very much needed.

Dedication

I will like to dedicate this work to my parents, Bamidele and Tinuola Arise, whose love and strength has been truly inspirational.

Preface

A version of Chapter 3 has been published [Abiola Osho, Colin Goodman, and George Amariucai."MIDMod-OSN: a microscopic-level information diffusion model for online social networks." International Conference on Computational Data and Social Networks. Springer, Cham, 2020.] I was the lead investigator, responsible for all major areas of concept formation, data collection and analysis, as well as manuscript composition. Colin Goodman was involved in the early stages of concept formation and contributed to data collection and processing. George Amariucai was the supervisory author on this project and was involved throughout the project in concept formation and manuscript composition/edits.

A version of Chapter 4 has been published [Abiola Osho, Caden Waters, and George Amariucai, "An Implicit Crowdsourcing Approach to Rumor Identification in Online Social Networks," 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2020, pp. 174-182, doi: 10.1109/ASONAM49781.2020.9381339.] I was the lead investigator, responsible for all major areas of concept formation, data collection and analysis, as well as manuscript composition. Caden Waters was involved in the early stages of concept formation and contributed to data processing. George Amariucai was the supervisory author on this project and was involved throughout the project in concept formation and manuscript composition/edits.

The study in Chapter 5 is a response to the ICWSM-2020 14th International Conference On Web And Social Media (ICWSM) Data Challenge in 2020 [Abiola Osho, Ethan Tucker, and George Amariucai. "Implicit crowdsourcing for identifying abusive behavior in online social networks."] I was the lead investigator, responsible for all major areas of concept formation, data collection and analysis, as well as manuscript composition. Ethan Tucker was involved in contributed to data collection and processing. George Amariucai was the supervisory author on this project and was involved throughout the project in concept formation and manuscript composition/edits.

Chapter 1

Introduction

1.1 Online Social Networks

Online social networks (OSNs) provide a medium where people build social relationships with other people who share similar personal or career content, interests, activities, backgrounds, or real-life connections. These social networking sites allow users to share ideas, photos, videos, posts and inform others about online or real-world activities and events with people within their social network.[4] Over the last decade, social media have grown from being an avenue for just social connections to the go-to media for the dissemination of information, news, and events around the world. With more than 4.2 billion active users in February 2022, OSN continues to experience growth at an expeditious rate.[5]

With the increasing popularity and ease of access, OSN relationships are seen to transgress location, culture, beliefs, personality traits, as well as many barriers limiting in-person social interactions. Similar to many cyber security challenges, users have to deal with the task of being potentially exposed to the billions of users on OSNs. Sifting through the many posts that users encounter daily can prove daunting to the point where it becomes difficult to identify what is real or not. Due to their wide reach, oversimplified conversations, and ability to provide quick blasts of information, OSNs have become an avenue for the spread of rumors, manipulation of public opinions, among other things. With the current political

and economic climate around the world, we continue to witness individuals, organizations, and governments exploit OSNs users in 280 characters or less.

1.2 Security Challenges Associated with OSNs

While the dissemination of accurate information may protect the general public and potentially save lives, the spread of false or inaccurate information proves detrimental to public health and safety in many contexts. During a time when targeted manipulation through inferential attacks is an increasingly serious problem, it is important to study the creation and spread of information, opinion formation, and how these affect the privacy of a user participating in OSNs. Through post engagement, users sometimes carve a niche for themselves by establishing interest and influence in certain areas. Influencers on OSNs build a reputation for their knowledge and expertise on a specific topic by making regular posts about that topic. They generate large followings of enthusiastic, engaged people who pay close attention to their points of view. These influencers have the power to affect the purchasing decisions of others because of their *authority* on the said topic. Organizations looking to gain public engagement look to these influencers who have built social relationship assets with which brands can collaborate to achieve their marketing objectives.

While social interaction and influence can be beneficial to users, they can also be detrimental to their privacy. This is because the activities they generate can be used to learn other latent (and sometimes sensitive) information, like their beliefs and orientations. Studies have shown that the online disclosure of certain personality traits can influence the hiring decisions of some U.S. employers who introduce biases through personal information posted by job candidates on social media sites.[6] As more users are trying to leverage social media to create a brand value and become more influential, spammers are luring such users to help manipulate their social reputation with the help of paid services or collusion networks.[7] Even though some users keep their networks limited to friends and associates, Wilcox et al.[8] pointed out that focusing on close friends may cause a momentary increase in self-esteem, leading those focused on strong ties to display less self-control on OSNs.

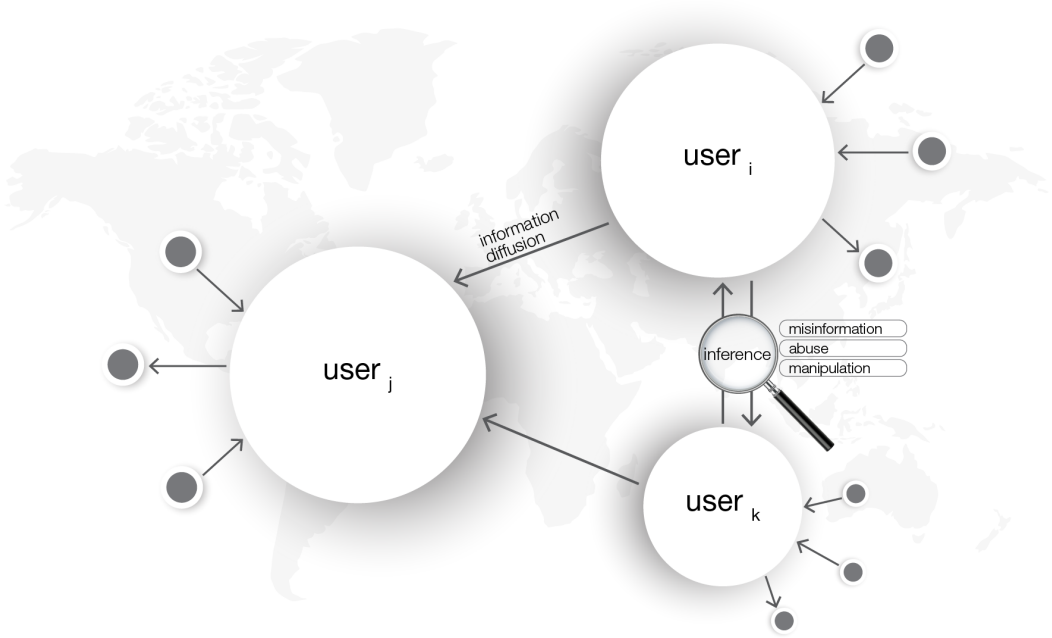

Figure 1.1: *Implication of active participation in online social networks.*

Although there are many security challenges originating from interactions between users and the posts they create, we examine three of these security problems: misinformation, abuse, and manipulation. We explore these problems by looking at how information spread between users can contribute to these security challenges, Figure 1.1.

1.3 Purpose of the Research

For users looking to maintain some level of anonymity by keeping their online activities separate from their personalities, their online behavior can still be mined by manipulators, spammers, and third-party recommender systems to suggest content to them. With the help of data mining and learning algorithms, these exploiters build profiles around the user's interactions, networks, topics of interest, etc., and use them to their benefit. The users are then bombarded with content (stories, news, and ads) that perfectly match their online behavior and profiles while limiting users' exposure to diverse content and information. Such content could also hide ulterior motives, like spreading rumors and hate throughout

the network.

The continued querying of OSNs to mine user behavior for manipulation purposes, targeted advertising, behavioral analysis, etc., has caused the repeated battle between privacy preservation and social influence, leaving users with the option to either limit their participation or accept that their information will continue to be mined. OSN creators looking to combat this challenge have provided privacy preservation techniques that require users to make their accounts private, making their contents restricted to only their followers, but this method does not address the influence challenge due to how restrictive the measures are.

The subjective nature of privacy brings about the challenge of having an approach that measures the degree of manipulation gain based on a user's preferences and propensity for disclosure, as what one user deems private might not be deemed private by another user. With such solutions, a user can make the decision whether there is a need for additional privacy protection measures based on their privacy needs. The search to find a balance between privacy preservation and social influence leads us to ask if users can be given control over their own privacy. This research aims to directly address this question by allowing the users to examine their likelihood for manipulation based on their social interactions, giving them insight into their degree of exposure so that they can choose how much protection they need based on their needs. User and message data collected from the Twitter API is used throughout this work.

The solution proposed first examines how a user's post will spread in the network based on the user's public profile, post creation and engagement, followers, and friends. This information serve as a basis for estimating possible exposure of users to abuse and misinformation in the network. Emphasis is placed on the information spreading process and exploring how it can be used for manipulation purposes. This gives insight into the properties that will make a user influential enough to the point where they can manipulate others. Additionally, the microscopic follower-followee relationship is explored to show how direct interactions can contribute to targeted manipulation based on inferential attack. Overall, the proposed model suggests a privacy preservation mechanism where the user can minimize their manipulation gain by adding noise to their profile in order to muddle up an attacker's opinion of them.

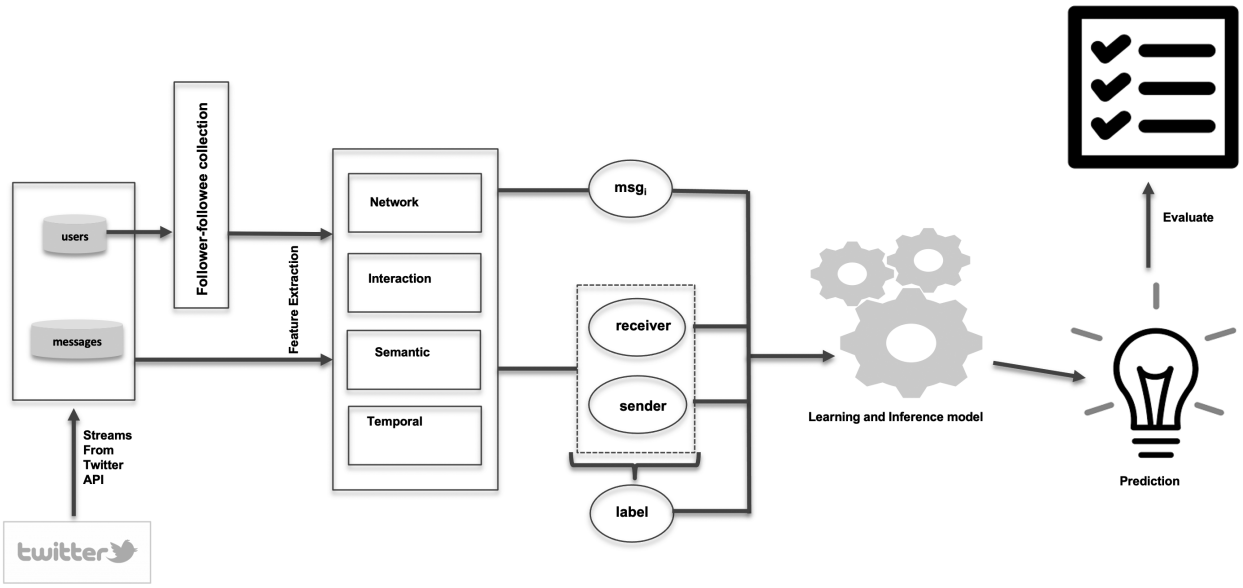

Figure 1.2: *Figure describing the blueprint of the models used throughout this work.*

The OSN user is given the opportunity to measure their manipulation gain when they choose to interact as they usually would and compare it with their manipulation gain when they add a layer of security to it. The privacy preservation technique is then constrained on the utility of social influence to ensure that there is a balance between the user's privacy and the gratification they derive from interacting with the network.

1.4 Concepts and Terms

Some concepts used throughout this work include:

- Sender: the user spreading a tweet to their followers. The tweet could be in form of a tweet the user creates, retweet, shared tweet, or a quoted tweet.
- Receiver: the user that follows the sender of a tweet and can view the sender's tweets in their timeline.
- Informative events: associated with topics relating to general knowledge and which have not attained viral status.

- Trending events: associated with topics that can be described as viral, breaking news, hot topics, or crises.
- Diffusion: a tweet is said to diffuse between a pair of users if the receiver engages with the sender's post in form of a retweet, like, favorite, quote.
- Latent attributes: are not directly observed from the collected data but are inferred from other directly observable attributes using mathematical models. Examples include sentiment, intent, and emotions of tweets.
- Misinformation: false or inaccurate information intended to deceive.
- Cyber abuse: online behavior that threatens, intimidate, harass, harm, or humiliate a person.
- Manipulation: the ability to unfairly control or influence a user's interactions with a post or topic.

1.5 Data Collection and Ethics

For the purpose of this research, Twitter user and tweet data are collected through the Twitter API[9] made available to developers and researchers. These are publicly available tweets and user profiles that can be crawled by either collecting tweets and associated user information or by collection tweets of a particular user. We use both methods by building a crawler based on Tweepy,[10] an open source library for Twitter API. For replicability, it should be noted that Twitter levies a 900 requests per 15 minutes rate limit and any requests above that would lead to an error. Additionally, there is a 3200 tweet limit per user, meaning that a crawler can only have access to the last 3200 tweets posted by a user. To address these restrictions, Tweepy provides modules that can be implemented into a crawler to allow for a rest period once the rate limit is reached and also limit data collection to only 3200 tweets per user.

For each dataset used, an instance of the data includes attributes of the sender, attributes of the receiver, and attributes of the message as the input variable. This input variable is then associated with a label defined by the learning task, Figure 1.2. The data is split into training and test sets but the challenge here is that we could have many instances of a sender and receiver pairs based on the spread of different messages. This could lead to overlap in the training and test datasets. It should be noted that for accurate evaluation of model capabilities, this overlap should be addressed during the data processing stage.

We maintain the privacy and confidentiality of Twitter users and their posts by adhering to Twitter developer agreement and policy document.[11] We do not share the dataset used in this work but only made the crawler available on GitHub.

1.6 Contributions

The following list describes the new contributions of this work:

- We build a tool to crawl the Twitter Search API using user IDs and results stored as JSON in a database encoding the key-value pairs with named attributes and associated values.
- The crawler is made publicly available on GitHub.
- In this work, we present a node-to-node feature analysis model to learn the diffusion process by combining a set of network, interaction, semantic and temporal features.
- We fit a stochastic model to the relationship between these features and the probability of diffusion.
- This research identifies the optimal subset of features needed to efficiently predict information diffusion in Twitter events.
- We draw conclusions regarding the time to tweet, as well as the most important user attributes that contribute to achieving maximum retweetability, and in turn maximum diffusion, in the network.

- Additionally, we demonstrate the value of crowdsourcing to predicting the virality of a post before this would even be possible by counting user reactions.
- We investigate credibility prediction by exploring rumor propagation founded on microscopic-level misinformation spread.
- We propose a model that predicts if a message is *True* or *False* by observing the latent attributes of the message, along with users and their reactions over the network.
- We examine the contribution of individual users to rumor propagation in OSNs, by investigating features of users (both the post sharer and receiver) and how these features influence the propagation of rumor.
- In this research, we investigate abusive behavior prediction by exploring abuse propagation founded on microscopic-level information spread.
- We then propose a model that predicts the abuse level associated with a tweet by observing the latent attributes of the message, along with those of the users, and their reactions over the network.
- We evaluate the role of user and message features in detecting the abuse level of a post, by measuring the contributions of individual users and their posts to the spread of abusive posts in OSNs.
- We fit a stochastic model to estimate a user's susceptibility to manipulation through inferential attack.
- We implement a privacy preservation mechanism controlled by the user based on their propensity for disclosure.
- We provide a metric for estimating manipulation gain based on implemented protection mechanism.
- We draw conclusions regarding the degree of change in disclosed attributes to minimize manipulation.

- We constrain manipulation gain over the social influence of a user in their network.

1.7 Overview

The remainder of this document is laid out as follows. Following a review of the appropriate literature in Chapter 2, Chapter 3 examines how different types of Twitter events impact the node-to-node influence dynamics associated with information spread. Chapter 4 presents a new paradigm for credibility prediction introducing a model that predicts if a message is *True* or *False* by observing the latent attributes of the message, along with those of the users interacting with it, and their reactions to the message. Chapter 5 presents a model to identify abusive posts through a detection mechanism that simply observes the natural interaction between users encountering the messages. Chapter 6 presents a model that allows the user to adjust their online persona to limit their susceptibility to manipulation based on their preferred disclosure threshold. Finally, Chapter 7 offers concluding remarks and several recommendations for future research.

Chapter 2

Literature Study

In considering the solution to privacy and security challenges associated with active participation in social networks, a discussion of the existing literature is warranted. Specifically, we address three relevant classes of security problems in this chapter; namely, misinformation, abuse, and manipulation. The solutions proposed in this research to address these topics are examined from the information diffusion point of view, and as such there is also a need to review the works done in this area. For each class of problem, models and features will be explored, relative to the type of problems we want to solve.

2.1 Information Diffusion and Prediction Models

The information diffusion process can be observed through the diffusion graph and rate of adoption of the information by the nodes in the graph. The diffusion graph shows influence in the network, which is important for viral marketing,[12–14] crisis communication,[15] and retweetability.[16] Generally, influence analysis models have focused on relationship strength based on profile similarity and interaction activity,[17] and the mechanisms responsible for network homogeneity.[18] Identifying influential users has been found to be useful when trying to select seed nodes in the community that will maximize the spread of information across the networks. For instance, Pei et al.[19] worked on finding the best spreaders in dissimilar

social platforms when the complete global network structure is unavailable. The work of Yingcai et al.[20] observed that (1) the authority of an influential user on social media which can be used to change the opinions of other users and (2) opinion similarity factors where users tend to accept an opinion that is similar to their own, are important factors when selecting seed nodes for information spread.

Predictive models like the independent cascade (IC) model[21] make use of submodular functions to approximate the selection of most influential nodes where people observe the choices of others while ignoring their personal knowledge. The linear threshold model (LT) described by Granovetter[22] deals with binary decisions where a node has distinctly mutually exclusive alternatives and an inactive node is activated by its already activated neighbors, if the sum of influence degrees exceeds its own threshold. Asynchronous IC and LT (AsIC and AsLT respectively) were defined by Saito et al.[23;24] and introduced a time delay parameter before a parent node can activate an inactive child node. In AsIC, if the child node remains inactive after the specified period δ, the parent node is given only a single chance to attempt activating the child node to eliminate the likelihood of a single node being simultaneously activated by multiple parent nodes. In AsLT, a node decides when to receive the information once the activation condition has been satisfied. Some other studies like that of Wang et al.[25] propose a model based on Partial Differential Equations (PDE) by introducing a diffusive logistic model to predict temporal and spatial patterns in Diggs, an online social news aggregation site. A Linear Influence Model was developed in the work of Yand and Leskovec,[26] focused on modeling the global influence of a node on the rate of diffusion through the implicit network by estimating an influence function to quantify the number of successive adoptions attributed to a node over time.

2.1.1 Feature Selection for Information Propagation

Guille et al.[27] introduced a variant of the AsIC model called the T-BAsIC framework that assigns a fixed value for a real time-dependent function for each link, without fixing the diffusion probability. The model relies on three different dimensions to compute the diffusion

probability: social, semantic, and time. The model was designed to predict the daily volume of tweets for a topic and variations in popularity of topics over time. They proceeded by identifying 2 types of users: (1) transmitters that pass along information and (2) stiflers that become dead-ends for information travel, with stiflers growing with time for a given topic. In the work of Ferrara et al.[28] the authors leverage a mixture of metadata, network, and temporal features in detecting users spreading extremist ideology and predict content adopters and interaction reciprocity in social media. They adopted logistic regression and random forests learning models with 52 features observed from Twitter data of over 25,000 accounts labeled as supportive of the Islamic State. Given the temporal relevance of tweets, Spasojevic et al.[29] propose finding the best times for a user to post on social networks in order to maximize the probability of audience response. They hypothesize that the probability that an audience member reacts to a message depends on factors such as his daily and weekly behavior patterns, his location and timezone, and the volume of other messages competing for his attention.

