Information Consumption on Social Media: Efficiency, Divisiveness, and Trust

A dissertation submitted towards the degree
Doctor of Engineering of
the Faculty of Mathematics and Computer Science of
Saarland University

by

Mahmoudreza Babaei

Saarbrücken 2020

Date of Colloquium:	14th June 2021
Dean of Faculty:	Prof. Dr. Thomas Schuster

Chair of the Committee:	Prof. Dr. Isabel Valera
First Reviewer:	Prof. Dr. Krishna P. Gummadi
Second Reviewer:	Prof. Dr. Meeyoung Cha
Third Reviewer:	Dr. Manuel Gomez Rodriguez
Fourth Reviewer:	Dr. Adish Singla
Academic Assistant:	Dr. Savvas Zannettou

ABSTRACT

Over the last decade, the advent of social media has profoundly changed the way people produce and consume information online. On these platforms, users themselves play a role in selecting the sources from which they consume information, overthrowing traditional journalistic gatekeeping. Moreover, advertisers can target users with news stories using users' personal data. This new model has many advantages: the propagation of news is faster, the number of news sources is large, and the topics covered are diverse. However, in this new model, users are often overloaded with redundant information, and they can get trapped in filter bubbles by consuming divisive and potentially false information. To tackle these concerns, in my thesis, I address the following important questions:

(i) How efficient are users at selecting their information sources? We have defined three intuitive notions of users' efficiency in social media: link, in-flow, and delay efficiency. We use these three measures to assess how good users are at selecting who to follow within the social media system in order to most efficiently acquire information.

(ii) How can we break the filter bubbles that users get trapped in? Users on social media sites such as Twitter often get trapped in filter bubbles by being exposed to radical, highly partisan, or divisive information. To prevent users from getting trapped in filter bubbles, we propose an approach to inject diversity in users' information consumption by identifying non-divisive, yet informative information.

(iii) How can we design an efficient framework for fact-checking? Proliferation of false information is a major problem in social media. To counter it, social media platforms typically rely on expert fact-checkers to detect false news. However, human fact-checkers can realistically only cover a tiny fraction of all stories. So, it is important to automatically prioritizing and selecting a small number of stories for human to fact check. However, the goals for prioritizing stories for fact-checking are unclear. We identify three desired objectives to prioritize news for fact-checking. These objectives are based on the users' perception of truthfulness of stories. Our key finding is that these three objectives are incompatible in practice.

KURZDARSTELLUNG

In den letzten zehn Jahren haben soziale Medien die Art und Weise, wie Menschen online Informationen generieren und konsumieren, grundlegend verändert. Auf Social Media Plattformen wählen Nutzer selbst aus, von welchen Quellen sie Informationen beziehen hebeln damit das traditionelle Modell journalistischen Gatekeepings aus. Zusätzlich können Werbetreibende Nutzerdaten dazu verwenden, um Nachrichtenartikel gezielt an Nutzer zu verbreiten. Dieses neue Modell bietet einige Vorteile: Nachrichten verbreiten sich schneller, die Zahl der Nachrichtenquellen ist größer, und es steht ein breites Spektrum an Themen zur Verfügung. Das hat allerdings zur Folge, dass Benutzer häufig mit überflüssigen Informationen überladen werden und in Filterblasen geraten können, wenn sie zu einseitige oder falsche Informationen konsumieren. Um diesen Problemen Rechnung zu tragen, gehe ich in meiner Dissertation auf die drei folgenden wichtigen Fragestellungen ein:

- **(i) Wie effizient sind Nutzer bei der Auswahl ihrer Informationsquellen?** Dazu definieren wir drei verschiedene, intuitive Arten von Nutzereffizienz in sozialen Medien: Link-, In-Flow- und Delay-Effizienz. Mithilfe dieser drei Metriken untersuchen wir, wie gut Nutzer darin sind auszuwählen, wem sie auf Social Media Plattformen folgen sollen um effizient an Informationen zu gelangen.

- **(ii) Wie können wir verhindern, dass Benutzer in Filterblasen geraten?** Nutzer von Social Media Webseiten werden häufig Teil von Filterblasen, wenn sie radikalen, stark parteiischen oder spalterischen Informationen ausgesetzt sind. Um das zu verhindern, entwerfen wir einen Ansatz mit dem Ziel, den Informationskonsum von Nutzern zu diversifizieren, indem wir Informationen identifizieren, die nicht polarisierend und gleichzeitig informativ sind.

- **(iii) Wie können wir Nachrichten effizient auf faktische Korrektheit hin überprüfen?** Die Verbreitung von Falschinformationen ist eines der großen Probleme sozialer Medien. Um dem entgegenzuwirken, sind Social Media Plattformen in der Regel auf fachkundige Faktenprüfer zur Identifizierung falscher Nachrichten angewiesen. Die manuelle Überprüfung von Fakten kann jedoch realistischerweise nur einen sehr kleinen Teil aller Artikel und Posts abdecken. Daher ist es wichtig, automatisch eine überschaubare Zahl von Artikeln für die manuellen Faktenkontrolle zu priorisieren. Nach welchen Zielen eine solche Priorisierung erfolgen soll, ist jedoch unklar. Aus diesem Grund identifizieren wir drei wünschenswerte Priorisierungskriterien für die Faktenkontrolle. Diese Kriterien beruhen auf der Wahrnehmung des Wahrheitsgehalts von Artikeln durch Nutzer. Unsere Schlüsselbeobachtung ist, dass diese drei Kriterien in der Praxis nicht miteinander vereinbar sind.

PUBLICATIONS

Parts of this thesis have appeared in the following publications.

- "Analyzing biases in perception of truth in news stories and their implicationsfor fact checkin". Mahmoudreza Babaei, Abhijnan Chakraborty, Juhi Kulshrestha, Elissa M. Redmiles, Meeyoung Cha,and Krishna P. Gummadi. In *the ACM Conference on Fairness, Accountability, and Transparency (FAT*)*, Atlanta, Georgia, US, January 2019. Extended version is accepted in *IEEE Transactions On Computational Social Systems, 2021*.

  (Full paper)

- "Promoting High Consensus News Selectively to Reach a Diverse Audience". Mahmoudreza Babaei, Baharan Mirzasoleiman, Jungseock Joo, and Adrian Weller. In *MAISoN Workshop on Responsible Social media mining (Res-AI)*.

  (Full paper)

- "Purple Feed: Identifying High Consensus News Posts on Social Media". Mahmoudreza Babaei, Juhi Kulshrestha, Abhijnan Chakraborty, Fabricio Benevenuto, Krishna P. Gummadi, Adrian Weller. In *Proceedings of AAAI/ACM Conference on Artificial Intelligence, Ethics & Society (AIES)*, New Orleans, USA, February 2018.

  (Full paper)

- "Information Consumption on Social Media: Efficiency, Trust, and Divisiveness" Mahmoudreza Babaei. *EuroCSS Doctoral Colloquium*, Cologne, Germany, December 2018".

  (Doctoral Colloqium Extended Abstract)

- "On the efficiency of the information networks in social media". Mahmoudreza Babaei, Przemyslaw Grabowicz, Isabel Valera, Krishna P. Gummadi, and Manuel Gomez-Rodriguez. In

InProceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM), pages 83-92, ACM, 2016.

(Full paper)

- “On the users’ efficiency in the twitter information network”. Mahmoudreza Babaei, Przemyslaw Grabowicz, Isabel Valera, and Manuel Gomez-Rodriguez. In *Proceedings of International AAAI Conference on Web and Social Media (ICWSM)*, Oxford, UK, May 2015.

  (Short paper)

Additional publications during doctoral studies.

- “Adversarial Graph Embeddings for Fair Influence Maximization over Social Networks”. Moein Khajehnejad, Mahmoudreza Babaei , Jessica Hoffman, Mahdi Jalili, Adrian Weller. In *International Joint Conferences on Artificial Intelligence 2020 (IJCAI).*

  (Full paper)

- “On the Fairness of Time-Critical Influence Maximization in Social Networks”. Junaid Ali, Mahmoudreza Babaei , Abhijnan Chakraborty, Baharan Mirzasoleiman, Krishna P. Gummadi, and Adish Singla

  Under review in Transactions on Knowledge and Data Engineering.

  It is also accepted in *Human-Centric Machine Learning Workshop at NeurIPS’19(HCML)* 2019.

  (Full paper)

- “On Microtargeting Socially Divisive Ads: A Case Study of Russia-Linked Ad Campaigns on Facebook”. Filipe N. Ribeiro*, Koustuv Saha*, Mahmoudreza Babaei, Fabricio Benevenuto, K Gummadi, Elissa M. Redmiles.

  In *the ACM Conference on Fairness, Accountability, and Transparency (FAT*)*, Atlanta, Georgia, US, January 2019 (Full paper)

- “Media Bias Monitor: Quantifying Biases of Social Media News Outlets at Large-Scale”. Filipe N. Ribeiro, Lucas Henrique, Fabricio Benevenuto, Abhijnan Chakraborty, Juhi Kulshrestha, Mahmoudreza Babei, Krishna P. Gummadi. In *Proceedings of AAAI International Conference on Web and Social Media (ICWSM)*, Stanford, USA, June 2018.

(Full paper)

- “The Road to Popularity: The Dilution of Growing Audience on Twitter”. Przemyslaw A. Grabowicz, Mahmoudreza Babaei, Juhi Kulshrestha, and Ingmar Weber. In *Proceedings of International AAAI Conference on Web and Social Media (ICWSM)*, Cologne, Germany, May 2016. (Short paper)

Dedicated to my family.

ACKNOWLEDGEMENTS

I feel extremely blessed to have had a strong support system during my Ph.D. life. I would like to dedicate this section to all the great people who were there for me in my darkest and brightest moments. I hope not to miss acknowledging any person, but if I do, I express my full-hearted gratefulness to whoever helped me out in times of need.

First of all, I would love to thank my advisor, Krishna Gummadi. I am grateful to him for providing me with the opportunity of graduate studies, supporting me throughout my Ph.D. life, and having faith in me. I have learned a lot from him about good research and traits of a good researcher.

I would like to thank Meeyoung Cha, Manuel Gomez Rodriguez, Adrian Weller, Jungseock Joo, and Adish Singla for sharing expertise, sincere and valuable guidance, and encouragement extended to me.

I have had the honor of collaborating with several wonderful researchers during my graduate studies. I would like to thank my co-authors and collaborators, Fabricio Benevenuto, Przemyslaw A. Grabowicz, Elissa M. Redmiles, Isabel Valera, Baharan Mirzasoleiman, and Ingmar Weber. They all encouraged and supported me throughout our collaborations and this Ph.D. thesis could not have been successful without them. I especially want to mention the collaboration experience and friendship with Juhi and Abhijnan. I learned a lot from both of them and it was an absolute pleasure to brainstorm during our long discussion sessions.

I was fortunate to have the opportunity of mentoring some very talented students, Moein Khajenejad and Xiaofeng Lin. Moreover, I am grateful to former and current MPI-SWS students, Saptarshi, Bimal, Riju, Mainack, Aastha, Nina, Till, Junaid, Moein, Heiko, Hai, Michael, Mahmoud, Mitra, Simin, Mehrshad, Johnnatan, Preethi, and Sophie, for their endless support. Furthermore, I would like to thank all my friends, and in particular Hassan, Hossein, Sepehr, Amir, Ghazaleh, Narges, and Mohammad. They all made my Ph.D. journey enjoyable and fascinating. Particularly, I want to especially thank Kiarash Tamaddon and Alireza Hosseini for being the best friends I could have asked for, and without their unwavering support, I could not have come out successful.

I am grateful to all the tireless technical support staff at MPI-SWS, especially Carina and Chris, for their efforts and help. I would also like to thank Rose for her fantastic English communication classes and her help with proofreading the papers and this thesis. Moreover, I am grateful to Justine for her help in proofreading this thesis. I would like to thank Claudia for her help with administrative matters as well as in general making my time in Germany significantly easier. Last but not least, thanks to Mary-Lou for providing all the information regarding submitting the dissertation and scheduling the defense.

Finally, I would like to thank my parents for their limitless support and unceasing encouragement, without whom I could not have succeeded in my graduate studies. I would like to thank my brothers and sisters in law, Alireza, Hamidreza, Kosar, and Sepideh, for supporting me spiritually throughout this long journey of Ph.D. and my life in general. At last, I couldn't be more thankful to God for the presence of my nephew, Mahyad, and niece, Nahal.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Traditionally, information and news were produced by professional journalists belonging to mass media organizations. These media organizations disseminate news to broad audiences using traditional forms of media broadcasts such as newspapers, magazines, radio, and television. However, with the advent of social media and further changes in the media environment and media markets, traditional media organizations have been facing a variety of challenges that threaten their survival [Murschetz and Friedrichsen, 2017, Picard, 2010, Barthelemy et al., 2011, Journalism and Media, 2014]. For instance, over the last 10 years, they have suffered a substantial decrease in daily circulation, and as a result they have lost advertising revenue proportional to their audience [Der, 2013]. Journalism and Media [2018] have observed that U.S. daily newspaper circulation is decreasing every year. The total estimated circulation of U.S. daily newspapers in 1980 was 62,645,000; by 2010, this fell to 48,000,000, and furthermore, by 2018, it dropped to 28,500,000. In a different survey [Journalism and Media, 2020, 2017], it has been shown that in 2017, 17% of U.S. adults (age 30 to 50) got their political news from print newspapers, while as of early 2020, this number has dropped to only 3% [Journalism and Media, 2020, 2017].

People used to use traditional media to recieve their news mostly from newspapers. Nowadays, this is changing, and people are getting news less from newspapers and more from websites/apps and television [Journalism and Media, 2020]. In addition, there has been an ongoing trend in whic people especially young people are getting more of their news from social media. Social media sites like Twitter and Facebook have increasingly become popular digital news and information marketplaces [Kim et al., 2017, Ribeiro et al., 2019]. A new survey by the Pew Research Center conducted between October 2019 and June 2020 finds that almost half (48%) of young U.S. people age 18 to 29, and 40%

of adults age 30 to 55 rely primarily on social media for political news, whereas only 25% of young people and 32% of adults use news websites/apps and television. [Journalism and Media, 2020].

The rapid adoption of social media sites has brought profound changes to the ways information is produced and consumed in our society. This paradigm shift in how content is shared —from mass media to online social media— offers new opportunities but also raises many new challenges and concerns. To better understand these issues, we now focus on discussing the differences between traditional media and social media.

- Traditional media is a one-way conversation, whereas social media is two-way . The typical flow of traditional media looks as follows: the professional publishers pitch the story, the reporter publishes the story and people are exposed to the story. Thus, the content and timing of information are completely out of the consumers' control [Gomez-Rodriguez et al., 2014]. However, with social media sites, users actively produce and consume a wide variety of information and ideas and can control the content and timing of the information they get by selecting their sources efficiently.

- Traditional forms of media generally offer a wider audience pool, while social media allows for more targeted distribution. Social media gives producers the opportunity to carefully target their audience, selecting everything from the demographics and geography of the audience to the time of day the post will go live [Sears and L. FREEDMAN, 1967, Stroud, 2011, Jomini, 2010, Flaxman et al., 2016].

- There are significant differences in ways we can use to assess the biases and accuracy of information produced by traditional mass media and social media. With traditional media, there were only a handful of news producers, making it easy for media watchdog groups to check the accuracy or biases in the presented news. However, the shift of news consumption to social media has allowed news and information to be produced and shared by members of social media themselves. The problem is, there is no accountability for the accuracy of the presented information. The consequence of this has been the proliferation of fake news, conspiracy theories, and hateful content [Shu et al., 2017, Chopra et al., 2017, Zhao et al., 2015].

While there are other changes due to information consumption moving from mass media to social media, such as, how retrieval (search and recommendation) algorithms influence users' information consumption [Bakshy et al., 2015, Flaxman et al., 2016], our main motivation is to understand and counteract the new challenges that arise due to these three main differences described above.

1.1 Challenges

Here we briefly enumerate three essential challenges that we address in this thesis.

Efficiency: Users on social media sites actively produce and consume a wide variety of information and ideas. On these sites, users typically choose their information sources, which in turn determine what specific information they receive, how much information they receive, and how quickly this information is shown to them. Since users can only select their information sources and don't have control over the context they are exposed to, a natural question that arises is how efficient social media users are at selecting their information sources. In this thesis, we propose a computational framework to quantify users' efficiency in selecting information sources in terms of the number of sources the user follows, the amount of (redundant) information she acquires, and the delay with which she receives the information.

Non-divisiveness: Many news stories are divisive, often posted by polarized publishers, eliciting different reactions from users with different political leanings or pre-existing views, e.g., conservatives or liberals. While various news sources publish divisive and non-divisive news that cover a given story, users often limit themselves to the divisive stories that reinforce their prior views [Sears and L. FREEDMAN, 1967, Stroud, 2011, Jomini, 2010, Flaxman et al., 2016, Himelboim et al., 2013, Babaei et al., 2018, Chiang and Knight, 2011]. This selective exposure and consumption of divisive information may lead to a more politically fragmented, less cohesive society [Cha et al., 2010] and the formation of filter bubbles or echo chambers [Bakshy et al., 2015, Bozdag, 2013, Flaxman et al., 2016, Pariser, 2011]. In this thesis, we propose a complementary approach by identifying and highlighting news posts that are likely to evoke similar reactions from the readers, irrespective of their political leanings. We hope that this approach could lead to less segregation in information consumption [Chakraborty et al., 2017].

Trust: Due to the massive amount of news produced every day on social media, there is no accountability for the presented data's accuracy, leading to the proliferation of fake news. Fighting falsehoods online is one of the great challenges today, particularly during a pandemic. Most social media platforms typically rely on expert fact-checkers to fight this proliferation of false news [Friggeri et al., 2014, Kwon et al., 2017, Ma et al., 2016]. However, human fact-checkers can realistically only cover a subset of stories being circulated, raising the question of how social media platforms should prioritize stories for fact-checking. Here, we present a framework for prioritizing stories, with three important objectives: (1) removing false news stories from circulation, (2) correcting the misperception of users, and (3) decreasing disagreement among different users' perceptions. We argue that current mechanisms are insufficient and only partially satisfy these objectives. We further propose new mechanisms to operationalize the objectives by gathering user perceptions using a novel truth perception test.

In summary, we focus on three important questions: i) How efficient are users at selecting their information sources?, (ii) How does the consumed information impact users and society? , and (iii) How do users perceive the truthfulness of information?

In the rest of the chapter, we briefly review the related work and then give a high-level overview of the specific contributions of this thesis.

1.2 State-of-the-art and related work

In this thesis, we study users' efficiencies in their information source selections. Then we focus on how the consumed information impacts users and society. Accordingly, we demonstrate how non-divisive news (high consensus) may lead to a less politically fragmented and more cohesive society. Finally, we design an efficient framework for fact-checking that helps people assess the credibility of information on social media.

We have three parts in this thesis as described above. The first part *i.e.,* efficiency, is something that people have never looked at before. People have analyzed how information is produced and consumed [Cha et al., 2012, Kwak et al., 2010, Wu et al., 2011] as well as the *amount* of information being exchanged between different groups of users, but this is not really related to the question of how efficient users are at selecting their information sources. Therefore, in this section we focus on the last two parts. We briefly outline current knowledge about the biases in the information consumed by social media users. We finish with a brief overview of prior work in the area of fake news detection.

1.2.1 Biases in social media news

News media organizations have a major impact on political issues and, conclusively, on our society [Groseclose and Milyo, 2005, Chiang and Knight, 2011]. Therefore, several contemporary works have focused on understanding how and to what extent news media outlets can impact people and society, such as in the cases of the White Helmets in Syria [Starbird et al., 2018] and the 2016 US presidential campaign [Rizoiu et al., 2018, Ribeiro et al., 2019]. We are interested in the divisiveness of news content as it relates to concepts of bias that prior work has explored in the context of traditional news organizations [Ahlers, 2006, Althaus and Tewksbury, 2000, Dutta-Bergman, 2004, Gentzkow, 2007, Kaye and Johnson, 2003, Newell et al., 2008]. The approach that has been typically taken is to assess a bias of each publisher (*e.g.,* the publishers are categorized as left-leaning or right-leaning) either by analyzing the audience (audience-based method), or the content (content-based method). The audience-based strategy involves analyzing a news outlet's readership, which assumes that a news outlet's content and attitudes drive its audience's biases. For example, Groseclose and Milyo [2005] assigned a political bias score to news media outlets by investigating the relationship between the news media and members of the US Congress regarding their co-citation of think tanks. The content-based strategy is based on the news, content for example important events covered by news media sources [Covert and Wasburn, 2007, Budak et al., 2016]. Budak et al. [2016] combined machine-learning and crowdsourcing techniques to study the selection and framing of political issues by news organizations.

Now by moving to social media, people have tried to extend these approaches to social media. Ribeiro et al. [2015] and think-tanks [Bakshy et al., 2015, Gentzkow and Shapiro, 2010, Zhou et al., 2011, Mitchell et al., 2014] use audience-based method to assess social media outlets. As there are significantly more news publishers and audiences on social media than in the traditional media, the audience-based strategies can only cover a limited number of news publishers, whereas content-based strategies are more suitable to extend to social media [Munson et al., Ribeiro et al., 2015, Bakshy et al., 2015, Zhou et al., 2011, Mitchell et al., 2014]. Recently, the dissemination of news on social networks has been investigated by several studies [Bhattacharya and Ram, 2012] in terms of publishers' biases [Chakraborty et al., 2016a] and the characteristics of spreaders [Hu et al., 2012]. However, what we are really interested in is not really the biases of publishers as a whole, but the biases of any individual article to see to what extent the article is divisive.

1.2.2 Strategies for detection and mitigation of fake news

Algorithms have become ubiquitous in curating and presenting information to users on online platforms. In recent years, a growing amount of effort has been put into detecting false information by analyzing large-scale digitally logged user behavioral and social network data on the web. False news can be divided into two different types, misinformation and disinformation.

The first kind is *misinformation* (*i.e.,* a piece of information that happens to be wrong). Here, researchers have long investigated online *rumors*, a term used to describe claims that have yet to be verified as 'true' [Qazvinian et al., 2011, Kwon et al., 2013a]. Based on theoretical studies on characterizing online rumor behaviors [Oh et al., 2013, Maddock et al., 2015], computer science researchers have developed rumor detection algorithms based on features describing both linguistic characteristics as well as diffusion patterns of rumors [Kwon et al., 2013b, Zhao et al., 2015]. Recently, a study [Kwon et al., 2017] compared classification capabilities across such multiple feature categories and built an algorithm that achieves high accuracy at identifying rumors in the early stages of rumor spreading. Another line of studies proposed deep learning approaches to detect rumors without a labor-intensive feature engineering. Ma et al. [2016] proposed an RNN-based algorithm to learn sequential information on online rumor spreading. From experiments on Twitter and Weibo, their approach outperformed existing feature-based algorithms and further tackled early detection problems. Other newly proposed deep learning models combine temporal activity patterns of spreaders and source characteristics with existing features. In particular, a model called CSI [Ruchansky et al., 2017] demonstrated state-of-the-art performance in detecting rumors on social media platforms.

The second information type is *disinformation* (*i.e.,* a piece of information that is intentionally manipulated or wrong). This refers to "fake news" which is intentionally and verifiably false [Shu et al., 2017]. Detecting news articles that contain false claims is a challenging task because human evaluators have shown only marginal improvements (66%) over random guesses (50%) in a crowd sourced study [Kumar et al., 2016]. Such findings justify the need for an automated fact-checking system. As a preliminary step, recent studies have focused on the fake news problem known as *clickbait articles* or *stance detection*, in which the news headline and the associated body text have a discordant relationship [Chen et al., 2015]. One study [Chakraborty et al., 2016b] developed a SVM model that predicts clickbait articles based on linguistic patterns. Using the same dataset, another group suggested a neural network approach that measures textual similarities between the headline and first paragraph [Rony et al., 2017].

The above studies discuss various methods to identify false news stories on the Web. Once detected, false stories are clearly labeled, either by reputable sources or by a large number of individuals, but no study has examined how false or true news stories may be *differently perceived* by people irrespective of their ground truth. Furthermore, these stories do not consider the case when a true story is perceived falsely by the audience and misses the chance to propagate properly. In this work, we take a step back and ask whether and how the veracity of news stories relates to their perceived quality by the public and measure to what extent false stories get perceived as true and vice versa.

1.3 Thesis research: Analyzing information consumption on social media along the dimentions of efficiency, divisiveness, and trust

Broadly speaking, our contributions in this thesis can be divided into three parts. (i) We propose a computational framework to quantify users' efficiency in selecting information sources. Our framework is based on the assumption that the goal of users is to acquire a set of unique pieces of information. We define three different notions of efficiency —link, in-flow, and delay— corresponding to the number of sources the user follows, the amount of (redundant) information she acquires and the delay with which she receives the information. (ii) We present an approach that identifies non-divisiveness (high consensus) posts that generate similar reactions from different readers regardless of their political leanings. We then try to see how we can break the filter bubble (exposing people to eaually to diverse news) that people get trapped in. (iii) We focus on examining how users perceive the truthfulness of the information they consume from their information sources. Finally based on users' perceptions, we introduce three orthogonal ranking objectives for fact-checking.

1.3.1 Our contributions

In this thesis, we analyze social media users' information consumption along the dimensions of efficiency [Babaei et al., 2015, 2016], divisiveness [Babaei et al., 2018], and trust [Babaei et al., 2019], which we briefly describe next as in the following.

Our research methodology follows a data-driven approach, with the ultimate goal of contributing to more informed system designs. We frequently conduct large-scale measurement studies to collect real data about information that users produce or consume on social media. In particular, to develop a computational method to measure users' efficiency at selecting information sources , we leverage tools and techniques from graph theory, combinatorial optimization, convex optimization, and machine learning. We further leverage data mining schemes to elicit knowledge about users' perception and judgments of various political content.

The insights obtained from the measurement and analysis studies will ultimately help in the design of more effective system. Lastly, we build and publicly deploy our proposed methodologies to provide usable and practical tools for users of existing social computing platforms (twitter-app.mpi-sws.org/purple-feed). In the rest of the document, we briefly describe each of the projects I have done in my PhD thesis.

1.3.2 Measuring the efficiency of users for selecting information sources [Babaei et al., 2015, 2016]

The task of selecting information sources from potentially tens to hundreds of millions of users poses serious challenges and raises important questions. For example, recent studies have observed that due tothe fear of missing out on important information, users tend to follow too many other users [Hodas and Lerman, 2012]. In the process, they receive a lot of redundant information [Babaei et al., 2015], become overloaded, and effectively miss the information they are interested in [Gomez-Rodriguez et al., 2014, Lerman and Hogg, 2014].

These observations raise questions about how efficient users are at choosing their information sources. In our work, [Babaei et al., 2015, 2016], we identified three notions of efficiency for a user: (i) link (i.e., number of sources the user follows), (ii) in-flow (i.e., the amount of (redundant) information she acquires), and (iii) delay (i.e., the delay with which she receives the information). We introduced a computational framework to quantify a user's efficiency, which estimated the optimal set of users from whom the user could have acquired the same pieces of information more efficiently, and compared this optimal set with the user's set of followees.

We observed that Twitter users exhibit sub-optimal efficiency across our three notions of efficiency. Moreover, we showed that this lack of efficiency is a consequence of the triadic closure mechanism by which users typically discover and follow other users on social media platforms. Finally, we developed a heuristic algorithm that enables users to be significantly more efficient at acquiring the same unique pieces of information.

1.3.3 Breaking filter bubbles and echo chambers [Babaei et al., 2018, 2021]

Having characterized how efficient users are at selecting information sources, we next focus on the impact of the consumed information on users and society. Within our society, there are many divisive topics (posted by polarized publishers) for which different subgroups hold opposing ideological positions (e.g., Republicans and Democrats in the U.S.). Here we first investigate how we can identify non-divisive news. Finally, we propose and quantify the benefits of a potential strategy to spread non-divisive news across readers with diverse political leanings.

Identifying non-divisive content from controversial topics [Babaei et al., 2018]

Social media platforms provide a wide variety of news sources covering the entire spectrum of political ideology (*e.g.,* Republicans and Democrats in the U.S.). Yet, many users mostly limit themselves to news stories posted by polarized news sources, which reinforces their pre-existing views. This selective exposure and consumption of divisive information may lead to a more politically fragmented, less cohesive society [Liu and Weber, 2014]. To minimize the possibility of social media users getting trapped in ‘echo chambers’ or ‘filter bubbles’, researchers have proposed more diversity to the news that users are consuming [Park et al., 2009a, Munson et al., 2013b, Keegan, 2017]. Such approaches, which highlight the most belief-challenging news, often results in users rejecting them, thereby defeating their purpose [Bail, 2014, Lord et al., 1979, Nyhan and Reifler, 2010, Taber and Lodge, 2006, Wood and Porter, 2019, Bail et al., 2018]. Here, we claim that while all publishers, including polarized (biased, partisan) publishers, may post many divisive news stories related to a specific divisive topic, they also post many non-divisive news stories related to that topic. By non-divisive news stories, we mean stories that evoke similar reactions from different readers, irrespective of their own political leanings. We propose a complementary approach to inject diversity in users’ news consumption by highlighting and identifying non-divisive news posts on divisive topics posted by polarized (divisive) publishers. Exposing users to non-divisive news may result in decreasing disagreement amongst them about the divisive topics.

Using our proposed consensus inference method, we publicly deployed "Purple Feed" – a system which highlights non-divisive (high consensus) posts from different news outlets on Twitter. With "Purple Feed", the users can view the non-divisive (high consensus) tweets posted by both Republican- and Democrat-leaning media outlets over the past week (deployed at http://twitter-app.mpi-sws.org/purple-feed/).

Promoting high consensus news selectively to reach a diverse audience [Babaei et al., 2021]
As we discussed in the previous section, even though online social media (OSN) helps expose people to diverse information and news, users may narrow their attention to posts that reinforce their pre-existing views, which could lead to a more fragmented society. Aiming to combat this, as duscussed earlier, we divided news on a given story into non-divisive (high consensus) and divisive (low consensus) posts based on how similar the reactions were across users with diverse political views. We compiled a Twitter dataset and make the following three key observations: (1) divisive (low consensus) news is more likely to remain within subgroups of users with similar political leanings, whereas high consensus news tends to spread across subgroups; (2) non-divisive (high consensus) news posted by neutral publishers spreads more equally across subgroups; and (3) users that get the information from other users instead of publishers, get even more biased exposure to news. Given the above observations we propose methods to propagate high consensus stories broadly across society, aiming to break filter bubbles (and thus potentially decreas polarization in society).

1.3.4 Utilizing users' truth perceptions to prioritize the news stories for fact-checking [Babaei et al., 2019]

Finally, in the last part of the thesis, we focus on addressing the important problem of detecting and mitigating the dissemination of false information. Recently, social media sites have been severely criticized for allowing fake news stories to spread unchecked on their platforms [Friggeri et al., 2014, Kwon et al., 2017, Ma et al., 2016]. To counter the proliferation of fake news, social media sites are relying on their users' perceptions of the truthfulness of news stories to select stories to fact check. However, to date, few studies have focused on understanding how users perceive truth in news stories or how biases in their perceptions might affect current strategies to detect and label fake news stories. Moreover, the goal of fact checking is also not clear. In the last part of the thesis, we describe three important goals for fact-checking news stories based on users' perceptions of the truthfulness of stories:

Goal 1: Removing false news stories from circulation: By using users' truth perception values as a proxy, stories that are flagged by more users as false would be selected for fact checking at a higher probability than stories that were not flagged or were flagged less often.

Goal 2: Correcting the misperception of users: Based on the biases in users' perceptions, stories in which users' perceived truth levels differ significantly from ground truth levels (which we define as the Total Perception Bias) would be prioritized for fact-checking. Using our metric of Total Perception Bias and combining it with user demographics, we designed an automated method for identifying news stories with high and low perception bias achieving an accuracy of 82%.

Goal 3: Decreasing the disagreement among users' perceptions of truth: For society to have fruitful debates in the public sphere, it is essential for common ground to exist among the different, possibly disagreeing sections of the society. Thus, it is desirable to decrease the disagreement among users' truth perceptions by prioritizing those stories for fact checking in which people disagree the most about the truth value of the stories. To achieve this goal, we prioritize stories with high values of disputability (i.e., high variance in users' truth perceptions of the story).

Through the analysis of a large set of user perceptions (N=15,000), we quantify biases in such perceptions and explore how partisan leanings can influence stories prioritized for fact-checking. We propose a framework for social media platforms to prioritize stories for fact checking by leveraging users' truth perceptions to achieve the three complementary objectives mentioned above. Using a combination of user perceptions elicited using our truth perception tests, users' demographic features, and supervised machine learning methods, we provide mechanisms for operationalization strategies that utilize users' truth perceptions to achieve the above objectives for prioritizing stories for fact-checking.

1.4 Organization of the thesis

As part of my thesis research I have mainly focused on analyzing social media users' information consumption along the three dimensions of efficiency [Babaei et al., 2015, 2016], divisiveness [Babaei et al., 2018, 2021], and trust [Babaei et al., 2019]. During my Ph.D. I have also worked on other projects such as: (i) analyzing the quantitative and qualitative characterization of the Russia-linked ad campaigns on Facebook [Ribeiro et al., 2019], (ii) studying the evolution of audiences' characteristics on social media over time [Grabowicz et al., 2016], and (iii) addressing the problem of influence maximization while enhancing group fairness conditions [Ali et al., 2019b, Khajehnejad et al., 2020b]. The rest of the thesis is organized as follows:

In Chapter 2, we introduce the three intuitive notions of users' efficiency in online information and social networks: link (the number of sources the user follows), in-flow (the amount of redundant information she acquires), and delay efficiency (the delay with which she receives the information). We then use these three measures to see how efficient users are at selecting their followees on Twitter. Finally, we propose a method to help users be more efficient at selecting their followees in terms of in-flow and delay efficiencies.

In Chapter 3, we first define non-divisive (high consensus) and divisive news. Then we perform a preliminary empirical study of low and high consensus posts on social media platforms to compare them. We propose a method to automatically identify non-divisive news from others (we also deploy the "Purple Feed" system in which we highlight non-divisive posts from different publishers). Finally, we propose and quantify the benefits of a potential strategy to spread non-divisive news to readers with diverse political leanings.

In Chapter 4, we investigate how users perceive the truthfulness of stories. Then based on users' perceptions, we identify three desired objectives to prioritize news for fact-checking. We finally propose a method to prioritize stories for fact-checking in terms of the three objectives that we introduced.

In Chapter 5, we conclude with a short discussion of the main findings of the thesis and their implications followed by a brief description of future work directions.

CHAPTER 2

Users' efficiency

As we described in the previous chapter, the advent of social media has profoundly changed the way people produce and consume information online in which people play both roles of producing and consuming information. Social media sites such as Twitter, Tumblr, or Pinterest have become global platforms for public self-expression and conversation. More formally, we can think of these sites as large information networks where nodes *i.e.,* users both create and consume information [Kwak et al., 2010]. In this context, people play the information curators' role by deciding which information to post, which other people in the network to follow, and which information posted by other nodes to forward [Romero and Kleinberg, 2010b, Christakis and Fowler, 2010].

However, the task of selecting information sources from potentially tens to hundreds of millions of users poses serious challenges and raises important questions that have not yet been addressed. For example, recent studies have observed that out of fear of missing out on important information, users tend to follow too many other users [Hodas and Lerman, 2012]. In the process, they receive a lot of redundant information [Babaei et al., 2015], become overloaded, and effectively miss the information they are interested in [Gomez-Rodriguez et al., 2014, Lerman and Hogg, 2014]. Moreover, it is very hard to ascertain the quality, relevance, and credibility of information produced by social media users [Agichtein et al., 2008, Castillo et al., 2011, Farajtabar et al., 2015]. Also, many users rely on their network neighborhood for discovering new sources of information, as observed by the large number of triadic closures [Simmel, 1950, Granovetter, 1973, Romero and Kleinberg, 2010a] in link creation.

In this scenario, we want to understand: 1) How efficient are the users of a social media site at selecting which other users to follow to acquire information of their interest? and, 2) can we propose methods to enable a user to acquire the same pieces of information from another set of users in the social media site more efficiently?

To answer these intertwined questions, we view the structure of the information networks in social media sites as the outcome of a network formation game [Kearns, 2012], where a node (*i.e.,* user) links to other nodes to solve a specific task (*i.e.,* acquire information relevant to the user). In this thesis, we propose a general computational framework to quantify and optimize the efficiency of links created by users to acquire information.

In the rest of this chapter, we begin by giving a brief overview of the background and related work in Section 2.1. We then introduce the new dataset to investigate users' efficiency in Twitter as a real information network. Next, to conduct our study, we needed to define the efficiency measurements. We propose a general computational framework to quantify and optimize the efficiency of links created by users to acquire information. Later, We also discuss the plausible reasons based on network structure that lead users to be inefficient in real information networks. Finally, we conclude our work.

2.1 Background

Since the network structure has a significant effect on users' efficiencies, we first briefly discuss the related work to network structure and network properties. Then, we explain the different types of networks categorized by [Newman, 2003].

2.1.1 Networks properties:

The advent of online social networks has seen an explosion of interest in networks' structure, in which researchers have made many studies about network structure and try to propose a model that could explain the structure. Thus, many works focus on social network properties that arise in networks as follows:

Power-law degree distributions :

An alternative way of presenting degree distribution is as follow [Newman, 2003]:

$$P_k = \sum_{k'=k}^{\infty} p'_k \tag{2.1}$$

which is the probability that the degree is greater than or equal to k. Many studies show that most of the degree distribution of real-world networks are right-skewed. In other words, they follow power-law in their tails: $p_k \sim k^{-\alpha}$, where, α is a constant exponent. Note that such power-law distributions show up as power laws in the cumulative distributions also, but with exponent $\alpha - 1$ rather than α:

$$P_k = \sum_{k'=k}^{\infty} k'^{-\alpha} \sim k^{-\alpha-1} \tag{2.2}$$

In empirical studies of directed graphs like the Web, researchers have usually been given only the individual distributions of in- and out-degree [Albert et al., 1999, Barabási et al., 2000, Broder et al., 2000]. Networks with power-law degree distributions sometimes are referred to as scale-free networks [Albert and Barabási, 2002, Barabási and Albert, 1999, Dorogovtsev and Mendes, 2002, Strogatz, 2001].

The Small-World effect or small diameter :

There is a popular experiment carried out by Stanely Milgram in the 1960s, in which the goal was to deliver the letter to a target by passing from person to person. The results show that the letter reaches the target very fast, and this is the first direct demonstration of the small-world effect. Consider an undirect network; the mean shortest path between node pairs in a network is as follows:

$$l = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}^{-1} \tag{2.3}$$

Where d_{ij} is the geodesic distance between nodes i and j. Many studies show the mean shortest path in real-world networks is small, and it is called a small-world effect [Albert and Barabási, 2002, Barabási, 2003, Fronczak et al., 2002, Newman, 2001].

Transitivity or Clustering :

If node u is connected to node v and node v to node w, then there is a high probability that node u will also be connected node w. In social networks, transitivity means that your friend's friend is also likely to be your friend. In terms of network topology, transitivity interprets whereby the presence of a heightened number of triangles in the network–sets of three nodes [Newman, 2003].

Transitivity is quantified by clustering coefficient C:

$$C = \frac{\text{3 number of triangles in the network}}{\text{number of connected triples of vertices}} \tag{2.4}$$

The clustering coefficient has been used in some other similar definitions widely in its own right in the sociological literature [Dorogovtsev et al., 2002, John, 2000, Szabó et al., 2003].

Homophily :

Homophily provides us with a first, fundamental illustration of how a network's surrounding contexts can drive the formation of its links [Easley et al., 2010]. We can interpret homophily as the tendency of individuals to associate and bond with similar others. In other words, people with the same properties, such as age, gender, and political leaning, connect to share interesting information with each other [Ferguson, 2017, McPherson et al., 2001, Krivitsky et al., 2009].

There are many other properties, such as navigability and network resilience in which explining them are out of the this thesis' scope.

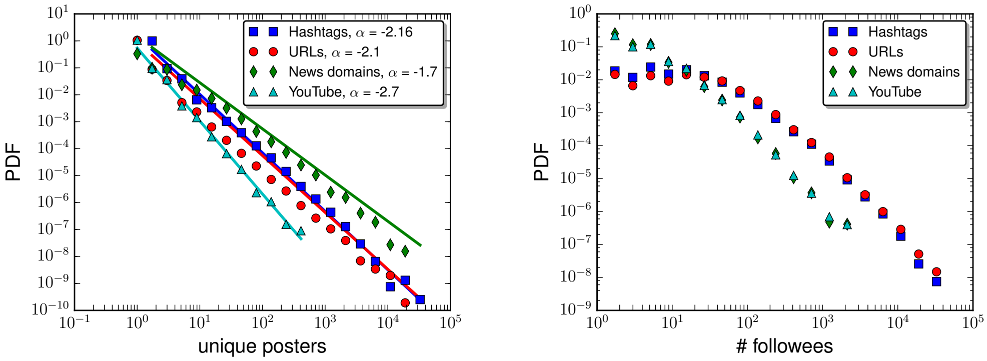
2.1.2 Real world networks:

Here we discuss and explain different types of networks along with their properties. Newman in [Newman, 2003] inspired by the paper by Watts and Strogatz [Watts and Strogatz, 1998] divide real world networks to four categories: social networks, information networks, technological networks, and biological networks.

Social networks: A social network is a set of people or groups of people that use internet-based social media sites to stay connected with friends, family, colleagues, customers, or clients with some pattern of contacts or interactions between them [John, 2000, Wasserman and Faust, 1994]. Social networking created based on friendship and social purpose [Rapoport and Horvath, 1961, Fararo and Sunshine, 1964], business purpose [Jones and Handcock, 2003, Klovdahl et al., 1994, Mirzasoleiman et al., 2012], sexual contacts [Liljeros et al., 2001], etc. have been studied largely in the past [Galaskiewicz, 2016, Galaskiewicz and Marsden, 1978, Mariolis, 1975]. One of the important experiments is the famous "small-world" experiment of Milgram [Milgram, 1967, Travers and Milgram, 1977] that tells us about network structure. The most popular social media sites are such as Facebook, Twitter, LinkedIn, and Instagram. The most important challenge to study of social media sites was their size. Thus, many researchers tried to investigate the social media sample with a smaller size to address this issue. However, those traditional social network studies often suffer from inaccuracy problems and subjectivity because of the small sample size. Accordingly, many researchers focus on relatively reliable smaller data such as collaboration networks [Adamic et al., 2000, Amaral et al., 2000, Newman et al., 2001, Watts and Strogatz, 1998]. Mislove *et al.* in [Mislove et al., 2007], for the first time, studied to examine multiple online social networks such as Flickr, YouTube, LiveJournal, and Orkut at scale to analyze the structure of multiple online social networks. They report many interesting observations. For instance, the networks contain a densely connected core of high-degree nodes; and that this core links small groups of strongly clustered, low-degree nodes at the fringes of the network.

Information networks: There are two classic examples of information networks. 1) citations network in which there is a directed link from article A to article B indicates that A cites B. In this network, papers can only cite other papers that have already been written. The structure and properties of the citation network have been studies by [Redner, 1998, Seglen, 1992]. One interesting observation by [Price, 1965] is, that the number of scientists who have written k papers falls off as $k^{-\alpha}$ for some constant α that means the distribution of the numbers of papers written by individual scientists follows a power law. 2) The World Wide Web network in which web pages containing information, linked to other pages by hyperlinks such as wikipeida[Huberman, 2001]. The Web has been very heavily studied by Albert *et al.* [Albert et al., 1999, Barabási et al., 2000], Kleinberg *et al.* [Kleinberg et al., 1999], and Broder *et al.* [Broder et al., 2000]. Other works discuss other properties, such as having power-law in- and out-degree distributions [Barabási et al., 2002, Albert et al., 1999, Broder et al., 2000, Flake et al., 2002, Kleinberg et al., 1999, Kumar et al., 2000].

Technological networks: This kind of network refers to such networks typically designed to distribute some commodity or resource. For instance, the electric power grid is considered a technological network in which the goal is to transmit electricity to every part of a country. The structure of this network is studied widely by [Amaral et al., 2000, Watts, 2000, Watts and Strogatz, 1998]. The other popular instances of technological networks that studied include the network of airline



(a) Users per meme

(b) The number of followees

Figure 2.1: The distributions of (a) the number of users posting a unique meme and (b) the number of followees posting a specific type of meme at least once. For (a), a power-law is fitted (solid lines) and the exponent α is given.

routes [Amaral et al., 2000], networks of roads [Mirzasoleiman et al., 2011, Babaei et al., 2011, Kalapala et al., 2003], railways [Latora and Marchiori, 2002, Sen et al., 2003], pedestrian traffic [Chowell et al., 2002], marker [Babaei et al., 2013], and river networks.

Biological networks: The last category of real networks is biological networks. The classic example of a biological network is the network of metabolic pathways, which represents metabolic substrates and products with directed edges joining them if a known metabolic reaction exists that acts on a given substrate and produces a given product. Another popular example is the protein interaction network that indicates the mechanistic physical interactions between proteins. Many studies discussed the structure of this kind of network [Ito et al., 2001, Jeong et al., 2001, Maslov and Sneppen, 2002, Solé and Pastor-Satorras, 2002].

Here in this thesis we are using social and information networks to address our concerns.

2.2 Dataset

We use a large Twitter dataset, as reported in previous work [Cha et al., 2010], which comprises the following three types of information: profiles of 52 million users, 1.9 billion directed follow links among these users, and 1.7 billion public tweets posted by the collected users. The link information of the network is based on a snapshot taken at the time of data collection, in September 2009. In our work, we limit ourselves to tweets published during one week, from July 1, 2009 to July 7, 2009, and filter out users that did not tweet before July 1, in order to be able to consider the social graph to be *approximately* static. After this filtering, we have 395,093 active users, 39,382,666 directed edges, and 78,202,668 tweets.

Then, we sample 10,000 users at random out of the 395,093 active users and reconstruct their timelines by collecting all tweets published by the (active) people they follow (among all the 395,093 users), build their ego-networks (i.e., who follows whom among the people they follow), and track all the unique memes they were exposed to during the observation period. We consider four different types of memes:

I. **Hashtags.** Hashtags are words or phrases inside a tweet which are prefixed with the symbol ”#”. They provide a way for a user to generate searchable metadata, keywords or tags, in order to describe her tweet, associate the tweet to a (trending) topic, or express an idea. Hashtags have become ubiquitous and are an integral aspect of the social Web nowadays [Romero et al., 2011].

II. **URLs.** We extract all URLs mentioned inside tweets [Mislove et al., 2007]. Since most of URLs in Twitter are shortened, we unwrap them by calling the API of the corresponding shortening service. Here, we considered seven popular URL shorteners: bit.ly, tinyurl.com, is.gd, twurl.nl, snurl.com, doiop.com and eweri.com, and discard any URL that could not be unwrapped. In general, URLs correspond to online articles, posts, links, or websites.

III. **News domains.** We extract all domain names mentioned inside tweets that correspond to mainstream media sites indexed by Google News [Leskovec et al., 2009]. News domains correspond to media outlets, which may be specializing in the coverage of some topics or perspectives.

IV. **YouTube videos.** We extract all URLs mentioned inside tweets that match the pattern `www.youtube.com/watch`. Here, each of these URLs corresponds to a different YouTube video.

The above memes provide different levels of granularity. For example, news domains are very generic, while YouTube videos are fairly specific. In more detail, the set of active users mention $286,219$ unique hashtags, $379,424$ URLs, 18,616 news domains, and $19,998$ YouTube videos. Figure 2.1a shows the distribution of the number of unique posters for different types of memes, which follows a power-law distribution. The tail of the distribution, as expected, is the heaviest for news domains, while the lightest for YouTube videos. Moreover, as shown in Figure 2.1b, the tail of the distribution of the number of followees tweeting at least one of the memes is also a power-law. In the remainder, we consider only such followees.[1] Also, we focus on users whose information network is fairly developed by filtering out any user following less than 20 followees.

Note that, although our methodology does not depend on the particular choice of meme, it does make two key assumptions. First, it assumes we can distinguish whether two memes are equal or differ. Distinguishing certain memes such as hashtags may be trivial but distinguishing others, such as ideas, may be very difficult. Second, it assumes that receiving several copies of the same meme from different users does not provide additional information, even if different users express different opinions about the meme. It would be interesting to relax the second assumption in future work.

Importantly, in 2009 Twitter did not have features such as “Lists” and “Personalized Suggestions”, so the main way users received and processed information was through their feed, for which we have complete data. The drawback of using older data is smaller number of users and social activity.

2.3 Different types of efficiency

In this work, we propose a computational framework to quantify users’ efficiency at selecting information sources. Our framework is based on the assumption that the goal of users is to acquire a set of unique pieces of information.

Our computational framework is based on the following key concept: given a set of unique ideas, pieces of information, or more generally, *memes* $\mathcal{I}$ spreading through an information network, there is an *optimal* set of nodes that, if followed, would enable us to get to know $\mathcal{I}$. Thus we need to illustrate that, what do we mean by an optimal set in which we consider the relevant notion of optimality for each type of them. In this section, we introduce three different notions of efficiency, namely, link, in-flow and delay efficiency. However, we could leverage this idea to define more complex notions of efficiency. For example, we could define efficiency in terms of diversity, i.e., it would be interesting to find the set of users that, if followed, would cover the same unique memes while maximizing the diversity of topics

[1] Considering all followees leads to qualitatively similar results, but lower absolute values of efficiency.

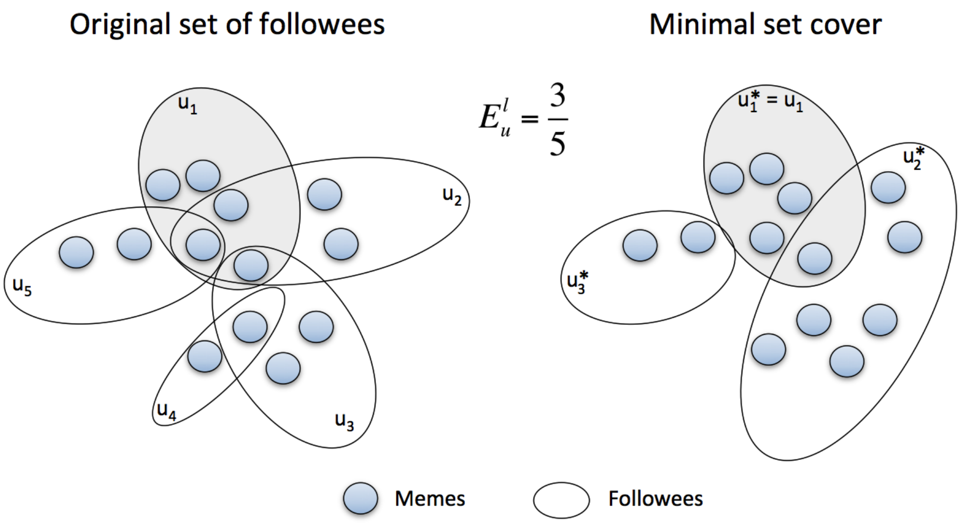


Figure 2.2: Our notion of link efficiency, E_u^l. We define link efficiency as $E_u^l = |\mathcal{U}^l(\mathcal{I}_u)|/|\mathcal{U}_u|$, where $\mathcal{I}_u$ is the set of (unique) memes (blue circles) a user u receives in her timeline by following a set of followees $\mathcal{U}_u = \{u_1, \ldots, u_5\}$ (left), and $\mathcal{U}^l(\mathcal{I}_u) = \{u_1^*, u_2^*, u_3^*\}$ (right) is the minimal set cover (of users) that, if followed, would provide the same set of memes $\mathcal{I}_u$. In the illustration, each user u_i posts the memes within the associated ellipsoid. Hence, in this example, the link efficiency value is $E_u^l = 3/5$.

or per. This would provide a framework to mitigate the effects of the filtering bubble and echo chamber present in current social media systems that we focus more on it in the next section of thesis.

For each type of efficiency, we provide a formal definition and propose a method to approximately compute it with provable guarantees. Then, we use the methods to investigate the efficiency of Twitter users at acquiring information as a real information network.

2.3.1 Link efficiency

As I mentioned above, to compute the efficiency we need to define the optimal set. Here, the optimal set $\mathcal{U}^l(\mathcal{I})$ is the one that contains the smallest number of users which cover the entire information the user receive. Then, we compute link efficiency by comparing the number of people a user follows, i.e., the number of *followees*, with the size of the optimal set $\mathcal{U}^l(\mathcal{I})$.

Finding the optimal set reduces to a minimum set cover problem, which can be solved using a well-known and efficient greedy algorithm with provable guarantees [Johnson, 1973].

Consider a user u and the set of unique memes $\mathcal{I}_u$ she is exposed to through her feed in a given time period, by following $|\mathcal{U}_u|$ users. Then, we define the optimal set $\mathcal{U}^{\mathrm{l}}(\mathcal{I}_u)$ as the minimal set of users that, if followed, would expose the user to at least $\mathcal{I}_u$, and define the link efficiency of a user u at

```
Input: set of all users $\mathcal{U}$; set of unique memes $\mathcal{I}_u$; followee set $\mathcal{U}_u$; set of memes $\mathcal{I}^v$ posted by
       user $v$
Set $\mathcal{U}^{\mathrm{l}} = \emptyset$;
Set $\mathcal{X} = \mathcal{I}_u$;
while $\mathcal{X} \neq \emptyset$ do
    Set $v^* = \arg\min_{v \in \mathcal{U} \backslash \mathcal{U}^{\mathrm{l}}} \frac{1}{|\mathcal{X} \cap \mathcal{I}^v|}$;
    Set $\mathcal{U}^{\mathrm{l}} = \mathcal{U}^{\mathrm{l}} \cup \{v^*\}$;
    Set $\mathcal{X} = \mathcal{X} \backslash \mathcal{I}^{v^*}$;
end
Output: $\mathcal{U}^{\mathrm{l}}$
```

Algorithm 1: Greedy set cover for estimating link efficiency

acquiring memes as

$$E_u^{\mathrm{l}} = \frac{|\mathcal{U}^{\mathrm{l}}(\mathcal{I}_u)|}{|\mathcal{U}_u|}, \tag{2.5}$$

where $0 \leq E_u^{\mathrm{l}} \leq 1$. If the number of users she follows coincides with the number of users in the minimal set, then her efficiency value is $E_u^{\mathrm{l}} = 1$. The larger the original number of followees in comparison with the size of the minimal set, the smaller the link efficiency. Figure 2.2 illustrates our definition of link efficiency.

Examples of link inefficiency: Our definition captures two types of link inefficiency, which we illustrate by two extreme examples. If a user u follows $|\mathcal{U}_u|$ other users, each of them mentioning different (disjoint) sets of memes, and there is another user $v \notin \mathcal{U}_u$ that cover all the memes the followees cover, then the user's efficiency will be $E_u^{\mathrm{l}} = 1/|\mathcal{U}_u|$. If a user u follows $|\mathcal{U}_u|$ other users and all these users mention exactly the same memes, then the user's efficiency will be $E_u^{\mathrm{l}} = 1/|\mathcal{U}_u|$ and $\lim_{|\mathcal{U}_u| \to \infty} E_u^{\mathrm{l}} = 0$. The former type of link inefficiency is due to following users that individually post too few memes, while the latter is due to following users that collectively produce too many redundant memes.

Computing link efficiency: In practice, computing E_u^l, as defined by Eq. 2.5, reduces to finding the minimal set of users $\mathcal{U}^{\mathrm{l}}(\mathcal{I}_u)$, which can be cast as the classical minimum set cover problem [Karp, 1972]. Although the minimum set cover problem is NP-hard, we can approximate $\mathcal{U}^{\mathrm{l}}(\mathcal{I}_u)$ using a well-known and efficient greedy algorithm [Johnson, 1973], which returns an $O(\log d)$ approximation of the minimum size set cover, where $d = \max_{v \in \mathcal{U}} |\mathcal{I}^v|$ is the maximum number of memes posted by any user. Refer to Algorithm 1 for a full description of our procedure to approximate link efficiency.

Link efficiency of Twitter users: We use now definitions of users' efficiency defined by Eqs. 2.5 to investigate how efficient Twitter users are at acquiring four different types of memes: hashtags, URLs, newsdomains and YouTube videos regarding link efficiency. we estimate the empirical probability density function[2] (PDF) for each type of efficiency and meme. We show the results in Figure 2.3, in which we find several interesting patterns. As figure 2.3 depicts people are suboptimal regarding link efficiency.

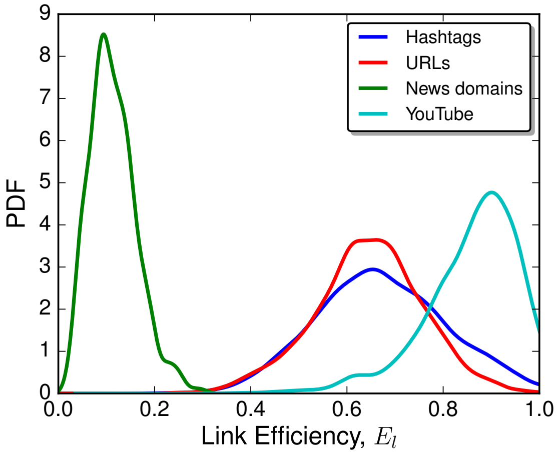


Figure 2.3: The distributions of link efficiencyfor the four types of memes.

2.3.2 In-flow efficiency

The optimal set $\mathcal{U}^f(\mathcal{I})$ is the one that provides the least amount of tweets per time unit. Then, we compute in-flow efficiency by comparing the amount of tweets per time unit a user receives from the people she follows with the amount of tweets per time unit she would have received by following the users in the optimal set $\mathcal{U}^f(\mathcal{I})$. Finding the optimal set reduces to a minimum weighted set cover problem, which again can be solved efficiently with provable guarantees [Johnson, 1973].

Definition: Consider a user u and the set of unique memes $\mathcal{I}_u$ she is exposed to through her feed in a given time period, by following $|\mathcal{U}_u|$ users. Then, we define the optimal set $\mathcal{U}^{\mathrm{f}}(\mathcal{I}_u)$ as the set of users that, if followed, would expose the user to, at least, $\mathcal{I}_u$, while providing the least amount of tweets per time unit, *i.e.,* the minimum tweet in-flow. In particular, we define the in-flow efficiency of a user u

[2] The PDFs have been empirically estimated using kernel density estimation [Bowman and Azzalini, 2004].

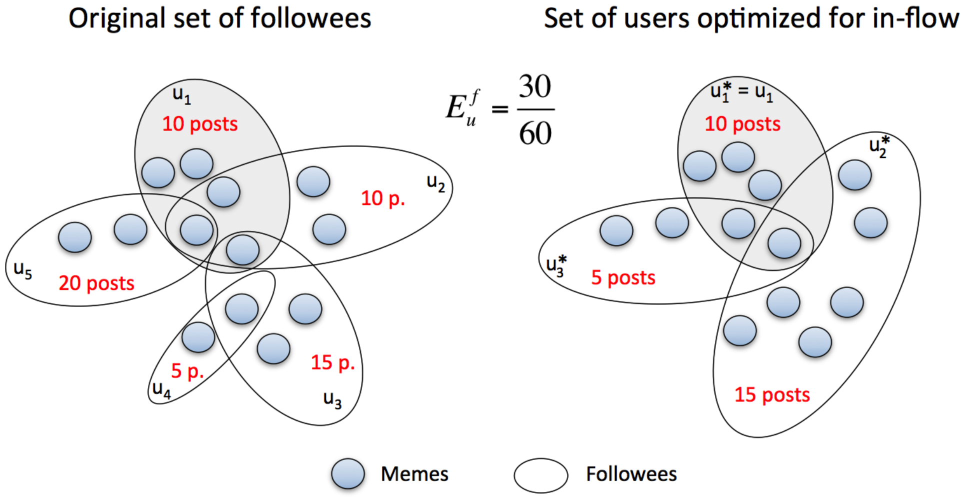


Figure 2.4: Our notion of in-flow efficiency, E_u^f. We define in-flow efficiency as $E_u^f = f(\mathcal{U}^f(\mathcal{I}_u))/f(\mathcal{U}_u)$, where $\mathcal{I}_u$ is the set of (unique) memes (blue circles) a user receives in her timeline by following a set of followees $\mathcal{U}_u = \{u_1, \dots, u_5\}$ (left), and $\mathcal{U}^f(\mathcal{I}_u) = \{u_1^*, u_2^*, u_3^*\}$ (right) is the set cover (of users) with the smallest associated in-flow $f(\mathcal{U}^f(\mathcal{I}_u))$ that, if followed, would provide the same set of memes $\mathcal{I}_u$. In the illustration, each user u_i posts the memes within the associated ellipsoid and the red values in the ellipsoid represent the in-flow of each user. Hence, in this case, the in-flow efficiency value is $E_u^{\text{f}} = 30/60 = 0.5$.

at acquiring memes as

$$E_u^{\text{f}} = \frac{f(\mathcal{U}^{\text{f}}(\mathcal{I}_u))}{f(\mathcal{U}_u)}, \tag{2.6}$$

where $f(\mathcal{U}_u)$ denotes the amount of tweets produced by the set of users $\mathcal{U}_u$ per time unit (user u's in-flow) and $0 \leq E_u^{\text{f}} \leq 1$. The in-flow efficiency $E_u^{\text{f}} = 1$ if user u's in-flow coincides with the amount of tweets per time unit posted by the users in the optimal set $\mathcal{U}^{\text{f}}(\mathcal{I}_u)$. Here, the larger is user u's in-flow in comparison with the amount of tweets per time unit posted by the users in the optimal set, the lower is her in-flow efficiency.

Figure 2.4 illustrates our definition of in-flow efficiency using an example.

Examples of in-flow inefficiency: As in the case of link inefficiency, this definition captures several types of in-flow inefficiency. First, it is easy to see that the example of extreme link inefficiency due to following users posting exactly the same memes, also leads to in-flow inefficiency. if we assume that each user produces the same amount of tweets per time unit. First, it is easy to see that the two examples of extreme link inefficiency, due to following users that cover too few memes or due to following redundant users, also lead to in-flow inefficiency, if we assume that each user produces the same amount of tweets per time unit. Second, there is another type of in-flow inefficiency, which we

```
Input: set of all users $\mathcal{U}$; set of unique memes $\mathcal{I}_u$; number of tweets $N^v$ posted by user $v$; set
       of memes $\mathcal{I}^v$ posted by user $v$
Set $\mathcal{U}^{\text{f}} = \emptyset$;
Set $\mathcal{X} = \mathcal{I}_u$;
while $\mathcal{X} \neq \emptyset$ do
    Set $v^* = \arg\min_{v \in \mathcal{U} \setminus \mathcal{U}^{\text{f}}} \frac{N^v}{|\mathcal{I}^v \cap \mathcal{X}|}$;
    Set $\mathcal{U}^{\text{f}} = \mathcal{U}^{\text{f}} \cup \{v^*\}$;
    Set $\mathcal{X} = \mathcal{X} \setminus \mathcal{I}^{v^*}$
end
Output: $\mathcal{U}^{\text{f}}$
```

Algorithm 2: Greedy set cover for estimating in-flow efficiency

illustrate by an additional extreme example. Consider user u that follows $|\mathcal{U}_u|$ other users and the amount of tweets produced by these followees has a divergent mean, *e.g.,* it has a Pareto distribution (a power law) with exponent $\alpha \leq 1$. Then, if there exists another set of $|\mathcal{U}_u|$ users mentioning the same unique memes and the amount of tweets produced by them has a non-divergent mean, *e.g.,* it is a Pareto distribution with exponent $\alpha > 1$, the user's efficiency will converge to zero as $|\mathcal{U}_u|$ increases, *i.e.,* $\lim_{|\mathcal{U}_u| \to \infty} E_u^{\text{l}} = 0$. Asymptotically, an infinite in-flow could be replaced with a finite in-flow that include the same set of unique memes.

Computing in-flow efficiency: In practice, computing the optimal set of users $\mathcal{U}^{\text{f}}(\mathcal{I}_u)$ reduces to solving the weighted set cover problem, which is also NP-hard. Analogously, we can find an approximate solution to $\mathcal{U}^{\text{f}}(\mathcal{I}_u)$ using a greedy algorithm [Johnson, 1973], which returns an $O(\log d)$ approximation to the set cover with minimum in-flow, where $d = \max_{v \in \mathcal{U}} |\mathcal{I}^v|$ is the maximum number of memes posted by any user. Refer to Algorithm 2 for a full description of our procedure to approximate in-flow efficiency with an approximation factor $O(\log d)$.

In-flow efficiency of Twitter users: We use now definitions of users' efficiency defined by Eqs. 2.6 to investigate how efficient Twitter users are at acquiring four different types of memes: hashtags, URLs, newsdomains and YouTube videos regarding inflow efficiency. As figure 2.5 depicts people are suboptimal regarding link efficiency.

2.3.3 Delay efficiency

The optimal set $\mathcal{U}^t(\mathcal{I})$ is the one that provides the memes as early as possible. Then, we compute delay efficiency by comparing the average delay per meme that the user achieves through the people she

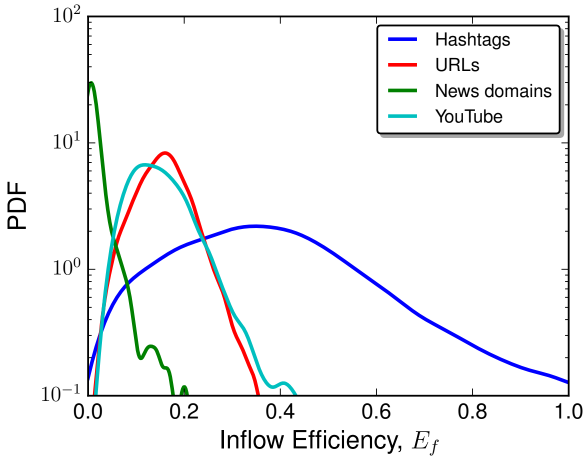


Figure 2.5: The distributions of link efficiencyfor the four types of memes.

follows with the average delay she would achieve by following the users in the optimal set. Here, we define the delay at acquiring a meme in a social media system as the difference between the time when the user received the meme in her timeline and the time when the meme was first mentioned by a user in the social media system. Finding the optimal set reduces to finding the set of users who made the first mention of each of the memes in the social media system.

Definition: Consider a user u and the set of unique memes $\mathcal{I}_u$ she is exposed to through her feed in a given time period, by following $|\mathcal{U}_u|$ users. Then, we define the optimal set $\mathcal{U}^{\mathrm{t}}(\mathcal{I}_u)$ as the set of users that, if followed, would expose the user to, at least, $\mathcal{I}_u$, with the smallest time delay. Here, we define the delay at acquiring a meme provided by a set $\mathcal{U}_u$ as the difference between the time when a user in $\mathcal{U}_u$ first mentions the meme and the time when the meme was first mentioned during the given time period by any user in the whole social media system. We then define the delay efficiency of a user u at acquiring memes as

$$E_u^{\mathrm{t}} = \frac{1}{1 + \langle t_i - t_i^0 \rangle_{i \in \mathcal{I}_u}}, \tag{2.7}$$

where t_i is the time a user in $\mathcal{U}_u$ first mentions meme i, t_i^0 is the time when the meme is first mentioned by a user in the whole social media system, and $\langle t_i - t_i^0 \rangle_{i \in \mathcal{I}_u}$ is an average delay over all memes received by user u, measured in days. **Examples of delay inefficiency:** The delay efficiency $E_u^{\mathbf{t}} = 1$ if the followees of the user u are the first to post the set of memes $\mathcal{I}_u$ in the whole system. The delay efficiency becomes lower than 1 when the user is exposed to the memes at later times than their time of

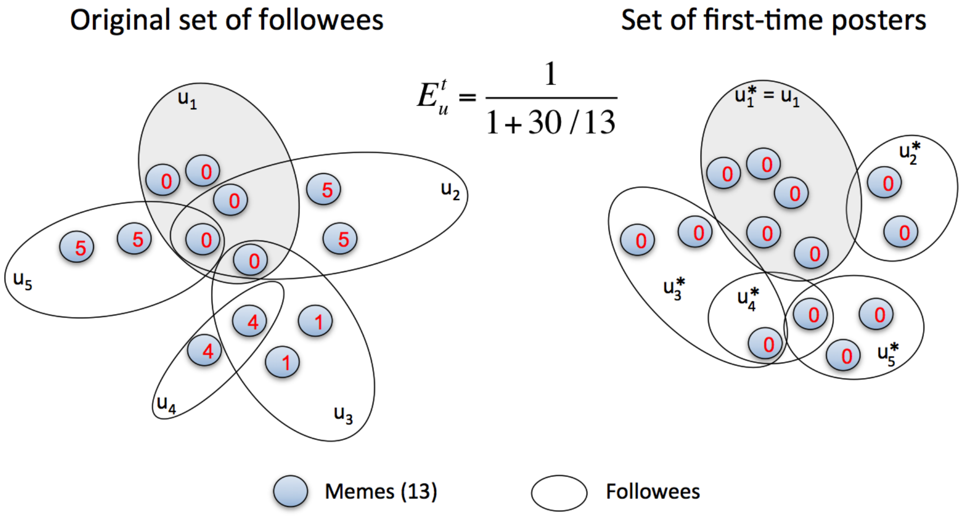


Figure 2.6: Our notion of delay efficiency, E_u^t. We define delay efficiency as $E_u^t = 1/(1 + \langle t_i - t_i^0 \rangle_{i \in \mathcal{I}_u})$, where t_i is the time in which a user receives meme i in her timeline, t_i^0 is the time when the meme is first mentioned by a user in the whole social media system, $\mathcal{I}_u$ is the set of (unique) memes (blue circles) a user receives in her timeline by following a set of followees $\mathcal{U}_u = \{u_1, \ldots, u_5\}$ (left), and $\mathcal{U}^t(\mathcal{I}_u) = \{u_1^*, \ldots, u_5^*\}$ (right) is the set cover (of users) that, if followed, would provide the same set of memes $\mathcal{I}_u$ as early as possible. In the illustration, each user u_i adopts the memes within the associated ellipsoid for the first time after a delay indicated by the red number. Hence, the delay efficiency is $E_u^t = 1/(1 + 30/13)$.

birth. The larger is the average delay of received memes, the smaller is the delay efficiency. Figure 2.6 illustrates our definition of delay efficiency using an example.

Computing delay efficiency: In this case, we can compute the delay efficiency directly by finding when each of the memes appeared for the first time in the system, without resorting to approximation algorithms, as in the case of link and in-flow efficiencies. One can query the first time of appearance for each meme in $O(1)$ by building a mapping between memes and their first time of appearance in a hashtable.

Delay efficiency of Twitter users: We use now definitions of users' efficiency defined by Eqs. 2.7 to investigate how efficient Twitter users are at acquiring four different types of memes: hashtags, URLs, newsdomains and YouTube videos regarding delay efficiency. As figure 2.7 depicts people are suboptimal regarding delay efficiency.

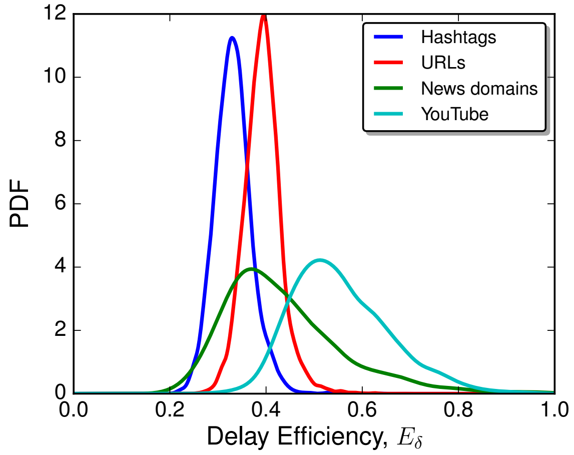


Figure 2.7: The distributions of link efficiencyfor the four types of memes.