2.2 Rumor Propagation

Research in political science explored the differential diffusion of true, false, and mixed (partially true, partially false) news stories on Twitter using the fact-checked rumor cascades that spread on Twitter over a 12-year period. In the work of Vosoughi et al.[30] it was observed that falsehood diffused faster, farther, deeper and more broadly than truth in all categories of information, with a more noticeable impact in false political news. The study also observed that false news are often more novel, inspiring fear, disgust and surprise in replies while true stories inspired anticipation, sadness, joy and trust. In like manner, Grinberg et al.[31] examined the spread of fake news on Twitter during the 2016 U.S. presidential election and observed that the exposure to fake news sources was extremely concentrated with seven fake news sources accounting for more than 50% of fake news exposures. The study showed that political affinity was associated with the sharing of content from fake news sources and that the sharing of content from fake news sources was positively associated with tweeting

about politics, and exposure to fake news sources. Computer scientists like the Friggeri et al.[32] examined the spread of rumors on Facebook and found that rumor cascades run deeper in the social networks. When rumor debunking posts are available, Takahashi et al.[33] and Friggeri et al.[32] reported that users will either delete a post, if it is confirmed to be rumor, or share otherwise. Additionally, Abdullah et al.[34] revealed that users spread the messages that they deem important and mostly retweet messages because of the need to retweet interesting tweet content or tweet creators.

To classify conversations within their formative stages, Sampson et al.[35] proposed a rumor classification method to leverage implicit links to classify emergent conversations when very little conversation data is available. They used implicit links formed with hashtag and web links to establish similarity between otherwise unlinked conversations. Wu and Liu[36] focused on the diffusion of information by inferring the embedding of social media users with social network structures; and utilize an LSTM-RNN model to represent and classify propagation pathways of a message.

2.2.1 Feature-Based Rumor Detection

To demonstrate the importance of features for rumor detection, Castillo et al.[37] extracted 68 features from tweets and categorized them as (1) message-based which considers characteristics of the tweet content, such as length of post, presence of exclamation, number of positive/negative sentiment words, (2) user-based which considers characteristics of Twitter users, such as registration age, number of followers, number of friends, and number of user posted tweets, (3) topic-based which aggregates the message-based and user-based features, and (4) propagation-based which considers characteristics related to the propagation tree that can be built from the retweet of the post. Subsequently, Liang et al.[38] explored rumor identification using users' behavior to differentiate between normal authors and rumormongers. Furthermore, Wu et al.[39] introduced the propagation tree, and used a random walk graph-kernel based hybrid SVM classifier to capture the high-order propagation patterns in addition to topic and sentiment features for rumor detection in Sina Weibo. In the work of

Yang et al.[40] the authors proposed two new features: (1) a client-based feature referring to mode of access – whether mobile or non-mobile – and (2) a location-based feature referring to the actual place where the event mentioned by the rumor-related microblogs happened – domestic (in China) or foreign. Kwon et al.[41] observed from rumor time series that rumors tend to have multiple and periodic spikes, whereas non-rumors typically have a single prominent spike, and proposed an automatic detection mechanism of rumor on Social Networks using Periodic External Shocks model. Mendoza et al.[42] analyzed the retweet network topology and found the diffusion patterns of rumors different from news. They also found that rumors tend to be questioned more than news by the Twitter community, suggesting that the Twitter community works as a collaborative filter of information. To show the role of emotional signals in fake news detection, Giachanou et al.[43] proposed a Long Short Term Memory (LSTM) model that incorporates emotional signals extracted from text to differentiate between credible and non-credible posts. Finally, Guo et al.[44] described a fake news detection model based on a dual emotion representations by simultaneously learning emotion representations for both the publishers and users of posts.

2.2.2 Crowdsourcing Techniques for Misinformation

CrowdFlower is a popular tool among researchers for labelling data for misinformation research. Zubiaga et al.[45] used CrowdFlower to get a team of journalists to manually label tweets, with the annotators identifying only one of their specified features to support the truth status of the post. They used a feature scheme labeled as: support, response-type, certainty, and evidentiality. Their experiment showed that around 65% of the replies to original tweets were in the form of comments, which added little to the veracity of stories, while around 85% of tweets annotated had no evidence about the content being a rumor. In the work of McCreadie et al.[46] the authors used CrowdFlower to label tweets as belonging to unsubstantiated information, disputed information, misinformation, reporting, linked disputes, or opinionated posts. Their analysis showed substantial disagreement in regard to posts that provide opinions, with a minority of assessors often describing them as containing

disputed information, or being ambiguous.

A tool designed to allow journalists to identify and understand rumours quickly after they begin spreading on social media, using flags like "Is this true?" is presented by Resnick et al.[47] These rumors are then displayed on a community website where users can up-vote them if they think they're worth investigating further. Tschiatschek et al.[48] sort to automatically limit the spread of fake news by leveraging flagging tools added by Facebook. They proposed a model that uses Bayesian inference for detecting fake news and jointly learn about users' flagging accuracy over time. They worked to determine posts that will impact potentially fake news and hand them off to experts to review and remove. Ghenai and Mejova[49] applied a combination of machine learning and crowdsourcing techniques to identify rumor spread on Zika virus, and proposed a model that combined sentiment analysis, linguistic, readability and unique medical domain features to distinguish between rumor and non-rumor tweets.

One of the challenges to crowdsourcing is to ensure workers provide objective and truthful reporting. To account for this, Yingjie et al.[50] proposed a bidding and incentive mechanism for mobile crowdsourcing. To guarantee trustworthy submissions, the authors applied Evolutionary Game Theory to ensure that the best strategy for workers was to submit trustworthy data. Each worker is assigned a reputation score, which begins at a maximum but is decreased if a worker submits untrustworthy data, and increased if the worker submits trustworthy data. Different tasks on the platform have different reputation thresholds, which workers must exceed to work on the task. This makes reporting trustworthy data the most stable strategy for workers.

2.3 Models for Abusive Behavior Detection

As OSNs become interesting targets for spammers and malicious users, Verma et al.[51] reviewed literature to identify features used for detecting spam and malicious users. They pointed out that spam detection algorithms commonly explore features categorized as user-based, content-based, and hybrid (combining user and content-based features. Badjatiya et al.[52] applied several deep learning models with pre-trained word embedding over a dataset of

16k labeled tweets to classify tweets as racist, sexist, or neither. The best results from their experiments were derived by training with Long Short Term Memory networks (LSTMs) and embedding with a Gradient Boosted Decision Trees (GBDT). The work of Nobata et al.[53] describes a supervised learning system for detecting abusive language in online comments. Token unigrams and bigrams, and character n-grams were extracted from a dataset of 2 million Yahoo finance and news article comments. They adopted a classifier based on linguistic and syntactic features to achieve an F-score of 0.795 for Finance comments and 0.817 for News comments. To classify Twitter users as aggressors, bullies, or spammers, Chatzakou et al.[54] developed a classifier based on random forests to identify aggressive and bullying accounts. The model used a combination of user, text, and network features for its identification task. Almaatouq et al.[55] presented an analysis of suspended spam accounts on Twitter. Using Gaussian Mixture Model, the authors discovered that there are two primary categories of spammers on Twitter with distinct behavior. They hypothesized that the first group mainly consists of fraudulent accounts, while the second is made up of legitimate accounts that have been compromised.

2.3.1 Crowdsourcing Techniques for Offensive Behavior Identification

CrowdFlower is a popular tool among researchers for labeling data for research requiring labeled data as ground truth. Burnap and Williams[56] used CrowdFlower to label 2000 tweets by having annotators answer the question "Is this text offensive or antagonistic in terms of race, ethnicity, or religion?". They used the labeled data in a machine learning classifier for identifying hateful and antagonistic content on Twitter. Founta et al.[57] also used CrowdFlower to annotate a large collection of tweets with a set of abuse-related labels. Their research covers different forms of abusive behavior in order to identify a robust and consistent set of labels - abusive, hateful, normal, and spam - to characterize abuse-related tweets. To distinguish between hate speech and everyday usage of potentially offensive language in tweets, Thomas et al.[58] presented an automated model to classify tweets as

hate speech, offensive language, or neither. They used labeled data crowdsourced using CrowdFlower and adopted a logistic regression model with L2 regularization to identify hate speech.

On examining the mislabeled hateful tweets in the work of Thomas et al.[58] they observed that some were possibly incorrectly labeled in the crowdsourcing step, and some contained few of the terms commonly associated with hate speech. Tweets with less common slurs were also frequently mislabeled. One of the challenges to crowdsourcing is to ensure workers provide objective and truthful reporting. A lot of sites rely on posters to crowdsource the identification of abusive content because it is impossible for moderators to identify all abusive content. The research of Ghosh et al.[59] presented an algorithm where users rate content based on a set of ratings and the users also get rated based on the probability that they will correctly label a contribution. To account for the trustworthiness of crowdsourced content, Wang et al.[50] proposed a bidding and incentive mechanism for mobile crowdsourcing. To ensure trustworthy submissions, the authors applied Evolutionary Game Theory to ensure that the best strategy for workers was to submit trustworthy data. Each worker is assigned a reputation score, which begins at a maximum but is decreased if a worker submits untrustworthy data. It is also increased if the worker submits trustworthy data. Different tasks on the platform have different reputation thresholds, which workers must exceed to work on the task. This makes reporting trustworthy data the most stable strategy for workers.

2.4 Privacy and Information Leakage

Users post and provide personal information without an understanding of how it might be used or accessed, leading to privacy and information leakage. Privacy controls are provided to limit access to user information but OSN default settings allow unlimited access, unless the controls are enabled by users. Krishnamurthy and Wills[60] found that between 55% and 90% of users in OSNs still allow their profile information to be viewable and 80% to 97% of users allow their set of friends to be viewed. To address privacy concerns, users can utilize privacy settings and hide sensitive information, but it has been shown by He

et al.[61] and Zheleva et al.[62] that such measures, even though promising, are not sufficient to protect the user's privacy due to the friendship relations, group memberships, or even participating in activities like mentions, tags (posts and photo), shares, and commenting, which can be harvested through 'screen-scraping'[63] or other means. To show that group membership can be increase information leakage, Zheleva et al.[62] proposed eight privacy attacks for sensitive attribute inference using a variety of classifiers and features to show ways in which an adversary can utilize links and groups in predicting private information. The problem of privacy leakage even under privacy control continues due to the underlying conflicts between privacy control and essential OSN functionalities.[64]

Information leakage can be viewed as the combined probability of sensitive attribute inference from the information available in immediate friends' profiles, and Talukder et al.[65] addressed that by presenting a friends rank component that finds the amount of match of sensitive attribute values between the user and friends. The user is then provided a self-sanitization component that shows which high ranked friend would cause more leakage than a friend who is ranked lower. To estimate Facebook users' ages, Dey et al.[66] exploited the underlying social network structure to design an iterative algorithm, which derives age estimates based on ages of friends and friends of friends, while Li et al.[67] inferred demographic information such as age, gender, education by observing users' exposed location profiles. Since users will sometimes have accounts over multiple social networks, Chen et al.[68] described the privacy leakage that arise from cross-network aggregation based on four real-world social network datasets. Since these networks offer various levels of privacy protection, the weakest privacy policies in the social network ecosystem determine how much personal information is disclosed online.[69]

Many security and privacy risks also emanate from publishing social network data sets, as these can be used for cross-network aggregation. Yin et al.[70] defined the attribute couplet attack where relationships between pairs of users and additional background information is used to unveil protected identities. Amiri et al.[71] proposed a community detection method for privacy preservation utilizing hierarchical clustering, with nodes divided iteratively based on learning automata. To address the privacy concerns emanating from neighborhood at-

tacks, Zhou and Pei[72] adopted k-anonymity and l-diversity models from relational data to social network data. Neighborhood attacks arise when an adversary uses knowledge about the neighbors of a target victim and the relationship among the neighbors to re-identify the victim from a social network even if the victim's identity is preserved using the conventional anonymization techniques. Liu et al.[73] explored edge-weight perturbing methods using Gaussian randomization multiplication and greedy perturbation algorithm for edges considered confidential to limit the risk of disclosure of confidential knowledge, while retaining the shortest path and the approximate cost of the path between pairs of nodes in the original network.

Chapter 3

MIDMod-OSN: A Microscopic-level Information Diffusion Model for Online Social Networks[1]

Information diffusion describes how information is transmitted between individuals. In online social network like Twitter, information can easily become viral because it allows strangers to filter, discuss and share information of common interest with networks of followers and through the use of hashtags. The ease of accessibility and the broad reach makes Twitter a strategic tool for businesses, interest groups, politicians and journalists and during crises and disasters. Information is said to propagate, or diffuse, when it flows from one individual or community in a network to another. In the case of Twitter, diffusion can be seen as an action to share a Tweet with a user's followers with (i) no other new content added, called Retweet or (ii) new content added, called a Quote. Most studies in analyzing information diffusion focus on the overall spread of information by focusing on event detection and the spread of the event across the network without comprehensively evaluating the diffusion process on a microscopic level – i.e., the factors that influence diffusion, differences in the spread of information in varying Twitter events and the information dissemination process. It is usually hard to assess why some information disseminates and other does not, but it is safe to

assume that the features and/or the contexts of messages that go "viral" and those that do not must differ to some extent. In crises/trending Twitter events, the volumes of messages and interaction grow exponentially within a short time. This kind of interaction explosion is expected to impact the prediction model in a different way than when the spread is over a longer period. We assume that building a temporal pattern of a user's online behavior – like the time of day when the user creates or reacts to tweets versus when the tweets get retweeted – is important, as this behavior can be exploited for targeted information spread. By successfully identifying the features that make a difference in determining the virality status of a post, organizations can identify attributes to look out for in nodes that will ensure maximum information spread and nodes to avoid in case of containment. After predicting the spreading behavior of a post from one node to another, one can extend the scope of prediction to community-wide and/or network-wide.

3.1 Overview

Existing models for predicting information diffusion observe diffusion on a holistic level across trending events or hashtags. Many of these studies are focused on finding super-seeders, or influential nodes, based on the assumptions that the influence of the feature vector will be static across event types. The feature vector is a combination of attributes, possibly specific to user, message, network, and/or interaction, that contribute to an account's online persona. In this study, we hypothesize that the features that contribute to information diffusion in online social networks are significantly influenced by the type of event being studied. Since Twitter is increasingly becoming a place to visit for trends and breaking news, as well as asking questions and gathering information on general topics, we classify Twitter events as (1) informative for topics relating to general knowledge and which have not attained viral status, and (2) trending for topics that can be described as viral, breaking news, hot topics, or crises. We describe a topic to be trending if there are observed sharp spikes in the rate of posts relating to the topic instead of a gradual growth observed over a period of time. Similar to studies on predicting extremism[28] and temporal dynamics[27] in social networks, we build

a model that predicts diffusion using features learned from Twitter data. We go the extra mile by exploring the node-to-node influence dynamics associated with information spread. We use machine learning models to observe the performance and effect of similar sets of features on the previously identified Twitter events types to understand how the pattern of discussion, diversity in opinion, urgency and timeliness of topics influence diffusion behavior. The proposed model is built on Bayesian Logistic Regression for learning and prediction and Random Forests model for feature selection. These two statistical models have been observed to perform sufficiently well in predicting information spread in online social networks.

3.2 Dataset Description

One of the biggest challenges to this research is access to data, as most of the datasets and tools available only provide part of the information (usually, tweets and network features) needed for academic research. Due to the number of features being examined, we needed the complete metadata of Tweet and user JSON (JavaScript Object Notation) objects. For the purpose of future research requiring Twitter JSON objects, we created a tool that crawls the Twitter Search API using the usernames or IDs of a set of seed users and made it publicly available on GitHub. The tool creates a relationship graph built around the seed users and their followers. Since it is almost impossible to have the complete Twitter graph, the sub-graph generated is as representative as it can be. For each of the 4 topics we are exploring, we randomly select 50 users and build a followership relationship around them for up to depth 2. The user (or node, used interchangeably throughout the remainder of this paper) information is then used to build a database crawled over a 30-day period, by collecting all the tweets created by users in the sub-graph during this time period.

The use of Twitter to report real-life events is steadily increasing and for this study, we classify these events into two categories: informative and trending. We then base our study on two different topics for each event. The topics defined are (1) Informative: (1.1) Health benefit of coffee, (1.2) Mental health and (2) Trending: (2.1) 2018 Kansas elections, (2.2) Government shutdown. The data and network distribution for the dataset can be found

Event type	Topic	No. of users	No. of edges	No. of tweets	Diffused/not diffused ratio
Informative	health benefits of coffee	50919	1100270	2958382	40/60
	mental health	29362	3224330	4030412	
Trending	2018 kansas general elections	15339	2509255	24188962	52/48
	government shutdown	12581	2549136	14513377	

Table 3.1: *Data distribution.*

in Table 3.1. We associate each topic with a bag of words that are deemed important to the topic by creating a list of words frequently used with or associated with the topic. A tweet is said to be relevant to a topic if and only if it contains one or more of the predefined keywords. For example, 60 key words were used to identify tweets belonging to the topic of *health benefit of coffee*. The data is split into (1) a training set used for parameter estimation and (2) a test set to assess the performance of the model. We limit the data collection to tweets created in English, and with the API location set to the United States. Since we are interested in the temporal characteristics of the data, the timestamp associated with the data is relative to the time zone of the crawler.

Information spread behavior: In a directed network $G = (V, E)$ with no self-links (communities within the graph might contain cycles), V is the set of nodes and $E(\subset V \times V)$ is the set of edges. For each node $v \in V$, we denote U as the set of v's followers and W as the set of v's friends, i.e., $U = \{u; (v, u) \in E\}$ and $W = \{w; (w, v) \in E\}$, respectively. Similar to Saito et al.,[23] we assume $AsIC$ with the time delay function associated with information diffusion along the edge. At time t each node v gets a chance to activate (get a reaction through retweet, favourite, quote or reply) its follower u. If node u is not activated by time $t + \delta$, then node v looses the competition for activating u to any other node v' that attempts to activate u between $t + \delta$ and the time of u's activation. For simplicity, we assume that activation is restricted to a node's interaction with the network, but in reality, this will not always be the case, as activation is not solely dependent on the network activities but could be from sources external to the network itself, thereby causing delay in activation.

3.3 Learning and Feature Estimation Models

The model we propose takes a pair of users with established followership relationship and extracts a set of attributes classified as: Network, Interaction, Semantic and Temporal. We adopt two off-the-shelf machine learning models: Bayesian Logistic Regression for prediction tasks and Random Forests for feature selection due to the good performance of both models in similar settings, observed in Guille et al.[27] and Ferrara et al[28]. We use the attributes described in Tables 3.2, 3.3, 3.4, and 3.5 to train our model based on Bayesian Logistic Regression (BLR). At the initial stage of experimentation, we explored the Random Forests model for prediction but observed model over-fitting which could be due to the similarities in the data. We went with the BLR model which has been proven efficient by Guille et al.[27] and Ferrara et al[28]. The prediction capability of the model is tested and evaluated before the feature selection phase. One challenge with high dimensional models is that as dimensionality increases, the space between data points becomes very large,[74] to the extent that it is difficult to produce reliable results. By removing features that are highly correlated and those with minimal effect on the predictability of the model, we select a subset of the original features by using Random Forests (RF) as a filter. The BLR model is then re-trained with the selected feature set and evaluated to determine the predictive abilities of the selected features.

We perform node-to-node influence analysis by examining feature performance between two users with established followership relationship. We extract attributes from our dataset and organize them as: Network, Interaction, Semantic and Temporal. The features are estimated for both the source and destination nodes, with an associated binary label depicting diffusion along the edge between them. For each user, we learn 27 features, and a social homogeneity (common to two users, showing an overlap in the sets of users they relate with, i.e. common friends and followers) by adopting the features of Guille et al.[27] (excluding the temporal feature) and introducing new ones. Since each observation is a pair of users given as source and destination, the input to the learned model is a vector of 55 features along with a diffusion label per data point. For the temporal dimension, we study the creation,

consumption and forwarding of content by splitting a 24-hour period into 6 hours interval (i.e., 0:00-5:59 am, 6:00-11:59 am, 12:00-5:59 pm, 6:00-11:59 pm) and learn a model for each time period. Overall, we learn 4 temporal models for each pair of users, to observe how the post and reactions to post behaviors change across different time periods in a day.

Feature	Description
avg number of followers	higher follower count depict higher reach
avg number of friends	average number a user follows
ratio of followers-to-friends	shows how balanced the user's network is

Table 3.2: *Network features extracted for each user to serve as input variables to the learning model.*

3.4 Model Evaluation

Each input is a vector set of 55 features, learned over 4 different time periods. The performance of the models are obtained using the k-fold cross validation technique, with $k = 10$ folds, and using the $80\% - 20\%$ training-test data split and averaging performance across the 10 folds. The prediction capabilities of the learned model are tested based on its abilities to predict if there is diffusion across an edge given the learned model. We employ standard machine learning evaluation metrics: Precision, Recall and F1 score, along with Area under the Receiver Operating Characteristics (ROC) curve to measure the predictability of the model.

3.5 The Diffusion Prediction Experiment

In this section, we describe our experimental setup, and the results obtained for each phase of our model. We evaluate the performance of the prediction and feature selection models, and then make comparisons with state-of-the-art prediction models. Finally, we discuss the time to tweet paradigm based on our observations.

Feature	Description
volume of tweets	normalized over account's lifetime
social homogeneity	shows two users' common friends and followers
ratio of directed tweets	percentage of his posts are directed at others
active interaction	binary value depicting established interaction between them
mention rate	gives volume of posts directed at the user
ratio of retweet-to-tweet	percentage of user's posts that have been retweeted
tweets with hashtags	how many of his original posts contain hashtags
retweets with hashtags	shows the user follows and reacts to posts containing hashtags
volume of retweets over account's lifetime	we assume the account is a forwarding bot if all his posts are retweets
avg tweets per day	gives insight into how active the user is
avg number of mentions excluding retweets	shows how interesting others find the user
ratio of mentions-to-tweet	includes posts where the user mentioned and retweeted other people's posts
tweets containing URL	shows how many of the user's original tweet contain URLs
retweets containing URL	shows the user follows and reacts to posts containing URLs
tweets containing media	shows how many of the user's original tweet contain media (photos, videos)
retweets containing media	shows the user follows and reacts to posts containing media
presence of user description	a boolean value showing if the user's profile has description (bio)
ratio of favorited-to-tweet	shows how many of the user's tweets have been endorsed by others

Table 3.3: *Interaction features extracted for each user to serve as input variables to the learning model.*

Feature	Description
presence of keywords	boolean value that shows if the user has tweeted about the topic
positive polarity of tweets	percentage obtained from running sentiment analysis on all the user's tweets
negative polarity of tweets	percentage obtained from running sentiment analysis on all the user's tweets

Table 3.4: *Semantic features extracted for each user to serve as input variables to the learning model.*

Feature	Description
ratio of tweets per time	percentage of all user's posts within a time period
ratio of tweets that got retweeted	percentage of original tweets that got retweeted within a time period
ratio of retweet per time period	percentage of reactions user produce within a time period
average time before retweet	estimates the average time elapsed before the user gets a reactions

Table 3.5: *Temporal features extracted for each user to serve as input variables to the learning model.*

3.5.1 Experimental setup

We perform a supervised learning task where we train the model using the attributes from a pair of nodes with an established followership relationship and label the interaction between them as either diffused or not diffused. An edge is said to be diffused if and only if the destination user (in Twitter terms: *follower*) has at any point forwarded his friend's (*followee*) messages on the topic being examined. The attributes learned are said to be representative of users' network, interaction, participation, role and importance in the spread of information to other nodes in the network. As previously stated, these attributes are learned over four different time intervals. After learning these features, we fit a regression function that maps the learned user attributes to the likelihood of diffusion between the nodes.

Given the directed nature of the Twitter graph, the learning task is non-deterministic, as switching the source and destination nodes may produce a different mapping between the input and output variables. Initially, we maintain an equally weighted feature space with the assumptions that each feature will influence the forwarding decision (reshare, reply or not) with equal magnitude. Subsequently, the feature selection framework is initialized to first learn a function with the same set of attributes, secondly rank the features in decreasing order of importance, and third retrain the model using the 15 most important features.

We evaluate the effectiveness of our model and methods on predicting diffusion between node pairs in the spread of information across the social network on selected topics using the methods described in Section 3.4. Also, we present our findings on the optimal subset of features necessary for maximized diffusion predictions, with discussions on the best time to

post given the event type. Experimental results show a significant improvement over state-of-the-art models both in accuracy of prediction and the ability of the model to differentiate between diffused and not diffused edges.

3.5.2 Diffusion prediction model

Firstly, we observed that the volume of tweets across a 30-day period varied widely for informative and trending events. As shown in Table 3.1, it can be established that even though the combined number of users observed in the trending events is 2.8 times less than the number of users across informative events, we were still able to record 5.5 times more tweets over informative events. We note that in our dataset, trending events generate up to 15 times more tweets than informative events with the same network size. This sort of data projection will be sufficiently affected by the impact of the topic. For instance, one can forecast such data growth for trending events with wide reach like political and health topics but not in lifestyle. Other factors that will impact the data projection include time of day, and external sources like coverage in traditional media.

In Table 3.6, we show the performance of our models, averaged out across topics in each event class, given the performance metrics previously highlighted. Using the F1 measure, the model achieved 93% accuracy in prediction in informative events and 86% in trending events. The simplified models, based on the 15 most important feature for training, showed a 90% prediction accuracy in informative events and 89% in trending events. Results in the present study are consistent with the prediction results for trending events in past literature.

Event type	Model	Precision	Recall	F1
Informative	55 features	0.91	0.96	0.93
	top-15	0.87	0.94	**0.91**
Trending	55 features	0.87	0.84	0.86
	top-15	0.89	0.90	**0.89**

Table 3.6: *Performance evaluation of MIDMod-OSN in predicting diffusion of posts from different event types.*

Furthermore, we compare both our prediction models with the state-of-the-art diffusion prediction model proposed by Guille et al.,[27] see Table 3.7, and observe that both models with

55 and top-15 features perform considerably better than the state of the art. Our hypothesis that increasing the feature vector space by extracting and learning more attributes from the Twitter JSON objects will make the predictive model more robust is proved correct as we were able to record a 7% increase from the model of Guille et al. It might be argued that a 7% increase is not enough to justify the increase in computation time and resources caused by the increase in feature space, however, we oppose this argument with the feature selection phase, introduced solely for maximizing diffusion prediction by utilizing the features that will directly impact the information spread. For a small cost in accuracy, reducing the input variables by 72% (top-15 features) will give a prediction accuracy of 91%, which is only a 2% reduction in predictive power (when compared with all 55 features). In like manner, an 81% reduction (top-10 features) yields a prediction accuracy of 87%, constituting a 6% reduction in accuracy. The trade-off in adopting the top-10 features is significant, and as such, we adopt the top-15 important features as the optimal set of features necessary for diffusion prediction without incurring expensive computational costs.

Event type	Model	F1	AUC-ROC
Informative	55 features	0.93	0.98
	top-15	0.91	0.96
	top-10	0.87	0.94
	Guille et al.(13 feat.)	0.86	0.94
Trending	55 features	0.86	0.94
	top-15	0.89	0.96
	top-10	0.88	0.94
	Guille et al.(13 feat.)	0.88	0.95

Table 3.7: *Prediction accuracy using proposed model with different number of features and state-of-the-art.*

Contrary to expectations, it is observed that learning all possible features in trending events impacts prediction accuracy negatively. Due to the consistently changing pattern of interactions and behavior in trending events, increasing the number of features learned brings about over-fitting caused by the exponential growth in the data needed for training. We are able to mitigate the impact of over-fitting in the model using the k-fold cross validation technique, with k set to 10. Nonetheless, it will be detrimental to suggest that learning these features is of no value, as we are convinced that feature selection over several topics

will be useful in building a template of attributes for a pre-trained prediction model. The accuracy of the prediction model is consistent with previous studies that have focused on Trending events.

Rank	Informative	Trending
1	dest (destination node) average url per tweet	Social homogeneity
2	src (source node) ratio of retweet per time period	dest active interaction between the nodes
3	src volume of tweets over account's lifetime	src avg number followers
4	dest ratio of tweets that got retweeted per time period	src ratio of favorited to tweet
5	social homegenity	ratio of common friends
6	dest avg number of media in retweets	src ratio of retweet per time period
7	src ratio of retweets to tweets	src volume of tweets over account's lifetime
8	src ratio of tweet per time	src active interaction between the nodes
9	src ratio of tweets that got retweeted per time period	src avg url per tweet
10	dest avg number of retweets with hastags	src ratio of retweets to tweets
11	dest ratio of retweet per time	src ratio of mentions to tweet
12	src avg number of retweets with hastags	src avg number of tweets
13	src average url per retweet	src avg number of mentions not including retweets
14	src avg number of tweets	dest avg number of mentions not including retweets
15	dest avg number of retweets	dest volume of tweets over account's lifetime

Table 3.8: *Ranking of the top 15 optimal features that should be maximized for maximum diffusion or minimized for containment.*

3.5.3 Cross testing between models

To further show that the performance of the models is not biased to topic domains, we tested the informative model with a political related topic and trending model with an health related topic. On testing both models with data from new topics (not used for training and in new topic domains), we observed results similar to those reported earlier with F1 score of 90.1%

for informative and 89% for trending events. This confirms that the models will perform comparably regardless of topic domain.

To ascertain that there is indeed a difference between the informative and trending models, we evaluated the informative model with data from trending topics and evaluated the trending model with data from informative topics. The objective is to test if the knowledge gained from one model can be used in making predictions in the other. The outcome of predicting the diffusion of trending posts using a trained informative model produced an F1 score of 82%, while we observed an F1 score of 78% from predicting informative posts using a trained trending model. This result is not totally surprising, due to the irregular pattern associated with posts and users contributing to trending topics.

3.6 Feature Selection Framework

One justification for using multivariate methods is that they take into account feature redundancy and yield more compact subsets of features, as features that are individually irrelevant may become relevant when used in combination, which also shows that correlation between sets of features does not necessarily imply redundancy.

Evaluating the Random Forests model using a 10 fold cross-validation technique achieved an AUC score of 99% in both informative and trending events using the complete set of features. Considering that the goal of the feature analysis task of this study is to identify the optimal set of features necessary to maximize diffusion prediction, we select the top 15 features, rather than the traditional top 10 (for reasons highlighted in 3.5.2). In Table 3.8, we report the ranking of the top 15 features in the two event types.

Given two users, we observed that the attributes of the followers (destination nodes) account for 40% of the optimal subset of features, in informative events, and for 20% in trending events. In recent happenings in online social networks, it has been observed that discussions and threads that impact trending events are not usually trending in nature. For instance, the much publicized propaganda campaign during the U.S. 2016 elections targeted users on both sides of the political divide by exposing them to opinions formulated over time,

using hashtags and shortened URLs. In real life, a considerable number of trending topics are indeed informative events that become trending due to a change triggered by an incident. Irrespective of the type of event, social homogeneity and source's (1) ratio of retweet per time period (2) volume of tweets over account's lifetime (3) ratio of retweets to tweets (4) average number of tweets, prove to be important in the information diffusion process.

We notice that the follower's features are powerful enough to impede diffusion in informative events but these abilities diminish as the event becomes trending. As topics become viral, the number of followers a user has ranks third in trending events. Even though this feature is previously deemed unimportant in informative posts, combining it with a high ratio of retweets to tweets, mention rate and active interaction from his follower will boost his reach. It is inadequate to assign importance to an account across all networks and topics, as seen in Rao et al.,[75] if the importance and authority it wields vary with changing topic, event and social network. It is paramount that the relevance of a user be decoupled across social networks, especially Twitter, since a considerable number of users maintain a level of anonymity. For instance, a user will not run a web search on an account to confirm the authenticity or authority of its posts before reacting on Twitter. Also, a user that is authoritative on health-related issues on Twitter might be an unreliable source of health-related posts on Facebook. It is inadequate to assign importance to an account across all networks and topics if the importance and authority it wields vary with changing topic, event and social network. Throughout this research, we demonstrated that the role of the followers in diffusion prediction is more than just a contribution to the follower count of the sender, and should combine the effectiveness of the interaction of each follower with their friend. Our results show that the influence a user wields in a network is an aggregate of his influence over each of the nodes in the network, thus combining all three centrality concepts as introduced by Freeman.[76]

3.7 Time to Tweet

The results from our experiment validate our assumptions that the extent to which messages diffuse will be significantly influenced by the time when they are created. As observed in the top-15 most important features, see Table 3.8, for both the follower and followee in the network, the time period where most of their messages (original tweets and reaction) fall are crucial to propagation. Experimental results show that more than 75% of informative posts fall into the 2nd (6:00-11:59 am) and 4th (6:00-11:59 pm) time periods, but those of trending posts are in the 3rd (12:00-5:59 pm) and 4th (6:00-11:59 pm) time periods. It is interesting to note that both Twitter event types got considerable attention during the 4th time period as this for most people is a time to catch up with the day's activities. However, we observed that the best time to tweet an informative message on Twitter for maximum diffusion is in the 2nd time period, while trending is in the 4th.

We speculate that the contrast in peak diffusion times can be due to the reactive nature of trending events, occurring mostly after the day's activities, unlike the active nature of informative events, where a user is mostly putting opinion out. Additionally, the many time zones in the United States could also contribute to this result as users could fall into different time periods. The goal of this analysis on time to tweet is to gain insight into the influence of time on the spreading process as against pinpointing the best time to tweet for maximum exposure. For maximum exposure, it will be important that the time period is kept shorter.

3.8 Crowdsourcing for Early Trending Topic Detection

In this section we discuss the concept of crowdsourcing in OSNs, and why it is important. We describe the experiment and experimental results on adopting MIDMod-OSN for crowdsourcing the early detection of trending topics.

3.8.1 The early detection of trending topics

Individuals and organizations looking to use Twitter as an advertising or political campaign platform will find it useful to know ahead of time if a newly created message or hashtag will become trending, in order for them to maximize the attention for personal gain or minimize negative exposure. Similarly, governmental or non-governmental organizations attempting to neutralize the spread of misinformation during crisis scenarios could monitor users' reactions to previously identified harmful-misinformation-carrying messages, and predict whether these messages will become viral before this determination can be done via standard methods, like counting tweets. This would enable them to effectively fight the further spread of the misinformation before it has a chance of becoming viral.

3.9 Using MIDMod-OSN for Crowdsourcing

In the past, individuals and organizations have used OSNs like Twitter as an avenue to obtain ideas in a crowdsourcing context. In crowdsourcing tasks, especially when backed by incentives, participants may introduce an implicit bias in the data brought about by the "presence" of an observer, leading to a change in behavior[77] or opinion and causing them to provide feedback that they feel is expected or sense what the "community" rewards, and comply. By contrast, our proposed crowdsourcing mechanism aims to observe users in the wild, making it independent of the bias introduced by conscious detection. In this study, we view a user's reaction to a post as an implicit contribution to crowdsourcing. Users' posts and reactions serve as criticism or validation to reports on crises and events, products and services, protests, or even political campaigns.

Users react uniquely to posts, and their reaction may or may not be correlated with the message's potential for becoming trending. While some users react to posts from all event types (trending and informative), others only react (share, quote, favorite, reply or retweet) to tweets that are trending or about to attain the trending status because of the need to share or contribute to hot topics. This kind of users can serve as discriminants in the model

that predicts the trending character of a message. The goal of the prediction task is to show that the diffusion behavior and OSN behavior of users is useful for predicting the trending character of a message when the reaction count is unavailable.

3.9.1 Experiment design and results

For this experiment, we are interested particularly in evaluating the usefulness of users' reactions to predicting message virality. It is for this reason that we must avoid (1) including specific message features in the classifier and (2) including – explicitly or implicitly – counts of tweets relating to a specific message. The first requirement fits naturally with our previous model, which only relies on user, rather than message, features. To satisfy the second requirement, we must construct an experiment that treats each user interaction with the message independently of all others. That is, we purposely make a prediction of virality from each user interaction, rather than combining all user interactions into a single model.

Our model predicts if a message will go viral or not, by including the diffusion property *diffuse/not diffuse* of the message as an independent variable during the training phase. We examine how users on Twitter relate with posts of their friends by building a classifier to distinguish user interactions based on the virality status of the message. For a message m, where $m \in \{1, \ldots, M\}$, spread over a network with n interactions, we train a model that predicts the virality status of the message based on the diffusion behavior observed along each one of the n links along which the message propagates. This results in n distinct predictions. The overall predicted output is calculated as the majority virality status observed across the n interactions. We select 1000 messages –500 each– from trending and informative event types and evaluate the MIDMod-OSN's ability to predict if a message will go viral or not. For instance, if a trending message is spread over 5 interactions and the model predicts the post to be Trending 3 out of 5 times, we accept the output as Trending and evaluate the model over its correct classification of M messages in the test collection.

We run the experiment with 10000 users. With this fraction of the network, we were able to show that to a certain degree that the diffusion behavior and OSN behavior of users is

useful for predicting the trending character of a message without having to count the number of reactions, see Table 3.9.

Model	Precision	Recall	F1
virality-predicting	0.65	0.78	0.70

Table 3.9: *Performance evaluation of MIDMod-OSN in predicting the trending status of a message without counting reactions.*

We should note here that when attempting to predict message virality, one should consider a more comprehensive model, including message attributes and a joint treatment of all user reactions to a specific message. Nevertheless, the results of this experiment demonstrate that crowdsourcing (at least part of) the detection mechanism is not without merit.