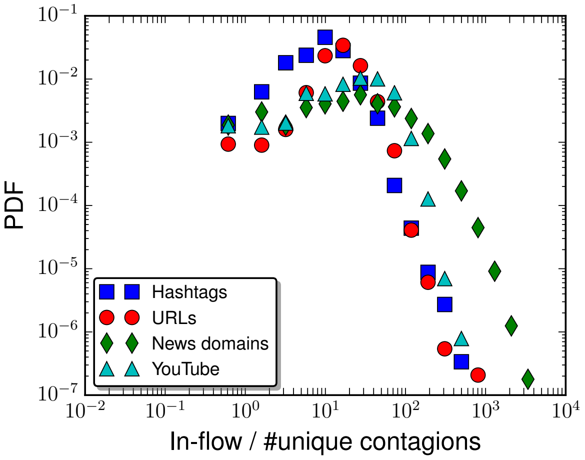
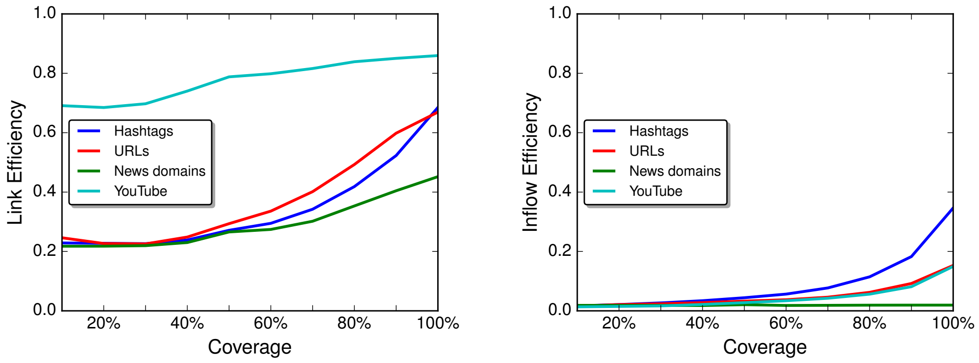


Figure 2.8: The distribution of the ratio between the number of received tweets and unique memes.

2.4 Intersting observations regarding Twitter users' efficiency

Once we have the three definitions of users' efficiency and investigate how efficient Twitter users are at acquiring four different types of memes: hashtags, URLs, news domains, and YouTube videos, we find several interesting patterns. First, all PDFs resemble a normal distribution, however, their peaks (modes) and widths (standard deviations) differ across efficiencies and type of memes. For most users and most types of memes, the efficiency value is significantly below one, giving empirical evidence that users are typically sub-optimal. Second, while the PDFs for link (Figure 2.3) and delay (Figure 2.7)



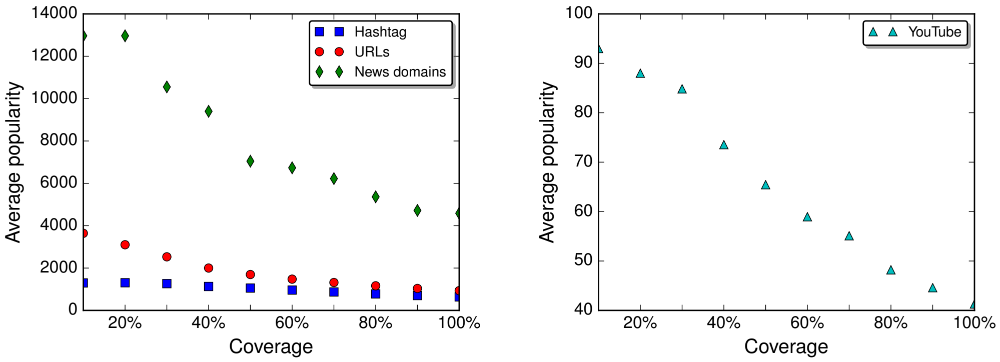
(a) Link efficiency

(b) Inflow efficiency

Figure 2.9: The average link and in-flow efficiencies versus the percentage of covered memes.

efficiencies look quite similar, the in-flow efficiency differs significantly (Figure 2.5). Third, users are most efficient at acquiring YouTube videos, followed by URLs and hashtags, and news domains. This order coincides with the ordering of the exponents of the corresponding power-law distribution of memes' popularity (Figure 2.1a), *i.e.,* the exponent of the power-law (its absolute value) is the highest for YouTube links, followed by URLs and hashtags, and finally for news domains. Note that the higher the exponent is, the higher the proportion of non-popular memes with respect to the popular ones, and thus one can conclude that users are more efficient at acquiring non-popular memes than popular memes. A plausible explanation is that users posting non-popular memes are likely to be included in the optimal set, since there is nobody else who posts these memes and, as a consequence, the optimal set differs less from the original set of followees. Moreover, note that in Figure 2.5, the in-flow efficiency of both news domains and YouTube videos is shifted to the left (*i.e.,* presents much lower efficiency values) compared to the link efficiency in Figure 2.3. This shift is due to the fact that, as shown in Figure 2.8, ratio between the total number of received tweets and unique memes is much larger for unique news domains and YouTube video memes than for hashtags and URLs, which in turn, translates into a lower in-flow efficiency.

In the above measurements, we estimated the probability density functions of user's efficiency considering full coverage of the received memes. Importantly, it is straightforward to extend our definitions of link and in-flow efficiencies to account for partial coverage, by simply considering a set of users that, if followed, would expose the user to, at least, a percentage of the unique memes $\mathcal{I}_u$, by stopping the greedy algorithm whenever the given percentage is reached. Note, however, that



(a) Hashtags, URLs, news

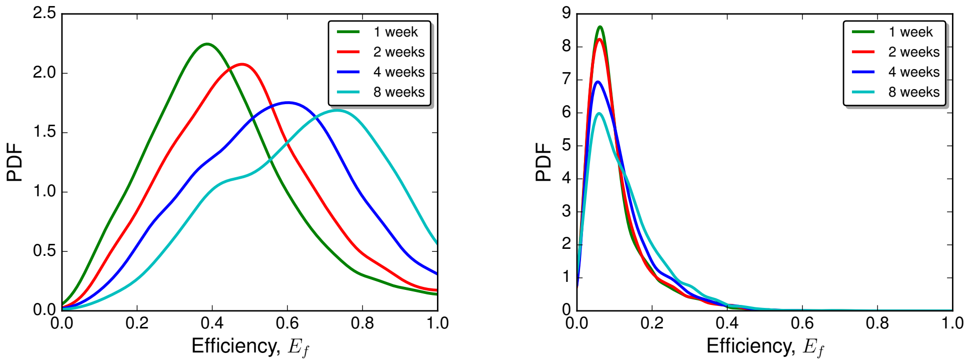
(b) YouTube videos

Figure 2.10: The average popularity of covered memes as a function of the percentage of covered memes.

computing the efficiency for a partial coverage based on the full set of followees would be unfair, since some of the followees in $\mathcal{U}_u$ may be not tweeting any of the covered memes. Thus, for the purpose of computing the efficiencies for partial coverage, we take into account only the users in $\mathcal{U}_u$ who tweet at least one of the covered memes.

Figure 2.9 shows the average link and in-flow efficiencies against coverage for the same four memes. As one may have expected, the higher the coverage, the higher the link and in-flow efficiency, since the memes that are covered first by the greedy algorithm are the popular ones, as shown in Figure 2.10. This result confirms that users are more efficient at acquiring less popular information, but less so at acquiring more popular information. A plausible explanation is that less popular information is produced by only a handful of users and so the optimization is limited to this set of users.

Finally, we investigate if our results are consistent across different time periods. In particular, we measure user efficiency based on time periods of different lengths: one, two, four and eight weeks. In Figure 2.11a, we can observe that as we increase the period, there are more unique memes there to cover, which results in an increase in the efficiency. However, the distribution of efficiency is nearly unchanged for 80% coverage (Figure 2.11b). Thus, the findings presented in this study are qualitatively robust to the choice of time period. Additionally, we find that the choice of the week does not influence the distributions of efficiency, however, these results are not shown due to space limitation.



(a) 100% Coverage

(b) 80% Coverage

Figure 2.11: In-flow efficiency measured for hashtags appearing in the time periods of different lengths. The results for other efficiencies and meme types are qualitatively the same.

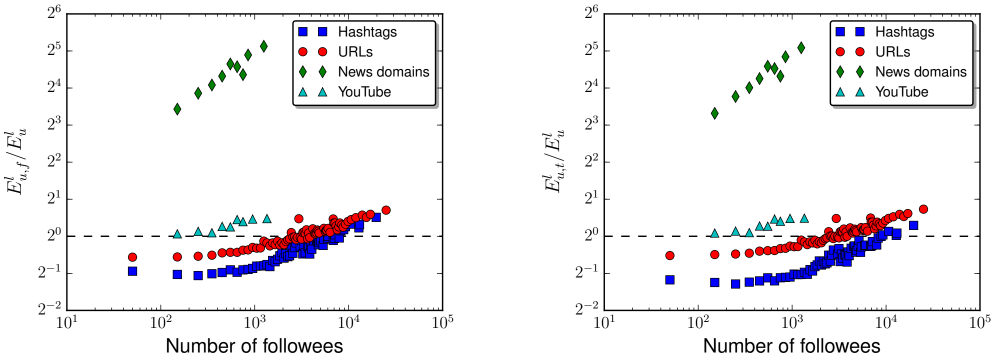
2.5 Cross efficiency

In our definitions of efficiency, the optimal set for a given user is the set of users that minimizes the number of links, in-flow or delay, while covering the same set of unique memes. However, this naturally raises the question as to how efficient the optimal sets for a given definition of efficiency are in terms of the other definitions. For example, how efficient is the link-optimal set with respect to in-flow or delay efficiency? In this section, we first address this question, then introduce the idea of finding sets of users that jointly optimize multiple notions of efficiency, and finally develop a heuristic algorithm that simultaneously improves both in-flow and delay efficiency of users.

2.5.1 Cross-efficiency of optimal sets

Given a user u and the set of unique memes $\mathcal{I}_u$ she is exposed to in a given time period, our definitions of efficiency compare the original set of followees with the optimal sets $\mathcal{U}^{\text{l}}(\mathcal{I}_u)$, $\mathcal{U}^{\text{f}}(\mathcal{I}_u)$ and $\mathcal{U}^{t}(\mathcal{I}_u)$ in terms of number of links, in-flow and average delay, respectively. Here, we assess the efficiency of the optimal sets for each definition of efficiency in terms of the other definitions, which we call *cross-efficiencies*.

More specifically, we compute them as follows:



(a) Effect of in-flow efficiency optimization on link efficiency

(b) Effect of delay efficiency optimization on link efficiency

Figure 2.12: The effect of optimization of in-flow and delay efficiencies on link efficiency, plotted as the ratio of the efficiency in the optimized network and the original network against the number of followees. The dashed line marks the ratio equal to 1, which corresponds to the lack of change in the efficiency due to the respective optimization.

We compute the link efficiency of the optimal sets for in-flow and delay efficiency, *i.e.,*

$$E^{\mathrm{l}}_{u,\mathrm{f}} = \frac{|\mathcal{U}^{\mathrm{l}}(\mathcal{I}_u)|}{|\mathcal{U}^{\mathrm{f}}(\mathcal{I}_u)|} \quad \text{and} \quad E^{\mathrm{l}}_{u,\mathrm{t}} = \frac{|\mathcal{U}^{\mathrm{l}}(\mathcal{I}_u)|}{|\mathcal{U}^{\mathrm{t}}(\mathcal{I}_u)|},$$

Since we would like to know if an efficiency of an optimized information network is increased in comparison with the efficiency of the original network we focus on measuring the ratio of an efficiency of the optimized and original networks. Here for link efficiency we compute the ratio as follows:

$$\frac{E^{\mathrm{l}}_{u,\mathrm{f}}}{E^{l}_{u}} \quad \text{and} \quad \frac{E^{\mathrm{l}}_{u,\mathrm{t}}}{E^{l}_{u}}$$

If the ratio is higher than one for the given optimization algorithm, then the corresponding efficiency is improved by that algorithm with respect to the original set of followees. If the ratio is below one then the respective efficiency is decreased by the optimization algorithm. As Figure 2.12a and b depict, the optimal sets for news domains are more efficient than the original sets. This observation may happen due to the following reason: news domains tend to be more popular than the other types of memes (as shown in Figure 2.1). As a consequence, a user may receive multiple copies of the same news domain from various followees, and it is very easy to find efficient sets in terms of in-flow; it is enough to simply remove some of their followees from the network to improve both link and in-flow efficiencies.[3] The ratio for other memes tends to be below one or close to one, which indicates that optimizing for inflow or delay efficiencies results in decreased link efficiency. We did the same think to compute the in-flow efficiency of the optimal sets for link and delay efficiency, *i.e.,*

$$E^{\mathrm{f}}_{u,\mathrm{l}} = \frac{f(\mathcal{U}^{\mathrm{f}}(\mathcal{I}_u))}{f(\mathcal{U}^{\mathrm{l}}(\mathcal{I}_u))} \quad \text{and} \quad E^{\mathrm{f}}_{u,\mathrm{t}} = \frac{f(\mathcal{U}^{\mathrm{f}}(\mathcal{I}_u))}{f(\mathcal{U}^{\mathrm{t}}(\mathcal{I}_u))},$$

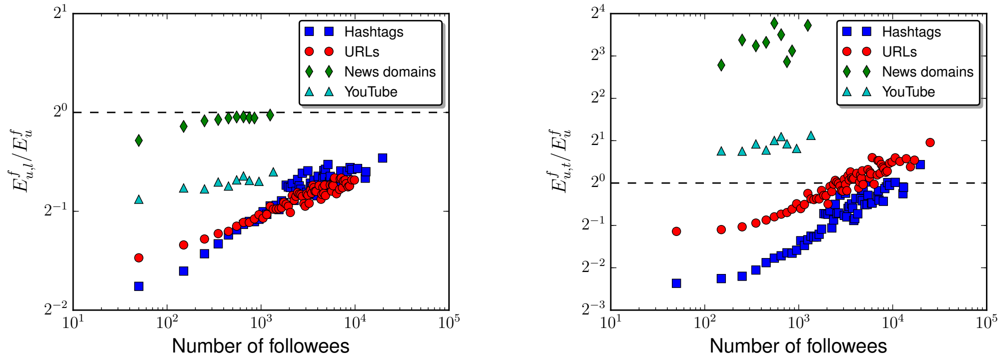
and the delay efficiency of the optimal sets for link and in-flow efficiency, *i.e.,*

$$E^{\mathrm{t}}_{u,\mathrm{l}} = \frac{1}{1 + \langle t^{\mathrm{l}}_i - t^0_i \rangle_{i \in \mathcal{I}_u}} \quad \text{and} \quad E^{\mathrm{t}}_{u,\mathrm{f}} = \frac{1}{1 + \langle t^{\mathrm{f}}_i - t^0_i \rangle_{i \in \mathcal{I}_u}},$$

where t^{l}_i is the time a user in $\mathcal{U}^{\mathrm{l}}(\mathcal{I}_u)$ first mentions meme i and t^{f}_i is the time a user in $\mathcal{U}^{\mathrm{f}}(\mathcal{I}_u)$ first mentions meme i.

To see if an efficiency of an optimized information network is increased in comparison with the efficiency of the original network we compute the ratio as follows:

$$\text{in-flow } \frac{E^{\mathrm{f}}_{u,\mathrm{l}}}{E^{f}_{u}} \quad \text{and} \quad \frac{E^{\mathrm{f}}_{u,\mathrm{t}}}{E^{f}_{u}} \quad \text{; delay } \frac{E^{\mathrm{t}}_{u,\mathrm{l}}}{E^{t}_{u}} \quad \text{and} \quad \frac{E^{\mathrm{t}}_{u,\mathrm{f}}}{E^{t}_{u}}$$

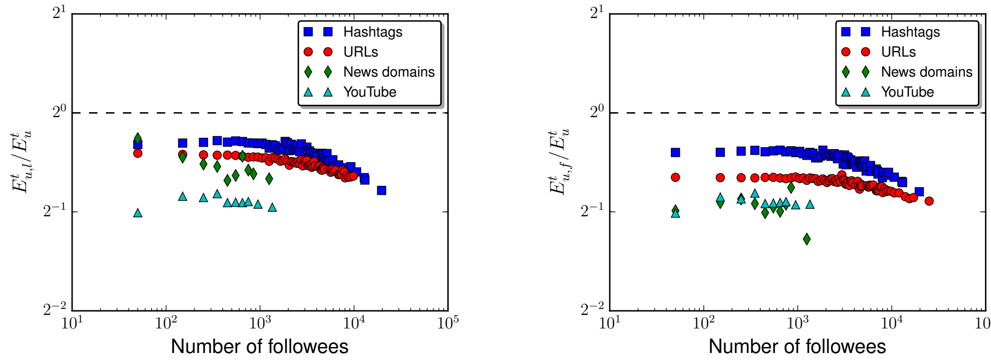


(a) Effect of link efficiency optimization on in-flow eff.

(b) Effect of delay efficiency optimization on in-flow eff.

Figure 2.13: The effect of optimization of link and delay efficiencies on in-flow efficiency

[3] In fact, over 85% of users receive less unique news domains than they have followees.



(a) Effect of link efficiency optimization on delay eff.

(b) Effect of in-flow efficiency optimization on delay eff.

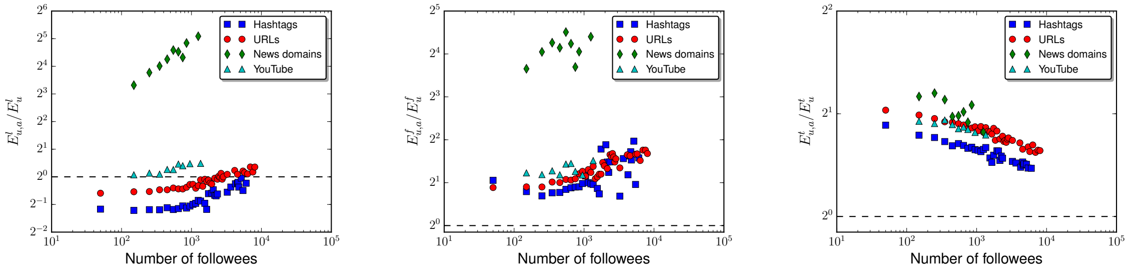
Figure 2.14: The effect of optimization of link and in-flow efficiencies on delay efficiency

As Figures 2.13 and 2.14 show, the ratio tends to be below one or close to one for most meme types, which indicates that optimizing for one definition of efficiency generally results in decreased efficiency with respect to the other two definitions.

However, there is one exception similar to effect on link efficiency, as Figure 2.13b show, the optimal sets for news domains are more efficient than the original sets. We discussed the plausible reason earlier. In Figures 2.13(a) and (b) we note that the improvement in the in-flow efficiency tends to grow with the number of followees due to the increased number of redundant (non-unique) information received by the users who follow many other people. In terms of delay efficiency, the optimal sets for link and in-flow efficiency for hashtags and URLs are more efficient than the original sets (blue and red points in Figures 2.14), however, they are less efficient for news domains and YouTube videos (green and teal points in Figures 2.14). In Figures 2.14, the improvement in the delay efficiency tends to drop with the number of followees because it is likely that users who receive many copies of the same meme receive it early on. Thus, for users with many followees, it is harder to improve the delay efficiency.

2.5.2 Joint-optimization of efficiencies

So far, we have looked for optimal sets of users in terms of a single efficiency (be it link, in-flow or delay). Moreover, in the previous section, we have shown that optimal sets in terms of a single efficiency typically decrease the other efficiencies. Therefore, one could imagine developing an algorithm a looking for sets of users that are optimized with respect to several efficiencies; in other words, a



(a) Effect on link efficiency

(b) Effect on in-flow efficiency

(c) Effect on delay efficiency

Figure 2.15: The effect of optimization of both in-flow and delay efficiencies on different types of efficiencies, plotted as the ratio of the affected efficiency of the optimized network and the original network against the number followees.

multi-objective algorithm. Given such an algorithm a, we could compute the efficiency of the optimal set $\mathcal{U}_a(\mathcal{I}_u)$ with respect to single quantities, *i.e.,*

$$E^{\mathrm{l}}_{u,\mathrm{a}} = \frac{|\mathcal{U}^{\mathrm{l}}(\mathcal{I}_u)|}{|\mathcal{U}^{\mathrm{a}}(\mathcal{I}_u)|}, \; E^{\mathrm{f}}_{u,\mathrm{a}} = \frac{f(\mathcal{U}^{\mathrm{f}}(\mathcal{I}_u))}{f(\mathcal{U}^{\mathrm{a}}(\mathcal{I}_u))} \text{ and}$$

$$E^{\mathrm{t}}_{u,\mathrm{a}} = \frac{1}{1 + \langle t^{\mathrm{a}}_i - t^0_i \rangle_{i \in \mathcal{I}_u}}$$

where t^{a}_i is the time a user in $\mathcal{U}^{\mathrm{a}}(\mathcal{I}_u)$ first mentions meme i. Ideally, we would like to find optimal sets that are efficient with respect to the considered quantities. Here, as a proof of concept, we next develop a heuristic method to find sets of users optimized with respect to both in-flow and delay.

Joint optimization of in-flow and delay efficiency

We leverage the greedy algorithm from the weighted set cover problem to design a heuristic method that finds sets of users with high in-flow and delay efficiencies, while delivering the same unique memes to the user (refer to Algorithm 3). In particular, in the heuristic method, the weights are powers of tweets in-flow N^{α}_v and average delay T^{β}_v over all unique memes produced by the user v. The exponents α and β can be readily adjusted to induce higher or lower in-flow efficiency and delay efficiency, respectively. Here, we experiment with $\alpha = 1$ and $\beta = 0.5$, which achieves a good balance between in-flow and delay efficiency.

We summarize the ratio of link, in-flow and delay efficiency of the set of users provided by our heuristic method and the original set of followees in Figure 2.15. We discus several interesting

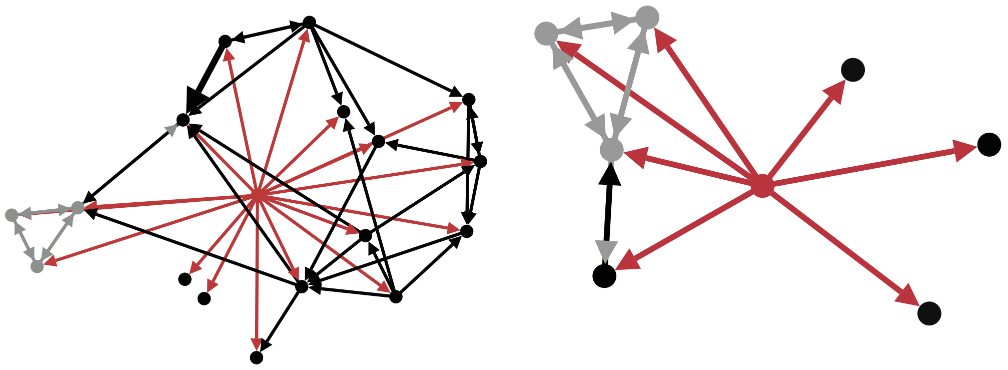
```
Input: set of all users $\mathcal{U}$; set of unique memes $\mathcal{I}_u$; set of memes $\mathcal{I}^v$ posted by user $v$
Set $\mathcal{U}^* = \emptyset$;
Set $\mathcal{X} = \mathcal{I}_u$;
while $\mathcal{X} \neq \emptyset$ do
    Set $v^* = \arg\min_{v \in \mathcal{U} \setminus \mathcal{U}^{\mathrm{f}}} \frac{N_v^\alpha T_v^\beta}{|\mathcal{I}^v \cap \mathcal{X}|}$;
    Set $\mathcal{U}^* = \mathcal{U}^* \cup \{v^*\}$;
    Set $\mathcal{X} = \mathcal{X} \setminus \mathcal{I}^{v^*}$
end
Output: $\mathcal{U}^*$
```

Algorithm 3: Greedy set cover for jointly optimizing in-flow and delay efficiencies

observations. First, since the algorithm does not optimize with respect to the number of links, the link efficiency is not improved by this algorithm, *i.e.,* the ratio between the link efficiency of the set provided by the heuristic method and the link efficiency of the original set of followees ratio is around or below 1 for three out of four meme types. Second, we find that both in-flow and delay efficiencies are significantly increased over the efficiency of the original set of followees for all types of memes. The in-flow efficiency is, on average, 7.4-times higher for news domains, 1.8-times higher for hashtags, 1.3-times higher for URLs, 1.2-times higher for YouTube videos. The delay efficiency is, on average, 1.8-times higher for news domains, 1.4-times higher for hashtags, 1.4-times higher for URLs, 1.2-times higher for YouTube videos. There is always an increase in in-flow and delay efficiencies independently of the number of followees that the users have originally. However, while the improvement in the in-flow efficiency tends to be larger for users with many followees, the improvement in the delay efficiency is larger for users with fewer followees. Thus, we conclude that our algorithm increases both in-flow and delay efficiency of users.

2.6 Sructure of ego networks: original vs. optimized

In the previous sections, we have introduced three meaningful definitions of efficiency and applied them to show that Twitter users tend to choose their information sources inefficiently. In this section, we investigate the rationale behind this sub-optimal behavior by comparing the structure of the user's ego-networks associated with both the original set of followees and the sets optimized for efficiency. Here, we define a user's ego-network as the network of connections (who-follows-whom) between the ego user and her followees.



(a) Original

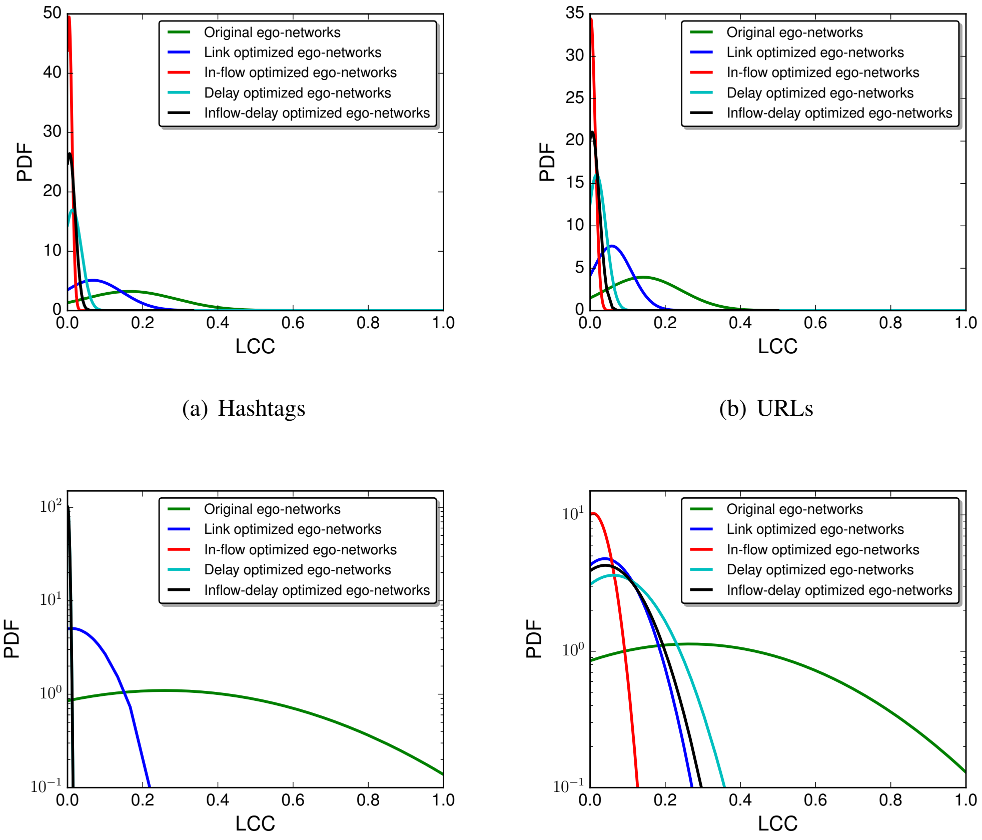
(b) Minimal set cover

Figure 2.16: Ego-networks for a Twitter user (red node). Some users (black nodes) only belong to one of the ego-networks while others (gray nodes) belong to both. The original ego-network contains more triangles than the ego-network induced by the minimal set cover, whose structure is closer to a star.

First, as an example, we take one particular user and illustrate the structure of her original ego-network and the ego-network of an optimal set in terms of link efficiency (Figure 2.16). By visual comparison of both ego-networks, we can see that while the ego-network induced by the optimal set displays a structure much closer to a star, the original ego-network contains many more triangles and higher clustering coefficient. Due to its proportionally lower number of triangles, the optimal set is not *discoverable* by triadic closure [Simmel, 1950, Granovetter, 1973, Romero and Kleinberg, 2010a] or information diffusion [Bakshy et al., 2012], which have been recently shown to be two major driving forces for link creation in social networks [Weng et al., 2013, Myers and Leskovec, 2014, Antoniades and Dovrolis, 2013].

Remarkably, this phenomenon happens systematically across all users, efficiency definitions, and types of memes, as displayed in Figure 2.17, which shows the distribution of local clustering coefficient (LCC) for the users' original ego-networks and the ego-networks induced by different optimized sets. We find that while the LCC distribution for the original ego-networks is well spread and centered at $0.15 - 0.30$, the LCC distributions for the ego-networks induced by the optimal sets are skewed towards zero.[4] One could still think that this is simply a consequence of differences in the number of followees, *i.e.,* the size of the ego-network. However, Figure 2.18 rules out this possibility by showing a

[4] Note that the distribution of clustering coefficient of inflow-delay optimized network is located between distributions of in-flow optimized and delay optimized ego-networks.



(a) Hashtags

(b) URLs

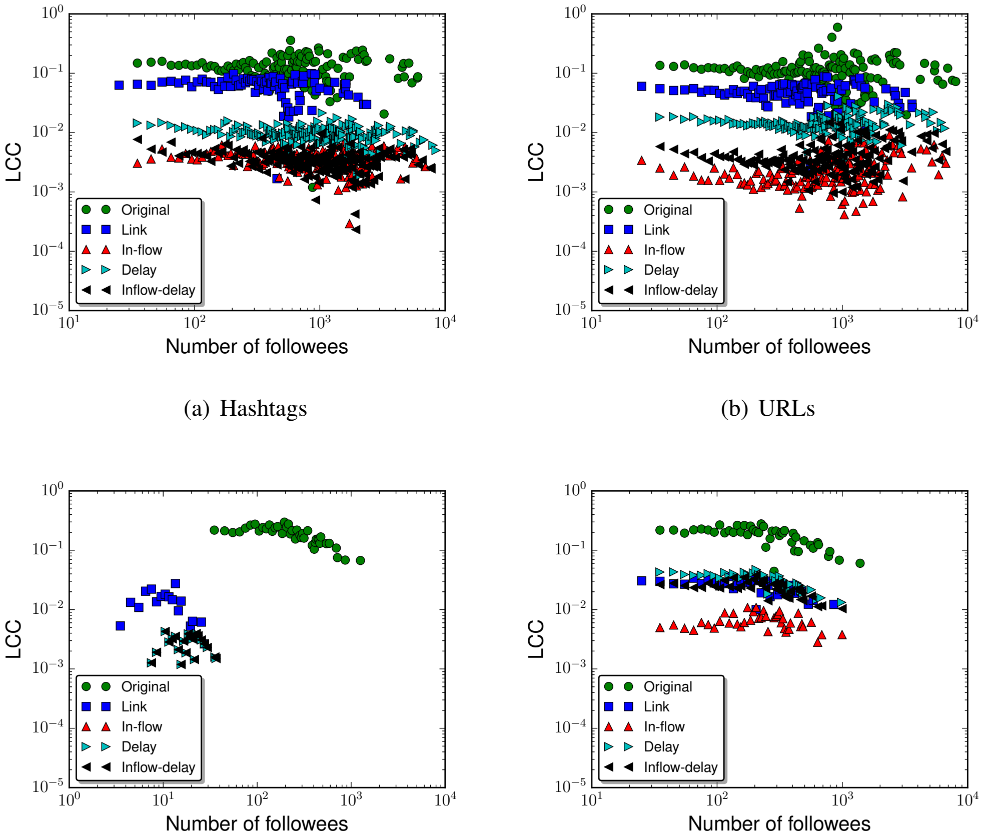
(c) News domains

(d) YouTube videos

Figure 2.17: The distributions of local clustering coefficient (LCC) of the original ego-network (green circles) and the ego-networks optimized for link (blue circles), in-flow (red circles), delay (teal circles), and inflow-delay efficiency (black circles) for: (a) hashtags, (b) URLs, (c) news domains, (d) and YouTube videos.

striking difference of several orders of magnitude between the LCC of the original ego-networks and the ego-networks induced by the optimal sets across a wide range of number of followees. These findings suggest that the way in which social media users discover new people to follow (*e.g.,* triadic closure or information diffusion) or receive recommendations (*e.g.,* pick people in a 2-hop neighborhood [Adamic and Adar, 2003]) can lead to sub-optimal information networks in terms of (link, in-flow and delay) efficiency.

We have argued that optimal sets typically differ from the original set of followees due to their low number of triangles in the associated ego-networks, and thus lack of discoverability. However, are optimal sets with higher number of triangles in efficient ego-networks easier to discover for users? Figure 2.19 answers this question positively by showing the average local clustering coefficient in the



(a) Hashtags

(b) URLs

(c) News domains

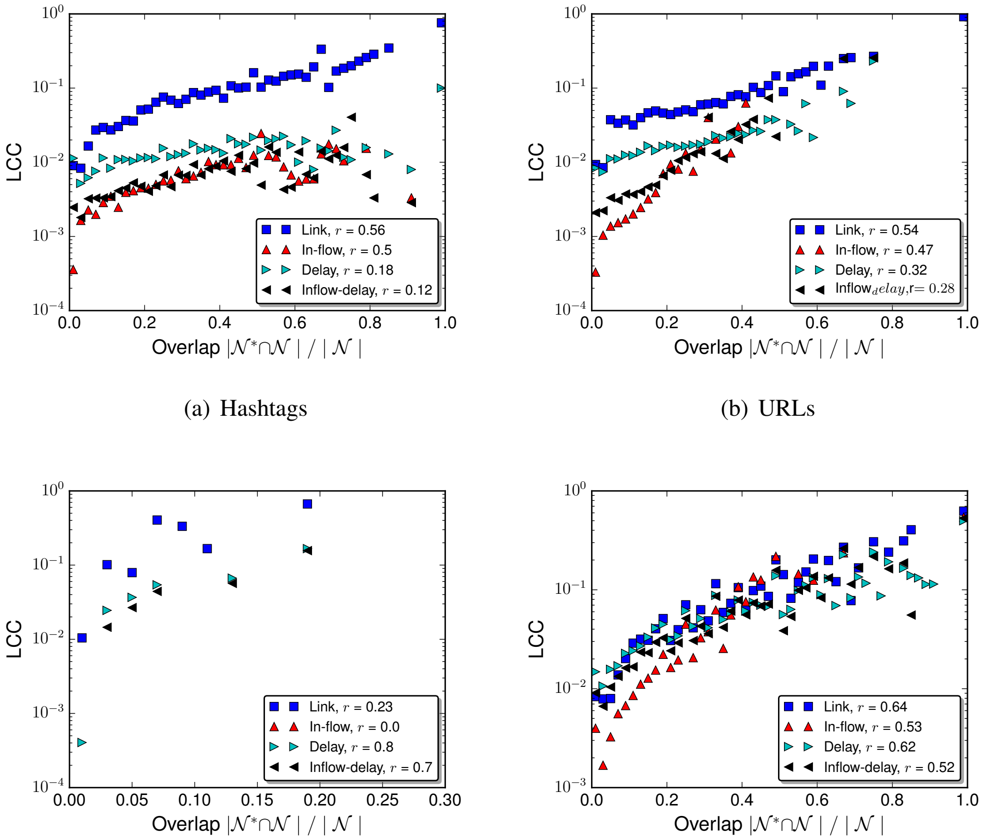
(d) YouTube videos

Figure 2.18: Average local clustering coefficient versus the number of followees in the original ego-network (green circles) and the ego-networks optimized for link (blue squares), in-flow (red triangles), delay (teal triangles), and inflow-delay efficiency (black triangles).

ego-network induced by the optimal set against the overlap between the users in the optimal set and the original set of followees. Here, by overlap we mean the fraction of users in the optimal set that are also in the original set of followees. In particular, we find a positive correlation (Pearson's $0.07 < r < 0.55$, $p < 10^{-10}$) between the local clustering coefficient and the overlap, which indicates that if the nodes in the optimal set are *discoverable* through triadic closure, the user may be more likely to find them and decide to follow them.

2.7 Summary

In conclusion, we propose a computational framework to quantify users' efficiency at selecting information sources. Our framework is based on the assumption that the goal of users is to acquire a set



(a) Hashtags

(b) URLs

(c) News domains

(d) YouTube videos

Figure 2.19: Local clustering coefficient of the optimized ego-networks versus the overlap between the original ego-network and the ego-networks optimized for link (blue squares), in-flow (red triangles), delay (teal triangles), and inflow-delay efficiency (black triangles).

of unique pieces of information. To quantify user's efficiency, we ask if the user could have acquired the same pieces of information from another set of sources more efficiently. We define three different notions of efficiency – link, in-flow, and delay – corresponding to the number of sources the user follows, the amount of (redundant) information she acquires and the delay with which she receives the information. Our definitions of efficiency are general and applicable to any social media system with an underlying information network, in which every user follows others to receive the information they produce. In our experiments, we measure the efficiency of Twitter users at acquiring different types of information. We find that Twitter users exhibit sub-optimal efficiency across the three notions of efficiency, although they tend to be more efficient at acquiring nonpopular pieces of information than they are at acquiring popular pieces of information. We then show that this lack of efficiency is a consequence of the triadic closure mechanism by which users typically discover and follow other users in social media. Thus, our study reveals a tradeoff between the efficiency and discoverability of information sources. Finally, we develop a heuristic algorithm that enables users to be significantly more efficient at acquiring the same unique pieces of information.