Chapter 4

An Implicit Crowdsourcing Approach to Rumor Identification in Online Social Networks[2]

The impartial and unrestrained spread of information in social networks can be of great value as observed in September 2015 where the US geological survey tracked earthquakes by simply following mentions of the term 'earthquake' usgs or the 2012-13 flu epidemic where researchers used tweet data to correlate the spread of the disease with a view to reducing its impact,[78] and in stock markets where consumer insights companies use social media data to predict shifts in consumer spending behaviors that translate to shifts in stock prices. However, the same social network features that offer these benefits can quickly become detrimental when the spreading information is false, like during hurricane Sandy where there were false tweets about the NYSE being flooded with up to 3 feet of water, which even got reported by some news outlets.[79]

According to deflationism,[80] assertions that predicate truth of a statement do not attribute a truth property to such a statement. Since there is no real-world truth label to posts (i.e., text, images, memes, etc.), OSN users simply decide to react to a post based on the perceived credibility of the message. A message intended to deceive might have concealed

meanings, emotions and sentiments even if it appears otherwise. The search for the truthfulness of a message might be lacking, depending on how accepting or prejudiced the user is towards a topic, especially when they are exposed to contradicting information from diverse sources. Since some rumors never completely die out, persisting with low frequencies with potential for flare-ups from time to time, detecting misinformation posts early on, before a flare-up, is more meaningful than detecting them when 90% of the total related post volume has already been consumed.[35;81]

4.1 Overview

In this study, we adopt an implicit crowdsourcing model for predicting the credibility of posts in OSNs, which works by simply observing users' interaction with these posts. The proposed model is implicit, in the sense that no undue influence is exerted upon the observed users, and hence guarantees that the users' posting and reaction behavior is completely natural. We introduce a new paradigm for credibility prediction predicated on the interaction between users encountering the messages. Seeing as feature design and selection strongly impact a machine learning model's accuracy much more than the model used,[82] we place emphasis on identifying the features that determine the spread of *True* posts, and those that determine the spread of *False* posts. We train a Bayesian Logistic Regression model by incorporating network, interaction and message features to measure the node-to-node influence dynamics to rumor propagation.

Existing research in rumor propagation and identification examine the behavior of misinformation posts over the network based on diffusion speed, depth, concentration, location, and sometimes combining features to differentiate posts. However, with access restrictions to the complete Twitter network graph and posts, it is important that we examine how individual users contribute to the diffusion of rumor posts and what features of the post sharer and receivers influence this paradigm. Since the spread of gossip is a uniform process, spreading from node to node,[83] it is essential to note that the diffusion process is influenced not only by the creator of the tweet, but also by the sharer of the tweet.

Here, we describe two research problems and adopt an implicit crowdsourcing approach to addressing them:

1. We investigate credibility prediction by exploring rumor propagation founded on microscopic-level misinformation spread. By observing the spreading behavior of rumors in online social networks, we propose a model that predicts if a message is *True* or *False* by observing the latent attributes of the message, along with users and their reactions over the network.

2. We examine the contribution of individual users to rumor propagation in OSNs, by investigating features of users (both the post sharer and receiver) and how these features influence the propagation of rumor.

Previous crowdsourcing-based approaches in rumor detection focus on conversation annotation for credibility detection. We introduce a novel approach that explores crowdsourcing as an automated tool for identifying rumor in online social networks. We classify users based on the types of posts they generally react to: (i) reacts to only *True* posts, (ii) reacts to only *False* posts and (iii) reacts to a mix of *True* and *False* posts. Users in class i and ii are good discriminators for both credibility detection and feature identification, while users in class iii do not serve as good discriminators in the prediction model.

4.2 Features for Rumor Propagation and Identification

Here, we describe a framework that given a tweet will predict (1) whether the tweet is *True* or *False* by observing user interaction with the tweet, (2) whether the followers of the spreader (could be the author or someone sharing) will react to the tweet in the form of a retweet, share, quote, like or favorite. We suggest 3 categories of features: message, interaction, network, and train a random forest classifier to rank the features in order of importance, then we build a Bayesian logistic regression model for classification. We adopt some of the features examined in the literature and suggest new ones, described below.

4.2.1 Network-based features

In microblogs such as Twitter, a *friend* is someone a user follows, and a user can see all of his friends' posts. In like manner, a follower is someone that follows and has direct access to all of a user's posts. We consider three features of the user's network: *followers count*, *friends count*, which have been extensively studied by Castillo et al.,[37] Liang et al.,[38] Yang et al.,[40] and *followers to friends ratio*, which was used by Wu et al.,[39] to establish opinion leaders. These attributes are important because a user's friends impact the kind and volume of messages that end up in his timeline and the higher the number of followers, the farther the possibility of reach. This is also reflected in policies by OSNs like Twitter and Instagram who attach value to the followers count, where users become verified once they cross a certain threshold, even if the account holder is not a celebrity or public figure. Table 4.1 describes the network features used in the model.

Feature	Description
followers count	higher count depict higher reach
friends count	# of accounts user follows
followers-friend	ratio to show influence in the network

Table 4.1: *Network-based features.*

4.2.2 Interaction-based Features

Since we are exploring rumor propagation as being dependent on the influence being wielded between users and taking propagation depth to be a factor of how messages cascade across the network, we examine the nitty-gritty of the followee-follower relationship to establish the features that influence the spread of rumor over the network. Here, we identify specific attributes of the user's online persona and posting behavior as determinant to being an influencer or influenced in the network. The assumption is that both the follower and followee contribute equally to the diffusion of a post, and an aggregate of network and message attributes tilt the reaction decision. Table 4.2 describes the 14 interaction attributes being considered. The last 5 features have been explored by Castillo et al.,[37] while we introduce 9

new features to the study of rumors in social networks.

Feature	Description
shared friends	common nodes they interact with
directed tweets	ratio of tweets directed at someone
dialogue	active interaction from user 1 to 2
retweet-to-tweet	ratio of user's tweets with retweet
tweet wit hashtag	ratio of user's tweets that contain hashtags
tweets with url	ratio of user's posts with URL
tweets with media	ratio of user's posts with media
avg favorite-tweet	ratio of posts that get favorited
avg tweets/day	shows how active the user is
has url	does user's profile have a URL
has description	does user's profile have description
is verified	is the account verified
status count	volume of tweets over account's lifetime
account age	# of days since account was created

Table 4.2: *Interaction-based features.*

4.2.3 Message-based Features

Twitter posts are very fluid, taking up various forms as feedback, news, marketing campaigns, etc., so it is expected that rumors in this medium come in all forms. We account for this variation and consider the concealed form and intents of posts. Previous work have focused on count of positive and negative words in a tweet, with some exploring the polarity of the message sentiment but we look to explore the latent attributes of the message by introducing new features encompassing the type of post and emotion it is meant to incite. We adopt paralleldots API to perform content analysis on tweets to reveal the sentiment, intent, emotion and abusive attributes. Paralleldots uses deep learning to provide analysis on a given text. Table 4.3 describe the message attributes - relating to the form, meaning and intent of the message, adopted in our model.

Feature	Description
quoted status	has post been quoted
is rt	has post been retweet
rt count	# of retweets
rt status	is post a retweet
favorited count	# of favorites
has hashtag	does post contain hashtags
has url	does post contain URL
has mentions	does post mention someone using "@"
has media	does post contain media
avg tweet length	length of tweet / 280 (max length)
positive sentiment	positive polarity of tweet
negative sentiment	negative polarity of tweet
neutral sentiment	neutral polarity of tweet
happy emotion	is post meant to incite happiness
fear emotion	is post meant to incite fear
sad emotion	is post meant to incite sadness
angry emotion	is post meant to incite anger
bored emotion	is post meant to incite boredom
feedback intent	is post meant to be a feedback
news intent	is post meant to be news
query intent	is post meant to be a query
spam intent	is post meant to be spam
marketing intent	is post meant for marketing
abusive	is post abusive

Table 4.3: *Message-based Features.*

4.3 Experiment Setup

In this section, we describe the data collection process, prediction models and the metrics for evaluation. The approach is to (1) identify topics labeled as *False* (in other words, rumor) or *True* using Snopes[84] – an online fact-checking site – and collect Twitter posts about the topic. (2) To each user, we associate a total of 17 features, to include 3 network and 14 interaction attributes; and to each message, we associate 24 attributes: 10 observable and 14 latent attributes. We then train a Bayesian logistic regression model based on the prediction task. In Figure 4.1, we present an abstraction of the experiment setup for the credibility prediction task.

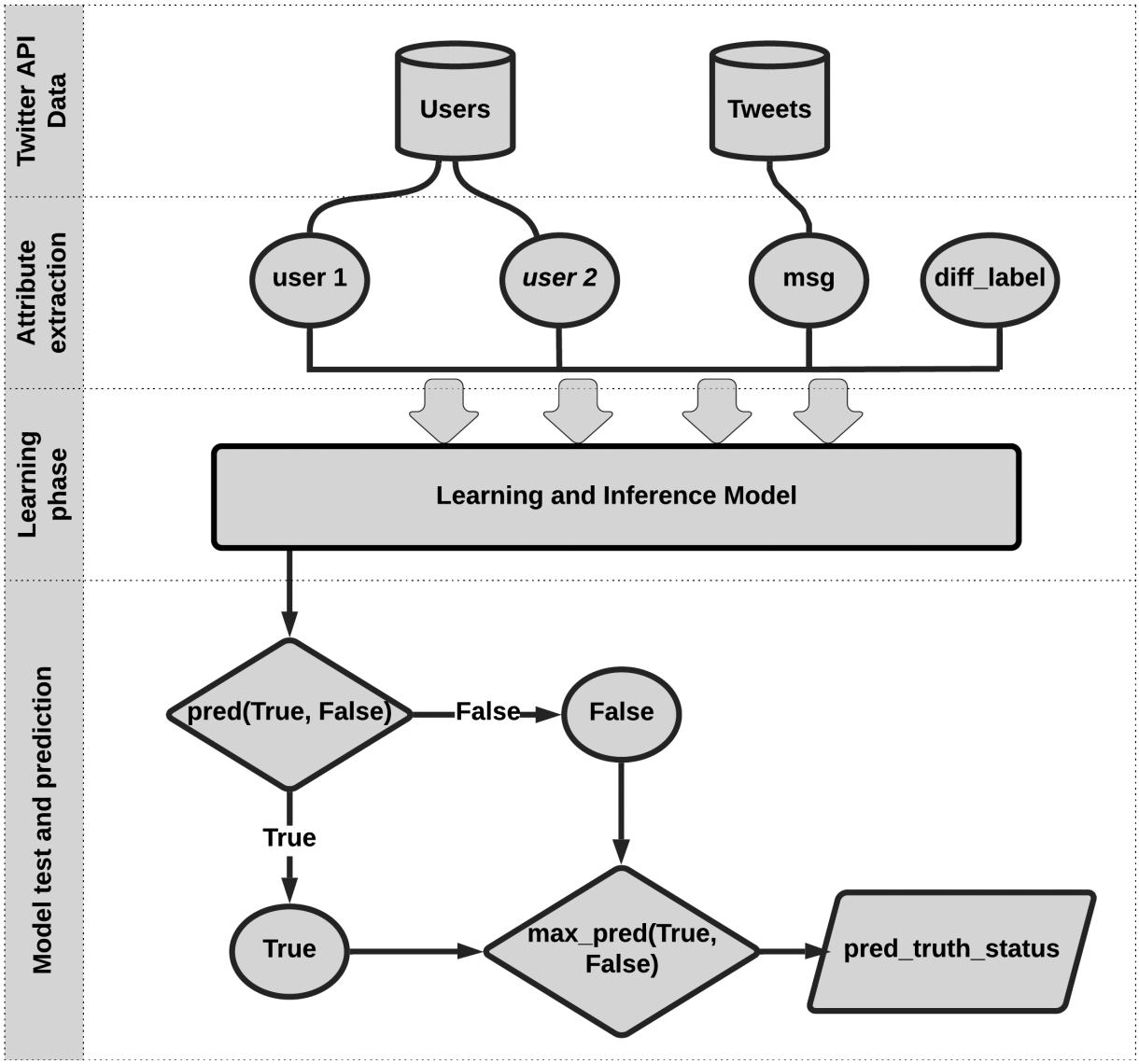

Figure 4.1: *An illustration of experiment setup for the credibility prediction task.*

4.4 Data Collection

We used Snopes[84] to identify topics that have been fact-checked and rated as *True* or *False*. Even though Snopes has different categories including those labeled "Mostly True" and "Mostly False", we restrict this research to those that are strictly labelled *True* or *False*. For each topic, we assign a set of keywords and crawl the Twitter search API using queries of the form $(K_1 \vee K_2 \vee K_3)$, similar to that described by Mathioudakis and Koudas,[85] but with K_i representing the conjunction of possible keyword combinations. For instance, the topic "In a

Category	Topic	Keywords	# tweets
False	Hillary Clinton said "we must destroy Syria for Israel"	Hillary, destroy, syria	96724
	the FBI discovered bones of young children in Jeffrey Epstein's private island	epstein, bones, children	189812
	odessa shooter was "a democrat socialist who had a Beto sticker on his truck"	odessa, shooter, beto, democrat, sticker	19491
True	blood spots visible in the left eye of Joe Biden during a CNN debate in Sept. 2019	Joe, Biden, blood, eye	38983
	anti-abortion rep. DesJarlais encouraged some women to have abortions	abortion, Desjarlais, mistress, republican	15591
	video shows air traffic over the US on 9/11 as thousands of flights were grounded after a terrorist attack	flights, grounded, after, 9/11	10355

Table 4.4: *Topics identified from Snopes, along with the associating keywords used in querying the Twitter seach API.*

leaked e-mail, Hillary Clinton said 'we must destroy Syria for Israel.'" had keywords "hillary, destroy, syria" and query $((hillary \wedge destroy \wedge syria) \vee (hillary \wedge destroy) \vee (hillary \wedge syria) \vee (destroy \wedge syria))$. Table 4.4 gives a breakdown of our topics, along with the associating keywords and number of tweets (including retweets).

We assign *True* or *False* label to each original tweets (no retweets or replies) by examining the content of the post, as tweets could still contain false content even though the topic is *True*. In the dataset, we found a large variation in the volume of tweets in the *True* and *False* collections, with *False* posts accounting for more than 80% of the entire dataset. Also, we observed that the propagation depth of *False* posts ran deeper, with an average retweet depth of 4, while *True* posts averaged a retweet depth of 2. Lastly, we observed a "diffused"/"not diffused" ratio of 35/65 for the tweets in the collection of *True* topics and 45/55 for tweets in the collection of *False* topics. This difference in diffusion rates reveals that *False* posts tend to have more reaction-to-post than *True* posts.

4.5 Prediction Models

Given a collection of messages and an associated user, we recreate the Twitter followership graph by connecting all of the user's followers. Based on the assumption that users will interact with their friends' messages uniquely, we assign the diffusion label as a function of the reaction observed per message and show that this microscopic-level information spread based on the latent message and user interaction attributes is sufficient to give insight to the credibility of a message. We perform two supervised learning tasks by adopting two off-the shelf machine learning models: Bayesian Logistic Regression and Random Forests for prediction and feature selection, respectively.

4.5.1 Predicting credibility of posts

We train a model that predicts if a message is *True* or *False*. We extract the features described previously, and additionally include the diffusion property as an independent variable during the training phase. More specifically, an edge is said to be *diffused* if and only if the destination user (in Twitter terms: *follower*) has reacted (reply, retweet, quote, like) to the friend's (*followee's*) post. We examine how users on Twitter relate with posts of their friends by building classifiers to distinguish user interactions based on the credibility of the message. For a message m, where $m \in \{1, \ldots, M\}$, spread over a network with n interactions, we train a model that predicts the truth status of the message based on the diffusion behavior observed along each one of the n links along which the message propagates. The predicted output is the majority truth status observed across the n interactions. For instance, if a *True* message is spread over 5 interactions and the model predicts the post to be *True* 3 out of 5 times, we accept the output to be *True* and evaluate the model over its correct classification of M messages in the test collection.

4.5.2 Predicting rumor propagation

To further demonstrate the differences in the propagation of *True* and *False* posts, we perform a node-to-node analysis between a pair of users, the spreader and receiver, examining each user's posting behavior, and their interactions to predict the receiver's reaction. Here, we aim to show that our model performs well in an established environment, in order to compare with previous models for propagation prediction. This task is valuable to strengthening our hypothesis that the propagation behavior is a significant attribute to predicting the credibility of a message based on how users in OSN interact with posts of varying veracity.

First, we build separate models for *True* and *False*, performed a supervised learning task using the Bayesian logistic regression by assigning diffusion label "diffused" between a spreader and his follower, if the follower has reacted to an identified tweet (in either case, *True* or *False*) and "not-diffused" otherwise. We adopt an 80-20 train-test split of the data and account for over-fitting by performing 10-fold cross validation. We make predictions on the capability of the model to correctly predict diffusion on the message type and take it a step further by investigating the model's ability to generalize across message type. Then, we build a Random Forests classifier to analyse the importance of the input features and perform selection on the best features for rumor propagation and identification tasks.

4.6 Baseline

We compared the performance of our proposed model to state-of-the-art models in predicting the credibility of posts in social networks.

4.6.1 Emotion-based

Guo et al.,[44] exploits the emotions of both the publisher and receiver of contents to classify posts as fake or not.

4.6.2 Implicit-link

Sampson et al.,[35] use hashtags and web linkage method to link conversations. We tested using the linking method without pruning.

4.6.3 User-behavior

Liang et al.,[38] describes a user behavior-based rumor identification scheme, in which the users' behaviors are treated as hidden clues to identify rumor posts in microblogs.

4.7 Experimental Results

In this section, we report the results obtained from each phase of the experiment.

4.7.1 Predicting credibility by implicit crowdsourcing

While some users react to posts of varying credibility, others only react to tweets that are precisely *True* or *False*. So training a model that learns to distinguish this interaction-reaction relationship is useful for identifying the credibility of a tweet by observing the reaction of a user based on the established interaction between the users. By incorporating the diffusion status of a tweet, we train a model to predict the credibility of the message. The objective of the task is to show that collating the implicitly sourced diffusion behavior between users is useful for predicting the credibility of a post. This implicit crowdsourcing approach is important in real-world situations where there is a need for the system to passively interact with the network. A passive interaction is crucial especially in systems requiring real-time and undetectable communication, for example, an automated rumor identification system for social networking websites.

The result from our experiment validate our assumption that the difference associated with the message, interaction and diffusion patterns of *True* and *False* posts can be exploited in predicting the credibility of messages. By combining these attributes and using the F-score as a measure of accuracy, we were able to achieve 91% accuracy in identifying whether

Model	Precision	Recall	F
crowdsourced	**0.919**	**0.903**	**0.911**
not-crowdsourced	0.838	0.801	0.823
emotion-based	0.798	0.832	0.815
implicit-link	0.861	0.713	0.780
user-behavior	0.753	0.873	0.809

Table 4.5: *Model performance for predicting credibility of a tweet using crowdsourcing techniques.*

messages are credible or not, see Table 4.5. It is important to note that the model is tested using labelled data with existing ground-truth. To show the impact of the diffusion attribute to the credibility prediction task, we carried out a parallel credibility identification task without the diffusion label and observed a performance of 82%. We also show that a comprehensive model exploiting the attributes of the network, interaction and message will perform better than those that use one or the other.

4.7.2 Features analysis for rumor propagation

Establishing a difference in the diffusion prediction models for *True* and *False* posts is amply dependent on showing that there exists a difference between these types of messages and the attributes that steer user reactions. For us to efficiently apply a crowdsourcing approach to the detection of misinformation, we need to differentiate the attributes of *False* posts from those of *True* posts, before we can demonstrate that they diffuse differently. Differentiating between this diffusion pattern is beneficial for the early detection of rumor to mitigate its spread and effect within the network. One justification for using multivariate methods is that they take into account feature redundancy and yield more compact subsets of features, as features that are individually irrelevant may become relevant when used in combination, which also shows that correlation between sets of features does not necessarily imply redundancy. Considering that the goal of the feature analysis task of this study is to identify the optimal set of features necessary to maximize diffusion prediction irrespective of credibility-status, we train a random forests model and then select the top 20 features for the rumor propagation tasks.