CHAPTER 3

Break the filter bubbles that users get trapped

With an increasing number of people relying on social media platforms to acquire their information, there have been growing concerns about the impacts such a shift can have on their news consumption. Although politically diverse news publishers post stories, it has been observed that readers often focus on the news, which reinforces their pre-existing views, leading to ‘filter bubble’ or ‘echo chamber’ effects. These biases in the news consumption may lead to an increase in societal polarization. To combat this, some recent systems such as Wall Street Journal’s ‘Blue Feed, Red Feed’ system nudge readers toward diverse stories with different points of view by showing both sides of a topic posted from biased publishers. However, recent work shows that exposure to opposing views on social media can increase political polarization [Bail et al., 2018]. Alternatively, we present a complementary approach which identifies non-divisive (high consensus) ‘purple’ posts that generate similar reactions from readers with different political leanings. We also propose and quantify the benefits of a strategy to spread high consensus news across readers with diverse political leanings. The aim is to spread high consensus news, and quantify the significant decrease in the disparity of users’ news exposure.

We begin by giving a brief overview of the related work and background. Then, in the first part of this chapter (Section 3.2), we discuss identifying non-divisive content. Finally, in the second part of this chapter (Section 3.3), we focus on studying the properties of non-divisive news to see if it is possible to propagate non-divisive news broadly across society (may help to break filter bubbles and lead to decrease polarization in society).

3.1 Background

We review diverse and polarized news dissemination and information propagation in social networks. We then briefly discuss different supervised learning methods and Indipendent Cascade model (IC) in which we consider them for detection of non-divisive news (high consensus) and our propagation model, respectively.

3.1.1 News consumption polarization on social media

Several recent studies have investigated the dissemination of news in social networks [Bhattacharya and Ram, 2012], focusing on biases [Chakraborty et al., 2016a], political news [An et al., 2014], and the characteristics of spreaders [Hu et al., 2012].

Traditionally, professional news organizations played a major role in spreading news by selectively presenting news stories to citizens [Shoemaker and Vos, 2009]. Accordingly, news media had a high impact on political issues and public opinions [Groseclose and Milyo, 2005, Chiang and Knight, 2011]. Several works have focused on understanding how and to what extent news media outlets can impact people and society, such as the White Helmets in Syria [Starbird et al., 2018] and the 2016 US presidential campaign [Rizoiu et al., 2018, Ribeiro et al., 2019].

By examining cross-ideological exposure through content and network analysis, [Himelboim et al., 2013] showed that political talk on Twitter is highly partisan and users are unlikely to be exposed to cross-ideological content through their friendship network. Other studies also report similar findings such as users' higher willingness to communicate with other like-minded social media users [Liu and Weber, 2014].

To understand the political bias in social media better, many researchers have studied political polarization on Twitter by analyzing different groups' behavior. [Conover et al., 2011] showed that Twitter users usually retweet the users who have the same political ideology as themselves, making the retweeting network structure highly partitioned into left- and right-leaning groups with limited connections between them.

Previous works have mostly investigated news media political bias, and the bias introduced in the content of the news, by different methods such as crowdsourcing and machine learning [Budak et al., 2016, Gentzkow and Shapiro, 2010, Babaei et al., 2018]. Here, we propose a complementary approach [Babaei et al., 2018], in which the goal is to inject diversity in users' information consumption by identifying non-divisive (high consensus) yet informative news, based on using features such as the publishers' political leaning.

We show that non-divisive and divisive (high and low consensus) posts are equally popular and cover broadly similar topics. Then, we investigate how non-divisive and divisive news spread through social media and their potential impacts on readers biased exposure, which is one of our main concerns in this thesis.

3.1.2 Collecting users' political leanings

We inferred every user's political leaning as a score between -1 and +1, using the method of [Kulshrestha et al., 2017], in which, we needed to collect their followees. Inferring the political leaning of a given Twitter user u is based on the following steps – (*i*) generating two representative sets of users who are known to have a democratic or republican bias, (*ii*) inferring the topical interests of u by looking at her followees, and (*iii*) examining how closely u's interests match with the interests of the representative sets of democratic and republican users. Formally,

$$leaning(u) = cos\_sim(I_u, I_D) - cos\_sim(I_u, I_R), \tag{3.1}$$

where I_u is the interest vectors of user u, and I_D, I_R are normalized aggregate interest vectors for the democrat seed set (I_D) and the republican seed set (I_R). Similarity between interest vectors are measured by cosine similarity.

3.1.3 Information propagation in social networks

The process of increasing information propagation and network diffusion by identifying and choosing the optimal set of individuals that utilize social influences to maximize adoption or reception of information in society has been widely studied [Goyal et al., 2013, Richardson and Domingos, 2002, Kempe et al., 2003, Hartline et al., 2008]. [Richardson and Domingos, 2002] considered influence maximization as an algorithmic problem. To solve it, the authors used a heuristic approach to find an initial set of nodes to maximize the number of further adapters. The effectiveness of these strategies is studied by Kempe et al. [2003] under different social contagion models such as Linear Threshold (LT) and Independent Cascade (IC) models. They showed that finding the optimal solution is NP-hard. Motivated by its hardness, Kempe et al. used influence function properties such as monotonicity and submodularity to obtain provable approximation guarantees. Since then, various related extensions have been studied [Goyal et al., 2013, Bharathi et al., 2007, Budak et al., 2011, Carnes et al., 2007].

Here our goal is to propagate news with less disparity among users with different political leanings. Recent studies on fair influence maximization [Ali et al., 2019a, Tsang et al., 2019, Khajehnejad et al., 2020a] are the most related ones to our work. However, their approaches may not be directly applicable to online networks such as Twitter. We show that in social media, the political leaning of the seeds can make a considerable bias in users' exposure to news. In particular, we observe that high consensus news posted by neutral publishers has the lowest disparity for spreading among all users (liberal, conservative, and neutral). We use the fair influence maximization method proposed by [Ali et al., 2019a] as a baseline.

3.1.4 Supervised learning methods

We use supervised learning approaches to identify whether a news tweet has high consensus or low consensus using the features described later. Thus, we applied three different categories of supervised learning.

I. Non-probabilistic classifier: Since model built by SVM training approach assigns new data to one class or the another, it is known as non-probabilistic classifier. In this work, our goal is to distinguish the high and low consensus tweets. In other words, we want to create a model based on our training data to detect a new tweet as a high or low consensus. Thus, SVM is one of the suitable classifiers for our goal.

II. Probabilistic classifier:

. Naive Bayes: This method is also known as posterior class probabilities. Since Naive Bayes can efficiently handle items with sparse features and also assumes the strong independency between the features, it is a popular baseline for text categorization. It also is also very helpful for tweet classification because there is 140-character limit for each tweet that means each tweet includes a few features. Thus, we consider this classifier for our goal.

. Logistic Regression: This classifier is suitable for dependent features. When the output contains two classes, then binary logistic regression helps efficiently. Since we are using different categories of features that some of them depend on others, then it is desirable to apply this method.

III. Ensemble classifier: Random forest is a kind of ensemble algorithms. It uses multiple machine learning algorithms to achieve better accuracy. Multiple decision trees are used to classify the data. To classify the test data, the algorithm put down the data in each tree and each tree votes for that. Then, it chooses the highest vote for classification . It assigns estimation and weight to each variable which present importance of each feature . Random forest achieves high accuracy and it runs efficiently on large data set

3.1.5 Independent cascade model (IC)

In the IC model, information propagates through every edge (v, w) with probability p_{vw}. We have a set of discrete time steps which we denote with $t = \{0, 1, 2, \cdots\}$. At $t = 0$, the initial seed set $S \subseteq \mathcal{V}$ is activated. At every time step $t > 0$, a node $v \in V$ which was activated at time $t-1$ can activate its inactivated neighbors w with probability p_{vw}. The model assumes that once a node is activated, it stays active throughout the whole process and each node has only one chance to activate its neighbors. The described process stops at time $t > 0$ if no new node gets activated at this time. We note that the IC model is a stochastic process, in which a node u can influence its neighbors w based on the Bernoulli distribution with success probability p_{uw}. A possible outcome of the process can be denoted via a set of timestamps $\{t_v \geq 0 : v \in \mathcal{V}\}$, where t_v represents the time at which a node $v \in V$ is activated.

3.2 Identifying non-divisive content from controversial topics

Within our societies, there are many topics for which different subgroups hold opposing ideological positions. For example, there are primarily two distinct political affiliations in the U.S.: Republicans (the 'red' group) and Democrats (the 'blue' group). Social media platforms provide a wide variety of news sources covering this ideological spectrum, yet many users largely limit themselves to news stories which reinforce their pre-existing views. This *selective exposure*, where red users read red news and blue users read blue news, leads to a more politically fragmented, less cohesive society [Liu and Weber, 2014]. Further, this selective exposure effect is often amplified by social media platforms which recognize users' preferences and thence recommend more red news to red users and more blue news to blue users. While this approach may work well for recommending consumer goods such as movies or music, there are concerns that such stilted news selections limit exposure to differing perspectives and lead to the formation of 'filter bubbles' or 'echo chambers' [Bakshy et al., 2015, Bozdag, 2013, Flaxman et al., 2016, Pariser, 2011], resulting in a worrying increase in *societal polarization* [Sunstein, 2002, Schkade et al., 2007].

To combat this polarization, a number of systems intended to promote diversity have been proposed. These systems deliberately expose users to different points of view by showing red news to blue users, and blue news to red users; or by showing both red and blue news to both red and blue user groups. The hope is to nudge users to read viewpoints which disagree with their own [Munson et al., 2013a, Park et al., 2009b]. A prominent example is the Wall Street Journal's 'Blue feed, Red feed' system [Keegan, 2017], which presents posts from the most extreme news publishers on Facebook, with the aim of showing diametrically opposed perspectives on news stories.

Unfortunately, however, such systems have had limited success. While some diversity-seeking users enjoy the added perspectives, many users either ignore or reject disagreeable points of view [Munson and Resnick, 2010]. Indeed, by confronting users with the most radical posts from the other ideological side, such systems may even *increase* polarization by encouraging users to retreat to a more entrenched version of their initial position [Lord and Ross, 1979, Miller et al., 1993, Munro and Ditto, 1997].

In this work, we propose a complementary approach by identifying and highlighting news posts which are likely to evoke similar reactions from the readers, irrespective of their political leanings. We define these non-divisive or 'purple' news posts to be those with *high consensus*, *i.e.,* having a general agreement in their readers' reactions to them. We propose that these high consensus purple stories could be recommended to both red and blue users, evoking a more unified response across society, which we hope might lead to lower segregation in information consumption [Chakraborty et al., 2017], and might help to promote greater understanding and cohesion among people. In Table 3.1, we show a sample of red, blue and purple news stories to highlight the differences between the three types of stories.[1]

Given this context, we investigate the following questions:

[1] See en.wikipedia.org/wiki/Dismissal_of_James_Comey.

Low Consensus **Conservative**	Fox News: Schieffer Slams Trump: Comey Firing Reminds Me ofJFK-Oswald ConspiraciesSource, 55 Retweets, 510 Replies, 149 Likes.
	Fox News: @POTUS: “All of the Democrats, I mean, they hated Jim Comey. They didn’t like him, they wanted him fired".https://t.co/1ebOtqfIOc 491 Retweets, 293 Replies, 2k Likes.
	Politico: Analysis: Is this a constitutional crisis? Legal experts size up the Comey firing. http://politi.co/2qPEN1c, 210 Retweets, 63 Replies, 254 Likes.
Low Consensus **Liberal**	The New York Times: What all the Russia investigations have done and what couldhappen nextSource, 131 Retweets, 52 Replies, 226 Likes.
	Salon: Report: Trump "revealed more information to the Russian ambassadorthan we have shared with our own allies" 51 Retweets, 14 Replies, 34 Likes
	CNN: Is Donald Trump the "little boy President"? A @CNNOpinioncontributor takes a closer look at his latest moveshttp://cnn.it/2pE1Uaq, 179 Retwees, 264 Replies, 468 Likes.
High Consensus	Fox News: @johnrobertsFox on firing of James Comey: "This came as a shockto literally everyone, including the @FBI Director." #TheFiveSource, 156 Retweets, 259 Replies, 574 Likes.
	AP: BREAKING: Senate intelligence committee invites fired FBI DirectorComey to appear in closed session next Tuesday, 2.7K Retweets, 176 Replies, 5k Likes
	Politico: James Comey told lawmakers he wanted more resourcesfor Russia probe http://politi.co/2r2HxpfSource, 132 Retweets, 40 Replies, 204 Likes.

Table 3.1: Samples of high and low consensus news posts. First and second rows include low consensus news with conservative and liberal leaning respectively.

1. How can we define the consensus of news posts in order to operationalize the identification of high consensus purple posts?

2. Do helpful purple news posts exist on social media?

3. How do purple posts compare with low consensus (blue or red only) posts?

4. Can we automate the identification of consensus of news posts on social media in order to discover purple posts?

Our work provides a fresh tool which we hope will help to break filter bubbles, encourage healthier interaction between population subgroups, and lead to a more cohesive society.

3.2.1 Consensus definition and measurement

A key step of our work is to understand whether news posts with high consensus exist in social media. To verify this, first we need to operationalize the concept of ‘consensus’ for news posts, i.e., to provide a definition for consensus that allows one to measure it, both empirically and quantitatively. Second, we need to construct ground truth datasets to measure consensus of real news posts in social media. Next, we describe how we performed these steps.

Operationalizing consensus for news posts

According to the Oxford English Dictionary, consensus is defined as “a general agreement”.[2] Inspired by this definition, we consider a post to have *high consensus if there is a general agreement in readers' reaction to it, irrespective of their own political leaning*. Specifically, in the context of US politics, a post would have high consensus if the reaction of Democrat readers to the post is similar to the reaction of Republican readers. For a given social media post, we measure the reaction of Democrats and Republicans as whether the readers agree or disagree with the content of a post. Formally, we measure the amount of consensus as

$$\text{consensus} = 1 - |\frac{\#D_{disagree}}{\#D} - \frac{\#R_{disagree}}{\#R}| \tag{3.2}$$

where $\#D_{disagree}$ and $\#R_{disagree}$ respectively denote the number of Democrats and Republicans who disagree with the post, while $\#D$ and $\#R$ are the total number of Democrats and Republicans.[3] A consensus value closer to 1 indicates that both Democrats and Republicans disagreed with it to similar extents, thereby indicating high consensus; while a value closer to 0 is indicative of low consensus.

Note that there is no unique way to measure consensus. In addition to Equation 3.2, it can also be measured in terms of *attitude polarization indices* such as coherence, divergence, intensity, and parity [Persily, 2015]. The common requirement for these indices is *attitude response data*, which in our context is provided by the support and negative response from the readers. We use our definition presented in Equation 3.2 due to its simplicity, while still effectively capturing the nuances of consensus as in the other measures.

[2] See en.oxforddictionaries.com/definition/consensus.

[3] Considering the fraction of readers from each side who disagree with a post implicitly takes into account the fraction of readers who have neutral or favorable reactions to it.

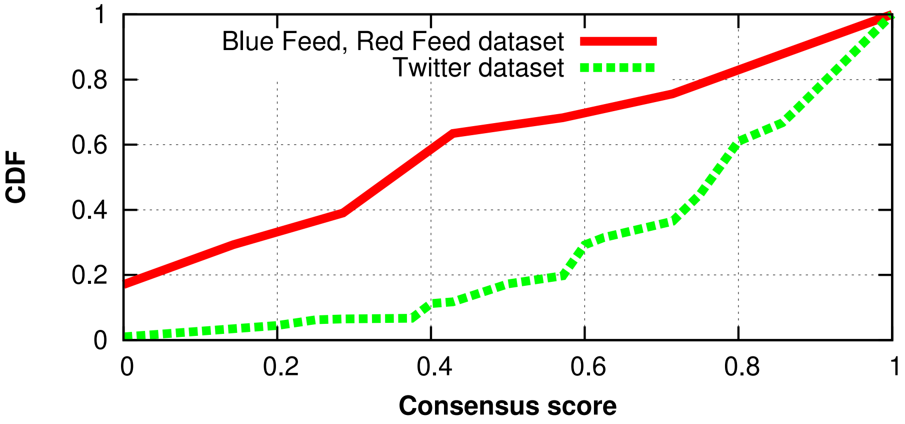


Figure 3.1: Distribution (CDF) of consensus values of news posts from the two datasets.

Measuring consensus of news posts on social media

Using our definition of consensus, we conducted an Amazon Mechanical Turk (AMT) experiment to quantify the consensus of two distinct datasets of news posts:

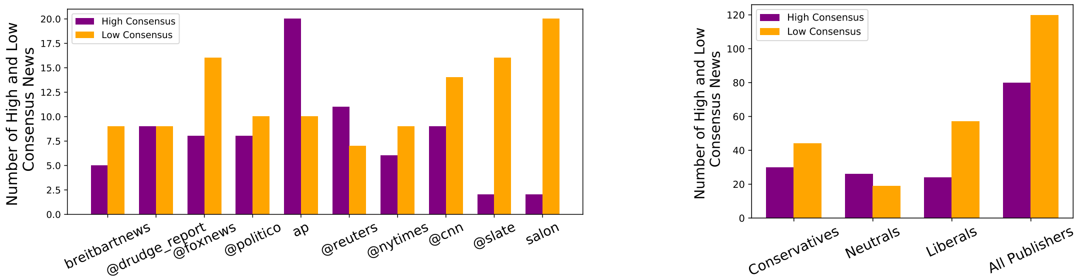
(i) *Blue Feed, Red Feed dataset* - Using the Wall Street Journal's 'Blue Feed, Red Feed' system [Keegan, 2017], we collected the top 10 posts from each of the liberal and conservative sides for the queries "Trump" and "healthcare", giving us a total of 40 news posts.

(ii) *Twitter dataset* - We also collected 40 news posts tweeted by each of the following 10 news publishers with well known political biases varying from liberal to neutral to conservative: Slate, Salon, New York Times, CNN, AP, Reuters, Politico, Fox News, Drudge Report, and Breitbart News; giving us a total of 400 posts. These news posts were collected during the one week period of 9th to 15th May, 2017.

In this experiment, we only recruited AMT workers from the US and at the end of the experiment, we also collected their political leanings. We showed every news post to workers and asked them for their reaction to the post by selecting one out of three options – agreement, neutral or disagreement.[4] After the experiment, applying Equation 3.2 to responses from an equal number (seven) of Democrat and Republican AMT workers, we computed the consensus values for the news posts in our two datasets.

Figure 3.1 shows the distribution (CDF) of consensus values for the news posts from our datasets. We observe that news posts from the 'Blue Feed, Red Feed' dataset are skewed towards lower values of

[4] We presented the questions in the following form:
Tweet: "Salon: Can anyone find an economist who thinks Trump's tax cuts will pay for themselves? https://t.co/Hz6JvQHvXB"
Question: Do you agree with this tweet?
(a) Agree, (b) Neither agree nor disagree, or (c) Disagree.



(a)

(b)

Figure 3.2: Number of high and low consensus news posted on Twitter during 9th-15th May, 2017. (A) shows the number of high and low consensus news for 10 selected publishers, and (B) shows the aggregated result for conservative, liberals, and neutral publishers.

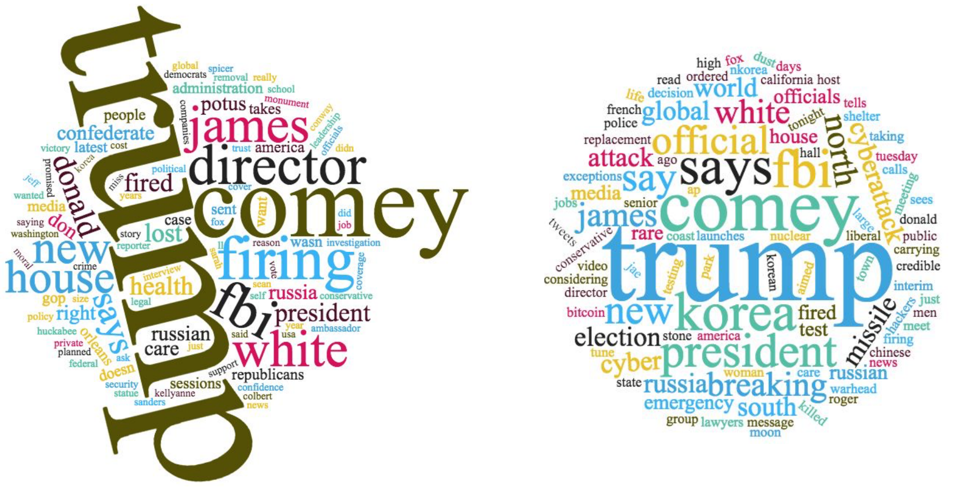
consensus, indicating that the readers from the two different parties have different reactions to them; whereas the random news posts from the 10 publishers from the Twitter dataset have a noticeable skew towards higher consensus. Our observations suggest that while news outlets on social media do publish posts with varying degrees of consensus, systems such as ‘Blue Feed, Red Feed’, which highlight posts from extremely biased news outlets, tend to pick lower consensus content which leads readers with different leanings to react differently. Figure 3.2(b) shows the total number of high consensus and low consensus news posted by the same 10 publishers, grouped into liberal, conservative, and neutral categories. It also shows the total number of high consensus and low consensus news posted by all the 10 publishers. We can see that the total number of low consensus posts are considerably higher than the total number of high consensus posts.

As we discuss later, low consensus posts have a much smaller chance of being received by users with different political leanings, which leads to a more politically fragmented society. On the other hand, high consensus posts have a better chance of spreading through communities with various political leanings, and can be utilized to break the filter bubbles.

In this work, we make a case to promote high consensus posts in order to increase exposure to ideologically cross-cutting content, which may help in lowering societal polarization.

3.2.2 Empirical study of consensus of news posts

Given that news posts with high consensus do exist, next we conduct an empirical study on consensus of news posts on social media. Our main goal in this section is to understand if news posts with high consensus are interesting to users (*i.e.,* do they become popular), if they cover a wide range of topics,



(a) Low consensus tweets

(b) High consensus tweets

Figure 3.3: Topical coverage of low & high consensus tweets.

and if they expose users to relatively more cross-cutting content. We refer to the 100 news posts with the highest consensus values in our ground truth Twitter dataset as *high consensus* news posts, and the 100 tweets with lowest consensus values as *low consensus* news posts.

To what extent do high and low consensus news posts become popular? To measure popularity, we count the number of retweets for high and low consensus tweets. On average, high consensus tweets are retweeted 158 times, whereas low consensus tweets are retweeted 177 times, i.e. slightly more often but the numbers are very close. We observe a similar pattern when we compare the median number of retweets of high consensus tweets (93) with low consensus tweets (89), indicating that both high and low consensus tweets engage their readers to similar extents. This suggests that recommendation systems which highlight high consensus tweets would feature content which is of similar popularity to that generated by systems which highlight low consensus tweets.

Do high and low consensus news posts cover different topics? To verify whether high and low consensus tweets cover similar (or very different) topics, we present the 100 most common words for both sets of tweets in Figure 3.3. From the figure, it is evident that although both sets do cover popular political topics (*e.g.,* 'Trump', 'Comey', 'FBI' and other topics associated with FBI director James Comey's dismissal), high consensus tweets are topically more diverse and also contain posts on non-US centric political topics (*e.g.,* 'North Korea') and other more niche topics (*e.g.,* 'jobs', 'cyberattack').

Do high consensus posts lead to more exposure to ideologically cross-cutting content? To investigate whether highlighting high consensus tweets leads to higher exposure to ideologically cross-cutting contents, we examine whether the higher consensus tweets have relatively more retweets from the users of opposite leaning (with respect to the publisher's leaning), when compared to lower consensus tweets. This analysis is motivated by the reasoning that as users of opposite leaning retweet the publisher's tweets, more opposite leaning users from these users' neighborhoods would get exposed to them, leading to higher exposure to cross-cutting content for users, and potentially lower polarized news consumption on social media.

To validate whether our reasoning holds, we consider a particular tweet to have *high cross-cutting exposure* if the number of opposite leaning retweeters for this tweet is higher than the baseline number of opposite leaning retweeters of its publisher (computed as the average across 100 random tweets of the publisher). When we rank the tweets by their consensus values and compare the top and bottom 10% tweets, we find that a much larger fraction (45%) of high consensus tweets have high cross-cutting exposure than low consensus tweets (30%), indicating that high consensus tweets indeed lead to higher exposure to cross-cutting content.

3.2.3 Identifying high and low consensus news posts on social media

After empirically exploring the consensus of social media news posts, we now turn our attention towards *automatically identifying* high and low consensus news posts, which can scale up to cover a large number of news publishers on Twitter. In this section, we first briefly discuss different features of social media posts that have been applied in prior prediction and classification tasks. Then, we propose and validate a novel class of *audience leaning based features* which are ideally suited for our consensus identification task.

3.2.4 Features used in prior work

Prior works on classification and prediction tasks for social media posts have mostly used two broad types of features: *publisher based*, and *tweet based* features. For instance, the political leaning of the publisher has been used to quantify the tweet's leaning [Kulshrestha et al., 2017], or the leaning of news story URLs being shared by them on Facebook [Bakshy et al., 2015]. Others have used tweet based features for predicting the relevance of a tweet for a topic [Tao et al., 2012], to rank tweets [Duan et al.,

Feature Category	**Features**
Publisher based	Number of followers/friends/tweets Average number of retweets/replies/favorites Political leaning, Language, Location
Tweet based	Bag of words, Creation time Number of retweets/replies/favorites

Table 3.2: Features used in prior work. The three most important features from each category are highlighted in blue.

2010], or to quantify to what extent a tweet is interesting [Naveed et al., 2011]. Many other studies have combined both publisher and tweet based features for various tasks including predicting future retweets [Petrovic et al., 2011, Suh et al., 2010], and even predicting users' personality traits [Golbeck and Hansen, 2011]. Table 3.2 shows the features from each class which we are aware were used previously.

3.2.5 Our proposed audience leaning based features

We propose a novel class of *audience leaning based* features, which to our knowledge have not previously been used for predicting and classifying tweet properties. We use these features to identify high and low consensus posts on Twitter.

For every tweet, there are three types of audience:

(i) *Followers* of the publisher of the tweet – they are the passive supporters of the post (on average 67% of followers are of the same political leaning as the publisher[5]),

(ii) *Retweeters* of the tweet – they are more active supporters of the post (on average 78% of retweeters are of the same leaning as the publisher), and

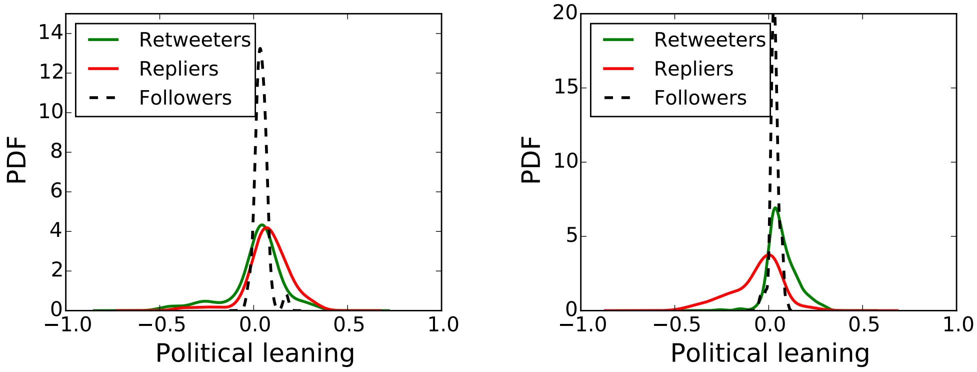
(iii) *Repliers* to the tweet – they are usually a mix of users supporting or opposing the news post (on average 35% of repliers are of the opposite leaning to the publisher). In Table 3.3, we show a random sample of replies from our Twitter dataset, and notice that many of them oppose either the news content or the publisher.

We hypothesize that we can use the political leaning distributions of the three audiences of a post to quantify whether different readers of a post are having similar reactions to it (*i.e.,* to measure consensus). To demonstrate our hypothesis, we select one high consensus and one low consensus post for which we

[5] Followers of famous politicians (e.g., President Trump) indeed include many users (such as journalists) from both ends of the political spectrum, who may not necessarily support him or his views.

@CNN You mean, like the UNFOUNDED claims of Russian collusion? You people are typically selective in your bias pro? https://t.co/CESkVpIZOk
@nytimes His actions were disgraceful. Being fired does not make him a sympathetic figure. He affected the outcome? https://t.co/bIbiuj2CJJ
@BreitbartNews I just wonder, what motivates these libtards... https://t.co/mzpBIKdPr4
@CNN hey fakenews do some homework, get out of office! Every illegal that get a drivers license is registered to vote dem! I'd card, regs!
@AP Jews are so desperate to take over Syria that they will make up anything.

Table 3.3: Random sample of replies for tweets in our dataset.



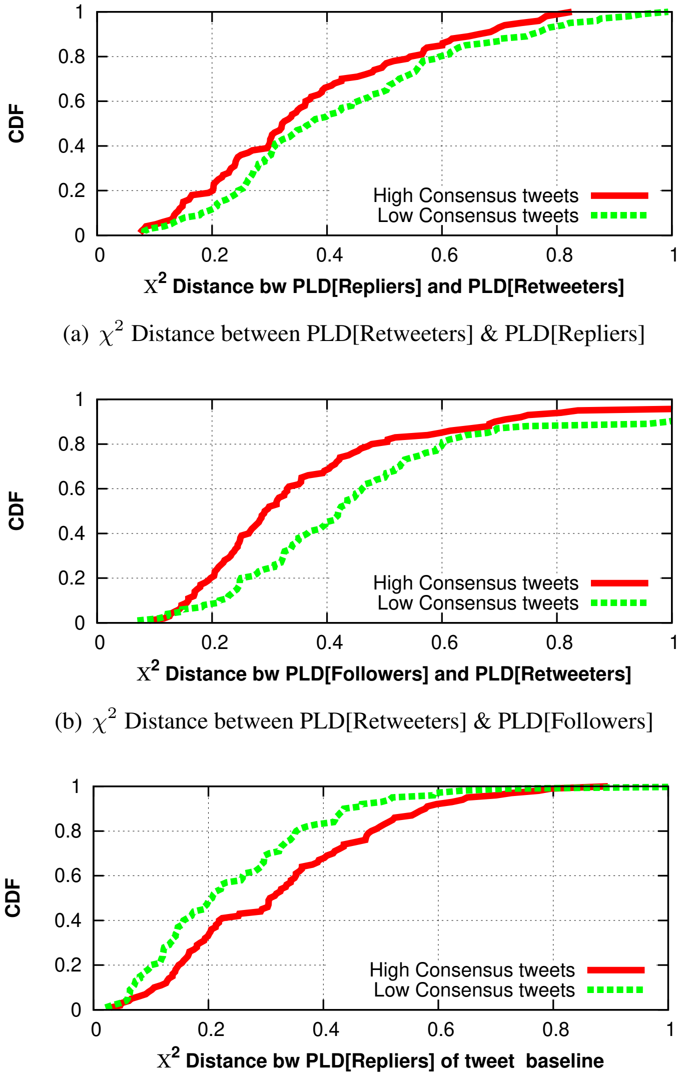
(a) High consensus tweet

(b) Low consensus tweet

Figure 3.4: Distributions of political leanings of different audiences for the following news posts: (A) High consensus: "Trump ordered emergency meeting after global cyber attack: official http://reut.rs/2r6Qkt8" posted by Reuters, (B) Low consensus: "Michelle Obama criticizes Trump administration's school lunch policy http://cnn.it/2qckHwZ" posted by CNN .

computed consensus values using AMT workers' judgments, and then computed the political leaning distributions of the three audiences.

Inferring political leaning of Twitter users is a research challenge on its own, and beyond the scope of this work. We adopt the methodology proposed in [Kulshrestha et al., 2017], which returns the political leaning of a Twitter user in the range $[-1.0, 1.0]$, with scores in $[-1.0, -0.03)$ indicating Republican leaning, $[-0.03, 0.03]$ indicating neutral and $(0.03, 1.0]$ indicating Democrat leaning. In Figure 3.4, we plot the political leaning distributions of the three audiences, for a high consensus and a low consensus post.



(a) χ^2 Distance between PLD[Retweeters] & PLD[Repliers]

(b) χ^2 Distance between PLD[Retweeters] & PLD[Followers]

(c) χ^2 Distance between PLD[Repliers] & publisher baseline PLD[Repliers]

Figure 3.5: Distributions of χ^2 distance between different audience political leaning distributions for 25% tweets with highest and lowest consensus values.

We can observe that there is a striking difference between the audience leaning distributions of high and low consensus tweets in Figure 3.4. For the high consensus tweet, these distributions are much more similar than for the low consensus tweet. More interestingly, retweeters typically being supporters, have similar political leaning distribution as the followers of the publishers (for both types of posts). However, for a lower consensus post, repliers being opposers, have a different distribution. Therefore, we find that the *degree of similarity of the leaning distributions of the audiences* of the post contains a useful signal to approximate the consensus for a post (*i.e.,* the similarity in reaction of readers of different leanings). We compute the χ^2 distances between the leaning distributions of the different audiences to capture their similarities. In Figure 3.5, we show the distribution of these χ^2 values for high and low tweets. The difference in the distributions for the high and low consensus posts give evidence for the discriminative power of these features.

Building upon these observations, we construct a number of audience leaning based features by utilizing the political leanings of the three types of audiences of a tweet. Table 3.5 lists all such features, which we use in this work.

3.2.6 Experimental evaluation

We first describe our experimental setup, then present our results for the aforementioned categories of features.

Experimental setup: We use supervised learning approaches to identify whether a news tweet has high consensus or low consensus using the features described in the previous section. For setting up the classifiers, we first need a ground truth dataset of high and low consensus tweets. We use the consensus values computed using AMT workers' judgments for the Twitter dataset described previously and label the top 25% consensus value tweets as high consensus, and bottom 25% tweets as low consensus tweets. We use this set of 200 labeled tweets as our ground truth dataset.

Using the features described earlier, we apply four different types of supervised learning classifiers for our task of tweet consensus classification: Linear SVM, Naive Bayes, Logistic Regression and Random Forest classifiers. While using textual features of the tweets, we follow a two step approach as described in [Chakraborty et al., 2016b, 2018]:

(i) first, we treat the textual features as bag-of-words and use Naive Bayes classifier to predict the class

Classifier	Different feature categories				
	Publisher based (P)	Tweet based (T)	P and T	Audience leaning based (A)	P, T, and A
Logistic Regression	0.58 ±0.008	0.58 ±0.008	0.68 ±0.009	0.72 ±0.012	0.72 ±0.011
Linear SVM	0.58 ±0.008	0.58 ±0.008	0.68 ±0.009	0.72 ±0.012	0.72 ±0.011
Naive Bayes	0.59 ±0.007	0.57 ±0.015	0.60 ±0.01	0.66 ±0.015	0.66 ±0.012
Random Forest	0.58 ±0.008	0.57 ±0.01	0.64 ±0.01	0.67 ±0.015	0.67 ±0.017

Table 3.4: Average accuracies and 90% confidence intervals for different categories of features used for predicting consensus of news tweets. Our proposed audience leaning based features perform best for this news post consensus classification task.

using these textual features, and (ii) then we input these prediction outputs of Naive Bayes classifier as features (along with our other features) to the different classifiers as the second step.

For training our classifiers, we use 5-fold cross-validation. In each test, the original sample is partitioned into 5 sub-samples, out of which 4 are used as training data, and the remaining one is used for testing the classifier. The process is then repeated 5 times, with each of the 5 sub-samples used exactly once as the test data, thus producing 5 results. The entire 5-fold cross validation was then repeated 20 times with different seeds used to shuffle the original dataset, thus producing 100 different results. The results reported are averages of the 100 runs, along with the 90% confidence interval. Also, we use feature ranking with recursive feature elimination that prunes out the insignificant features by obtaining their importance from the supervised techniques.[6]

Experimental results: We successively implemented the different classifiers first using features from each category separately, and then by combining the features from different categories. Accuracies are shown in Table 3.4. We observe that the tweet based features have the worst performance. This poor performance is most likely due to the short size of the tweets, which often means that there is very little information in the tweet text and it is hard to understand them without also inspecting the content of weblink, photograph or video included in the tweet. The performance of publisher based features is better than that of tweet based features. The political leaning of the publisher is found to be the most important feature for this category, and while it helps, it does not perfectly capture the notion of consensus. When we combine publisher and tweet based features, there is improvement in performance.

Next, we examine the performance of our proposed audience leaning based features and find it to perform the best amongst the three categories of features. Digging deeper, we find that we correctly classified 74% of high consensus tweets and 70% of low consensus tweets. We find χ^2 distance between

[6] See http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html.

Category	Features
Followers	# Dem/Rep/Neu, Sum/Avg/Median/Skew of PL Sum(PL) of Dem/Rep/Neu, PLD
Retweeters	# Dem/Rep/Neu, Sum/Avg/Med/Skew of PL Sum(PL) of Dem/Rep/Neu, PLD of baseline Avg #Dem/Rep/Neu in baseline, PLD χ^2 Distance bw PLD[Retweeters] of tweet & baseline
Repliers	# Dem/Rep/Neu, Sum/Avg/Med/Skew of PL Sum(PL) of Dem/Rep/Neu, PLD of baseline Avg #Dem/Rep/Neu in baseline, PLD χ^2 Distance bw PLD[Repliers] of tweet & baseline
Combination	χ^2 Distance bw PLD[Repliers] and PLD[Retweeters] χ^2 Distance bw PLD[Repliers] and PLD[Followers] χ^2 Distance bw PLD[Retweeters] and PLD[Followers]

Table 3.5: Audience leaning based features. In the table, Dem, Rep, and Neu denote Democrat, Republican, and Neutral respectively, PL denotes political leaning, and PLD denotes the distribution of political leanings. Baselines are computed by taking average of PLD across all tweets. Most important features are highlighted in blue.

the repliers' and retweeters' leaning distribution to be the most important feature, matching the intuition we built earlier in the paper. In fact, even when we combine the three categories of features, we do not find a performance gain over using the audience leaning based features alone. This is because when we inspect the 10 most important features out of all the categories, the top 7 most important features (highlighted in Table 3.5 in blue) are from our proposed category of audience leaning based features, highlighting how well suited they are for our consensus identification task.

3.3 Promoting non-divisive news selectively to reach a divers audience

Highlighting high consensus news that elicits similar responses from both sides could act as a soothing balm to help bring people together, despite initial ideological differences. As we discussed we proposed such a complementary approach to increase diversity in users' information consumption by identifying high consensus, yet interesting information. Our system recommends high consensus "purple" posts to both red (conservatives) and blue (liberals) users, hoping to increase users' exposure to cross-cutting news posts, leading to lower societal polarization and lower *segregation* in information consumption [Chakraborty et al., 2017]. Nevertheless, it still remains unclear how such information is spread across users in a network and how individuals choose to react to it. Here, we investigate users' willingness to share and spread such posts, or the reach of high and low consensus news stories across a diverse audience. We also examine the newsworthiness [Galtung and Ruge, 1965, Weber, 2014] of both high and low consensus news. Overall, we ask two fundamental questions on a Twitter network: (1) Can high consensus posts help to break filter bubbles (and thus potentially decrease polarization in society)? (2) Can we propose methods to propagate high consensus stories broadly across society? We highlight the following contributions:

I. We compile a novel dataset, which reveals how Twitter users with similar or different political leanings are connected to each other. To do so, we consider a dataset of 400 news tweets posted by 10 publishers containing 80 high and 80 low consensus posts [Babaei et al., 2018]. For every high or low consensus news post, we collected a subset of its 100 random retweeters and for each retweeter we collected a random set of their 100 followers. We compute the political leaning of the 1,616,000 users who either retweeted a high or low consensus news story or were exposed to it. Moreover, to simulate the spread of news in Twitter, we crawl a network of more than 100 million Twitter users. This allows us to compute the political leanings of 69,687 users connected by 2,907,026 links.

II. Using our dataset, we study how individuals with different political leanings get exposed to and retweet high and low consensus news posted by users from various political perspectives. We observe that low consensus news tends to proliferate primarily only amongst users with a particular political learning. In contrast, high consensus news has a higher chance of spreading through the entire network. Importantly, high consensus news posted by a set of neutral publishers spreads more equally across liberal and conservative users than if posted by the same number of a mix of non-neutral publishers.

III. Based on the above observations, we propose a strategy that seeds neutral publishers to expose roughly equal fractions of people with different political leaning to high consensus news with the minimum cost (hoping this may help to break filter bubbles which can trap users). We show that our proposed strategy is more effective than seeding the most influential nodes without taking the political leanings into account.

Our work provides new insights and a complementary tool which may help to reduce filter bubbles, encourage healthier interaction between population subgroups, and lead to a more cohesive society.

3.4 Dataset

Here, we consider the dataset of 400 news tweets posted by 10 publishers that we discussed earlier. The dataset contains 80 low and 80 high consensus news posts.

To obtain the political leanings of the users who either retweeted a high or low consensus news or were exposed to it, for every high or low consensus news post in the dataset, we collected a random set of its 100 retweeters. Then for each retweeter, we collected a random set of his 100 followers. Finally, for each of these 1,616,000 users we collected their followees to compute their political leaning.

As we discuss later, low consensus posts have a much smaller chance of being received by users with different political leanings, which leads to a more politically fragmented society. On the other hand, high consensus posts have a better chance of spreading through communities with various political leanings, and can be utilized to break the filter bubbles.

3.4.1 Collecting users' political leanings

For every news in our set of 80 high and 80 low consensus news posts, we collected a random set of its 100 retweeters. Then for each retweeter, we collected a random set of its 100 followers. Thus we have 1,616,000 twitter users.

We then inferred every user's political leaning, as a score between -1, +1, using the method of [Kulshrestha et al., 2017], where we discussed earlier 3.1.2.

For retweeters with certain political leaning, we calculated the expected fraction of their liberal, conservative, and neutral followers, as is shown in Table 3.6.

	Liberal	**Conservative**	**Neutral**
Liberal	0.76	0.04	0.2
Conservative	0.045	0.85	0.1
Neutral	0.3	0.27	0.43

Table 3.6: Expected fraction of liberal, conservative, and neutral followers of retweeters with various political leanings. Rows and columns correspond to retweeters and followers.

We also estimated the conditional probability that users with different political leanings retweet high consensus and low consensus news post from liberal, conservative, and neutral publishers (given that they retweet) in Table 3.7. It can be seen that users with a certain political leaning retweet low consensus posts from the publishers with the same political leaning with a very high probability.

Interestingly, users retweet high consensus news posts from the publishers with the same political leaning with a smaller probability. On the other hand, there is a very small chance that users with a certain political leaning retweet low consensus posts from the publishers with different political leanings. For high consensus news this probability is larger.

		Retweeters		
		Liberal	**Conservative**	**Neutral**
Publishers	**Liberal**	H: 0.65 L: 0.85	H: 0.08 L: 0.04	H: 0.27 L: 0.11
	Conservative	H: 0.12 L: 0.08	H: 0.68 L: 0.85	H: 0.2 L: 0.07
	Neutral	H: 0.34 L: 0.38	H: 0.33 L: 0.37	H: 0.33 L: 0.25

Table 3.7: Conditional probability of retweeting a high and low consensus news post (indicated H and L in the table) by users from various political leanings (given that they retweet). Rows and columns correspond to publishers and retweeters. For instance, in the first cell (first row and column), the probability that liberal users retweet high/low consensus news posts published by liberals publishers is 0.65/0.85.

3.5 The Gap between proliferation of high and low consensus news

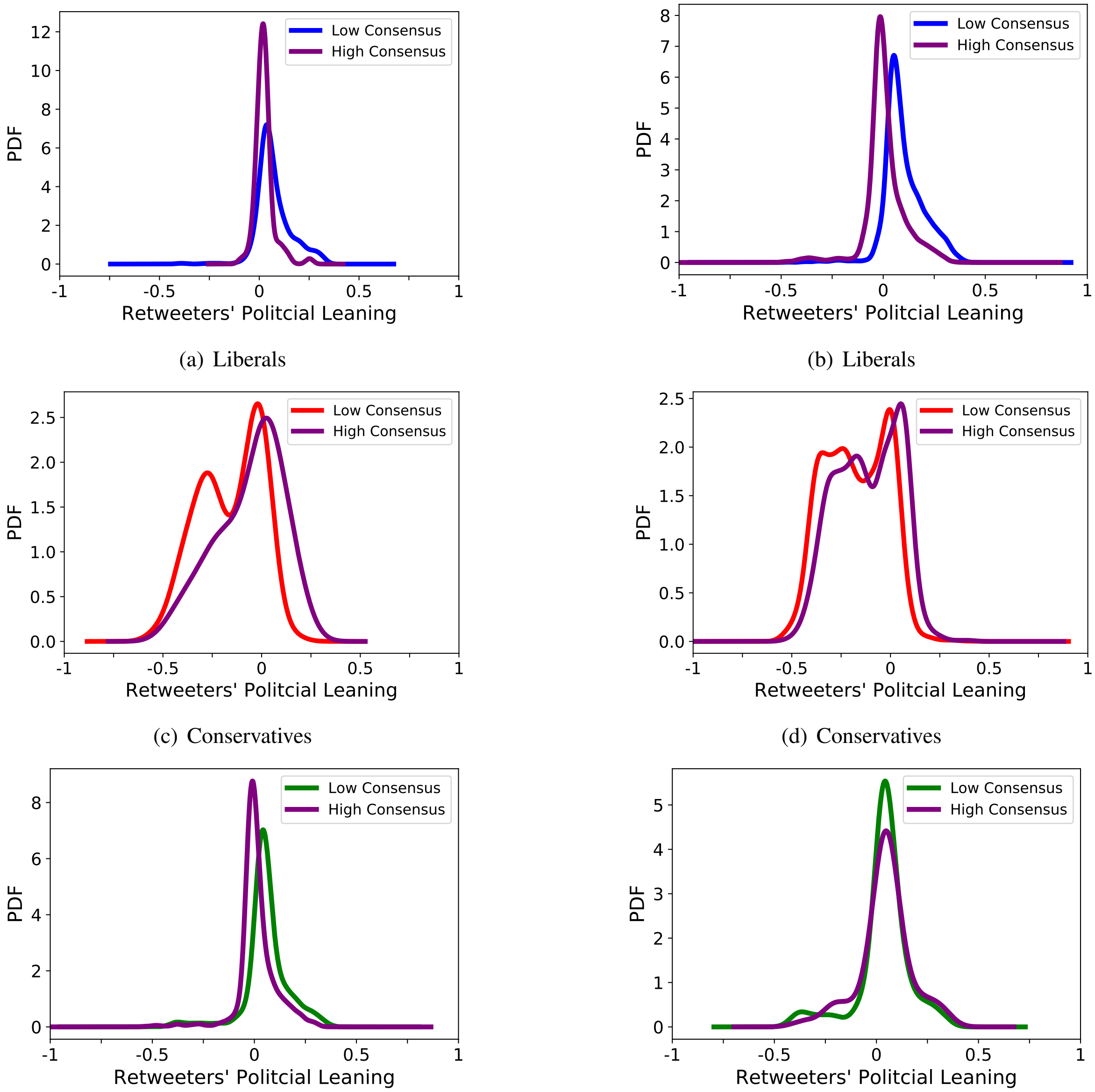
In this section, we investigate how high and low consensus news posts spread among users with various political leanings in Twitter. In particular, our goal is to answer the following key questions:

How do individuals with certain political leaning (liberal, conservative, and neutral) get exposed to high and low consensus news posts?

Studying the above key question allows us to understand the gap between proliferation of high and low consensus news, and develop strategies to decrease the polarization in the society by breaking the filter bubbles that trap users. We start by investigating users' behavior in retweeting high and low consensus news posts. Then, we discuss how the confirmation bias in retweeting behavior makes the filter bubbles grow larger and promote social polarization.

3.5.1 Confirmation bias in retweeting behavior

First, we study how users with different political leaning *share* high and low consensus news post. Specifically, we compare how users with different political leanings retweet low and high consensus news posts from publishers with different political perspectives.



(a) Liberals

(b) Liberals

(c) Conservatives

(d) Conservatives

(e) Neutrals

(f) Neutrals

Figure 3.6: Distribution of retweeters' political leanings for low and high consensus news posted by publishers with different political perspectives. Distribution of political leanings for a random high and a random low consensus news posted by (a) liberal, (c) conservative, and (e) neutral publishers. Distribution of average political leanings for 100 high and low consensus news posted by (b) liberal, (d) conservative, and (f) neutral publishers. Distribution of political leanings for high consensus news (purple) is more symmetric and centered around 0.

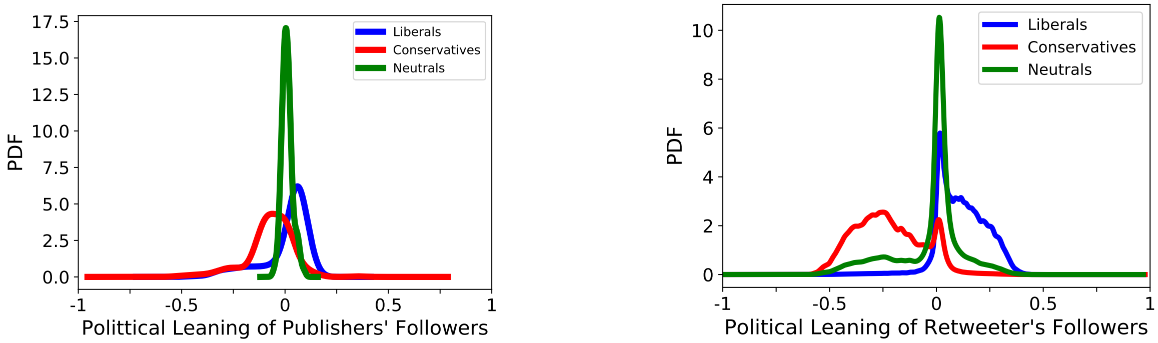
Figures 3.6(a), 3.6(c), 3.6(e) show the distribution of the political leanings of all retweeters for one random low consensus and one random high consensus news posted by CNN (liberal publisher), FoxNews (conservative publisher), and Reuters (neutral publisher)[7]. Notice that the distribution of retweeters' political leanings in Figure 3.6(a), 3.6(e) has more density in the right (liberal leaning) for the low consensus news. On the other hand, the distribution of retweeters' political leanings in Figure 3.6(c) has considerably more density in the left (conservative leaning) for the low consensus news. Importantly, the distribution of retweeters' political leanings for high consensus news (purple curve) is more symmetric in all the Figures. Moreover, the mean of the distribution for high consensus news is close to 0.

Next, we consider 80 low consensus and 80 high consensus news posted by the 10 publishers, ranging from liberals to neutrals to conservatives: Slate, Salon, New York Times, CNN, AP, Reuters, Politico, Fox News, Drudge Report, and Breitbart News. For each news post, we consider a set of its 100 retweeters chosen at random. Figures 3.6(b), 3.6(d), 3.6(f) show the distribution of the *expected* political leanings of retweeters of all the low consensus and high consensus news posted by liberal, conservative, and neutral publishers, respectively. Again, the distribution of retweeters' political leanings for high consensus news (purple curve) is more symmetric, and is centered around 0 in all the Figures. In particular, the distribution of retweeters' political leanings for high consensus news posted by neutral publishers has the most symmetric shape around 0.

We summarize our key observations as follows:

- I. Low consensus news posted by publishers with a specific political leaning (liberal/conservative) are mostly retweeted by users with similar political leanings (Figures 3.6(b), 3.6(d)).
- II. High consensus news posted by publishers with a specific political leaning (liberal/conservative) are retweeted by users with various political leanings (liberal/conservative/neutral) (Figures 3.6(b), 3.6(d)).
- III. While low and high consensus news posted by neutral publishers spread with lower disparity among users with different political leanings, high consensus news posted by neutral publishers have the highest probability to be spread with minimum disparity among users (Figure 3.6(f)).

[7] The PDFs have been empirically estimated using kernel density estimation [Bowman and Azzalini, 2004]



(a) Publisher's followers

(b) Retweeter's followers

Figure 3.7: Distribution of political leanings for (a) followers of liberal, conservative, and neutral publishers, and (b) followers of retweeters of liberal, conservative, and neutral publishers. As we get farther away from the publishers, the distribution of liberal and conservative followers becomes significantly more skewed (filter bubbles grow larger).

3.5.2 The growth of filter bubbles in twitter

Next, we investigate how individuals with different political leanings get *exposed* to high and low consensus news posted by liberal, conservative, and neutral publishers.

Figure 3.7(a) depicts the distribution of political leanings for followers (level 1) of liberal, conservative, and neutral publishers. We observe that users with conservative or liberal leanings are mostly exposed to news posted by publishers with the same political leaning (the bubble effect). Therefore, the distribution of political leaning for followers of liberal and conservative publishers are skewed to the left and right, respectively. Nevertheless, followers of conservative publishers has a more skewed distribution. This is resulted from the fact that the conservative community is denser, and has fewer connections to liberals and neutrals in Twitter (*c.f.* Table 3.6). On the other hand, the distribution of political leanings for neutral users is very symmetric and is centered at 0. Hence, neutral users get similar exposure to liberal and conservative view points.

Figure 3.7(b) shows the distribution of political leanings for followers of retweeters (level 2) of liberal, conservative, and neutral publishers. We observe that while the distribution of political leanings for followers of retweeters of neutral publishers is symmetric and centered around 0, the distribution of political leanings for followers of retweeters of liberal or conservative publishers are extremely skewed. As expected, followers of retweeters of conservative publishers have a more skewed distribution. Interestingly, the skewness of the distributions for followers of retweeters (level 2) is much larger compared to the skewness of distributions for followers of publishers (level 1). This means that filter bubbles in level 2 are larger than those in level 1. Our experiments show that as we get farther away from the publishers, filter bubbles grow even larger (Figure 3.12).

We summarize our key observations as follows:

- I. Conservatives and liberals get a biased exposure to the news posted in Twitter, while neutrals get similar exposure to liberal and conservative view points (Figure 3.7(a)).
- II. Users who get the news from retweeters get a significantly more biased exposure, compare to users who get the news from publishers. In other words, as we get farther away from the news publishers, the filter bubbles grow larger (Figure 3.7(b)).

3.5.3 Breaking filter bubbles

To break filter bubbles, we aim for all individuals to get similar exposure to news stories. Our proposed strategy to break the bubble effect is based on the three key observations discussed earlier: (1) while low consensus news are more likely to proliferate amongst the users with a particular political leaning, high consensus news has a much higher chance of spreading among users with different political leanings; (2) high consensus news posted by neutral publishers has the lowest disparity for spreading among liberal and conservative users; and (3) as users get farther away from publishers, they get a more biased exposure to news. Based on the above observations, we conjecture:

High consensus news posted by neutral users help break the filter bubbles.

High consensus news posted by neutral users achieve high spread with little disparity regarding political leaning. We confirm our conjecture and show the effectiveness of our proposed strategy through an extensive set of experiments later in the paper.

In the following section, we first formulate the problem of information diffusion in social networks. Then, we discuss the problem of finding a near-optimal set of neutral users to seed spreading high consensus news and break the filter bubbles.

3.6 Problem formulation: Information diffusion

We start by formulating the information diffusion problem to model the spread of news among individuals with various political leanings in Twitter. We simulate the proliferation of news by assuming that each user can be a publisher. Then we select a set of users and involve them to post news. We represent Twitter by a directed graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of directed edges between the nodes. The nodes in the network are partitioned into three disjoint groups $\mathcal{V} = \{\mathcal{V}_d, \mathcal{V}_r, \mathcal{V}_n\}$, where $\mathcal{V}_d, \mathcal{V}_r, \mathcal{V}_n$ represent users with liberal, conservative, and neutral leanings, respectively. A directed edge (v, u) exists if user v follows user u. When users post tweets, their followers can retweet and spread the tweets in the network. To model spread of information, e.g. news or tweets in Twitter, two well-known classical diffusion models are introduced in the literature [Pastor-Satorras and Vespignani, 2001]: (1) Independent Cascade model (IC) and (2) Linear Threshold (LT) model. In this work, we consider the IC model (section 3.1.5).

3.6.1 Information diffusion with low disparity

Our goal is to find the smallest seed set of users that when post a tweet, it spreads through at least a fraction $Q_p \in [0, 1]$ of liberals ($\mathcal{V}_l$), conservatives ($\mathcal{V}_c$), and neutrals ($\mathcal{V}_n$) in Twitter, where $p \in \{l, c, n\}$. We formulate the problem as follows:

$$\min_{S \subseteq \mathcal{V}} |S| \quad \text{subject to} \tag{3.3}$$

$$\sum_{p \in \{l,c,n\}} \min\left(f_p(S),\ Q_p \cdot |\mathcal{V}_p|\right) \geq \sum_{p \in \{l,c,n\}} Q_p,$$

where $f_l(.), f_c(.), f_n(.)$ determine the total number users among liberals ($\mathcal{V}_l$), conservatives ($\mathcal{V}_c$), and neutrals ($\mathcal{V}_n$) that are activated as a result of selecting the seed set S. We call $\mathcal{V}_p$ “saturated" by S when $\min\left(f_p(S),\ Q_p \cdot |\mathcal{V}_p|\right) = Q_p \cdot |\mathcal{V}_p|$. When a certain fraction Q_p of individuals with a particular political leaning p are exposed to a news (activated), any new activated individual with political leaning p cannot further improve the utility. This will give individuals with different political leanings a higher chance of being exposed to the news.

We note that the utility function, i.e., $f_p : 2^{\mathcal{V}_p} \rightarrow \mathcal{Z}^+$, is a non-negative, monotone, submodular set function [Kempe et al., 2003]. The submodularity is an intuitive notion of diminishing returns, stating that for any sets $A \subseteq A' \subseteq V$ and any node $a \in V \setminus A'$, it holds that:

$$f(A \cup \{a\}) - f(A) \geq f(A' \cup \{a\}) - f(A').$$

Although problem (3.3) is NP-hard in general [Wolsey, 1982], for maximizing a submodular function the following greedy algorithm provides a logarithmic approximation guarantee. The greedy algorithm starts from an empty set, add a new node to the set which provides the maximal marginal gain in terms of utility, and stops whenever the desired Q_p fraction of individuals with political leaning p are activated.

3.6.2 Spreading through neutrals to break the bubbles

In Problem (3.3), the fraction Q_p can be arbitrary for individuals with different political leanings. However, to break the filter bubbles we wish individuals with different political leanings to get a similar exposure to various news. In other works, we assume similar values for Q_l, Q_c, Q_n. Moreover, the news posted by individuals with neutral leanings have a higher chance of spreading among individuals with liberal and conservative political leanings. Therefore, to break the filter bubbles we aim at finding the smallest subset $S \subseteq \mathcal{V}_n$ that when post a news, at least a fraction Q_p of individuals with political leaning p get exposed to the news. Formally, we have

$$\begin{aligned} &\min_{S \subseteq \mathcal{V}_n} |S| \quad \text{subject to} \\ &\sum_{p \in \{l,c,n\}} \min\left(f_p(S),\ Q_p \cdot |\mathcal{V}_p|\right) \geq \sum_{p \in \{l,c,n\}} Q_p. \end{aligned} \tag{3.4}$$

3.7 Experimental results

In this section we investigate the effect of spreading high consensus news posted by neutral users among individuals with conservative, liberal, and neutral leanings in Twitter. In particular, we show that our proposed strategy is very effective in spreading information among individuals with various political

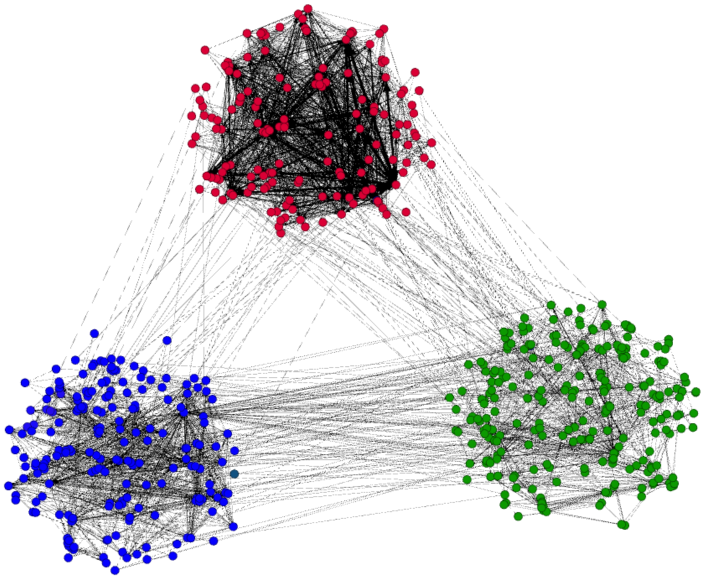
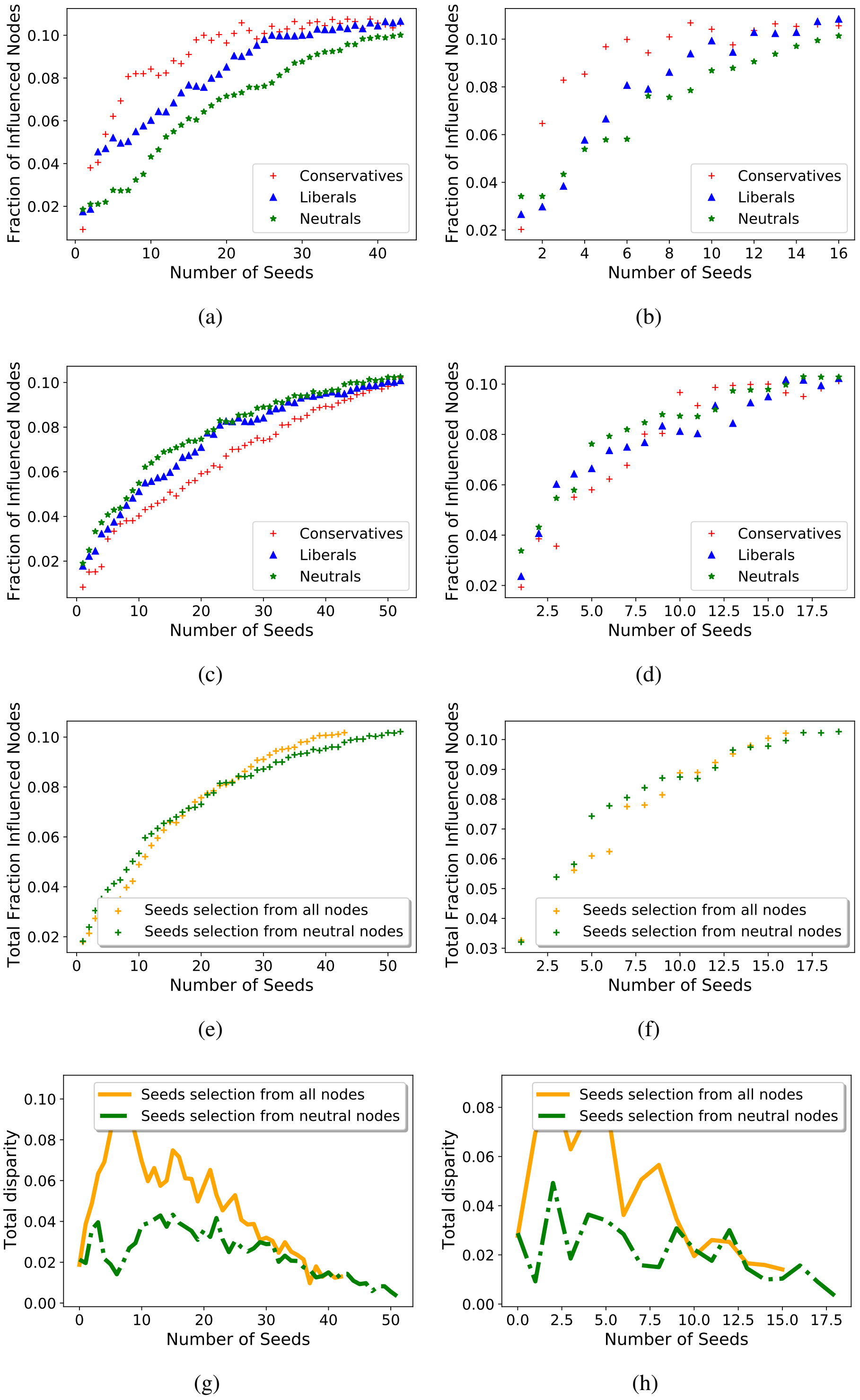


Figure 3.8: The sample graph from the real Twitter data set collected in 2009. Blue, red, and green nodes indicate users with liberal, conservative, and neutral political leanings.

leanings and lowering societal polarization for news consumption. We first describe our instance of Twitter network. We then explain our experimental setup, and present our findings.

Twitter network. Our network is collected from Twitter in September 2009 [Babaei et al., 2016, Cha et al., 2010], and includes: 52 million user profiles, 1.9 billion directed follow links among the users, and 1.7 billion public tweets posted by the users. In order to obtain a static network, we consider the tweets published on July 1, 2009, and filter out users that did not tweet before July 1. After this filtering, we have 70,000 active users. We then extract the strongest connected community, including 69,687 users and 2,907,026 link between them, yielding 19162, 3449, 47076 nodes with liberal, conservative, and neutral leanings, respectively. The average degree of network is 41.5. Figure 3.8 shows an induced random sample from our final Twitter network.

Sampled Twitter network. We also created a smaller network by sampling 10% of nodes uniformly at random from our original Twitter network, and connecting the users if they have a connection in the original network. The strongest connected community includes 3,753 users and 6,993 connections with average degree of 1.83. Our sampled Twitter network includes 812 liberals, 186 conservatives, and 2,755 neutrals. Note that the structure of the original Twitter network is very different than the sampled Twitter network. In particular, the sampled Twitter network is significantly sparser than the original Twitter network.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Figure 3.9: Fraction of individuals with liberal, conservative, and neutral political leanings who are exposed to a high consensus news. Left column shows the result on our Twitter network, and the right column shows the result on the smaller sampled Twitter network. (a), (b) show the fraction of exposed individuals when the seeds are selected from the entire network by solving Problem (3.3). (c), (d) show the fraction of exposed individuals when the seeds are selected from neutral users by solving Problem (3.4). (e), (f) compare the fraction of exposed individuals when the initial seed is selected from the entire network vs. neutrals. (g), (b) compare the disparity of diffusion when the initial seed set is selected from the entire network vs. neutrals.

Experimental setup. For a pair of users $u \in \mathcal{V}_i$ and $w \in \mathcal{V}_j$, we calculate the success probability of activation p_{uw} as the expected fraction of users with political leaning j who retweeted news posted by users with political leaning i. The retweeting probabilities are listed in Table 3.7.

We apply the greedy algorithm to find a near optimal subset of users that can spread a news over a certain fraction $Q_l = Q_c = Q_n = 0.1$ of liberals ($\mathcal{V}_l$), conservatives ($\mathcal{V}_c$), and neutrals ($\mathcal{V}_n$) in the Twitter network. To evaluate the utility function $f_p(.)$ in Problem 3.3 and Problem 3.4, we estimate it by using Monte Carlo sampling [Hastings, 1970]. We used 200 samples for this estimation, which yielded a stable estimation of the utility function.

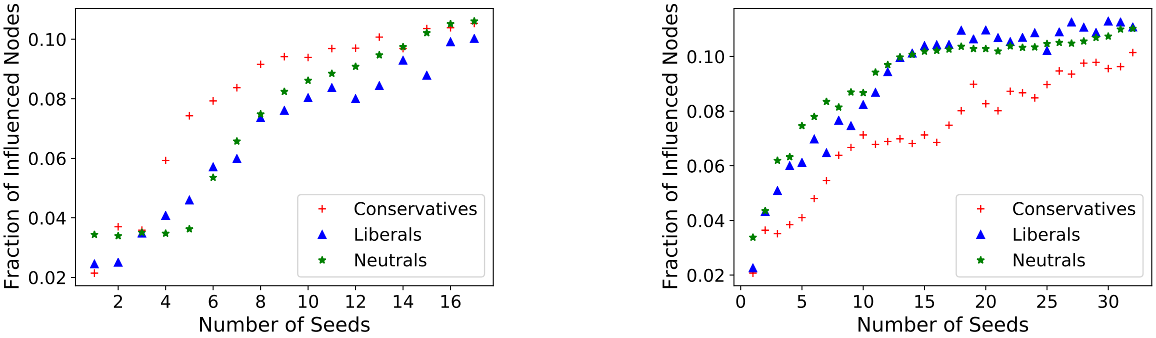
Note that using equal values for Q_l, Q_c, Q_n in Problem (3.3), we retrieve the fair influence maximization formulation proposed by [Ali et al., 2019a]. In our experiments, we compare our proposed strategy to fair influence maximization.

3.7.1 Neutrals can break filter bubbles

In our first set of experiments, we apply the greedy algorithm to Problem (3.3) and Problem (3.4) to find the initial set of users to spread news in Twitter. Figure 3.9 compares the fraction of individuals with liberal, conservative, and neutral political leanings who got exposed to a high consensus news spread through an initial seed set obtained by solving Problem (3.3) vs. Problem (3.4). The goal is to expose $Q_p = 10\%$ of individuals with liberal, conservative, and neutral leanings to the news. The top row shows the result on our original Twitter network, and the bottom row shows the result on the smaller sampled Twitter network. Note that the sampled network is much sparser than the original Twitter network.

Figures 3.9(a), 3.9(b) show the fraction of exposed individuals when the seeds are selected from the entire network by solving Problem (3.3). Figures 3.9(c), 3.9(d) show the fraction of exposed individuals when the seeds are selected from the users with neutral leanings by solving Problem (3.4). We note that as more individuals are added to the initial seed set by the greedy algorithm, the disparity in the number of exposed users with different political leanings is much smaller in Figures 3.9(c), 3.9(d) compared to Figures 3.9(a), 3.9(b). This clearly confirms the effectiveness of our proposed strategy in breaking the filter bubbles.

We note that if we do not take into account the different pattern of diffusion among users of various political leanings, the neutral users may not be the ones that can maximize the spread of information.



(a) Seeds with various leanings

(b) Seeds with neutral leanings

Figure 3.10: Fraction of users who were exposed to a high-consensus news story, when all users propagate the news with the same probability of 0.1, irrespective of their political leanings in Twitter. (a) shows the result when seeds are selected from the entire network, (b) shows the result when seed are selected only from the neutral users.

Figure 3.10(a), 3.10(b) compare the fraction of users who were exposed to a high-consensus news, when the seeds are selected from the entire network vs. only the neural users. Here, we assume that all the users spread the news with the same probability of 0.1 irrespective of their political leanings. We see that the news posted by neutral publishers have more disparity and reaches a smaller number of users.

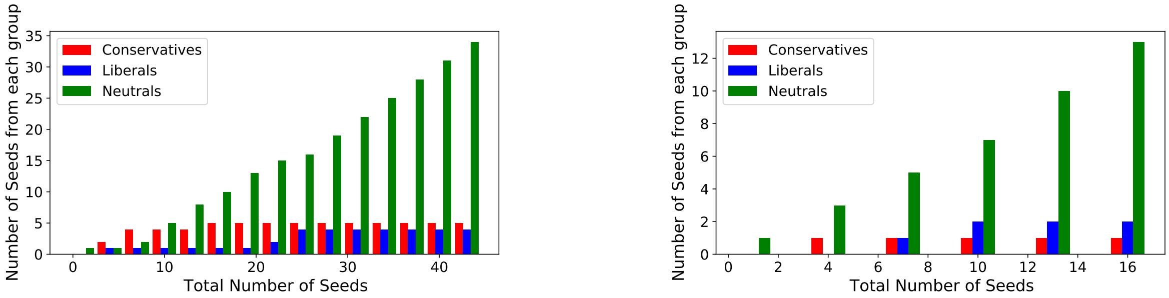
3.7.2 Neutrals can widely spread the news

Figures 3.9(e), 3.9(f) compare the fraction of exposed individuals when the initial seed is selected from the entire network vs. neutrals. There are two interesting observations: the initial seeds selected from neutral users can spread the news even more than users selected from the entire network. Moreover, as we continue the selection process, selected neutral seeds can spread the news as well as the seed set selected from the entire network. This interesting observation confirms the power of neutral users in spreading news in Twitter. As Table 3.8 depicts, the average number of retweeting of a tweet posted by liberal, conservative, and neutrals are almost equal. There is a interesting observation. High consensus tweets posted by neutrals are retweeted with many democrats and republicans in addition to neutrals. Interestingly, news posted by neutrals is retweeted by an even larger number of users compared to news posted by liberal or conservative publishers.