Rank	False	True
1	MSG is RT	MSG is RT
2	MSG favorited count	social homogeneity
3	MSG has mentions	MSG favorited count
4	dest tweet with hashtag	src tweets with URL
5	src retweet-to-tweet	MSG feedback intent
6	MSG news intent	MSG positive sentiment
7	src followers count	src directed tweet
8	MSG has URL	MSG has URL
9	src followers-friends	src avg favorite-tweet
10	src account age	src avg tweet/day
11	src tweets with URL	src followers count
12	MSG fear emotion	MSG has mentions
13	dest directed tweet	src account age
14	src status count	dest retweet-to-tweet
15	src friends count	src retweet-to-tweet
16	social homogeneity	src has URL
17	MSG RT count	dest follower-friends
18	dest friends count	src status count
19	MSG positive sentiment	MSG has hashtag
20	MSG negative sentiment	MSG RT status

Table 4.6: *Top 20 features for efficient diffusion prediction of True and False posts selected using Random Forest classifiers.*

From the ranked features in 4.6, we see that in tweets with *False* status (Rumor), the attributes of the message account for 45% of the ranked features with the combination of network and interaction accounting for 55%, while message attributes account for 40% of top ranked features for *True* posts. As anticipated, the latent attributes of the message rank in the top features for both *True* and *False* models, confirming that the meaning, intention and emotions of messages influence users' decisions in the diffusion process. From the ranked features, we can infer that rumor posts masked as news, meant to incite fear will diffuse better than others. However, it is surprising that the diffusion of rumor posts cannot be strictly tied to their sentiment as we observed that both negative and positive sentiments contribute equally to the performance of the model. Even though it ranks differently in both models, social homogeneity ranking well in both models shows that a user will most likely respond to the post of someone with interests similar to his own.

4.7.3 Predicting Rumor Propagation

We focus on the problem of predicting the diffusion decision (to react or not) of a user based on his perception of the message and interaction with the spreader of the information. In this model, we do not take into account the effect of previous exposure to similar posts, or the popularity of the message, we simply make an inference on whether a user will retweet, share, quote or favorite a tweet by estimating the probability of diffusion.

In Table 4.7, we show the performance of the model across message type, using the performance metrics previously highlighted. The model achieved 91.6% and 89.9% prediction accuracy for message with *True* and *False* status respectively.

Model	Precision	Recall	F
False	0.897	0.902	0.899
True	0.908	0.925	0.916

Table 4.7: *Model performance for predicting diffusion of True and False posts of a post.*

To show that the proposed model can be effectively transferred across topics and credibility status, we tested our model's performance over topics outside the training list. The results for inter-topic and inter-credibility prediction tasks are reported in Table 4.8. For inter-topic test, we observed performance of similar magnitude in diffusion prediction capabilities when the models are exposed to topics outside the training list. As observed from the table, there is a difference for inter-credibility test and we believe this is due to the difference in the features that influence diffusion for the message types. This result piques our interest because it shows that the properties of *True* and *False* posts are distinct enough that either model can discriminate significantly between each type of post.

Model	Precision	Recall	F
False	0.887	0.889	0.882
True	0.899	0.919	0.908
False model-True test	0.856	0.821	0.838
True model-False test	0.849	0.921	0.884

Table 4.8: *Model performance for inter-topic, inter-credibility diffusion prediction.*

Chapter 5

Implicit Crowdsourcing for Identifying Abusive Behavior in Online Social Networks[3]

Not all messages shared with abusive intent are written crudely. A message intended for cyber abuse might be concealed in sarcasm, emotions, and sentiments even if it appears otherwise. We hypothesize that there is a difference in the diffusion of abusive posts in OSNs, a difference that can be leveraged by using a crowdsourcing approach to predict the abusive label associated with these posts. We believe that some users are more likely than others to create, share, and/or react to posts meant for cyber abuse. These users will serve as discriminators in the detection model to sieve out outliers who do not contribute much to the detection task. We adopt an implicit crowdsourcing model by simply observing users' interaction with posts of varying abuse levels to ensure that the user's posting and reaction behavior is as natural as it can be.

5.1 Overview

In this study, we introduce an automated model for predicting the abuse level associated with a tweet predicated on the interaction between users encountering the messages. We describe two types of posts, *normal* and *abusive*. A post is said to be *abusive*, if and only if the content or context associated aligns with the intent for cyber abuse. Seeing as feature design and selection strongly impact a machine learning model's accuracy much more than the model used[82]. We train a Bayesian Logistic Regression model by incorporating user, message, and propagation features to estimate the node-to-node influence dynamics to the propagation of abusive posts. We describe two tasks in identifying abusive behavior online and adopt supervised machine learning models in addressing them.

1. We investigate abusive behavior prediction by exploring abuse propagation founded on microscopic-level information spread. By observing the spreading behavior of posts of varying abuse levels in online social networks, we propose a model that predicts the abuse level associated with a tweet by observing the latent attributes of the message, along with those of the users, and their reactions over the network.

2. We evaluate the role of user and message features in detecting the abuse level of a post, by measuring the contributions of individual users and their posts to the spread of abusive posts in OSNs.

Previous crowdsourcing-based approaches in abuse detection in social networks focus on conversation or account annotation for abuse detection. To the best of our knowledge, this is the first research that explores crowdsourcing as an automated tool for identifying abusive behavior in online social networks. We classify users based on the types of posts they generally react to: (i) reacts to only *normal* posts, (ii) reacts to only *abusive* posts and (iii) reacts to a mix of *normal* and *normal* posts. Users in class i and ii are good discriminators for both abuse level detection and feature identification, while users in class iii do not serve as good discriminators in the prediction model.

5.2 Model and Method

We propose a framework that given a tweet will predict the abuse level by observing the user interaction with the tweet – we leverage "the wisdom of the crowd" as it is often used in a crowdsourcing approach to assigning a label to the post. The proposed model differs from current crowdsourcing techniques in that it makes an inference from a supervised learning task and does not require a human annotator. Since this is a supervised learning task, the model requires labeled data and makes use of manually annotated tweets for learning and inference. To each user, we associate a total of 16 features, including 3 network and 13 interaction attributes; and to each message, we associate 11 attributes. We then train a Bayesian logistic regression model based on the prediction task.

5.3 Data Description

In this study, we make use of the ICWSM 2020 task 2 dataset made publicly available by Founta et al.[57] The dataset contains 100k annotated tweets associated with inappropriate speech labeled as abusive and hateful speech, as well as normal interactions and spam. For annotation, Founta et al.[57] defined the labels as:

- Abusive Language: Any strongly impolite, rude or hurtful language using profanity, that can show a debasement of someone or something, or show intense emotion.
- Hate Speech: Language used to express hatred towards a targeted individual or group, or is intended to be derogatory, to humiliate, or to insult the members of the group, on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender.
- Spam: Posts consisted of related or unrelated advertising / marketing, selling products of adult nature, linking to malicious websites, phishing attempts and other kinds of unwanted information, usually executed repeatedly.
- Normal: all tweets that do not fall in the defined categories.

Because the model we propose makes use of attributes of the user for inference, we used a tweet Hydrator[86] - an Electron-based desktop application for hydrating Twitter ID datasets. The Hydrator helps us turn tweet IDs back into JSON, retrieving information contained in the Tweet and User Objects.

Since tweets get deleted from time to time, by the user or Twitter, some tweets were no longer available through the Twitter API, as such, we had a reduced number of tweets after hydration. The dataset contained over 69k tweets with 56k unique users. Tweets labeled normal made up 62% of the dataset, abusive tweets accounted for 20%, spam tweets constituted 14%, while hateful tweets formed 4% of the data.

We recreate the Twitter followership graph for the available dataset by associating an edge between two users if there is a follower-followee relationship between them. Based on the assumption that users will interact with their friends' messages uniquely, we assign the diffusion label as a function of the reaction observed per message and show that this microscopic-level information spread based on the latent message and user interaction attributes is sufficient to give insight to the abuse level of a message.

5.4 Task 1: Implicit Crowdsourcing for Predicting the Abuse Level of a Tweet

First, we demonstrate the differences in the propagation of *Abusive* and *Normal* posts, we perform a node-to-node analysis between a pair of users, the spreader and receiver, examining each user's posting behavior, and their interactions to predict the receiver's reaction. Here, we aim to show that our model performs well in an established environment, to compare with previous models for propagation prediction. This task is valuable to strengthening our hypothesis that the propagation behavior is a significant attribute to predicting the abuse level of a message based on how users in OSN interact with posts of varying veracity. We believe that a tweet that is abusive or hate speech will stir up reaction from many users in the network, causing it to propagate farther than a normal or spam post will. Then, we train

a model that predicts the abusive label associated with a message. We extract the features described, and additionally include the diffusion property as an independent variable during the training phase.

More specifically, an edge is said to be *diffused* if and only if the destination user (in Twitter terms: *follower*) has reacted (reply, retweet, quote, like) to the friend's (*followee's*) post. We examine how users on Twitter relate with posts of their friends by building classifiers to distinguish user interactions based on the abusive label associated with the message. For a message m, where $m \in \{1, \ldots, M\}$, spreading over a network with n interactions, we train a model that predicts the abusive label associated with the message based on the diffusion behavior observed along each one of the n links along which the message propagates. The predicted output is the majority abusive label observed across the n interactions. For instance, if an *abusive* message is spread over 5 interactions and the model predicts the post to be *abusive* 3 out of 5 times, we accept the output to be *abusive* and evaluate the model over its correct classification of M messages in the test collection.

By comparison, the non-crowdsourced model relies solely on the features of the message and those of the original creator of the message in making a decision on the truth-status of the tweet.

5.5 Task 2: Estimating Features Contributing to Abusive Tweet Propagation

We perform a supervised learning task where we train the model using the attributes from a pair of nodes with an established followership relationship and label the interaction between them as either diffused or not diffused. The attributes learned are said to be representative of users' network, interaction, participation, role, and importance in the spread of information to other nodes in the network. As previously stated, these attributes are learned over four different time intervals. After learning these features, we fit a regression function that maps the learned user attributes to the likelihood of diffusion between the nodes. In this task, we

describe two sub-tasks:

1. We build a persona for the user to evaluate the user's tendency to post or react to posts of different abuse levels.

2. We explore the node-to-node relationship between users and seek to identify features that cause abusive posts to propagate.

For each user in the dataset, we assign an abusive, hate, spam, or normal score computed as a ratio of their post that is labeled as such. For each label i, where $i \in \{abusive, hate, spam, normal\}$

$$score_i = \frac{count_i}{N} \tag{5.1}$$

where N is the total number of tweets the user has in the collection.

5.5.1 Task 2.1: Evaluating a user's tendency to post or reactive to abusive posts

For the purpose of this data challenge, we assume a user's total tweets to be limited to the data in the collection. However, for a more robust prediction task, it is important that the score estimated in Eq (5.1) is estimated over the tweets shared on the user's timeline. For each user, we create an online persona by combining the user features, message-based features (over all of the user's messages) and the estimated abusive scores. We perform regression analysis on this behavioral pattern and then train a Random Forest classifier to rank the features that directly impact the probability that a user will post or react to a message that is labeled abusive or hateful. Currently, we limit the prediction task to focus on estimating a user's probability to post or react to abusive and/or hateful posts as these kinds of behavior are not as widely studied as spam.

5.5.2 Task 2.2: Identifying features for abusive post propagation

We perform a supervised learning task where we train the model using the attributes from a pair of nodes with an established followership relationship and label the interaction between them as either diffused or not diffused. The attributes learned are said to be representative of the user's profile, messages, network, interaction, participation, role, and importance in the spread of information to other nodes in the network. After learning these features, we fit a regression function that maps the learned user attributes to the likelihood of diffusion between the nodes. Then, we present a ranking of the features that contribute to the likelihood of abusive and hate tweets diffusing over the network.

5.6 Results

5.6.1 Predicting Abuse Level using Implicit Crowdsourcing

By simply observing the reaction generated between users in the network, we train a model that learns to distinguish the interaction-reaction relationship. The model's ability to effectively distinguish the uniqueness of this relationship over messages of different abuse levels is useful in detecting the abuse label associated with a message by observing the reaction and in turn, propagation of the message over the network. As previously stated, the objective of this task is to show that collating the implicitly sourced diffusion behavior between users is useful for detecting the abusive behavior of a post. This implicit crowdsourcing approach is important in real-world situations where there is a need for the system to passively interact with the network.

We carry out prediction tasks to detect the abuse levels associated with a tweet and we included an additional task to predict if a post is offensive (abusive OR hate). We create an offensive set of the data by combining posts previously labeled as abusive or hate. In Table 5.1, we show the performance of the model using the evaluation metrics described in Section 3.4, with the dataset split in a 60-30-10 train-validation-test ratio. We present the precision, recall, and F1 scores for the crowdsourced model (CRO - *) and differentiate

Abuse level	CRO- Precision	CRO- Recall	CRO- F1	nonCRO- Precision	nonCRO- Recall	nonCRO- F1
Abusive	0.85	0.82	**0.83**	0.65	0.71	**0.68**
Hate	0.82	0.89	**0.85**	0.61	0.63	**0.62**
Spam	0.90	0.92	**0.91**	0.67	0.69	**0.68**
Normal	0.95	0.90	**0.92**	0.80	0.85	**0.82**
Offensive	0.94	0.97	**0.95**	0.74	0.79	**0.76**

Table 5.1: *Model performance in abuse level of a post.*

it from the non-crowdsourced (nonCRO - *) model. From the results, we observed that prediction tasks using the crowdsourced model performed considerably better recording over 20% improvement than the non-crowdsourced model. The prediction result for posts labeled as *normal* is unsurprising because the model had more data to learn from than the other labels. For a model to implicitly predict the abuse level associated with a tweet, there is a need to learn from the user's prior interactions with the network. We further argue that a user's likelihood to create or react to an offensive post will influence their future interactions with similar posts. One can also argue that increased exposure of a user to offensive posts in their network will increase his chances of posting the same.

5.6.2 Features for Abusive Behavior Propagation

Here, we model the user's participation in the spread of abusive posts on Twitter and use the knowledge to measure the contribution of individual user in the creation and spread of abusive posts in OSNs.

Features impacting user's propensity for abusive posts

On Twitter, a user can show his/her interest in a topic by contributing to the topic through the creation of posts, retweets, replies, quotes, etc. By contributing to a given topic, users give little hints into their interests, possibly patterns to their behavior and expected reactions. We group tweets labeled as abusive and hate together as abusive posts. Even though the data is heavily skewed towards normal posts, result from the experiment shows that the sentiment around the topic plays a major role in whether a user will post something abusive

about it. Following closely to sentiment is the abusive score, this is unsurprising because a user with a high abusive score will most likely keep posting abusive tweets. As expected, the normal score ranks side by side the abusive score as they complement each other. Ranked next to that are the user favorite and average tweet per day. We observed that users with more friends than followers are more likely to exhibit abusive behaviors.

These features are only descriptive of the user's own tendencies towards abusive posts. It is important to note that identifiable events in the network can also contribute to a user's disposition to share abusive posts at a particular point in time. The presence of media (such as memes, emojis, and sometimes images with text) poses a challenge to this task as some of these media might contain offensive content that the model is unable to interpret. Due to the fluidity of the Twitter interaction, language, and user interests, we believe this task will perform better as a semi-supervised learning task where the model learns to adapt to the dynamic nature of the Twitter network.

Features impacting abusive post propagation

Establishing a difference in the diffusion prediction models for *abusive* and *normal* posts is amply dependent on showing that there exists a difference between these types of messages and the attributes that steer user reactions. In previous tasks, we have shown that the detection models differ from one abusive label to another, here, we show that the messages propagate differently by providing evidence that the attributes contributing to diffusion differ between *abusive* and *normal* posts. Please recall that *abusive* posts are described as tweets in the data associated with abusive and hate labels.

To further validate our assumption that there is a difference associated with the message, interaction and diffusion patterns of *abusive* and *normal* posts, we use random forest classifiers to provide the top-10 features, see Table 5.2 that aid in the propagation of *abusive* and *normal* messages. In this task, we model the diffusion of posts from one user to another and observe the reaction of the receiving user. We learn a function that maps what features of the source and destination users cause a reaction or otherwise. We measure the model's

Rank	Abusive	Normal
1	dest friends count	MSG is RT
2	dialogue	MSG has mentions
3	MSG has mentions	MSG favorited count
4	dest tweet with hashtag	src tweets with URL
5	src retweet-to-tweet	dest retweet-to-tweet
6	src status count	MSG sentiment score
7	src followers count	src directed tweet
8	MSG sentiment score	dialogue
9	src followers-friends	src avg favorite-tweet
10	MSG has hashtag	dest follower-friends

Table 5.2: *Top 10 features for predicting propagation of abusive and normal posts selected using Random Forest classifiers.*

ability to correctly predict a user's reaction based on the learned function.

From the ranked features, we see that the sentiment score and established dialogue is deemed important in the diffusion of posts (either abusive or normal) between two users. The influence of the source user greatly impacts propagation of abusive posts as we see from the network features ranking in the Top 10. One thing to note here is that a single user can act as both the source or destination node in the network, depending on his role at a particular point in time. Additionally, the presence of hashtag(s) in a tweet has an effect on its likelihood to get a reaction.

Chapter 6

Heuristic Gradient Optimization Approach to Controlling Susceptibility to Manipulation in Online Social Networks

The search to find a balance between privacy preservation and social influence leads us to ask if users can be given control over their own privacy. Our research aims to directly address this question by allowing the users to examine their likelihood for manipulation based on their social interactions, giving them insight into their degree of exposure so that they can choose how much protection needs to be implemented based on their privacy needs. The solution proposed suggests attributes that can be tweaked to minimize the user's exposure. This solution can then bring about questions from that user's network about how to identify authenticity in their network. But this is a whole different challenge centered around identifying fake profiles, bots, spammers, and misinformation spreaders.

We examine manipulation gain in terms of a user's susceptibility to targeted manipulation through inferential attack in a single tweet. We propose a model that first measures a user's probability of engaging with a post in a neutral environment and then measures the degree

of deviation of this probability when a profile and posts from that profile are targeting them. By doing this, the user has an idea of how much a particular friend in their network can cause them to change the way they interact. This change can be seen as manipulation gain because it is caused by how much information can be inferred from their activities. We describe 2 categories of features: user and message, that describe the user's network and the messages they create and interact with. We examine a scenario where the sender of a message tries to mislead the receiver by optimizing their (sender) attributes and those of the message to mimic what the receiver will typically show interest in. By doing this, the receiver is gradually manipulated to engage with a post that they would otherwise ignore. The receiver is said to be manipulated if the probability of engagement with the targeted message deviates from what it would have been if the message were produced in the absence of inferred knowledge about that specific user.

Existing models for preserving privacy through inferential attacks in OSNs focus on data sanitizing and anonymization. Privacy protection technique by Talukder et al.[65] proffer self-sanitization as a way to address information leakage but the sanitization itself is to be done by the user's friends. With little efforts in limiting manipulation in the wild, it is essential that the user has more control on how accurately spammers, learning models, and third-party vendors make inferences about them. We hypothesize that the profile and posts of the user are representative of their true self and can be used to make observations about them. We use a Gaussian Process Classification model to learn the user's probability to react to a post shared by their friend, and then use gradient optimization methods to heuristically search for attributes that can be optimized to limit their likelihood to respond to a targeted post. The proposed model provides a module for users to protect themselves by including noise in their profiles to minimize their susceptibility to this targeted attack. In the context of social networks, targeting can be in the form of carefully crafted messages, accounts (bots and trolls),[87] and in some cases both. In this study, we provide OSN users the opportunity to measure their manipulation gain when they choose to interact as they usually would, and compare it with their manipulation gain when they add a layer of security to muddle up an attacker's opinion of them.