Figure 3.11(a) shows the number of users selected with liberal, conservative, and neutral leanings for varying number of seeds selected greedily to solve Problem (3.3). Figure 3.11(a) shows the result on the Twitter network, and Figure 3.11(b) shows the result on the sampled Twitter network. We see

		Retweeters			
		Liberal	**Conservative**	**Neutral**	**Sum**
Publishers	**Liberal**	H: 76 L: 104	H: 9 L: 5	H: 32 L: 15	H:117 L:124
	Conservative	H: 9 L: 10	H: 58 L: 94	H: 18 L: 9	H:87 L:113
	Neutral	H: 45 L: 49	H: 43 L: 47	H: 43 L: 32	H:131 L:128

Table 3.8: Average number of retweeting a high and low consensus news post (indicated H and L in the table) by users from various political leanings. Rows and columns correspond to publishers and retweeters. For instance, in the first cell (first row and column), liberal users on average retweets high/low consensus news posts published by liberals publishers 76 times. We note that news posted by neutrals is retweeted by an even larger number of users, compared to news posted by liberal or conservative publishers.



(a) Twitter network

(b) Sampled Twitter network

Figure 3.11: Number of users selected with liberal, conservative, and neutral leanings for varying number of seeds selected greedily to solve Problem (3.3). Figure shows the results on (a) the Twitter network, and (b) the smaller sampled Twitter network.

that in the set of seeds greedily selected from the entire network, the majority of the users have neutral leanings. This further shows that neutral users are highly effective in spreading information in Twitter. This is consistent with our initial observation, that the news posted by neutrals has a higher probability of spreading among users with different political leanings.

3.7.3 Neutrals spread news with low disparity

Figures 3.9(g), 3.9(h) compare the total disparity of diffusion when the initial seed set is selected from the entire network vs. neutrals. We define the total disparity as the sum of all disparity (differences) between exposure for each pair of political leanings. Formally, we have: Total disparity=

$$\left|\frac{f_l(S)}{\mathcal{V}_l} - \frac{f_c(S)}{\mathcal{V}_c}\right| + \left|\frac{f_l(S)}{\mathcal{V}_l} - \frac{f_n(S)}{\mathcal{V}_n}\right| + \left|\frac{f_c(S)}{\mathcal{V}_c} - \frac{f_n(S)}{\mathcal{V}_n}\right|.$$

We observe that the total disparity is much smaller when the initial seed set is selected from users with neutral political leanings (Problem (3.3)) compared to the case when the initial seed set is selected from the entire network (Problem (3.4)). The difference is larger when the size of the initial seed set is smaller.

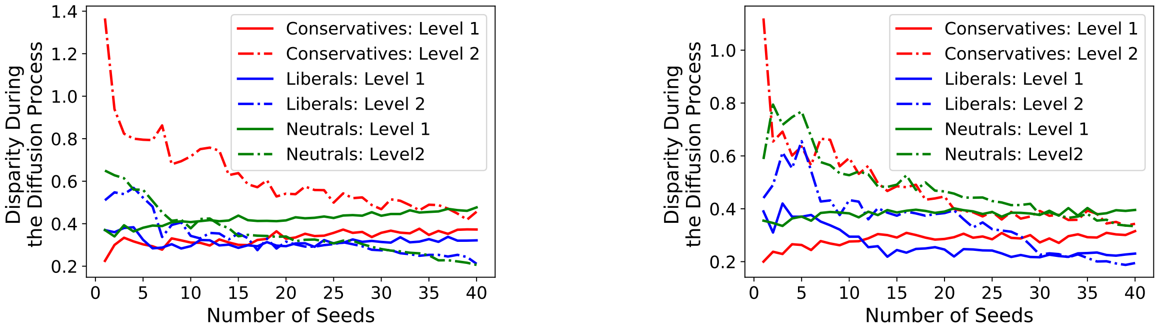
3.7.4 Filter bubbles grow larger over time

Figure 3.12 shows the fraction of individuals with various political leanings who got exposed to a high consensus news during the diffusion process (IC), for varying number of seeds. More precisely, for a given seed set information diffusion proceeds in discrete time steps $t = \{0, 1, 2, \dots, \}$. Figure 3.12 compares the fraction of users with various political leanings who received the news in the first time-step, $t = 1$, and second time-step $t = 2$ in our original Twitter network. Figure 3.12(a) shows the result when the seeds are selected from the entire network, by solving Problem (3.3). Figure 3.12(b) shows the result when the seeds are selected from the users with neutral political leanings, by solving Problem (3.4). It can be seen that when seeds are selected from the entire network, the disparity becomes larger as the diffusion process continues. On the other hand, the disparity is much smaller when seeds are selected from neutral users.

The above result confirms our observation that the filter bubbles grow larger as the diffusion continues over time. In other words, when the seeds are selected from the entire network, as the we get farther away from the initial set of seeds, the disparity in the number of users with different political leanings who are exposed to the news becomes larger. On the other hand, when diffusion is originated from neutral seeds, users with different political leanings get exposed to the information at the same time. This is crucial while spreading time-critical information, such as health-related information or emergency warnings, in the network.

3.8 Discussion

Since we propose increasing exposure to high consensus news, we would like to check that such stories carry important information for public discourse. Babaei et al. [2018] compared low and high consensus posts on social media by empirically analyzing their properties. They showed that both types of posts are equally popular and cover similar topics. We checked this by analyzing 400 randomly selected posts



(a)

(b)

Figure 3.12: Fraction of users who are exposed to the news from different groups in first and second time step of propagation process in the Twitter network. (a) Shows the result when seed are selected from the entire network (Problem 3.3), (b) shows the result when seed are selected from neutral users (Problem 3.4).

including examples of high and low consensus news, along with their sources, number of retweets, replies, and likes. See Table 3.9 for details. We highlight the following observations:

I. For both types of news posts, a variety of news sources exists across the ideological spectrum. Figure 3.2 also shows several publishers with different political leaning that posts both types of high and low consensus news.

II. On average, high and low consensus tweets are retweeted 158 and 177 times respectively. On average high consensus tweets are liked 532 times, whereas, low consensus news are liked 488 times. Thus, high and low consensus news stories have similar popularity.

III. Galtung and Ruge [1965] introduce newsworthiness theory in which they propose several news factors such as frequency, meaningfulness, continuity, etc. Eilders [2006] showed that these factors impact news' worthiness. Weber [2014] proposes the following hypothesis: "The news factors of a news item influence the level of participation in commenting in an article's comments section". Weber also noted several other factors, such as having a high social impact or being controversial, that may attract more comments as participation [Weber, 2012]. Weber emphasized that if a news story attracts more comments, then it has higher worthiness. Here we can consider the number of replies as participating comments. On average, high consensus and low consensus news stories received 100 and 114 replies (comments), respectively, suggesting that both types of news have similar worthiness.

In summary, we observe that high and low consensus news are similar along multiple dimensions, including variety of news source, popularity, topic covering, and worthiness.

3.9 Summary

To summarize, we propse an approach to inject diversity in users' consumption by defining and operationalizing the concept of consensus of news posts in terms of general agreement in readers' reaction, irrespective of their own political leanings. We then use human judgments to generate a ground truth dataset of non-divisinve (high consensus) and divisive (low consensus) news posts on social media, and observe that a substantial amount of high consensus purple posts are posted by news publishers on social media (perhaps surprisingly, even by politically extreme publishers).

We also find that both types of tweets are equally popular with users (*i.e.,* garner similar number of retweets) and also cover similar topics. Further, we observe that high consensus purple posts tend to provide more cross-cutting exposure to views than low consensus posts. To identify high consensus purple news posts automatically, we propose a novel class of features of social media posts on Twitter, which we term *audience leaning based features*. These features describe the distribution of the political leanings of audience subgroups interacting with a post – namely the retweeters and repliers of a post. Intuitively, retweeters are more likely to be supportive of it, while repliers have a higher likelihood of opposing it. Additionally, the followers of the publisher of the post also form a passive audience subgroup for the post. We use these audience leanings as features to capture the degree of consensus that a social media post is likely to have. We present an evaluation showing that our proposed features are well suited to help identify high and low consensus tweets automatically with high accuracy, leading to significantly better performance than can be achieved using previously proposed publisher based and content based features.

We then propose and quantify the benefits of a potential strategy to *spread* high consensus news across readers with diverse political leanings. We first compile a dataset and make the following three key observations: (1) low consensus news is more likely to remain within subgroups of users with similar political leanings, whereas high consensus news spreads more across subgroups; (2) high consensus news posted by neutral publishers spreads more equally across subgroups; and (3) users that get the information from other users instead of the publishers, get an even more biased exposure to

High Consensus News	Low Consensus News
BREAKING: Senate intelligence committee invites fired FBI Director Comey to appear in closed session next Tuesday . Source: AP, 2.7K Retweets, 176 Replies, 5k Likes	Report: Trump "revealed more information to the Russian ambassador than we have shared with our own allies" Source: Salon, 51 Retweets, 14 Replies, 34 Likes.
@johnrobertsFox on firing of James Comey: "This came as a shock to literally everyone, including the @FBI Director." #TheFive Source: Fox News, 156 Retweets, 259 Replies, 574 Likes.	Orrin Hatch makes clear the conservative case against Obamacare: Once the public "is on the dole, they'll take eve.. Source: Salon, 113 Retweets, 12 Replies, 10 Likes.
White House calls emergency meetings as global cyberattack spreads http://politi.co/2qgNnW1 Source: Politico, 80 Retweets, 31 Replies, 72 Likes.	Why are Republicans attacking the Census Bureau? Because they don't want an accurate count of Americans Source: Salon, 691 Retweets, 22 Replies, 680 Likes.
The Latest: US says Russia should be worried about N. Korea missile launch; Japan, US, South Korea discuss threat. http://apne.ws/2r5iNQ1 Source: AP, 167 Retweets, 12 Replies, 93 Likes.	Is Donald Trump the "little boy President"? A @CNNOpinion contributor takes a closer look at his latest moves http://cnn.it/2pE1Uaq Source: CNN, 179 Retwees, 264 Replies, 468 Likes.
White House wants the FBI to complete its investigation into Russia interference in the 2016 election. http://apne.ws/2qso0kS Source: AP, 179 Retweets, 68 Replies, 97 Likes.	@SarahHuckabee: "@POTUS over the last several months lost confidence in Director Comey. The DOJ lost confidence in Director Comey." Source: Fox News, 122 Retweets, 94 Replies, 593 Likes.
Donald Trump's lawyers say he doesn't have any Russian money "with a few exceptions" http://dlvr.it/P7QvTv Source: Salon, 125 Retweets, 10 Replies, 18 Likes.	Schieffer Slams Trump: Comey Firing Reminds Me of JFK-Oswald Conspiracies Source: Fox News, 55 Retweets, 510 Replies, 149 Likes.
North Korea's Sunday missile test is what one researcher is calling an "extended middle finger to Trump" Source: CNN, 212 Retweets, 63 Replies, 326 Likes.	Sarah Huckabee Sanders went on Fox last night and, wait for it, said it's "time to move on" from Russia probe: http://slate.me/2qsCqBO Source: Slate, 28 Retweets, 35 Replies, 48 Likes.
"He knows the last three days have not been good for him": Sean Spicer's make-or-break briefing http://politi.co/2r1t8MQ Source: Politico, 59 Retweets, 37 Replies, 112 Likes.	"A fresh start will serve the FBI": Republicans provide cover for Donald Trump http://ift.tt/2q6C1BM Source: Salon, 107 Retweets, 61 Replies, 88 Likes.
San Diego police: Teen shot and killed left suicide note http://fxn.ws/2ps2UPO #FOXNewsUS Source: Fox News, 52 Retweets, 21 Replies, 84 Likes.	Acting FBI Director contradicts White House claim that fired director James Comey had lost support. http://apne.ws/2r5GJjr Source: AP, 357 Retweets, 56 Replies, 497 Likes.
@HillaryClinton launches Onward Together PAC. Read more: http://fxn.ws/2qlcwz0 Source: Fox News, 144 Retweets, 574 Replies, 94 Likes.	Democrats are now openly talking about impeaching Donald Trump Source: Salon, 105 Retweets, 22 Replies, 153 Likes.
Sources: James Comey told lawmakers he wanted more resources for Russia probe http://politi.co/2r2Hxpf Source: Politico, 132 Retweets, 40 Replies, 204 Likes.	It appears that Trump may have just falsely accused himself of wiretapping himself: http://slate.me/2r9agb8 Source: Slate, 198 Retweets, 25 Replies, 303 Likes.
Trump says 'his decision' to fire FBI chief, calls him 'showboat': NBC interview http://reut.rs/2r6gUjg Source: Reuters, 69 Retweets, 62 Replies, 83 Likes.	What all the Russia investigations have done and what could happen next Source: The New York Times, 131 Retweets, 52 Replies, 226 Likes.
Condoleezza Rice: 'When you're not credible about Syria, you're not credible about North Korea' http://fxn.ws/2pudYMd Source: Fox News, 288 Retweets, 80 Replies, 806 Likes.	Republicans are allowing states to drug test people applying for unemployment benefits Source: Salon, 19 Retweets, 10 Replies, 16 Likes.
Senior US official says Trump administration has approved weapons for Kurds. http://apne.ws/2qYRLac Source: AP, 180 Retweets, 45 Replies, 97 Likes.	VP Mike Pence defends firing of FBI Director James Comey, says Trump 'made the right decision at the right time.' http://apne.ws/2q3fl7Q Source: AP, 65 Retweets, 83 Replies, 76 Likes.
When men and women finish school and start working, they're paid pretty much equally. But then it all changes. Source: The New York Times, 854 Retweets, 120 Replies, 1.1K likes.	Pres. Trump's firing of FBI Director James Comey is a "grotesque abuse of power," legal analyst Jeffrey Toobin says http://cnn.it/2q1FQd4 Source: CNN, 803 Retweets, 169 Replies, 1.3 Likes.

Table 3.9: Examples of high and low consensus news including the source, number of retweets, replies and likes.

news. Then, we propose a strategy that spreads high consensus news through neutral publishers, and quantify the significant decrease in the disparity of users' news exposure. Our extensive experiments on Twitter shows that seeding high consensus information with neutral publishers is an effective way to achieve high spread with little disparity regarding political leaning.

CHAPTER 4

Analyzing biases in perception of truth in news stories and their implications for fact checking

4.1 Introduction

Technologists, policymakers, and media watchdog groups are criticizing social media sites like Facebook and Twitter for allowing misinformation to spread unchecked on their platforms [Barrabi, 2018]. Recently, the 'PizzaGate' conspiracy theory has seen anew on teen-loved TikTok app [Ovide, 2020]. The spread of such "fake news" has been linked to foreign meddling in political elections [Allcott and Gentzkow, 2017], riots [Taub and Fisher, 2018], mass displacement, and even loss of human lives [Hogan and Safi, 2018]. Studies have proposed methods and tools to automatically detect fake news [Grinberg et al., 2019, Guess et al., 2019], for example, by identifying linguistic features employed by fake news creators [Shu et al., 2017, Chopra et al., 2017, Zhao et al., 2015, Bourgonje et al., 2017, Chakraborty et al., 2016b, Bhatt et al., 2017, Kumar et al., 2016], by analyzing the propagation patterns of such news in social media [Qazvinian et al., 2011, Ruchansky et al., 2017, Kwon et al., 2017, Kim et al., 2017, Oh et al., 2013, Vicario et al., 2019], or by checking new content against a database of known fake and real news [Ciampaglia et al., 2015, Kumar et al., 2017, Ruchansky et al., 2017]. Could we try to keep 3 refs at max per sentence?

Until reliant fully automated detection mechanisms for online misinformation arrive, social media platforms at large will remain dependent on human supervision to understand the news context [Graves, 2018]. Many platforms rely on crowdsourced reports as well as dedicated fact-checking outlets[1] like Snopes, PolitiFact, FullFact, and FactCheck [Lyons, 2018]. Stories deemed false by the fact-checkers

[1] https://www.facebook.com/help/1952307158131536

who follow principled methods such as Poynter's Code of Principles[2] and then ranked lower in users' news feeds or timelines, significantly limiting their future views [Lyons, 2018].

Fact checking by human experts is a highly resource-constrained process: it is not possible to fact check every news story circulated on social media. Thus, platform providers need to prioritize stories for fact checking. The most pertinent question that emerges in this context is *how should the platform prioritize 'check-worthy' news stories?* Social media sites currently encourage users to report any news they encounter on the platform they perceive to be fake. Stories reported as fake by numerous people are then prioritized for fact checking. In essence, to counter the proliferation of fake news, social media platforms are relying on their *users' perceptions of the truthfulness of news* to prioritize stories for fact checking.

Despite this reliance on user perceptions, no prior study has focused on understanding how the crowd perceives truth in news stories, and how these perceptions affect the detection and possible correction of online falsehoods. In this work, we perform the first in-depth analysis of users' truth perceptions of news stories – rather than news outlets [Pennycook and Rand, 2019] – by designing and validating a novel truth perception test. Using this test, we solicit users' truth perceptions for 150 stories that have already been fact checked, allowing us to compare users' perceptions to a known ground truth level determined by fact checkers.

Our comparison of users' perceptions of truth and actual ground truth reveals several discrepancies. To illustrate them, consider the following six stories:

(S1) Jared Kushner registered to vote as a woman in New York — *Fact-checked as False*

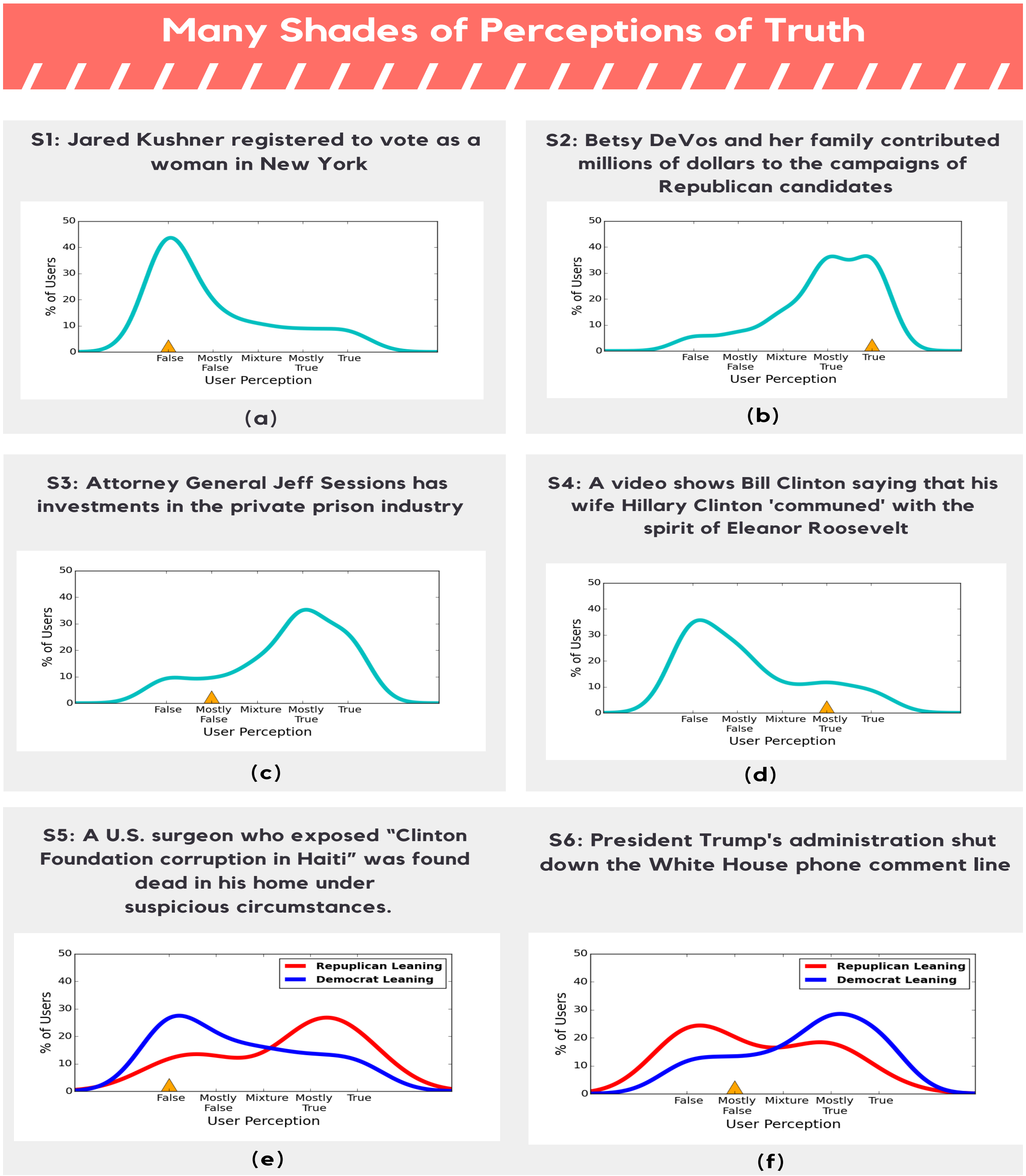
(S2) Betsy DeVos and her family contributed millions of dollars to the campaigns of Republican candidates — *Fact-checked as True*

(S3) Attorney General Jeff Sessions has investments in the private prison industry — *Fact-checked as Mostly False*

(S4) A video shows Bill Clinton saying that his wife Hillary Clinton 'communed' with the spirit of Eleanor Roosevelt — *Fact-checked as Mostly True*

(S5) A U.S. surgeon who exposed "Clinton Foundation corruption in Haiti" was found dead in his home under suspicious circumstances — *Fact-checked as False*

[2] **https://www.poynter.org/international-fact-checking-network-fact-checkers-code-principles**


Many Shades of Perceptions of Truth

S1: Jared Kushner registered to vote as a woman in New York

(a)

S2: Betsy DeVos and her family contributed millions of dollars to the campaigns of Republican candidates

(b)

S3: Attorney General Jeff Sessions has investments in the private prison industry

(c)

S4: A video shows Bill Clinton saying that his wife Hillary Clinton 'communed' with the spirit of Eleanor Roosevelt

(d)

S5: A U.S. surgeon who exposed "Clinton Foundation corruption in Haiti" was found dead in his home under suspicious circumstances.

(e)

S6: President Trump's administration shut down the White House phone comment line

(f)

Figure 4.1: Ground truth and perceived truth levels for six different news stories. Here, ground truth level (shown as orange triangles on x-axis) of each news story is obtained from Snopes, and the perceived truth levels are inferred by gathering the truth perceptions of 100 surveyed users.

(S6) President Trump's administration shut down the White House phone comment line — *Fact-checked as Mostly False*

Fig. 4.1 shows users' truth perceptions for these six stories, along with their fact-checked ground truth levels, as determined by Snopes. The difference between the ground truth and perceived truth levels highlights the need to account for different perception biases.

First, the majority of users correctly inferred the truthfulness of stories $S1$ and $S2$. Since story $S1$ is perceived to be false by most users, it may get reported by many and is thus likely to be prioritized by the social media platforms to be fact checked. However, we assert that there is *little* to be gained by fact checking stories whose truth value is already correctly judged by the crowds, just as there is little use in fact checking claims by news satire outlets like The Daily Show[3] and The Onion (theonion.com).

On the contrary, the figure shows that there exist biases in the truth perceptions of the users for stories $S3$ and $S4$, with significant differences between the truth levels perceived by the users and the stories' actual ground truth levels. $S3$ reveals *gullibility* of the users, where people over-estimate the truth level of the story (*i.e.,* false positive bias). In contrast, $S4$ reveals users' *cynicality* – people under-estimate the truth level of the story (*i.e.,* false negative bias). Interestingly, $S4$ is more likely to be reported by users and fact checked with higher priority than S_3. In fact, on today's social media platforms, the higher the false positive bias in the perceptions of a story, the less likely it is to be reported and, consequently, become a subject for fact checking. Worse, current social media platforms do not have mechanisms to reassure users about the credibility of a true story like S_4 that is mistakenly perceived by many users as false (*i.e.,* high false negative bias), even after the story is fact checked.

Figs. 4.1(a-d) also highlight disagreements between users about the truthfulness of individual stories. These disagreements are highly correlated with their political leaning. Fig. 4.1(e) and 4.1(f) show that users with different political ideologies (*e.g.,* Democrat and Republican-leaning users) indeed perceive truth in news stories differently. People are more likely to trust stories that confirm their political beliefs, while they are more likely to distrust stories that contradict their beliefs. Story $S6$, which attacks Trump's administration, is 'Mostly False' as determined by the expert fact checkers. However, the majority of users who identify themselves as Democrats perceive this story to be accurate, while most Republican users label it as false. On the other hand, the story $S5$, which raises questions

[3] http://www.cc.com/shows/the-daily-show-with-trevor-noah

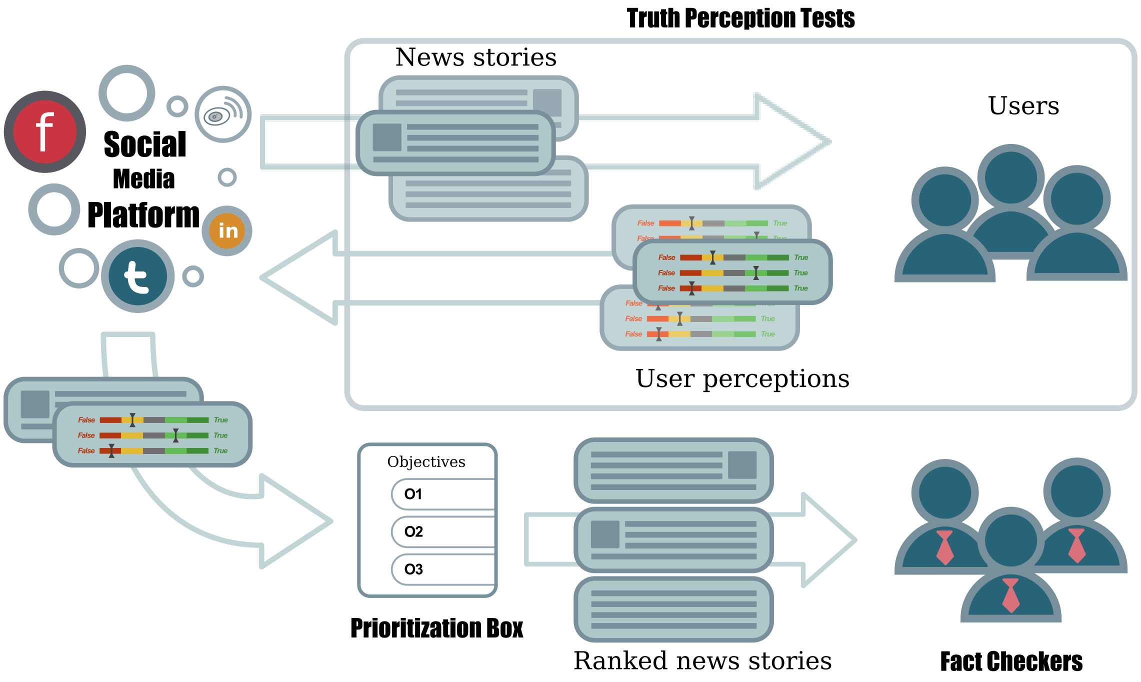


Figure 4.2: Overview of our proposed framework for social media platforms to prioritize stories for fact checking by leveraging users' truth perceptions. Social media platforms first gather users' perceptions of truth for the different news stories being shared on the platform using our Truth Perception Tests (RQ1). Then the platforms pass the news stories along with the users' truth perceptions to the prioritization box (RQ2) and specify the prioritization objective. The prioritization box would then output a ranked list of news stories that should be prioritized for fact checking based on the platform's chosen objective (RQ3).

against Clinton, is 'False' according to the expert fact-checkers, while the majority of Republican-leaning users perceive it to be accurate, and Democrat users perceive it as false.

These examples highlight the pitfalls of ignoring biases in the crowds' truth perceptions when using them to prioritize stories for fact checking, and suggest the need for a clear definition of objectives for the prioritization of stories, to ensure that the power of the crowd is being used appropriately to meet these objectives. This research proposes a framework (shown in Fig. 4.2) for social media platforms to prioritize stories for fact checking by more effectively leveraging users' truth perceptions to satisfy three important objectives:

O1: *Removing false news stories from circulation*

To restrict their circulation on social media platforms, any stories that are false need to be fact checked with higher priority. Intuitively, this objective has been the primary focus of social media platforms. Since the truth values of stories are not known beforehand, prior research efforts [Shu et al., 2017, Kumar et al., 2016, Chen et al., 2015, Rony et al., 2017, Qazvinian et al., 2011, Kwon et al., 2013a, Oh et al., 2013, Maddock et al., 2015, Kwon et al., 2013b, Zhao et al., 2015, Kwon et al., 2017, Ma et al., 2016, Ruchansky et al., 2017] have focused on automatically detecting potentially false stories. Such potential false stories can then be prioritized for fact checking.

O2: *Correcting the misperception of users*

While must of the prior work has argued for the removal of false stories from social media platforms, legal experts and free speech campaigners have compared it to censorship[4]. To address such concerns, social media platforms may want to prioritize and fact check stories for which users' perceived truth levels are far from their ground truth levels, and flag these stories rather than removing them altogether.

O3: *Decreasing disagreement among different users' perceptions*

For the society to have fruitful debates in the public sphere, it is essential to set a common ground for different sections of the society. To ensure the existence of a common ground, it is essential to identify topics that incur a significant degree of disagreement, and the platforms can then prioritize them for fact checking to let people know the *objective* truth value of the stories. In our experiments, such stories have a significant variance in truth perceptions reported by different users, especially when these users have different ideological leanings.

Given this context, we answer the following three research questions in this paper:

- **RQ1:** How can we collect users' perceptions of truth in news stories in a robust manner?
- **RQ2:** How do the three objectives for fact checking compare to one another? Can they be satisfied simultaneously?
- **RQ3:** If a platform chooses an objective for prioritizing stories for fact checking, how can the objective be implemented by leveraging users' perceptions of truth in news stories?

4.2 Background

Over the recent years, the meaning of fake news has evolved and become compatible with disseminating false information. For instances, Allcott and Gentzkow [2017] define it as "a news article that is intentionally and verifiably false" and Golbeck et al. [2018] describe it as "information presented as a news story that is factually incorrect and designed to deceive the consumer into believing it is true".

[4] **https://www.theguardian.com/media/2018/apr/24/global-crackdown-on-fake-news-raises-censorship-concerns**

There are several definitions or types of fake news such as fabricated content, misleading content, imposter content, manipulated content, false connection, and false context [Wardle, 2017]. Sharma et al. [2019] generalize the fake news definition as "A news article or message published and propagated through media, carrying false information regardless the means and motives behind it" to capture the different types of fake news. In this section, we review the detection of fake news from four viewpoints:

4.2.1 Content-based methods:

One way to assess the authenticity of news is to evaluate the content of news, such as text or images. In this section, we briefly discuss fake news detection methods based on the content of the information. **Text and image features**: Traditional machine learning frameworks (supervised, semi-supervised, or unsupervised) use a set of manually selected features at various language levels such as lexicon, syntax, semantic, and discourse-level to detect fake news [Feng et al., 2012, Pérez-Rosas et al., 2017, Zhou et al., 2019, Chen and Guestrin, 2016]. Later, by embedding text and images as news content to word-level [Mikolov et al., 2013] or pixel matrix, well-trained neural networks such as VGG-16/19 [Simonyan and Zisserman, 2014], Text-CNN [Kim, 2014], RNNs [Hochreiter and Schmidhuber, 1997], GRUs [Cho et al., 2014], and BRNNs [Schuster and Paliwal, 1997], and the Transformer [Devlin et al., 2018, Vaswani et al., 2017] are used to extract latent textual and visual features of news content. Finally, given news is classified as true or fake news. **Fact-checking**: Experts check the news produced by traditional media to assess news authenticity. Social media platforms currently rely on human experts and dedicated fact-checking outlets, such as Snopes (`snopes.com`), PolitiFact (`politifact.com`), Full Fact (`fullfact.org`), and FactCheck (`Fact Check.org`) [Lyons, 2018]. These websites provide content such as what is false and why is it false. They also provide invaluable insights for identifying check-worthy content [Kumar et al., 2017] and explainable fake news detection [Shu et al., 2019]. Stories deemed false by the fact checkers, who follow principled methods such as Poynter's Code of Principles [5], are then ranked lower in users' news feeds or timelines, significantly limiting their future views [Lyons, 2018]. Human supervision limits the number of claims of news that is fact-checked. It is not possible to fact check every news story circulated on social media. Thus, platform providers need to prioritize stories for fact-checking. The most pertinent question that emerges in this

[5] **https://www.poynter.org/international-fact-checking-network-fact-checkers-code-principles**

context is *how should the platform prioritize 'check-worthy' news stories?* Currently, social media sites encourage users to report any news stories reported as fake by numerous people are then prioritized for fact-checking. In essence, to counter the proliferation of fake news, social media platforms are relying on their *users' perceptions of the truthfulness of news* to prioritize stories for fact-checking.

4.2.2 Propagation-based methods:

Malicious spreaders can easily manipulate the content-based methods that are being used for detecting fake news. Thus, several studies focus on other methods; for example, [Jin et al., 2018, Zhou et al., 2019] claim that fake news has different patterns compared to true news, such as having high informality and diversity as well as being more emotional. Vosoughi et al. [2018] have observed that fake news spreads through social media with different patterns compared to true news. Several cascade features such as cascade size, cascade breadth, cascade depth, structural virality (Average distance among all pairs of nodes in a cascade), node degree, spread speed, and cascade similarity are used to classify the news as fake or true [Castelo et al., 2019, Vosoughi et al., 2018, Wu et al., 2015]. Bian et al. [2020] and Ma et al. [2018] develop recursive neural networks based on news cascades to classify the news.

4.2.3 Source-based methods:

Untill now, we focused on the authenticity of the news; however, one can detect fake news by focusing on the credibility of its source. By source of news, we mean the sources that create and publish the news as well as the sources that spread the news stories [Shu et al., 2020, Zhang et al., 2018, Zhou and Zafarani, 2019]. We just need to assess a few outlets' credibilities for traditional mass media or popular news publishers in social media. Sitaula et al. [2020] construct the collaboration network of news authors in which they show that the networks are homogeneous. Network homogeneity means that fake-news authors are more densely connected. True news authors are also strongly connected, while there is a weak connection across the groups. There are also many resources that show the ground truth on the credibility of news sources such as Meida/Fact Check [6] or NewsGuard [7], which provide the list of news sources with their credibility based on a different point of view such as political leaning. However, in social media, every user can be a news publisher, and they can be malicious users (they

[6] https://mediabiasfactcheck.com/

[7] http://www.newsguardtech.com/

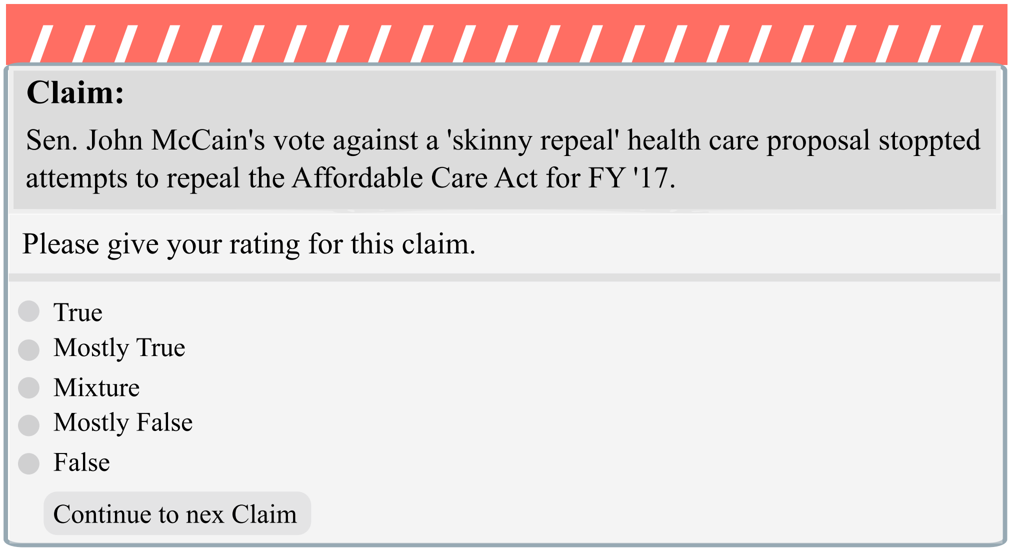
intentionally spread the fake-news same as bots) or vulnerable normal users (spread the fake-news unintentionally without recognizing the falsehood). Several works detect malicious or bots using groups of features such as network, user, friend, temporal, content, sentiment [Cai et al., 2017, Morstatter et al., 2016, Shao et al., 2018].