6.1 Model

6.1.1 Dataset

To generate the features required for the model, we collect the metadata of Tweet and User JSON (JavaScript Object Notation) objects. We adopted a previously created tool made publicly available on GitHub, that crawls the Twitter Search API using the usernames or IDs. In collecting Twitter dataset aimed at general conversations, bias can be introduced due to events happening in the real world. For the purpose of this research, we adopted the crawler to collect streams of tweets and associated user profiles instead of specifying usernames, IDs, topics or demographics, with the data collected over multiple 7-day periods.

We remove accounts with no followers and/or activities, and then focus on the set of users with established followership relationship by creating a relationship graph by connecting the users in our dataset. We also limit our data collection to posts in English language. Table 6.1 shows the distribution of the dataset after cleaning and pre-processing. For each tweet in the dataset, we associate the profiles of the sender (associated account), receiver (the sender's follower), and a binary *engagement label*, see Eq 6.1, if the receiver has generated some reaction to the post.

Data Statistics	Count
Total Number of Users	12200
Follower-Followee relationships	46400
Total Number of Tweets	81500

Table 6.1: *Data Distribution.*

6.1.2 Attributes

The proposed framework makes use of 2 categories of features, see Table 6.3: one for the user and the other for the message. For each user (sender or receiver), we learn 8 directly observable attributes and a social homogeneity feature that is common to both of them. In microblogs such as Twitter, a *friend* is someone a user follows and that user can see all of their friends' posts. We consider features describing the user's network: *followers*

Notation	Description
τ_i	response from user u_i
θ_i	attributes of user u_i
θ'_i	disclosed attributes of user u_i
M_{ij}	message attributes from user u_i to user u_j
M'_{ij}	optimized message from user u_i to user u_j
ϵ	measure of manipulation gain
β	probability of response to random message
γ	probability of response to targeted message

Table 6.2: *Summary of Notation.*

count, *friends count*, and features describing the user's interactions with the network. These attributes have been extensively studied in the works of by Castillo et al.,[37] Liang et al.,[38] and Yang et al.[40] These attributes are important because a user's friends impact the kind and volume of messages that end up in their timeline and the higher the number of followers, the farther the possibility of reach. This is also reflected in policies by OSNs like Twitter and Instagram who attach value to the followers count, where users become verified once they cross a certain threshold, even if the account holder is not a celebrity or public figure.

Twitter posts are very fluid, taking up various forms and content, we learn 6 attributes of the message that describes the content, popularity and sentiment associated with them.

6.1.3 Learning Model

In adversarial search, the adversary interacts with a network with the purpose of getting nodes to behave in a predetermined way, even if it deviates from how the nodes will normally act. In inferential attacks, adversaries will first learn their opponent's behavior and try to mislead the nodes in the network into accepting the wrong hypothesis by modifying the way they use their resources. In this work, the sender of a message (adversary) learns the behavior of their network by observing activities generated by their followers in terms of topics their followers are interested in, inferring latent attributes and demographic attributes that might not have previously been shared by their followers. The sender can then use this learned information to adapt their profile and posts into one that their followers will find interesting

	Feature	Description
	User Features	
1	tweets containing URL	number of user's tweets containing URLs
2	presence of user description	shows user's profile has description (bio)
3	user verified	shows if user's account is verified
4	number of followers	higher follower count means higher reach
5	number of friends	average number a user follows
6	account age	account age in days
7	status count	total number of posts over account's lifetime
8	user favorites count	number of user's tweets endorsed by others
9	social homogeneity	depicts common friends
	Message Features	
10	presence of hashtags	shows if a tweet contains hashtags
11	presence of URLs	shows if a tweet contains URLs
12	presence of media	shows if a tweet contains media
13	tweet favorites count	favorites count for tweet
14	retweet count	retweet count for tweet
15	sentiment score	sentiment score for tweet

Table 6.3: *Attribute Description.*

enough to interact with. The receiver (the adversary's follower) is gradually manipulated into engaging with an account that they would generally not interact with because the posts coming from such an account mimic what the receiver will typically show interest in. We define manipulation in terms of a user's engagement with a post caused by inferences made on the user's behavior. Engagement is simply the generation of a reaction from the receiver of the post in the form of a reply, retweet, favorite, share, or like. A user is said to be manipulated if the probability of engagement with a targeted message deviates from what it would have been if it were a message produced in the absence of any knowledge about the specific user. It should be noted that this deviation can be positive or negative and it is only meant to show that there is a change from the user's regular behavior.

The objective of the sender is to vary the attributes of its messages to maximize the probability of engagement from the receiver of the message. The assumption is that both the sender and receiver are unaware of the true state of the network and the directly observable attributes are taken to be the true states of either of the two parties. The receiver reacts by varying its published features in accordance with the goal of an adver-

sary/manipulator/sender. The receiver aims to minimize the absolute difference between their probability of interaction with a targeted message and the probability of interaction with a random post on their timeline. At each point of optimization, the sender/receiver takes the observable attributes as the true attributes of the other. Both parties can only modify their attributes before disclosing them and cannot make changes after they are disclosed. The receiver (user B) can only modify their features, while the sender (user A) can optimize over either their own attributes or those of the message.

We start by studying a user's interaction with posts in their timeline, Eq (6.1), and learn the probability that the user will respond to a random post on their timeline.

$$\tau_i = \begin{cases} 1, & \text{if engagement is observed} \\ 0, & \text{otherwise.} \end{cases} \tag{6.1}$$

The probability that user B (receiver) with attributes θ_B will engage with a random message m_{AB} from user A (sender) with attributes θ_A, given the true attributes of user B is:

$$\beta = P[\tau_B = 1|\theta_A, M_{AB}, \theta_B]. \tag{6.2}$$

If user B chooses to modify their disclosed attributes, thus disclosing θ'_B in place of θ_B, then from user A's perspective, the probability that user B will engage with a message m_{AB} is given by:

$$P[\tau_B = 1|\theta_A, M_{AB}, \theta'_B]. \tag{6.3}$$

Since user A is only privy to the disclosed attributes, user A optimizes over their own attributes θ_A and over the message attributes M_{AB} to ensure engagement with the post, such that:

$$[\theta'_A(\theta'_B), M'_{AB}(\theta'_B)] = argmax_{\theta_A, M_{AB}} P[\tau_B = 1|\theta_A, M_{AB}, \theta'_B]. \tag{6.4}$$

In order to minimize the chances of being manipulated, user B needs to publish an

optimal set of attributes $\theta^*(B)$, such that:

$$\theta^*_B = argmin_{\theta'_B} P[\tau_B = 1|\theta'_A(\theta'_B), M'_{AB}(\theta'_B), \theta_B]. \tag{6.5}$$

This leads to user A finding the optimal

$$[\theta^*_A(\theta^*_B), M^*_{AB}(\theta^*_B)] = argmax_{\theta_A, M_{AB}} P[\tau_B = 1|\theta_A, M_{AB}, \theta^*_B]. \tag{6.6}$$

The challenge with this is that the true probability of engagement is based on user B's true attributes, that is:

$$\gamma = P[\tau_B = 1|\theta'_A, M'_{AB}, \theta_B]. \tag{6.7}$$

The model is intended such that user B sets a threshold ϵ on the maximum allowable deviation of γ from β.The user then works to ensure that

$$|\gamma - \beta| \leq \epsilon \tag{6.8}$$

at every point in time. Manipulation is said to be successful if $|\gamma - \beta| > \epsilon$. The goal of the adversary (user A) is to maximize the LHS of Eq (6.8) while user B focuses on minimizing manipulation gain, ϵ arising from inferential attacks by shrinking that value as much as possible.

6.2 Experiment

The task is to minimize the probability of the user reacting to a targeted message, so that it is similar to the likelihood that the user will react to a randomly crafted message. Since test predictions take the form of class probabilities, we adopt Gaussian process classification[88] to learn the probability of engagement for each data point.

We learn a function F, over the attributes of user A, disclosed attributes of user B and message attributes m_{AB} that produces the probability of engagement P_r

$$F(\theta_A, \theta_B, m_{AB}) \rightarrow P_r. \tag{6.9}$$

6.2.1 Gaussian Process Classification Model

The Gaussian Process Classification (GPC) model is capable of making fine distinctions in the sense that it models $p(y|x)$ as a fixed Bernoulli distribution. In the GPC model, inference is made from the latent function f given the observed data $D = \{(x_i, y_i)|i = 1, ..., n\}$, with $f_i = f(x_i)$, $f = [f_1, f_2, ..., f_n]$, $X = [x_1, x_2, ..., x_n]$, and $y = [y_1, y_2, ..., y_n]$, where X is the collection of inputs and y are the class labels. x_i is a vector representing the sender, receiver and message attributes, while y_i is a binary value depicting if the receiver has responded to the message from the sender. The GPC model requires specifying a kernel that observes the inputs X and class labels y and defines the covariance function of the data. Inference is then made by computing the distribution of the latent variable corresponding to a test case, and subsequently using this distribution over the latent function f to produce a probabilistic prediction.

In GPC problems, the posterior presents to be analytically intractable and inference involves adopting approximation techniques. We adopt a GaussianProcessClassifier[89] that implements the logistic link function. The integral of this function cannot be computed analytically but is easily approximated in the binary case such as ours.

For probabilistic predictions, we performed RBF kernel search with different choices of hyperparameters and found that parameters set at $ConstantKernel(1.0) * Matern(length\_scale = 1, nu = 1.5)$ performed best for the dataset. The GaussianProcessClassifier then approximates the non-Gaussian posterior with a Gaussian based on the Laplace approximation technique.

The complexities associated with predicting probability estimates pose concerns on how confident we are about the calibration of the model. We adopt the Expected Calibration Error (ECE) as a way to measure miscalibration. ECE is defined as difference in expectation between confidence and accuracy, this is estimated by simply taking the weighted average

over the absolute accuracy/confidence difference.[90–92] Larger ECE values show larger difference between output confidence and actual model accuracy of the prediction — larger miscalibration, while smaller ECE values indicate less miscalibration. We write

$$ECE = \sum_{i=1}^{K} P(i).|o_i - e_i|, \tag{6.10}$$

where K is the number of bins, o_i is the true fraction of positive instances in bin i, e_i is the mean of the post-calibrated probabilities for the instances in bin i, and $P(i)$ is the empirical probability (fraction) of all instances that fall into bin i.[90] For a properly calibrated model, we test the Platt scaling and isotonic calibration techniques. Platt scaling fits a univariate logistic regression model over the data by transforming classification output into probability distribution. The isotonic calibration technique is similar to Platt scaling but it is a non-parametric regression technique that makes no assumptions on the form or relationship between variables.

6.2.2 Gradient Optimization

We are faced with a minimization and maximization problem where the receiver looks to minimize their probability of reaction while the sender is looking to maximize this probability. In finding these best values, we explore the gradient method of optimization with the search directions defined by the gradient of the function at the current point, i.e., descent for receiver and ascent for the sender. Due to the intractability of the GPC model for numerical computation, we treat the function f as a black-box oracle where at each iteration, we provide the data point and receive the output which is a partial derivative of the attribute vector derived by searching the attribute space to find the values that move us closer to the solution.

We adopt an iterative gradient optimization approach, where at each iteration, we perform a gradient ascent or descent. The gradient is estimated with respect to the features and move in the direction of the gradient in maximization tasks (gradient ascent), but in the opposite direction of the gradient in minimization tasks (gradient descent). The learning

rate is an arbitrary value that controls how quickly we ascend or descend. We set multiple learning rates of 0.01, 0.001, and 0.0001 but discovered that a rate of 0.001 performed best in the model. Since we are performing a greedy search, the first solution is always accepted as optimal.

Mixed discrete-continuous variables pose a bound constraint to search in gradient-based optimization approach since discrete variables often derived from categorical (or binary) values have no ordering and offer no meaning to the learning model. This constraint makes computing gradient values for continuous and discrete variables simultaneously challenging as a 0.001 step to a binary value does not carry the same as it does for a continuous variable. To address this constraint, we adopt a search method that does not update both continuous and discrete variables simultaneously.[93;94] Instead, the search algorithm performs a search in the continuous space, and then searches in the space of discrete variables to find the optimal gradient. In our case, our discrete variables are binary, so a flip in the discrete sub-vector space was needed for computation. This variable flipping is done per variable in the sub-vector to find the direction of the gradient.

6.2.3 Optimization over Multiple Connections

The task outlined in Section 6.2.2 describes a one-to-one relationship between a user and their connection. In reality, a user ideally has many followers and friends, and it is expected that they will make optimizations to their profiles based on these relationships and not just one. For user A with a set of followers $\mathcal{B} = \{B_1, B_2, B_3, ..., B_n\}$, user A needs to find an optimal θ_A, M_{AB} that will maximize the probability of reaction over their set of followers. It is important to note that this optimal θ_A, M_{AB} might not be the optimal for each B_i in $\mathcal{B}$, but it is considered optimal in the sense that it causes a noticeable deviation over the entire set. This optimal value θ_A, M_{AB} can be estimated from Eq (6.4) as:

$$[\theta'_A(\{(\theta'_{B_i})\}), M'_{AB}(\{(\theta'_{B_i})\})] = argmax_{\theta_A, M_{AB}} F(\{P[\tau_{B_i} = 1|\theta_A, M_{AB}, \theta'_{B_i}]; 1 \leq i \leq n\}) \tag{6.11}$$

where $F(p_1, p_2, ..., p_n)$ is the aggregate function over $\mathcal{B}$ and can be the max, mean, or median; but we opt for the mean function in our case.

Each user B_i, is looking to minimize its probability of reaction by disclosing a set of attributes:

$$\theta'_{B_i} = argmin_{\theta_B} P[\tau_{B_i} = 1|\theta'_A(\{(\theta'_{B_i})\}), M'_{AB}(\{(\theta'_{B_i})\}), \theta_B] \tag{6.12}$$

but the optimal θ'_{B_i} can only be learned by examining the disclosed attributes of other B_is,

$$\theta'_{B_i} = F_i(\{\theta'_{B_j}, j \neq i\}, \theta_{B_i}), \tag{6.13}$$

that is:

$$\begin{aligned}
\theta'_{B_1} &= F_1(\theta'_{B_2}, \theta'_{B_3}, ...\theta'_{B_n}, \theta_{B_1}) \\
\theta'_{B_2} &= F_2(\theta'_{B_1}, \theta'_{B_3}, ...\theta'_{B_n}, \theta_{B_2}) \\
\theta'_{B_3} &= F_3(\theta'_{B_1}, \theta'_{B_2}, ...\theta'_{B_n}, \theta_{B_3}) \\
&\vdots \\
\theta'_{B_n} &= F_n(\theta'_{B_1}, \theta'_{B_2}, ...\theta'_{B_{n-1}}, \theta_{B_n}).
\end{aligned} \tag{6.14}$$

The optimal solution for Eq (6.14) is a fixed point. Similar to the setup in Section 6.2.2, user A seeks to maximize the probability of reaction by disclosing the attributes θ'_A and M'_{AB} estimated with respect to the disclosed attributes of their followers. Unlike single user targeting where the receiver simply optimize their own attributes given the sender and message attributes, in the multi connection optimization, the receiver, user B_i, has to make assumptions on the attributes of their neighbors (other followers of user A). The user B_i discloses attribute θ'_{B_i} that is aggregated over their true attributes θ_B and the disclosed

attributes of their neighbors. At every iteration, the user B_i continues to estimate θ'_{B_i} by observing changes in their neighbors. Convergence occurs when user B_i finds a fixed point when a change to θ_{B_i} does not lead to a noticeable change in the probability of reaction.

6.2.4 Constraints over Social Influence

In previous sections, we attempt to fulfill the privacy needs of OSN users by minimizing their propensity for manipulation. In the process of introducing some noise into their online personae, user B is exposed to the risk of losing their social influence. From OSN interactions, it is safe to establish that while some users are concerned about manipulation gain, they would not be willing to implement privacy preservation mechanisms if it will have a negative impact on their social influence. For some, this influence translates to a monetary value often described as "social currency" [95] which is a brand's followers, likes, comments, shares and views. It is the extent to which people basically share the brand information — or lifestyle, for those influencers who share part of their everyday lives.

Here we describe social influence, ρ, as the percentage of reactions observed on user B's post. From user B's perspective, a privacy preservation mechanism will minimize their manipulation gain ϵ, from Eq (6.8), and change in social influence. A change in social influence $\delta\rho$ defined as:

$$\delta\rho = \rho_\gamma - \rho_\beta \tag{6.15}$$

where ρ_β is social influence estimated in a neutral network before any optimization, and ρ_γ is social influence estimated after the user has optimized their profile to minimize manipulation gain through inferential attack. A negative $\delta\rho$ indicates influence loss and a $\delta\rho = 0$ is desired.

To compute ρ, we extend the crowdsourcing module of the MIDMod-OSN model described in Section 3.8 to project the reactions of user B's followers. The value of ρ is then calculated as the percentage of positive reactions received. Note that ρ_β is computed using the true attributes of user B, while ρ_γ is estimated using the disclosed (optmized) attributes of user B.

6.3 Results

6.3.1 GPC Model & Calibration

To ensure the inferences drawn from the model align with what is expected, we perform a goodness-of-fit test using the ECE metric to confirm that the model fits the sets of observations as it should. The concern here is that the margin of error observed in the GPC model would impact the optimization task. As mentioned previously that a small ECE values indicate less miscalibration, leading to more confidence in the gradient search results. We use additional calibration techniques (Platt scaling and isotonic scaling[91;92]) in hopes to find a better calibrated model.

Calibration Technique	Expected Calibration Error (ECE)
Model as-is	**0.0161**
Platt scaling	0.0350
Isotonic scaling	0.0218

Table 6.4: *Expected Calibration Error (ECE) reported for GPC model using various calibration techniques.*

Table 6.4 reports the ECE values observed from using the model as it is, as against when er use additional calibration techniques. We see that the model without additional calibration techniques performed best, and this is based on the fact that off-the-shelf GPC models already have some calibration implemented in them. We accept this as favorable and adopt the GPC model as-is to be the baseline for further experimentation.

6.3.2 Gradient Optimization and Manipulation

By giving users the power to change certain attributes about their online personae, we cause a deviation to the accuracy of assumptions drawn about them. Accomplishing this means controlling the users' susceptibility to manipulation as messages targeted at them would not accurately model their interest. These changes can take the form of noise where the user introduce random behavior, or it could be in the form of making adjustments to their profile, for example creating posts on new topics. Before the user introduces noise into their profile,

we see a probability of reaction of 0.488. By changing certain attributes about themselves, the user optimizes their disclosed attributes to confuse manipulators and models that would have otherwise enhanced their profile or messages to target the user.

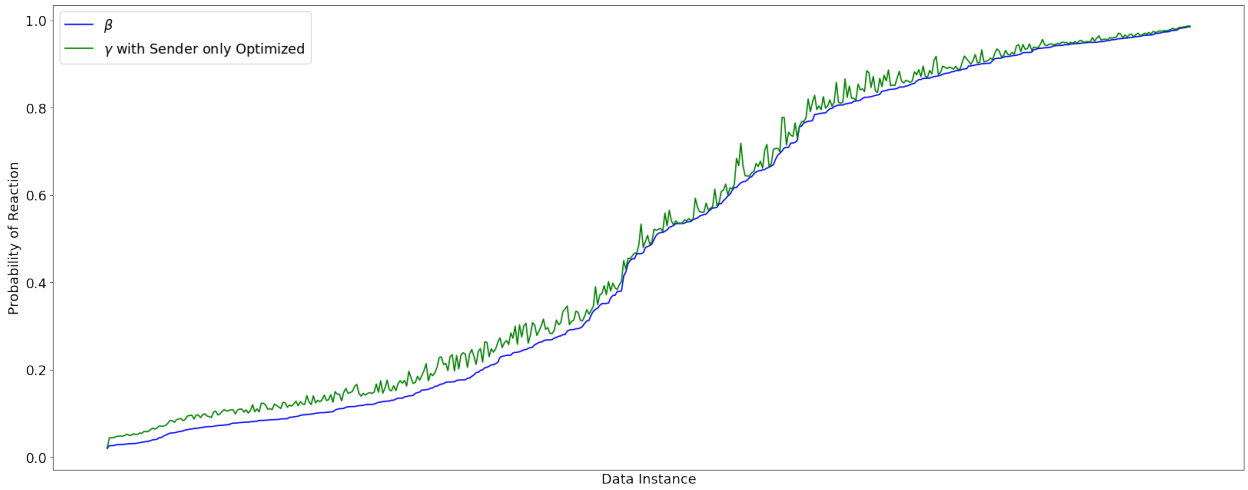

Figure 6.1: *Probability of response when receiver responds to random post and when sender and message attributes are optimally chosen to manipulate receiver.*

In Figure 6.1, each point along the x-axis represents a data instance representing the attributes of user A, user B, and the message M_{AB}. The data instances are first ordered with increasing probability of reaction in a neutral environment and this order is maintained throughout experimentation. We see the effect of targeted manipulation on the receiver by showing how their probability of engaging with an optimized profile and targeted message differs from what their probability of reaction would be if no targeting is being done. Since the goal on the part of the sender (adversary) is to increase the probability of reaction, it is not surprising that the observed change in receivers with probability > 0.8 is much lower than others. By optimizing their attributes and that of the message, the sender is able to increase the average probability of reaction to 0.510 and standard deviation of 0.343.

With the receiver defending themselves from possible attacks by optimizing their attributes, Figure 6.2, we observe a reduction in the probability of reaction with an average of 0.505. Even though we still observe a deviation from the original probability values (blue line in the graph), we recall that the receiver gets to set a deviation value as their acceptable threshold. The difference in these probability values, as described in Eq (6.8), can be

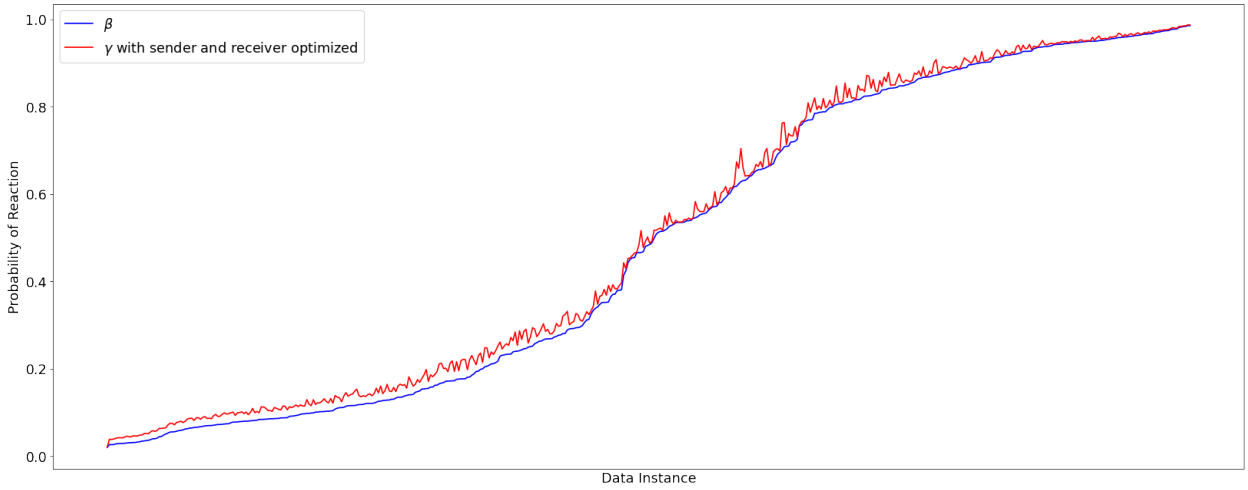

Figure 6.2: *Probability of response when receiver responds to random post and when receiver optimize their attributes while being targeted.*

seen in Figure 6.3. From this result, the receiver is able to observe the effect of the privacy preservation mechanism and compare the manipulation gain value with their set threshold, ϵ. This observation can serve as a guide for the receiver when making decisions on how much protection they intend to put on their profiles. The average manipulation gain of 0.022 from a single tweet can be reduced to 0.016 when the receiver decided to protect themselves using the protection mechanism. Even though we do not see drastic changes in the probability values given that targeting is done over a single tweet, we are able to show that it is possible for manipulation to occur and to what extent. However, this results would become more interesting when there are several messages targeting the same user.

6.3.3 Attribute Disclosure

Once the gradient search converges and we find the optimal points, note that these optimal points are not necessarily global, the receiver can then view the needed changes to be made to their account. For use in the learning model, the data directly observed from Twitter is first pre-processed by normalization and at each step of the gradient optimization, we move back to the original semantic domain and account for errors associated with quantization by ensuring to check for boundary conditions in the model, and approximating to

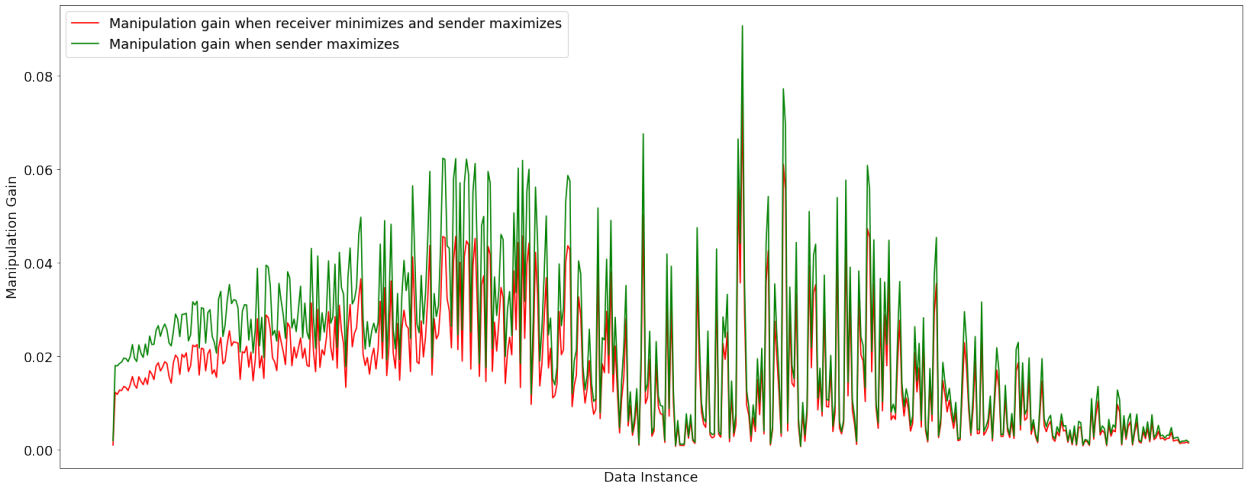

Figure 6.3: *Estimated manipulation gain when receiver does nothing but sender and message attributes are optimized compared with manipulation gain when receiver optimizes their attributes while being targeted.*

the nearest integer values (for integer variables) in the original semantic domain. Being a greedy approach it is expected that these optimum values might just be local and having a metaheuristic function like simulated annealing can aid in ensuring we always find the global optimum.

Foremost, we looked to compare the differences between receivers with lower and higher manipulation gain. This difference is performed by comparing receivers of the same sender as this gives better understanding into their behavior. After looking at the different set of followers, one key finding is that receivers with higher ratio of retweets to original posts tend to experience higher manipulation gain. Additionally, from the social homogeneity score, we observed that users (both sender and receiver) sharing more common friends tend to give off more about their interests, thereby leading to higher risk of effective targeting and ultimately, manipulation gain. Finally, senders creating controversial and alarming posts will generate more interaction from their followers, and through this they can learn the stance or opinion of their followers.

In Table 6.5, we report the observed change in the users' (both sender and receiver) disclosed attributes post-optimization. We see that the changes involved in the sender's attributes require more interaction with the network: increased posts, number of follower

Features	Sender	Receiver
tweets containing URL	No change	No change
presence of user desc	No change	Turned off
user verified	Needed	No change
number of followers	Increase 20%	Increase 50%
number of friends	Increase 5%	Reduce 15%
account age	Increase 10%	Increase 10%
status count	Increase 5%	Reduce 5%
user favorites count	Increase 30%	Reduce 20%
social homogeneity	Increase 8%	Reduce 8%

Table 6.5: *Observed changes in disclosed attributes post-optimization.*

& friends, with an older account. This would be supported by the need to be more visible to generate trust and influence in the network for manipulation purposes. The receiver on the other hand needs to reduce their exposure through the friend's count but maintain an increase in their influence even though there needs to be a reduction in how often they engage with the network leading to a steady decrease in the volume of posts. In reality these changes are gradual as some of them are reliant on others. A change like user verified, for someone who is not a public figure, is dependent on their influence which can be through volume of posts, engagement, and/or endorsement from their network. Additionally, an increase in account age translates to having an older account relative to other accounts in the user's network. It is unsurprising that the sender needs to interact more while the receiver needs to reduce engagement but one thing to note here is that this optimization model gives insight into the degree of change that will make an impact. The receiver looking to make minimal changes for privacy preservation will need not cut off interaction completely, but rather make the needed changes based on their need and loss threshold.

6.3.4 Optimization over Multiple Connections

Realistically, if the adversary looks to manipulate many users at the same time, it is expected that the attributes they disclose and the messages posted are intended to generate some reaction from the majority of their followers. However, the downside to this is that the amended version of their attributes might translate to a negative change in some followers'

reactions.

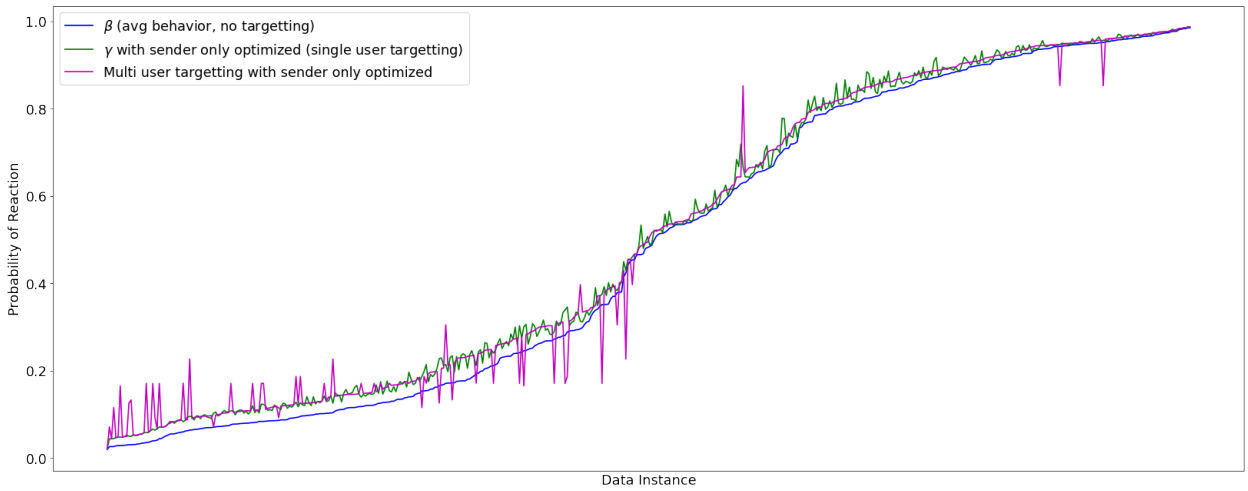

Figure 6.4: *Probability of response when sender optimize attributes over multiple followers compared with when receiver responds to random post and when sender and message attributes are optimally chosen to manipulate receiver.*

As observed in Figure 6.4, for some receivers, the sender's attribute enhancement leads to increase in the probability of reaction notably higher than when the sender optimizes for just that one user. Nonetheless, we see that for 75% of the receivers, the observed probability of reaction values in the case of multi connection targeting are below the values when there is direct targeting on a single receiver. It is surprising though, that about 10% of points are seen to be below the original probability observed when there are no optimizations being carried out in the network. A closer look into these points showed that the receiver profiles are disimilar to others and achieving convergence for them is not trivial.

Similarly, the results in Figure 6.5 where the receiver makes adjustments by observing the disclosed attributes of other followers of the sender are consistent with the change observed with the sender only optimizing over multiple connections. The only difference is an average reduction of 5% in the probability score. The challenge here is that the optimizing receiver, user B_i, is unaware if the disclosed attributes of neighboring followers, Eq (6.14) are the true attributes. In Figure 6.6, we see a higher peak in manipulation gain value than for single user targeting but unlike single user targeting, the recorded gain values are more closely distributed.

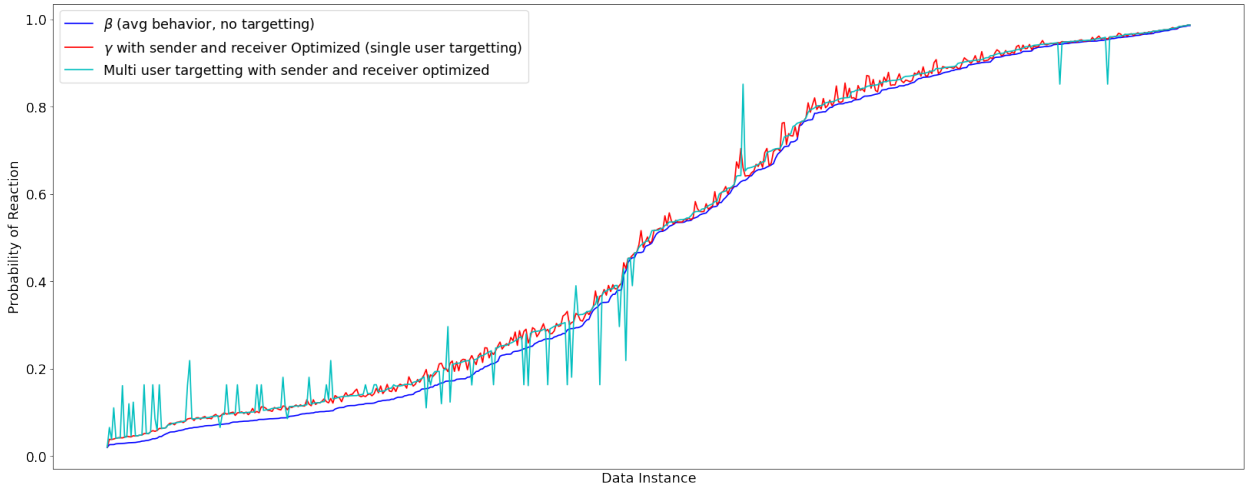

Figure 6.5: *Probability of response when receiver optimize attributes by considering multiple connections compared with when receiver responds to random post and when receiver optimize their attributes while being targeted.*

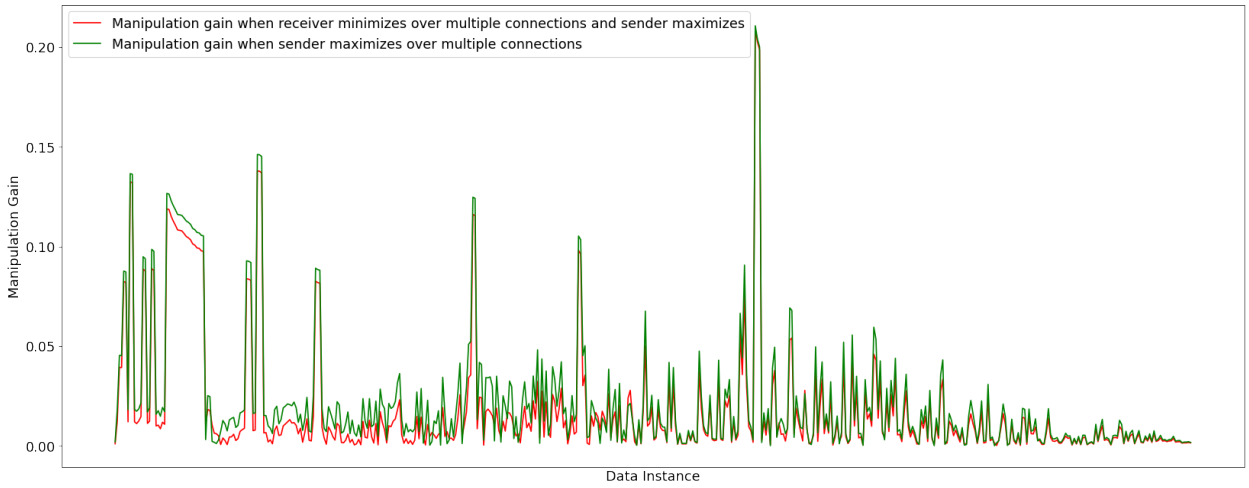

Figure 6.6: *Estimated manipulation gain when optimization is done over multiple connections.*

Post optimization, the changes observed in attributes are reported in Table 6.6. While there is no difference in some attribute adjustments, we see a more subtle change to the receiver's profile as against when there is single user targeting.

Features	Sender	Receiver
tweets containing URL	No change	No change
presence of user desc	No change	Turned off
user verified	Needed	No change
number of followers	Increase 21%	Increase 7%
number of friends	Increase 2%	Reduce 5%
account age	No change	Reduce 4%
status count	Increase 5%	Reduce 10%
user favorites count	Increase 4%	Increase 14%
social homogeneity	No change	No change

Table 6.6: *Observed changes in disclosed attributes with multi-connection targeting post-optimization.*

6.3.5 Constraints over Social Influence

While trying to minimize susceptibility to privacy through inferential attacks, the user (receiver) risks losing some of their social influence depending on how restrictive the implemented privacy preservation mechanism is. The privacy preservation technique suggested in this chapter is as restrictive as the user intends based on a preset ϵ value. Deciding what this value is could prove difficult since it is unsure what that value means. As a way to determine an ϵ relatable to the user, having a way to measure the worth of different ϵ values is important.

Estimating social influence ρ using the crowdsourcing model saw an F-1 score of 78%. For a single user targeting scenario where the sender optimizes for just the receiver and the receiver also makes adjustments to their profile to minimize their manipulation gain, we observed an average reduction of 16% in social influence when we use the receiver's disclosed attributes in the predictive model. In multi-connection optimization scenario, we observed an average reduction of 24%. It is surprising that the average reduction in the multi-connection situation is larger than in a single user targeting situation but this can be due to the big

difference in attributes changes that need to be made, especially in the number of followers.

Chapter 7

Conclusion and Recommendations

7.1 Summary and Conclusions

This research studied the privacy and security implications of participating in online social networks premised on social interaction and the diffusion of information between users. The research explored three main security problems; namely, misinformation, abuse and manipulation. For each of the security problems, the node-to-node interaction between users was explored to show how the microscopic level of relationship contributes to the identified security problem.

Information Diffusion. The MIDMod-OSN model was proposed to gain insight into how different Twitter events, classified as *Trending* or *Informative*, spread from node to node. MIDMod-OSN model was trained using 55 features extracted directly from the Twitter REST API and outperformed the prediction power of state-of-the-art models. It was established that a prediction model based on the top-15 most important features, selected by our feature selection framework, is optimal in correctly predicting diffusion, achieving an AUC score of 96% in both event types. The theoretical contribution of MIDMod-OSN is distinguishing between Informative and Trending Twitter events, and teasing out differences in information diffusion patterns. Even though they are generally overlooked, informative posts make up a big chunk of messages shared on social networks. It was shown that there are

differences between the patterns of interaction between users when exchanging these kinds of posts and trending posts. Additionally, a divergence in features influencing the reaction to post was established, with 40% of the top ranked features belonging to the followers in informative events and 20% in trending events. From these results, it can be inferred that an influence maximization model needs to combine centrality concepts for control, efficiency and activity.