The above studies discuss various methods to identify false news stories on the Web. Once detected, false stories may be treated in several different manners, which are non-orthogonal:

- **Strategy 1: Remove false news stories from circulation**. Upon detection, a hard-line policy is to remove them entirely to block their spread on social platforms. Alternative soft-line policies would be to down-rank contents or label them as "false." Previous research has shown that contents labeled as a rumor will less likely to spread further, indicating efficacy of labeling [Friggeri et al., 2014] Various independent fact-checking agencies such as snopes.com and politifact.com act as distributed data sources for news platforms.

- **Strategy 2: Correct the misperception of the users**. Beyond reducing the circulation of false news stories, a more active mitigation strategy is to "correct" for its impact on social networks. While there exists a number of reputable fact-checking sources and studies, no study has examined how false or true news stories may be differently perceived by people irrespective of their ground truth. For example, false urban legends may spread even when people are aware of their veracity, simply because they are amusing. Perception towards political news stories may vary depending on one's underlying political belief. In correcting the misperception of users, one needs to decide which stories to prioritize (i.e., the current study) and also design an effective methods for correction, which is beyond the scope of this paper.

- **Strategy 3: Decrease the disagreement between the users' perceptions of truth**. In light of building a healthy public sphere that allows diverse ideologies, it is necessary to set a common ground on the intent and knowledge of news stories. Common ground can be achieved by helping to decrease the disagreement amongst users' truth perceptions on news stories. For this, one needs to identify news stories of upmost disagreement to prioritize (i..e, the current study) and design an effective methods for mitigation (i.e., out of scope of this paper). This paper will discuss an effective method to identify news articles of upmost disagreement.



Claim:

Sen. John McCain's vote against a 'skinny repeal' health care proposal stoppted attempts to repeal the Affordable Care Act for FY '17.

Please give your rating for this claim.

○ True
○ Mostly True
○ Mixture
○ Mostly False
○ False

Figure 4.3: An example of the survey question that we used for performing Truth Perception Tests for the news claims in our dataset.

The remainder of this paper will introduce data and methodology, expand on the needs for these mitigation strategies, and suggest specific algorithms for each strategy.

RQ1: Designing truth perception tests

To address our first research question, we designed *Truth Perception Tests* (TPTs) that can be used to assess how users implicitly perceive truth in news stories — *i.e.,Perceived Truth Level (PTL)*. Using this methodology, we solicit users' truth perceptions for a set of 150 stories that have already been verified by expert fact-checkers at Snopes.com, and thus have a known *Ground Truth Level (GTL)* to which we can compare users' truth perceptions.

We perform TPTs as online surveys. While we did not limit our respondents to a specific time frame, we strongly encouraged them to respond rapidly by giving them the following instructions at the start of the test: "Please do not conduct any web search or use any online/offline resources for verifying or validating the claim presented to you. Please use your best judgment (your instinctive gut based guess within a few seconds) to label the claims." On average, our respondents gave their truth perception responses for each claim within 10 seconds.

To gather truth perceptions, we showed respondents a news claim and asked them to label the claim as either 'True', 'Mostly True', 'Mixture', 'Mostly False' and 'False', as shown in the example depicted in Fig. 4.3. We mapped these five perceived truth level (PTL) choices to a scale between -1.0

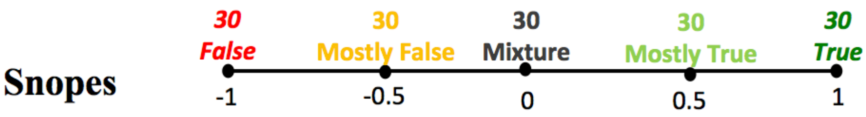


Figure 4.4: Mapping truth labels of news stories on a scale between -1.0 and + 1.0. The number of stories collected for each ground truth label are also indicated.

and +1.0. By aggregating the answers given by each user *u* for a news story *S*, $PTL_u(S)$, we compute the aggregate perceived truth level, $PTL(S)$, of the story as follows:

$$\text{PTL(S)} = \frac{1}{N}\sum_{u=1}^{N} PTL_u(\text{S}) \tag{4.1}$$

where *N* is the total number of users whose truth perceptions for the story *S* are being aggregated.

The news claims utilized in the TPTs were drawn from news stories that had been professionally fact-checked by Snopes and thus we know their ground truth level. Snopes uses the same set of labels that we used as our answer choices to categorize news stories: ‘False’, ‘Mostly False’, ‘Mixture’, ‘Mostly True’, and ‘True’. Again, we mapped these truth categories on a scale between -1.0 and +1.0, as shown in Fig. 4.4. In January 2018, from the claims labeled under the Politics topic category by Snopes, we selected 30 recently fact checked news stories for each truth category to get a total of 150 stories. The ground truth level for each story *S*, $GTL(S)$, is given by the value of the truth category assigned by Snopes for that story.

Test design validation

To ensure that our TPTs are maximally robust to variations in deployment and a broad set of potential survey biases, we conducted multiple micro-experiments. In these micro-experiments we evaluated how, if at all, different test designs may influence our results. We evaluated three types of effects:

- *Sample Effects*: Survey methodology literature [Peer et al., 2014] reports that less-naive respondents (*e.g.,* experts) may answer certain survey questions differently than naive respondents. Additionally, demographic composition of the survey sample is known to affect the generalizability of results[AAPOR, 2010]. Therefore, to account for such sample effects, we compared the results of our tests: (i) when run using Amazon Mechanical Turk (MTurk) Masters [AMT, The Mechanical Turk Blog 2011] vs. naive MTurk workers, both from the US; and, (ii) when run using MTurk Masters from the US vs. a census-representative sample of US participants recruited by Survey Sampling International (`surveysampling.com`).

The survey variations we compare in the context of sample effects are:

- Running our test using MTurk Masters vs. naive MTurkers.
- Running our test using census-representative sample of participants recruited by Survey Sampling International(SSI participants) vs. MTurk Masters(experts participants).

Surveys	Chi square dependency of Dist-ANS	Chi square dependency of Acc	Correlation of TPB
MTurk Masters & MTurk naive	Chi-value:0.0 p-value=1.0	Chi-value:0.0 p-value=1.0	0.9
MTurk naive & SSI workers	Chi-value:0.0 p-value=1.0	Chi-value:0.0 p-value=1.0	0.89

Table 4.1: Sample effects: We evaluate the similarity between the distribution of answers to each survey using a X^2 test of independence.

Table 4.1 depicts that both distribution of answering and accuracy of judgments are independent (fail to reject H0) of types of survey respondents. Last column also shows that there is a significantly high correlation between TPB of claims of different surveys with different workers samples. This means that a particular claim has a close value of TPB in different surveys which confirm that our measure is robust against the sample effects.

Where, **Total Perception Bias (TPB)** of a story S captures the total error (gullibility or cynicality) in the users' perceptions of truth levels of the story, and is given by

$$\text{TPB(S)} = \frac{1}{N}\sum_{u=1}^{N} |PTL_u(\text{S}) - GTL(\text{S})| \tag{4.2}$$

where N is the total number of users whose truth perceptions of the story S are being aggregated.

- *Answer Choice Effects*: It has been reported previously [Redmiles et al., 2017] that Likert scale length (e.g., even or odd numbers of answer choices, where scales with an odd number of answer choices include a “middle” neutral option), may effect the strength of participants' responses. We compared the effects of using a 6 and 7 point Likert item scale. Additionally, the text labels of the Likert answer choices may also affect respondents' answers to survey questions[8]. To examine this effect, we compared the effect of using the Snopes' labels (see Fig. 4.3) with an alternate 7 point scale ("I can confirm it to be true", "Very likely to be true", "Possibly true", "Can't tell", "Possibly false", "Very likely to be false", and "I can confirm it to be false") and 6 point scale (which excluded the "Can't tell" option from the 7 point scale). We evaluated the answer choice effects by comparing 6, 7 point scale with Snopes' 5 point scale. Table 4.2 depicts that both distribution of answering and accuracy of judgments are independent (fail to reject H0) of the types of answer choices in the surveys. Significant high correlation between TPB of claims of different surveys with different answer choices is shown in last column.

Surveys	Chi square dependency of Dist-ANS	Chi square dependency of Acc	Correlation of TPB
7-pt scale & 6-pt scale	Chi-value:0.0 p-value=1.0	Chi-value:0.0 p-value=1.0	0.94
7-pt scale & 5-pt scale	Chi-value:0.0 p-value=1.0	Chi-value:0.0 p-value=1.0	0.98

Table 4.2: Answer choice effects

- *Satisficing and Incentive Effects*: Satisficing [Krosnick et al., 1996] is a commonly observed survey response effect in which respondents select what they consider to be the minimum acceptable answer, without fully considering their true feelings. Surveys such as our TPTs may be at particular risk of satisficing because they encourage quick responses. Thus, we explored the effect of incentivizing participants to provide correct answers to evaluate whether satisficing may be affecting our test results.

To investigate the impact of satisficing and incentives, we designed a survey in which we gave respondents incentives for answering correctly. At the beginning of the survey, we told the participants: "In addition to the amount promised for the task, for each of your judgements which CORRECTLY matches the actual truth status of the claims, we will pay you 5 cents as a bonus. For example, if you judge a claim to be 'True', or 'Mostly True', and the claim is actually true, then you'll get 5 cents for

[8] Prior work shows that it is always best practice to have text labels on Likert item points [Krosnick and Fabrigar, 1997], thus we do not examine the omission of text labels.

Surveys	Chi square dependency of Dist-ANS	Chi square dependency of Acc	Correlation of TPB
5-pt scale & incentive and 5-pt scale	Chi-value:0.0 p-value=1.0	Chi-value:0.0 p-value=1.0	0.85
7-pt scale & incentive and 7-pt scale	Chi-value:0.0 p-value=1.0	Chi-value:0.0 p-value=1.0	0.92

Table 4.3: Satisficing and Incentive effects

the claim. Similarly, to get the bonus for an actual false claim, it should be judged by you as 'False' or 'Mostly False'. Finally if you judged the claim as 'Mixture' and the claim actually is mixture or mostly true/mostly false you will earn bonus." To ensure that participants do not use online or offline resources to estimate the truthfulness of the claims we showed a timer in each page and told them: "If your judgment for each question takes more than 15 seconds then there would not be any bonus, even if you answer the question correctly."

To see if incentivizing has any effect or not we compare the incentivized survey with the unincentivized survey. Table 4.3 depicts that incentivizing has no effect.

In brief, we found no statistically significant differences across the survey variations for the proportion of correct answers. Additionally, we observed statistically significant high-correlation between our proposed measure of TPB, computed for our survey variations, with the Pearson correlation coefficients ranging from 0.90 to 0.96. Figures 4.5 depicts that that wisdom of crowds (accuracy of judging by users) is very similar across different survey variations.

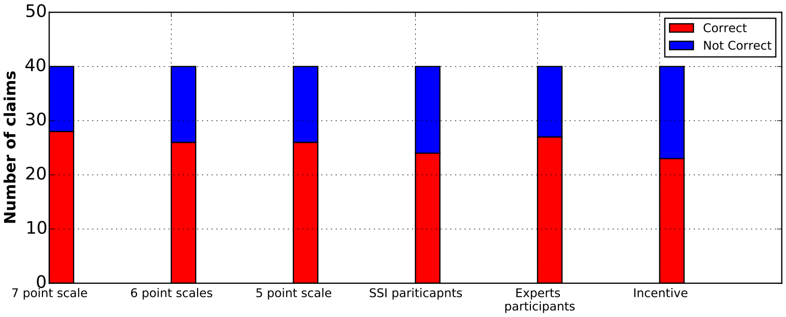


Figure 4.5: Accuracy of judgments(wisdom of crowds) for different design surveys.

Figures 4.6 summarizes the results of comparing TPB test variations, which show very similar results across variants. We thus conclude that our test is relatively robust and consequently useful for application in industry settings and future research on content misperceptions.

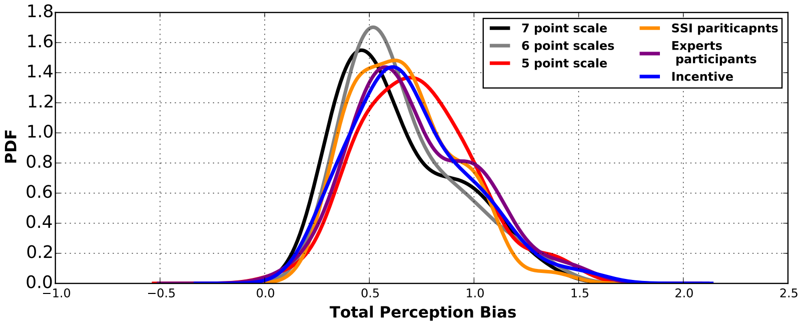


Figure 4.6: PDF of Total Perception Bias(TPB) for different design surveys.

Data collection

We ran our validated truth perception tests on MTurk during May-June 2018 collecting a total of 15,000 responses. Each MTurk worker saw 50 claims and no worker could take the survey more than once. Any Mturk worker over the age of 18 who resided in the US was eligible to participate in our survey.

Limitations

While we validate our truth perception tests extensively to ensure they are robust against design variations, our method does have some limitations which we discuss here. When users encounter and flag false news stories on the social media platforms, they are not only exposed to the claim or headline, but also to the source of the article, the images from the article, summary snippet or text of the article, and additional context for instance likes or shares for the story etc. Our controlled experiments do capture the effect of the claim (or headline) of the news stories on the users, but they do not capture the effects of other factors as yet, and a promising direction of future work would be to design controlled experiments to measure the impact of the other factors.

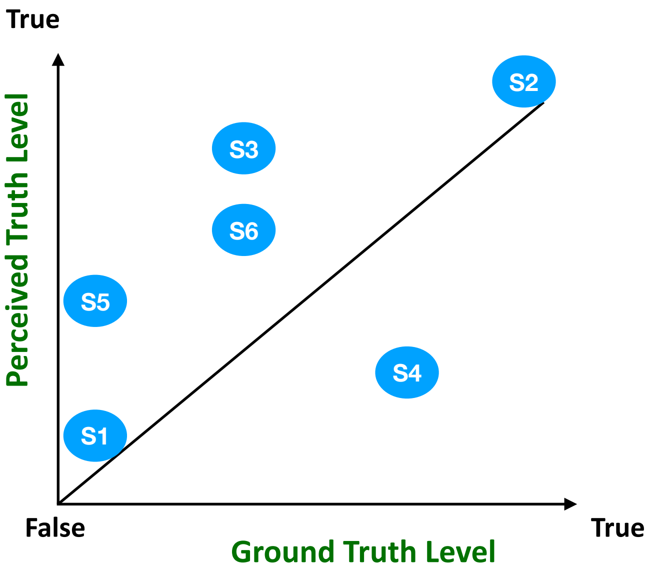


Figure 4.7: Sample stories from Fig. 4.1. Perceived Truth Level is determined by averaging truth perceptions of 100 AMT workers. Ground Truth Level is determined by Snopes.

RQ2: Comparing the prioritization objectives

As discussed earlier, social media sites today prioritize stories based on the number of reports they receive from users flagging a piece of content as false. This approach assumes that false stories will receive more reports from users than true stories and hence will be fact checked with higher priority than true stories. Fig. 4.7 depicts the perceived truth levels of the six stories mentioned in Introduction, versus their ground truth levels as determined by Snopes.

Under the current strategy, the priority order would be S1, S4, S5, S6, S3, and then S2, while the desired ranking to satisfy objective O1 (removing false news stories by ranking according to GTL) would be S1, S5, S6, S3, S4, and then S2. Thus the current strategy does not satisfy this objective satisfactorily.

As described earlier, based on our analysis of users' truth perceptions, we identified two additional objectives for fact checking stories: O2 (correcting users' misperceptions) and O3 (decreasing disagreement among users). Thus, we also seek to evaluate: *Does the current strategy satisfy O2 or O3? Are these three objectives (O1, O2 and O3) compatible and can one strategy address them simultaneously?*

In this section, we compare objectives O1, O2, and O3 to see if they are addressed by the current reporting-based strategy and to evaluate whether these three objectives can be satisfied simultaneously. For performing these comparisons, we need to first prioritize the six stories according to every objective.

	O1 and O2	O1 and O3	O2 and O3
Spearman's ρ	0.31	-0.05	-0.01

Table 4.4: Correlation between rankings to satisfy different objectives.

Now we compare objectives O1, O2, and O3. For objective O2, we use the metric of Total Perception Bias (TPB) to rank the stories. Intuitively, TPB captures the aggregate deviation of perceived truth level (aggregated over N users) from ground truth level of a story S that we discussed it earlear.

For ranking according to O3, we can either rank news stories using Disputability (*i.e.,* the variance in the individual truth perceptions of users) or according to Ideological Mean Perception Bias (IMPB) which captures the difference in truth perceptions of different ideological groups (Democrats and Republican in this case), given by

$$IMPB(S) = |MPB_{Dem}(\mathrm{S}) - MPB_{Rep}(\mathrm{S})| \tag{4.3}$$

where, $MPB(S)$ = PTL(S) - GTL(S), measures the error in the collective perceptions of users in assessing the truth level of a story.

When we rank the example stories based on the three objectives using these metrics, we get different priority orders:

Priority order to satisfy O1: S1,S5,S6,S3,S4,S2

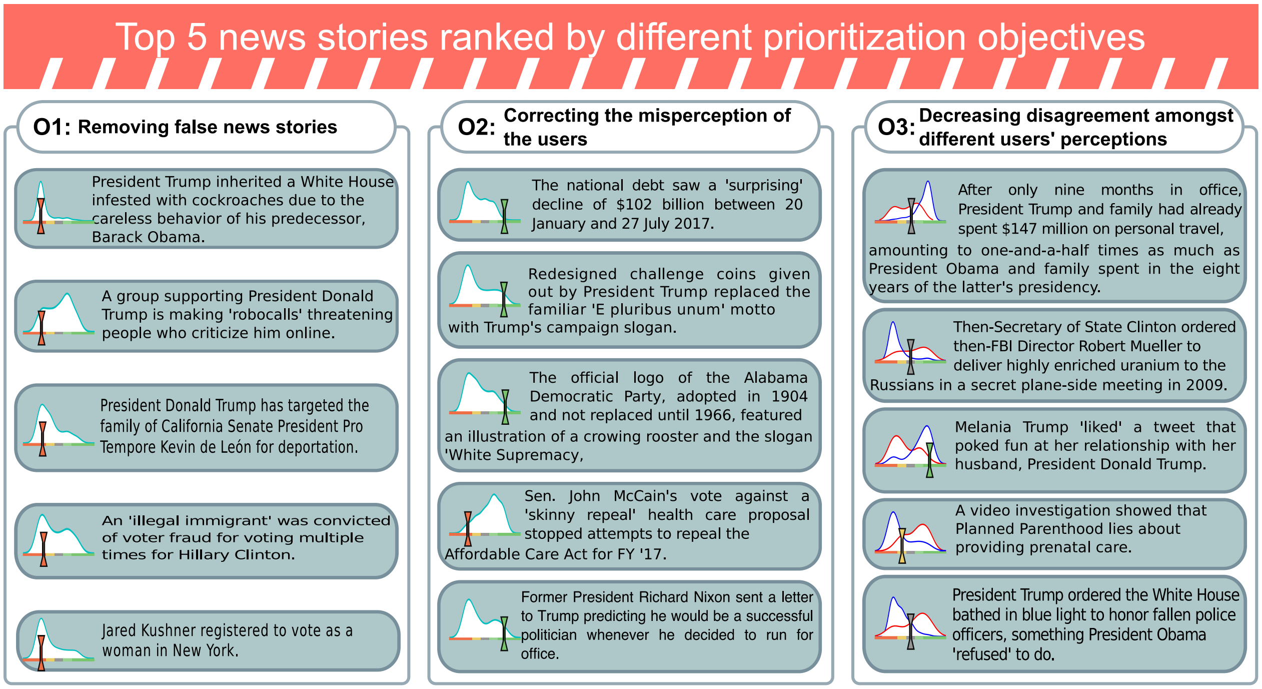
Priority order to satisfy O2: S4,S5,S3,S6,S2,S1

Priority order to satisfy O3 (Disputability): S3,S5,S6,S2,S1,S4

Priority order to satisfy O3 (IMPB): S5,S6,S1,S4,S3,S2

Moreover, when we consider the full dataset of all 150 news stories and rank them according to each objective, we observe little correlation between the rankings achieved when satisfying these different objectives. While we can observe some association of ranks between O1 and O2, there is almost no association (ρ close to 0) between the other pairs of objectives. Table 4.4 presents the Spearman's rank correlation coefficient ρ (ranging between +1.0 and -1.0) between stories ranked by different objectives. Thus, we can conclude that *these three objectives are incompatible, and can not be satisfied simultaneously*.

Thus, platform providers must chose one objective over the others to prioritize stories. Each choice of objective will necessitate that an entirely different set of potentially "fake" news stories willl remain


Top 5 news stories ranked by different prioritization objectives

O1: Removing false news stories

President Trump inherited a White House infested with cockroaches due to the careless behavior of his predecessor, Barack Obama.

A group supporting President Donald Trump is making 'robocalls' threatening people who criticize him online.

President Donald Trump has targeted the family of California Senate President Pro Tempore Kevin de León for deportation.

An 'illegal immigrant' was convicted of voter fraud for voting multiple times for Hillary Clinton.

Jared Kushner registered to vote as a woman in New York.

O2: Correcting the misperception of the users

The national debt saw a 'surprising' decline of $102 billion between 20 January and 27 July 2017.

Redesigned challenge coins given out by President Trump replaced the familiar 'E pluribus unum' motto with Trump's campaign slogan.

The official logo of the Alabama Democratic Party, adopted in 1904 and not replaced until 1966, featured an illustration of a crowing rooster and the slogan 'White Supremacy,

Sen. John McCain's vote against a 'skinny repeal' health care proposal stopped attempts to repeal the Affordable Care Act for FY '17.

Former President Richard Nixon sent a letter to Trump predicting he would be a successful politician whenever he decided to run for office.

O3: Decreasing disagreement amongst different users' perceptions

After only nine months in office, President Trump and family had already spent $147 million on personal travel, amounting to one-and-a-half times as much as President Obama and family spent in the eight years of the latter's presidency.

Then-Secretary of State Clinton ordered then-FBI Director Robert Mueller to deliver highly enriched uranium to the Russians in a secret plane-side meeting in 2009.

Melania Trump 'liked' a tweet that poked fun at her relationship with her husband, President Donald Trump.

A video investigation showed that Planned Parenthood lies about providing prenatal care.

President Trump ordered the White House bathed in blue light to honor fallen police officers, something President Obama 'refused' to do.

Figure 4.8: The top 5 ranked news stories prioritized according to the three objectives of social media platforms for selecting stories for fact checking. The low overlap between the three ranked lists highlights the complementary nature of the objectives.

unverified. To illustrate this effect, Fig. 4.8 displays the top five news stories, ranked by each objective. Thus, special care needs to be taken by the platforms to finalize the design of their fact checking exercise.

RQ3: Operationalizing objectives using truth perceptions

Having identified and compared three potential objectives that a social media platform could have for prioritizing stories for fact-checking, we now focus on how the platform can operationalize these objectives by leveraging users' perceptions of truth in news stories.

O3: Decreasing disagreement among different users' truth perceptions

We start by describing the easiest objective to operationalize (O3). The goal is to prioritize stories that have the highest disagreement in user truth perceptions. We quantify the disagreement in users' perceptions as the *disputability* of news stories, *i.e.,* the variance in the individual truth perceptions of users. The platform can collect users' truth perceptions and rank the stories according to their disputability to satisfy O3.

If the ideological leanings of the users assessing the stories are known then the stories which have a maximum disagreement in the perceptions of users with different ideologies can be prioritized. We capture such differences in assessment as the Ideological Mean Perception Bias (IMPB) of a story, as defined earlier. Most social media platforms, such as Facebook and Twitter, have detailed information about their users via users' explicit inputs or behavior on these platforms, including information on their potential ideological leaning. Therefore, platforms could compute the IMPB of a story to assist in fact checking prioritization.

Further, even in the absence of such information about the ideological leanings of users, it is possible to achieve O3. We found that the disputability of stories is moderately correlated (Pearson Correlation: 0.38) with IMPB. Thus, prioritizing stories by disputability also prioritizes stories with higher variation in perception between users with different ideological leanings.

O2: Correcting the misperception of users

To correct users' misperceptions, we need to quantify the extent to which users incorrectly perceive the truth of a story. To do so, we use the previously defined Total Perception Bias (TPB) metric to measure the aggregated error (gullibility or cynicality) in users' perceptions of a story S. Ranking stories by TPB prioritizes misperceived stories: stories where users' perceived truth levels (PTL) differ widely from the ground truth level (GTL) of the story. However, to compute TPB, we must know GTL, which is not available in practice. Here, we propose an alternative approach: training a supervised learning classifier that classifies a story as having either high or low TPB. To design such a classifier, we need the GTL of a small set of stories that have been labeled as high or low TPB for generating the training data. Then, TPB can be predicted for a larger set of stories for which GTLs may not be known.

As an illustration, we construct a classifier to predict the TPB values for the 150 stories we studied in this work. We label a news story to have 'High TPB' if it has a TPB value above the median TPB value, or 'Low TPB' if it has a value lower than the median. We split our dataset of 120 claims, and consider 80% of the data (96 claims) as the training dataset and the remaining 20% (24 claims) as the test dataset. Using this ground truth dataset, we train four types of classifiers (Linear SVM, Naive Bayes, Logistic Regression, and Random Forest). Our feature set includes the mean, median, variance, and skew of perceptions of users with different demographic features such as 'Political Ideology', 'Age', 'Gender', 'Education', 'Employment', 'Income', and 'Marital status' (Table 4.5). Applying feature

Demographic attribute	**Attribute values**
Political ideology	Conservative, Moderate, Liberal
Age	18-24, 25-34, 35-44, 45-54, 55-64, 65-74
Gender	Female, Male
Education degree	College graduate bs/ba or other 4year degree, Postgraduate training or professional schooling after college toward a masters degree or PhD law or medical school, Post-graduate training or professional schooling, Some college associate degree no 4 year degree, High school graduate grade 12 or certificate, Technical trade or vocational school after high school'
Employment	In full-time work permanent, In full-time work temp contract, Retired, Unemployed, In part-time work permanent, In part timework temp contract, Student only, Part-time work part-time student, Self-employed
Income	Under 10000, 10000-20000, 30001-40000, 40001-50000, 50001-60000, 60001-70000, 70001-100000, 100001-150000, 150001 or more
Marital status	Married, Living with partner, Divorced, Widowed, Separated, Single

Table 4.5: Demographic attributes collected from the Amazon Mechanical Turk workers who took our Truth Perception Tests.

ranking with recursive feature elimination, we observed that the best set of features includes 'Political Ideology', and 'Income'.

For training our classifiers, we use 5-fold cross-validation. In each test, the original sample is partitioned into 5 sub-samples, out of which 4 are used as training data, and the remaining one is used for testing the classifier. The process is then repeated 5 times, with each of the 5 sub-samples used exactly once as the test data, thus producing 5 results. The entire 5-fold cross-validation was then repeated 20 times with different seeds used to shuffle the original dataset, thus producing 100 different results. The results reported are average accuracies across these 100 runs, along with the 90% confidence interval.

We observe an average prediction accuracy of 82% (using Linear SVM & Random Forest classifiers), with 90% confidence interval of 0.09%, illustrating the potential for satisfying O2 given a small ground truth dataset. In second row of Table 4.6, we depict the performance as the average accuracy across the 100 runs along with the 90% confidence interval of the four types of supervised classifiers for our prediction task, using the best set of features (including 'Political ideology' and 'Income') determined by feature ranking with recursive feature elimination. As shown in the table, we achieve maximum accuracy of 82%.

Table 4.6: Prediction results using different types of supervised methods for the two tasks of predicting GTL and TPB. Performance of each classifier is reported as the average accuracy across the 100 runs along with the 90% confidence intervals.

	Linear SVM	**Naive Bayes**	**Logistic Regression**	**Random Forest**
Predicting GTL	0.7 ± 0.007	0.67 ± 0.008	0.68 ± 0.010	0.7 ± 0.009
Predicting TPB	0.82 ± 0.009	0.78 ± 0.008	0.79 ± 0.008	0.82 ± 0.010

Note that our prediction algorithm for TPB of news stories is based only on users' truth perceptions and their basic demographic attributes. We believe, predictive performance could be further improved by including more detailed demographic and behavioral features, typically available to social media platforms.

O1: Removing false news from circulation

Finally, to operationalize O1, social media platforms need to prioritize false stories for fact-checking. We examined two methods that leverage the users' truth perceptions (PTL) to estimate the ground truth levels of news stories. For both the methods, we need a labeled ground truth dataset, so we label all the stories annotated to be 'True' or 'Mostly True' by Snopes to be 'True', while labeling all stories annotated to be 'False' or 'Mostly False' by Snopes as 'False'. Ignoring the stories labeled 'Mixture', we were left with a labeled dataset of 60 'True' stories and 60 'False' stories.

We first took a "wisdom of crowds" approach and estimated the GTL using the average PTL value for the 100 workers who assessed the story. We considered stories with a positive average PTL to be 'True', while negative ones to be 'False'. We observed that we correctly assess the truth labels for 67% of stories in our ground truth labeled dataset. Additionally, when we rank stories by PTL and GTL, respectively, we observe a moderate ranking correlation of 0.4.

Alternatively, similar to O2, we trained supervised classifiers to predict the truth value ('True' or 'False') of a story. Using the same set of classifiers, feature set and experimental setup as O2, we achieve an average accuracy of 70% (using Linear SVM & Random Forest classifiers) across the 100 runs, with a 90% confidence interval of 0.7%. In first row of Table 4.6, we depict the performance as the average accuracy across the 100 runs along with the 90% confidence interval of the four types of supervised classifiers for our prediction task.

Operationalizing O1 proved to be very challenging, as also demonstrated by the amount of prior research on identifying "fake" news stories in recent times [Shu et al., 2017, Chopra et al., 2017, Zhao et al., 2015, Bourgonje et al., 2017, Chakraborty et al., 2016b, Bhatt et al., 2017, Kumar et al., 2016, Qazvinian et al., 2011, Ruchansky et al., 2017, Kwon et al., 2017, Kim et al., 2017, Oh et al., 2013, Ciampaglia et al., 2015, Kumar et al., 2017, Ruchansky et al., 2017]. While we only achieve limited success in operationalizing O1, further improvements could be potentially made in the future, if we can gather more information such as the network structure [Kumar et al., 2016, Qazvinian et al., 2011, Kim et al., 2017, Ciampaglia et al., 2015] or engagement of users while sharing the news [Kumar et al., 2016, Ruchansky et al., 2017, Kwon et al., 2017, Kim et al., 2017].

Summary

In summary, we make three primary contributions in this paper.

1. *Methodological:* We developed a new method for assessing users' truth perceptions (N=15,000) of content (e.g., news stories). Our test asks users to *rapidly* assess (*i.e.,* at the rate of a few seconds per story) how truthful or untruthful the claims in a news story are. We conducted our truth perception tests on-line and gathered truth perceptions of 100 Amazon Mechanical Turk (AMT) workers from the USA [AMT, The Mechanical Turk Blog 2011] for each story.

2. *Empirical:* Our exploratory analysis of users' truth perceptions yielded several interesting findings. For instance, (i) for many stories, the collective wisdom of the crowd (average truth rating) differs significantly from the actual truth of the story, *i.e.,* wisdom of crowds is inaccurate, (ii) across different stories, we find evidence for both false positive perception bias (*i.e.,* a gullible user perceiving the story to be more true than it is in reality) and false negative perception bias (*i.e.,* a cynical user perceiving a story to be more false than it is in reality), and (iii) users' political ideologies influence their truth perceptions for the most controversial stories (those stories with high variance in truth perception between users), it is frequently the result of users' political ideologies (*i.e.,* whether they support democrats vs. republicans) influencing their truth perceptions.

3. *Practical:* Our predictive analysis of users' perception biases reveals the limitations of current strategies for selecting a small set of news stories to fact check based on how many users report the story as fake. We provide a proof of concept simulation for how our truth perception test and classifier can be used to achieve the three goals stated above for prioritizing stories for fact checking. However, please note that design of mechanisms to signal the fact checked label to the users such that they are receptive to them is out of scope of this study.

CHAPTER 5

Concluding discussion

A growing number of people rely on social media platforms, such as Twitter and Facebook, for their news and information needs [Lichterman, 2010, Teevan et al., 2011], where users themselves play a role in selecting the sources from which they consume information, overthrowing traditional journalistic gatekeeping [Shoemaker et al., 2009]. Since users can just select their information sources, they don't have full control over the content they receive. Moreover, it is very hard to ascertain the quality, relevance, and credibility of information produced by social media users [Agichtein et al., 2008, Castillo et al., 2011, Farajtabar et al., 2015]. To tackle these concerns, in this thesis, we first address the question of how efficient users are at selecting their information sources.

We have defined three intuitive notions of user's efficiency in social media – link, in-flow and delay efficiency – to assess how *good* users are at selecting who to follow within the social media system to acquire information. Our framework is general and applicable to any social media system where every user *follows* others within the system to receive the information they produce. We have then leveraged our notions of efficiency to help us in understanding the relationship between different factors, such as the popularity of received information and the users' ego-networks structure.

Here, we have focused on three definitions of efficiency (link, in-flow, and delay). However, we could leverage this idea to define more complex notions of efficiency. For example, we could define efficiency in terms of diversity, *i.e.,* it would be interesting to find the set of users that, if followed, would cover the same unique memes while maximizing the diversity of topics or perspectives that are delivered with the memes, and then compare this set with the original set of followees in terms of diversity. This would provide a framework to mitigate the effects of the filtering bubble and echo chamber present in current social media systems. Moreover, some of the memes could be treated preferentially over other memes. This could be achieved by means of covering a list of non-unique memes favoring repetitions of a preferential subset of memes, *e.g.,* memes matching the user's interests should be delivered to the user more often. Remarkably, these more complex notions of efficiency can often be expressed as integer linear programs, similarly to the minimal set cover problem, which can be solved using relaxation methods with provable guarantees [Vazirani, 2001].

Additionally, we have introduced a heuristic method that improves both in-flow and delay efficiency of users, while still delivering them the same unique memes. Similar heuristics can be naturally designed to optimize efficiency with respect to multiple quantities (be it link, in-flow, delay, or diversity). In this context, it would be very interesting to design methods with provable guarantees to find sets of users that are optimal with respect to multiple quantities.

Our work also opens other interesting venues for future work. For example, we have defined and computed a measure of efficiency for each user independently. However, one could also think on global notions of efficiency for the Twitter information network as a whole, perhaps using a multi set cover approach. We have evaluated user's efficiency at acquiring four different types of memes. However, a systematic comparison of user's efficiency at acquiring many types of memes appears as a interesting research direction. Since we have applied our framework to study information efficiency only on Twitter, it would be interesting to study information efficiency of other microblogging services (Weibo, Pinterest, Tumblr) and social networking sites (Facebook, Google+). Finally, it would be worth to investigate how users' efficiency relates to their levels of activity and engagement within the online social media system.

Having characterized how efficient users are at selecting information sources, we then focus on how can we break the filter bubbles that users get trapped in. To minimize the possibility of social media users getting trapped in 'echo chambers' or 'filter bubbles', prior works have proposed to introduce diversity in the news that users are consuming [Munson et al., 2013a, Park et al., 2009b, Keegan, 2017]. Often, such approaches which highlight the most belief challenging news, increase the chances of users rejecting them, thereby defeating the original purpose [Munson and Resnick, 2010, Lord and Ross, 1979, Miller et al., 1993, Munro and Ditto, 1997]. In this thesis, we propose a complementary approach to inject diversity in users' news consumption by highlighting news posts which evoke similar reactions from different readers, irrespective of their own political leanings.

Towards that end, to our knowledge, we made the first attempt to define and operationalize consensus of news posts on social media. Subsequently, we compared several properties of high and low consensus news posts and found them to be equally popular, and covering similar topics. Additionally, we observed that high consensus posts lead to higher cross-cutting exposure for the users. Next, utilizing our proposed novel class of audience leaning based features, we developed a method to automatically infer the consensus of news posts on Twitter. Using our proposed consensus inference method, we publicly deployed "Purple Feed" – a system which highlights high consensus posts from different news outlets on Twitter. With "Purple Feed", the users can view the high consensus tweets posted by both Republican-leaning and Democrat-leaning media outlets during the last one week.[1] Users can also view both high and low consensus posts posted by individual publishers.[2]

We then studied the diffusion of news in Twitter. We investigated how users with various political leanings (liberals, conservatives and neutrals) get exposed to low and high consensus news posted by different publishers (e.g. CNN, FoxNews, etc.). We found that (1) while low consensus news stories are more likely to proliferate amongst the users with a particular political leaning, high consensus news has a much higher chance of spreading among users with different political leanings; (2) high consensus news posted by neutral publishers has the lowest disparity for spreading among liberal and conservative users; and (3) as users get farther away from the publishers, they get a more biased exposure to the news. Based on the above observations, we studied the effect of spreading high consensus news through neutral users on decreasing the disparity in users' exposure. Our extensive simulation experiments on Twitter showed that our proposed strategy can be highly effective in decreasing the disparity of information across users with differing views. Our findings may be helpful for breaking filter bubbles and reducing fragmentation in online social media.

In future, we plan to conduct a large scale characterization study of news posts and publishers on social media, and evaluate the impact of showing high consensus news posts on the users. We believe that our work on identifying high consensus news posts could be integrated with different information retrieval mechanisms on social media, and could be useful for designing mechanisms for mitigating filter bubble and echo chambers, for reducing fragmentation in news consumption, and for encouraging healthy debate on diverse issues on social media platforms.

[1] Available at http://twitter-app.mpi-sws.org/purple-feed/.

[2] For instance, high and low consensus tweets posted by New York Times can be viewed at: http://twitter-app.mpi-sws.org/purple-feed/app-tweet-1.php?query=NYTimes.

Finally we address the concern regarding how do the users perceive the truthfulness of information? We deeply examined how users perceive truth in news stories by developing novel and robust truth perception tests, where users are asked to rapidly assess how true or false the claims in a news story are. We validated our tests against deployment variations and common survey biases such as sample effects, answer choice effects, and satisficing and incentive effects. For our dataset of 150 news claims collected from Snopes.com, we performed our truth perception tests online on the AMT platform to collect users' perceptions of truth in news stories (N=15,000). Leveraging users' truth perceptions, we propose a novel framework for prioritizing stories for fact checking, with three potential, competing objectives: (i) removing false news stories from circulation, (ii) correcting the misperception of the users, and (iii) decreasing the disagreement between different users' perceptions of truth. Using a combination of user perceptions elicited using our truth perception tests, users' demographic features, and supervised machine learning methods we provide mechanisms for operationalization strategies that utilize users' truth perceptions to achieve the above objectives for prioritizing stories for fact checking.

Our findings can help inform the design of mechanisms for selecting stories to fact check, and can aid social media platform providers and fact checking organizations to combat fake news more efficiently.

BIBLIOGRAPHY

Newspaper crisis hits germany. **https://www.spiegel.de/international/germany/circulation-declines-hit-german-papers-a-decade-after-america-a-915574.html**, 2013.

Get better results with less effort with mechanical turk masters, http://mechanicalturk.typepad.com/blog/2011/06/get-betterresults-with-less-effort-with-mechanical-turk-masters-.html, The Mechanical Turk Blog 2011.

AAPOR. Research synthesis: Aapor report on online panels. *Public Opinion Quarterly*, 74(4):711–781, 2010.

L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

L. A. Adamic, B. A. Huberman, A. Barabási, R. Albert, H. Jeong, and G. Bianconi. Power-law distribution of the world wide web. *science*, 287(5461):2115–2115, 2000.

E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of the 1st International Conference on Web Search and Data Mining*, pages 183–194, 2008.

D. Ahlers. News consumption and the new electronic media. *Harvard International Journal of Press/Politics*, 11(1):29–52, 2006.

R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

R. Albert, H. Jeong, and A.-L. Barabási. Diameter of the world-wide web. *nature*, 401(6749):130–131, 1999.

J. Ali, M. Babaei, A. Chakraborty, B. Mirzasoleiman, K. P. Gummadi, and A. Singla. On the fairness of time-critical influence maximization in social networks. *ArXiv*, abs/1905.06618, 2019a.

J. Ali, M. Babaei, A. Chakraborty, B. Mirzasoleiman, K. P. Gummadi, and A. Singla. On the fairness of time-critical influence maximization in social networks. *arXiv preprint arXiv:1905.06618*, 2019b.

H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. Technical Report 2, National Bureau of Economic Research, 2017.

S. L. Althaus and D. Tewksbury. Patterns of internet and traditional news media use in a networked community. *Political communication*, 17(1):21–45, 2000.

L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley. Classes of small-world networks. *Proceedings of the national academy of sciences*, 97(21):11149–11152, 2000.

J. An, D. Quercia, M. Cha, K. Gummadi, and J. Crowcroft. Sharing political news: the balancing act of intimacy and socialization in selective exposure. *EPJ Data Science*, 3(1):12, 2014.

D. Antoniades and C. Dovrolis. Co-evolutionary dynamics in social networks: A case study of twitter. *arXiv preprint arXiv:1309.6001*, 2013.

M. Babaei, H. Ghassemieh, and M. Jalili. Cascading failure tolerance of modular small-world networks. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 58(8):527–531, 2011.

M. Babaei, B. Mirzasoleiman, M. Jalili, and M. A. Safari. Revenue maximization in social networks through discounting. *Social Network Analysis and Mining*, 3(4):1249–1262, 2013.

M. Babaei, P. Grabowicz, I. Valera, and M. Gomez-Rodriguez. On the users' efficiency in the twitter information network. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

M. Babaei, P. Grabowicz, I. Valera, K. P. Gummadi, and M. Gomez-Rodriguez. On the efficiency of the information networks in social media. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 83–92. ACM, 2016.

M. Babaei, J. Kulshrestha, A. Chakraborty, F. Benevenuto, K. P. Gummadi, and A. Weller. Purple feed: Identifying high consensus news posts on social media. 2018.

M. Babaei, A. Chakraborty, J. Kulshrestha, E. M. Redmiles, M. Cha, and K. P. Gummadi. Analyzing biases in perception of truth in news stories and their implications for fact checking. In *FAT*, page 139, 2019.

M. Babaei, B. Mirzasoleiman, J. Joo, and A. Weller. Promoting high consensus news selectively to reach a diverse audience. 2021.

C. A. Bail. *Terrified: How anti-Muslim fringe organizations became mainstream*. Princeton University Press, 2014.

C. A. Bail, L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.

E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528, 2012.

E. Bakshy, S. Messing, and L. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 2015. ISSN 0036-8075. doi: 10.1126/science.aaa1160. URL **http://science.sciencemag.org/content/early/2015/05/06/science.aaa1160**.

A. Barabási, H. Jeong, Z. Néda, R. Ravasz, A. Schubert, and T. Vicsek. Adamic, la (1999). the small world web. in proceedings of the third euro-pean conference on research and advanced technology for digital libraries (pp. 443–452). springer-verlag. albert, r., & barabasi, al (2000). topology of evolving networks: local events and universality. phys rev lett, 85, 5234–5237. *reviews of modern physics*, 74: 47–97, 2002.

A.-L. Barabási. Linked: The new science of networks, 2003.

A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

A.-L. Barabási, R. Albert, and H. Jeong. Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A: statistical mechanics and its applications*, 281(1-4):69–77, 2000.

T. Barrabi. Facebook, twitter face congressional hearings on political bias, fake news. **https://www.foxbusiness.com/technology/facebook-twitter-face-congressional-hearings-on-political-bias-fake-news**, 2018.

S. Barthelemy, M. Bethell, T. Christiansen, A. Jarsvall, and K. Koinis. The future of print media. *Retrieved Jan*, 4:2015, 2011.

S. Bharathi, D. Kempe, and M. Salek. Competitive influence maximization in social networks. In *International workshop on web and internet economics*, pages 306–311. Springer, 2007.

G. Bhatt, A. Sharma, S. Sharma, A. Nagpal, B. Raman, and A. Mittal. On the Benefit of Combining Neural, Statistical and External Features for Fake News Identification. *arXiv preprint arXiv:1712.03935*, 2017.

D. Bhattacharya and S. Ram. Sharing news articles using 140 characters: A diffusion analysis on twitter. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 966–971. IEEE, 2012.

T. Bian, X. Xiao, T. Xu, P. Zhao, W. Huang, Y. Rong, and J. Huang. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 549–556, 2020.

P. Bourgonje, J. M. Schneider, and G. Rehm. From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles. In *Proceedings of the EMNLP Workshop on Natural Language Processing meets Journalism*, 2017.

A. W. Bowman and A. Azzalini. *Applied smoothing techniques for data analysis*. Clarendon Press, 2004.

E. Bozdag. Bias in algorithmic filtering and personalization. *Ethics and Information Technology*, 15(3):209–227, Sept. 2013. ISSN 1388-1957. doi: 10.1007/s10676-013-9321-6. URL **http://dx.doi.org/10.1007/s10676-013-9321-6**.

A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer networks*, 33(1-6):309–320, 2000.

C. Budak, D. Agrawal, and A. El Abbadi. Limiting the spread of misinformation in social networks. In *WWW*, 2011.

C. Budak, S. Goel, and J. M. Rao. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1):250–271, 2016. doi: 10.1093/poq/nfw007. URL **+http://dx.doi.org/10.1093/poq/nfw007**.

C. Cai, L. Li, and D. Zeng. Detecting social bots by jointly modeling deep behavior and content information. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1995–1998, 2017.

T. Carnes, C. Nagarajan, S. M. Wild, and A. Van Zuylen. Maximizing influence in a competitive social network: a follower's perspective. In *EC*, pages 351–360. ACM, 2007.

S. Castelo, T. Almeida, A. Elghafari, A. Santos, K. Pham, E. Nakamura, and J. Freire. A topic-agnostic approach for identifying fake news pages. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 975–980, 2019.

C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, pages 675–684, 2011.

M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 10–17, 2010.

M. Cha, F. Benevenuto, H. Haddadi, and K. P. Gummadi. The world of connections and information flow in twitter. *IEEE Trans. Systems, Man, and Cybernetics, Part A*, 42:991–998, 2012.

A. Chakraborty, S. Ghosh, N. Ganguly, and K. P. Gummadi. Dissemination biases of social media channels: On the topical coverage of socially shared news. In *ICWSM*, pages 559–562, 2016a.

A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly. Stop clickbait: Detecting and preventing clickbaits in online news media. In *Proceedings of the ASONAM*, 2016b.

A. Chakraborty, M. Ali, S. Ghosh, N. Ganguly, and K. P. Gummadi. On quantifying knowledge segregation in society. *arXiv preprint arXiv:1708.00670*, 2017.

A. Chakraborty, M. Luqman, S. Satapathy, and N. Ganguly. Editorial algorithms: Optimizing recency, relevance and diversity for automated news curation. In *The 2018 Web Conference Companion*, WWW'18. ACM, April 2018.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of KDD*, 2016.

Y. Chen, N. J. Conroy, and V. L. Rubin. Misleading online content: Recognizing clickbait as false news. In *Proceedings of the ACM Workshop on Multimodal Deception Detection*, 2015.

C.-F. Chiang and B. Knight. Media bias and influence: Evidence from newspaper endorsements. *The Review of Economic Studies*, 78(3):795–820, 2011.

K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

S. Chopra, S. Jain, and J. M. Sholar. Towards Automatic Identification of Fake News: Headline-Article Stance Detection with LSTM Attention Models, 2017.

G. Chowell, J. Hyman, S. Eubank, and C. Castillo-Chavez. Analysis of a real world network: The city of portland. Technical report, Technical Report BU-1604-M, Department of Biological Statistics, 2002.

N. A. Christakis and J. H. Fowler. Social network sensors for early detection of contagious outbreaks. *PloS one*, 5(9):e12948, 2010.

G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini. Computational fact checking from knowledge networks. *PloS one*, 10(6), 2015.

M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political polarization on twitter. In *In Proceedings of AAAI ICWSM*, 2011.

T. J. A. Covert and P. C. Wasburn. Measuring media bias: A content analysis of time and newsweek coverage of domestic social issues, 1975–2000. *Social science quarterly*, 88(3), 2007.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

S. N. Dorogovtsev and J. F. Mendes. Evolution of networks. *Advances in physics*, 51(4):1079–1187, 2002.

S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. Pseudofractal scale-free web. *Physical review E*, 65(6):066122, 2002.

Y. Duan, L. Jiang, T. Qin, M. Zhou, and H.-Y. Shum. An empirical study on learning to rank of tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 295–303, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL **http://dl.acm.org/citation.cfm?id=1873781.1873815**.

M. J. Dutta-Bergman. Complementarity in consumption of news types across traditional and new media. *Journal of broadcasting & electronic media*, 48(1):41–60, 2004.

D. Easley, J. Kleinberg, et al. *Networks, crowds, and markets*, volume 8. Cambridge university press Cambridge, 2010.

C. Eilders. News factors and news decisions. theoretical and methodological advances in germany. *Communications*, 31(1):5–24, 2006.

M. Farajtabar, M. Gomez-Rodriguez, N. Du, M. Zamani, H. Zha, and L. Song. Back to the past: Source identification in diffusion networks from partially observed cascades. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, 2015.

T. J. Fararo and M. Sunshine. A study of a biased friendship network, 1964.

S. Feng, R. Banerjee, and Y. Choi. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 171–175, 2012.

N. Ferguson. The false prophecy of hyperconnection: How to survive the networked age. *Foreign Aff.*, 96:68, 2017.

G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee. Self-organization and identification of web communities. *Computer*, 35(3):66–70, 2002.

S. Flaxman, S. Goel, and J. M. Rao. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80(S1):298–320, 2016. doi: 10.1093/poq/nfw006. URL **+http://dx.doi.org/10.1093/poq/nfw006**.

A. Friggeri, L. A. Adamic, D. Eckles, and J. Cheng. Rumor cascades. In *ICWSM*, 2014.

A. Fronczak, P. Fronczak, and J. Holyst. Exact solution for average path length in random graphs. Technical report, 2002.

J. Galaskiewicz. *Social organization of an urban grants economy: A study of business philanthropy and nonprofit organizations*. Elsevier, 2016.

J. Galaskiewicz and P. V. Marsden. Interorganizational resource networks: Formal patterns of overlap. *Social science research*, 7(2):89–107, 1978.

J. Galtung and M. H. Ruge. The structure of foreign news: The presentation of the congo, cuba and cyprus crises in four norwegian newspapers. *Journal of peace research*, 2(1):64–90, 1965.

M. Gentzkow. Valuing new goods in a model with complementarity: Online newspapers. *American Economic Review*, 97(3):713–744, 2007.

M. Gentzkow and J. Shapiro. What drives media slant? evidence from u.s. daily newspapers. *Econometrica*, 78(1):35–71, 2010. URL **https://EconPapers.repec.org/RePEc:ecm:emetrp:v:78:y:2010:i:1:p:35-71**.

J. Golbeck and D. Hansen. Computing political preference among twitter followers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1105–1108, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0228-9. doi: 10.1145/1978942.1979106. URL **http://doi.acm.org/10.1145/1978942.1979106**.

J. Golbeck, M. Mauriello, B. Auxier, K. H. Bhanushali, C. Bonk, M. A. Bouzaghrane, C. Buntain, R. Chanduka, P. Cheakalos, J. B. Everett, et al. Fake news vs satire: A dataset and analysis. In *Proceedings of the 10th ACM Conference on Web Science*, pages 17–21, 2018.

M. Gomez-Rodriguez, K. Gummadi, and B. Schoelkopf. Quantifying Information Overload in Social Media and its Impact on Social Contagions. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, pages 170–179, 2014.

A. Goyal, F. Bonchi, L. V. Lakshmanan, and S. Venkatasubramanian. On minimizing budget and time in influence propagation over social networks. *Social network analysis and mining*, 3(2):179–192, 2013.

P. Grabowicz, M. Babaei, J. Kulshrestha, and I. Weber. The road to popularity: The dilution of growing audience on twitter. In *International AAAI Conference on Web and Social Media*, ICWSM '16, May 2016. URL **https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13134/12790**.

M. S. Granovetter. The strength of weak ties. *American journal of sociology*, pages 1360–1380, 1973.

L. Graves. Understanding the promise and limits of automated fact-checking. Technical report, 2018.

N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer. Fake news on twitter during the 2016 u.s. presidential election. *Science*, 363(6425):374–378, 2019. ISSN 0036-8075. doi: 10.1126/science.aau2706. URL **https://science.sciencemag.org/content/363/6425/374**.

T. Groseclose and J. Milyo. A measure of media bias. *The Quarterly Journal of Economics*, 120(4): 1191–1237, 2005. ISSN 00335533, 15314650. URL **http://www.jstor.org/stable/25098770**.

A. Guess, J. Nagler, and J. Tucker. Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science Advances*, 5(1), 2019. doi: 10.1126/sciadv.aau4586. URL **https://advances.sciencemag.org/content/5/1/eaau4586**.

J. Hartline, V. Mirrokni, and M. Sundararajan. Optimal marketing strategies over social networks. In *Proceedings of the 17th international conference on World Wide Web*, pages 189–198, 2008.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 04 1970. doi: 10.1093/biomet/57.1.97. URL **https://doi.org/10.1093/biomet/57.1.97**.

I. Himelboim, S. McCreery, and M. Smith. Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on twitter. *Journal of Computer-Mediated Communication*, 18(2):40–60, 2013.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

N. Hodas and K. Lerman. How visibility and divided attention constrain social contagion. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing*, pages 249–257, 2012.

L. Hogan and M. Safi. Revealed: Facebook hate speech exploded in myanmar during rohingya crisis. **https://www.theguardian.com/world/2018/apr/03/revealed-facebook-hate-speech-exploded-in-myanmar-during-rohingya-crisis**, 2018.

M. Hu, S. Liu, F. Wei, Y. Wu, J. Stasko, and K.-L. Ma. Breaking news on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2751–2754, 2012.

B. A. Huberman. The laws of the web, 2001.

T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences*, 98(8):4569–4574, 2001.

H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

C. Jin, P. Netrapalli, and M. I. Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Conference On Learning Theory*, pages 1042–1085. PMLR, 2018.

S. John. Social network analysis: A handbook. *Contemporary Sociology*, 22(1):128, 2000.

D. S. Johnson. Approximation algorithms for combinatorial problems. In *Proceedings of the fifth annual ACM symposium on Theory of computing*, pages 38–49. ACM, 1973.

S. N. Jomini. Polarization and partisan selective exposure. *Journal of Communication*, 60(3):556–576, 2010.

J. H. Jones and M. S. Handcock. An assessment of preferential attachment as a mechanism for human sexual network formation. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1520):1123–1128, 2003.

P. R. C. Journalism and Media. State of the News Media 2014: News Video on the Web. **https://www.journalism.org/files/2014/03/News-Video-on-the-Web.pdf**, 7 2014.

P. R. C. Journalism and Media. Americans' online news use is closing in on TV news use. **http://www.pewresearch.org/fact-tank/2017/09/07/americans-online-news-use-vs-tv-news-use/**, 7 2017.

P. R. C. Journalism and Media. Newspapers Fact Sheet. **http://www.journalism.org/fact-sheet/newspapers/**, 6 2018.

P. R. C. Journalism and Media. Americans' online news use is closing in on TV news use. **https://www.journalism.org/2020/07/30/americans-who-mainly-get-their-news-on-social-media-are-less-engaged-less-knowledgeable/**, 7 2020.

V. Kalapala, V. Sanwalani, and C. Moore. The structure of the united states road network. *Preprint, University of New Mexico*, 2003.

R. M. Karp. *Reducibility among combinatorial problems*. Springer, 1972.

B. K. Kaye and T. J. Johnson. From here to obscurity?: Media substitution theory and traditional media in an on-line world. *Journal of the American Society for Information Science and Technology*, 54(3): 260–273, 2003.

M. Kearns. Experiments in social computation. *Communications of the ACM*, 55(10):56–67, 2012.

J. Keegan. Blue feed, red feed - see liberal facebook and conservative facebook, side by side. http://graphics.wsj.com/blue-feed-red-feed/, 2017.

D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD*, 2003.

M. Khajehnejad, A. Asgharian Rezaei, M. Babaei, J. Hoffmann, M. Jalili, and A. Weller. Adversarial graph embeddings for fair influence maximization over social networks. In C. Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4306–4312. International Joint Conferences on Artificial Intelligence Organization, 7 2020a. Special track on AI for CompSust and Human well-being.

M. Khajehnejad, A. A. Rezaei, M. Babaei, J. Hoffmann, M. Jalili, and A. Weller. Adversarial graph embeddings for fair influence maximization over social networks. *arXiv preprint arXiv:2005.04074*, 2020b.

J. Kim, B. Tabibian, A. Oh, B. Schölkopf, and M. Gomez-Rodriguez. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. *arXiv preprint arXiv:1711.09918*, 2017.

Y. Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins. The web as a graph: measurements, models, and methods. In *International Computing and Combinatorics Conference*, pages 1–17. Springer, 1999.

A. S. Klovdahl, J. J. Potterat, D. E. Woodhouse, J. B. Muth, S. Q. Muth, and W. W. Darrow. Social networks and infectious disease: The colorado springs study. *Social science & medicine*, 38(1): 79–88, 1994.

P. N. Krivitsky, M. S. Handcock, A. E. Raftery, and P. D. Hoff. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social networks*, 31(3):204–213, 2009.

J. A. Krosnick and L. R. Fabrigar. Designing rating scales for effective measurement in surveys. *Survey measurement and process quality*, pages 141–164, 1997.

J. A. Krosnick, S. Narayan, and W. R. Smith. Satisficing in surveys: Initial evidence. *New directions for evaluation*, 1996(70):29–44, 1996.

J. Kulshrestha, M. Eslami, J. Messias, M. B. Zafar, S. Ghosh, K. P. Gummadi, and K. Karahalios. Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 417–432, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4335-0. doi: 10.1145/2998181.2998321. URL **http://doi.acm.org/10.1145/2998181.2998321**.

R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 57–65. IEEE, 2000.

S. Kumar, R. West, and J. Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of WWW*, 2016.

S. Kumar, R. West, and J. Leskovec. Toward Automated Fact-Checking: Detecting Check-Worthy Factual Claims by ClaimBuster. In *Proceedings of ACM SIGKDD*, 2017.

H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web (WWW '10)*, 2010.

S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang. Aspects of rumor spreading on a microblog network. In *Proceedings of the SocInfo*, 2013a.

S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang. Prominent features of rumor propagation in online social media. In *Proceedings of the ICDM*, 2013b.

S. Kwon, M. Cha, and K. Jung. Rumor detection over varying time windows. *PloS one*, 12(1):e0168344, 2017.

V. Latora and M. Marchiori. Is the boston subway a small-world network? *Physica A: Statistical Mechanics and its Applications*, 314(1-4):109–113, 2002.

K. Lerman and T. Hogg. Leveraging position bias to improve peer recommendation. *PLoS One*, 9(6), 2014.

J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506, 2009.

J. Lichterman. New Pew data: More Americans are getting news on Facebook and Twitter. **http://www.niemanlab.org/2015/07/new-pew-data-more-americans-are-getting-news-on-facebook-and-twitter/**, 2010.

F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Åberg. The web of human sexual contacts. *Nature*, 411(6840):907–908, 2001.

Z. Liu and I. Weber. Is twitter a public sphere for online conflicts? a cross-ideological and cross-hierarchical look. In *International Conference on Social Informatics*, pages 336–347. Springer, 2014.

C. G. Lord and L. Ross. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, pages 2098–2109, 1979.

C. G. Lord, L. Ross, and M. R. Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, 37 (11):2098, 1979.

T. Lyons. Hard questions: What's facebook's strategy for stopping false news?, facebook newsroom. **https://newsroom.fb.com/news/2018/05/hard-questions-false-news**, 2018.

J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha. Detecting Rumors from Microblogs with Recurrent Neural Networks. In *Proceedings of the IJCAI*, 2016.

J. Ma, W. Gao, and K.-F. Wong. Rumor detection on twitter with tree-structured recursive neural networks. Association for Computational Linguistics, 2018.

J. Maddock, K. Starbird, H. J. Al-Hassani, D. E. Sandoval, M. Orand, and R. M. Mason. Characterizing online rumoring behavior using multi-dimensional signatures. In *Proceedings of the CSCW*, 2015.

P. Mariolis. Interlocking directorates and control of corporations: The theory of bank control. *Social Science Quarterly*, pages 425–439, 1975.

S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296 (5569):910–913, 2002.

M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

S. Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

A. G. Miller, J. W. McHoskey, C. M. Bane, and T. G. Dowd. The attitude polarization phenomenon: Role of response measure, attitude extremity, and behavioral consequences of reported attitude change. *Journal of Personality and Social Psychology*, 64(4):561, 1993.

B. Mirzasoleiman, M. Babaei, M. Jalili, and M. Safari. Cascaded failures in weighted networks. *Physical Review E*, 84(4):046114, 2011.

B. Mirzasoleiman, M. Babaei, and M. Jalili. Immunizing complex networks with limited budget. *EPL (Europhysics Letters)*, 98(3):38004, 2012.

A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42, 2007.

A. Mitchell, J. Gottfried, J. Kiley, and K. Matsa. Political polarization and media habits. *Pew Research Center*, 21, 2014.

F. Morstatter, L. Wu, T. H. Nazer, K. M. Carley, and H. Liu. A new approach to bot detection: striking the balance between precision and recall. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 533–540. IEEE, 2016.

G. D. Munro and P. H. Ditto. Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information. *Personality and Social Psychology Bulletin*, 23(6):636–653, 1997. doi: 10.1177/0146167297236007. URL **https://doi.org/10.1177/0146167297236007**.

S. Munson, S. Chhabra, and P. Resnick. BALANCE - Tools for improving your news reading experience. *http://balancestudy.org/*.

S. Munson, S. Lee, and P. Resnick. Encouraging reading of diverse political viewpoints with a browser widget. In *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013*, pages 419–428, Boston, USA, 2013a. AAAI press.

S. A. Munson and P. Resnick. Presenting diverse political opinions: How and how much. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1457–1466, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-929-9. doi: 10.1145/1753326.1753543. URL **http://doi.acm.org/10.1145/1753326.1753543**.

S. A. Munson, S. Y. Lee, and P. Resnick. Encouraging reading of diverse political viewpoints with a browser widget. In *ICWSM*, 2013b.

P. C. Murschetz and M. Friedrichsen. Does online video save printed newspapers? online video as convergence strategy in regional printed news publishing: The case of germany. In *Digital transformation in journalism and news media*, pages 115–128. Springer, 2017.

S. A. Myers and J. Leskovec. The bursty dynamics of the twitter information network. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 913–924, 2014.

N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi. Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proceedings of the 3rd International Web Science Conference*, WebSci '11, pages 8:1–8:7. ACM, 2011.

J. Newell, J. J. Pilotta, and J. C. Thomas. Mass media displacement and saturation. *The International Journal on Media Management*, 10(4):131–138, 2008.

M. E. Newman. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences*, 98(2):404–409, 2001.

M. E. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

M. E. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical review E*, 64(2):026118, 2001.

B. Nyhan and J. Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, 2010.

O. Oh, M. Agrawal, and H. R. Rao. Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *Mis Quarterly*, 37(2), 2013.

S. Ovide. A tiktok twist on ‘pizzagate’, nytimes. **https://www.nytimes.com/2020/06/29/technology/pizzagate-tiktok.html**, 2020.

E. Pariser. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Group , The, 2011. ISBN 1594203008, 9781594203008.

S. Park, S. Kang, S. Chung, and J. Song. Newscube: Delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, 2009a.

S. Park, S. Kang, S. Chung, and J. Song. NewsCube: delivering multiple aspects of news to mitigate media bias. In *In Proceedings of ACM CHI*, 2009b.

R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.

E. Peer, J. Vosgerau, and A. Acquisti. Reputation as a sufficient condition for data quality on amazon mechanical turk. *Behavior research methods*, 46(4):1023–1031, 2014.

G. Pennycook and D. G. Rand. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7):2521–2526, 2019.

V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*, 2017.

N. Persily. *Solutions to political polarization in America*. Cambridge University Press, 2015.

S. Petrovic, M. Osborne, and V. Lavrenko. RT to win! predicting message propagation in twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, ICWSM'11. AAAI, May 2011.

R. G. Picard. The future of the news industry. *Media and society*, 5:365–379, 2010.

D. J. D. S. Price. Networks of scientific papers. *Science*, pages 510–515, 1965.

V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the EMNLP*, 2011.

A. Rapoport and W. J. Horvath. A study of a large sociogram. *Behavioral science*, 6(4):279–291, 1961.

E. M. Redmiles, Y. Acar, S. Fahl, and M. L. Mazurek. A summary of survey methodology best practices for security and privacy researchers. Technical report, 2017.

S. Redner. How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 4(2):131–134, 1998.

F. N. Ribeiro, F. B. Lucas Henrique, A. Chakraborty, J. Kulshrestha, M. Babei, and K. P. Gummadi. Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *Proceedings of the 12th International AAAI Conference of Web and Social Media*, ICWSM '18, May 2015.

F. N. Ribeiro, K. Saha, M. Babaei, L. Henrique, J. Messias, F. Benevenuto, O. Goga, K. P. Gummadi, and E. M. Redmiles. On microtargeting socially divisive ads: A case study of russia-linked ad campaigns on facebook. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 140–149, 2019.

M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *KDD*, 2002.

M.-A. Rizoiu, T. Graham, R. Zhang, Y. Zhang, R. Ackland, and L. Xie. # debatenight: The role and influence of socialbots on twitter during the 1st 2016 us presidential debate. *arXiv preprint arXiv:1802.09808*, 2018.

D. M. Romero and J. Kleinberg. The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, 2010a.

D. M. Romero and J. Kleinberg. The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, 2010b.

D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, pages 695–704, 2011.

M. M. U. Rony, N. Hassan, and M. Yousuf. Diving Deep into Clickbaits: Who Use Them to What Extents in Which Topics with What Effects? *arXiv preprint arXiv:1703.09400*, 2017.

N. Ruchansky, S. Seo, and Y. Liu. CSI: A Hybrid Deep Model for Fake News Detection. In *Proceedings of the CIKM*, 2017.

D. Schkade, C. R. Sunstein, and R. Hastie. What happened on deliberation day? *California Law Review*, 95(3):915–940, 2007.

M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

D. Sears and J. L. FREEDMAN. Selective exposure to information: A critical review. 31, 06 1967.

P. O. Seglen. The skewness of science. *Journal of the American society for information science*, 43(9): 628–638, 1992.

P. Sen, S. Dasgupta, A. Chatterjee, P. Sreeram, G. Mukherjee, and S. Manna. Small-world properties of the indian railway network. *Physical Review E*, 67(3):036106, 2003.

C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer. The spread of low-credibility content by social bots. *Nature communications*, 9(1):1–9, 2018.

K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, and Y. Liu. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(3):1–42, 2019.

P. J. Shoemaker and T. Vos. *Gatekeeping theory*. Routledge, 2009.

P. J. Shoemaker, T. P. Vos, and S. D. Reese. Journalists as gatekeepers. *The handbook of journalism studies*, 73, 2009.

K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 395–405, 2019.

K. Shu, D. Mahudeswaran, S. Wang, and H. Liu. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 626–637, 2020.

G. Simmel. *The Sociology of Georg Simmel*. Free Press of Glencoe, 1950.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

N. Sitaula, C. K. Mohan, J. Grygiel, X. Zhou, and R. Zafarani. Credibility-based fake news detection. In *Disinformation, Misinformation, and Fake News in Social Media*, pages 163–182. Springer, 2020.

R. V. Solé and R. Pastor-Satorras. Complex networks in genomics and proteomics. *Handbook of Graphs and Networks*, pages 145–167, 2002.

K. Starbird, A. Arif, T. Wilson, K. Van Koevering, K. Yefimova, and D. Scarnecchia. Ecosystem or echo-system? exploring content sharing across alternative media domains. In *ICWSM*, pages 365–374, 2018.

S. H. Strogatz. Exploring complex networks. *nature*, 410(6825):268–276, 2001.

N. Stroud. *Niche News: The Politics of News Choice*. Oxford University Press., 2011.

B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, SOCIALCOM '10, pages 177–184. IEEE Computer Society, 2010.

C. R. Sunstein. The law of group polarization. *Journal of Political Philosophy*, 10(2):175–195, 2002. ISSN 1467-9760. doi: 10.1111/1467-9760.00148. URL **http://dx.doi.org/10.1111/1467-9760.00148**.

G. Szabó, M. Alava, and J. Kertész. Structural transitions in scale-free networks. *Physical Review E*, 67(5):056102, 2003.

C. S. Taber and M. Lodge. Motivated skepticism in the evaluation of political beliefs. *American journal of political science*, 50(3):755–769, 2006.

K. Tao, F. Abel, C. Hauff, and G.-J. Houben. What makes a tweet relevant for a topic? In *#MSM-Workshop on Making Sense of Microposts*, volume 838 of *CEUR Workshop Proceedings*, pages 49–56, 2012.

A. Taub and M. Fisher. Where countries are tinderboxes and facebook is a match. **https://www.nytimes.com/2018/04/21/world/asia/facebook-sri-lanka-riots.html**, 2018.

J. Teevan, D. Ramage, and M. R. Morris. #twittersearch: A comparison of microblog search and web search. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 35–44, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0493-1. doi: 10.1145/1935826.1935842. URL **http://doi.acm.org/10.1145/1935826.1935842**.

J. Travers and S. Milgram. An experimental study of the small world problem. In *Social Networks*, pages 179–197. Elsevier, 1977.

A. Tsang, B. Wilder, E. Rice, M. Tambe, and Y. Zick. Group-Fairness in Influence Maximization. *arXiv preprint arXiv:1903.00967*, 2019.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

V. V. Vazirani. *Approximation algorithms*. Springer, 2001.

M. D. Vicario, W. Quattrociocchi, A. Scala, and F. Zollo. Polarization and fake news: Early warning of potential misinformation targets. *ACM Trans. Web*, 13(2), Mar. 2019. ISSN 1559-1131. doi: 10.1145/3316809. URL **https://doi.org/10.1145/3316809**.

S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380): 1146–1151, 2018.

C. Wardle. Fake news. it’s complicated. *First Draft*, 16, 2017.

S. Wasserman and K. Faust. Social network analysis cambridge university press, 1994.

D. Watts. Small worlds princeton university press, princeton, 1999; ds callaway, mej newman, sh strogatz, dj watts. *Phys. Rev. Lett*, 85:5468, 2000.

D. J. Watts and S. H. Strogatz. Collective dynamics of âĂŸsmall-world’ networks. *nature*, 393(6684): 440–442, 1998.

P. Weber. The virtual get-together. determinants interpersonal - "o public communication on news websites. *Social Media and Web Science. Frankfurt am Main: DGI*, pages 457–459, 2012.

P. Weber. Discussions in the comments section: Factors influencing participation and interactivity in online newspapers’ reader comments. *New media & society*, 16(6):941–957, 2014.

L. Weng, J. Ratkiewicz, N. Perra, B. Gonçalves, C. Castillo, F. Bonchi, R. Schifanella, F. Menczer, and A. Flammini. The role of information diffusion in the evolution of social networks. In *Proceedings* *of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 356–364, 2013.

L. A. Wolsey. An analysis of the greedy algorithm for the submodular set covering problem. *Combinatorica*, 2(4):385–393, 1982.

T. Wood and E. Porter. The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, 41(1):135–163, 2019.

K. Wu, S. Yang, and K. 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.

S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who says what to whom on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 705–714, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0632-4. doi: 10.1145/1963405.1963504. URL **http://doi.acm.org/10.1145/1963405.1963504**.

J. Zhang, L. Cui, Y. Fu, and F. B. Gouza. Fake news detection with deep diffusive network model. *arXiv preprint arXiv:1805.08751*, 2018.

Z. Zhao, P. Resnick, and Q. Mei. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the WWW*, 2015.

D. X. Zhou, P. Resnick, and Q. Mei. Classifying the political leaning of news articles and users from user votes. In *In Proceedings of AAAI ICWSM*, 2011.

X. Zhou and R. Zafarani. Network-based fake news detection: A pattern-driven approach. *ACM SIGKDD Explorations Newsletter*, 21(2):48–60, 2019.

Y. Zhou, J. Lu, Y. Zhou, and Y. Liu. Recent advances for dyes removal using novel adsorbents: a review. *Environmental pollution*, 252:352–365, 2019.