Misinformation. The aim was to show that there is a difference in the spreading behavior of rumor and truthful information in OSNs. A model based on Bayesian logistic regression was presented to predict the credibility status of a message by simply crowdsourcing the interaction and propagation behaviors of similar messages. The crowdsourcing detection model integrates information diffusion by using the diffusion label ("diffused" or "not diffused") associated with the node-to-node interaction between a pair of users. This diffusion label is then combined with the user and message attributes to predict the credibility status for that edge. The credibility status of a particular message is aggregated over all the edges in the network. The result from experiment showed that rumor is mostly masked as news content, meant to incite fear emotions in the reader with mixed sentiments, and that the diffusion attribute is significant to predicting the credibility of a tweet. To identify the credibility of a post especially in the conversation emergent stage where there are not enough posts on the topic or a veracity source, users who interact with specific types of posts serve as good discriminators of credibility. A system looking to efficiently identify the truth status associated with a message will benefit from a comprehensive model exploiting the attributes of the network, interaction and message rather than focusing on just the content of the post.

Abuse. The research in identifying abuse in OSNs established that the abuse levels - abusive, hate, spam, normal, associated with a post can be predicted by simply crowdsourcing the interaction and propagation behaviors of similar messages. The crowdsourcing detection model integrates information diffusion by using the diffusion label ("diffused" or "not diffused") associated with the node-to-node interaction between a pair of users. Results from this experiment show an improvement of about 20% over models that are non-crowdsourced.

Manipulation. The study on manipulation presents a model that limits a user's sus-

ceptibility to targeted manipulation through inferential attack. The model is designed such that it utilizes the user's probability of engaging with a post as a way to measure their susceptibility to manipulation gain and provides the user with the ability to make small changes to themselves in order to confuse a manipulator about who they are or what their interests are. The proposed model showed that even though there will be costs to participating in OSNs, as little bits about the user might still be exploited, these costs can be minimized depending on threshold set by the user as their maximum manipulation gain. Additionally, constraining the manipulation gain on social influence gives insight into the change in the user's percentage of response.

7.2 Limitations and Future Work

From this work, several observations and questions have arisen that are suitable for future research.

Information Diffusion. Future works may include more complex prediction tasks, involving the use of latent user and message attributes for predicting user reactions to posts based on the user's perceived veracity of the post in OSNs.

Misinformation. An interesting area to explore will be to adapt the model on mixed content topics. In mixed content topics where the topics are partially true and false, observing the mixed veracity paradigm on the model will prove valuable. An additional goal to consider will be to estimate the degree of truthful of a tweet.

Abuse. One way to extend this work will be to assign an abusive score for a user based on the emotions their posts incite in the network and the responses the messages get i.e., a user will be deemed more abusive if they incites abusive and/or hateful responses within the network.

Manipulation. One aspect not considered in this work is how the influence of a user adopting the privacy preservation metric is impacted when they minimize their manipulation gain over multiple friends, that is, when the receiver tries to minimize over multiple user A. Also, it will be interesting to see whether setting a social influence reduction threshold can

play a role in privacy preservation.

Bibliography

[1] Abiola Osho, Colin Goodman, and George Amariucai. MIDMod-OSN: A microscopic-level information diffusion model for online social networks. In *Computational Data and Social Networks*, pages 437–450, Cham, 2020. Springer International Publishing. ISBN 978-3-030-66046-8.

[2] Abiola Osho, Caden Waters, and George Amariucai. An implicit crowdsourcing approach to rumor identification in online social networks. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 174–182, 2020. doi: 10.1109/ASONAM49781.2020.9381339.

[3] Abiola Osho, Ethan Tucker, and George Amariucai. Implicit crowdsourcing for identifying abusive behavior in online social networks. *arXiv preprint arXiv:2006.11456*, 2020.

[4] Wikipedia. Social networking service, 2022. URL https://en.wikipedia.org/wiki/Social_networking_service. Last accessed March 2022.

[5] Statista. Social media - Statistics. https://www.statista.com/topics/1164/social-networks/#dossierKeyfigures, 2022.

[6] Alessandro Acquisti and Christina Fong. An experiment in hiring discrimination via online social networks. *Management Science*, 66(3):1005–1024, 2020.

[7] Anupama Aggarwal. Detecting and mitigating the effect of manipulated reputation on online social networks. In *Proceedings of the 25th international conference companion on World Wide Web*, pages 293–297, 2016.

[8] Keith Wilcox and Andrew T Stephen. Are close friends the enemy? online social

networks, self-esteem, and self-control. *Journal of Consumer research*, 40(1):90–103, 2013.

[9] Twitter. Twitter api, 2022. URL https://developer.twitter.com/en/docs/twitter-api. Last accessed April 2022.

[10] Joshua Roesslein. tweepy documentation. *Online] http://tweepy. readthedocs. io/en/v3*, 5, 2009.

[11] Twitter. Developer agreement and policy – twitter developers — twitter developer platform, 2022. URL https://developer.twitter.com/en/developer-terms/agreement-and-policy. Last accessed April 2022.

[12] Mani R Subramani and Balaji Rajagopalan. Knowledge-sharing and influence in online social networks via viral marketing. *Communications of the ACM*, 46(12):300–307, 2003.

[13] Wei Chen, Chi Wang, and Yajun Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1029–1038. ACM, 2010.

[14] Pedro Domingos. Mining social networks for viral marketing. *IEEE Intelligent Systems*, 20(1):80–82, 2005.

[15] Adam Acar and Yuya Muraki. Twitter for crisis communication: lessons learned from japan's tsunami disaster. *International Journal of Web Based Communities*, 7(3):392–402, 2011.

[16] Venkata Kishore Neppalli, Murilo Cerqueira Medeiros, Cornelia Caragea, Doina Caragea, Andrea H Tapia, and Shane E Halse. Retweetability analysis and prediction during hurricane sandy. In *ISCRAM*, 2016.

[17] Rongjing Xiang, Jennifer Neville, and Monica Rogati. Modeling relationship strength

in online social networks. In *Proceedings of the 19th international conference on World wide web*, pages 981–990. ACM, 2010.

[18] Kevin Lewis, Marco Gonzalez, and Jason Kaufman. Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences*, 109(1): 68–72, 2012.

[19] Sen Pei, Lev Muchnik, Jr Andrade José S, Zhiming Zheng, and Hernán A. Makse. Searching for superspreaders of information in real-world social media. *Scientific reports*, 4(1):5547, Jul 3, 2014. doi: 10.1038/srep05547. URL https://www.ncbi.nlm.nih.gov/pubmed/24989148.

[20] Yingcai Wu, Shixia Liu, Kai Yan, Mengchen Liu, and Fangzhao Wu. Opinionflow: Visual analysis of opinion diffusion on social media. *IEEE transactions on visualization and computer graphics*, 20(12):1763–1772, 2014.

[21] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[22] Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443, 1978.

[23] Kazumi Saito, Masahiro Kimura, Kouzou Ohara, and Hiroshi Motoda. Generative models of information diffusion with asynchronous timedelay. In *Proceedings of 2nd Asian Conference on Machine Learning*, pages 193–208, 2010.

[24] Kazumi Saito, Ryohei Nakano, and Masahiro Kimura. Prediction of information diffusion probabilities for independent cascade model. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 67–75. Springer, 2008.

[25] Feng Wang, Haiyan Wang, and Kuai Xu. Diffusive logistic model towards predicting information diffusion in online social networks. In *Distributed Computing Systems*

Workshops (ICDCSW), 2012 32nd International Conference on, pages 133–139. IEEE, 2012.

[26] Jaewon Yang and Jure Leskovec. Modeling information diffusion in implicit networks. In *2010 IEEE International Conference on Data Mining*, pages 599–608. IEEE, 2010.

[27] Adrien Guille, Hakim Hacid, and Cécile Favre. Predicting the temporal dynamics of information diffusion in social networks. *arXiv preprint arXiv:1302.5235*, 2013.

[28] Emilio Ferrara, Wen-Qiang Wang, Onur Varol, Alessand Flammini, and Aram Galstyan. Predicting online extremism, content adopters, and interaction reciprocity. In *International conference on social informatics*, pages 22–39. Springer, 2016.

[29] Nemanja Spasojevic, Zhisheng Li, Adithya Rao, and Prantik Bhattacharyya. When-to-post on social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2127–2136. ACM, 2015.

[30] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[31] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on twitter during the 2016 us presidential election. *Science*, 363 (6425):374–378, 2019.

[32] Adrien Friggeri, Lada Adamic, Dean Eckles, and Justin Cheng. Rumor cascades. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[33] Tetsuro Takahashi and Nobuyuki Igata. Rumor detection on twitter. In *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems*, pages 452–457. IEEE, 2012.

[34] Nor Athiyah Abdullah, Dai Nishioka, Yuko Tanaka, and Yuko Murayama. User's action and decision making of retweet messages towards reducing misinformation spread during disaster. *Journal of Information Processing*, 23(1):31–40, 2015.

[35] Justin Sampson, Fred Morstatter, Liang Wu, and Huan Liu. Leveraging the implicit structure within social media for emergent rumor detection. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, page 2377–2382, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340731. doi: 10.1145/2983323.2983697. URL https://doi.org/10.1145/2983323.2983697.

[36] Liang Wu and Huan Liu. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *WSDM 2018 - Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, WSDM 2018 - Proceedings of the 11th ACM International Conference on Web Search and Data Mining, pages 637–645. Association for Computing Machinery, Inc, February 2018. doi: 10.1145/3159652.3159677. 11th ACM International Conference on Web Search and Data Mining, WSDM 2018 ; Conference date: 05-02-2018 Through 09-02-2018.

[37] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011.

[38] Gang Liang, Wenbo He, Chun Xu, Liangyin Chen, and Jinquan Zeng. Rumor identification in microblogging systems based on users' behavior. *IEEE Transactions on Computational Social Systems*, 2(3):99–108, 2015.

[39] Ke Wu, Song Yang, and Kenny Q Zhu. False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st international conference on data engineering*, pages 651–662. IEEE, 2015.

[40] Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, page 13. ACM, 2012.

[41] Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. Prominent

features of rumor propagation in online social media. In *2013 IEEE 13th International Conference on Data Mining*, pages 1103–1108. IEEE, 2013.

[42] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: Can we trust what we rt? In *Proceedings of the first workshop on social media analytics*, pages 71–79. ACM, 2010.

[43] Anastasia Giachanou, Paolo Rosso, and Fabio Crestani. Leveraging emotional signals for credibility detection. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 877–880, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361729. doi: 10.1145/3331184.3331285. URL https://doi-org.er.lib.k-state.edu/10.1145/3331184.3331285.

[44] Chuan Guo, Juan Cao, Xueyao Zhang, Kai Shu, and Huan Liu. Dean: Learning dual emotion for fake news detection on social media. *arXiv preprint arXiv:1903.01728*, 2019.

[45] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Kalina Bontcheva, and Peter Tolmie. Crowdsourcing the annotation of rumourous conversations in social media. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, page 347–353, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450334730. doi: 10.1145/2740908.2743052. URL https://doi.org/10.1145/2740908.2743052.

[46] Richard McCreadie, Craig Macdonald, and Iadh Ounis. Crowdsourced rumour identification during emergencies. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, page 965–970, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450334730. doi: 10.1145/2740908.2742573. URL https://doi.org/10.1145/2740908.2742573.

[47] Paul Resnick, Samuel Carton, Souneil Park, Yuncheng Shen, and Nicole Zeffer. Rumor-

lens: A system for analyzing the impact of rumors and corrections in social media. In *Proc. Computational Journalism Conference*, volume 5, 2014.

[48] Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. Detecting fake news in social networks via crowdsourcing. *arXiv preprint arXiv:1711.09025*, 2017.

[49] Amira Ghenai and Yelena Mejova. Catching zika fever: Application of crowdsourcing and machine learning for tracking health misinformation on twitter. *CoRR*, abs/1707.03778, 2017. URL http://arxiv.org/abs/1707.03778.

[50] Yingjie Wang, Zhipeng Cai, Guisheng Yin, Yang Gao, Xiangrong Tong, and Guanying Wu. An incentive mechanism with privacy protection in mobile crowdsourcing systems. *Computer Networks*, 102:157–171, 2016.

[51] Monika Verma and Sanjeev Sofat. Techniques to detect spammers in twitter-a survey. *International Journal of Computer Applications*, 85(10), 2014.

[52] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760, 2017.

[53] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153, 2016.

[54] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*, pages 13–22, 2017.

[55] Abdullah Almaatouq, Erez Shmueli, Mariam Nouh, Ahmad Alabdulkareem, Vivek K Singh, Mansour Alsaleh, Abdulrahman Alarifi, Anas Alfaris, et al. If it looks like a

spammer and behaves like a spammer, it must be a spammer: analysis and detection of microblogging spam accounts. *International Journal of Information Security*, 15(5): 475–491, 2016.

[56] Pete Burnap and Matthew L Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242, 2015.

[57] Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[58] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*, 2017.

[59] Arpita Ghosh, Satyen Kale, and Preston McAfee. Who moderates the moderators? crowdsourcing abuse detection in user-generated content. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 167–176, 2011.

[60] Balachander Krishnamurthy and Craig E. Wills. Characterizing privacy in online social networks. In *Proceedings of the First Workshop on Online Social Networks*, WOSN '08, page 37–42, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581828. doi: 10.1145/1397735.1397744. URL https://doi.org/10.1145/1397735.1397744.

[61] Jianming He, Wesley W Chu, and Zhenyu Victor Liu. Inferring privacy information from social networks. In *International Conference on Intelligence and Security Informatics*, pages 154–165. Springer, 2006.

[62] Elena Zheleva and Lise Getoor. To join or not to join: the illusion of privacy in so-

cial networks with mixed public and private user profiles. In *Proceedings of the 18th international conference on World wide web*, pages 531–540, 2009.

[63] Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42, 2007.

[64] Yan Li, Yingjiu Li, Qiang Yan, and Robert H Deng. Privacy leakage analysis in online social networks. *Computers & Security*, 49:239–254, 2015.

[65] Nilothpal Talukder, Mourad Ouzzani, Ahmed K. Elmagarmid, Hazem Elmeleegy, and Mohamed Yakout. Privometer: Privacy protection in social networks. In *2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010)*, pages 266–269, 2010. doi: 10.1109/ICDEW.2010.5452715.

[66] Ratan Dey, Cong Tang, Keith Ross, and Nitesh Saxena. Estimating age privacy leakage in online social networks. In *2012 Proceedings IEEE INFOCOM*, pages 2836–2840, 2012. doi: 10.1109/INFCOM.2012.6195711.

[67] Huaxin Li, Haojin Zhu, Suguo Du, Xiaohui Liang, and Xuemin Shen. Privacy leakage of location sharing in mobile social networks: Attacks and defense. *IEEE Transactions on Dependable and Secure Computing*, 15(4):646–660, 2018. doi: 10.1109/TDSC.2016.2604383.

[68] Huaxin Li, Qingrong Chen, Haojin Zhu, Di Ma, Hong Wen, and Xuemin Sherman Shen. Privacy leakage via de-anonymization and aggregation in heterogeneous social networks. *IEEE Transactions on Dependable and Secure Computing*, 17(2):350–362, 2020. doi: 10.1109/TDSC.2017.2754249.

[69] Danesh Irani, Steve Webb, Kang Li, and Calton Pu. Modeling unintended personal-information leakage from multiple online social networks. *IEEE Internet Computing*, 15(3):13–19, 2011. doi: 10.1109/MIC.2011.25.

[70] Dan Yin, Yiran Shen, and Chenyang Liu. Attribute couplet attacks and privacy preservation in social networks. *IEEE Access*, 5:25295–25305, 2017. doi: 10.1109/ACCESS.2017.2769090.

[71] Fatemeh Amiri, Nasser Yazdani, Heshaam Faili, and Alireza Rezvanian. A novel community detection algorithm for privacy preservation in social networks. In *Intelligent Informatics*, pages 443–450. Springer, 2013.

[72] Bin Zhou and Jian Pei. The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks. *Knowledge and information systems*, 28(1):47–77, 2011.

[73] Lian Liu, Jie Wang, Jinze Liu, and Jun Zhang. Privacy preservation in social networks with sensitive edge weights. In *proceedings of the 2009 SIAM International Conference on Data Mining*, pages 954–965. SIAM, 2009.

[74] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *International conference on database theory*, pages 217–235. Springer, 1999.

[75] Adithya Rao, Nemanja Spasojevic, Zhisheng Li, and Trevor Dsouza. Klout score: Measuring influence across multiple social networks. In *Big Data (Big Data), 2015 IEEE International Conference on*, pages 2282–2289. IEEE, 2015.

[76] Linton C Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978.

[77] R. Jhangiani, I.-C. A. Chiang, C. Cuttler, and D. C. Leighton. *Research Methods in Psychology (4th ed.)*. Kwantlen Polytechnic University, Surrey, BC, 2019. URL https://kpu.pressbooks.pub/psychmethods4e/.

[78] David A Broniatowski, Michael J Paul, and Mark Dredze. National and local influenza surveillance through twitter: an analysis of the 2012-2013 influenza epidemic. *PloS one*, 8(12), 2013.

[79] CNN-Business. Man faces fallout for spreading false sandy reports on twitter, 2012. URL https://www.cnn.com/2012/10/31/tech/social-media/sandy-twitter-hoax/index.html. Last accessed 21 January 2020.

[80] Scott Soames. The truth about deflationism. *Philosophical Issues*, 8:1–44, 1997.

[81] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D11-1147.

[82] Mark Andrew Hall. *Correlation-based feature selection for machine learning*. PhD thesis, University of Waikato Hamilton, 1999.

[83] Boris Pittel. On spreading a rumor. *SIAM Journal on Applied Mathematics*, 47(1): 213–223, 1987.

[84] snopes.com. Snopes.com, 1994. URL https://www.snopes.com/. Last accessed November 2019.

[85] Michael Mathioudakis and Nick Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 1155–1158. ACM, 2010.

[86] DocNow. Docnow/hydrator: Turn tweet ids into twitter json from your desktop!, 2016. URL https://github.com/DocNow/hydrator. Last accessed 20 March 2020.

[87] Adam Badawy, Emilio Ferrara, and Kristina Lerman. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 258–265, 2018. doi: 10.1109/ASONAM.2018.8508646.

[88] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.

[89] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[90] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[91] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[92] Jonathan Wenger, Hedvig Kjellström, and Rudolph Triebel. Non-parametric calibration for classification. In *International Conference on Artificial Intelligence and Statistics*, pages 178–190. PMLR, 2020.

[93] Robert Michael Lewis and Virginia Torczon. Pattern search algorithms for bound constrained minimization. *SIAM Journal on optimization*, 9(4):1082–1099, 1999.

[94] Tommaso Giovannelli, Giampaolo Liuzzi, Stefano Lucidi, and Francesco Rinaldi. Derivative-free methods for mixed-integer nonsmooth constrained optimization. *arXiv preprint arXiv:2107.00601*, 2021.

[95] Forbes. The importance of social currency, 2020. URL https://www.forbes.com/sites/theyec/2020/03/03/the-importance-of-social-currency/?sh=1779524e1678. Last accessed March 2022.